Deliverable 1

Project Members (Name, Email, Github, Class Year):

- *Taesung Yoon, tsyoon@bu.edu, @yoontsy2000, Class of 2024*
- *Jin Young Bang, jybang@bu.edu, @jinyoungbang, Class of 2024*
- *Minh Le: minhle@bu.edu, @lqm181, Class of 2024* - Team Rep
- *Katherine Rimey: rimey@bu.edu, @KTRimey, Class of 2024*
- *Duc Minh Nguyen, nguymi01@bu.edu, @nguymi01, Class of 2023*

Introduction

In this deliverable, we have performed a preliminary analysis of the MBTA Bus Departure Dataset, focusing on understanding the data and answering two key questions relevant to our project. Our analysis involved data collection, preprocessing, exploratory data analysis (EDA), and the generation of insights related to bus travel times and service disparities.

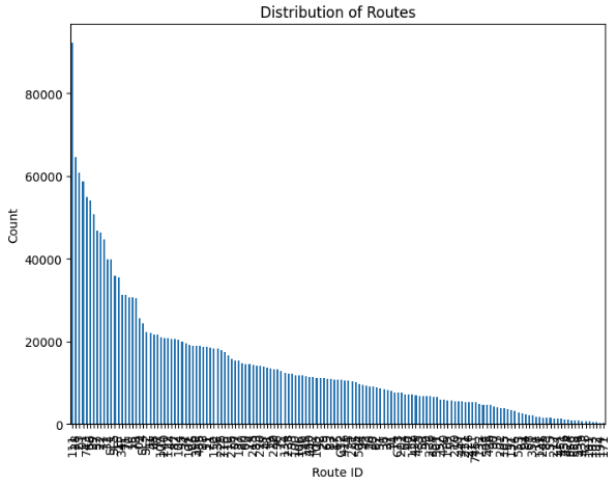
Data Collection and Preprocessing

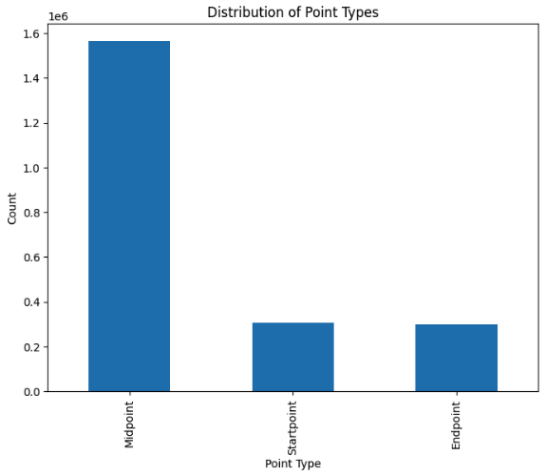
We initiated the deliverable by collecting and preprocessing a preliminary batch of data, focusing on the MBTA Bus Arrival Departure Times 2022 Dataset. This dataset contains information about bus departures, including scheduled and actual departure times, routes, directions, and various other attributes. The data was collected from January 1, 2022, to January 31, 2022, in line with our project's timeframe.

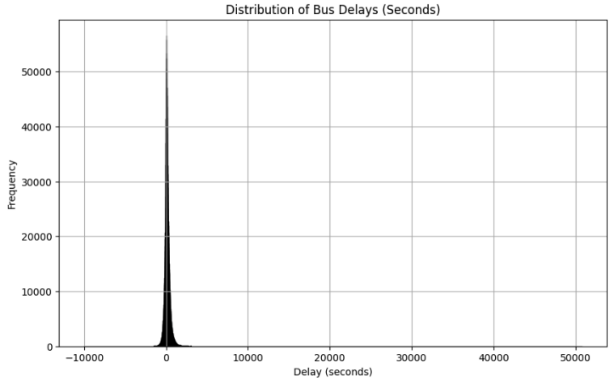
We performed the following data preprocessing steps:

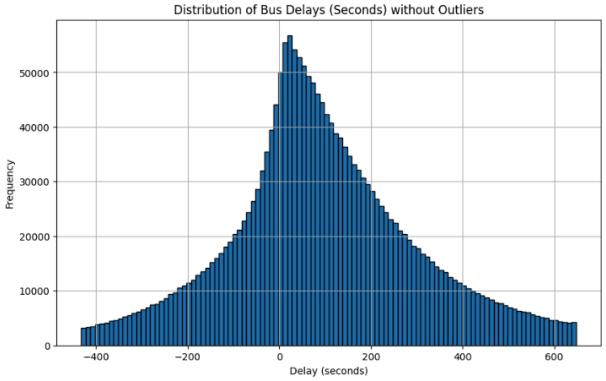
- Conversion of date and time columns to datetime objects.
- Removal of rows with missing scheduled or actual departure times.
- Calculation of delay in minutes and seconds by comparing scheduled and actual departure times.

Through our preliminary EDA, we generated some plots:





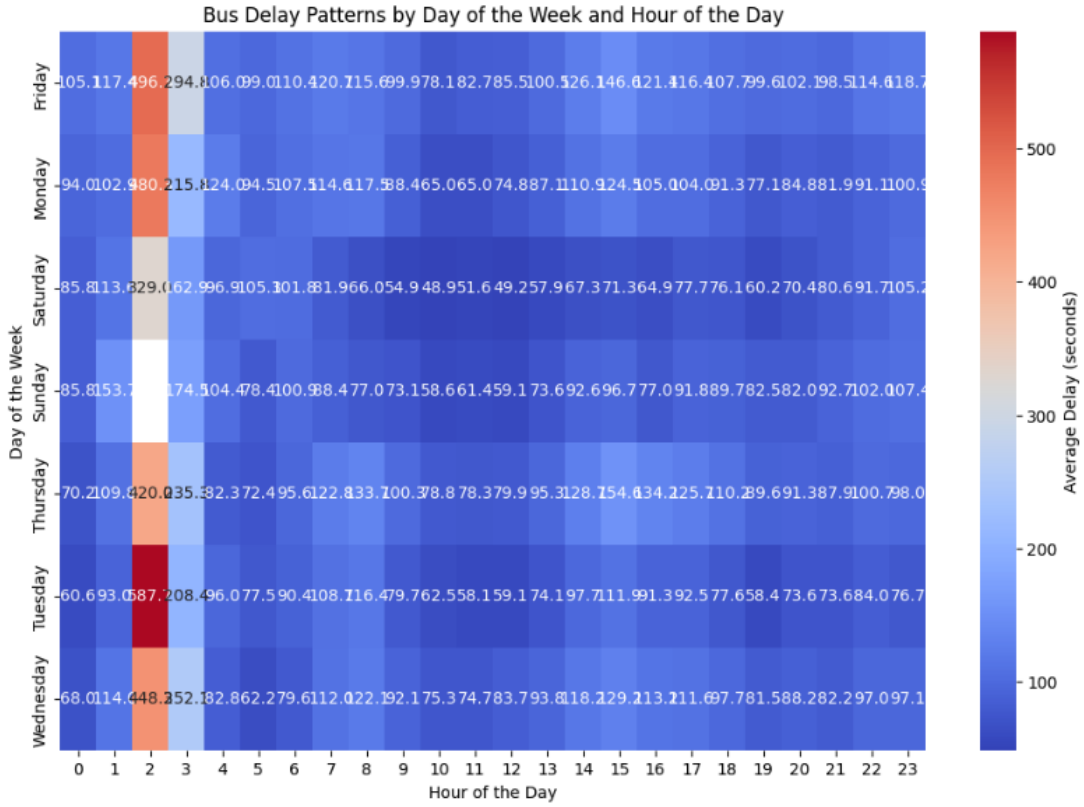




Some plots were not suitable for interpretation. For example, recognizing the challenges in interpreting the plot on the left side, we chose to enhance clarity by visualizing the data through interquartile ranges and grouping them in ten-second intervals. This approach facilitates a more accessible and informative visualization.

In our data analysis, we employed Seaborn to create a heatmap aimed at uncovering patterns in bus delay times based on the day of the week and the hour of the day. This visualization provides valuable insights into potential outliers in our dataset.

Upon closer examination of the heatmap, it becomes evident that Hour 2 exhibits a substantial average delay. This anomaly may be attributed to irregular data recording practices, potentially skewing our results. Further investigation is warranted to ensure data accuracy and reliability



Base Question 1: What are the end-to-end travel times for different bus routes?

To address the research question, our team initiated a preliminary investigation to identify suitable datasets related to MBTA Bus Routes and travel times. Subsequent research led us to discover a dataset containing bus arrival and departure events. We proceeded to download and meticulously analyze this dataset, determining that it contained sufficient data to effectively address the initial and secondary research inquiries.

Since we knew the timeframe focus of this project was January 1, 2022 - January 31, 2022 (mentioned in the overview document), we specifically downloaded the MBTA Bus Arrival Departure Times 2022.

Analysis Steps:

- We processed the data to determine the end-to-end travel times for each bus route, considering both inbound and outbound directions.
- The data was further analyzed to compute raw, average, and median travel times for each route.

Key Findings:

- The analysis revealed variations in travel times for different bus routes.
- We observed that routes closer to Downtown Boston had shorter end-to-end travel times, while routes farther from the center exhibited longer commute durations.

Base Question 2: Disparities in Service Levels of Different Routes

Analysis Steps:

- Similar to question 1, we processed the data to calculate delay times for each bus route.
- We then analyzed the data to identify the most delayed routes, both inbound and outbound.

Key Findings:

- The analysis highlighted significant disparities in service levels across different bus routes.
- We identified routes with the longest average delay times, indicating potential areas for improvement.

Data Visualizations

Throughout the analysis, we created data visualizations to provide insights into the patterns and disparities in bus travel times and service levels. These visualizations included bar charts, box plots, and interactive maps generated using libraries such as Matplotlib, Seaborn, Folium, and Geopandas. These visualizations enhanced our understanding of the data and made our findings more accessible.

Conclusion

In this deliverable, we conducted a preliminary analysis of the MBTA Bus Departure Dataset, aiming to address two key questions related to bus travel times and service disparities. Our findings offer insights into the performance of different bus routes and highlight areas for further investigation and potential improvements in the public transportation system.

The analysis and visualizations presented in this deliverable serve as a foundation for our ongoing research, which will continue to explore the dynamics of the MBTA bus network, identify patterns, and propose data-driven recommendations for enhancing service quality.

For more detailed information, code, and visualizations, please refer to the `main.ipynb` Jupyter Notebook provided in the pull request.