UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA

GIANCARLO LIMA TORRES

UMA ABORDAGEM DE CIÊNCIA DE DADOS EM UMA ANÁLISE SOCIOECONÔMICA DE PREÇOS PARA VIAGENS DE TRANSPORTE POR APLICATIVO UBER

Maceió – AL
2022

GIANCARLO LIMA TORRES

UMA ABORDAGEM DE CIÊNCIA DE DADOS EM UMA ANÁLISE SOCIOECONÔMICA DE PREÇOS PARA VIAGENS DE TRANSPORTE POR APLICATIVO UBER

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas, como requisito para obtenção do grau de Mestre em Informática.

Orientador:
Professor Dr. Bruno Almeida Pimentel

Maceió – AL
2022

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

T693a Torres, Giancarlo Lima.
Uma abordagem de ciência de dados em uma análise socioeconômica de preços para viagens de transporte por aplicativo Uber / Giancarlo Lima Torres. – 2022.
119 f. : il. color.

Orientador: Bruno Almeida Pimentel.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2022.

Bibliografia: f. 110-117.
Apêndice: f. 119.

1. UBER (Aplicatico de transporte). 2. Ciência de dados. 3. Dados socioeconômicos. I. Título.

CDU: 004

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação/UFAL
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

GIANCARLO LIMA TORRES

UMA ABORDAGEM DE CIÊNCIA DE DADOS EM UMA ANÁLISE SOCIOECONÔMICA DE PREÇOS PARA VIAGENS DE TRANSPORTE POR APLICATIVO UBER

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 17 de agosto de 2022.

Banca Examinadora:

Prof. Dr. BRUNO ALMEIDA PIMENTEL
UFAL – Instituto de Computação
Orientador

Prof. Dr. EVANDRO DE BARROS COSTA
UFAL – Instituto de Computação
Examinador Interno

Documento assinado digitalmente
gov.br RAFAEL DE AMORIM SILVA
Data: 23/08/2022 16:26:40-0300
Verifique em https://verificador.iti.br

Prof. Dr. RAFAEL DE AMORIM SILVA
UFAL – Instituto de Computação
Examinador Interno

Prof. Dr. DIEGO CARVALHO DO NASCIMENTO
Universidade do Atacama - CHILE
Departamento de Matemática
Examinador Externo

“ Pois é Deus quem efetua em vocês tanto o querer quanto o realizar, de acordo com a boa vontade Dele. ”

Bíblia Sagrada. Filipenses 2:13.

AGRADECIMENTOS

Agradeço a Deus, criador de tudo que existe, por toda sabedoria e persistência que me concedeu durante esta pesquisa. Durante essa jornada, descobri que devemos sempre tentar, mesmo se as circunstâncias parecerem impossíveis, pois para Deus nada é impossível. É passando por dificuldades que nos aperfeiçoamos e, por isso, Deus muitas vezes permite que passemos por situações desfavoráveis.

Creio que a Ciência é uma dádiva concedida por Deus para que possamos evoluir como seres humanos. Nesse sentido, devemos compartilhá-la e aperfeiçoá-la sempre que possível. O trabalho em equipe é de suma importância para o crescimento da humanidade. Assim, quero agradecer ao meu orientador Professor Bruno que teve muita paciência e sabedoria ao me conduzir pelos caminhos da Ciência de Dados, fazendo com que enxergasse perspectivas ainda não conhecidas por mim.

Gratidão aos colegas que conheci durante o programa de mestrado da instituição, pois me ajudaram em momentos difíceis em algumas disciplinas, além de poder contribuir com eles para que conseguissem obter sucesso.

Agradecimento aos meus amigos que volta e meia me ouviam falar sobre as dificuldades que enfrentava durante a pesquisa e me incentivavam a prosseguir.

Também não poderia deixar de agradecer aos meus colegas de trabalho que sempre me apoiaram durante esse período de aulas e pesquisa em que executaram algumas atividades minhas para que eu pudesse continuar estudando.

Por fim, não poderia deixar de registrar os agradecimentos a minha amada esposa que sempre depositou confiança em mim, incentivando a prosseguir mesmo estando cansado das atividades do trabalho e da própria pesquisa.

A Deus toda honra e glória!

RESUMO

Estudos que utilizam dados da empresa de transporte por aplicativo Uber evidenciaram que há fatores que contribuem para o aumento de preços dos seus serviços de viagens. Nesse contexto, esta pesquisa teve como objetivo analisar rotas de viagens de usuários de baixa renda e contribuir na redução desses preços. Para isso, buscou-se responder: Se um centro financeiro estivesse mais próximo de bairros economicamente mais pobres, haveria mudança nos preços médios dessas viagens? Essa mudança poderia melhorar financeiramente a vida das pessoas de baixa renda? A proposta de nossa pesquisa para responder a esses questionamentos foi a de averiguar em regiões territoriais essas concentrações financeiras por meio de um processo de Ciência de Dados, analisando preços e dados socioeconômicos da cidade sul-americana de Fortaleza, localizada no país Brasil e da cidade norte americana de Boston, localizada no país Estados Unidos da América. Assim, seria possível evidenciar se os usuários da Uber que moram em bairros mais pobres financeiramente e utilizam esse serviço de viagens acabam pagando mais caro do que os usuários dos bairros mais ricos, quando o destino é o centro financeiro. As análises e os resultados obtidos para Boston serviram de validação por analogia para os resultados obtidos para Fortaleza. A base de dados analisada para Boston se refere a um conjunto de dados real, disponível na comunidade *online Kaggle*. A base de dados analisada para Fortaleza foi construída durante nosso trabalho e também está disponível na comunidade *online Kaggle*, podendo servir de ferramenta para analises futuras em outras pesquisas. Para construção dessa base, foram utilizadas informações de Fortaleza sobre tráfego, horários de pico, dias da semana, quantidade de viagens e o simulador de preços da Uber. Para alcançar o objetivo da pesquisa, a Metodologia empregada consistiu nas etapas de Obtenção e Construção de preços, Obtenção de Dados Socioeconômicos, Análise Exploratória de Dados, Limpeza e Tratamento de Dados, Construção de Modelos de Aprendizado de Máquina e Análise entre os Dados Socioeconômicos e os preços de viagens para as cidades em estudo. Como resultados obtidos, observou-se que, em um cenário mais desconcentrado de centro financeiro, os usuários de baixa renda da Uber em Fortaleza poderiam ter os preços das viagens reduzidos em cerca de 43,07%. Essa redução representaria uma economia mensal de cerca de 18,82% de suas Rendas Médias Pessoais. Para usuários que vivem em bairros ricos (alta renda), essa descentralização aumentaria os custos de viagens para pouco mais de 100%. No entanto, esse aumento representaria 6,71% de suas Rendas Médias Pessoais. Futuras pesquisas podem expandir os resultados aqui obtidos, otimizando a base de dados criada e modificando o processo de Ciência de Dados utilizado.

Palavras – chave: Transportes por Aplicativo, Dados Socioeconômicos, Ciência de Dados, Análise Exploratória de Dados.

ABSTRACT

Studies that use data from the transport company Uber showed that there are factors that contribute to the increase in prices of its travel services. In this context, this research aims to analyze travel routes for low-income users and contribute to reducing these prices. For this, we sought to answer: If a financial center were closer to economically poorer neighborhoods, would there be a change in the average prices of these trips? Could this change financially improve the lives of low-income people? The purpose of our research to answer these questions was to investigate this factor of financial concentration in territorial regions through a Data Science process, analyzing prices and socioeconomic data in the South American city of Fortaleza, located in the country Brazil and from the North American city of Boston, located in the United States of America. Thus, it would be possible to show whether Uber users who live in financially poorer neighborhoods and use this travel service end up paying more than users in richer neighborhoods, when the destination is the financial center. The analyzes and results obtained for Boston served as a validation by analogy for the results obtained for Fortaleza. The analyzed database for Boston refers to a real dataset, available in the Kaggle online community. The database analyzed for Fortaleza was built during our work and is also available on the Kaggle online community, which can serve as a tool for future analysis in other research. To build this base, information from Fortaleza about traffic, peak times, days of the week, number of trips and the Uber price simulator were used. To achieve the objective of the research, the methodology used consisted of the steps of Obtaining and Construction of prices, Obtaining Socioeconomic Data, Exploratory Data Analysis, Cleaning and Processing of Data, Construction of Machine Learning Models and Analysis between the Socioeconomic and travel prices to the cities under study. As results obtained, it was observed that, in a more decentralized scenario of a financial center, low-income users of Uber in Fortaleza could have their trip prices reduced by about 43.07%. This reduction would represent a monthly savings of around 18.82% of their Average Personal Income. For users living in wealthy (high-income) neighborhoods, this decentralization would increase travel costs to just over 100%. However, this increase would represent 6.71% of their Average Personal Income. Future research can expand the results obtained here, optimizing the created database and modifying the Data Science process used.

Keywords: Transport by Application, Socioeconomic Data, Data Science, Exploratory Data Analysis.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS & SIGLAS

AED	Análise Exploratória de Dados
AM	Aprendizagem de Máquina
API	Application Programming Interface
ATTS	App-based Third-party Taxi Service
BP&DARD Boston	Agência de Pesquisa da Divisão de Planejamento e Desenvolvimento de
CV	Cross Validation
EQM	Erro Quadrático Médio
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
IDH – B	IDH a nível de Bairro para Fortaleza
IDH-R	Índice de Desenvolvimento Humano - Renda
IDM - M	IDH Municipal
IPCA	Índice Nacional de Preços ao Consumidor Amplo
IPLANFOR	Instituto de Planejamento de Fortaleza
IQR	Interquartile Range
K-S	Kolmogorov-Smirnov
LGPD	Lei Geral de Proteção de Dados
MTE	Ministério do Trabalho e Emprego
NAU	Nova Agenda Urbana
NTU	Associação Nacional das Empresas de Transportes Urbanos
ODS	Objetivos de Desenvolvimento Sustentável
ONU	Organização das Nações Unidas
PlanMob	Plano de Mobilidade
PNAD	Pesquisa Nacional por Amostra de Domicílios
PNUD	Programa das Nações Unidas
REQM	Raiz do Erro Quadrático Médio
RMP	Renda Média Pessoal Mensal
SDK	kits de desenvolvimento de software
SER	Secretarias Executivas Regionais
SGD	Gradiente Descendente
SIT-FOR	Sistema Integrado de Transporte de Fortaleza
SMDE	Secretaria Municipal de Desenvolvimento Econômico
UDH	Unidades de Desenvolvimento Humano
URR	Uber Ride Request

SUMÁRIO

INTRODUÇÃO

Este capítulo apresenta a introdução do trabalho, informando a motivação, os objetivos e a organização dos textos subsequentes.

1.1 Motivação

No âmbito das “Smart Cities” (Cidades Inteligentes), os governos e as organizações realizam investimentos em capital humano, social, infraestrutura e em tecnologias que possuem a capacidade de criar novos produtos e serviços com possibilidade de desestabilizar concorrentes que antes dominavam o mercado. Dessa forma, seria possível promover um crescimento econômico sustentável, melhorando a qualidade de vida em geral. Nesse contexto, os transportes por aplicativos podem ser considerados exemplos de serviços em Cidades Inteligentes (OBEDAIT ET AL, 2019).

De acordo com Quick (2020), os transportes por aplicativos - conhecidos também por táxi por aplicativo e carona remunerada - são serviços digitais de transporte de passageiros, transporte de refeições e delivery de itens diversos. A participação de pessoas neste mercado, com objetivo de adquirir uma renda extra ou até encontrar um emprego que seja mais rentável, é um fenômeno mundial. Para uma pessoa prestar serviços nesse âmbito, é preciso possuir um veículo e vincular-se a uma ou mais empresas de aplicativos. Algumas empresas desse nicho são: Uber, InDriver, Lyft, 99, Cabify, Rappi e iFood.

Nesse contexto, existem fatores que podem influenciar a oferta e a demanda dos transportes por aplicativos, como características urbanas, demográficas, renda, concorrência, disponibilidade de outros meios de transporte, fluxo turísticos peculiares a cada cidade, dentro outros (QUICK,2020). Nessa perspectiva, pesquisas recentes buscaram encontrar relações entre características socioeconômicas de alguma região e alguns aspectos da empresa de transporte por aplicativo Uber (SILVA,2020). Esses aspectos são: acessibilidade; tempo de espera; indicadores de qualidade de vida de bairros; indicadores de habitabilidade para cidades e bairros; e planejamento urbano. Todavia, segundo Silva (2020), esses estudos pouco têm

explorado a dimensão preço na relação entre o serviço ofertado pela Uber e as características socioeconômicas dos locais de embarque e/ou desembarque. O autor observou que o tempo e a distância estão relacionados ao processo de precificação do serviço de viagens da Uber para a cidade de Natal, localizada no Brasil, e que isso possibilitaria uma melhora nas estratégias de oferta e demanda desse serviço. Porém, esse processo pode apresentar outros fatores que possam vir a contribuir na concepção desses preços. Esse fator seria a grande concentração de atividades comerciais em uma porção territorial, afetando a oferta e a demanda, ocasionando uma elevação de preços para usuários que residem longe desses centros financeiros.

No intuito de averiguar esse fator de concentração, nosso trabalho analisou preços, considerando dados socioeconômicos da cidade sul-americana de Fortaleza, localizada no país Brasil e da cidade norte americana de Boston, localizada no país Estados Unidos da América, no intuito de evidenciar que os usuários que moram em bairros mais pobres financeiramente e utilizam o serviço de viagens da Uber acabam pagando mais caro do que os moradores dos bairros mais ricos, quando o destino é o centro financeiro. Cabe salientar que as análises e os resultados obtidos para Boston serviram de validação por analogia para os resultados aqui obtidos para a cidade de Fortaleza. Dessa forma, foi possível dar mais justificativa para pesquisar nesse contexto e focalizar nosso trabalho na cidade de Fortaleza. Considerando isso e tendo acesso a essa base de dados de Boston, optou-se por realizar experimentos com Modelos de Aprendizado de Máquina, no intuito de analisar a predição de distância com base em preços ofertados. Dessa forma, seria possível recomendar políticas públicas baseadas em Ciência de Dados, obtendo informações de possibilidade de redução de custos.

Essas políticas poderiam ser tomadas de decisões no âmbito de propostas de descentralizações de centros econômicos, possibilitando amenizar gastos da renda dos usuários mais pobres, sem afetar de maneira muito significativa os usuários de regiões mais ricas. Nesse sentido, nosso trabalho tentou comparar as diferenças de custos de deslocamento dos usuários que utilizam o serviço de viagem de transporte por aplicativo Uber para alguns bairros e regiões das cidades acima mencionadas. Para alcançar esse objetivo, foi realizada uma Revisão da Literatura em que abordamos assuntos no âmbito de Transportes por Aplicativos, Dados Socioeconômicos, Ciência de Dados, Técnicas de Avaliação de Desempenho, Análise Estatística, Medidas de Centralidade para Grafos e Trabalhos Relacionados. Após esse levantamento, foi esquematizada uma Metodologia que consistiu na elaboração de Estratégias que serviram para aplicar o conteúdo levantado na Revisão da Literatura, bem como para a Obtenção e Construção de preços para as cidades de Fortaleza e Boston. Além disso, ainda nessa etapa metodológica foi realizado o Processo de Obtenção de Dados Socioeconômicos

para as cidades em estudo. Outra etapa da Metodologia, foi a esquematização de procedimentos para realizar uma Análise Exploratória dos Dados obtidos, bem como procedimentos para Limpeza e Tratamento dos Dados. A última etapa metodológica foi o procedimento para Construção dos Modelos de Aprendizado de Máquina para a cidade de Boston, bem como o procedimento para a Análise dos Dados Socioeconômicos com os preços de viagens para essas cidades.

Durante a realização de toda a nossa pesquisa, não foram encontrados dados públicos de empresas de viagens de transporte por aplicativo para a cidade de Fortaleza. Devido a legislações vigentes, como a Lei Geral de Proteção de Dados (LGPD), a divulgação de informações nesse contexto é restrita. Por conta disso, optamos por simular uma base de dados para essa cidade, considerando a plataforma de simulação de preços da própria Uber e as pesquisas realizadas sobre horários de pico, demanda e oferta do serviço por semana, tráfego e preços praticados. Nesse sentido, foram realizados alguns passos para concepção dessa base de dados: Determinação do tipo de serviço Uber que seria objeto de estudo, Criação de trajetos de viagens para os bairros em estudo, Considerações acerca da quantidade de solicitações de viagens, Utilização do simulador de preços da Uber, Obtenção das faixas de valores dos preços e reajuste dos mesmos quando necessário, e, por fim, a Criação da base de dados com informações dos preços, trajetos, horários e dias da semana. Importante salientar que na construção dessa base, a quantidade de viagens estipulada para os trajetos levou em consideração os achados na Revisão da Literatura o que ocasionou uma distribuição dessa quantidade de forma não equiparada, sendo estabelecidos porcentagens distintas de quantidade de viagens a depender do dia da semana e do horário do dia.

Por outro lado, para a cidade de Boston foi possível encontrar uma base de dados real de viagens de transporte por aplicativo Uber. Os preços se referem ao serviço UberX. Essa base está disponível na comunidade *online Kaggle*. O nome da base de dados se chama *"rideshare_kaggle"* e pode ser obtida por meio do link <https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma>. Nossa base de dados criada para a cidade de Fortaleza se chama *"uber_fortaleza"* e está disponível também nessa comunidade, podendo ser obtida por meio do link < https://www.kaggle.com/datasets/giancarlolimatorres/uber-fortaleza>.

A escolha da cidade de Fortaleza foi devido à facilidade de encontrar dados socioeconômicos a nível de bairro e por pertencer a região Nordeste do Brasil, com características socioeconômicas parecidas com outras capitais dessa região: trânsito movimentado, clima predominantemente quente e desigualdade de renda (SILVEIRA, 2020). Nesse mesmo caminho, foi realizada a escolha da cidade de Boston, pois foi a cidade que foi encontrada uma base de dados mais completa e coerente a nível de bairro, embora os dados socioeconômicos não sejam de fácil acesso como os encontrados para Fortaleza.

Como resultados obtidos pela Análise Exploratória de Dados, para ambas as cidades, houve convergência no comportamento de alta de preços em viagens para centros comerciais. Para a cidade de Boston, o Modelo de Aprendizado de Máquina de Floresta Aleatória obteve maior desempenho em relação aos demais Modelos provenientes dos demais algoritmos regressores abordados nesta pesquisa. Uma justificativa para isso é o fato de que a concepção de preços apresenta um comportamento não linear, sendo mais perceptível para algoritmos não lineares como a Floresta Aleatória. Com esse resultado, seria possível recomendar políticas públicas no sentido de redução de custos para usuários desse tipo de serviço de viagens. Para a cidade de Fortaleza, as Análises Estatísticas evidenciaram a diferença representativa de preços entre trajetos de bairros mais pobres para bairros mais ricos financeiramente. Em um cenário de concentração de centro financeiro para essa cidade, o impacto dos custos dessas viagens para usuários mais pobres economicamente é alto em relação a suas Rendas Médias Pessoais. Em contrapartida, os usuários mais ricos são pouco afetados economicamente em suas Rendas considerando esse cenário. Por outro lado, caso esse cenário fosse mais desconcentrado, os usuários de baixa renda teriam redução nos custos dessas viagens, aumentando suas Rendas e os usuários mais ricos gastariam um pouco mais, mas não teriam suas Rendas afetadas de maneira elevada.

Considerando essa conjuntura, esta pesquisa pode contribuir na análise de dados e propor recomendações de políticas públicas para redução de custos de pessoas de baixa renda. Além disso, pesquisas futuras poderão estender os resultados aqui obtidos, otimizando a base de dados criada para Fortaleza e propondo melhorias no processo de Ciência de Dados adotado.

1.2 Objetivo da Pesquisa

1. Objetivo Geral

Esta pesquisa tem como objetivo analisar os trajetos de viagens de usuários de baixa renda e contribuir na redução dos preços dessas viagens em transporte por aplicativo Uber, considerando as cidades de Fortaleza e Boston como regiões de estudo. Para isso, buscou-se responder: Se um centro comercial fosse mais próximo de bairros economicamente mais pobres, haveria mudança na média de preços? Essa mudança poderia melhorar financeiramente a vida das pessoas de baixa renda?

2. Objetivos Específicos

Para atingir o objetivo geral supracitado, os seguintes objetivos foram considerados:

- Entender as relações entre as dimensões “preço” e “distância” em um contexto de viagens de transporte por aplicativo Uber;
- Analisar dados socioeconômicos das cidades de Fortaleza e Boston que utilizam esse tipo de transporte;
- Avaliar metodologias de relacionamento entre as dimensões citadas e os dados socioeconômicos dessas cidades;
- Desenvolver uma estratégia de coleta de informações de viagens desse seguimento e de informações socioeconômicas para as cidades estudadas;
- Analisar as informações obtidas das viagens por meio de Análise Estatística, Medidas de Centralidade para Grafos e Ciência de Dados;
- Obter relações entre os dados socioeconômicos e as informações analisadas das viagens;
- Estabelecer uma representação dos resultados obtidos;
- Averiguar as relações evidenciadas.

1.3 Organização do Texto

O restante dessa dissertação está estruturado da seguinte forma:

- **Capítulo 2 – Revisão da Literatura:** aborda informações acerca de transportes por aplicativos, enfatizando a empresa Uber e sua importância no mercado. Nesse capítulo também são abordados alguns dados socioeconômicos para as cidades de Fortaleza e Boston a nível de bairro que serviram de fundamento para as análises posteriores. Em seguida, algumas informações foram apresentadas acerca de Ciência de Dados, pois parte das análises precisou de um processo de coleta, limpeza, análise e criação de Modelos de Aprendizagem de Máquina, bem como obtenção de *insights* e respaldo científico para tomadas de decisão. Também foram abordados assuntos que fazem referência a algumas Técnicas de Avaliação de Desempenho para os modelos criados. A Análise Estatística foi abordada devido à necessidade de validação da representatividade das amostras em estudo e de averiguar que tipo de distribuição elas seguiam. A abordagem sobre Medidas de Centralidade para Grafos auxiliou nas análises sobre quantidade de viagens e importância de alguns trajetos. Por fim, os Trabalhos Relacionados abordam alguns estudos que evidenciaram relações entre características socioeconômicas de alguma região e alguns aspectos da empresa de transporte por aplicativo Uber.

- **Capítulo 3 – Metodologia:** aborda a estratégia de aplicação dos conteúdos levantados na Revisão da Literatura, para realizar as análises e o procedimento de representação das observações encontradas. Nesse sentido, o capítulo aborda o processo de obtenção de dados socioeconômicos para ambas as cidades em estudo, além de abordar o procedimento de construção dos modelos de Aprendizagem de Máquina para Boston, bem como a estratégia de obtenção dos preços para as viagens de Fortaleza. O capítulo também aborda o procedimento de análise entre os dados socioeconômicos e os preços das viagens de ambas as cidades.

- **Capítulo 4 – Resultados e Discussões:** capítulo em que são feitas as análises sobre os dados das viagens para ambas as cidades, considerando os trajetos entre bairros de classe pobre, média e rica, em termos de Renda Média Pessoal mensal. O capítulo analisa os dados

socioeconômicos dessas rendas a nível de bairro no intuito de evidenciar as possíveis relações entre os preços das viagens e as regiões de alta e baixa concentração comercial.

- **Capítulo 5 – Conclusão:** este capítulo formaliza uma síntese dos resultados obtidos nas análises, bem como algumas limitações da pesquisa e futuras análises que podem ser feitas.

REVISÃO DA LITERATURA

Neste capítulo são abordados os seguintes assuntos: Transportes por Aplicativos, Dados Socioeconômicos a nível de bairro, Ciência de Dados, Técnicas de Avaliação de Desempenho, Análise Estatística, Medidas de Centralidade para Grafos e Trabalhos Relacionados.

2.1 Transportes por Aplicativos

O transporte público pode ser dividido em coletivos e em individuais. De acordo com a Lei nº 12.587 / 2012 (Política Nacional de Mobilidade Urbana), transporte coletivo inclui serviços públicos para toda a população, enquanto transporte individual pode ser definido como transporte de passageiros que não é aberto ao público e é realizado por meio de viagens personalizadas.

O transporte individual de passageiros possui relação com o transporte urbano. Nesse sentido, o estudo do tráfego individual pode ser considerado um catalisador para o melhoramento da mobilidade urbana, pois pode contribuir na redução do número de veículos particulares e, assim, conseguir o compartilhamento, ajudando a otimizar o uso dos veículos urbanos. Dentre as vantagens do transporte individual, a autonomia e a praticidade se destacam, pois o deslocamento ocorre de acordo com os desejos dos passageiros, inclusive nas rotas definidas por eles mesmos (FARIAS, 2016).

No contexto do transporte urbano, o desenvolvimento e o uso de novas tecnologias mudaram o dominante mercado de táxis. A coexistência de serviços de táxi tradicionais e os novos "*app-based third-party taxi servisse*" (ATTS) baseados em aplicativos transformaram a indústria em um mercado de economia compartilhada. O ATTS é fundamentalmente diferente dos táxis tradicionais em termos de barreiras à entrada e às políticas tarifárias, trazendo concorrência para o mercado. Quando a demanda de viagens aumenta, sua tarifa aumentará em valor variável com base no tempo e distância estimados de viagem (QIAN e UKKUSURI, 2017).

De acordo com Zhang Jia et al. (2016), o mercado tradicional de táxis está ameaçado pelos novos e "rebeldes táxis feitos sob medida". Rael et al. (2016) afirmam que esse tipo de serviço é denominado serviço de "*ride-hailing*", assim como os serviços prestados por empresas como Uber e Lyft, baseados em aplicativos móveis e sob demanda. Nesse contexto, o papel desse meio de transporte no âmbito do transporte urbano tem gerado polêmica. De acordo com os autores, os motoristas de táxis acreditam que os serviços de transporte por aplicativos são ilegais porque contornam a legislação existente e se envolvem em concorrência desleal.

Nesse cenário, a Uber fornece serviços como compartilhamento ponto a ponto, transporte de táxi, entrega de comida e compartilhamento de bicicletas para cerca de 3 milhões de motoristas e aproximadamente 75 milhões de passageiros em mais de 600 cidades ao redor do mundo. São mais de 15 milhões de viagens, em média, concluídas na Uber todos os dias, e os dados de cada reserva e viagem são registradas pela própria empresa. Portanto, este é um mercado de mão dupla em que a plataforma Uber deve mediar para calibrar oferta e demanda, garantindo que ambas as partes (cliente e empresa) obtenham altos níveis de satisfação (BEZERRA,2019).

A Uber coleta grandes quantidades de dados de passageiros e motoristas, fornecendo acesso seletivo a esses dados por meio de *Application Programming Interface* (API) e kits de desenvolvimento de software (SDKs) disponíveis em seu site de desenvolvedor. Um desses serviços para desenvolvedores é a API *Uber Ride Request* (URR). URR fornece acesso a muitas funções básicas do aplicativo móvel Uber, por exemplo, a seleção do tipo de serviço (*Uber X, Uber Select, Uber Black, Uber Pool,* etc.), especificação de locais de coleta, tempo estimado de chegada, dados de preços estimados e solicitações de viagens.

Segundo Bezerra (2019), as pesquisas que utilizam Uber APIs são geralmente divididas em alguns grupos temáticos: compreensão dos fenômenos de preços, otimização de itinerários e recompensas, pesquisa do consumidor, pesquisa de competitividade e construção de novos produtos e serviços. Devido a isso, presume-se que os dados da Uber podem fornecer um indicador de habitabilidade de uma cidade de maneira simples, rápida, de baixo custo e sensível ao tempo e ao contexto.

Devido à natureza, à escala e à cobertura das operações da Uber, a empresa fornece uma fonte única de dados em tempo real para a interação entre os residentes urbanos e a infraestrutura de transporte. Essas informações permitem a comparação de vários níveis de dados em cidades, regiões e comunidades, mas também fornecem dados sensíveis ao contexto que fornecem uma visão sobre o impacto de outros fatores que afetam os motoristas da Uber e as viagens diárias, incluindo acidentes de trânsito, clima, e outros eventos (BEZERRA, 2019).

De acordo com a Uber (2021), o preço estimado que os usuários da Uber veem no aplicativo consiste na soma de três partes: a tarifa básica, o número de quilômetros percorridos e o número de minutos gastos na viagem. Quando a soma desses componentes for menor que o preço mínimo, a soma será substituída por este, que será igual ao preço de cancelamento do serviço.

Além disso, existem preços padrão que complementam a estrutura básica de preços. A Uber também utiliza as chamadas "tarifas dinâmicas" nas quais o preço básico muda de acordo com uma relação específica de oferta e demanda. O modelo utilizado pela Uber para construir tarifas dinâmicas é implementado por meio de um algoritmo de precificação complexo de código fechado (LOBEL, 2015).

Essa tarifa dinâmica funciona aumentando o preço do consumidor em um múltiplo. Por exemplo, 1,4 vezes significa que o preço da viagem é 40% mais alto do que o preço padrão. Esse múltiplo é baseado no nível de demanda associado a uma determinada oferta de motoristas disponíveis no entorno calculado (DELOITTE, 2016).

2.2 Dados Socioeconômicos

Esta seção informa os dados socioeconômicos pertinentes para a elaboração da pesquisa, considerando as duas cidades em estudo, Fortaleza e Boston. Fortaleza foi escolhida devido à facilidade de encontrar dados socioeconômicos a nível de bairro e por pertencer a região Nordeste do Brasil, com características socioeconômicas parecidas com outras capitais dessa região: trânsito movimentado, clima predominantemente quente e desigualdade de renda (SILVEIRA, 2020). Nesse mesmo caminho, foi realizada a escolha da cidade de Boston, pois foi a cidade que foi encontrada uma base de dados mais completa e coerente a nível de bairro, embora os dados socioeconômicos não sejam de fácil acesso como os encontrados para Fortaleza.

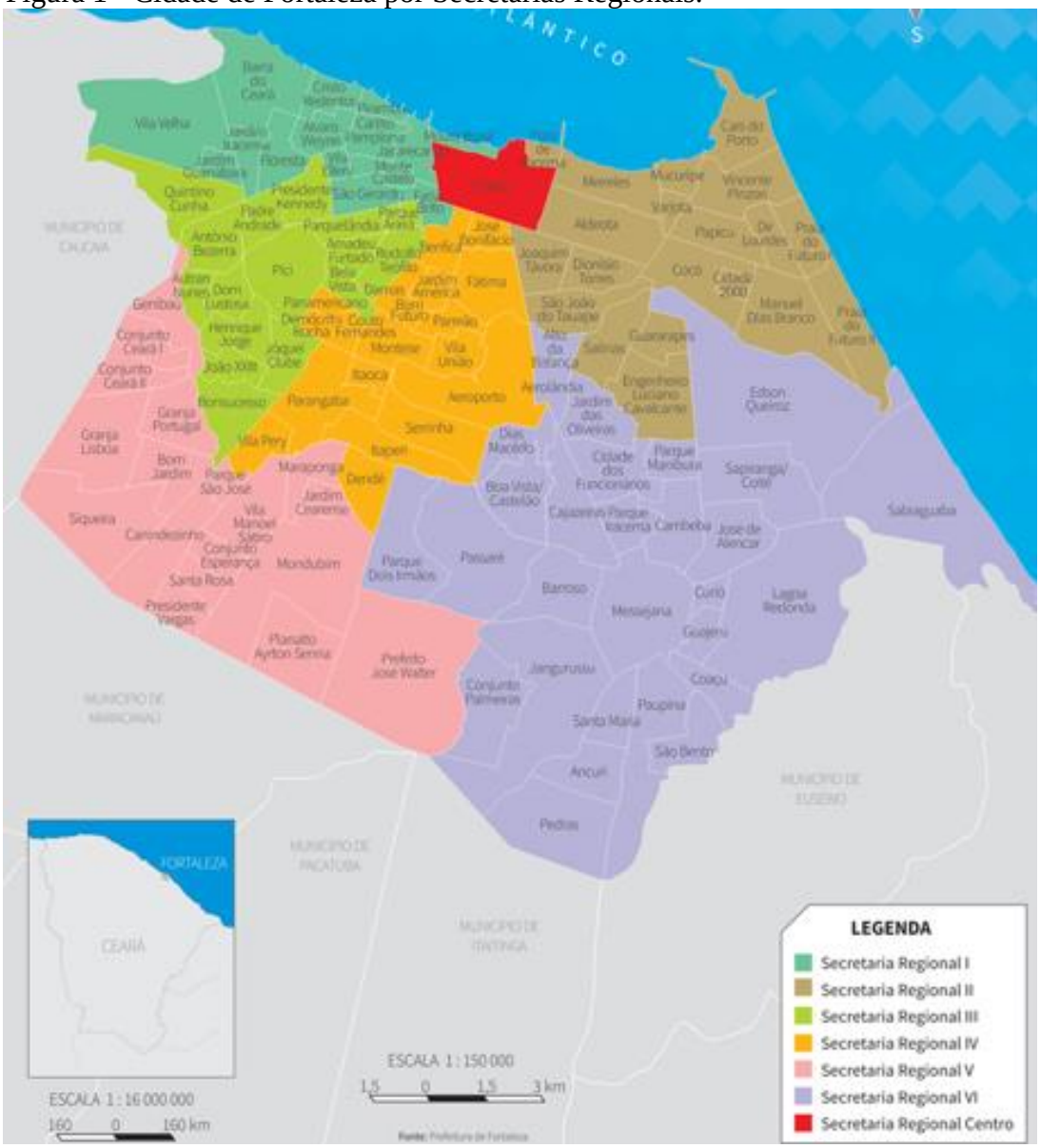
2.2.1 Fortaleza

A cidade de Fortaleza, capital do estado do Ceará, está localizada nas margens do Oceano Atlântico e situada na região Nordeste do Brasil, em uma zona de clima tropical, marcada pela elevada umidade. A cidade abriga 121 bairros, 39 Territórios Administrativos e doze Secretarias Executivas Regionais (SER), que são suas subprefeituras, órgãos públicos de gestão direta de cada área. A economia local é bastante diversificada, com importantes atividades secundárias e terciárias. Seu território é muito procurado por turistas, em razão da presença de belas praias e da rica cultura local (GOVERNO DO CEARÁ, 2022).

Em 2015, o Instituto de Planejamento de Fortaleza (IPLANFOR), por meio de um estudo sobre mobilidade chamado de Plano de Mobilidade (PlanMob), observou que as pessoas que moram em bairros periféricos necessitam viajar para o centro comercial diariamente, pois a cidade ainda possui uma grande concentração comercial na porção norte da cidade. Porção essa que abrange quase todo o centro comercial. A pesquisa também cita as viagens relacionadas à educação que também são realizadas em grande parte para o centro comercial. Devido a isso, há uma sobrecarga no sistema viário e no sistema de transporte, pois acarreta congestionamentos nas vias, excesso de fluxo de pessoas nos terminais urbanos e ônibus operando acima da capacidade.

O IPLANFOR destaca que os bairros que mais produzem viagens estão localizados nas regiões Oeste e Sul da cidade, locais que concentram a maior parte dos bairros de baixa renda (regiões periféricas). Por outro lado, os bairros de maior atração de viagens coincidem com as áreas de maior oferta de emprego (centro comercial) que estão localizadas nas regiões Norte e Nordeste, correspondendo com a maioria dos bairros de alta renda. Essa análise explica os grandes volumes de pessoas se deslocando diariamente da periferia para a região Norte e Nordeste da cidade, gerando congestionamentos e a superlotação de algumas linhas de transporte coletivo em horários de pico. A Figura 1 ilustra o mapa da cidade e destaca as Secretarias Executivas Regionais.

Figura 1 - Cidade de Fortaleza por Secretarias Regionais.



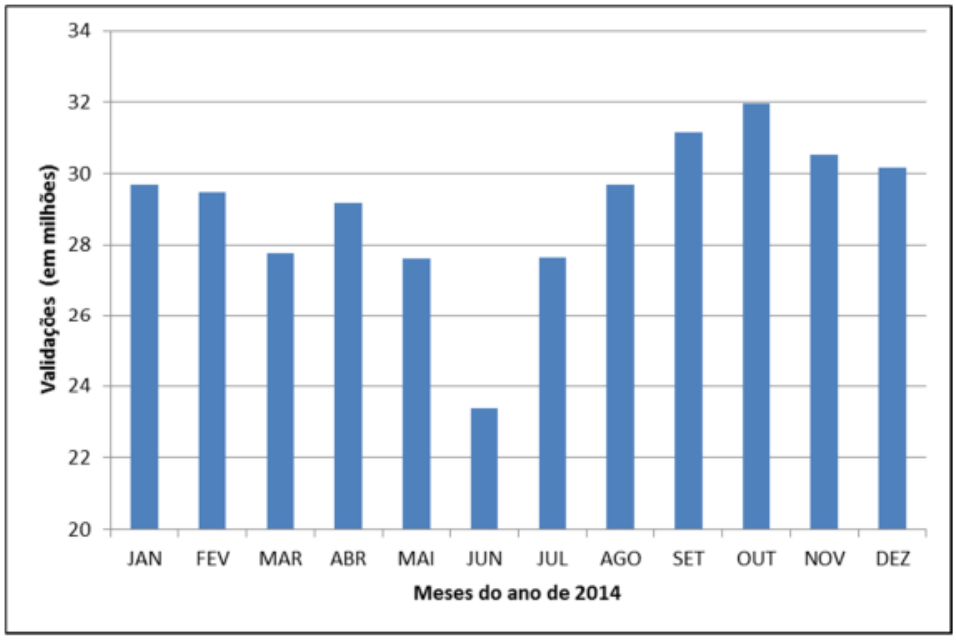
Fonte: (IPLANFOR, 2015).

A Secretaria Regional V possui a maior quantidade de bairros de baixa renda da cidade. A Secretaria Regional Centro e a Secretaria Regional II concentram todo o centro comercial, coincidindo com a maior parte dos bairros de alta renda.

De acordo com o estudo do instituto, a espinha dorsal do transporte coletivo em Fortaleza é o sistema regular operado por ônibus que é integrado ao sistema complementar operado por micro-ônibus. Esses transportes são fiscalizados e gerenciados pelo Sistema Integrado de Transporte de Fortaleza (SIT-FOR). Estima-se que, em 2015, o SIT-FOR era composto por 295 linhas regulares e 22 linhas complementares, realizando o transporte de aproximadamente um milhão de passageiros por dia. Além disso, foi observado que, dentre outras variáveis, a superlotação dos transportes coletivos é considerada alta devido ao concentrado de pessoas que se deslocam dos bairros periféricos (bairros de baixa renda) para o centro comercial da cidade (bairros de alta renda). Além disso, essa problemática é multidisciplinar, podendo haver outras variáveis influenciadoras.

O IPLANFOR destacou em sua pesquisa, considerando o ano de 2014 e o transporte coletivo da cidade, o número de validações das viagens realizadas durante cada mês para aquele ano. Validação significa o tipo de viagem realizada pelo usuário no que se refere ao pagamento: Inteira, Gratuidade ou Meia. O Gráfico 1 indica que o mês com menor número de validações correspondeu ao mês de junho com 23.372.027 validações, já o mês de outubro foi o que apresentou maior número de validações com 31.968.460.

Gráfico 1 - Distribuição das validações no ano de 2014.

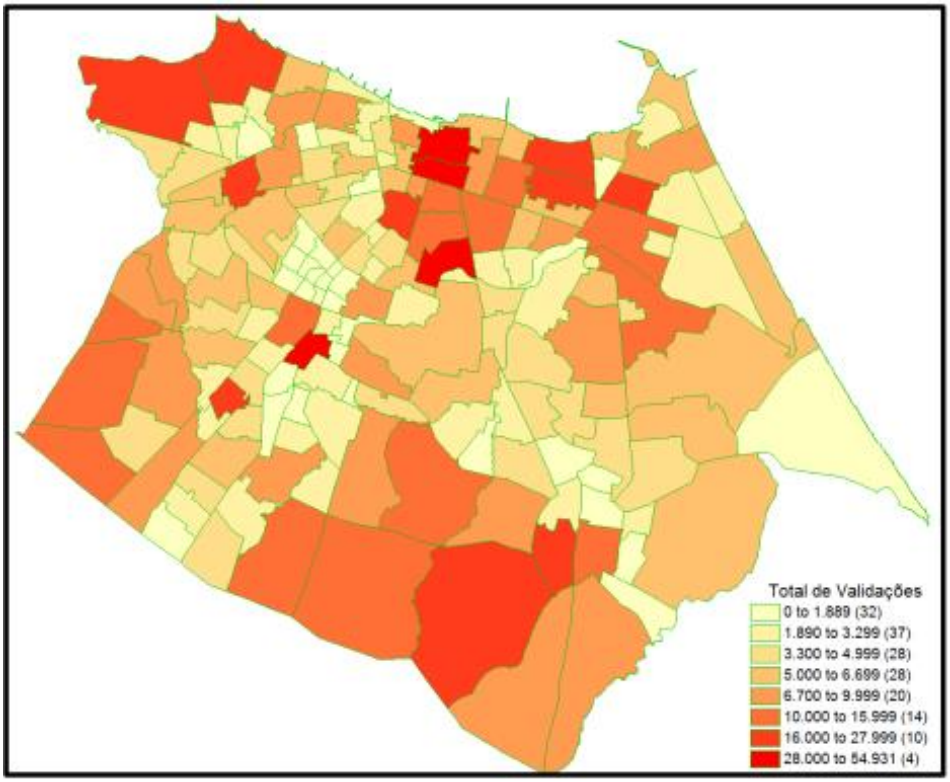


Fonte: (IPLANFOR, 2015).

A média mensal de validações foi de 29.015.178 com um desvio padrão de 2.143.967, indicando elevada variação mensal, ou seja, o sistema não segue uma distribuição previsível para ocorrência de validações.

O estudo do instituto também observou a distribuição espacial da quantidade de viagens realizadas por bairro para o ano de 2014. A Figura 2 ilustra um mapa de carregamento de viagens por bairro. Quanto mais escuro, maior o número de produção de viagens ocorridas no bairro. As áreas da cidade com maior número de viagens são regiões periféricas localizadas ao sul e ao leste da cidade, bem como a região central ao norte, coincidindo com a maior parte dos bairros pobres (periferia) e dos bairros ricos (centro comercial) da cidade. De acordo com o estudo, este comportamento é esperado uma vez que essas regiões periféricas possuem maior densidade populacional e a região central possui maior quantidade de serviços, empregos e comércios.

Figura 2 - Total de validações por bairro em 2014.



Fonte: (IPLANFOR, 2015).

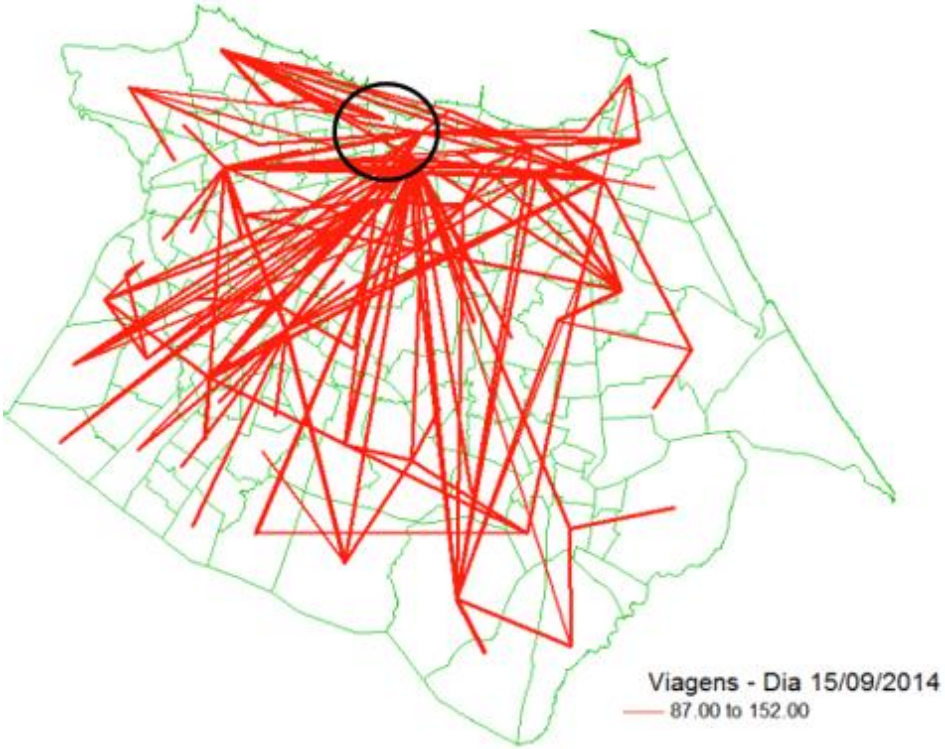
A pesquisa informa que devido a isso Fortaleza é uma cidade com grande dependência da área central. Além disso, o instituo observou que há relação de correlação entre a produção de viagens com a densidade, renda, e outras variáveis socioeconômicas. Isso evidencia que a

renda é um fator de impacto no que se refere ao deslocamento das pessoas na cidade de Fortaleza.

A produção de viagens nas regiões periféricas tem início mais cedo do que nas regiões melhor centralizadas. Acredita-se que isso se deve ao fato de as distâncias percorridas pelos usuários das regiões periféricas serem maiores do que as distâncias das regiões mais próximas ao centro. Portanto, o início das viagens nas áreas periféricas acaba ocorrendo mais cedo.

A pesquisa do instituto informa acerca de um trabalho sobre matriz de origem-destino dos deslocamentos dos usuários dos transportes coletivos a partir dos dados de Bilhetagem Eletrônica. Por essa matriz foi possível observar as principais linhas de desejo, ou seja, preferência de trajeto. A Figura 3 ilustra esse comportamento.

Figura 3 - Linhas de desejo considerando 87 a 152 viagens/dia.

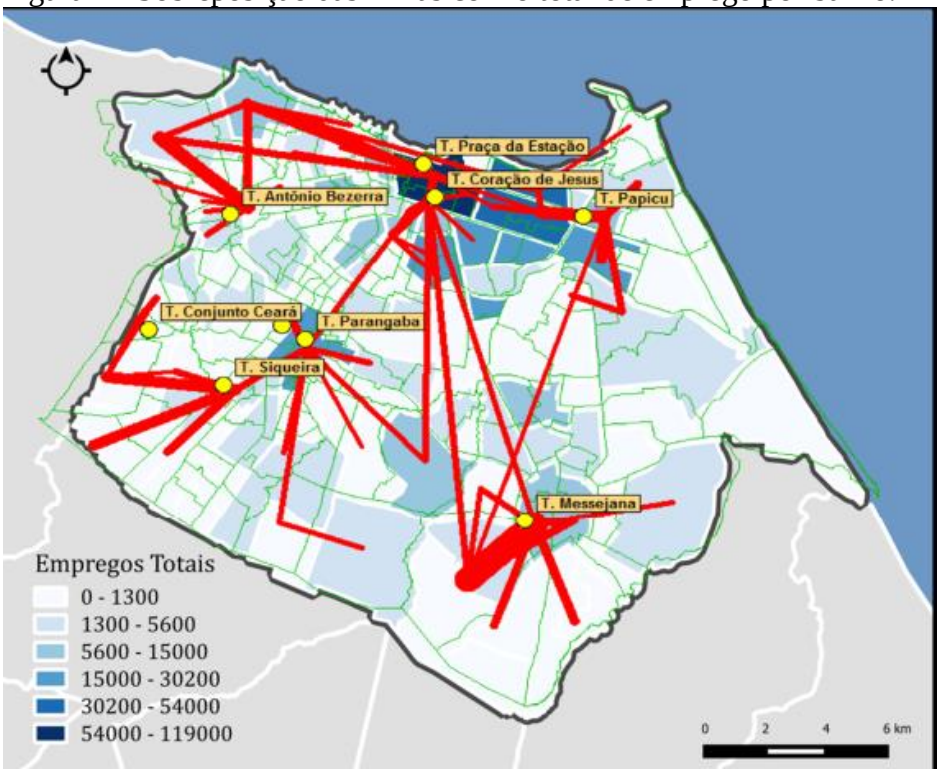


Fonte: (IPLANFOR, 2015).

Há uma grande tendência de deslocamentos em direção ao centro e aos terminais de integração, o que mostra a grande dependência da região central.

Dados do Ministério do Trabalho e Emprego (MTE) evidenciaram que as linhas de viagens coincidem com as zonas de maior número de empregos que estão localizadas nos bairros mais ricos da cidade, inclusive o centro comercial. A Figura 4 ilustra esse comportamento entre as linhas de viagens e o total de emprego por bairro.

Figura 4 - Sobreposição das linhas com o total de emprego por bairro.



Fonte: (IPLANFOR, 2015).

O IPLAFOR salienta que as observações foram prioritariamente sobre viagens que se destinaram ao trabalho e à educação. Além disso, grande parte da população da cidade é dependente do transporte público e são residentes de áreas periféricas, muito distantes dos principais pontos de atividades comerciais, obrigando-os a iniciar seus deslocamentos logo nas primeiras horas do dia.

O estudo do instituto cita como um dos problemas de mobilidade urbana da cidade a alta concentração de renda e a consequente segregação geográfica da sociedade: as pessoas de maior poder aquisitivo ocupando a zona Leste e Norte da cidade (concentra a porção mais nobre dos bairros e o centro comercial) e as de menor poder aquisitivo ocupando as zonas Sul e Oeste. Além disso, observou-se uma superlotação diária do sistema de transporte público por ônibus e congestionamento nas principais vias. Grande parte desse congestionamento é explicado pela concentração da maior parte das atividades econômicas e consequentemente da maior quantidade de empregos nas zonas Norte e Leste. Em contraste, observou-se a maioria da população residindo na zona Oeste e Sul da cidade, por isso as viagens de deslocamento a grandes distâncias.

Considerando essa conjuntura, dados do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2010, por meio da Pesquisa Nacional por Amostra de Domicílios (PNAD),

revelaram que no item renda - um dos três principais itens do Índice de Desenvolvimento Humano (IDH) – o valor médio per capita de Fortaleza teve um aumento de 85,18% nas duas últimas décadas, passando de R$ 457,04 em 1991 para R$ 610,48 em 2000 e R$ 846,36 em 2010. Esse índice é chamado de IDH-R e representa a renda média mensal per capita por bairro. A pobreza extrema (medida como a proporção de pessoas com renda domiciliar per capita inferior a R$ 70,00) passou de 15,25% em 1991 para 9,02% em 2000 e para 3,36% em 2010.

De acordo com Silveira (2020), o IDH – R é considerado um indicador do potencial médio dos residentes de um bairro para obter bens e serviços. Ele é usado como um indicador da capacidade das pessoas de garantir um padrão de vida que possa atender às suas necessidades básicas. A Tabela 1 mostra 5 bairros de cada classe de renda (ricos, pobres e médios) para a cidade de Fortaleza em termos de IDH – R em 2010. Os valores se referem a pessoas com 10 anos ou mais de idade.

Tabela 1 - IDH - R de alguns bairros de classes de renda diferentes em Fortaleza.

RICOS	**IDH – R**	**RENDA MÉDIA PESSOAL R$**	**MÉDIA DA RENDA MÉDIA PESSOAL R$**
Meireles	0,953	3659,54	3.099,72
Aldeota	0,778	2901,57	
Dionísio Torres	0,722	2707,35	
Mucuripe	0,732	2742,25	
Guararapes	0,950	3488,25	
POBRES			
Conjunto Palmeiras	0,010	239,25	301,88
Parque Presidente Vargas	0,014	287,92	
Canindezinho	0,025	325,47	
Genibaú	0,027	329,98	
Siqueira	0,026	326,8	
MÉDIOS			
Autran Nunes	0,032	349,74	591,76
Dendê	0,115	633,44	
Parque Dois Irmãos	0,093	557,84	
Cajazeiras	0,155	768,93	
Messejana	0,120	648,89	

Fonte: Adaptado de Secretaria Municipal de Desenvolvimento Econômico de Fortaleza com base nos dados do Censo Demográfico de 2010.

Os bairros em pior situação no quesito IDH – R são: Conjunto Palmeiras, Parque Presidente Vargas, Canindezinho, Siqueira e Genibaú. Um dos fatores que contribuiu para esses números foi o desemprego em uma crescente nos últimos anos. A taxa de desemprego atingiu 12,7% da população da região metropolitana de Fortaleza no primeiro trimestre de 2019. Essa

taxa foi considerada a sétima maior entre qualquer outra região metropolitana do país no mesmo período. Segundo dados do PNAD em 2019, a cidade concentrou cerca de 10,8% da população desocupada do país. Em 2021, esse número de desempregados atingiu 14,4%, percentual superior ao recorde histórico do primeiro trimestre de 2019.

De acordo com os dados obtidos pelo PNAD em 2021, a taxa de desemprego no estado refletiu na deterioração do mercado de trabalho em meio à nova Pandemia do Coronavírus. A recuperação gradual da economia, iniciada no primeiro trimestre de 2017 em função da crise 2015-2016, melhorou levemente a taxa de participação do estado do Ceará em 2019. No entanto, a pandemia da Covid-19 reverteu a tendência de recuperação das atividades econômicas, resultando em uma queda significativa na taxa de crescimento em 2020.

2.2.2 Boston

A cidade de Boston, capital de Massachusetts nos Estados Unidos da América, segundo o Censo desse país em 2019, possuía 684.379 mil habitantes, porém, no ano de 2011, esse número chegava a 609.942 mil habitantes. Segundo Melnik (2011), a taxa de desemprego era mais baixa que as taxas nacionais e estaduais, tanto que em janeiro de 2011, essa mesma taxa de desemprego era de 7,8%, dois pontos abaixo da média nacional. Isso ocorreu devido à força de trabalho da cidade ter apresentado um crescimento de 18,5% de 2000 a 2010, com recessões nos anos de 2001 e 2008, a qual elevou a população a dobrar o índice de desemprego.

Na categorização do desemprego, de acordo com Melnik (2011), incluem-se uma porcentagem maior de residentes não brancos que também apresenta um alto índice entre afro-americanos da cidade. Além disso, jovens com menos de 20 anos também fazem parte deste cenário, ocupando 24,5% da taxa de pessoas desempregadas. Em contraste, Boston possui uma das populações mais instruídas entre as principais cidades americanas que, de acordo com o *American Community Survey* de 2010, 44,3% da população adulta em Boston possuía pelo menos um diploma em bacharel, classificando a cidade em 4º lugar entre as 25 maiores cidades do país nesse quesito.

A Agência de Pesquisa da Divisão de Planejamento e Desenvolvimento de Boston (BP&DARD) de 2017 realizou uma pesquisa formalizada em um documento chamado de Perfis dos Bairros. A partir disso foi possível obter informações socioeconômicas de alguns bairros dessa cidade.

Segundo a pesquisa, o bairro de Back Bay fica próximo do centro financeiro da cidade com pouco menos de 2 km de distância. Esse bairro é considerado um dos mais caros da cidade. Dados de 2015 da pesquisa indicam que as principais ocupações dos residentes desse bairro são nas áreas de Administração, Negócios e Finanças. Além disso, para esse mesmo ano, a renda média familiar anual foi de $ 88.469, valor maior do que a renda média de Boston ($ 55.777). A proporção de famílias que possuem carro caiu de 62% em 2000 para 53% em 2015.

O bairro de Beacon Hill também fica próximo do centro financeiro, estando a pouco mais de 1 km de distância. É uma região de predominância residencial. Dados de 2015 da pesquisa indicam que as principais ocupações dos moradores estão nas áreas de Negócios, Finanças, Gestão e Vendas. A renda familiar média em Beacon Hill em 2015 foi $ 93.033, superior à média de Boston de $ 55.777. Além disso, 42% das famílias de Beacon Hill possuem pelo menos um carro, abaixo da média da cidade de 65%.

A Universidade de Boston possui alguns campus. O campus principal está situado às margens do rio Charles, nos bairros Fenway e Allston, enquanto o Boston University Medical Campus fica no bairro de South End. O campus principal, que foi objeto de análise para essa pesquisa, fica um pouco mais distante do centro financeiro da cidade com quase 4 km de distância. Com aproximadamente 4 mil professores e mais de 30 mil alunos, a Universidade de Boston é a quarta maior universidade privada do país e o quarto empreendimento que mais emprega na cidade (BP&DARD, 2017).

O bairro de Fenway também fica um pouco distante do centro financeiro da cidade com pouco mais de 2 km de distância. Além disso, há alguns fatores que atraem muitas pessoas para lá. Um deles é o fato de ser o bairro onde o Boston Red Sox, time de Baseball local, sedia os seus jogos durante a temporada, no famoso Estádio Fenway Park. Outro motivo importante de atração de pessoas é o fato de ser uma região que possui muitos pontos de encontros, como casas noturnas, bares e instituições culturais (BP&DARD, 2017). Nesse sentido, essa região possui 59% de residentes entre 18 e 24 anos de idade, uma porção que se compara aos 17% da cidade para a mesma faixa etária.

Haymarket Square é uma região localizada dentro do bairro North End que fica a menos de 1 km do centro financeiro da cidade. Esse bairro é considerado um dos mais antigos da cidade e devido a isso há vários pontos turísticos. Segundo (BP&DARD, 2017), 81 % dos moradores desse bairro com 16 anos ou mais de idade participam da força de trabalho. Essa taxa é superior a da cidade (68%). Além disso, as principais ocupações dos habitantes são Negócios e Operações Financeiras, Ocupações em Educação, Jurídico, Serviço Comunitário, Artes e Mídia. A renda familiar média em North End em 2015 foi $ 82.965, superior à média

de Boston de $ 55.777. Em 2015, 47% dos domicílios do North End tinham pelo menos um veículo, em comparação com 65% de todos os lares de Boston.

A North Station é uma conhecida estação de metrô subterrânea na cidade e fica a pouco mais de 1 km de distância do centro financeiro. Ela está localizada sob a Haverhill Street e edifícios adjacentes no quarteirão entre a Causeway Street e a Valenti Way. Por estar entre várias regiões, seu fluxo de pessoas é alto.

A Northeastern University (Universidade do Nordeste) é uma universidade que fica próxima do bairro Fenway, considerando seu campus principal. Sua distância do centro financeiro é de pouco mais de 2 km. A área também é conhecida como o Distrito Cultural Fenway.

A South Station também é uma conhecida estação da cidade que fica a menos de 1 km de distância do centro financeiro da cidade. Ela é bastante conhecida por ser multimodal, ou seja, possuir diversos tipos de meios de terminais de transportes, como um terminal rodoviário, um terminal ferroviário, uma estação de metrô e uma estação de ônibus municipais (BP&DARD, 2017).

O Theatre District é um teatro que também fica próximo do centro financeiro com quase 1 km de distância. Esse teatro fica localizado no bairro de Downtown. O bairro possui uma força de trabalho que corresponde a 63% de sua população de 16 anos ou mais de idade. Uma porcentagem um pouco a baixo se comparada com a cidade (68%). As principais ocupações de seus moradores são nas áreas de Gestão, Negócios e Finanças, Preparação de Alimentos e Serviços. A renda média familiar desses moradores em 2015 foi de $ 65.090, superior à média de Boston de $ 55.777. Além disso, 42% dos domicílios possuem pelo menos um carro, abaixo da média da cidade de 65% (BP&DARD, 2017).

Por fim, o bairro de West End fica a pouco mais de 1 km do centro financeiro da cidade. Esse bairro também é considerado uma região nobre, possuindo diversos pontos conhecidos da cidade como o Museu de Ciência, arena esportiva onde recebe famosos jogos de hóquei, basquete, além de grandes shows. Também nesse bairro possui o Museu de West End, famoso por seu grande acervo histórico (BP&DARD, 2017). Além disso, muitos dos trabalhadores que vivem nessa região estão alocados nas áreas de Saúde e Assistência Social. A renda familiar média nesse bairro em 2015 foi de $ 90.694, significativamente maior do que a média de Boston de $ 55.777. Menos da metade (46%) das residências em 2015 tinham um veículo, em comparação com 65% de todos os lares de Boston.

2.3 Ciência de Dados

A Ciência de Dados surgiu como um campo de atuação de competências interdisciplinares devido às novas ideias de acadêmicos estatísticos e à propagação da Ciência da Computação em vários contextos da sociedade. Pode-se atribuir à Ciência de Dados a extração de informação útil a partir de imensas bases de dados complexas, dinâmicas, heterogêneas e distribuídas (BUGNION ET AL, 2017).

De acordo com Provost et. Al (2013), em Ciência de Dados há 3 domínios de conhecimento que se relacionam: Programação de Computadores; Estatística e Matemática; e Domínio do Conhecimento. Neste sentido, existem três pressupostos:

1. Os especialistas em Ciência de Dados devem apresentar habilidades na área de Ciência da Computação, visto que os dados são armazenados, manipulados e transmitidos por computadores. Neste contexto, os ambientes computacionais para o Desenvolvimento de Software são ferramentas essenciais para promover a Curadoria Digital, implementar os algoritmos de Aprendizagem de Máquina e a construção das interfaces de Visualização da Informação. É de suma importância saber utilizar essas tecnologias de modo a acessar e transformar os dados para abstrair e representar informação útil.
2. O conhecimento sobre Matemática e Estatística é necessário para a realização de atividades de Análise de Dados. Ou seja, os profissionais de Ciência de Dados devem entender o funcionamento dos algoritmos de Aprendizagem de Máquina, bem como saber interpretar os resultados estatisticamente. Interdisciplinarmente, a atividade de interpretação é facilitada pela Visualização da Informação, a qual privilegia a utilização de elementos de representação gráfica da informação.
3. O Domínio do Conhecimento do problema deve ser disponível e amplamente utilizado no Processo de Tomada de Decisão. Neste sentido, as soluções de Ciência de Dados são voltadas para a formulação de hipóteses e a aquisição de informação aderente como insumo no processo decisório.

A Ciência de Dados configura-se como um suporte metodológico ao processo de Tomada de Decisão, facilitando a obtenção de informação contextualizada a explicitação de fenômenos ocultos contidos nos dados ou a refutação/confirmação de hipóteses previamente estabelecidas. Esse processo é denominado como Tomada de Decisão Guiada por Dados (PROVOST ET AL, 2013).

2.3.1 Ciclo de vida de Ciência de Dados

Existem várias interpretações do que se pode formalizar sobre um ciclo de vida de Ciência de Dados. As etapas deste ciclo podem variar de acordo com o contexto de aplicação, bem como a ponderação sobre qual ponto de granularidade os cientistas de dados estão dispostos a alcançar (MARTINEZ ET AL, 2021). Nesse sentido, uma abordagem de ciclo de vida nesse âmbito pode ser descrita em 5 etapas nas quais podem subsidiar um caminho a ser traçado por quem deseja realizar um projeto de Ciência de Dados. Essas etapas são descritas a seguir e podem ser representadas pela Figura 5 abaixo.

Figura 5 - Exemplo de um ciclo de vida de Ciência de Dados.

Fonte: (GONÇALVES, 2022).

Primeira Etapa – Entendimento do Problema: É essencial a compreensão do problema e seu contexto. Nesse sentido, uma abordagem que pode ser adotada é a procura de respostas para algumas perguntas chaves que podem ser elencadas por essas: Quanto ou quantos? Qual categoria? Qual grupo? Está se comportando de maneira diferente para o contexto? Qual opção deve ser adotada?. Nessa etapa, também é importante identificar os objetivos centrais, obtendo informações sobre as variáveis que precisam ser previstas.

Segunda Etapa – Coleta de Dados: Essa etapa consiste na seleção dos dados pertinentes para o projeto, bem como o domínio ao qual pertencem. Além disso, nessa etapa é adotada a

estratégia de obtenção e a maneira mais eficiente de armazenar os dados obtidos. Isso é preciso porque muitas vezes os dados precisam ser tratados e até mesmo criados em algumas ocasiões.

Terceira Etapa – Processamento de Dados: Essa etapa engloba a limpeza e o tratamento dos dados obtidos da etapa anterior. Muitas vezes essa fase demanda mais tempo devido a existência de vários cenários possíveis que precisam, geralmente, serem filtrados. Além disso, existe a possibilidade de haver inconsistências em uma coluna de uma tabela dos dados, bem como inconsistências em tipos de dados e que precisam ser tratados. Também nessa etapa são vistas questões relacionadas aos dados ausentes e que podem gerar muitos erros na criação de modelos de Aprendizagem de Máquina, além da necessidade de realizar análises extras que envolvam estatística ou outros tipos de medidas matemática.

Quarta Etapa – Exploração de Dados: Também chamada de Análise Exploratória de Dados (AED), essa etapa consiste na análise sobre os dados tratados e limpos provenientes da etapa anterior. Pode-se deduzir que essa etapa é um tipo de *brainstorming* da análise de dados, tornando possível o entendimento de possíveis padrões nos dados considerados. Uma das maneiras mais utilizadas para essa análise é por meio de gráficos, como histogramas, gráficos de linhas, séries temporais, curvas de distribuição para ver tendências gerais, etc. A partir dessa etapa é possível formar hipóteses sobre os dados.

Quinta Etapa – Comunicação de Resultados: Essa etapa consiste no início da criação de valor para os interessados dos dados em estudo. Nesse sentido, é nessa etapa que os Cientistas de Dados demonstram a capacidade de externar da maneira mais clara possível os resultados obtidos para que isso possa auxiliar na tomada de decisão.

Sexta Etapa – *Feedback*: Essa etapa trata da troca de informações entre os envolvidos em um projeto de Ciência de Dados. Nessa fase há possibilidade de rever os passos anteriores, podendo acrescentar mais análises ou retificar alguma etapa. É também nessa fase que o valor criado é utilizado em benefício dos usuários no intuito de respaldar decisões em vários em vários seguimentos.

De acordo com Hotz (2022), as etapas de um Ciclo de Vida de Ciência de Dados não são exaustivas, sendo reduzidas ou ampliadas a depender do contexto de aplicação. Além disso, cada etapa pode ser substituída por outra em termos de sequência ou maneira de executar, haja vista que o domínio de aplicação pode se adaptar melhor se as etapas forem postas de outra maneira.

2.3.2 Aprendizagem de Máquina

Como descrito na seção 2.3, a Estatística é um dos pilares da Ciência de Dados. Entretanto, ela não é suficiente para extrair informações úteis das grandes bases de dados que a maioria dos negócios possuem atualmente. Nesse contexto surge a Aprendizagem de Máquina (AM) que pode ser conceituada como um processo algorítmico que pode aprender à medida que consome mais dados (QUICK, 2020). A Ciência de Dados manipula grandes quantidades de dados e esses algoritmos são cada vez mais importantes nesse contexto. Pode-se dizer que a Aprendizagem de Máquina auxilia no processo de automação do processo de descoberta de padrões desconhecidos de um grande volume de dados, além de propiciar uma elevação da qualidade e do desempenho da Ciência de Dados, possibilitando um novo modo de gerenciamento (VASCONCELOS ET AL, 2017).

Um dos objetivos da AM é desenvolver informações eficientes de reconhecimento de padrões que serão capazes de generalizar além dos exemplos de um conjunto de treinamento, para que possam – de acordo com o problema em questão e os dados disponíveis para análise – obter bons resultados (ROZA, 2016). AM pode ser empregado em diferentes campos, como pesquisas na web, filtragem de spam, sistemas de recomendação, publicidade, detecção de fraude, classificação de imagem, etc.

Para entender como a AM gera conhecimento e aprende a partir de padrões e dados, precisa-se entender a estrutura desse tipo de aprendizado, que poderá ser por um processo de indução. Isso significa uma forma de inferir lógica para obter conclusões gerais sobre um determinado conjunto de exemplos. A indução é o recurso mais comum usado pelo cérebro humano para adquirir novos conhecimentos (MONARD ET AL, 2003).

Nesse contexto, para que determinado conceito seja aprendido na indução, várias hipóteses de conhecimento serão geradas no exemplo analisado, podendo essas hipóteses geradas ser verdadeiras ou não. O processo de indução é apropriado para uma pequena quantidade de dados. Além disso, se uma amostra de dados não for devidamente selecionada, assim como as variáveis (atributos), a hipótese obtida pode ser de pouco valor. Assim, na AM precisa-se receber um grande número de amostras para poder aprender e extrair as informações relacionadas ao problema a ser resolvido. As variáveis neste conjunto de dados agregam valor e geram um maior número de hipóteses para um determinado algoritmo aprender (CARVALHO ET AL, 2011).

Um dos objetivos da Aprendizagem de Máquina é desenvolver métodos eficazes de reconhecimento de padrões que possam generalizar exemplos de conjuntos de treinamento para que possam ser usados com conformidade (ROZA, 2016). Nesse sentido, pode-se citar duas possibilidades de aprendizagem: não supervisionado e supervisionado. De modo geral, a aprendizagem não supervisionada não utiliza informações das variáveis de saída. Ou seja, não existem resultados pré-definidos para o modelo utilizar como referência para aprender. Por outro lado, na aprendizagem supervisionada há um supervisor. O supervisor é dado registrando um valor da variável de saída. Esse valor é a variável que pretende prever a partir dos dados existentes. Como resultado, é obtido um modelo que descreve o conjunto de dados utilizado e espera-se prever o comportamento de saída de novos valores.

Nesse contexto, a regressão é uma das técnicas que permite investigar e compreender a relação entre variáveis de resposta e variáveis específicas por meio da construção de modelos supervisionados. A análise de regressão pode ser usada como um método descritivo de análise de dados para vários fins, como descrever a relação entre as variáveis para entender um processo, prever um valor de uma variável a partir do valor de outras variáveis e observar o valor de uma outra variável. Ademais, pode-se ainda controlar o valor de uma variável em um intervalo de interesse. A seguir, são descritos alguns tipos de regressão.

1. Regressão Linear

No contexto da Aprendizagem de Máquina, a Regressão Linear é um algoritmo supervisionado usado para predição de uma variável alvo contínua. Nesse sentido, a Regressão Linear simples é a menor estimativa quadrática de um modelo de Regressão Linear com uma única resposta. Em outras palavras, essa regressão tem por objetivo obter uma equação matemática da reta que represente o melhor relacionamento numérico linear entre o conjunto de pares de dados em amostras selecionadas de dois conjuntos de variáveis (AMARAL, 2016).

A Análise de Regressão é usada principalmente para modelagem preditiva e permite estimar as relações entre as variáveis, além de ser um dos métodos de formulação de modelos estatísticos. Portanto, é uma forma de modelagem preditiva que analisa a relação entre as variáveis dependentes (γ) e as variáveis independentes (X), para encontrar conexões entre elas por parâmetros β. Esta relação pode ser demonstrada pela equação (1) abaixo:

$$\gamma \sim f(X, \beta) \tag{1}$$

A função representada por *f* pode assumir várias formas e depende essencialmente do número de variáveis independentes, do tipo de variável dependente e da forma de distribuição dos dados a serem expressos pela linha de regressão. Adicionalmente, os modelos mais comuns nesse contexto são a Regressão Linear e Regressão Logística. A Regressão Logística é usada quando γ é binário ou dicotômico e X é categórico ou não categórico por natureza. Por outro lado, na Regressão Linear, a variável dependente é contínua e a variável independente pode ser contínua ou discreta, sendo que a linha de regressão é linear.

2. Regressão SGD

No contexto da Aprendizagem de Máquina, o processo de classificação pode ser visto como a determinação de uma função *f* que mapeia amostras para uma determinada classe. Uma função de perda associa um custo às inferências erradas nesse tipo de atividade. Logo, surge a necessidade por um coeficiente de *f* que minimize a margem de custo. O Gradiente Descendente (SGD) é um algoritmo que busca cumprir essa missão, testando diferentes valores para o coeficiente em questão e analisando o custo para cada classificação do conjunto de treinamento (BARRADAS ET AL, 2015).

Por ser uma tarefa dispendiosa, já que o valor do coeficiente só é atualizado depois de cada iteração em que toda a base de treinamento foi percorrida, o SGD surge como uma alternativa para contornar esse problema, atualizando o coeficiente para cada amostra do conjunto e não apenas no fim das iterações (DIAB, 2019).

3. Árvore de Decisão

Na perspectiva da Aprendizagem de Máquina, Árvores de Decisão são algoritmos que podem ser utilizados para classificar dados. Podem ser usadas em conjunto com a indução de regras e apresentar os resultados hierarquicamente, ou seja, com priorização. Nelas, um atributo importante é apresentado na árvore como o primeiro nó, e os atributos menos relevantes são mostrados nos nós subsequentes (GAMA, 2021). Uma vantagem das Árvores é a tomada de decisões levando em consideração os atributos mais relevantes, além de serem de fácil compreensão para as pessoas. Ao escolher e apresentar os atributos em ordem de importância, as Árvores de Decisão permitem aos usuários conhecer quais fatores mais influenciam os seus dados (GARCIA, 2000).

As Árvores de Decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados (GARCIA, 2000). Uma Árvore pode utilizar uma estratégia chamada "dividir-para-conquistar". Nessa estratégia, um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (GAMA, 2002). Nesse sentido, a capacidade de discriminação de uma Árvore de Decisão advém das características de divisão do espaço definido pelos atributos em subespaços e da associação de uma classe a cada subespaço.

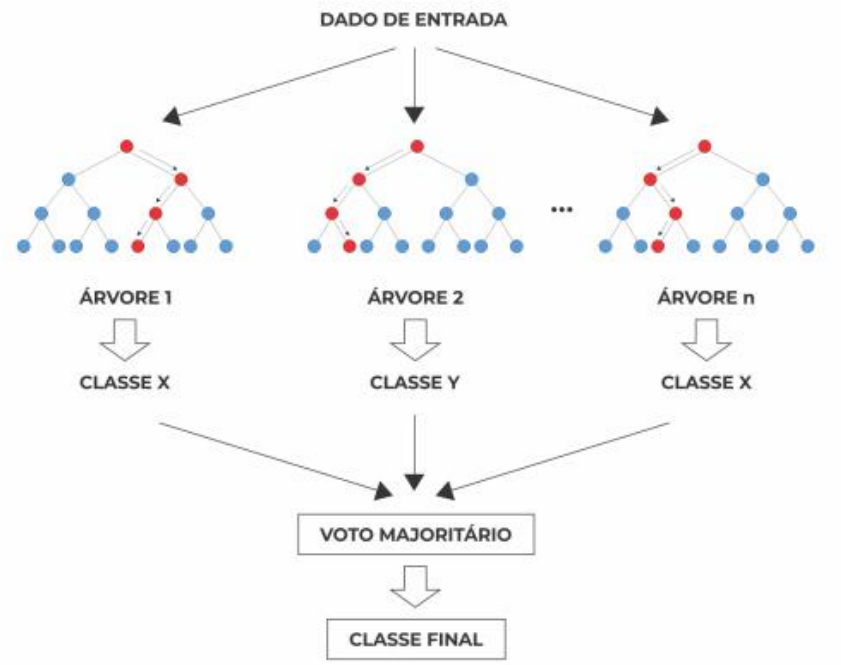
Segundo Garcia (2000), as Árvores consistem de: nodos (nós), que representam os atributos, e de arcos (ramos), provenientes desses nodos e que recebem os valores possíveis para esses atributos (cada ramo descendente corresponde a um possível valor desse atributo). Nas árvores existem nodos folha (folha da árvore) que representam as diferentes classes de um conjunto de treinamento, ou seja, cada folha está associada a uma classe. Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

4. Floresta Aleatória

No âmbito da Aprendizagem de Máquina, uma Floresta Aleatória (*Random Forest*) pode ser definida como um método de criar várias Árvores de Decisão e combiná-las para obter previsões de maior precisão e estabilidade. Dessa forma, podem ser utilizadas para classificação e regressão. Sua operação é alcançada pela definição de muitas Árvores de Decisão, sendo cada uma diferente das demais e responsável por um conjunto diferente de regras (DONGES, 2018).

Quando há dados de entrada, todas as Árvores classificam os dados selecionando-os aleatoriamente e, assim, são capazes de fornecer seus resultados, definindo a classificação final. Dessa maneira, o resultado é dado pelo modelo estatístico de todas as categorias dos resultados ou suas médias numéricas (SOUZA, 2017). Na Figura 6 é ilustrado um *Random Forest* genérico com *n* Árvores.

Figura 6 - Random Forest genérico.



Fonte: Adaptação de (DIMITRIADIS ET AL, 2018).

Observa-se que existem *n* Árvores com diferentes classes em que há uma convergência mediante um voto majoritário. Dessa forma, é possível obter uma classe final.

2.4 Técnicas de Avaliação de Desempenho

Para avaliar o desempenho de algoritmos é importante o uso de técnicas de avaliação de desempenho para verificar se os resultados das predições são satisfatórios. Algumas destas técnicas são apresentadas a seguir.

1. K-Fold Cross Validation

Cross Validation (CV) é uma técnica muito utilizada para avaliação de desempenho em modelos de Aprendizagem de Máquina (DA SILVA LOCA, 2020). O CV consiste em particionar os dados em conjuntos (partes), em que um conjunto é utilizado para treino e outro conjunto é utilizado para teste e avaliação do desempenho do modelo. A utilização do CV possibilita detectar se o modelo está ajustado de maneira excessiva aos dados de treinamento, ou seja, sofrendo *overfitting*. *Overfitting* pode ser traduzido como um “sobre ajuste”. É um

termo usado para descrever quando um modelo se ajusta em demasia em relação a um conjunto de dados anteriormente observado.

K-Fold envolve a divisão aleatória da base de dados em K subconjuntos (onde K é definido anteriormente), cada um dos quais com aproximadamente o mesmo tamanho de amostra. Em cada iteração, k-1 subconjunto é usado para o conjunto de treinamento e o subconjunto restante como o conjunto de teste. Dessa forma, gera-se resultados de medição para avaliação. Este processo garante que cada subconjunto será usado para teste em algum ponto durante a avaliação de um modelo (DA SILVA LOCA, 2020).

2. Z – Score

Identificar observações como *outliers* – dados que se distanciam radicalmente de todos os outros – depende do pressuposto de que os dados são normalmente distribuídos. Nesse sentido, para conjuntos de dados univariados, assume-se que eles seguem uma distribuição normal. Se não apresentarem uma distribuição normal, então a determinação dos *outliers* pode ser devido a não normalidade, ao invés de serem realmente *outliers* (FILIBEN, 2013).

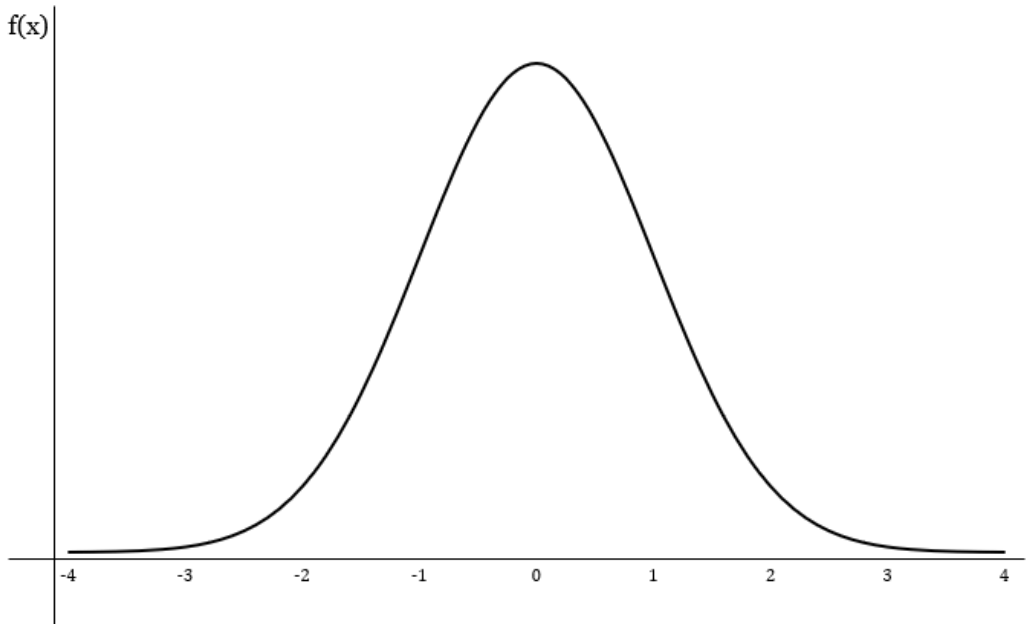
Para identificar *outliers*, o Z – Score fornece a indicação da distância numérica entre um ponto e a média da amostra. Esta distância é baseada no desvio padrão. Nesse contexto, o Z – Score pode ser considerado um método de descrever um ponto com base na relação entre a média e o desvio padrão deste ponto e um conjunto de pontos, ambos do mesmo espaço amostral. Ele mapeia os dados para uma distribuição com a média definida como 0 e o desvio padrão definido como 1.

Segundo Gorrie (2016), um outro objetivo do Z – Score é mitigar a influência da localização e do tamanho dos dados, permitindo a comparação direta entre diferentes bases de dados. Como os dados são centralizados e redimensionados, qualquer ponto que esteja muito longe de zero pode ser considerado um *outlier*, ou seja, dados que se distanciam. O valor crítico do score Z aceito na literatura é um valor menor que -3 ou maior que +3.

2.5 Análise Estatística

A Análise Estatística pode ser definida como uma ciência para coletar, explorar e apresentar dados no intuito de descobrir padrões, além de possibilitar encontrar tendências (SANTOS, 2022). Nesse contexto, alguns métodos nesse âmbito podem ser utilizados para averiguar se algumas amostras possuem características de determinadas distribuições. Uma distribuição bastante conhecida é a Distribuição Normal que é uma das mais importantes em estatística. Ela é uma distribuição de probabilidade contínua e simétrica em ambos os lados da média, de modo que o lado direito é uma imagem espelhada do esquerdo. O Gráfico 2 representa o formato padrão da curva de uma Distribuição Normal.

Gráfico 2 - Curva padrão de uma Distribuição Normal.



Fonte: Autor (2022).

Os valores dos dados nessa distribuição, em sua maioria, tendem a se agrupar em torno da média. As caudas são assintóticas, ou seja, teoricamente, se estende de – infinito a + infinito, sem tocar o eixo horizontal x. Além disso, para uma distribuição perfeitamente normal, a média, mediana e moda terão o mesmo valor, visualmente representados pelo pico da curva do Gráfico 2 (GONÇALVES, 2014).

Na prática, muitos dados contínuos exibem essa curva quando representados graficamente. Um exemplo seria a seleção aleatória de 100 indivíduos em que seria esperado

observar essa distribuição para muitas variáveis contínuas, como altura, pressão arterial e quociente de inteligência.

Nessa perspectiva, muitos testes estatísticos dependem dessa distribuição, como os testes paramétricos. Entretanto, para usá-los, os dados devem apresentar Distribuição Normal. Caso isso não aconteça, é preciso corrigir essa falta de normalidade ou usar testes não paramétricos (WAINER, 2022). Assim, um tipo de teste para verificar a normalidade de uma distribuição é o teste de Normalidade de Kolmogorov-Smirnov (K-S).

De acordo com Lopes at et (2013), o teste K-S pode fornecer o parâmetro valor de prova (valor-p, p-value ou significância), que pode ser interpretado como a medida do grau de concordância entre os dados e a hipótese nula (H0), sendo H0 correspondente à Distribuição Normal. Quanto menor for o valor-p, menor é a consistência entre os dados e a hipótese nula. Nesse sentido, a regra de decisão adotada para saber se uma distribuição é Normal ou não é rejeitar H0. Em outras palavras, considerando α como um valor arbitrado de confiança, se valor-p $\leq \alpha$, rejeita-se H0, ou seja, não se pode admitir que o conjunto de dados em questão tenha Distribuição Normal. Por outro lado, se valor-p $> \alpha$, não se rejeita H0, ou seja, a distribuição Normal é uma distribuição possível para o conjunto de dados em questão.

Nessa perspectiva, é importante saber que o estudo da normalidade dos dados é um pré-requisito para análises futuras. Diante disso, é possível escolher outros testes que servem para analisar comportamentos das amostras, sendo estes testes diferentes a depender do comportamento anteriormente encontrado: ser uma Distribuição Normal ou não. Caso os dados em estudo sejam do tipo Distribuição Normal, é possível utilizar testes como T-Student e as análises fatoriais exploratória. Caso não sigam uma Distribuição Normal, então é recomendado utilizar testes que satisfaçam as exigências desses dados (BOGONI, 2022).

Diante desse contexto, os testes não paramétricos são testes de hipótese que não requerem que a distribuição da população seja caracterizada por determinados parâmetros. Ou seja, alguns testes de hipóteses pressupõem que uma amostra segue uma Distribuição Normal com parâmetros de média e desvio padrão. Os testes não paramétricos não têm essa suposição, de forma que eles são úteis quando os dados são fortemente não normais (CONTADOR ET AL, 2016). Estudos com amostras pareadas são comuns em diversas áreas do conhecimento e consiste em realizar mais de uma medida em uma mesma unidade amostral. Dessa forma, é possível verificar se houve diferença entre essas medidas, em que a primeira informação será pareada com a segunda informação, com a terceira e assim sucessivamente (CAPP ET AL, 2020).

Nesse sentido, o teste de Wilcoxon serve para verificar se existe diferença significativa de uma variável numérica entre dois grupos de interesse, comparando as medianas. Ele é um teste não-paramétrico em que as amostras precisam ser dependentes. Segundo Moreno et al (2020), no teste de Wilcoxon são calculados os valores numéricos da diferença entre cada par, sendo possível três condições: aumento (+), diminuição (-) ou igualdade (=). Uma vez calculadas todas as diferenças entre os valores obtidos para cada par de dados, essas diferenças são ordenadas pelo seu valor absoluto (sem considerar o sinal), substituindo-se então os valores originais pelo posto que ocupam na escala ordenada. Assim, o teste da hipótese de igualdade entre os grupos é baseado na soma dos postos das diferenças negativas e positivas. Nesse sentido, de acordo com os autores, este teste considera o valor dessas diferenças ao invés de considerar apenas o sinal das diferenças entre os pares, tornando-o um dos melhores testes não paramétricos nesse contexto.

O objetivo do teste dos sinais de Wilcoxon é comparar as performances de cada sujeito (ou pares de sujeitos) no sentido de verificar se existem diferenças significativas entre os seus resultados nas duas situações. Os resultados da situação B são subtraídos dos da situação A e a diferença resultante (d) é atribuído o sinal mais (+) ou, caso seja negativa, o sinal menos (-). Estas diferenças são ordenadas em função da sua grandeza (independentemente do sinal positivo ou negativo). O ordenamento obtido é depois apresentado separadamente para os resultados positivos e negativos. O menor dos valores deste segundo, dá-lhe o valor de uma "estatística" designada por W, que pode ser consultada em uma tabela de significância apropriada (MORENO ET AL, 2020).

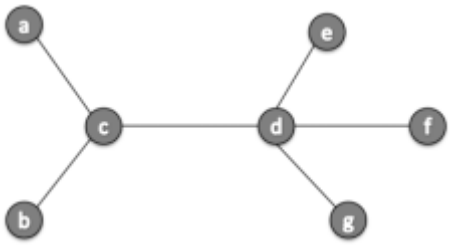
2.6 Grafos

Pode-se definir grafos como uma forma de abstração que pode ser utilizada para representar vários comportamentos do mundo, como trajetos, redes em vários âmbitos de conhecimento, problemas de otimização, etc. (BORBA, 2013). Nesse sentido, um grafo G pode ser representado por um par ordenado (V, E), em que V é um conjunto de vértices e E é um conjunto de arestas. Cada aresta e pertencente ao conjunto E pode ser denotado por $e = (v, w)$ sendo este um par de vértices. Os vértices v e w são os extremos da aresta e são denominados

vértices adjacentes ou vizinhos. A aresta e é dita incidente a ambos os vértices v e w, ou seja, a aresta parte de um e chega ao outro vértice.

Existem algumas classificações quanto aos tipos de grafos. Um deles é o grafo não direcionado em que a relação (v, w) é simétrica. Neste caso, existe uma aresta direcionada que une v e w, sendo que o contrário também ocorre. Devido a isso, pode-se dizer que as arestas que ligam os vértices não possuem orientação. A Figura 7 representa um exemplo de grafo não direcionado formado pelo conjunto de vértices $V = \{a, b, c, d, e, f, g\}$. Cada par de vértices que se conectam formam as arestas pertencentes ao conjunto $E = \{(a, c), (b, c), (c, d), (d, e), (d, f), (d, g)\}$.

Figura 7 - Exemplo de um grafo não direcionado.



Fonte: Autor (2022).

Outro tipo de classificação para grafos são os ditos conexos. Um grafo $G(V, E)$ é conexo quando existir um caminho entre cada par de vértices, de outro modo, é dito desconexo. Assim, se existir pelo menos um par de vértices que não esteja conectado através de um caminho, o grafo é desconexo. Na Figura 8 é representado um grafo conexo em (a), pois pode-se encontrar um caminho entre quaisquer pares de vértices escolhidos. Em (b) isso não acontece, pois se for escolhido o nó 3 e o nó 1, verifica-se que não existe nenhum caminho conectando esse par de nós.

Figura 8 - (a) Grafo conexo. (b) Grafo desconexo.

Fonte: Autor (2022).

Há pesquisas que indicam a importância do estudo dos grafos no âmbito dos transportes. Um exemplo disso, são estudos sobre transporte público que tiveram como um dos objetivos

investigar propriedades estatísticas por meio da Teoria de Grafos. Numa análise mais detalhada, foram dedicados esforços na tentativa de compreender e propor topologias que diminuíssem algumas vulnerabilidades de redes de transportes. Dessa forma, seria possível evidenciar mecanismos de crescimento dessas redes, bem como vulnerabilidades diante de ataques e situações de falhas em cascata nesse contexto (SOUSA, 2016).

2.6.1 Medidas de Centralidade

No âmbito de aplicação de grafos em redes, surgem as Medidas de Centralidade. Em uma rede, os vértices mais centrais são aqueles a partir dos quais pode-se atingir qualquer outro com mais facilidade ou rapidez. As Medidas de Centralidade identificam a posição de um ponto relativamente a outros na sua rede. Nesse sentido, a centralidade pode traduzir uma ideia de poder. Assim, quanto mais central um ponto de interesse, maior influência e poder terá na sua rede (GAMA, 2012).

Há vários tipos de Medidas de Centralidade. Pode-se destacar a Centralidade de Grau, a Centralidade de Proximidade e a Centralidade de Intermediação que são amplamente utilizadas em vários contextos de grafos que envolvem redes (BORBA, 2013).

A Centralidade de Grau pode ser conceituada como o número de contatos diretos que um vértice possui. Um ponto de interesse que se encontra numa posição que permita o contato direto com muitos outros é visto pelos demais como um canal maior de informação, razão pela qual ser dito mais central. Dessa forma, a Centralidade de Grau é a contagem do número de adjacências de um vértice.

Seja G um grafo qualquer (conexo ou não) com n vértices e seja x um vértice de G, então a Centralidade de Grau (Degree (D)) de x denotada por $\sigma_D(x)$, é o número de arestas incidentes a x. Sendo uma matriz de Adjacência (A) uma matriz booleana com colunas e linhas indexadas pelos vértices desse grafo, então, por meio dessa matriz A e do grafo G, tem-se que:

$$\sigma_D(x) = \sum_{i=1}^{n} a_{ix} \tag{2}$$

Onde a_{ix} são elementos da matriz de Adjacência A(G).

Por outro lado, a Centralidade de Proximidade consiste na soma das distâncias de um vértice em relação aos demais vértices do grafo. Nesse contexto, não estar longe demais dos restantes dos elementos do grafo é mais importante que ter muitas ligações.

Assim, considerando um grafo G conexo com n vértices e seja x um vértice de G. A Centralidade de Proximidade de x é dada pelo inverso da soma das distâncias de x a todos os outros vértices do grafo. Dessa forma, tem-se que:

$$\sigma_c(x) = \frac{1}{\sum_{i=1}^{n} d_G(x,i)} \tag{3}$$

Onde $d_G(x,i)$ se refere à distância de x a i.

Por fim, a Centralidade de Intermediação permite medir a capacidade que um ponto de interesse tem de poder influenciar os seus pares em uma rede. Nesse sentido, um nó importante faz parte de muitos caminhos, como por exemplo em um ponto que é integrado em várias rotas, pois provavelmente terá vantagens estratégicas a depender do contexto de aplicação.

Desse modo, considerando um grafo G (conexo ou não) com n vértices e seja x um vértice de G, então a Centralidade de Intermediação de x pode ser formalizada por:

$$\sigma_B(x) = \sum_{i=1}^{n} \sum_{i<j}^{n} \frac{g_{ij}(x)}{g_{ij}}, i, j \neq x \tag{4}$$

Onde g_{ij} significa o número de caminhos mínimos do vértice i ao vértice j e $g_{ij}(x)$ indica a quantidade desses caminhos mínimos que passam por x.

2.7 Trabalhos Relacionados

Esta seção descreve alguns estudos que buscaram relacionar características socioeconômicas de algumas regiões com alguns aspectos do serviço de viagem da empresa Uber. Esses aspectos foram a acessibilidade ao serviço de viagem, o tempo de espera do serviço de viagem e o preço do serviço de viagem.

No âmbito dos transportes, a acessibilidade pode ser entendida como a possibilidade das pessoas se locomoverem sem obstáculos físicos e financeiros (WANG ET AL, 2018). Nesse

sentido, trata-se de algo relacionado não apenas ao acesso físico aos veículos, mas também aos valores cobrados pelo serviço.

De acordo com Wang et al (2018), a acessibilidade não reflete apenas o desenvolvimento espacial, a rede de transporte e sua distribuição, mas também pode ser interpretada como uma medida temporal. Os autores informam que a medida do tempo pode ser mais sensível do que as medidas baseadas em um local, ao tentar espelhar restrições demográficas, sociais, econômicas e culturais.

Nesse contexto, segundo Wang et al (2018), estudos têm utilizado o tempo de viagem como uma medida comparativa para entender o desiquilíbrio ou equilíbrio entre emprego e moradia, bem como disparidades raciais, econômicas e gênero em áreas urbanas.

Considerando essa perspectiva, Wang et al (2018) delinearam sua pesquisa em duas questões: o tempo de espera como um *proxy* (intermediário) para uma medida de acessibilidade no serviço de viagens da Uber; e, considerando a Uber como uma infraestrutura de transporte virtual, levanta a questão se a empresa está relacionada à polarização socioespacial em um bairro ou um acesso mais equitativo, independentemente dos perfis socioeconômicos.

Para alcançar as respostas dos questionamentos, Wang et al (2018) utilizaram a API de desenvolver da Uber. Os autores focaram nas informações relacionadas a tempo de espera estimada das viagens na cidade de Atlanta, Estados Unidos da América, durante um mês no ano de 2016. A pesquisa centralizou os dados para dois modelos de serviço de viagens da empresa: UberX e UberBlack. Esses modelos são diferenciados por capacidade e preço. O UberX é um serviço mais acessível, de baixo custo e é o mais popular. Por outro lado, UberBlack é uma opção premium, com preços mais elevados e deve atender a algumas especificidades para o veículo e motorista. Dessa forma, a pesquisa coletou dados de tempo estimado de espera para esses dois tipos de serviços a fim de fornecer uma relação entre acessibilidade de viagens da Uber e perfis socioeconômicos por bairros da cidade de Atlanta.

Os tempos de espera estimados foram coletados considerando pontos específicos de cada bairro a cada 30 minutos durante um mês, totalizando cerca de 360 mil pontos de dados. Já os dados socioeconômicos levaram em consideração pesquisas a nível de bairro que evidenciaram informações sobre densidade populacional, tempo médio de deslocamento para o trabalho, taxa de desemprego, ter veículo ou não ter e renda per – capita. Também incluíram uma taxa que denominaram de Taxa de Minoria que representa a relação entre o número da população negra de um bairro e o número total de população do mesmo bairro. Adicionalmente, coletaram informações acerca do transporte público, além de informações sobre as estradas que possibilitou realizar o cálculo da densidade viária. Nesse sentido, as amostras finais, contendo

dados de tempo de espera estimado e dados socioeconômicos por bairro, contiveram 101 amostras.

Wang et al (2018) empregaram quatro variáveis dependentes em modelos de regressão para explorar as relações entre disparidades socioeconômicas e acessibilidade ao Uber. Assim, consideraram o valor médio dos tempos de espera estimados como uma expectativa e o desvio padrão para medir a variabilidade. Nesse sentido, estabeleceram o tempo médio de espera por bairro para cada tipo de serviço de viagem da Uber (UberX e UberBlack) e a partir das amostras coletadas foi calculado o valor de tendência central de acessibilidade, que reflete a média de tempo de espera para aquele serviço naquele bairro. Nessa perspectiva, os usuários da Uber em bairros mais acessíveis teriam um tempo de espera menor quando solicitassem o serviço. Da mesma forma, o desvio padrão da estimativa dos tempos de espera para cada bairro por tipo de serviço foi calculado como medida de variabilidade de acessibilidade, indicando a dispersão ou a flutuação dos tempos de espera. Um bairro mais acessível é menos variável à flutuação dos tempos de espera e devido a isso possui um desvio padrão menor.

Wang et al (2018) ao utilizar quatro variáveis dependentes, consideraram elas como sendo: média da estimativa de tempo de espera e desvio padrão de espera estimado. Essas variáveis foram consideradas para cada tipo de serviço da Uber em estudo, por isso o total se deu em um valor igual a quatro. Em relação a medição das disparidades socioeconômicas, os autores utilizaram a média da renda per – capita dos usuários por bairro como *proxy* da riqueza e a taxa de minoria como *proxy* de raça. Além disso, os autores incluíram as chamadas variáveis de controle, pois a acessibilidade pode estar relacionada a mais fatores do que apenas os *proxies* hipotéticos citados. Essas variáveis são: densidade populacional, tempo médio de deslocamento ao trabalho, taxa de desemprego, taxa de ter ou não ter veículo, número de pontos de transporte público e densidade da rede viária.

Wang et al (2018) utilizaram um *framework* de análise de regressão adaptado de Anselin (2013). Primeiro ajustaram os dados com um estimador de mínimos quadrados ordinários. Em seguida, uma matriz de pesos foi padronizada para que os resíduos do estimador fossem examinados no intuito de obter uma correlação espacial com auxílio do índice de Moran Global. Depois, utilizaram dois testes de Multiplicador de Lagrange para averiguar a dependência espacial e ajudar a selecionar ou o modelo de erro espacial ou o modelo de atraso espacial. Os testes indicaram que modelo de atraso espacial deveria ser utilizado em todas as variáveis dependentes em estudo. Assim, o modelo de atraso espacial é estimado utilizando estimadores de máxima verossimilhança.

Os resultados de Wang et al (2018) indicaram que, para a UberX, o tempo médio de espera estimado é entorno de 3 a 10 minutos, com desvio padrão em torno de 1 a 3 minutos. Para a UberBlack, o tempo médio de espera estimado é entorno de 3 a 13 minutos, com desvio padrão em torno de 1 a 3 minutos. Os autores ressaltam que, para a UberX, a média possuir uma distribuição mais concentrada com um valor médio menor do que o outro serviço não é surpreendente, pois o UberX é um serviço mais popular e econômico, o que provavelmente resulta em mais serviços desse tipo no mercado. Além disso, UberBlack apresentou um custo de pelo menos 3 vezes maior por minuto se comparado ao serviço UberX e 4 vezes maior em custo por milha. Os autores também evidenciaram que nem o valor da renda per – capita, nem a taxa de minoria exercem um impacto significativo na acessibilidade da Uber. Em outras palavras, não há evidências que o serviço de viagens da Uber está associado ao agravamento de polarização socioespacial nos bairros de Atlanta. Pelo contrário, geralmente oferece acesso equitativo a todos os bairros, independentemente dos perfis socioeconômicos (exemplificados por renda per – capita e raça).

Por outro lado, Wang et al (2018) evidenciaram que a densidade populacional e a densidade da rede viária estão associadas a uma melhor acessibilidade do serviço de viagens da Uber nos bairros daquela cidade. Por exemplo, um aumento de 1% na densidade da rede viária corresponde a uma diminuição de cerca de 0,15% no tempo médio de espera da Uber, com redução de cerca de 0,20% no desvio padrão correspondente a esse serviço. Entretanto, afeta de maneira negativa a acessibilidade o comportamento de deslocamentos, como o tempo médio de viagem para o trabalho. Um aumento de 1% no tempo de viagem para o trabalho está associado a 0,54% a mais no tempo médio de espera do serviço de viagem da Uber.

De acordo com Wang et al (2018), para os bairros da cidade de Atlanta, a acessibilidade ao serviço de viagens da Uber não é restrita aos brancos e ricos. Maior densidade de rede viária e maior densidade populacional estão associados a uma melhor acessibilidade do serviço de viagens da Uber. Isso revela o quão é importante a estrutura urbana na acessibilidade de serviços desse seguimento, pois define onde os carros podem ir. Além disso, uma rede viária mais densa pode significar melhor conexão em bairros. Assim, a média do tempo de espera estimado e o desvio padrão tendem a serem menores nessas proximidades. Consequentemente, a área será mais acessível aos serviços da Uber. Além disso, quanto maior o tempo médio de viagem para o trabalho, pior a acessibilidade do Uber. Isso acontece porque esse tempo possui relação com o equilíbrio entre o emprego, a moradia e ao congestionamento na hora do *rush.* Longo tempo de viagem ao trabalho implica uma distribuição relativamente desequilibrada da proporção do número total de empregos para contagem de domicílios e sistemas de transportes ineficientes,

em que o tempo de espera dos serviços Uber pode ser usado para delinear tal comportamento e funções urbanas (WANG ET AL, 2018).

De acordo com Wang et al (2018), algumas limitações da pesquisa são levantadas. A primeira está na unidade de análise que foi considerada por bairro de uma cidade. Estudos em outra escala pode não chegar as mesmas conclusões. Devido a isso, considerando o contexto geográfico, pode-se utilizar modelos multiníveis e de nível misto. Outra limitação seria o estudo ter concentrado as análises em uma única cidade. Como a Uber existe em vários locais do mundo, seria interessante explorar como fatores culturais e políticos afetam as disparidades espaciais de acessibilidade. Além disso, existe a limitação da representação da raça por uma única variável: cor. Isso pode não refletir outras facetas das disparidades sociais. Por último, há a limitação dos dados coletados que são de 2016. Seria importante revisar o trabalho para dados mais atuais.

Estudos indicam que a habitabilidade tem se mostrado como um atributo chave para a sociedade e para o planejamento urbano em todo o mundo (BEZERRA, 2019). Nesse sentido, ela foi incluída como um dos princípios da Nova Agenda Urbana (NAU) adotada pela Organização das Nações Unidas (ONU). A NAU representa um compromisso de política internacional em apoio aos Objetivos de Desenvolvimento Sustentável (ODS).

De acordo com Bezerra et al (2019), habitabilidade pode ser definida como um conjunto de atributos de um lugar, englobando aspectos como moradia, vizinhança e região que contribuem para a qualidade de vida e bem-estar dos moradores. Nesse sentido, pode-se elencar alguns indicadores importantes de habitabilidade que são concebidos considerando vários domínios políticos. Esses indicadores são: ambiente natural, crime e segurança, educação, emprego e renda, saúde e serviços sociais, habitação, lazer e cultura, alimentação e outros bens, espaço público aberto, transportes, coesão social e democracia local.

Além disso, segundo Bezerra et al (2019), uma cidade habitável requer que as necessidades dos habitantes estejam alinhadas com a infraestrutura e ecossistemas construídos que fornecem os bens e serviços dos quais a vida e os meios de subsistência na cidade dependem. Nessa perspectiva, os indicadores de habitabilidade podem variar em termos de granularidade geográfica, por exemplo, país, cidade e bairro, bem como em termos peculiares, como transporte, saúde, grupos populacionais e partes interessadas (governo, indivíduo, indústria, etc.). Devido a essa complexidade, a medição da habitabilidade consume muitos recursos, tempo e propicia um elevado custo financeiro para sua obtenção. Nesse contexto, o trânsito tem sido um fator organizador central em torno das comunidades construídas

(BEZERRA ET AL, 2019). Os sistemas de transporte e de trânsito apresentam indicadores de habitabilidade e são utilizados pela NAU nesse sentido.

Considerando essa conjuntura, Bezerra et al (2019) exploraram o uso de dados de tempo de chegada estimado de solicitações de viagens da Uber como um indicador simples de habitualidade urbana. Devido a sua natureza, escala e cobertura, a Uber fornece uma fonte de dados objetivos sobre a interação entre os habitantes de uma cidade e sua infraestrutura, em particular, infraestrutura de transporte. Dessa forma, é possível comparar dados em vários níveis, como cidades e bairros, mas também fornece dados sensíveis ao contexto, propiciando *insights* sobre o impacto de outros fatores que afetam os motoristas e viagens da Uber, incluindo incidentes de trânsito, clima e outros eventos. Nessa perspectiva, Bezerra et al (2019) levantaram como pesquisa a possibilidade de utilizar os dados da Uber para fornecer um indicador de habitabilidade urbana simples, rápido, de baixo custo, sensível ao tempo e ao contexto. Para isso, os autores consideraram a API *Uber Ride Request* (URR) para a cidade brasileira de Natal.

Bezerra et al (2019) utilizaram como meio de demonstração de viabilidade dessa proposta de pesquisa uma abordagem orientada a dados fundada em Análise Exploratória de Dados, explorando os dados da Uber. Nesse sentido, o principal objetivo do estudo foi mostrar que o serviço de viagens da Uber reage inerentemente às características e dinâmicas de uma cidade, possibilitando que esses dados sejam usados como fonte para um novo indicador de habitabilidade. De acordo com os autores, para a cidade de Fortaleza, as regiões Norte e Oeste têm a maior população residente (e as maiores taxas de crescimento populacional), mas também são caracterizadas por baixos níveis de renda per capita. A região Leste é o centro da cidade com bairros mais antigos e é caracterizada por alta densidade populacional e crescimento, enquanto a região Sul é uma adição recente à cidade e contém muitos dos mais novos, mais modernos e sofisticados restaurantes, hotéis e serviços sociais associados. Devido a essas características, os autores consideraram essa cidade como um contexto empírico adequado, uma vez que possui um mercado Uber ativo resultante de suas atividades turísticas, possui bairros com diferenças socioeconômicas e populacionais claramente definidas e é uma cidade relativamente pequena.

Bezerra et al (2019) utilizaram a API URR para recuperar dados de tempo de chegada estimado de viagens da Uber. Esse tempo foi considerado como o tempo em segundos entre uma solicitação ao sistema Uber para um motorista e o horário de início da viagem correspondente. Além disso, foi considerado diferentes bairros de Natal, onde para cada bairro foram definidos 10 locais de coleta, considerando os seguintes critérios: 5 pontos de interesse

como escolas, praças e parques; 4 coordenadas aleatórias para as quais a API conseguia responder as requisições de viagens; e o centroide geográfico de cada bairro. O estudo considerou 2 tipos de serviço de viagens da Uber: UberX e Uber Select. Os dados de interesse de tempo de chegada foram obtidos considerando um período de 1 mês, fevereiro de 2018. Para isso, os autores utilizaram uma infraestrutura dedicada de coleta e a cada 10 minutos o tempo de chegada estimado médio foi calculado a partir dos dados coletados para cada bairro. Os resultados foram anexados a um conjunto de dados de séries temporais. Dessa forma, o conjunto de dados resultante compreende uma série temporal de resolução de 10 minutos.

Bezerra et al (2019), por considerarem a natureza preliminar da pesquisa, utilizaram como técnica analítica inicial para descoberta orientada a dados a Análise Exploratória de Dados (AED). Nesse sentido, detectaram padrões, tendências, correlações e relações entre os dados para gerar *insights*, verificando a coerência dos dados como escala e formato esperado, além da identificação de dados ausentes e *outliers*. Dessa forma, empregaram vários métodos gráficos para apresentar resultados de análise de dados de maneira intuitiva, para que diferentes representações pudessem levar a indicações adicionais. A análise englobou dados de qualidade de vida, séries temporais e análises espaciais dos tempos de chegada. O passo seguinte foi correlacionar esses dados com dados socioeconômicos na tentativa de encontrar facetas do serviço de viagens e avaliar o potencial do tempo de chegada estimando da Uber como um novo indicador de habitabilidade. Para isso, os autores utilizaram a linguagem Python, bem como algumas bibliotecas para manipulação e plotagem dos gráficos.

Bezerra et al (2019) observaram uma diferença na média geral dos tempos de chegada entre os dois tipos de serviço de viagens, UberX e Uber Select, tanto para os períodos normais quanto para horários de pico, considerando os bairros selecionados. Isso sugere que a escolha do tipo de serviço pode ter mais impacto no tempo de chegada do que a escolha de fazer uma viagem em horário normal ou de pico. Além disso, foi observado que o tempo de chegada médio nos horários de pico pode ser explicado pelo o aumento de tráfego, congestionamentos e demanda pelo serviço de viagens da Uber. Como esperado, observaram também que um grande aumento no tempo de chegada no final da noite pode indicar uma menor demanda ou baixa disponibilidade de motoristas. Também foi evidenciado que diferentes regiões da cidade possui um perfil de tempo de chegada diferente e devido as curvas agrupadas e distintas nos gráficos de séries temporais com detecção de picos é uma indicação de que os dados são sensíveis ao contexto. Ou seja, os dados estão relacionados a eventos em toda a cidade, como Carnaval e eventos climáticos. O aumento dos tempos de chegada no Carnaval pode ser resultado de picos de demanda de turistas. Por outro lado, o aumento dos tempos em relação ao clima, como em

dias chuvosos pode indicar problemas de infraestrutura ou de transporte público na cidade. Nesse sentido, essa análise, segundo os autores, pode destacar tanto o potencial quanto os limites dos dados da Uber para pesquisa de habitabilidade urbana.

Segundo Bezerra et al (2019), por meio de análises espaciais, foi possível observar que os bairros Norte e Oeste estão sujeitos a maiores tempos de chegada. Essa análise consistiu na construção de mapas coropléticos que ilustram regiões com foco em cores. Nesse sentido, os bairros foram coloridos de acordo com suas médias de tempo de chegada. Além disso, a partir das análises realizadas, avaliaram os dados da Uber como um potencial indicador de habitabilidade urbana. Para isso, os autores combinaram informações de outras fontes de estudo que incluíam dados demográficos (população e densidade) e dados socioeconômicos de qualidade de vida, para construir um conjunto de dados de qualidade de vida de Natal. Ao correlacionarem esse conjunto de dados com os tempos de chegada da Uber em estudo, foi possível avaliar preliminarmente a utilidade desses dados da Uber como um indicador de habitabilidade urbana. Para isso, criaram uma matriz de correlação que possibilitou indicar fortes correlações entre as médias dos tempos de chegada e os indicadores de domínios políticos discutidos inicialmente.

Nesse contexto, foi possível evidenciar que os tempos médios de chegada, para ambos os tipos de serviço de viagens, estão fortemente correlacionados negativamente com os indicadores políticos relativos à infraestrutura ambiental urbana e aspectos socioeconômicos. Isso sugere que os bairros com menor nível de infraestrutura ambiental urbana e com aspectos socioeconômicos baixos possuem um tempo de espera mais longo para o serviço de viagens da Uber. Por outro lado, foi observado que os tempos possuem uma correlação fraca e negativa com a insegurança e a violência. Os autores acharam surpreendente essa observação, uma vez que o bom senso dita locais menos seguros como menos atraentes para os motoristas. Esse estudo de correlação evidenciou que aspectos infra estruturais e socioeconômicos de um bairro são potencialmente mais determinantes no desempenho de tempos de chegada do que todos os outros aspectos dos domínios políticos de habitabilidade. Além disso, considerando uma pontuação final em termos de domínios políticos de habitabilidade de cada bairro, foi possível observar que os tempos médios de chegada foram correlacionados negativamente com esses domínios por bairro, sugerindo que bairros com pontuações melhores são mais propensos a receber um serviço de viagem da Uber em menos tempo.

Considerando os resultados obtidos, Bezerra et al (2019) consideram que os dados da Uber podem refletir e destacar determinadas dinâmicas urbanas, além de potencial indicador de habitabilidade urbana. Entretanto, os autores destacam algumas limitações de sua pesquisa.

Uma delas foi a consideração de apenas uma métrica, o tempo de chegada estimado. Outros dados da Uber podem fornecer informações adicionais. Esses dados poderiam ser o preço de viagem e os tempos de durações das viagens. Outra limitação seria a concentração do estudo em apenas uma cidade e o período de coleta de dados de apenas 1 mês. A expansão tanto do número de cidades quanto no período de observações seria enriquecedora. Comparar os resultados com outros índices de habitabilidade e sustentabilidade, em conjunto com um período mais longo de tempo no estudo, poderá permitir uma maior exploração de como o clima, eventos, grandes incidentes e outras atividades impactam os dados da Uber e, portanto, a habitabilidade urbana. Por fim, os autores destacam que, a formulação de políticas baseadas em evidências parece ser uma tendência a ser parte central da governança em todo o mundo. Entretanto, essa formulação muitas vezes é prejudicada pela falta de dados confiáveis e oportunos, além de motivações políticas.

No intuito de testar a hipótese de que a precificação dos serviços de viagens da Uber se relaciona com características socioeconômicas dos lugares de embarque dessas viagens, Silva et al (2020) realizaram um estudo para a mesma cidade de Natal. Para alcançar esse objetivo, coletaram dados de preços de viagens do tipo de serviço UberX durante todo o ano de 2018, além de dados socioeconômicos a nível de Unidades de Desenvolvimento Humano fornecidos pelo Atlas do Desenvolvimento Humano no Brasil. Com os dados obtidos, foi possível construir modelos preditivos, utilizando técnicas de Aprendizagem de Máquina, para que posteriormente fossem submetidos a análises de regressão. Dessa forma, seria possível obter resultados.

De acordo com Silva et al (2020), para alcançar os resultados desejados, primeiramente obtiveram os dados de preços da UberX utilizando a API de simulação de preços da própria Uber. Esse recurso possibilita simular uma estimativa de preço mínimo e máximo a partir das localidades de origem e destino. A base de dados contendo essas informações foi concebida por meio de uma infraestrutura dedicada para essa tarefa durante todo o período de coleta. A aplicação era responsável por a cada 10 minutos, em paralelo, simular o preço de viagens da UberX para cada um dos 36 bairros selecionados da cidade. Para cada bairro, 10 coordenadas foram escolhidas com base nos seguintes critérios: conjunto fixo de 5 coordenadas de interesse; conjunto fixo de 4 coordenadas aleatórias; e o centroide geográfico do bairro. O resultado final do conjunto de dados foi de cerca de 5 milhões de linhas, considerando um intervalo de janeiro a dezembro de 2018.

De acordo com Silva et al (2020), os dados socioeconômicos dos bairros em estudo levaram em consideração apenas os locais de embarque e se referem às UDHs que, por sua vez, se baseia nas informações obtidas pelo Censo Demográfico brasileiro, gerenciado pelo Instituto

Brasileiro de Geografia e Estatística (IBGE). De acordo com os autores, as UDHs foram construídas para buscar agregar os dados de forma que gerem áreas mais homogêneas e que captem melhor as condições socioeconômicas. Nessa perspectiva, as variáveis socioeconômicas selecionadas se dividiram em 3 categorias: características de condição de vida, características de atividade econômica e *proxy* de demanda do serviço da UberX.

Os autores também adotaram uma abordagem baseada na Análise Exploratória dos Dados (AED). Nesse sentido, para o conjunto de dados de preços da UberX, a exploração das características do conjunto de dados aconteceu por meio de gráficos de suas variáveis quantitativas e mapas coropléticos para analisar relações espaciais. Nesse contexto, Silva et al (2020) observaram que as UDHs de Natal aparentam ser um fator decisivo na determinação do preço da Uber X. As 36 UDHs apresentaram uma média de preço diferente, dependendo da localização de origem. As UDHs mais afastadas do centro da cidade e, principalmente, aquelas que estão localizadas nas zonas mais ao Norte apresentaram uma média de preço mais elevada. Entretanto, as UDHs mais próximas do centro mostraram médias mais baixas. Segundo os autores, isso pode ocorrer por diversos motivos, entre eles o deslocamento de trabalhadores da zona Norte para as zonas Centro/Sul, devido essas áreas centralizarem as principais atividades econômicas do município. Além disso, foi evidenciado que o horário de uso do serviço também é um influenciador do preço. Para os dados obtidos, há quatro períodos principais de tempo durante o dia onde ocorrem picos no preço médio da Uber X, atingindo o maior valor entre 17 horas e 18 horas. Entretanto, a diferença entre o preço normal e durante um horário de pico não foi bastante significativa.

Na etapa de limpeza e tratamento dos dados, Silva et al (2020) consideraram de extrema importância sua realização devido ao fato de que os dados sem processamento não eram uniformes nem previsíveis. Nesse sentido, o primeiro passo foi tratar os dados ausentes, uma vez que muitas técnicas de Aprendizagem de Máquina não conseguem processar com êxito os dados ausentes. Para os dados da UberX coletados, os dados faltantes apresentaram origens distintas. Por esse motivo e pelo fato desses dados ausentes representarem uma pequena porção da base de dados, os autores optaram por remover todas as linhas que possuíam dados ausentes. Em seguida, Silva et al (2020) observaram a presença de *Outliers* no conjunto dados. Nesse contexto, os *outliers* são os preços das viagens altamente discrepantes dos demais valores, seja para valores excessivamente altos ou baixos. Para auxiliar nessa detecção, os autores utilizaram a técnica estatística denominada de *Interquartile Range* (IQR). Essa técnica utiliza quartis ao contrário de métodos que utilizam média ou o desvio padrão. A última etapa dessa parte consistiu no reescalonamento de variáveis numéricas para que não interferissem nos modelos

preditivos. No conjunto de dados considerado, a única variável que necessitou reescalar foi o código das UDHs de origem e destino, uma vez que era composto por 15 dígitos. A codificação escolhida pelos autores foi a codificação ordinal, uma vez que funciona tanto para dados categóricos quanto para dados numéricos e, de modo geral, designa um número inteiro para cada valor ou categoria presente no conjunto de dados.

Silva et al (2020), após a etapa de limpeza e tratamento, realizaram a criação dos modelos preditivos. Para analisar o relacionamento entre as variáveis socioeconômicas e o preço de Uber X, a pesquisa utilizou a técnica de Análise de Regressão. Nesse sentido, as variáveis socioeconômicas seriam algumas das variáveis utilizadas para estimar o valor de preço dos serviços da Uber X. Para atingir esse objetivo, foram criados modelos de Regressão Linear e não linear baseado em Árvores de Decisão. Os autores utilizaram o algoritmo de *Gradient Boosting*, fornecido pela biblioteca XGBoost que está disponível para várias linguagens de programação. A escolha desse algoritmo se deu pelo fato do tamanho da base de dados que inviabilizou utilizar a maioria dos outros algoritmos que não conseguiam treinar os modelos de regressão em um período de tempo viável. Para os modelos de regressão, foi utilizado 80% da base de dados como dados de treinamento e os 20% restantes foram utilizados como dados de teste para avaliar o desempenho do modelo.

Como resultados, Silva et al (2020) obtiveram no modelo de Regressão Linear um Erro Quadrático Médio (EQM) de 47.48 e Raiz do Erro Quadrático Médio (REQM) de 6.89 no conjunto de testes utilizado para avaliar o desempenho do modelo. Essas duas medidas indicam o erro do modelo em relação ao valor que era esperado. Isso significa que a diferença entre o preço estimado pelo modelo linear e o preço real da UberX é, aproximadamente, de R$ 6,89 a mais ou a menos. Dessa forma, para o modelo de Regressão Linear, os resultados mostraram que todas as variáveis sociodemográficas do local de embarque, com exceção da taxa de atividade econômica, se relacionaram com o preço da UberX. Considerando os coeficientes de regressão encontrados, os autores puderam determinar o grau e o sentido do relacionamento entre as variáveis e os preços. Valores dos coeficientes muito altos ou muito baixos indicam que as variáveis possuem alta significância na determinação dos preços. Enquanto coeficientes bem próximos de zero indicam variáveis que não se relacionaram bem e que tiveram pouca influência no preço. Nesse contexto, o coeficiente das variáveis indicativas de condição de vida mostrou que um aumento de 0.1 no índice de Gini (indica desigualdade) e IDH (indica qualidade de vida) fez o preço da Uber X aumentar, respectivamente, cerca de R$ 0,61 e R$ 0,45. Isso evidenciou que a desigualdade pode ser um fator que influencia mais para o aumento do preço que a qualidade de vida.

Por outro lado, para o modelo de Regressão não Linear baseado em Árvore de Decisão, Silva et al (2020) conseguiram estimar o preço da Uber X com maior precisão se comparado ao modelo linear. O modelo não linear obteve EQM de 11.34 e REQM de 3.25 no conjunto de dados de teste, o que representa uma melhoria de cerca de 52% em relação ao outro modelo. As variáveis socioeconômicas mais importantes para essa melhoria foram a taxa de atividade econômica, as variáveis de proxy de demanda do serviço e o índice de Gini. Os autores consideraram como peso de importância das variáveis o percentual de vezes que cada variável contribuiu para estimar o preço com maior precisão. Isso mostrou que as variáveis socioeconômicas também se relacionaram com o preço para o modelo de Regressão não Linear. Uma característica que, se mantendo constante em ambos os modelos, serviu para enfatizar a importância das variáveis socioeconômicas na estimativa de preço. Além disso, de maneira análoga ao modelo linear, os autores ajustaram alguns valores para determinadas variáveis no intuito de observar o comportamento dos preços. Nesse sentido, aumentar os valores da variável de taxa de atividade econômica para valores maiores que 72% fez com que a estimativa de preço da UberX aumentasse em média até 1290%.

Silva et al (2020) mostraram que variáveis socioeconômicas de locais de embarque podem se relacionar com preços de viagens da UberX. Além disso, esse relacionamento se mostrou um relacionamento não linear visto que o desempenho exibido pelo modelo de Regressão Linear foi inferior ao desempenho do modelo não linear. Nessa perspectiva, os autores realçam a importância dos dados de preço de viagens da Uber como um potencial indicador de características socioeconômicas. Pelo fato de ser sensível a questões de mobilidade urbana, os dados de viagens da Uber podem ser utilizados como possíveis indicadores de problemas de mobilidade. Além disso, os autores relatam algumas limitações dessa pesquisa. Uma delas está no conjunto de dados relacionados ao preço de viagens. São dados simulados e podem não representar 100% a realidade dos custos das viagens. Outra limitação foi o fato de contemplarem apenas um tipo de serviço de viagens da Uber, a UberX. Utilizar dados de outros serviços da empresa e até mesmo dados de outros serviços de transporte pode gerar novos ou até melhores resultados. Além disso, pode-se destacar outra limitação que se refere aos dados socioeconômicos utilizados, pois fazem referência ao ano de 2010. No período de 10 anos entre cada censo, os dados socioeconômicos podem mudar bastante e não refletir de maneira mais destacada a realidade da população. Revisar a pesquisa utilizando dados mais recentes pode melhorar o desempenho dos modelos criados e gerar novos resultados.

Os trabalhos supracitados buscaram evidenciar relações entre informações acerca de viagens da Uber com características socioeconômicas em determinadas localidades. Essas

relações abrangeram aspectos sobre comportamento de preços e de distâncias quando analisados com fatores sociais que poderiam indicar influência na oferta e demanda dos serviços de viagens da Uber. Os achados dessas pesquisas indicaram que algumas características socioeconômicas de locais podem influenciar na concepção de preços.

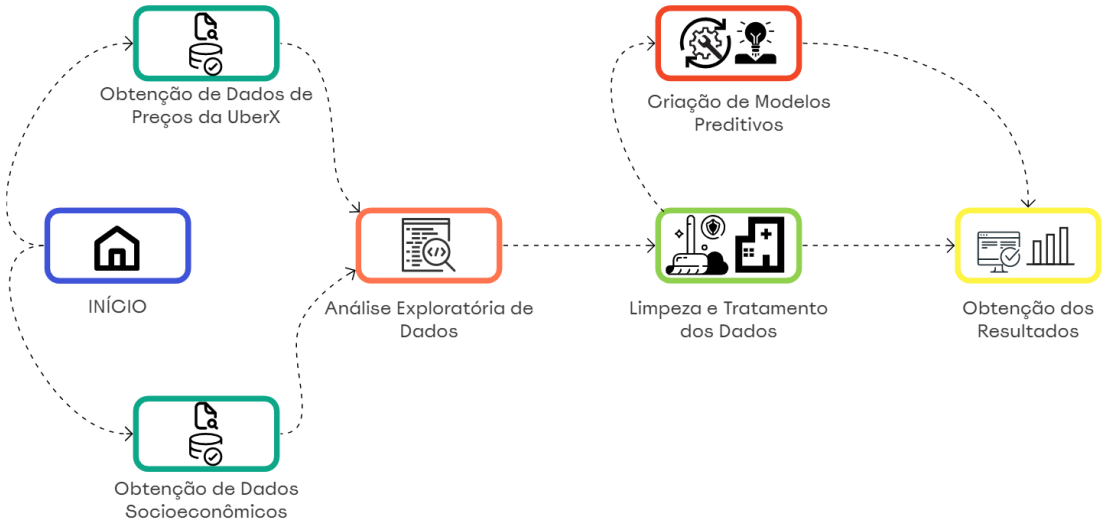
Adicionalmente, os Trabalhos Relacionados levantados nessa pesquisa não focaram no comportamento de preços entre trajetos de viagens para centros financeiros, considerando bairros de diferentes classes de renda. Os estudos salientaram a acessibilidade ao serviço de viagem, o tempo de espera desse serviço, bem como o comportamento do preço quando uma viagem era solicitada. Entretanto, não enfatizaram em suas analises se esses preços poderiam estar tendo uma forte elevação ou não para usuários de baixa ou alta renda e se esses preços apresentavam alguma tendência de aumento para esses usuários quando o destino era o centro financeiro de alguma região. Além disso, não há análise nos estudos levantados que demonstre quais seriam os valores de um possível impacto desses aumentos de preços na Renda Média Pessoal dos usuários de diferentes classes de renda.

Considerando essa conjuntura, há possibilidade de se analisar o comportamento desses preços quando se leva em consideração os trajetos de viagens entre regiões com rendas financeiras distintas, quando o destino dessas solicitações de viagens é o centro financeiro de alguma região. Nessa perspectiva, o presente estudo busca encontrar indicativos de que uma não concentração de centros financeiros pode contribuir na redução de custos dos preços de viagens da Uber para usuários residentes em regiões mais pobres economicamente. Analisar outras regiões com outros contextos econômicos poderia levar a uma indicação de que o processo de precificação de viagens da Uber seria influenciado por uma concentração de atividades comerciais.

METODOLOGIA

Neste capítulo são descritas as etapas de ciclo de vida de um processo de Ciência de Dados adaptado da seção 2.3.1, para auxílio na obtenção dos resultados. O processo adotado nessa estratégia de execução foi ajustado considerando o contexto da pesquisa, que envolve Dados Socioeconômicos e preços de viagens em transporte por aplicativo Uber. As etapas foram: Obtenção de Dados de Preços da UberX, Obtenção de Dados Socioeconômicos, Análise Exploratória de Dados, Limpeza e Tratamento dos Dados, Criação de Modelos Preditivos e Obtenção dos Resultados. A Figura 9 ilustra o fluxo dessas etapas.

Figura 9 - Fluxo do processo utilizado na pesquisa.



Fonte: Autor (2022).

Pelo fluxo de processo proposto acima, nota-se que as etapas são utilizadas em vários contextos de aplicação. Entretanto, para esse estudo, as etapas de Obtenção de Dados de Preços da Uber e de Limpeza e Tratamento dos Dados foram adaptadas para a realidade dos objetivos da pesquisa. Além disso, há o acréscimo da etapa de Obtenção de Dados Socioeconômicos que consiste em uma etapa de levantamento de informações sociais relacionadas a um local ou região. Há várias dimensões para essa etapa, como Educação, Idade, Renda, Etnia, Emprego, entre outras. Para a etapa de Obtenção de Dados de Preços da Uber, foi elaborada uma estratégia

para criação de uma base de dados que contivesse preços de viagens da Uber para a cidade de Fortaleza.

Esses preços necessitavam ser condizentes com a realidade praticada pelo serviço. Devido a isso, foram feitas pesquisas com o objetivo de obter informações que respaldassem a lógica de criação desses preços. Por outro lado, os preços de viagens para Boston não precisaram ser simulados, uma vez que foi possível encontrar uma base de dados real que continha essas informações. Para a etapa de Limpeza e Tratamento dos Dados, foi realizada uma análise Estatística no intuito de validar a base de dados criada para a cidade de Fortaleza. Nesse sentido, seria possível respaldar estatisticamente os valores simulados dos preços.

Para a cidade de Boston, essa análise não foi realizada, uma vez que os preços contidos em sua base provinham de viagens reais. Além disso, para essa etapa foi realizada uma análise em grafos, utilizando Medidas de Centralidade. Essas medidas auxiliaram nas análises de impacto dos trajetos de viagens para diferentes rotas. Dessa forma, seria possível averiguar se determinado bairro teria um impacto maior ou menor no valor da média de preços das viagens, caso uma trajetória fosse alterada.

Nesse contexto, para a cidade de Fortaleza, foi selecionado para experimento 5 bairros de cada classe rica, média e pobre. O critério de escolha levou em consideração a parte financeira e a localização. Nesse sentido, foram utilizadas as informações da Renda Média Pessoal dos residentes com 10 anos ou mais de idade obtidos na seção 2.2.1. Assim, foi extraído 5 representantes de cada classe, considerando as distâncias da região rica (centro comercial englobando Norte e Leste geográfico da cidade), da região mediana (centro geográfico da cidade) e da região periférica (Sul e Sudeste geográfico da cidade). Como levantado na Revisão da literatura, a cidade de Fortaleza concentra os bairros ricos nas regiões Norte e Leste que coincidem com o centro comercial e possui residentes com alta renda per-capita. A região do centro geográfico da cidade coincide com bairros de renda per-capita intermediária e a região periférica (Sul e Sudeste) representa a maioria dos bairros de baixa renda.

Para a cidade de Boston, foram selecionados alguns bairros e regiões pertencentes a bairros. A escolha levou em consideração a distância geográfica deles em relação ao centro financeiro da cidade. O intuito foi observar o comportamento de alta e baixa de preços, quando as solicitações de viagens partiam de bairros ou regiões mais afastadas ou próximas desse centro financeiro. Com os resultados obtidos dessa análise, seria possível evidenciar alguma tendência de preços semelhante ao observado para a cidade de Fortaleza. Os bairros e regiões escolhidos levaram também em consideração a facilidade de disponibilidade de informações de renda per-

capita a nível de bairro ou região da cidade. O total de bairros e regiões da cidade foi de 12 unidades que foram descritas na seção 2.2.2.

3.1 Obtenção de Dados de Preços da Uber

Os preços analisados foram para o tipo de serviço UberX. Isso foi determinado devido a algumas considerações:

- Como observado anteriormente, UberX é o serviço mais popular da empresa Uber, fornecendo preços mais acessíveis para a população. Isso possibilita que usuários de várias faixas de renda tenham acesso a esse tipo de serviço, aumentado o escopo de usuários do serviço.
- A base de dados com preços de viagens reais obtida para a cidade de Boston contém informações do serviço UberX. Devido a isso, as simulações para criação da base de dados de Fortaleza levaram em consideração o mesmo tipo de serviço, permitindo mais homogeneidade nas observações para ambas as cidades em estudo.

Para determinação dos preços das viagens que serviram para compor a base simulada de Fortaleza, foi utilizado o simulador de preços da Uber e as pesquisas sobre características particulares da cidade. Dessa forma, foi possível gerar amostras de preços de viagens entre bairros, inclusive de viagens com origem e destino para o mesmo bairro. Além disso, as pesquisas levantaram evidências de que, em determinados dias e horários, os preços das viagens sofriam algumas variações. Nesse sentido, os preços, quando necessário, foram reajustados de acordo com os horários e os dias, já que em horários de pico e em determinados dias, o preço tende a aumentar. A Tabela 2 abaixo informa os horários com seus respectivos reajustes de preços.

Tabela 2 - Reajuste na concepção dos preços, quando necessário.

HORÁRIO	REAJUSTE (R$)
07:00 – 09:59	**+ [1,00 a 2,00]**
17:00 – 19:59	**+ [1,00 a 2,00]**
00:00 – 06:59	**- [1,00 a 2,00]**
20:00 – 23:00	**- [1,00 a 2,00]**
Outro Horário	**Sem Reajuste**

Fonte: Autor (2022).

O próprio simulador de preços da Uber inclui as regras de política nos cálculos dos preços da empresa, levando em consideração a localidade, a taxa base, a taxa dinâmica e as variações de oferta e demanda. Entretanto, para tornar a lógica de precificação mais realista, foi preciso considerar informações inerentes aos locais em estudo. Devido a isso, mediante as pesquisas realizadas, foi observado que as viagens da Uber na cidade ocorrem com uma divisão entre bairros de cerca de 4%. Ou seja, há uma divisão aproximada de 4 viagens para trajetos entre bairros pobres-ricos, pobres-médios, médios-ricos e demais trajetos possíveis entre eles. Nesse sentido, foi criada uma função chamada "*get_destination*" que implementa uma lógica de deslocamento. O Algoritmo 1 exibe um trecho de código dessa função.

Algoritmo 1 - Parte da lógica para divisão de deslocamento de viagens.

```
#ALGORITMO 1
#MEDIOS
    if (x in medios) and y <= 0.48:
        return random.choices(medios)
    if (x in medios) and y > 0.48 and y <= 0.64:
        return random.choices(ricos)
    if (x in medios) and y > 0.64 and y <= 0.8:
        return random.choices(guararapes)
    if (x in medios) and y > 0.8 and y <= 0.96:
        return random.choices(palmeiras)
    if (x in medios) and y > 0.96:
        return random.choices(pobres)
```

Fonte: Autor (2021).

Para codificação da função supracitada, foi considerada uma semente de aleatoriedade fornecida pela biblioteca *Random* disponível para a linguagem *Python*. Essa semente serviu para equilibrar as quantidades de escolhas dos bairros, considerando o número de viagens levantado em pesquisa e os destinos mais solicitados (IPLANFOR, 2015), (G1 CE, 2021). Dessa forma, as viagens tendem inicialmente a acontecerem para dentro do próprio bairro, por isso uma parcela maior de probabilidade dessa ocorrência.

Devido a isso, foi convencionado, por exemplo, que viagens originadas de bairros médios possuem uma possibilidade maior de ter como destino o próprio bairro, sendo definido assim uma porcentagem maior para esses casos de cerca de 48% de chance de isso acontecer. Ou seja, 48% das viagens tendem a acontecer para dentro do mesmo bairro. Esse valor garante que a quantidade maior de viagens aconteça para a mesma região. Essa porcentagem poderia ser outra, desde que fosse um valor grande o suficiente para estabelecer que as viagens em sua maioria aconteçam para uma mesma região. Nas outras alternativas, o destino de um usuário desse bairro poderá ser qualquer outro bairro da cidade, sendo essas quantidades definidas em uma porção 16% do total de viagens para as demais possibilidades (48% a 64%, 64% a 80%, 80% a 96%). A Tabela X abaixo informa os trajetos considerados com suas respectivas quantidades de viagens em porcentagem.

Tabela 3 - Distribuição de quantidade de viagens em porcentagem.

TRAJETO	48%	16%	14%
RICO -> RICO	**X**		
RICO -> POBRE		**X**	
RICO -> MÉDIO			**X**
POBRE -> POBRE	**X**		
POBRE -> RICO		**X**	
POBRE -> MÉDIO			**X**
MÉDIO -> MÉDIO	**X**		
MÉDIO -> RICO		**X**	
MÉDIO -> POBRE			**X**

Fonte: Autor (2022).

Além disso, é importante frisar que os bairros Guararapes e Conjunto Palmeiras possuem um comportamento de quantidade de viagens e preços diferentes dos demais bairros, por isso a necessidade de tratá-los separadamente.

Como o preço está relacionado ao dia e ao horário, também foi preciso incluir uma lógica que fosse condizente com a realidade da cidade. Nesse sentido, foi levantado em pesquisa que as viagens da Uber acontecem em maior parte nos finais de semana e um maior número de solicitações acontecem entre os horários das 7 às 14 horas e das 17 às 20 horas (IPLANFOR, 2015), (G1 CE, 2021). O Algoritmo 2 e

Algoritmo 3 exibem trechos de codificação para os horários e para os dias, respectivamente.

Algoritmo 2 - Parte da lógica para os horários.

```
#ALGORITMO 2

def get_horario_fds():
    x = random.random()
    if x <= 0.4:
        return random.randint(0,10)
    if x > 0.4 and x <= 0.8:
        return random.randint(17,24)
    else:
        return random.randint(10,17)

def get_horario_semana():
    x = random.random()
    if x <= 0.20:
        return random.randint(7,10)
    if x > 0.20 and x <= 0.4:
        return random.randint(12,14)
    if x > 0.4 and x <= 0.6:
        return random.randint(17,20)
    if x > 0.6 and x <= 0.75:
        return random.randint(20,24)
    if x > 0.75 and x <= 0.9:
        return random.randint(10,12)
    if x > 0.9 and x <= 0.95:
        return random.randint(14,17)
    else:
        return random.randint(0,7)
```

Fonte: Autor (2021).

Basicamente, a função "*get_horario_fds*" dá mais prioridade aos horários que são mais utilizados nas altas procuras do serviço de viagens para os finais de semana e a função "*get_horario_semana*" prioriza os horários das viagens que acontecem durante a semana. As

faixas de horários foram inseridas considerando todos os períodos de utilização que podem demandar uma quantidade significativa de solicitações, podendo ser um horário de final de semana ou um horário de dia da semana.

Algoritmo 3 - Lógica de escolha dos dias da semana.

```
#ALGORITMO 3

horario = []
for i in Dia_Semana:
    if i == "Sábado" or i == "Domingo":
        horario = np.append(horario, get_horario_fds())
    else:
        horario = np.append(horario, get_horario_semana())
```

Fonte: Autor (2021).

A depender do dia da semana, o preço pode sofrer influências diferentes na cidade. Devido a isso, o

Algoritmo 3 exibe o comportamento que a base deve tomar caso a semente de aleatoriedade escolha um dia da semana ou um dia de final de semana. Para isso, é preciso utilizar a lógica apresentada nas funções "*get_horario_fds*" e "*get_horario_semana*".

Paralelamente a esquematização dessas informações, foram realizadas simulações com o simulador de preços da empresa para cada dia da semana e a cada 15 minutos foi coletada uma amostra de preço entre todos bairros considerados nesse estudo. Esse processo foi realizado por um período de 2 semanas, entre o final do mês de agosto e início do mês de setembro de 2021. O experimento também levou em consideração uma época sem feriados e sem eventos de grandes proporções na cidade. Após isso, foi calculada a média aritmética para cada trajeto. A

Tabela 7 que contém os preços das simulações pode ser vista no Apêndice.

Com os valores simulados em mãos, foi possível obter uma faixa de valores de preços para cada trajeto em estudo. Com as ponderações supracitadas em relação ao dia da semana e o horário, foi possível construir uma base de dados mais condizente com a realidade da cidade, contendo os preços de viagens simuladas, além de outros atributos. Ao total, foi gerada uma amostra de tamanho de 100.000 de preços de viagens. O Algoritmo 4 exibe parte da codificação responsável pela lógica de precificação construída.

Algoritmo 4 - Parte da lógica de precificação.

```
#ALGORITMO 4

def get_prices(x,y):
    if x == y:
        return round(random.uniform(5.42,7.42),2)
    if x in ricos and y in ricos:
        return round(random.uniform(6.42,11.97),2)
    if x in ricos and y in pobres:
        return round(random.uniform(24.73,34.95),2)
    if x in pobres and y in ricos:
        return round(random.uniform(22.29,35.67),2)
    if x in pobres and y in pobres:
        return round(random.uniform(6.42,24.88),2)
```

Fonte: Autor (2021).

A função "*get_prices*" é responsável pela concepção dos preços da base. No algoritmo supramencionado observa-se, por exemplo, que se os bairros de origem e destino forem os mesmos, isso significa que o preço ficará em torno de R$ 5,42 a R$ 7,42 da média obtida pelo simulador de preços da Uber, uma vez que a distância entre locais da mesma região costuma ser menor do que um deslocamento entre regiões distintas. Por outro lado, caso a origem e o destino forem distintos, então a faixa de preço será correspondente aos demais preços simulados

pelo simulador de preços da Uber. Embora essa lógica se repita para os demais trajetos e faixas de valores, é preciso considerar o fator de reajuste de preços que é uma questão peculiar de cada localidade. Devido a isso, foi implementada uma função de reajuste chamada "*reajuste_preco*" que é chamada quando é criada a lista de preços correspondente a cada trajeto de viagem. O Algoritmo 5 exibe essa implementação.

Algoritmo 5 - Implementação de reajuste de preço.

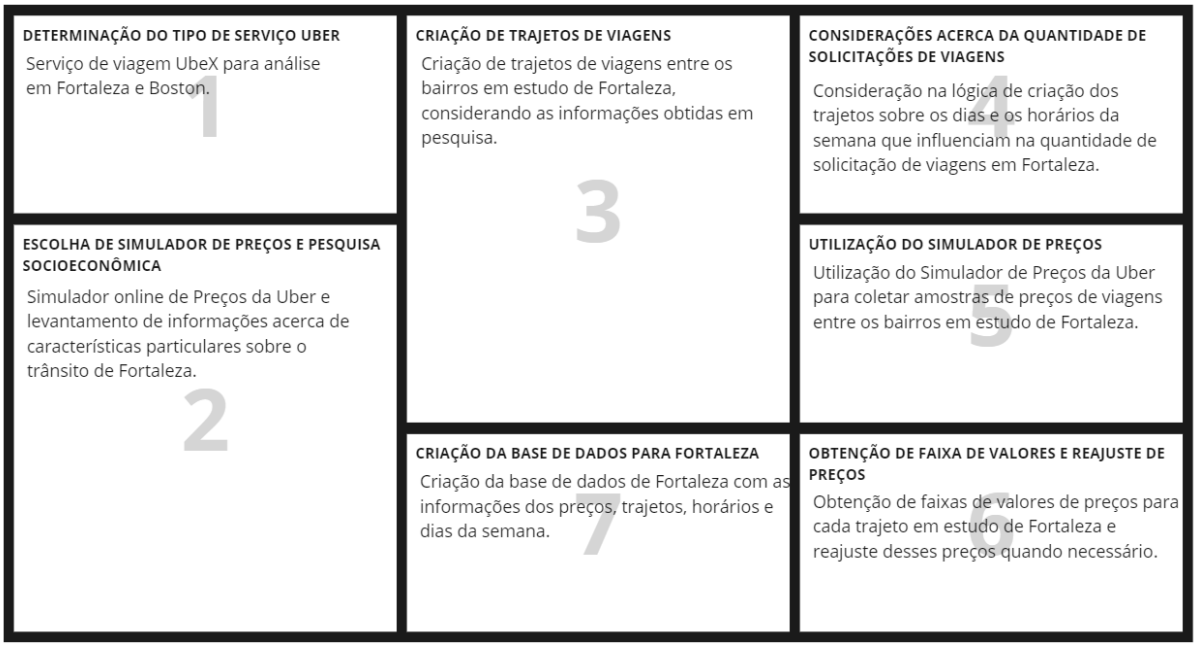
```
#ALGORITMO 5

def reajuste_preco(x,y):
    if x >= 7 and x < 10:
        a = y + round(random.uniform(1,2),2)
        return a
    if x >= 17 and x < 20:
        a = y + round(random.uniform(1,2),2)
        return a
    if x >= 0 and x < 7:
        a = y - round(random.uniform(1,2),2)
        return a
    if x >= 20 and x <= 23:
        a = y - round(random.uniform(1,2),2)
        return a
    else:
        return y

for i in range(100000):
    price[i] = reajuste_preco(horario[i],price[i])
```

Fonte: Autor (2021).

Pelas pesquisas realizadas neste trabalho, Fortaleza apresenta peculiaridades nas variações do preço de viagens da Uber a depender do horário. Devido a isso, a lógica de reajuste

levou em consideração um fator de aumento entre R$ 1,00 e R$ 2,00 que são os valores médios encontrados para esse tipo de variação. No algoritmo supramencionado, a lógica se baseia na verificação dos intervalos de horários de pico que acontece durante um dia para aquela cidade. A depender do valor da semente de aleatoriedade, o dia será de final de semana ou não, e isso fará com que essa função de reajuste seja chamada, quando assim pertinente for. A Figura 10 ilustra um resumo dos passos seguidos para construção da base de dados de Fortaleza.

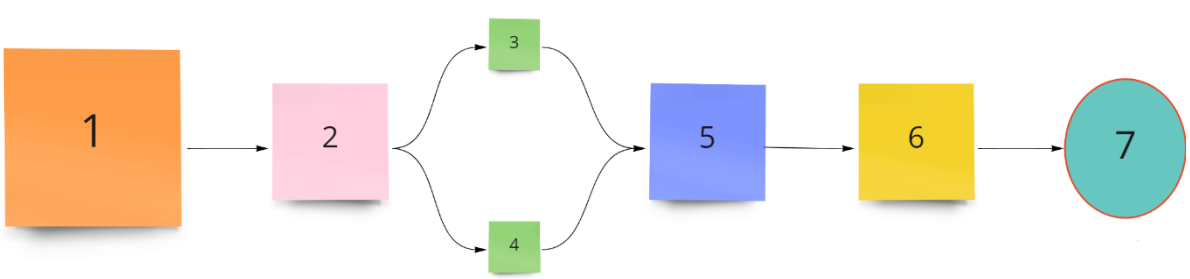
Figura 10 - Passos seguidos para construção da base de dados de Fortaleza.



Fonte: Autor (2022).

A Figura 11 abaixo ilustra um fluxograma desses passos.

Figura 11 - Fluxograma para construção da base de dados de Fortaleza.



Fonte: Autor (2022).

Para a cidade de Boston, os preços já estão registrados por meio da base de dados real disponibilizada pela plataforma *Kaggle*. Essa base possui 46 colunas, contendo informações de origem e destino de viagens, clima, mês, distância de viagem, horário da viagem, dia, preço e outras informações acerca de uma viagem. A base contém pouco mais de 600 mil linhas e se refere a dados coletados para o ano de 2018. Os preços das viagens da base dessa cidade podem fornecer mais uma indicação de alta de preços para trajetos de viagens que se destinam a centros financeiros de uma região. Dessa forma, é possível fundamentar por meio de outro contexto econômico (região de país desenvolvido) que há indicativo de alta de preços de viagens da Uber, quando o destino tende a ser centros financeiros.

3.2 Obtenção de Dados Socioeconômicos

A Uber não fornece informações socioeconômicas dos usuários de seus serviços de viagens. Nesse sentido, um meio alternativo de obter esses dados seria por meio da utilização das características do local de embarque das solicitações desses serviços.

Neste estudo, foram utilizadas prioritariamente informações referentes a renda das pessoas por bairro. Dessa forma, seria possível analisar os impactos financeiros para os usuários que partem dos bairros que residem, considerando trajetos entre bairros de diferentes classes. Além disso, foram ponderadas nas análises as outras informações gerais dos locais em estudo vistas na Revisão da literatura. Nesse contexto, para a cidade de Fortaleza, utilizou-se dados do Índice de Desenvolvimento Humano (IDH). Esses dados se baseiam no Censo Demográfico brasileiro realizado no ano de 2010 (último realizado a nível de bairro).

De acordo com a Secretaria Municipal de Desenvolvimento Econômico (SMDE) de Fortaleza, a metodologia de mensurar o desenvolvimento econômico é utilizada pela Organização das Nações Unidas (ONU) desde a década de 1990. Assim, é possível avaliar anualmente o grau de desenvolvimento dos países utilizando o IDH, que representa um indicador sintético composto por 3 dimensões: Renda, Educação e Longevidade.

Em nível municipal para Fortaleza, o Programa das Nações Unidas (PNUD) calculou para o ano de 2010 o IDH Municipal (IDH - M), objetivando avaliar o padrão de vida (dimensão Renda), acesso ao conhecimento (dimensão Educação) e as condições de longevidade e saúde

(dimensão Longevidade). Entretanto, a metodologia empregada para o cálculo do IDH – M possui alterações daquela utilizada para o IDH a nível de país. Isso acontece devido a alguns fatores, como as bases de dados utilizadas e outros indicadores para o cálculo do índice. Nesse sentido, o cálculo do IDH a nível de bairro para Fortaleza (IDH - B) possui adaptações à metodologia do IDH (FORTALEZA, 2014).

A classificação do IDH varia de 0 a 1. Quanto mais próximo de 1 melhor o grau de desenvolvimento, quanto mais próximo de 0 pior o grau. Nesse mesmo sentido, dá-se a classificação dos componentes do índice (Renda, Educação e Longevidade). Assim, considerando os cálculos realizados pela SMDE, foi possível obter o IDH – B de Fortaleza que serviu de base para que calculasse o IDH da renda dos respectivos bairros (IDH - R). Esse índice de renda considera a informação da renda média mensal das pessoas de 10 anos ou mais de idade.

Para fins de pesquisa, foram utilizadas informações de IDH – R, IDH – B e renda média pessoal de 5 bairros de cada classe para a cidade de Fortaleza: rica, pobre e média. As localizações geográficas foram escolhidas considerando uma divisão que evidenciasse distinções das rendas dos residentes desses bairros, uma vez que, para essa cidade, quanto mais longe do centro comercial um bairro está, menor é a sua renda. A Tabela 1 da seção 2.2.1 exibe as informações de IDH – R e a renda média pessoal. As informações dos IDHs de todos os bairros da cidade podem ser obtidas em (FORTALEZA, 2022).

Os dados dessa renda média pessoal, por serem do ano de 2010, foram convertidos para valores reais de 2021. Isso foi realizado porque as simulações dos preços das viagens foram realizadas nesse período. Dessa forma, foi possível mitigar as diferenças de inflação, juros e mudanças econômicas nesse intervalo de tempo. Para alcançar esse objetivo, foi utilizada a Calculadora do Cidadão disponibilizada pelo Banco Central do Brasil que possibilita a conversão de valores monetários em Reais de um ano para valores monetários em Reais de outro ano (BANCO CENTRAL DO BRASIL, 2022). Nessa perspectiva, para a conversão dos valores pela calculadora foi considerado o Índice Nacional de Preços ao Consumidor Amplo (IPCA) para o ano de 2021, pois, segundo a literatura, é o mais utilizado para realizar correção monetária. O IPCA retrata o âmbito mais geral da inflação da economia. Assim, foi possível realizar uma análise mais realista dos valores monetários envolvidos no contexto do experimento no intuito de comparar as diferenças de renda entre as classes de bairros, bem como o impacto desses valores em mudanças de trajetos de viagens.

As informações socioeconômicas obtidas para a cidade de Boston ficaram restritas aos dados colhidos pela Agência de Pesquisa da Divisão de Planejamento e Desenvolvimento de

Boston (BP&DARD) de 2017. Essa agência realizou uma pesquisa formalizada em um documento chamado de Perfis dos Bairros. Dessa forma, foi possível extrair alguns bairros e regiões da cidade para que fosse possível obter evidências sobre tendência de alta ou baixa de preços, considerando como ponto de referência o centro financeiro da cidade. As informações nesse contexto podem ser consultadas na seção 2.2.2.

3.3 Análise Exploratória dos Dados

De acordo com Natrella (2010), pode-se conceituar a Análise Exploratória de Dados (AED) como uma abordagem para analisar dados a qual é empregada técnicas gráficas no intuito de maximizar o discernimento, bem como a estrutura e variáveis importantes. Além disso, pode-se obter outros fatores que podem evidenciar comportamentos em um conjunto de dados.

Nessa pesquisa, a exploração das características dos conjuntos dos dados acontece por meio de gráficos, para obter relações entre as variáveis de interesse. Nesse sentido, para a cidade de Fortaleza, foram gerados gráficos que externassem informações sobre características socioeconômicas dos bairros em estudo, considerando o IDH com enfoque na dimensão renda, bem como sobre informações de preços, horários, trajetos e dias da semana de viagens do serviço UberX.

Por sua vez, para a cidade de Boston, foram gerados gráficos que externassem informações acerca da quantidade de viagens por preço e por distância. Também foram plotados gráficos que mostraram o valor médio e mediano dos preços das viagens, bem como aqueles preços que estavam distantes do valor médio, ou seja, *Outliers*. Além disso, gráficos que exibiam informações acerca do número de viagens por clima e a média de preços das viagens por clima também foram plotados. Plotou-se um gráfico para exibir informações entre a variação dos preços das viagens do UberX e a distância do trajeto, além de outro gráfico que relacionou o preço médio com a hora do dia. Por fim, foram gerados gráficos que relacionaram origem e destino das viagens, bem como a média de preço por destino da viagem e a correlação entre as variáveis da base de dados. Isso serviu para analisar o comportamento dos preços, considerando a classe de renda do bairro que solicitou uma viagem.

3.4 Limpeza e Tratamento dos Dados

Antes de obter as evidências provenientes da AED e da criação de modelos preditivos, é importante realizar uma limpeza e tratamento dos dados. De acordo com Haughton et al (2003), essa etapa é relevante, pois dados sem processamento costumam ser não uniformes e não previsíveis. Além disso, os dados podem elevar custos, quando utilizados de maneira bruta.

Nessa perspectiva, para a cidade de Boston, foram analisados os dados faltantes, também chamados de dados omissos. As análises desses dados possibilitaram observar que tratavam de origens distintas que possuíam ausência de valores para determinados atributos. Por este motivo e devido a quantidade total de dados faltantes representarem uma pequena fração de todo o conjunto da base de dados, optou-se por eliminá-los. Por outro lado, para a cidade de Fortaleza, não existiu dados ausentes, uma vez que a base foi gerada em ambiente controlado, por meio de simulações, pesquisas e cálculos matemáticos.

Outro ponto importante no que se refere à limpeza e ao tratamento de dados é a observação dos *Outliers* (pontos fora da curva), pois podem gerar dados inválidos que não retratam a realidade. Esses pontos apresentam um grande afastamento numérico das outras observações. Em nosso contexto de dados de preços de viagens da UberX, os *Outliers* surgem como preços muito elevados em relação à média e a mediana, para a cidade de Boston. Nesse sentido, para detectar esses pontos, existem técnicas estatísticas, como o Z – Score. Como levantado na Revisão da literatura, essa técnica fornece a indicação da distância numérica entre um ponto e a média da amostra. Nesse sentido, é baseado no desvio padrão e tenta mitigar a influência da localização e do tamanho dos dados. Como a base de dados de Boston é na ordem das centenas de milhares de linhas, optou-se por essa técnica. Por outro lado, a base de dados de Fortaleza não apresenta *Outliers*, uma vez que as simulações realizadas pelo simulador da empresa forneceram faixas de valores próximas das médias obtidas posteriormente.

Um ponto também importante nesse contexto de tratamento de dados é a verificação da distribuição dos dados amostrais. Como visto na seção 2.5, muitos métodos estatísticos assumem que os dados seguem uma Distribuição Normal. Porém, a maioria dos dados não são normais e acabam violando alguns desses testes, como o teste T – Student. Nesse sentido, foram verificados se os preços das viagens dos trajetos em estudo seguiam uma Distribuição Normal, para a base de dados de Fortaleza. Os trajetos em estudo foram: pobre – rico, pobre – médio e médio – rico. Após a verificação para cada trajeto se seguiam uma Distribuição Normal, seguiu-se para a realização do teste de Normalidade de Kolmogorov-Smirnov para validação. Os

resultados confirmaram que os preços das viagens para essa base de dados simulada não seguiam uma Distribuição Normal. Devido a isso, foi utilizado o teste não paramétrico de Wilcoxon.

Considerando esse cenário e como visto na seção 2.5, o teste de Wilcoxon é comumente tratado como uma versão não-paramétrica do teste T - Student para amostras pareadas. Ele testa se a distribuição das diferenças entre duas amostras é simétrica e centrada em zero. Nesse sentido, os pares considerados foram os trajetos supracitados pobre – rico e pobre – médio. Dessa forma, considerando um nível de confiança de 95% e nível de significância Alpha de 5%, foi determinado na execução do teste que se verificasse a hipótese alternativa (teste unilateral a direita) de que a média dos preços do trajeto pobre – rico é maior do que a média dos preços do trajeto pobre – médio. Assim, seria possível verificar se haveria evidência estatisticamente significante de que a média de preços do trajeto pobre – rico é maior do que a média dos preços do trajeto pobre – médio. Em outras palavras, se os grupos diferem estatitistacamente de maneira representativa.

Como a literatura afirma que geralmente dados de prestação de serviços que envolvem valores monetários não são previsíveis pela própria natureza da concepção desses valores, por exemplo, viagens de transporte por aplicativo, os testes estatísticos foram realizados somente para a base de dados Fortaleza, uma vez que foi uma base simulada que necessitou dessa verificação para ratificar a realidade desses tipos de dados. Devido a isso, a base de Boston não foi testada estatisticamente nesse contexto, não só por esses motivos, mas também por ser uma base real de preços de viagens de transporte por aplicativo que serviu como apoio no fundamento das observações realizadas para a cidade de Fortaleza.

Considerando as ponderações das bases de dados em estudo e pelo fato da base de dados de Fortaleza ser simulada, também foi tratado para essa base algumas medidas de centralidade para grafos no intuito de ratificar a realidade de seus dados e as pesquisas realizadas para essa cidade. Nesse sentido, primeiramente, os dados da base de dados de Fortaleza foram convertidos em um grafo ponderado em que o peso das arestas são as médias de preços das viagens entre os bairros considerados para estudo. Os trajetos entre esses bairros levaram em consideração a lógica de criação da base vista anteriormente nesse capítulo. Os bairros estão registrados na Tabela 1 da seção 2.2.1.

Para a conversão dessa base em um grafo, foi preciso utilizar a biblioteca "*networkx*" a qual permitiu a manipulação em nível de *nós* e *arestas*, bem como a atribuição do preço como sendo o fator de ponderação. Além disso, com essa biblioteca foi possível renderizar o grafo de

forma que os trajetos entre os bairros em estudo fossem visíveis, considerando seus respectivos pesos.

A partir dessa manipulação, foi possível utilizar as Métricas de Centralidade para grafos. Os métodos que permitiram realizar o cálculo das métricas estão disponíveis também na biblioteca "*networkx*". As métricas utilizadas nessa pesquisa foram: Centralidade de Grau, Centralidade de Proximidade e Centralidade de Intermediação. A biblioteca "*nxviz"* foi utilizada para criar os gráficos com os resultados das Métricas de Centralidade. O Algoritmo 6 exibe parte da implementação que converte um filtro de viagens da base de dados de Fortaleza.

Algoritmo 6 - Parte da implementação para conversão dos dados em grafo G.

```
#ALGORITMO 6

edges = viagens_geral[['Origem','Destino','Preço','TIPO_VIAGEM']]
nodes = viagens_geral[['Origem']]
edges.columns = ['source','target','weight','TIPO_VIAGEM']
nodes.columns = ['index']
edges = edges.round(2)
data = nodes.set_index('index').to_dict().items()
G = nx.from_pandas_edgelist(edges, edge_attr=['weight'])
G.add_nodes_from(data)
```

Fonte: Autor (2021).

Pelo Algoritmo 6, observa-se que as arestas (*"edges"*) e os nós (*"nodes"*) são tratados separadamente para que fossem manipulados de maneira que possibilitasse a distinção entre os trajetos. Para isso, foi criado um atributo chamado "*TIPO_VIAGEM*" que assume valores que representam o tipo de trajeto possível: viagens de bairros de classe pobre para bairros de classe média, viagens de bairros de classe pobre para bairros de classe rica e viagens de bairros de classe média para bairros de classe rica. Essas possibilidades de trajetos permitiram visualizar as diferenças de preços entre bairros de diferentes classes. As outras possibilidades de trajetos também foram analisadas e implementadas para permitir uma visão mais ampla das possíveis diferenças que poderiam ocorrer. Nota-se também a inclusão do preço como atributo de ponderação entre os trajetos. Isso pode ser visto na linha acima que utiliza o método *"from_pandas_edgelist"* para converter as arestas em um elemento integrante do tipo grafo, considerando o preço (*"weight"*) como peso. Os nós são convertidos em elemento integrante do tipo grafo pela utilização do método *"add_nodes_from"*.

Ainda em relação as métricas de centralidade, a Centralidade de Grau foi obtida pela utilização do método *"degree"*. O Algoritmo 7 exibe parte da implementação dessa métrica.

Algoritmo 7 - Parte da implementação da Centralidade de Grau.

```
#ALGORITMO 7

degree = G.degree(weight='weight')
max_degree = max(dict(degree).values())
```

Fonte: Autor (2021).

Observa-se pelo algoritmo acima que para realizar o cálculo de maneira consistente, é preciso considerar o peso das arestas que, no caso dessa pesquisa, é a média de preço de cada trajeto. A Centralidade de Grau indica que um *nó* importante está conectado com muitos *nós*. Como nessa pesquisa é considerado um peso para os trajetos possíveis, sendo este peso a média de preço, então isso significa que quanto mais alto o valor dessa métrica para um bairro (representado por um *nó*), então evidencia que esse bairro possui não só uma frequência alta nas ocorrências das viagens, mas também aquele que minimiza a média de preço para determinado trajeto. Isso significa que um nó importante para essa métrica e nesse contexto leva em consideração o custo do preço da viagem e não apenas a questão geográfica.

A Centralidade de Proximidade foi obtida pela utilização do método *"closeness_centrality"*. O Algoritmo 8 exibe parte da implementação dessa métrica.

Algoritmo 8 - Parte da implementação da Centralidade de Proximidade.

```
#ALGORITMO 8

closeness_centrality_weighted=nx.closeness_centrality(G,
distance='distance')
```

Fonte: Autor (2021).

Observa-se pelo algoritmo acima que o método recebe como parâmetros o grafo G e a distância (*"distance"*). Essa distância faz referência ao atributo preço da base de dados que foi introduzido anteriormente como elemento de ponderação no grafo. Nesse sentido, essa métrica

evidencia que um *nó* importante está próximo dos outros *nós*. Isso significa, para o contexto dessa pesquisa, que bairros próximos um dos outros apresentarão um valor alto para essa métrica. Como a ponderação é feita baseada no preço, o resultado dessa métrica evidencia essa proximidade considerando as médias de preços das viagens. Os valores obtidos podem mostrar que por meio dos preços de viagens é também possível indicar a proximidade entre locais por uma perspectiva de preços e não somente por uma perspectiva geográfica.

A Centralidade de Intermediação foi obtida pela utilização do método *"betweenness_centrality"*. O Algoritmo 9 exibe parte da implementação dessa métrica.

Algoritmo 9 - Parte da implementação da Centralidade de Intermediação.

```
#ALGORITMO 9

betweenness_centrality_weighted = nx.betweenness_centrality(G,
weight='weight')
```

Fonte: Autor (2021).

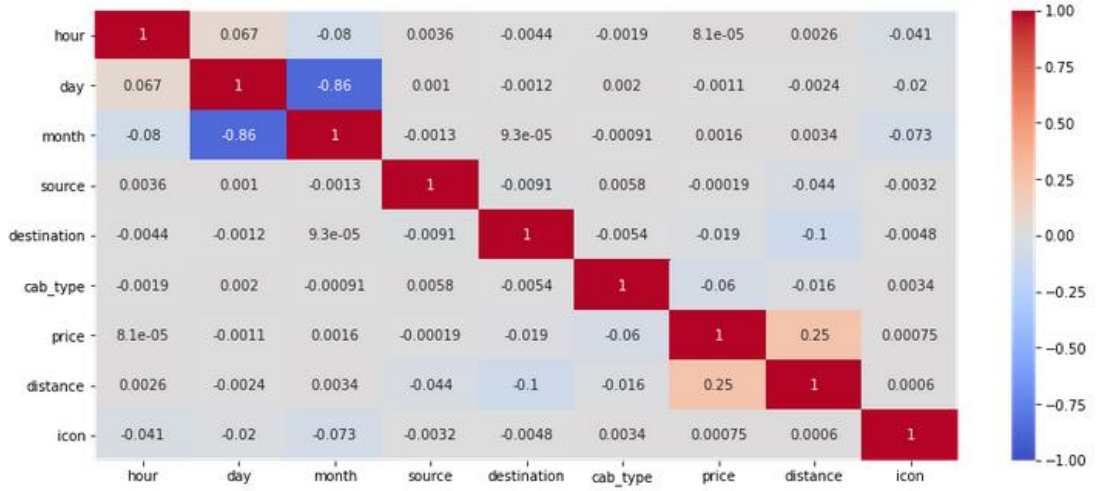
Observa-se pelo algoritmo acima que o peso também foi passado como parâmetro para o método considerado. Essa métrica indica o número de vezes que um *nó* age como ponte ao longo de um trajeto. Isso significa que um *nó* importante faz parte de muitos trajetos. No contexto dessa pesquisa, um valor alto para essa métrica indica que um bairro serve como ponte para vários trajetos de viagens de outros bairros.

3.5 Criação de Modelos Preditivos

Como informado anteriormente, a base de dados para a cidade de Boston é real. Ou seja, um compilado de informações acerca de várias viagens da UberX entre os bairros dessa cidade. As análises feitas nessa base servem para fundamentar os estudos sobre as viagens da UberX em Fortaleza, no sentido de possibilitar evidências de que mesmo em outro contexto (outro país com condições socioeconômicas, clima e outros aspectos diferentes) muitos comportamentos relacionados ao preço podem ser semelhantes independentemente da localidade. Nessa perspectiva, a base de dados de Boston possui alguns atributos para cada viagem, como hora, dia, mês, origem, destino, preço, distância e clima.

Considerando que a pesquisa foca nos preços das viagens, então foi realizado uma análise por meio de uma Matriz de Correlação na qual serviu para verificar quais pares de atributos se correlacionavam mais. A Figura 12 ilustra essa matriz.

Figura 12 - Matriz de Correlação para a base de dados de Boston.



Fonte: Autor (2021).

A Matriz de Correlação foi obtida por meio do método "*corr*" disponível no módulo "*LabelEncoder*" da biblioteca *"sklearn"*. Essa matriz evidencia que quanto mais próximo de 1 for a intersecção entre 2 atributos, mais correlacionados eles são. De outro modo, quanto mais próximo de -1, menos correlacionados são. Mediante isso, observa-se que os atributos preço (*"price"*) e distância (*"distance"*) é o par de maior correlação entre os demais. Mediante isso e as pesquisas encontradas para a cidade de Fortaleza, observou-se que o preço e a distância são grandes influenciadores do serviço de viagens da UberX em detrimento de outros atributos, como clima, hora, dia e outros.

Considerando essa conjuntura e as análises para verificar a convergência de tendência para os preços de ambas as cidades, foi proposta uma nova funcionalidade para a empresa Uber. Isso foi levantado porque observou-se que as viagens desse serviço na Uber oferecem as viagens com seus preços já estipulados. Entretanto, nesse seguimento de mercado existem algumas empresas que adotam o mecanismo inverso, no qual o usuário oferta um preço e o aplicativo de transporte retorna as melhores distâncias para aquele preço ofertado. Um exemplo desse tipo de estratégia é adotado pela empresa *InDriver* (2022). Nesse sentido, foi proposta a criação de Modelos de Aprendizagem de Máquina para previsão da variável dependente "distância"

baseada na variável independente "preço" e em outras variáveis independentes na base de dados de Boston. Dessa forma, seria possível fornecer mais liberdade ao cliente para escolher o preço de uma viagem e, consequentemente, possibilitar uma redução de custos.

Os algoritmos de Aprendizagem de Máquina utilizados para criação destes modelos foram escolhidos por serem bastante utilizados pela literatura no âmbito da regressão, considerando o domínio de dados utilizado nesta pesquisa. Os algoritmos são: Regressão Linear Simples, Regressão SGD, Árvore de Decisão e Floresta Aleatória.

Para construção dos modelos supracitados, foi utilizada a biblioteca *"sklearn"* que possui vários métodos que implementam variados regressores, além de métricas, validação, pré-processamentos e modos de seleção para modelos nesse contexto. O método utilizado para realizar a divisão dos dados para treino e teste foi o *"train_test_split"*. Os métodos para criação dos modelos em estudo foram, respectivamente, *"LinearRegression"*, *"DecisionTreeRegressor"*, *"SGDRegressor"* e *"RandomForestRegressor"*, para Regressão Linear Simples, Árvore de decisão, Regressão SGD e Floresta Aleatória. O método para verificação da acurácia dos modelos foi o *"r2_score"*. O método utilizado para padronizar as variáveis da base foi *"StandardScaler"*. O método para realizar a validação cruzada foi o *"cross_val_score"*.

Os parâmetros utilizados nos métodos seguiram os recomendados pela literatura no que se diz respeito ao domínio dos valores envolvidos, bem como o tamanho da base de dados considerada. Nesse sentido, para o método *"train_test_split"* foi utilizado um *"test_size"* de 20% e 80% para treino. Além disso, utilizou um *"random_state"* com fator de 42. No mesmo caminho, para o método *"cross_val_score"* foi utilizado o *"scoring"* com fator em R^2 e o parâmetro cv em 5 que indica o número de partições (*"folders"*).

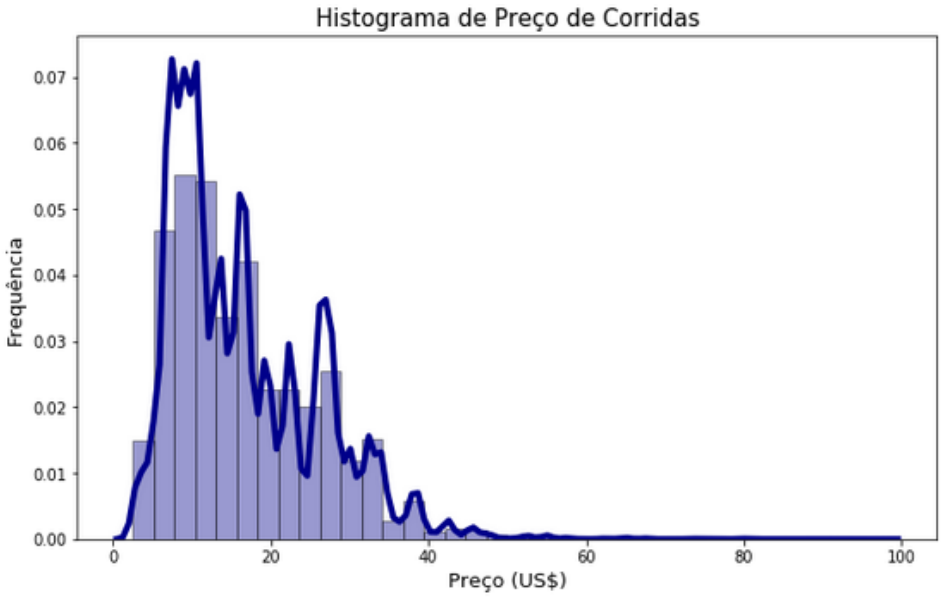
RESULTADOS E DISCUSSÕES

Neste capítulo são descritos os resultados e as discussões acerca da Análise Exploratória de Dados (AED) utilizada nas cidades em estudo. Também são descritos e discutidos os resultados obtidos para os modelos de predição de distância para a cidade de Boston, bem como o comportamento dos preços considerando as mudanças nos trajetos de viagens para a cidade de Fortaleza. Ainda para essa última cidade, são descritos os resultados e discussões acerca das análises estatísticas e medidas de centralidade. Por fim, é discutida uma possível mudança nos preços das viagens da Uber em Fortaleza, caso houvesse uma descentralização de seu centro financeiro.

3.6 Boston

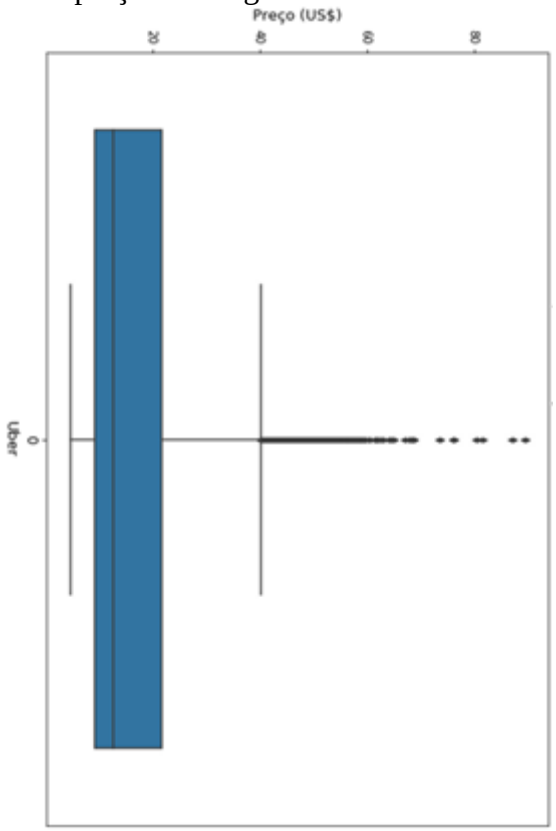
Primeiro plotou-se os histogramas das variáveis *"price"* (preço) e *"distance"* (distância). O Gráfico 3 e o Gráfico 4 ilustram a relação entre frequência e preço, e o Gráfico 5 entre frequência e distância, respectivamente.

Gráfico 3 - Relação entre frequência e preço para a cidade de Boston



Fonte: Autor (2021).

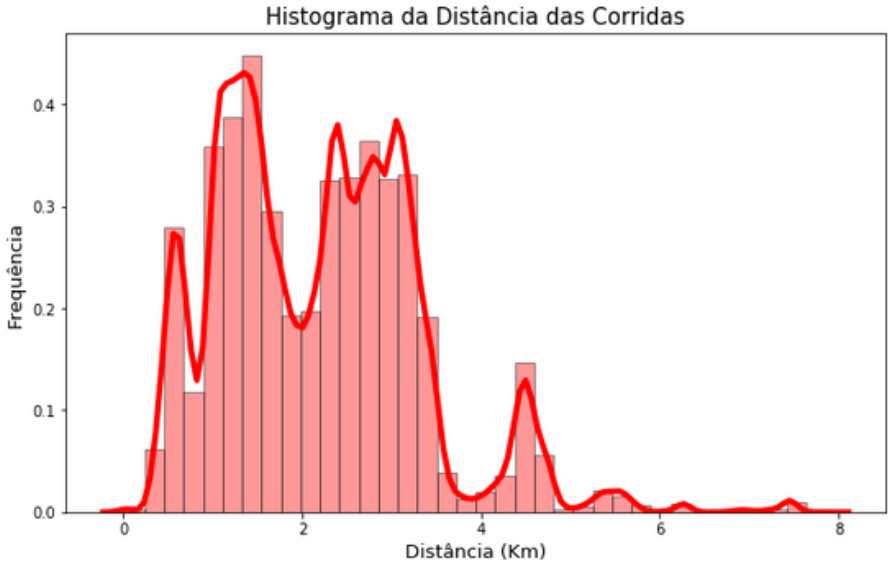
Gráfico 4 - Boxplot de preços de viagens Uber em Boston.



Fonte: Autor (2021).

Pelo *boxplot* acima, percebe-se a existência de *outliers* (pontos fora da curva) da variável preço, o que corrobora a análise realizada anteriormente pelo histograma. Além disso, esses dois gráficos ilustram a mesma informação, sendo que no *boxplot* a distribuição empírica está sendo vista por quartis. É possível verificar que o valor da mediana do preço da empresa Uber fica abaixo dos $ 20,00 para a cidade de Boston.

Gráfico 5 - Relação entre frequência e distância para a cidade de Boston.

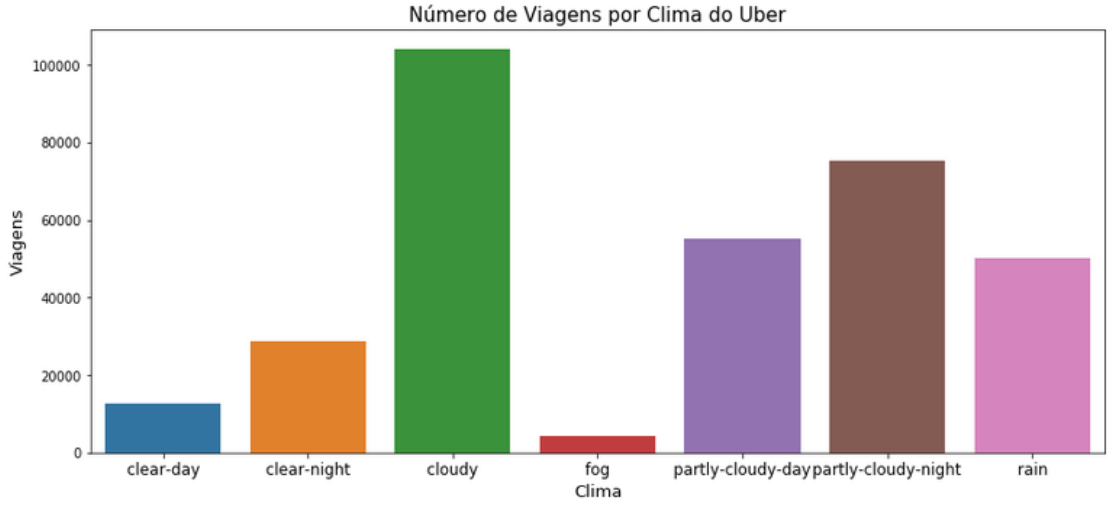


Fonte: Autor (2021).

Nota-se que ambas as variáveis preço e distância possuem uma assimetria negativa (mais puxada para esquerda), porém, a primeira apresenta valores de frequência menor, o que indica tamanha dispersão dos valores nos dados se comparada com a segunda. Os gráficos também revelam que há uma tendência em viajar menos para distâncias maiores, pois o preço tende a aumentar.

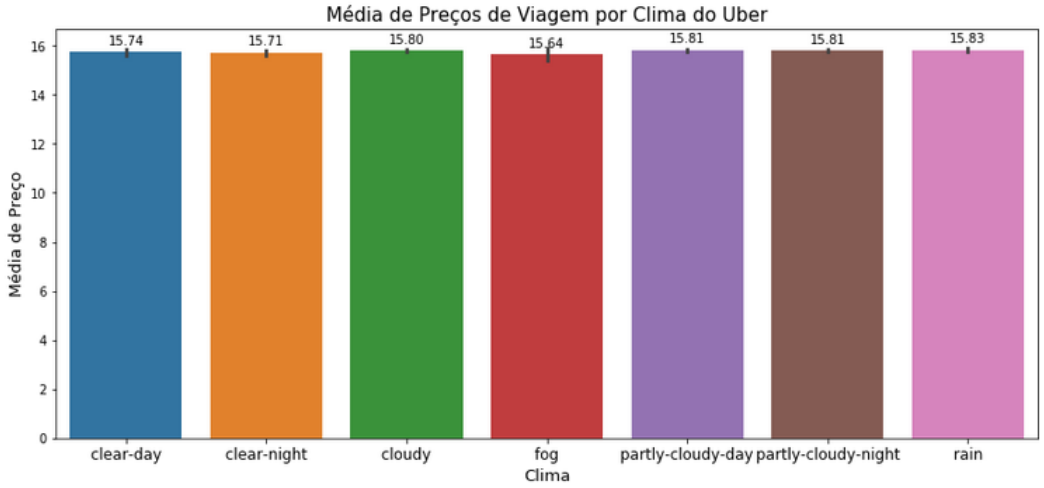
Em seguida, foram plotados dois gráficos de barras: o primeiro (Gráfico 6) com a quantidade de viagens solicitadas de acordo com o clima e o segundo (Gráfico 7) com a média de preços de viagens por clima.

Gráfico 6 - Quantidade de viagens por clima em Boston.



Fonte: Autor (2021).

Gráfico 7 - Média de preços de viagens por clima em Boston.

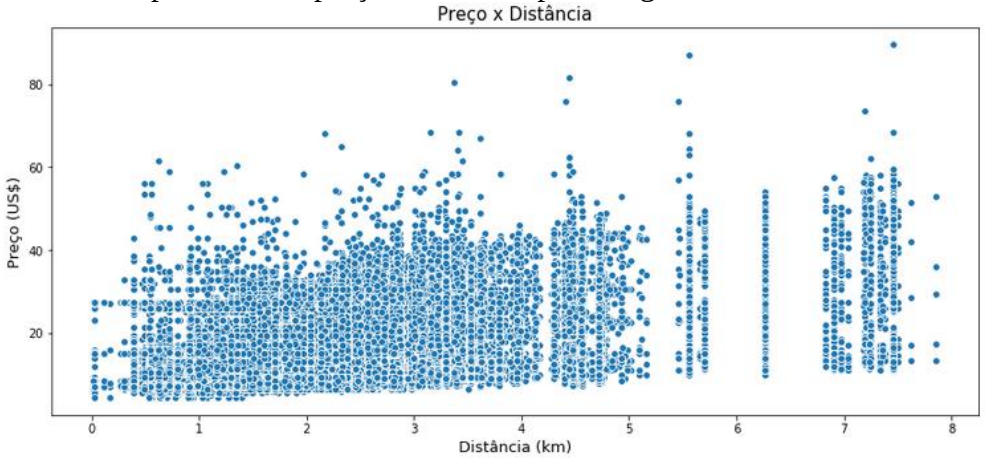


Fonte: Autor (2021).

No primeiro gráfico, foi observado que as viagens são solicitadas em sua maioria quando o clima está nublado ou chuvoso. No segundo gráfico, percebe-se que o clima não afetou a média de preços das viagens da Uber.

Um gráfico de dispersão (Gráfico 8) também foi plotado, para mostrar a variabilidade do preço em relação à distância das viagens. Dessa forma, seria possível verificar o quanto os *Outliers* estavam presentes na base, bem como a coerência entre distância e preço no sentido de seguirem ou não uma proporcionalidade crescente de valores.

Gráfico 8 - Dispersão entre preço e distância para viagens da Uber em Boston.

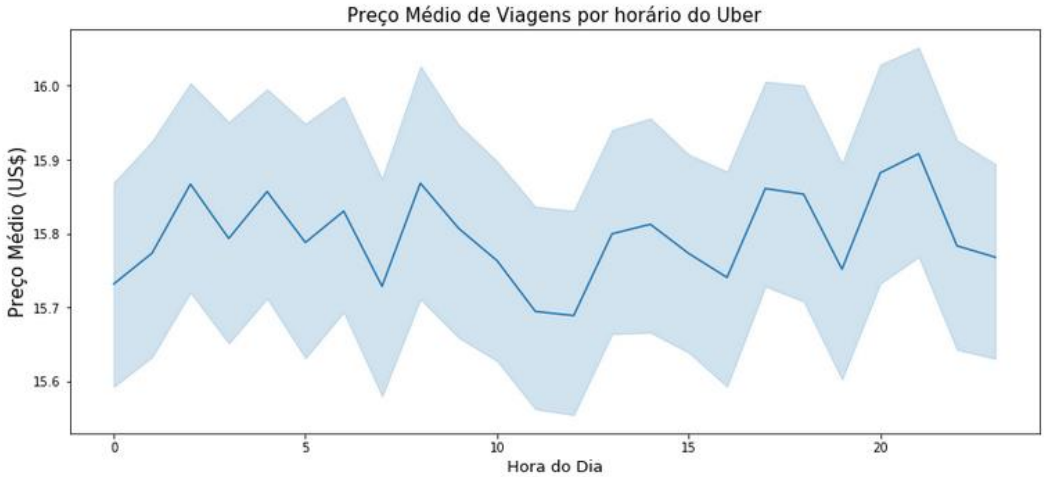


Fonte: Autor (2021).

Nota-se que há viagens na Uber com distâncias maiores e preços menores, mas também há viagens com distâncias menores e preços maiores. Os resultados observados no capítulo anterior pela Matriz de Correlação ajudam a entender esse comportamento. Distância e preço, entre os valores de todos os pares de variáveis da base, são as mais correlacionadas.

O último gráfico referente a AED foi um de série temporal (Gráfico 9) que ilustra o comportamento do preço médio das viagens ao longo do dia.

Gráfico 9 - Variabilidade do preço ao longo do dia.



Fonte: Autor (2021).

O Gráfico 9 evidencia que geralmente é mais vantajoso para o cliente utilizar o serviço Uber em horários em que ocorre menor tráfego para a cidade de Boston. Nesse sentido, em um horário onde exista maior demanda de usuários, o preço é reajustado devido à intensidade das procuras por deslocamentos.

O Gráfico 10 abaixo indica a média de preços por destino de solicitação. Por esse gráfico, percebe-se que os preços das viagens tendo como destino os bairros Fenway, Finacial District, Boston University e Northeastern University tendem a ser mais caros em relação aos demais bairros.

Gráfico 10 - Média de preços por Destino de Viagem.

Fonte: Autor (2021).

O Financial District é o centro financeiro da cidade, o que pode explicar o maior valor no preço. Como observado na seção 2.2.2, as universidades de Boston e Northeastern atraem muitos estudantes que utilizam os serviços de transporte por aplicativo. Devido a isso, embora essas universidades possuam um preço um pouco elevado em relação aos demais bairros de destino, elas possuem uma média de preço menor do que as viagens para o centro financeiro. Isso pode indicar que os bairros que estão um pouco mais distantes do centro financeiro, podem possibilitar uma diminuição dos preços de viagens se comparados com os preços praticados para o centro financeiro, mesmo estes bairros estando um pouco mais distantes do centro financeiro e possuindo fatores de atração como universidades, hotéis, museus, etc.

Outro exemplo de região um pouco mais afastada do centro financeiro e com o preço um pouco elevado, porém menor se comparado com a média de preço do centro financeiro é o

bairro de Fenway, um bairro mais pobre se comparado a maioria dos bairros analisados. Esse bairro está a pouco mais de 2 km de distância do centro, uma distância alta se comparada as demais regiões em estudo. Entretanto, sua média de preço é um pouco alta, mas menor do que a média de preço de viagens para o centro financeiro. Fenway apresenta diversos locais de atração de pessoas, como estádios de jogos, bares e restaurantes. Isso pode indicar que uma descentralização de centro financeiro tende a diminuir preços de viagens.

Por outro lado, regiões como Haymarket Square e North End estão localizados bem próximos do centro financeiro da cidade e são regiões nobres, como constatado na seção 2.2.2, indicando que os deslocamentos em termos de distância são menores, influenciando a baixa na média dos preços. Para as demais localidades analisadas nesse estudo para Boston, o comportamento de alta de preços para regiões um pouco mais afastadas foi o mesmo do observado para as regiões supracitadas.

Após a AED e as ponderações da Matriz de Correlação, procedeu-se para a criação dos modelos de Aprendizagem de Máquina, considerando a variável distância como dependente em relação as outras da base. Como os dados manipulados são de domínio contínuo e estão sendo tratados por regressores, então obteve-se os resultados dos Coeficientes de Determinação (R^2) para cada um dos regressores. Para isso, utilizou-se a técnica de Cross Validation em que se determinou um número de *folds* igual a 5 e o parâmetro de avaliação sendo o R^2 que faz referência ao coeficiente supracitado. Nesse sentido, foram obtidos 5 valores de R^2 para cada regressor em estudo os quais foram calculadas suas respectivas médias. Os resultados podem ser observados na Tabela 4 abaixo.

Tabela 4 – Resultados dos Coeficientes de Determinação Médio.

TÉCNICA	R^2 MÉDIO
REGRESSÃO LINEAR	**8%**
REGRESSÃO SGD	**7%**
ÁRVORE DE DECISÃO	**91%**
FLORESTA ALEATÓRIA	**94%**

Fonte: Autor (2021).

A Floresta Aleatória é um dos algoritmos de Aprendizagem de Máquina mais utilizados nas regressões de valores em conjuntos de dados de domínio contínuo e nas classificações de valores binários. Quando esse algoritmo é utilizado para combinar a predição de um conjunto

de Árvores de Decisão para obter uma única resposta como saída, então tende a apresentar melhor desempenho que a obtida com cada árvore de um modelo isoladamente devido à possibilidade de redução de variância (ALVARENGA, 2018). Nesse sentido, observa-se pela tabela acima que para este algoritmo foi obtido um R^2 Médio de 94%, a maior dentre as demais técnicas utilizadas.

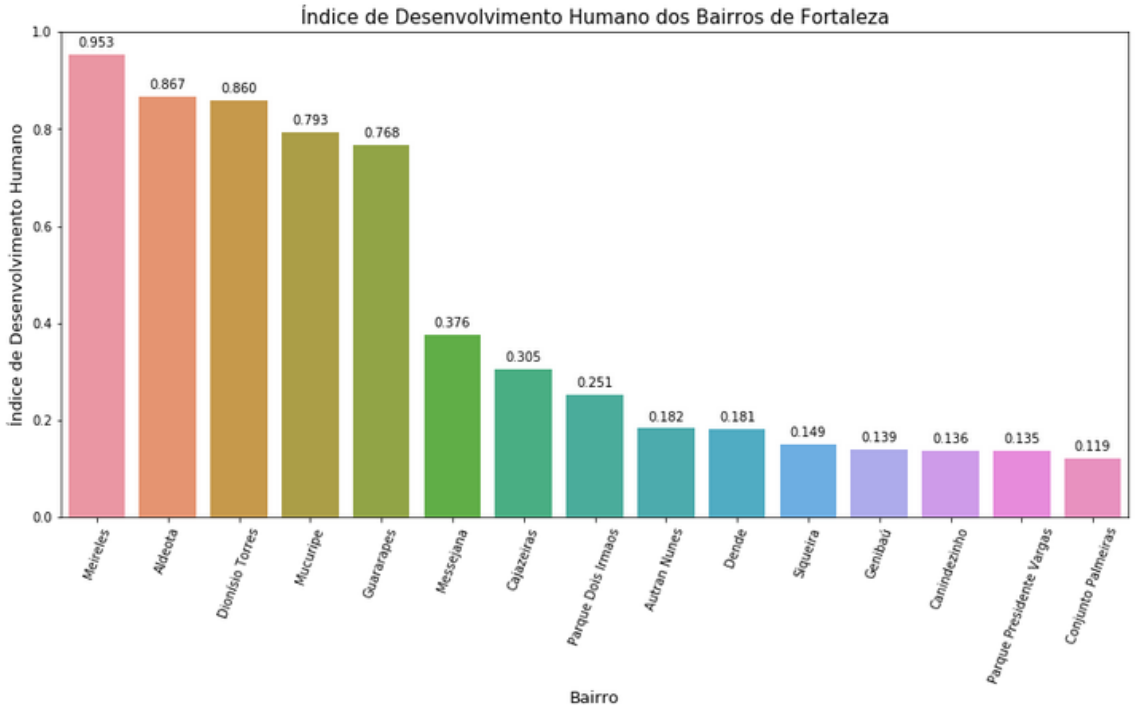
Como levantado na Revisão da literatura, a concepção de preços de viagens segue um comportamento não linear. Essa característica faz com que o desempenho dos modelos não lineares (Árvore de Decisão e Floresta Aleatória) seja maior em relação aos demais algoritmos lineares utilizados nesta pesquisa. Além disso, isso significa que o modelo de Floresta Aleatória se ajusta bem aos dados e as variáveis preditoras explicam 94% dos dados de distância previstos.

Importante salientar que essa predição da distância é apenas uma análise que pode servir para recomendações de políticas públicas, de modo que mais estudos podem ser utilizados no intuito de validar os benefícios dessa predição para usuários do serviço de viagens da Uber.

3.7 Fortaleza

Um dos principais fatores que determinam as disparidades socioeconômicas entre os bairros em estudo de Fortaleza é o IDH. No Gráfico 11, nota-se a diferença dos índices entre os mais desenvolvidos para os menos desenvolvidos.

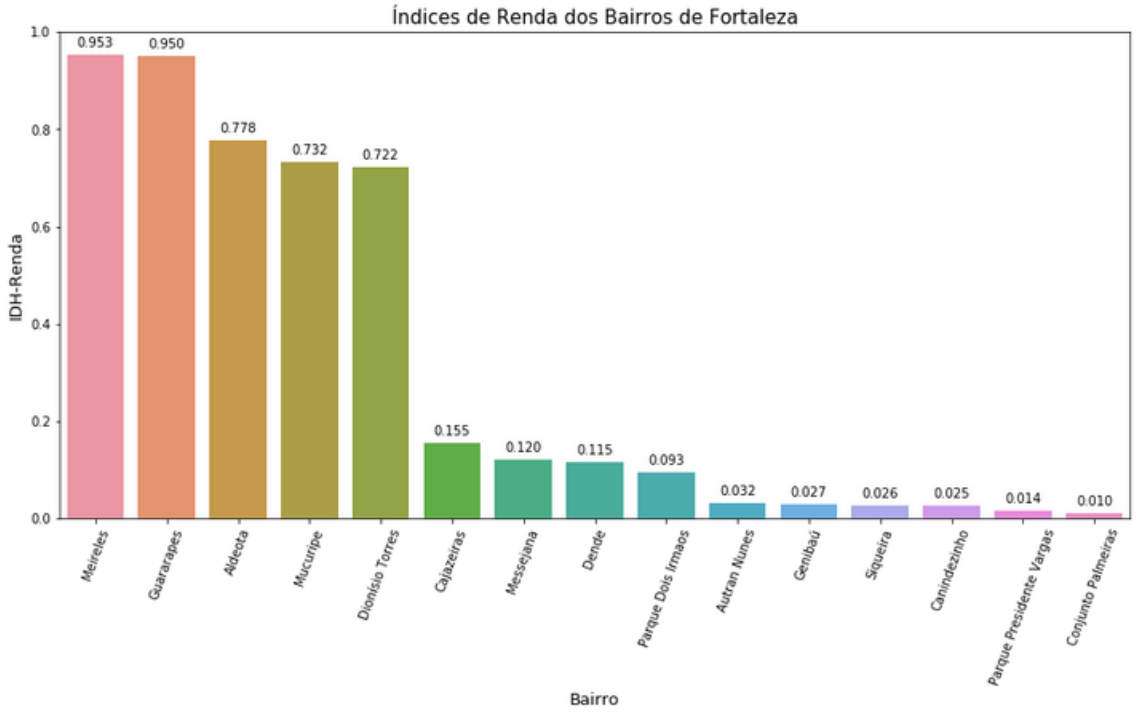
Gráfico 11 - IDH dos bairros em estudo para Fortaleza.



Fonte: Autor (2021).

Semelhante ao gráfico do IDH, observa-se o Gráfico 12 que ilustra as diferenças da renda média pessoal dos habitantes com 10 anos de idade ou mais de cada bairro em estudo.

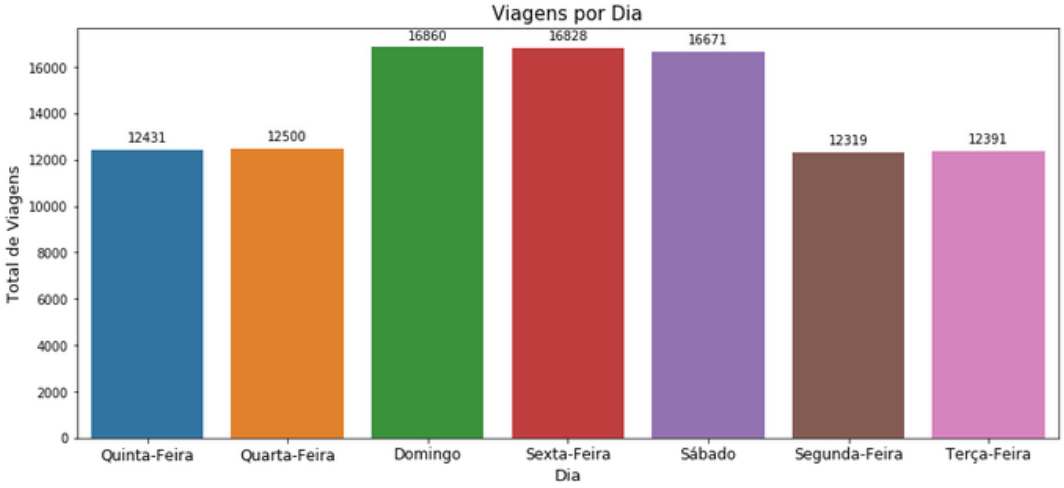
Gráfico 12 - IDH - Renda médio de habitantes de alguns bairros em Fortaleza.



Fonte: Autor (2021).

No gráfico a seguir (Gráfico 13), observa-se que os dias de sextas-feiras e finais de semana estão entre os dias preferidos dos clientes da Uber para utilização do serviço.

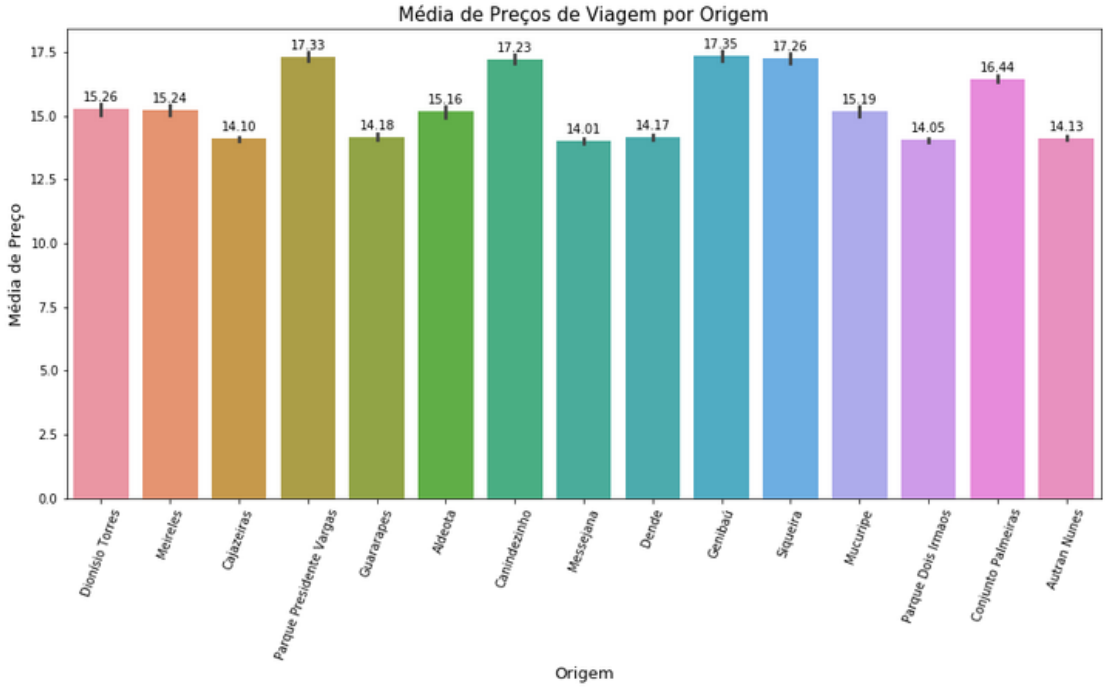
Gráfico 13 - Divisão de viagens por em Fortaleza.



Fonte: Autor (2021).

No Gráfico 14, percebe-se a disparidade das médias dos preços das viagens de acordo com a origem.

Gráfico 14 - Preço médio por bairro de origem da solicitação do serviço.

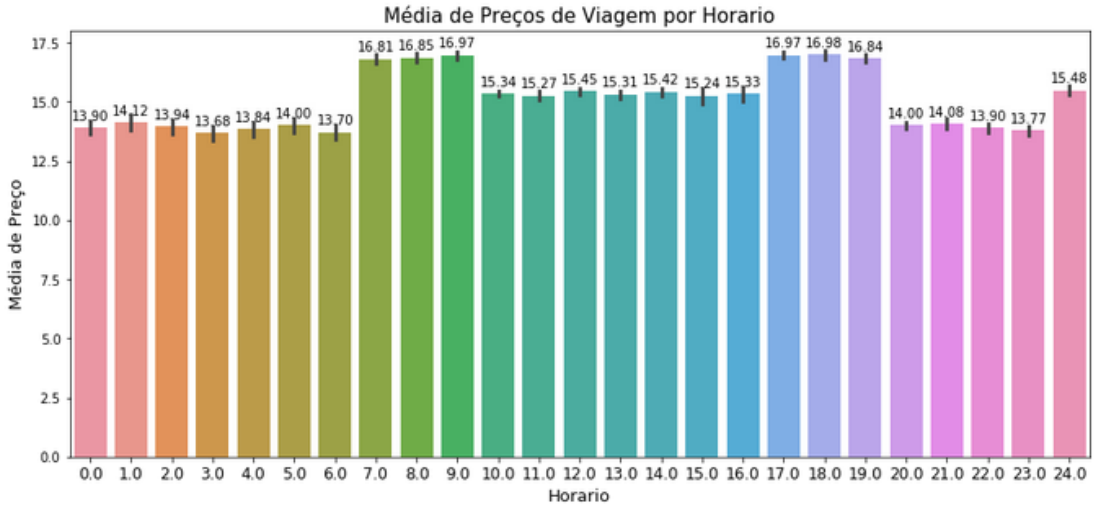


Fonte: Autor (2021).

Nota-se que usuários de bairros mais pobres e com menos desenvolvimento em Fortaleza acabam gastando mais com deslocamento via transporte de aplicativo do que aqueles que partem de bairros mais ricos. Por exemplo, considerando o bairro pobre Canindezinho, a média de preço de um usuário que solicita uma viagem desse bairro como origem é de R$ 17,23 sendo o destino qualquer um dos outros bairros em estudo.

O Gráfico 15 ilustra o preço médio gasto de acordo com o horário das viagens. Pode-se ter uma noção de quanto é mais vantajoso utilizar o serviço da Uber a noite a partir das 20 horas em Fortaleza.

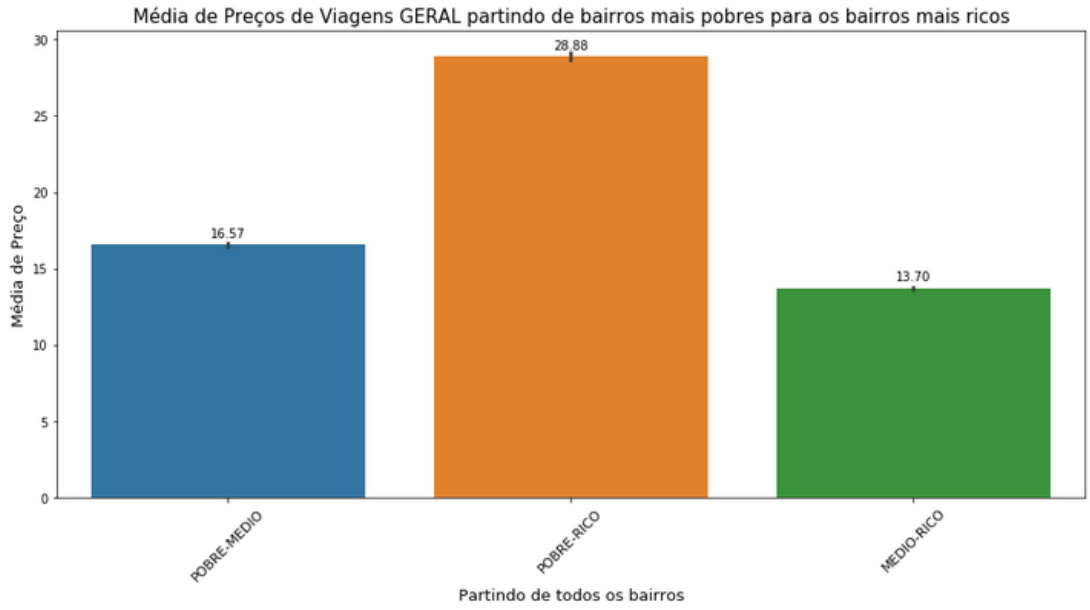
Gráfico 15 - Média de preços das viagens por horário em Fortaleza.



Fonte: Autor (2021).

O Gráfico 16 ilustra a média de preços considerando cada classe dos bairros em estudo, ou seja, pobre, médio e rico. Assim, é possível analisar o comportamento dos preços por uma visão mais holística, distinguindo os trajetos de viagens por classe de bairro.

Gráfico 16 - Média de preços por trajeto de bairros mais pobres para ricos.

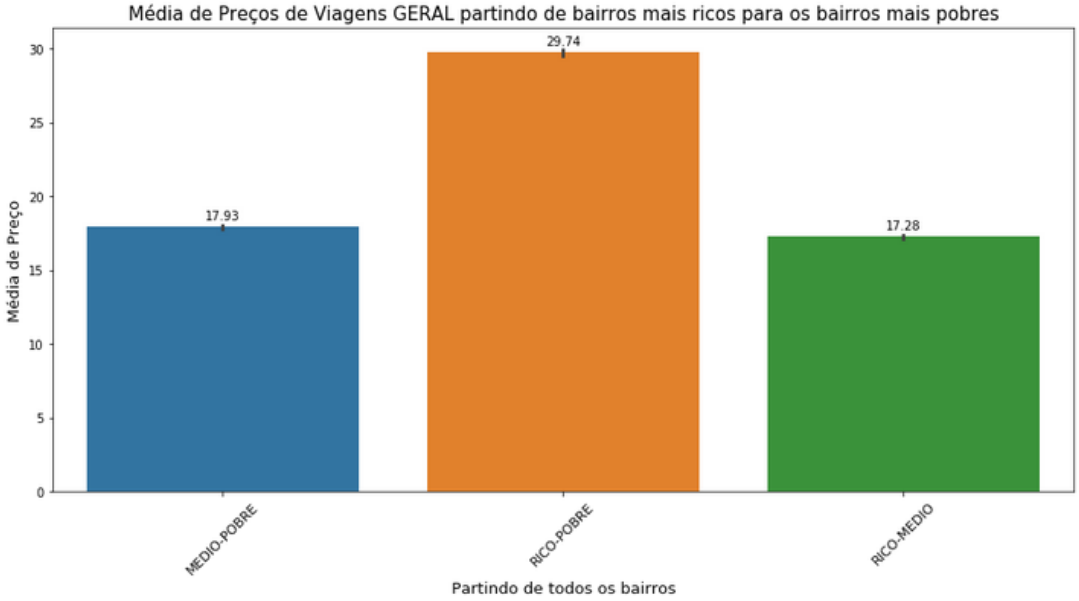


Fonte: Autor (2022).

Nota-se pelo gráfico acima que o trajeto que envolve a classe de bairros pobre e rico possui a média de preços mais elevada. Em contrapartida, o trajeto entre os bairros de classe média e rica possui a menor média. Isso corrobora com as pesquisas para essa cidade no sentido de que os bairros médios, além de estarem mais próximos geograficamente do centro comercial e dos bairros mais ricos, possuem um custo menor ao solicitarem viagens da Uber em detrimento dos bairros pobres.

De maneira complementar ao Gráfico 16, plotou-se o Gráfico 17 abaixo que ilustra o trajeto inverso do demonstrado anteriormente, ou seja, viagens de bairros mais ricos para bairros mais pobres.

Gráfico 17 - Média de preços por trajeto entre bairros mais ricos para pobres.

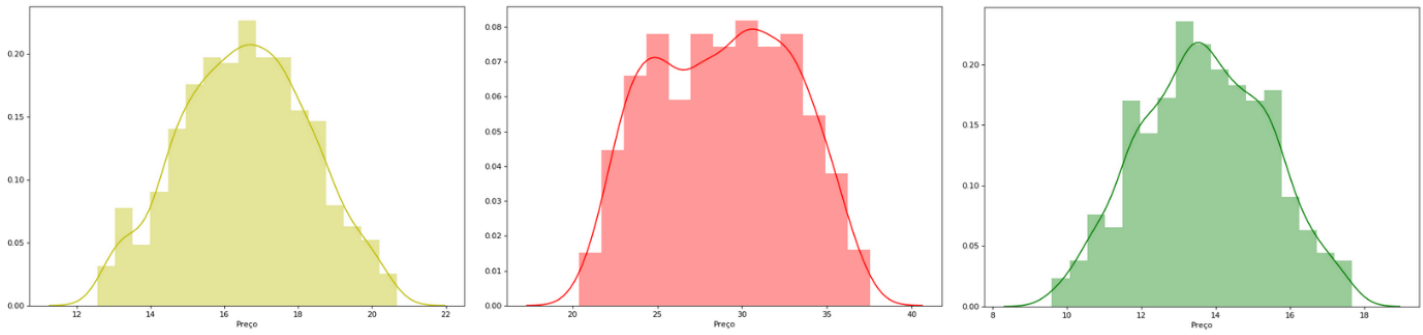


Fonte: Autor (2022).

Observa-se pelo gráfico acima que o comportamento de alta de preços se repete, porém, com um aumento nas médias. Isso pode evidenciar que o retorno dos usuários de bairros mais pobres geralmente paga ainda mais caro se comparado com as viagens de ida. Isso pode estar relacionado a questões de segurança, mobilidade, além da oferta e da demanda.

Após a AED, foi realizada uma análise estatística para essa base de dados de Fortaleza. O intuito dessa análise consistiu em verificar a validade dessa base simulada em termos estatísticos, para que fosse possível extrair informações mais consistentes com a realidade. Nesse sentido, foi verificado se as amostras dos trajetos em estudo seguiam uma Distribuição Normal. Cada amostra consistiu em 1000 viagens, considerando o filtro desejado de trajeto: Pobre – Rico, Pobre – Médio e Médio – Rico. O primeiro trajeto verificado foi o pobre – rico. O Gráfico 18 a seguir ilustra o comportamento das amostras para os trajetos Pobre-Médio, Pobre-Rico e Médio-Rico, respectivamente.

Gráfico 18 - Distribuições das amostras para os trajetos Pobre-Médio, Pobre-Rico e Médio-Rico.



Fonte: Autor (2022).

Nota-se pelo gráfico acima um comportamento com tendência simétrica em relação ao eixo das coordenadas. Entretanto, isso não é suficiente para determinar se as amostras seguem uma Distribuição Normal. Devido a isso, foi realizado o Teste de Normalidade de Kolmogorov-Smirnov para validação.

Considerando os parâmetros discutidos na seção 2.5, o valor de *p* foi menor do que de *alpha* (α). Isso significa que a hipótese é rejeitada e essa distribuição não é Normal. Os demais trajetos em estudo também apresentaram o mesmo comportamento, seguindo uma distribuição não Normal. Seus gráficos podem ser consultados no Apêndice. A Tabela 5 contém os valores *p* para os trajetos em estudo, considerando o valor *alpha* atribuído.

Tabela 5 - Valores de p para os trajetos em estudo.

TRAJETO DE VIAGEM	VALOR DE *P* (*α de* 5%)
POBRE PARA MÉDIO	**3.9E-06**
POBRE PARA RICO	**2.7E-55**
MÉDIO PARA RICO	**1.3E-06**

Fonte: Autor (2022).

Como o valor *p* foi menor do que o nível de significância *alpha*, então a probabilidade de obter dados como estes é muito pequena. Assim, pode-se concluir que os valores dos preços das viagens não seguem uma Distribuição Normal. Em outras palavras, como os resultados obtidos acima mostraram que os preços das viagens não seguem uma Distribuição Normal, foi

utilizado o Teste não Paramétrico de Wilcoxon. Um exemplo de par considerado para esse teste foram os trajetos pobre-rico e pobre-médio.

Considerando um nível de confiança de 95% e nível de significância *alpha* de 5%, foi determinado na implementação que se testasse a hipótese alternativa (teste unilateral a direita) de que a média dos preços das viagens do trajeto pobre-rico fosse maior do que a média dos preços do trajeto pobre-médio. Como resultado, obteve-se um valor *p (1.7E-16)* menor do que o valor *alpha*, evidenciando de maneira estatisticamente significativa que a média de preços das viagens de pobre para rico é maior do que a média de preços das viagens de pobre para médio. A Tabela 6 abaixo informa os valores de *p* para todas as combinações de pares de trajetos.

Tabela 6 - Valores de p para o Teste de Wilcoxon.

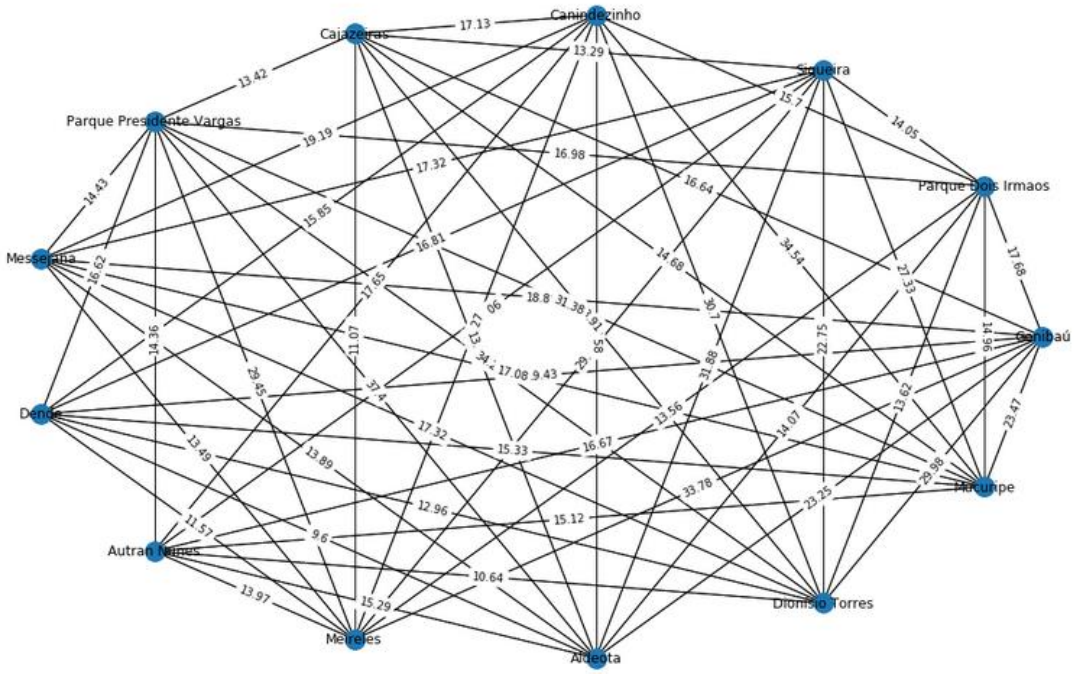
PAR DE TRAJETO	VALOR DE P (α *de* 5%)
POBRE-RICO / POBRE-MÉDIO	**1.7E-16**
POBRE-RICO / MÉDIO-RICO	**1.7E-16**
POBRE-MÉDIO / MÉDIO-RICO	**1.0E-13**

Fonte: Autor (2022).

Em outras palavras, os grupos diferem estatitistacamente de maneira representativa e corroboram com os histogramas de média de preços das viagens entre bairros ricos, pobres e médios (Gráfico 16 e Gráfico 17).

Em seguida, foi realizada uma análise dessas amostras para algumas medidas de centralidade para grafos. Essa análise teve importância no sentido de enfatizar o quão importante um bairro pode ser na influência da determinação de preços em viagens de transporte por aplicativo. Nesse sentido, primeiramente foram convertidas as amostras em estudo em objeto do tipo grafo considerando a discussão do capítulo metodologia. Após a conversão, foi possível plotar os trajetos em termos de suas ligações de viagens com suas respectivas médias de preços. A Figura 13 ilustra o grafo obtido.

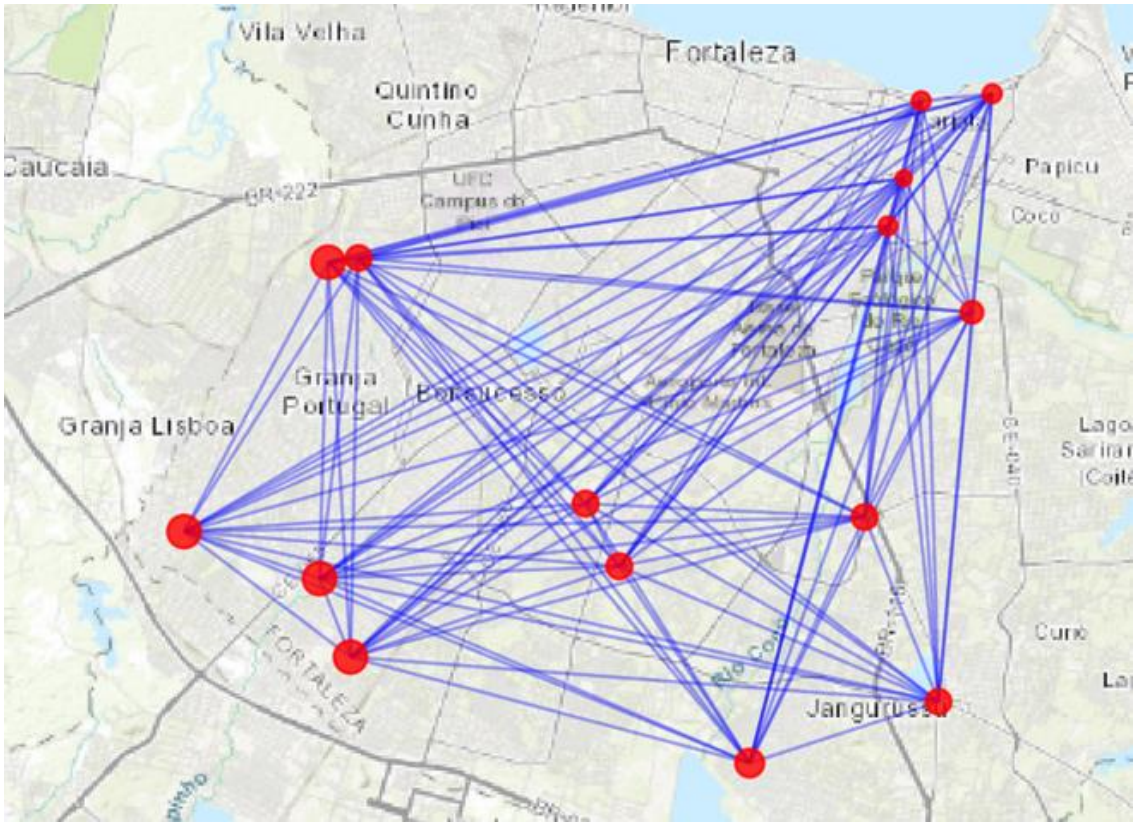
Figura 13 - Grafo das ligações entre bairros em estudo.



Fonte: Autor (2022).

Para um posicionamento mais condizente com a realidade, foi plotado um grafo desses bairros sobre um mapa da cidade de Fortaleza. Para isso, foram coletadas as informações de latitude e longitude de cada bairro no intuito de manter a realidade das distâncias e a proporcionalidade. Dessa forma, seria possível visualizar de maneira geográfica as distâncias entre os bairros. A Figura 14 ilustra essa representação.

Figura 14 - Grafo dos bairros em estudo sobre mapa de Fortaleza.

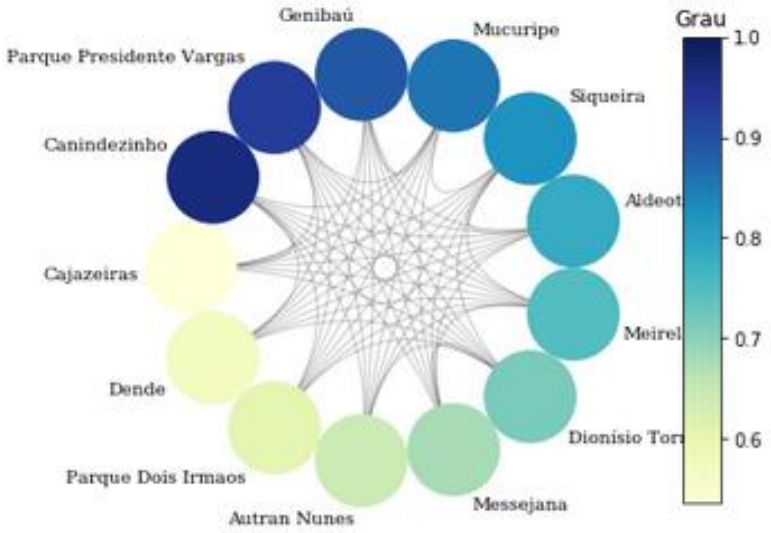


Fonte: Autor (2022).

Pela figura acima, nota-se a grande distância entre os bairros pobres localizados na parte Sudoeste do mapa e os bairros ricos situados na porção Nordeste. Por sua vez, os bairros médios estão representados pelos nós na linha transversal no mapa.

Após a conversão das amostras em um grafo, foi possível determinar as medidas de centralidade. As medidas utilizadas foram: Centralidade de Grau, Centralidade de Proximidade e Centralidade de Intermediação. O Gráfico 19 a seguir indica a Centralidade de Grau para os dados em análise.

Gráfico 19 - Centralidade de Grau para os dados em análise.

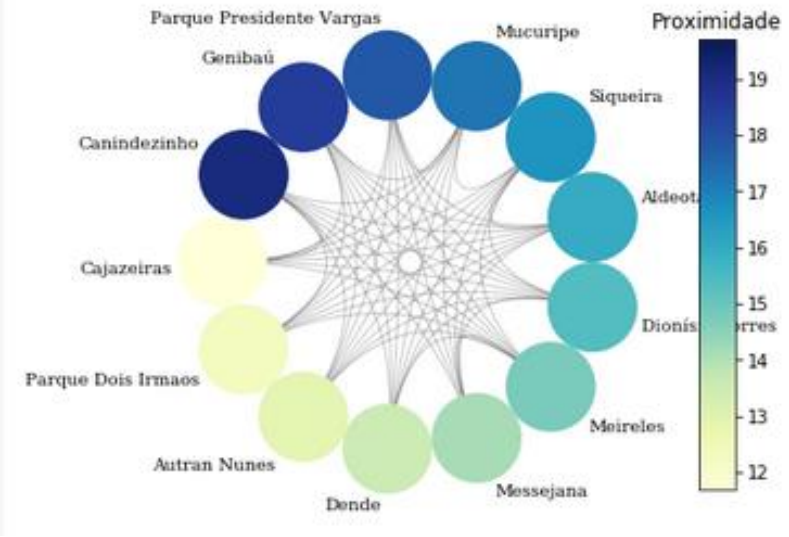


Fonte: Autor (2022).

Nota-se pelo gráfico acima que os bairros mais pobres em estudo (Canindezinho, Parque Presidente Vargas) apresentam maior Centralidade de Grau. Isso significa que esses bairros possuem não só uma frequência alta de solicitações de viagens, mas também aqueles que minimizam a média de preços para determinado trajeto. Isso leva em consideração o contexto da pesquisa que possui como peso nas arestas as médias de preços, bem como a indicação de que um *nó* importante em um grafo está conectado a muitos *nós*. Dessa forma, isso evidencia que para esse contexto, um bairro importante leva em consideração o custo do preço de uma viagem e não apenas a questão geográfica.

A próxima medida analisada foi a Centralidade de Proximidade. O Gráfico 20 abaixo ilustra seu comportamento.

Gráfico 20 - Centralidade de Proximidade para os dados em análise.



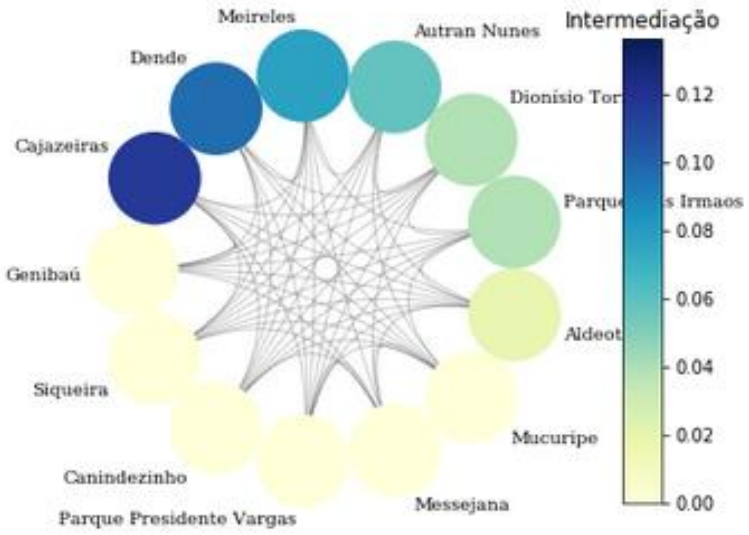
Fonte: Autor (2022).

Nota-se novamente altos valores para os bairros mais pobres. Isso acontece pelo fato de que muitos bairros pobres estarem próximos um dos outros em termos de média de preços. Ou seja, considerando o bairro pobre de Canindezinho como exemplo, isso indica que ele está próximo da maioria dos bairros pobres analisados. Em outras palavras, uma média mais próxima dos demais. Por outro lado, Cajazeiras (bairro de classe média) é um bairro que apresenta uma maior distância em média de preços entre os bairros médios analisados.

No contexto dessa pesquisa, isso significa que essa proximidade não se restringe a questão geográfica. Os bairros pobres considerados, por estarem próximos geograficamente, apresentaram médias de preços próximas, por isso um alto valor dessa medida para os bairros pobres e valores menores para os bairros médios e ricos, uma vez que estes estão mais distantes entre si. Nesse sentido, é importante salientar que o peso considerado nos experimentos foi a média de preços, o que nos leva a evidenciar que por meio dos preços de viagens também é possível indicar a proximidade entre locais por uma perspectiva de preços e não somente por uma perspectiva geográfica.

Por último, foi analisada a Centralidade de Intermediação. O Gráfico 21 a seguir ilustra seu comportamento.

Gráfico 21 - Centralidade de Intermediação para os dados em análise.



Fonte: Autor (2022).

Pelo gráfico acima, nota-se autos valores para bairros de classe média (como Cajazeiras e Dendê, além de outros de classe rica como Meireles). Essa métrica indica o número de vezes que um *nó* age como ponte ao longo de um trajeto. Como geograficamente os bairros de classe média ficam na porção central da cidade, isso evidencia o quanto eles são ponte entre os bairros pobres que ficam na porção territorial Sul e Sudoeste e os bairros ricos que ficam centralizados na porção Norte e Nordeste. A presença do bairro rico Meireles com um valor alto para essa métrica, evidencia sua proximidade geográfica entre os próprios bairros ricos e o centro comercial, indicando que muitos trajetos passam por ele.

Importante salientar que como o peso foi passado como ponderação no grafo e se refere ao preço da viagem, então isso significa que um bairro ponte pode estar em uma distância maior do que outro bairro considerado ponte, haja vista a possibilidade de acontecer uma viagem com preço menor, porém com distância maior a depender do dia, horário e demanda do serviço de transporte por aplicativo.

Após as análises sobre os dados da base de dados de Fortaleza, bem como o estudo socioeconômico para essa cidade formalizado na seção 2.2.1, observou-se que há uma a alta concentração de renda e a consequente segregação geográfica da população: as pessoas de maior poder aquisitivo ocupando a zona Leste da cidade (que concentra a porção mais nobre dos bairros e o centro comercial) e as de menor poder aquisitivo ocupando as zonas Sul e Oeste.

Além disso, observou-se uma superlotação diária do sistema de transporte público por ônibus e congestionamento nas principais vias. Grande parte desse problema é explicado pela concentração da maior parte das atividades econômicas e consequentemente da maior

quantidade de empregos nas zonas Norte e Nordeste. Em contraste, observou-se a maioria da população residindo na zona Oeste e Sul da cidade, por isso os desejos de deslocamento a grandes distâncias.

Como levantado na Revisão da literatura, Fortaleza apresenta uma alta Densidade Demográfica, característica essa bastante comum em capitais brasileiras. Por outro lado, Boston apresenta uma Densidade menor em relação a outras cidades Norte Americanas, principalmente se comparada a grandes cidades brasileiras. Em uma visão mais holística, muitas cidades Norte Americana, região de país desenvolvido, apresentam uma Densidade Demográfica menor que a grande maioria das grandes cidades Latino Americana. Países como o Brasil tendem a serem enxadas, fazendo com que as populações mais pobres financeiramente morem em periferias e ocasionem viagens a grandes distâncias, pois o centro financeiro das cidades costuma estar distante dessas periferias. Em outras palavras, uma alta Densidade Demográfica pode influenciar no aumento de preços de viagens em transporte por aplicativo Uber.

Considerando essa conjuntura, foi levantado dados de 2010 referente ao Rendimento Médio Mensal das pessoas de 10 anos ou mais de idade por bairros para a cidade de Fortaleza. Os dados desse rendimento consideram todos os bairros da cidade e são demonstrados a seguir:

Valores correspondentes a 2010:

- Média da Renda Pessoal de todos os bairros ricos: R$ 1.603,14.
- Média da Renda Pessoal de todos os bairros médios: R$ 633,84.
- Média da Renda Pessoal de todos os bairros pobres: R$ 411,44.

Dessa forma, seria possível estipular alternativas para os preços das viagens desse seguimento de transporte, considerando mudanças nos trajetos das viagens simuladas. Entretanto, como esses dados socioeconômicos são de 2010, é importante fazer um ajuste em seus valores no intuito de condizer mais com a realidade dos preços praticados atualmente no mercado. Nesse sentido, considerando que as simulações das viagens da UberX em Fortaleza foram realizadas no ano de 2021, então foi realizada uma conversão desses valores de renda.

A conversão dos valores do Rendimento Médio Mensal foi realizada utilizando o índice de inflação geral IPCA para o ano de 2021, considerando as ponderações vistas na Revisão da literatura. Após a conversão, os valores ficaram como a seguir:

Conversão dos valores para 2021:

- Média da Renda Pessoal de todos os bairros ricos: R$ 3089,31.
- Média da Renda Pessoal de todos os bairros médios: R$ 1221,43.
- Média da Renda Pessoal de todos os bairros pobres: R$ 792,86.

Uma pesquisa realizada em 2019 pela Associação Nacional das Empresas de Transportes Urbanos (NTU), que reúne mais de 500 companhias de ônibus urbanos e metropolitanos em todo o Brasil, mostra que os transportes por aplicativos estão atraindo mais as pessoas que usavam habitualmente somente o transporte público do que as pessoas que faziam uso frequente de carros próprios (DIÁRIO DO TRANSPORTE, 2020). De acordo com a pesquisa, mais de 60% dos usuários dos aplicativos vieram do transporte público. O levantamento focou em dez capitais brasileiras, incluindo Fortaleza. A pesquisa se deu por meio de entrevistas eletrônicas por questionários em redes sociais. 1.410 questionários foram analisados no período de 16 de outubro a 22 de novembro de 2019. Foi observado que a maioria dos entrevistados (52%) utiliza o transporte por aplicativo de 2 a 4 vezes por semana, uma média de 3 viagens por semana. A pesquisa evidenciou que o trabalho é o principal motivo dos passageiros que utilizam esse tipo de serviço todo dia. Os usuários esporádicos ou semanais utilizam o serviço, principalmente, para lazer.

Considerando que a maioria das pessoas da cidade de Fortaleza necessitam se deslocar para o centro comercial, tendo como origem, em sua maioria, os bairros periféricos, e considerando que essas viagens tenham como principais causas a educação e o trabalho, então, para essa pesquisa, foi adotada uma quantidade média de viagens por semana de 3 vezes para pessoas que moram em algum bairro de baixa renda (periférico) e para pessoas que moram em algum bairro de alta renda (próximo do centro comercial). Considerando que a maioria dos meses do ano possui 4 semanas, a quantidade de viagens mensal seria de 12 vezes, em média.

Nesse sentido, considerando o transporte por aplicativo Uber, a pesquisa evidenciou que a média de preço das viagens dos bairros de baixa renda (pobres) entre si (R$ 13,54) é menor se comparada com a média de preço das viagens de bairros pobres para bairros de maior renda (mais ricos), ou seja, trajeto pobre-médio: R$ 16,57 e trajeto pobre-rico: R$ 28,88.

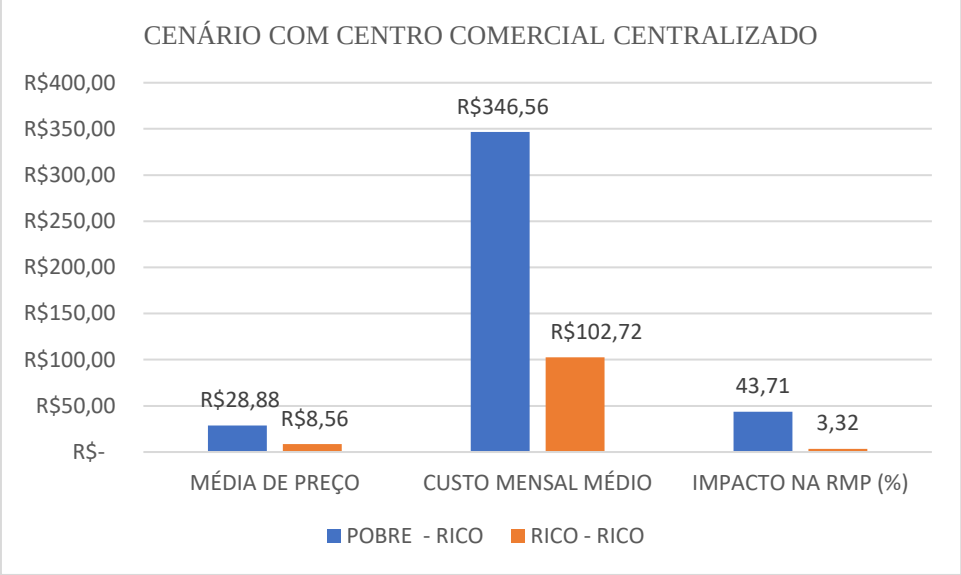
Uma proposta para tentar uma redução desses custos das viagens de Uber para os usuários dos bairros mais pobres seria a de uma maior descentralização do comércio da cidade, englobando mais os bairros médios e pobres, uma vez que a cidade possui uma grande concentração comercial distribuída na região Norte e Nordeste. Nessa perspectiva, foi evidenciado que o centro comercial está em torno de bairros ricos. Os usuários dos bairros pobres gastam em média R$ 28,88 (somente ida, pois o usuário pode retornar ao seu bairro por outro meio de transporte) para chegar ao centro comercial. Se a quantidade de viagens para o centro comercial de um usuário de bairro pobre for em média de 12 vezes em um mês, então daria um custo mensal de R$ 346,56. Isso corresponde a cerca de 43,71% da renda média pessoal de uma pessoa que reside em um bairro de baixa renda.

Por outro lado, os usuários que moram nos bairros ricos gastam em média R$ 8,56 por viagem ao centro comercial, região que concentra maior parte de seu território dentro desses mesmos bairros (viagens de rico para rico). Considerando também uma quantidade média de viagens de 12 vezes em um mês, então daria um custo mensal de R$ 102,72. Isso corresponde a cerca de 3,32% da renda média pessoal de uma pessoa que reside em um bairro de alta renda. Isso significa que os usuários dos bairros pobres gastam com viagens da Uber em Fortaleza um valor de cerca de 3 vezes mais do que os usuários que moram nos bairros ricos, quando o destino é o centro comercial da cidade.

De maneira complementar, os usuários de bairros pobres gastam em média R$ 16,57 em viagem para bairros médios. Isso significa uma diferença de R$ 12,44 se comparada com a média de preços das viagens de pobre para rico. Isso daria por mês, considerando a mesma quantidade de viagens de 12, um valor de R$ 149,28, correspondendo a 18,82% da renda média pessoal de uma pessoa de baixa renda. Uma redução de custo de 43,07% em comparação com o trajeto pobre-rico. Já os usuários de bairros ricos gastam em média R$ 17,28 em viagem para bairros de classe média. Isso significa um aumento de R$ 8,72 se comparada com a média de preços das viagens de rico para rico (R$ 8,56). Isso daria por mês, considerando a mesma quantidade de viagens de 12, um valor de R$ 207,36. Um aumento de custo de 101,8% em comparação com o trajeto rico-rico. Entretanto, esse valor corresponderia a 6,71% da renda das pessoas que residem nesses bairros.

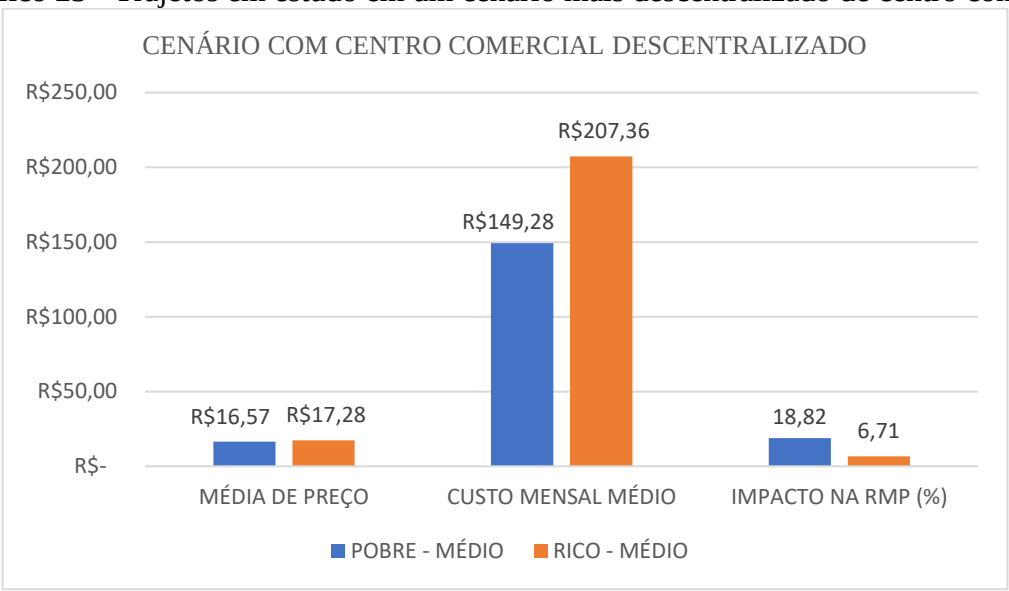
Os usuários dos bairros pobres, viajando para os bairros de classe média, teriam um gasto que representaria 18,82% de suas rendas. Ou seja, embora haja uma redução de custos para os residentes de bairros pobres, o gasto para essa parcela da população, nesse contexto, ainda seria de cerca de 3 vezes maior do que as pessoas que residem nos bairros ricos, mas teriam um impacto muito menor se comparado com o trajeto pobre-rico que representava quase metade da renda média pessoal (43,07%). O Gráfico 22 e o Gráfico 23 exibem, respectivamente, os valores das médias dos preços, custo mensal médio e o valor em porcentagem do impacto na Renda Média Pessoal (RMP) de trajetos de viagens de bairros pobres para ricos (POBRE – RICO) e de trajetos de bairros ricos para ricos (RICO – RICO) para o cenário atual em que há concentração do centro comercial na cidade de Fortaleza e para um possível cenário em que o centro financeiro fosse mais abrangente, englobando pelo menos os bairros de classe média.

Gráfico 22 - Trajetos em estudo em um cenário centralizado de centro comercial.



Fonte: Autor (2022).

Gráfico 23 - Trajetos em estudo em um cenário mais descentralizado de centro comercial.



Fonte: Autor (2022).

Se o centro comercial for descentralizado de uma forma que fique mais próximo pelo menos dos bairros médios, então poderia haver uma redução de custos de viagens para os usuários de bairros pobres de pelo menos de 43,07%. Isso seria uma economia de cerca de R$ 149,28 por mês ou R$ 1.791,36 por ano. Já para os usuários que moram nos bairros ricos, a descentralização do comércio para os bairros médios elevaria os custos das viagens a pouco mais de 100%. Para os bairros pobres, essa economia mensal representaria 18,82% do valor da renda média pessoal mensal. Para os bairros ricos, esse aumento representaria 6,71% do valor

da renda média pessoal mensal. Isso significa que o impacto do aumento de preços para os usuários de bairros ricos continuaria sendo menor se comparado com o impacto que há para os bairros pobres, quando partem de seus locais para um centro comercial localizado mais nos bairros de classe média. Entretanto, para os bairros mais pobres, haveria uma maior economia se comparado com a atual localização do centro comercial e a renda dos mais ricos seria afetada em menos de 7%.

Considerando os indicativos encontrados, uma possibilidade de tomada de decisão seria a de o governo federal brasileiro elaborar políticas púbicas no intuito de promover o comércio em regiões mais pobres financeiramente. Como evidenciado para a cidade de Boston, algumas regiões de menor renda possuem uma capacidade de atrair pessoas pelo fato de possuírem tipos de atividades comerciais, como bares, estádios esportivos, parques e restaurantes. Embora essas regiões de Boston não possuam o poder financeiro do centro comercial da cidade, elas ainda conseguem mitigar um pouco a alta de preços dos serviços de transportes no contexto da Uber. Nesse sentido, para a cidade de Fortaleza, poderia ser proposto incentivos fiscais aos empresários para que empreendessem em regiões de média e baixa renda, promovendo o incentivo econômico dessas regiões e possibilitando uma redução de custos de viagens em transportes por aplicativo.

Não obstante, o governo estadual poderia trabalhar em conjunto com as políticas promovidas pelo governo federal no intuito de aumentar a eficiência prática dessas medidas de incentivos. Dessa forma, os impactos financeiros poderiam ter mais chances de se concretizarem, reduzindo os gastos das rendas média pessoal mensal dos residentes de baixa renda e possibilitar uma redução, no quesito econômico, no âmbito das desigualdades sociais.

CONCLUSÃO

Neste trabalho foram feitas análises a partir de um processo de Ciência de Dados que considerou dados Socioeconômicos e preços de viagens do transporte por aplicativo Uber. Mais precisamente, estudou-se o comportamento dos preços, quando comparados com trajetos distintos de viagens dessa empresa. Dessa forma, foi possível observar o impacto desses preços na Renda Média Pessoal dos usuários residentes em regiões de baixa e alta renda.

Para entender melhor o comportamento dos preços nesse seguimento de transporte, foram estudados trabalhos que evidenciaram relações entre preços de viagens da Uber e algumas características dos locais de embarque dos usuários, bem como estudos que levantaram informações sobre concentração financeira em porções territoriais específicas. Diante disso, também foi possível observar que a distância é um atributo que influencia no processo de precificação do serviço de viagens da Uber e que nem sempre uma distância maior indica um aumento proporcional no preço, evidenciando que o tipo de localidade também possui influência nesse processo.

Na análise feita no capítulo Resultados e discussões, observou-se os resultados obtidos da Análise Exploratória de Dados. A partir desses resultados, foi observado que os modelos de predição de distância para a cidade de Boston seriam uma alternativa ao processo de precificação existente atualmente para a Uber. Assim, os usuários teriam maior liberdade na escolha da distância que melhor lhes convém mediante ao preço que estão dispostos a pagar. Importante salientar que essa predição da distância é apenas uma análise que pode servir para recomendações de políticas públicas, de modo que mais estudos podem ser utilizados no intuito de validar os benefícios dessa predição para usuários do serviço de viagens da Uber.

Como o enfoque da pesquisa foi para a cidade de Fortaleza, os resultados das análises estatísticas e das medidas de centralidade foram importantes para validar a base de dados simulada para essa cidade no sentido de comprovar por meio de testes e medidas que as amostras selecionadas eram pertinentes para as análises. Nesse sentido, foi possível realizar comparações entre alguns trajetos de viagens nessa cidade, no intuito de registrar as diferenças de médias de preços dos bairros em estudo, para que fosse possível levantar uma hipótese sobre diminuição de custos a partir de uma maior descentralização do centro comercial. Soma-se a isso o comportamento de preços para ambas as cidades em estudo, pois demonstraram

convergência de alta, quando o destino da maioria das viagens se destinava para o centro comercial.

A partir dessas observações foi proposto uma tomada de decisão que consiste em uma política de maior incentivo fiscal para empreendedores brasileiros, no intuito de fomentar o comércio em regiões de baixa e média renda, promovendo o incentivo econômico dessas regiões, possibilitando uma redução de custos de viagens em transportes por aplicativo. Dessa forma, uma não concentração financeira, englobando regiões mais pobres financeiramente, poderia impactar na redução de preços de viagens de transporte por aplicativo Uber, reduzindo os gastos da Renda Média Pessoal de usuários mais pobres financeiramente sem, contudo, impactar de maneira elevada a Renda Média Pessoal dos usuários mais ricos.

Considerando as evidências encontradas nesse trabalho, é possível adotar a metodologia aqui aplicada em outras cidades. Para isso, a etapa de Obtenção de Dados de Preços da Uber precisa ser concebida considerando as cidades em estudo, uma vez que a lógica de precificação depende do contexto individual de cada cidade. Caso seja possível encontrar uma base de dados real de preços para a cidade em estudo, então essa etapa pode ser resumida ao tratamento dos valores dos preços existentes. A etapa de Obtenção de Dados Socioeconômicos precisa ser determinada de acordo com a dimensão que a pesquisa deseja seguir. Há várias dimensões socioeconômicas que podem ser estudadas, como Educação, Idade, Emprego, entre outras. A etapa de Limpeza e Tratamento dos Dados também precisa ser adaptada a cidade em estudo, pois se os preços forem simulados, haverá a necessidade de realizar uma análise Estatística das simulações. Independentemente de os preços serem ou não simulados, é importante realizar uma análise no âmbito das Medidas de Centralidade para grafos, pois a partir das informações obtidas dessa análise, será possível estabelecer rotas de viagens que impactem na média de preços e, com isso, obter evidências de mudanças de valores.

O desenvolvimento dessa pesquisa também pode contribuir para orientações no desenvolvimento de políticas públicas no âmbito das desigualdades sociais, principalmente no quesito econômico. Os resultados obtidos revelam que há indicativo de redução de custos para uma população mais pobre economicamente, caso os centros comerciais sejam mais distribuídos em uma região. Entretanto, os resultados dessas análises precisam de mais estudos que possam detalhar esse comportamento, uma vez que essas desigualdades possuem características multifacetadas.

3.8 Limitações da Pesquisa

Um dos principais desafios em pesquisas ou experimentos que envolvem dados socioeconômicos é a subjetividade inerente desse âmbito. Muitos problemas nesse sentido geralmente envolvem vários aspectos da sociedade, limitando a eficiência de uma análise que considera apenas algumas perspectivas. Nesse sentido, essa pesquisa limitou-se na análise socioeconômica do IDH Renda de bairros. Além disso, existe a questão temporal desses dados que por serem provenientes do Censo Demográfico, possuem um período de atualização de cerca de 10 anos. Isso pode impactar a realidade atual das condições socioeconômicas de uma região em estudo. Entretanto, para tentar minimizar isso, foi realizada uma conversão dos valores monetários envolvidos para a cidade de Fortaleza, no sentido de homogeneizar os valores das rendas com os preços simulados das viagens.

Soma-se a isso, a questão das bases de dados, que por questões legislativas, as empresas não mais podem fornecê-las para fins de pesquisa. Entretanto, para tentar criar a base de dados de Fortaleza de forma mais realista, foram realizadas pesquisas no intuito de tentar tornar a base o mais condizente possível com a realidade, considerando aspectos como tráfego por dia, horários de pico e a quantidade média de viagens que os usuários utilizam o transporte Uber.

3.9 Trabalhos Futuros

Tendo em vista que as análises foram realizadas em contextos específicos, é interessante investigar outros cenários no intuito de validar os resultados já obtidos. Pode-se também explorar outros serviços de transportes, como os serviços de Moto Taxistas e o Transporte Coletivo Público, para explorar outras caracterizações de mudanças de preços e avaliar a representatividade das mesmas.

Uma análise que pode ser feita sobre os dados já obtidos dessa pesquisa é sobre a investigação de outra cidade do Nordeste brasileiro, bem como de outra cidade Norte Americana. Por serem regiões que possuem grandes diferenças econômicas, podem ser estudados mais afundo no intuito de encontrar mais relações que possam vir a colaborar na diminuição das desigualdades sociais. Os dados levantados no capítulo da Revisão da literatura

poderiam aumentar as probabilidades de encontrar padrões que representassem diferentes contextos, para mais de um domínio de análise.

Também existe a possibilidade de aprimorar a metodologia adotada nessa pesquisa, inserindo mais validações estatísticas e outras medidas de centralidade ou até mesmo outros parâmetros de validação. Uma forma de alcançar esses objetivos poderia ser o acréscimo de um componente espacial para os modelos de Aprendizado de Máquina lineares utilizados nesta pesquisa. Dessa forma, seria possível analisar o comportamento obtido em relação aos valores dos Coeficientes de Determinação obtidos aqui.

Além disso, pode-se tentar aplicar outras abordagens de Ciência de Dados de outros trabalhos sobre os dados aqui encontrados, seja para maior detalhamento da AED ou para detectar outros padrões. Com isso, seria possível comparar alguns resultados obtidos por outros estudos com os já obtidos aqui.

REFERÊNCIAS

ABRAHAM, Ajith; GUO, He; LIU, Hongbo. **Swarm Intelligence**: Foundations, Perspectives and Applications. v. 28, 2006. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-33869-7_1#citeas. Acesso em: 25 jun. 2022.

ALVARENGA JUNIOR, Wagner José. **Métodos De Otimização Hiperparamétrica: um Estudo Comparativo Utilizando Árvores De Decisão E Florestas Aleatórias Na Classificação Binária**. 2018. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia, Universidade de Minas Gerais, Belo Horizonte, 2018.

AMARAL, Fernando. **Introdução à Ciência de Dados**: mineração de dados e big data. 1. ed. Rio de Janeiro: Alta Books, 2016.

ANSELIN, Luc. Spatial econometrics: methods and models. **In**: Studies in Operational Regional Science, v. 4, 2013. Disponível em: https://link.springer.com/book/10.1007/978-94-015-7799-1#about-book-content. Acesso em: 25 jun. 2022.

AVRIM, Blum,; HOPCROFT, John; KANNAN, Ravi. **Foundations of Data Science**. Cambridge University Press: 2018. 432 p.

BANCO CENTRAL DO BRASIL: **Calculadora do Cidadão**.
Disponível em: https://www.bcb.gov.br/acessoinformacao/calculadoradocidadao. Acesso em: 25 jun. 2022.

BARRADAS FILHO, A. O *et al.* Application of artificial neural networks to predict viscosity, iodine value and induction period of biodiesel focused on the study of oxidative stability. **Fuel**, v. 145, p. 127-135, abr., 2015.

BEZERRA, A. C. T. **Uber: a gestão do relacionamento em novos modelos de negócio.** 2017. Trabalho de Conclusão de Curso (Bacharelado em Relações Públicas) – Universidade Federal da Paraíba, João Pessoa, 2017.

BEZERRA, Aguinaldo *et al*. A preliminary exploration of uber data as an indicator of urban liveability. **In**: International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA). IEEE, 2019. p. 1-8.

BOSTON PLANNING & DEVELOPMENT AGENCY RESEARCH DIVISION – BP&DARD. **Neighborhood Profiles**. 2017. Disponível em: bostonplans.org/research-maps. Acesso em: 28 jun. 2022.

BORBA, Elizandro. **Medidas de Centralidade em Grafos e Aplicações em redes de dados.** 2013. Dissertação (Mestrado em Matemática aplicada) – Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

BPS Opportunity Index. **Boston Public Schools**. Disponível em: https://www.bostonpublicschools.org/domain/2301. Acesso em: 25 jun. 2022.

BUGNION, Pascal; MANIVANNAN, Arun; NICOLAS, Patrick. **Scala: Guide for Data Science Professionals**. Birmingham: Packt Publishing, 2017.

CARVALHO, André *et al.* **Inteligência Artificial**: Uma abordagem de Aprendizado de Máquina. 2. ed. Rio de Janeiro: Ltc, 2011.

COHEN, B.; KIETZMANN, J. Ride On! Mobility Business Models for the Sharing Economy. **Organization & Environment**, v. 27, n. 3, 2014, p. 279-296. Disponível em: http://journals.sagepub.com/doi/abs/10.1177/1086026614546199. Acesso em: 25 jun. 2022.

CONTADOR, J.; SENNE, E. Testes não paramétricos para pequenas amostras de variáveis não categorizadas: um estudo. **Gest. Prod**., São Carlos, v. 23, n. 3, p. 588-599, 2016. Disponível em: http://old.scielo.br/pdf/gp/v23n3/0104-530X-gp-0104-530X357-15.pdf. Acesso em: 25 jun. 2022

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada:** para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007.

DE SOUZA, Allan Mariano; VILLAS, Leandro Aparecido. Vem Tranquilo: Rotas Eficientes baseado na Dinâmica Urbana Futura com Deep Learning e Computação de Borda. *In*: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS (SBRC), 38. , 2020, Rio de Janeiro. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 351-364. ISSN 2177-9384.

SIMES, Ric. Economic effects of ridesharing in Australia: A report for uber. **Delloite**. 2016. Disponível em: https://www2.deloitte.com/au/en/pages/economics/articles/economic-effects-ridesharing-australia-uber.html. Acesso em: 28 jun. 2022.

DIAB, Shadi. Optimizing Stochastic Gradient Descent in Text Classification Based on Fine-Tuning Hyper-Parameters Approach. A Case Study on Automatic Classification of Global Terrorist Attacks. **International Journal of Computer Science and Information Security (IJCSIS),** v. 16, n. 12, dez., 2018.

DIMITRIADIS, S. I.; LIPARAS, D. How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database. **Neural Regeneration Research**, v. 13, n. 6, 2018, p. 962-970.

DIÁRIO DO TRANSPORTE. Disponível em: https://diariodotransporte.com.br/2020/01/30/mais-de-60-dos-usuarios-dos-aplicativos-vieram-do-transporte-publico-e-preco-esta-entre-os-principais-motivos-da-troca/. Acesso em: 07 fev. 2022.

DONGES, Niklas. Random Forest Algorithm: A Complete Guide. **Builtin**. jul., 2021. Disponível em: https://builtin.com/data-science/random-forest-algorithm. Acesso em: 28 jun. 2022.

DUSI, L. A. **O uso de aplicativos para smartphone no transporte individual: 99Taxis e Uber.** 2016. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Civil) - Universidade de Brasília. Brasília, 2016.

FACELI, Katti *et al*. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC. 2011.

FARIAS, M. V. F. **Avaliação da percepção de qualidade da prestação do serviço de transporte individual de passageiros do Distrito Federal**: Taxi e Uber. Dissertação (Mestrado em Transportes). Universidadede Brasília. Brasília, 2016.

G1 CEARÁ. **Trânsito de Fortaleza fica 68% mais lento em horários de pico, diz pesquisa**. 2018. Disponível em: https://g1.globo.com/ce/ceara/noticia/transito-de-fortaleza-fica-68-mais-lento-em-horarios-de-pico-diz-pesquisa.ghtml Acesso em: 25 jun. 2022.

GAMA, João. **Árvores de decisão**. Disponível em: http://www.liacc.up.pt/~jgama/Mestrado/ECD1/Arvores.html. Acesso em: 21 nov. 2021.

GAMA, João *et al*. **Extração de Conhecimento de Dados/Data Mining**. 3. ed. Lisboa: Edições Sílabo, 2017. 436 p.

GARCIA, S. C. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde.** Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática. Universidade Federal do Rio Grande do Sul. Porto Alegre, 2003.

GONÇALVES, Pollyana. Afinal, como se desenvolve um projeto de Data Science? **TechBlogHotmart, set., 2018.** Disponível em: https://medium.com/techbloghotmart/afinal-como-se-desenvolve-um-projeto-de-data-science-233472996c34. Acesso em: 27 jun. 2022.

GONÇALVES, Paulo. **Uma abordagem da distribuição normal através da resolução de uma situação problema com a utilização do software Geogebra.** Dissertação (Mestrado em Matemática). Universidade Federal de Goiás. Jataí, 2014.

GOVERNO DO ESTADO DO CEARÁ. Disponível em: https://www.ceara.gov.br/ Acesso em: 25 jun. 2022.

HAUGHTON, D. *et al.* (2003). Effect of dirty data on analysis results. **Eighth International Conference on Information Quality**. p. 64 - 79. 2003.

HOTZ, Nick. What is a Data Science Life Cycle? **Data Science Process Alliance**, mai. 2022. Disponível em: https://www.datascience-pm.com/data-science-life-cycle/ Acesso em: 25 jun. 2022.

InDriver. Disponível em: https://indriver.com/pt/city Acesso em: 08/05/2022.

INSTITUTO DE PLANEJAMENTO DE FORTALEZA - IPLANFOR. **Plano de Mobilidade de Fortaleza** (PlanMob). 2015

INSTITUTO DE PLANEJAMENTO DE FORTALEZA - IPLANFOR. **Mobilidade**. Disponível em: https://catalogodeservicos.fortaleza.ce.gov.br/categoria/mobilidade. Acesso em: 27 jun. 2022.

KONAR, Amit. **Artificial Intelligence and Soft Computing**: Behavioral and Cognitive Modeling of the Human Brain. Crc Press. 1999.

LOCA, Antonio Luiz da Silva. **Uma metodologia experimental para avaliar abordagens de aprendizado de máquina para diagnóstico de falhas com base em sinais de vibração.** Dissertação (Mestrado em Informática) –Centro Tecnológico, Universidade Federal do Espírito Santo. Vitória, 2020.

LOPES, Manuela *et al.* Utilização dos testes estatísticos de Kolmogorov-Smirnov e Shapiro-Wilk para verificação da normalidade para materiais de pavimentação. **Transportes**, v. 21, n. 1, abr., 2013. Disponível em: https://www.revistatransportes.org.br/anpet/article/view/566. Acesso em: 25 jun. 2022.

MANCUSO, Aline. **Métodos Bayesianos em Metanálise**. 2010. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Federal do Rio Grande do Sul, 2010. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/29108/000775678.pdf?sequence=1. Acesso em: 26 jun. 2022.

MARTINEZ, I; VILES, E; OLAIZOLA, I. Data Science Methodologies: Current Challenges and Future Approaches. **Big Data Research**, v. 24, mai., 2021. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S2214579620300514. Acesso em: 24 jun. 2022.

MELNIK, M. Demographic and Socio-economic Trends in Boston: What we've learned from the latest Census data**. Boston Redevelopment Authority,** v. 29, nov., 2011.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre Aprendizado de Máquina. **In**: MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Sistemas Inteligentes Fundamentos e Aplicações. Barueri, SP: Manole Ltda, 2003. p. 39-56.

MORENO, L; MORCILLO, A. Comparação de dois grupos pareados Teste de Wilcoxon. **Statistics Biostatistics**, São Paulo, jan., 2020.

MUNAKATA, Toshinori**. Fundamentals of the New Artificial Intelligence**: Neural, Evolutionary, Fuzzy and More. 2. ed. Heidelberg: Springer. 2008.

NATRELLA, M. (2010). NIST/SEMATECH e-Handbook of Statistical Methods. **NIST**. Disponível em: https://www.itl.nist.gov/div898/handbook/. Acesso em: 28 jun. 2022.

OBEDAIT, A.A.; YOUSSEF, M.; LJEPAVA, N. Citizen-centric approach in delivery of smart govern- ment services. **Smart Technologies and Innovation for a Sustainable Future**, jan., 2019, 73–80.

PNAD. PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS. Disponível em: https://cidades.ibge.gov.br/brasil/ce/fortaleza/panorama. Acesso em: 13 nov. 2021.

PREFEITURA MUNICIPAL DE FORTALEZA. **Desenvolvimento Humano, por bairro, em Fortaleza.** Disponível em: https://www.fortaleza.ce.gov.br/noticias/prefeitura-apresenta-estudo-sobre-desenvolvimento-humano-por-bairro. Acesso em: 25 jun. 2022.

PROVOST, Foster; FAWCETT, Tom. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, n. 1, v. 1, fev., 2013.

QUICK, Bruno. Ideias e Negocio: transporte por aplicativo. **Sebrae**. 2020. Disponível em: https://bibliotecas.sebrae.com.br/chronus/ARQUIVOS_CHRONUS/IDEIAS_DE_NEGOCIO/PDFS/510.pdf. Acesso em: 25 jun. 2022.

ROZA, F. S. da. **Aprendizagem de máquina para apoio à tomada de decisão em vendas do varejo utilizando registros de vendas.** 2016. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Controle e Automação) – Universidade Federal de Santa Catarina. Florianópolis, 2016. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/171569/PFC_2016-1%20Felippe_Roza.pdf?sequence=1. Acesso em: 25 jun. 2022.

SANTOS, Virgílio. Análise estatística: qual é a importância da técnica nos negócios? **FM2S Educação e Consultoria.** 2022. Disponível em: https://www.fm2s.com.br/analise-estatistica/ Acesso em: 25 jun. 2022.

SARKAR, S.K.; MIDI, H. Importance of assessing the model adequacy of binary logistic regression. **Journal Applied Science**, v. 10, n. 6, 2010.

SCHWIETERMAN, J. P. Uber Economics: Evaluating the Monetary and Travel Time Trade-Offs of Transportation Network Companies and Transit Service in Chicago, Illinois. **Transportation Research Record**, v. 2673, n. 4, p 295-304, abr., 2019.

SILVA, J., LIMA, L.; BEZER, I. Uma metodologia orientada a dados sociodemográficos para predição de preços do Uber X. **In**: Congresso Brasileiro de Automática-CBA, v. 2. n. 1., dez., 2020.

SILVA, Pedro; BOGONI, Juliano. **Introdução à estatística básica.** Florianópolis: 2015. Disponível em: http://www.liaaq.ccb.ufsc.br/files/2013/10/Aula-4.pdf Acesso em: 25 jun. 2022.

SILVEIRA, Daniel. **Desigualdade de renda cresce no Nordeste e diminui nas demais regiões, aponta IBGE**. **G1**. Rio de Janeiro, 2020.

SOUSA, S. **Estudo das propriedades e robustez da rede de transporte público de São Paulo**. Dissertação (Mestrado em Ciências) – Universidade de São Paulo. São Paulo, 2014.

UBER. Disponível em: https://www.uber.com/global/pt-br/price-estimate/ Acesso em: 13/10/2021.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. **Bioestatística quantitativa aplicada.** Faculdade de Medicina. Programa de Pós Graduação em Ciências da Saúde: Ginecologia e Obstetrícia; organizadores: Edilson Capp e Otto Henrique Nienov. Porto alegre: UFRGS, 2020. 260 p.

VASCONCELOS, José Braga; BARÃO, Alexandre. **Ciência dos dados nas organizações**. FCA: 2017. 336 p.

WAINER, Jacques. Análise estatística 1. **IC Unicamp**. Disponível em: https://www.ic.unicamp.br/~wainer/cursos/2s2021/430/aula2.html. Acesso em: 25 jun. 2022.

WANG, Mingshu; MU, Lan. Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. **Computers, Environment and Urban Systems**, v. 67, p. 169-175, 2018.

WHITE, C. *et al.* Statistical Prediction of Sealant Modulus Change due to Outdoor Weathering. Natural and Artificial Aging of Polymers. **NLST**. v. 6, n. 24, p. 65-72, set., 2013.

APÊNDICE

Tabela 7 - Simulação de preços pelo simulador da Uber.

	Destino														
Origem	Meireles	Aldeota	Dionísio Torres	Mucuripe	Guararapes	Conjunto Palmeiras	Parque Presidente Vargas	Canindezinho	Genibaú	Siqueira	Autran Nunes	Dendê	Parque Dois Irmãos	Cajazeiras	Messejana
Meireles	7,91	6,42	6,85	6,88	11,17	23,56	31,78	32,46	29,7	34,26	17,28	16,63	16,07	18,36	19,34
Aldeota	7,31	6,67	11,13	8,26	15,53	27,49	32,88	29,3	24,73	31,29	16,11	15,83	15,81	17,55	17,86
Dionísio Torres	7,36	6,42	6,44	11,97	12,64	19,66	28	28,02	25,75	29,8	16,03	15,12	13,51	17,12	16,71
Mucuripe	7,42	9,29	10,23	6,42	11,05	24,98	25,65	42,72	31,08	34,95	18,66	19,47	17,34	18,78	21,14
Guararapes	14,57	12,12	11,01	10,92	6,42	20,88	29,84	30,28	28,01	32,07	16,55	13,87	14,76	15,23	15,78
Conjunto Palmeiras	18,42	16,63	15,98	19,31	17,56	6,42	16,3	19,35	20,35	19,84	17,53	16,37	14,24	16,87	15,95
Parque Presidente Vargas	24,58	17,83	22,29	25,62	23,29	14,03	6,42	6,42	15,31	11,52	15,27	14,85	14,34	15,65	16,36
Canindezinho	24,28	22,78	24,1	25,22	35,67	14,58	6,42	4,55	12,79	9,55	14,97	15,07	14,2	16,17	15,85
Genibaú	26,66	23,66	26,7	37,9	26,02	24,88	17,21	23,51	6,42	12,95	9,87	13,64	15,49	16,34	18,69
Siqueira	29,13	25,69	25,48	30,36	28,09	20,83	12,64	10,37	17,87	6,42	14,65	16,44	18,16	18,91	20,45
Autran Nunes	13,75	12,56	11,95	14,78	14,12	19,75	16,48	15,73	10,86	18,43	7,87	12,39	15,74	16,97	18,26
Dendê	12,84	11,66	10,83	15,48	10,39	17,51	16,94	16,13	17,38	16,5	13,47	8,56	11,41	13,69	16,41
Parque Dois Irmãos	13,41	10,57	9,96	13,72	11,48	16,87	16,11	15,88	18,34	17,61	18,81	15,14	6,82	10,47	12,73
Cajazeiras	15,61	14,45	15,13	14,36	12,54	10,15	17,74	18,08	19,15	18,86	21,74	19,83	14,69	9,67	12,41
Messejana	16,67	15,84	12,96	17,41	13,29	11,07	18,87	19,12	21,23	20,47	19,81	12,36	11,39	10,88	8,72
LEGENDA DAS CORES:															
BAIRROS RICOS															
BAIRROS POBRES															
BAIRROS MÉDIOS															

Fonte: Autor (2021).