XV CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

FERRAMENTA DE ANÁLISE DE DADOS PARA O TRANSPORTE PÚBLICO

José Ivan Silva da Cruz Júnior[1], Nazareno Ferreira de Andrade [2]

RESUMO

Com o crescimento cada vez maior dos centros urbanos e da consequente demanda pelo transporte público, cresceu o interesse por analisar os dados gerados na busca de ter uma melhor compreensão da dinâmica do sistema de transporte e dos padrões de mobilidade das pessoas. Esse conhecimento pode contribuir para a melhora do serviço oferecido e possibilitar a cidadãos e a gestores compreender como habitamos nossas cidades e como elas nos afetam. Devido a essa diversidade de possibilidades de aplicação, foram criados ferramentas de software para auxiliar pesquisadores e gestores no trabalho de obter informações relevantes, porém nem todas estão disponíveis para uso em outras pesquisas. Neste artigo, propomos uma ferramenta de software, implementada na linguagem R, que oferece um conjunto de funções úteis para análise de dados de transporte público, incluindo: a análise espacial das origens e destinos das viagens; a análise da lotação dos ônibus por rota; e a detecção de outliers na velocidade das viagens realizadas. A ferramenta de software proposta foi desenvolvida considerando a compatibilidade com vários conjuntos de dados e a usabilidade.

Palavras-chave: Transporte Público, Análise de Dados, Ferramenta de Software

DATA ANALYSIS TOOL FOR PUBLIC TRANSPORT

ABSTRACT

With the increasing growth of urban centers and the consequent demand for public transport, interest in analyzing the data generated in the search for a better understanding of the dynamics of the transport system and the mobility patterns of people has grown. This knowledge can contribute to the improvement of the service offered and enable citizens and managers to understand how we inhabit our cities and how they affect us. Due to this diversity of application possibilities, software tools were created to assist researchers and managers in the work of obtaining relevant information, but not all of them are available for use in other research. In this article, we propose a software tool, implemented in the R language, which offers a set of useful functions for analyzing public transport data, including: the spatial analysis of travel origins and destinations; the analysis of bus capacity by route; and the detection of outliers in the speed of the trips made. The proposed software tool was developed considering compatibility with various data sets and usability.

Keywords: Public Transport, Data Analysis, Software Tool

1. INTRODUÇÃO

Nos dias atuais, aproximadamente metade da população mundial vive em regiões urbanas. Cerca de oito por cento vive em cidades com mais de dez milhões de habitantes [1]. Estima-se que o número de pessoas que vive em cidades atingirá cinco bilhões até 2030 [2]. Sendo assim, muitas pessoas precisam usufruir de serviços urbanos que, dada a quantidade de pessoas e o tamanho das cidades, são complexos [3]. Todos os dias, com o propósito de trabalhar, de estudar ou de divertir-se, milhões de pessoas utilizam o transporte público para se deslocar.

O relatório, divulgado pelo Moovit [4] de 2016 (aplicativo usado diariamente por 350 milhões de passageiros de todo o mundo), sobre o cenário do transporte público no mundo, mostra que aproximadamente um terço dos usuários se desloca por mais de 2 horas diárias em grandes cidades como São Paulo, Cidade do México e Londres. Além disso, quase 40% dos usuários esperam mais de 20 minutos por dia em uma estação de ônibus. O relatório também mostra, para várias cidades onde o aplicativo é usado, a distância média que as pessoas percorrem em uma única viagem pela cidade, a qual varia de 3,6 km em Campina Grande, no Brasil, para 11,2 km em Hong Kong, por exemplo.

Nas últimas décadas, a diversidade e a quantidade de dados coletados diariamente pelos sistemas de informação cresceram vertiginosamente. Estima-se que só de 2013 a 2015 foram produzidos dados equivalentes a todos os anos anteriores da história. Tecnologias como big data, computação ubíqua, Internet das Coisas e como diversas outras surgiram para facilitar a produção e a análise de grandes quantidades de dados. Particularmente, no contexto de cidades, hoje, tem-se dados abundantes sobre locais, sobre as pessoas que habitam nesses locais, sobre vias, sobre trânsito e sobre outros variados especificidades.

Outro fenômeno relevante nesse contexto é a recente popularização e a evolução de métodos de análise de grandes quantidades de dados, através da Mineração de Dados, ou Ciência de Dados. A aplicação de técnicas de mineração de dados às grandes massas de dados sobre as cidades possibilita os cidadãos e os gestores compreenderem como habitam as cidades e como essas afetam a população em uma escala sem precedentes.

Sendo assim, este trabalho tem como objetivo aplicar técnicas de ciência de

dados e visual analytics a dados do transporte público, a fim de desenvolver o MobilityHelp, um ferramental útil a desenvolvedores de software para gestão de transporte público, visando facilitar a exploração e análise deste tipo de dado por gestores e tomadores de decisão. Para validação da ferramenta proposta, foi conduzido um estudo de caso com dados de transporte público da cidade de Curitiba (Paraná, Brasil).

2. METODOLOGIA

A metodologia adotada para analisar os dados é o KDD (Knowledge-Discovery in Databases), que é um processo de extração de informações de base de dados. Esse método consiste na imersão no domínio da aplicação para compreendê-lo de uma forma mais eficiente.

O processo KDD foi adotado com as seguintes particularidades:

- Compreensão do domínio da aplicação: essa etapa foi contemplada através de estudos bibliográficos sobre o transporte público no Brasil e, em especial, da cidade de Curitiba;
- Seleção de dados: os dados encontram-se disponíveis e foram repassados pela Prefeitura Municipal de Curitiba.

Os dados históricos sobre o transporte coletivo de Curitiba foram disponibilizados pela administradora local - URBS (Urbanização de Curitiba S/A) em parceria com a Universidade Tecnológica Federal do Paraná - UTFPR. São eles: (i) descrição da organização e de funcionamento do serviço de transporte público da cidade no formato GTFS; (ii) dados de geolocalização dos ônibus (GPS); (iii) registros de bilhetagem eletrônica dos ônibus e dos terminais de Curitiba.

Os dados utilizados na pesquisa foram produzidos por *Braz*, que desenvolveu uma Matriz Origem-Destino a partir dos dados de bilhetagem da cidade de Curitiba-PR [5]. Os dados das viagens utilizados nas análises correspondem ao período de 01/05/2017 até 17/07/2017.

Durante a etapa de processamento e análise dos dados, foi utilizada a plataforma de desenvolvimento RStudio. Nela o desenvolvimento se deu principalmente utilizando a linguagem de programação R, que é voltada para análise estatística e criação de visualizações de dados. Adicionalmente,

utilizou-se a linguagem de marcação Markdown, uma linguagem simples de marcação que possibilita a transformação das análises em relatórios.

Pré-processamento dos dados

Essencialmente, um *DataFrame* é uma estrutura de dados bidimensional, composta por linhas e colunas, remetendo a uma planilha. Ele pode ser criado a partir de arquivos, páginas da web ou dados gerados por código.

A partir da criação do DataFrame, torna-se possível realizar manipulações nos dados de maneira simples, além de fornecer informações úteis para análises exploratórias a serem realizadas com esses dados. Entretanto, possibilitar que o dataframe possua dados que sejam relevantes e simples de manipular exige um processo de transformação de dados. Para isso, utilizamos o Tidyverse, que é um pacote do R, cuja função é carregar outros pacotes do R como dplyr e o tidyr, para transformarmos a antiga base de dados na base de dados que desejamos.

Logo, o DataFrame que inicialmente coletamos foi submetido a uma etapa de pré-processamento para que resultasse em informações objetivas e passíveis de melhor manipulação. Novas variáveis foram adicionadas, derivadas a partir de combinações das variáveis originais. Por exemplo, originalmente, havia a informação de latitude e logitude do embarque e desembarque da viagem. Após o processamento, a informação de distância percorrida em quilômetros foi obtida.

Para o pré-processamento dos dados foi desenvolvido um script na linguagem de programação estatística R. O procedimento de transformação das variáveis consistiu nas seguintes etapas:

- Para obtermos a *duração mediana das viagens*, processamos as informações do horário de embarque e desembarque;
- Processamos a informação de cada viagem individualmente para calcular a *quantidade total de viagens*. Cada viagem era uma linha no dataframe. Logo, agregando por rota, obtivemos a quantidade total;
- *A distância percorrida* (em quilômetros) foi calculada a partir das coordenadas (latitude e longitude) do embarque e desembarque e, partir desses dados, foi calculada a mediana da distância.
- Para a *velocidade*, usamos a distância percorrida (em quilômetros) e a

duração da viagem (em hora);

- *O código do ônibus* onde a viagem foi feita foi obtido através do seu identificador.

3. DESENVOLVIMENTO

Inicialmente o trabalho se debruçou no pré-processamento dos dados, havendo dedicação de aproximadamente dois meses até que esse passo fosse completamente terminado. Após isso, houve o desenvolvimento de análises que poderiam trazer respostas relevantes sobre o transporte público de Curitiba, onde, a partir das mesmas, novas análises poderiam ser pensadas para desenvolvimento, com vistas a sua relevância. Esse processo durou cerca de três meses. Após alguns meses, notou-se a importância do uso dessas análises por aqueles que quisessem analisar o sistema de transporte público da sua localidade. Posteriormente, o trabalho se deu no desenvolvimento de um ferramental que pudesse ser usado independente do lugar a ser analisado.

No desenvolvimento da ferramenta, o maior desafio foi o de tornar as análises pensadas possível, visto a base de dados existente. Cada análise exigiu uma manipulação na base de dados para se chegar ao resultado almejado. Nesse processo, foi nítido ver a importância de uma boa base de dados para que as análises sejam viáveis.

4. RESULTADOS E DISCUSSÕES

4.1. Análise espacial de origens e destinos finais e dos horários de pico

A primeira análise proposta diz respeito à distribuição da movimentação nos ônibus ao longo de todo o dia. Isto é, o número de viagens que são feitas em cada intervalo de tempo, do início até o fim do dia. Esse intervalo de tempo será definido pelo usuário de acordo com a sua preferência. Assim, ele terá a possibilidade de definir sua própria granularidade temporal (comprimento do intervalo), por meio do código fonte disponibilizado, e poderá trabalhar com os dados aplicados a sua realidade e contexto.

No caso deste estudo de caso, para a cidade de Curitiba-PR, o horário definido foi de 4h às 23h, com intervalos de 1h, conforme pode ser

visto na Figura 1.

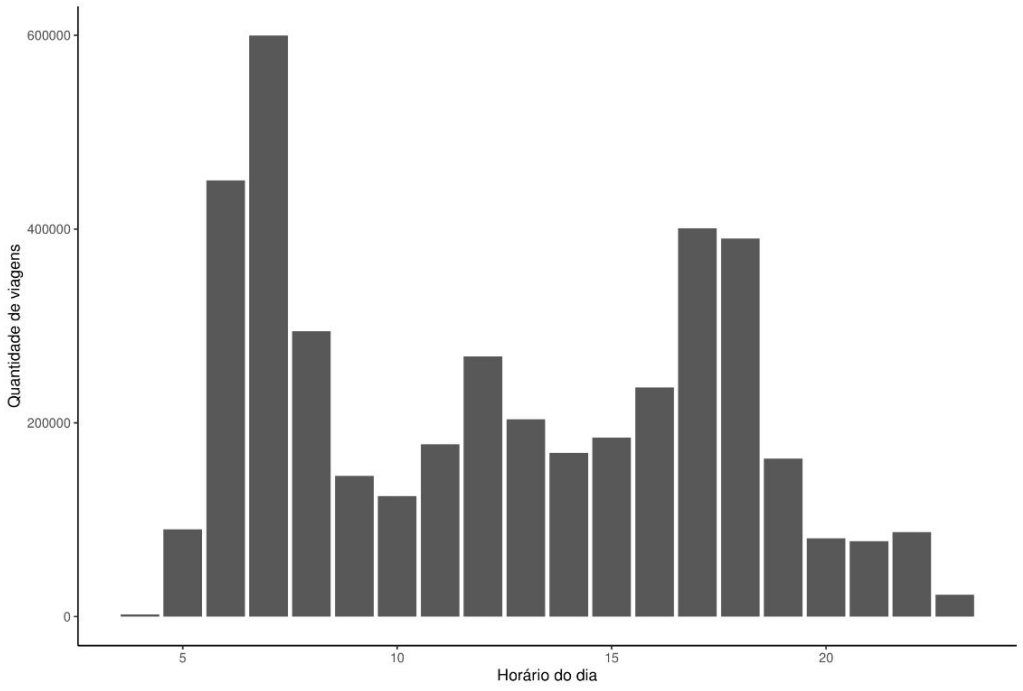

Figura 1: Quantidade de viagens por hora em Curitiba-PR.

Observa-se também os horários de pico, ou seja, quais são as faixas de horário em que a movimentação é maior. Os horários de pico demandam uma maior quantidade de viagens e, consequentemente, uma maior quantidade de ônibus e funcionários ativos. A melhor alocação desses recursos é de grande importância para que eles sejam utilizados de forma eficaz, evitando desperdícios e oferecendo o melhor serviço possível.

Como é apresentado na Figura 1, observamos os horários de pico compreendendo os horários de 6h às 8h, 11h às 13h e 17h às 19h, respectivamente. Denominamos esses horários como sendo *manhã*, *tarde* e *noite,* respectivamente.

No caso de Curitiba, o horário da *manhã* é o que concentra a maior quantidade de viagens feitas, seguido da *noite* e da *tarde*.

Por fim, é proposta uma análise espacial de origens e destinos finais das viagens. Ou seja, essa análise possibilita a visão de quais são os lugares da cidade mais demandados e em quais locais as pessoas mais embarcam. Essa análise também é feita de acordo com os horários escolhidos pelo pesquisador, gestor ou usuário. Assim, o usuário da ferramenta pode analisar a distribuição das origens e destinos de acordo com o seu interesse, incluindo os horários de pico ou não.

Vale ressaltar que a análise é a de origens e destinos finais do passageiro, Isto é, todas as viagens intermediárias que um passageiro faz até chegar seu destino, passando por diferentes pontos ou terminais de ônibus, não são considerados. Assim, é possível observar, onde, de fato, os passageiros embarcaram em seu destino inicial e onde desembarcaram para o seu destino final.

Para a faixa de horário com mais viagens em Curitiba, das 6h às 8h, como mostra a Figura 2, o embarque mostra-se bem distribuído em bairros periféricos e distantes do centro, onde geralmente se encontram bairros residenciais. Constata-se também, pela Figura 3, que, no mesmo horário, o desembarque se mostra mais acentuado e concentrado na região central da cidade. Por outro lado, na faixa de horário das 17h às 19h, o desembarque se acentua na região periférica da cidade (Figura 5) e, o embarque, na região central (Figura 4). Isso indica que na faixa de horário da manhã as pessoas tendem a sair de seus bairros em direção a região Central e bairros comerciais da cidade para trabalhar, estudar, etc, e, a noite, voltam para as suas casas.

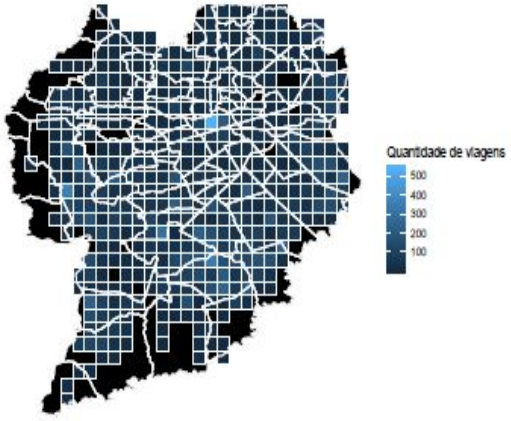

Figura 2: Embarque das viagens das 6:00 às 8:00

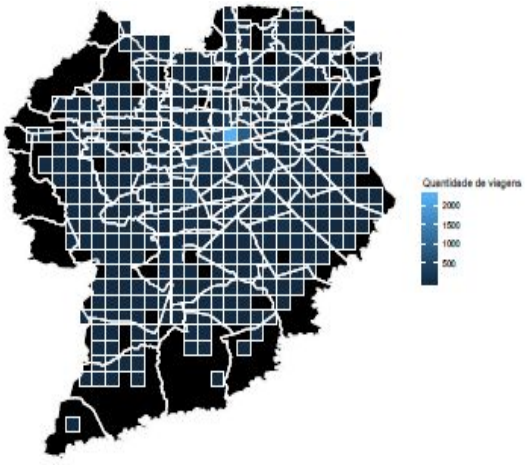

Figura 3: Desembarque das viagens das 6:00 às 8:00

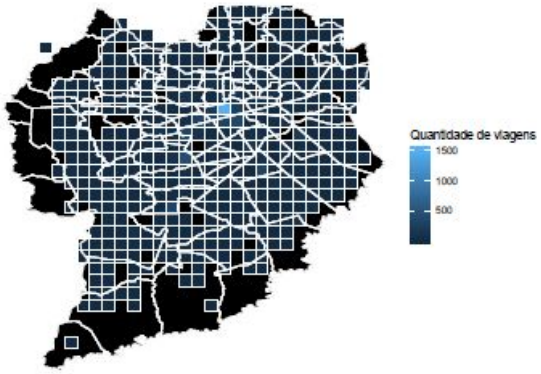

Figura 4: Embarque das viagens das 17:00 às 19:00.

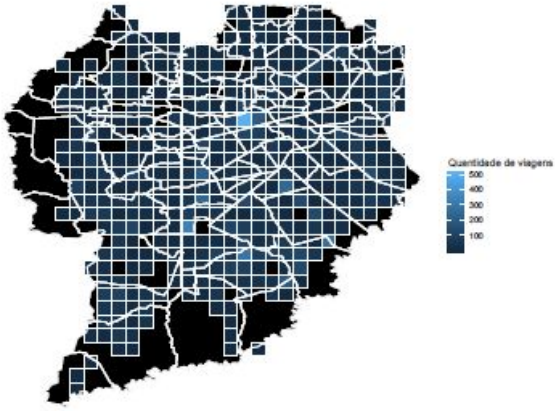

Figura 5: Desembarque das viagens das 17:00 às 19:00.

4.2. Lotação dos ônibus por dia, horário e rota

Uma realidade do transporte público é a sua propensão a atingir o limite da lotação em determinados horários e itinerários. Em muitos momentos, devido a isso, os usuários demoram mais tempo do que planejaram para fazer uma viagem e os gestores precisam lidar com maiores desafios de gestão em seus sistemas de transporte.

Uma outra análise proposta pelo presente trabalho é prover um panorama sobre a lotação dos ônibus seja qual for a linha, o horário ou dia pretendido. Sendo assim, o usuário da ferramenta pode verificar a lotação dos ônibus de uma determinada rota em qualquer dia ou horário que for do seu interesse, sendo possível verificar a quantidade de passageiros em cada ônibus que esteve fazendo viagem para uma linha específica em qualquer horário pretendido.

Dessa forma, os gestores terão uma forma mais fácil de fiscalizar a lotação por horário, obtendo maior capacidade de planejar e dispor os recursos demandados para o oferecimento de um serviço de melhor qualidade. Ao observar um desequilíbrio na quantidade de passageiros para os ônibus de uma mesma linha no horário, ele poderá procurar soluções para uma melhor distribuição de passageiros nos ônibus, evitando uma desigualdade relevante na quantidade de passageiros transportados para o mesmo destino na mesma faixa de horário.

Ressalta-se que a análise não é realizada indicando a quantidade

de passageiros viajando naquele exato momento, mas sim aqueles que fizeram check-in na rota no horário especificado

Para Curitiba, em um exemplo de aplicação da ferramenta para a rota 303 na faixa de horário de 18h às 19:59min do dia 02/05/2017, observa-se na Figura 6 que, no geral, os ônibus seguem uma média de quantidade de passageiros parecida no decorrer do horário, com a exceção de dois ônibus que apresentam uma quantidade de passageiros transportados bem maior que os demais, chegando a ter mais de 230 passageiros embarcando nos ônibus durante a faixa de horário.

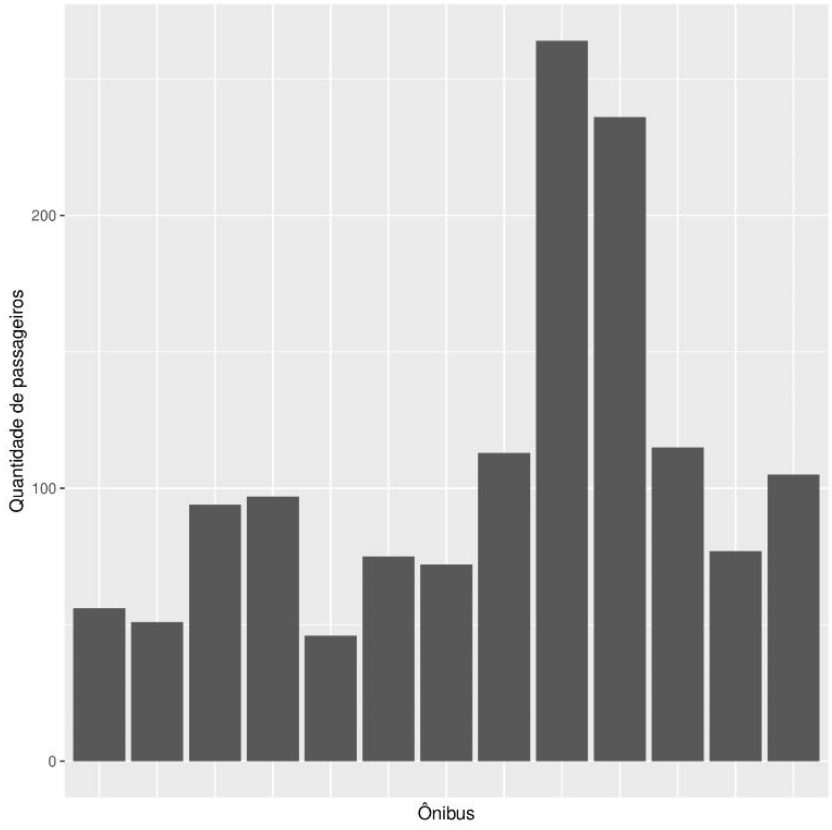

Figura 6: Quantidade de passageiros transportados nos ônibus na linha 303 de 18:00min

O gestor, ao observar a distribuição temporal dos ônibus e a quantidade de passageiros transportados, poderá enxergar através da transição de um ônibus para outro o desequilíbrio que pode haver entre a lotação dos dois. É importante destacar que os ônibus estão ordenados pelo primeiro check-in realizado na linha dentro do horário especificado.

4.3. **Detecção de outliers (viagens lentas)**

Os outliers são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva. Em outras palavras, um outlier é um valor que foge da normalidade e que pode causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise.

Detectar outliers se apresenta como uma tarefa de extrema importância para descoberta de conhecimento e mineração de dados, especialmente ao lidarmos com problemas em ciências aplicadas. E embora seja um tema abordado a um razoável tempo através de métodos estatísticos, se renova como um tópico de pesquisa de grande relevância nos dias atuais, devido ao grande crescimento na disponibilidade de dados para os pesquisadores e indústria. Em diversos cenários, os dados são tantos que realizar o processamento de todo o conjunto disponível é impraticável ou até mesmo indesejável. Assim, métodos capazes de selecionar aqueles dados com alto grau de distinção em meio a todo esse volume despertam grande interesse.

Diante disso, o presente trabalho sugere uma ferramenta de detecção de outliers objetivando buscar as viagens mais lentas realizadas. A intenção é dispor também ao usuário da ferramenta uma possibilidade de pré-processamento dos dados em função daqueles que fogem do comportamento esperado.

Nos dados de Curitiba-PR, podemos ver na Figura 7 que o padrão encontrado na relação entre distância percorrida e duração da viagem é diretamente proporcional. Isso quer dizer que, quanto maior for a distância da viagem, mais tempo o usuário demora para chegar a seu destino. Nosso objetivo é analisar as viagens que fogem desse padrão e se mostram como mais lentas que o normal. O critério utilizado para definir uma viagem como lenta foi a sua distância da nuvem de dados da Figura 7 que concentra a maior quantidade de viagens.

A ferramenta concede ao usuário as seguintes possibilidades de detecção dos outliers:

- Observar a quantidade de viagens lentas por dia da semana, mostrando assim quais os dias da semana onde as viagens tendem a ser mais lentas (Figura 8);

- Escolher um dia da semana específico e observar quais rotas apresentam a maior quantidade de viagens lentas. Nesse caso de uso, o dia escolhido foi a quinta-feira, porém a ferramenta permite a escolha de qualquer dia da semana (Figura 9).

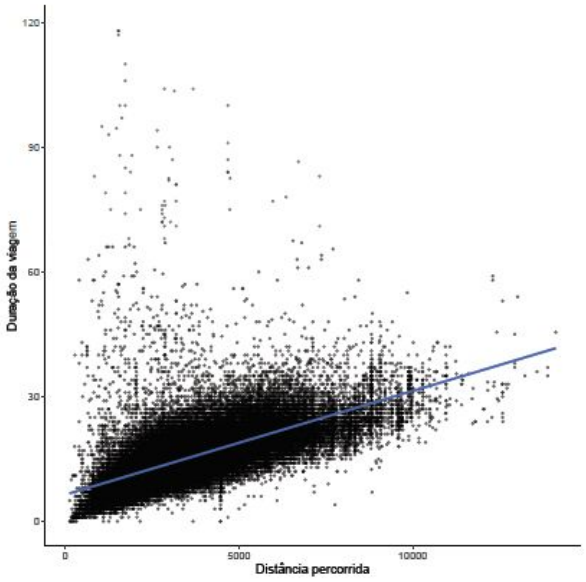

Figura 7: Relação da distância e da duração das viagens

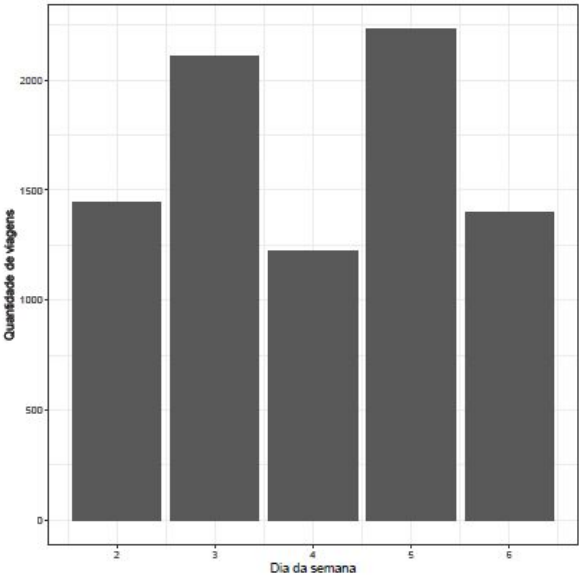

Figura 8: Quantidade de viagens lentas por dia da semana.

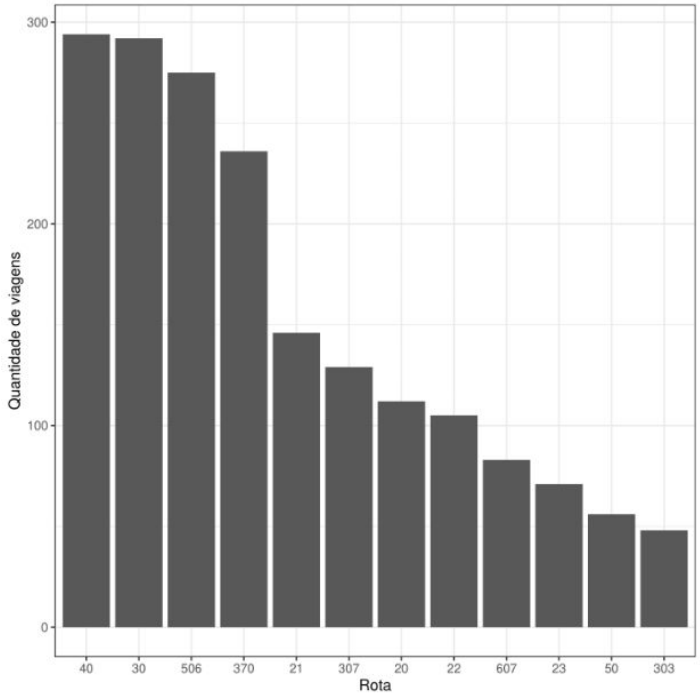

Figura 9: Quantidade de viagens lentas por rota (quinta-feira)

A possibilidade de detectar as viagens mais lentas no transporte público, nas diversas formas apresentadas acima, concede ao gestor uma maior capacidade de perceber onde há a necessidade de serem feitas correções e onde é necessário uma maior investigação para saber as razões que levaram à lentidão de determinadas linhas em determinados horários. A ferramenta é útil também para entender o funcionamento do transporte nos dias atípicos, ou seja, aqueles onde grandes eventos são realizados na cidade, exigindo uma dinâmica diferente da oferta de ônibus e da demanda de viagens em um horário específico.

5. CONCLUSÃO

Para a implementação da ferramenta, foi escolhida a linguagem de programação R tanto pela grande popularidade no desenvolvimento de análises, manipulação, e visualização de dados como pela usabilidade, facilidade na codificação, legibilidade e reutilização em projetos de análises de dados. O desenvolvimento se deu pela escolha de três análises que pudessem ser aplicadas em qualquer contexto de pesquisa no transporte público.

O maior desafio encontrado foi pensar e concluir análises que fossem realmente relevantes para os gestores e aqueles que trabalham com pesquisa no transporte público, em vistas da base de dados disponível. Pensar na utilidade das análises exigiu reuniões e refinamentos constantes com o orientador objetivando o esclarecimento de quais poderiam, de fato, agregar e ser útil para aqueles que buscassem por informações nesse contexto.

Em relação a aprimoramentos que podem ser realizados na ferramenta em trabalhos futuros, podem ser listados:

- Expansão da detecção dos outliers para a demanda de locais de destino;
- Visualizações mais avançados para as análises espaciais;
- Visualizações mais avançados para as análises de lotação dos ônibus;
- Análise da lotação dos ônibus em tempo real.

6. AGRADECIMENTOS

Gostaria de agradecer a oportunidade de trabalhar 12 meses no PIBITI/CNPq-UFCG, à Deus pelo dom da vida, ao meu professor orientador Nazareno que sempre foi solícito e dedicado a me ajudar no desenvolvimento do trabalho, ao meu amigo Tarciso pelas suas preciosas contribuições e por último à minha família que sempre batalhou para que eu pudesse ter acesso a universidade.

7. REFERÊNCIAS

1. Demographia. Demographia world urban areas, 2016.
2. United Nations Population Fund. United nations population fund - urbanization, 2016.
3. Bjorn Johnson. Cities, systems of innovation and economic development. ¨ Innovation, 10(2-3):146–155, 2008.
4. Moovit. Global cities public transit usage report, 2016.
5. BRAZ T. Inferring passenger-level bus trip traces fromschedule, positioningand ticketing data: Methodsand applications.Universidade Federal de CampinaGrande (UFCG), 2019