PROJECT REPORT

City of Boston: Transit & Performance: Team E

Members: Taesung Yoon, Jin Young Bang, Minh Le, Katherine Rimey, Duc Minh Nguyen

Problem Statement

Servicing over a million people daily, the MBTA contributes an estimated annual economic value of around $11.5 billion to the greater Boston area. Public transport is a linchpin for Massachusetts and Boston residents, impacting economic development, environmental considerations, and equity. While public transit falls under the purview of the MBTA as a state agency, the City of Boston can influence decisions, particularly regarding bus routes, on behalf of the community.

The project aims to analyze MBTA bus data, uncovering service performance trends across geographical areas. Focusing on potential disparities by neighborhood and demographic factors, our goal is to provide insights into the impact of bus performance on Boston residents, emphasizing economic development, environmental sustainability, and equity. This analysis is pivotal for enhancing public transportation, residents' quality of life, and environmental benefits, offering valuable guidance on optimizing the allocation of buses in different regions of Boston.

Data Collection and Cleaning

The analysis commenced with the collection and preprocessing of an initial set of data, specifically focusing on the MBTA Bus Arrival Departure Times 2022 Dataset. This dataset, encompassing information regarding bus departures from January 1, 2022, to January 31, 2022, aligns with the timeframe established for our project. Key attributes within the dataset include scheduled and actual departure times, routes, directions, and various other pertinent details.

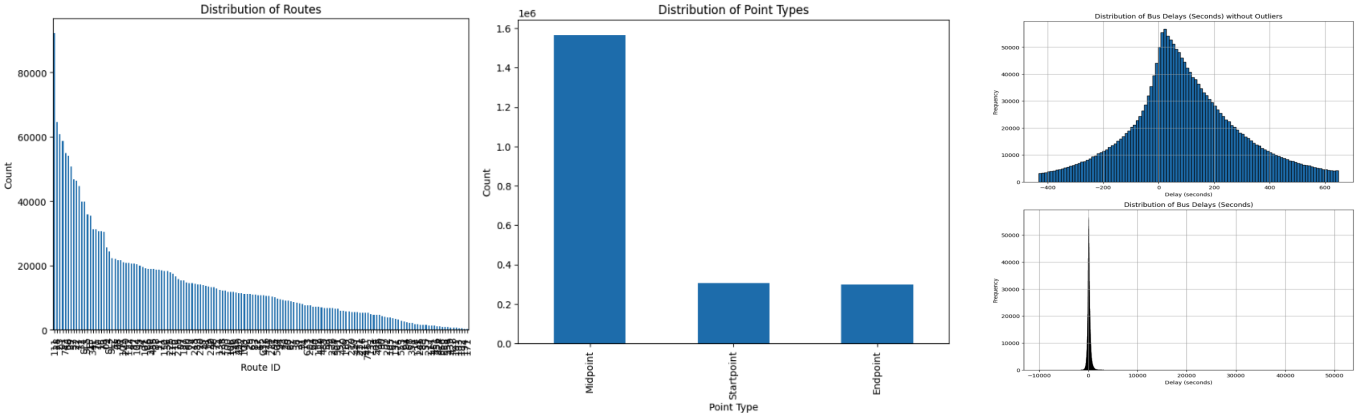
We utilized the MBTA V3 API for bus stop metadata, Analyze Boston's Census Data for insights to answer Base Question 3, and downloaded MBTA bus stops and Boston neighborhood shape files to improve map visualizations.

In the preprocessing phase, we implemented several crucial steps to enhance the data's suitability for analysis:

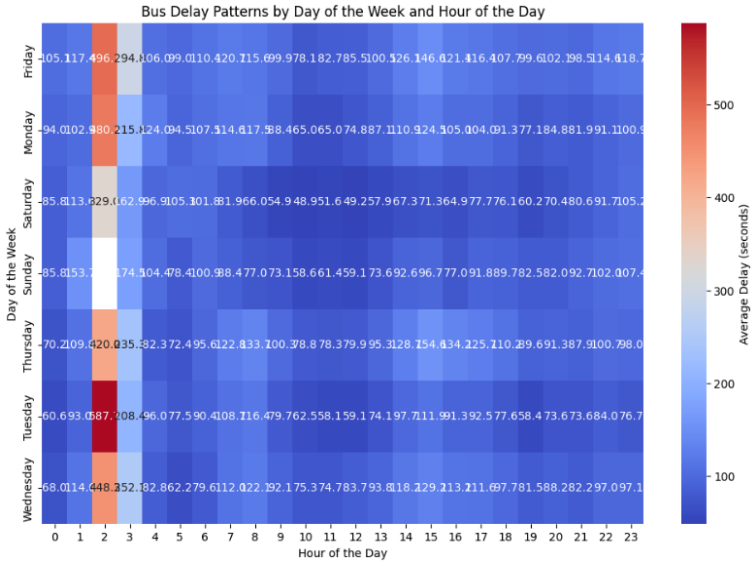
- Conversion of date and time columns to datetime objects: This transformation ensures a consistent and standardized format for temporal information, facilitating accurate analysis of time-related patterns.
- Removal of rows with missing scheduled or actual departure times
- Calculation of delay in minutes and seconds: By computing the delay between scheduled and actual departure times, we gained valuable insights into the temporal efficiency of bus departures. This metric serves as a foundational element for understanding the performance and punctuality of the bus services.

These preprocessing steps laid the groundwork for a robust analysis of the MBTA Bus Arrival Departure Times 2022 Dataset, setting the stage for subsequent exploratory data analysis and visualization to uncover meaningful insights and patterns within the data.

Exploratory Data Analysis



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.

Key Base Questions

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.

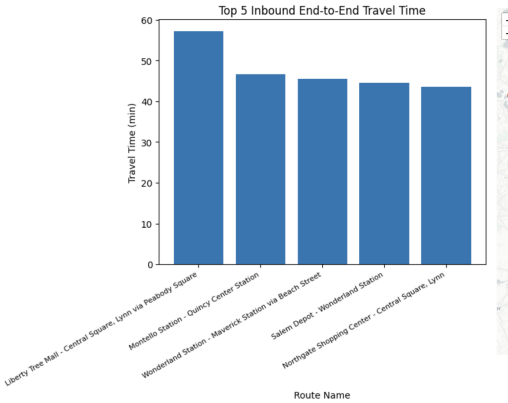
Since we knew the timeframe focus of this project was January 1, 2022 - January 31, 2022 (mentioned in the overview document), we specifically downloaded the MBTA Bus Arrival Departure Times 2022.

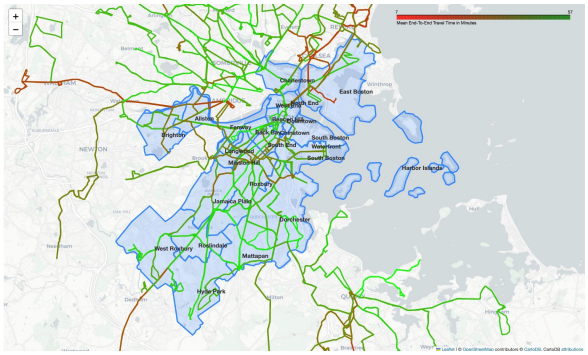
Analysis Steps:

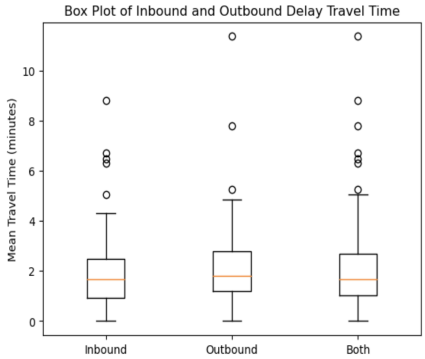
- We processed the data to determine the end-to-end travel times for each bus route, considering both inbound and outbound directions.
- The data was further analyzed to compute raw, average, and median travel times for each route.

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.







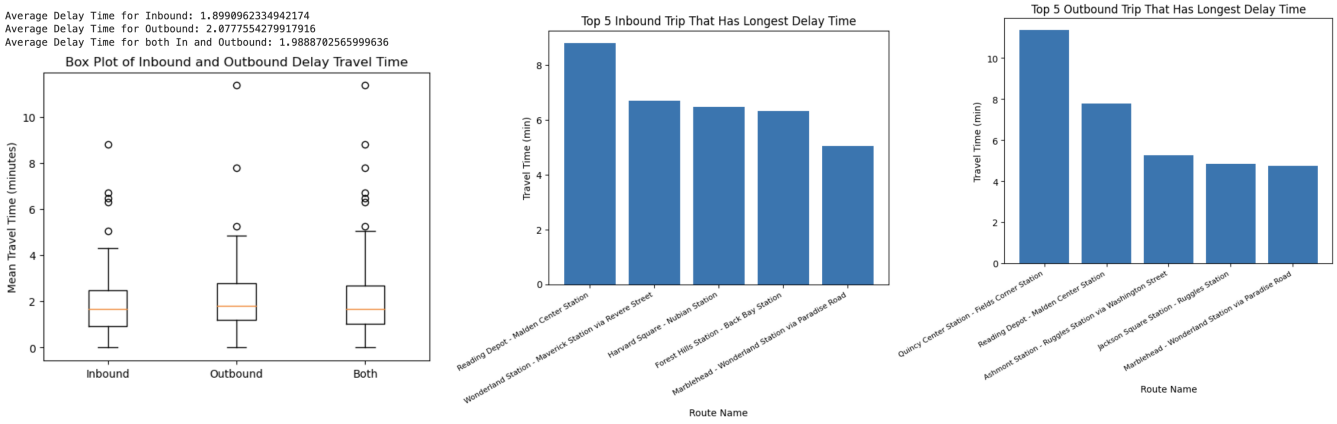
Base Question 2: Disparities in Service Levels of Different Routes

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.



<u>Base Question 3: Population Sizes/Characteristics of the Communities Serviced by Different Bus Routes; DIfferences in Characteristics of People Most Impacted</u>

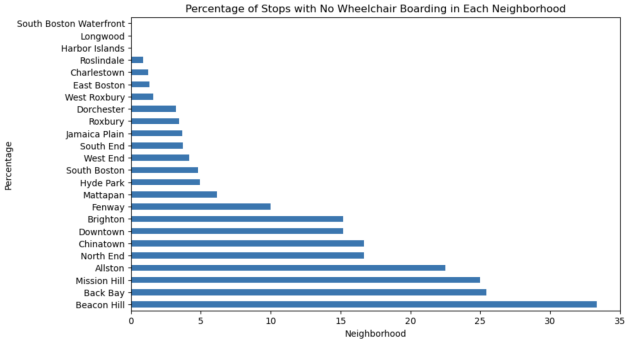
To address Base Question 3, our team conducted an analysis focusing on the population sizes and characteristics of communities served by different bus routes in Boston. Due to the absence of neighborhood information in the MBTA API, we utilized Shapefiles, converted CRS coordinates, and intersected routes, stops, and boundaries to associate each bus stop with its respective neighborhood.

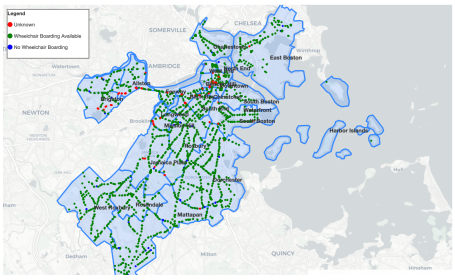
Methodology:

1. Geospatial Mapping:
    a. Utilized Shapely and PyProj to convert latitude and longitude coordinates to Boston neighborhood boundaries.
    b. Assigned each bus stop to its corresponding neighborhood, facilitating a community-centric analysis.
2. Demographic Characteristics:
    a. Integrated census data to understand population characteristics within neighborhoods, including total population and racial demographics.
3. Wheelchair Boarding Accessibility:
    a. Investigated the correlation between neighborhood demographics and wheelchair boarding accessibility.
4. Route-to-Stop Ratio:
    a. Calculated the ratio of bus routes to stops for each neighborhood.

Key Findings:

1. Demographic Disparities in Wheelchair Boarding:
    a. Neighborhoods with a significant Asian population show a higher prevalence of stops without wheelchair boarding facilities, indicating a potential accessibility concern for individuals with mobility challenges.
2. Route-to-Stop Ratio Disparities:
    a. POC-dominant neighborhoods exhibit a lower ratio of bus routes to stops, suggesting a potential need for improved public transportation infrastructure in these areas.





Extension Proposal

Extension Pitch	To explore the impact of seasonal times and rush hours on transportation patterns. By scrutinizing data across weekdays, weekends, and specific times over a longer time frame we seek to uncover nuanced insights that can inform urban planning and traffic management strategies.
Rationale	Understanding variations in transportation patterns during rush hours and seasonal changes is crucial for optimizing urban mobility. This extension provides an opportunity to reveal hidden trends within the dataset, aiding in resource allocation and infrastructure enhancements.
Questions for Analysis	● How do travel patterns differ during rush hours and non-rush hours? ● Are there distinct variations between weekdays and weekends? ● What trends emerge during different times of the day? ● How do external factors like weather impact transportation during these periods? ● Are there differences in delays/travel times for different months?
Data Sets & Sources	We will use the existing transportation dataset and process our data for the entire year of 2022, incorporating timestamps, days of the week, and weather conditions. Supplementary data on events, road closures, or public transportation schedules may also be considered.
Data Visualizations	Hourly Traffic Heatmap (GIF Animations); Weekday vs Weekend Chart; Line Graph of Delay Times over differents month (Time-Series Analysis)
Additional Information	Consideration of external factors, like weather and local events, will provide context for observed trends. Collaboration with transportation authorities and experts will contribute to actionable insights for urban planners and policymakers.

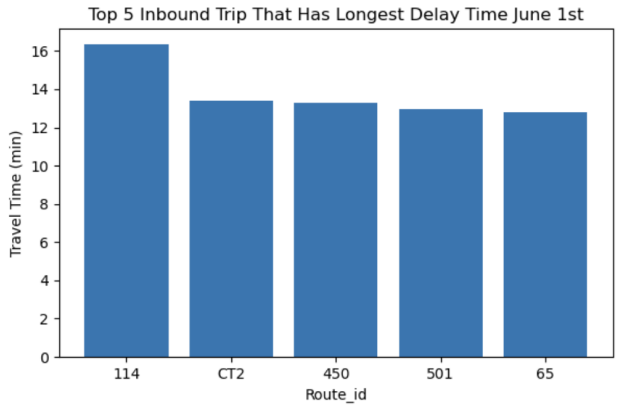
Visualization and Insights for Extension Proposal

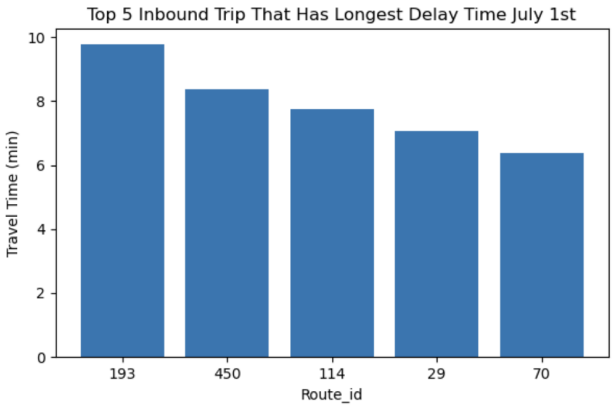
Differences in Months

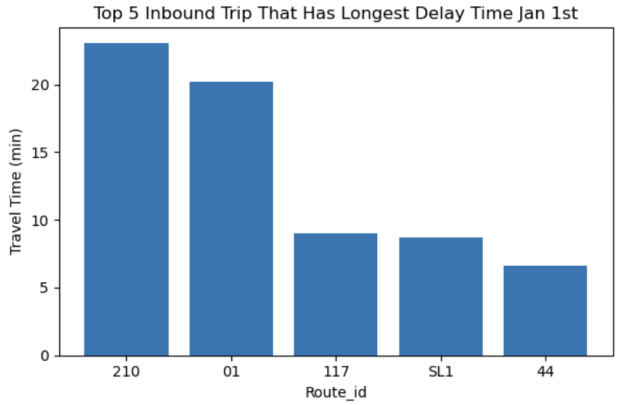
We began our analysis by focusing on the impact of seasonal changes, specifically comparing transportation patterns in January, June, and July. The code processed the data for January 1st, June 1st, July 1st, and additional days for each month, considering factors such as scheduled and actual timestamps, service dates, and day of the week.

We computed the average delay times for inbound trips, categorizing them by route and direction. This information was then used to create visualizations showcasing the top 5 inbound trips with the longest delay times for each month.

The bar charts revealed variations in delay times across different months. For example, the top 5 inbound trips with the longest delay times in June differed from those in January and July. This suggests that seasonal factors play a role in transportation delays.







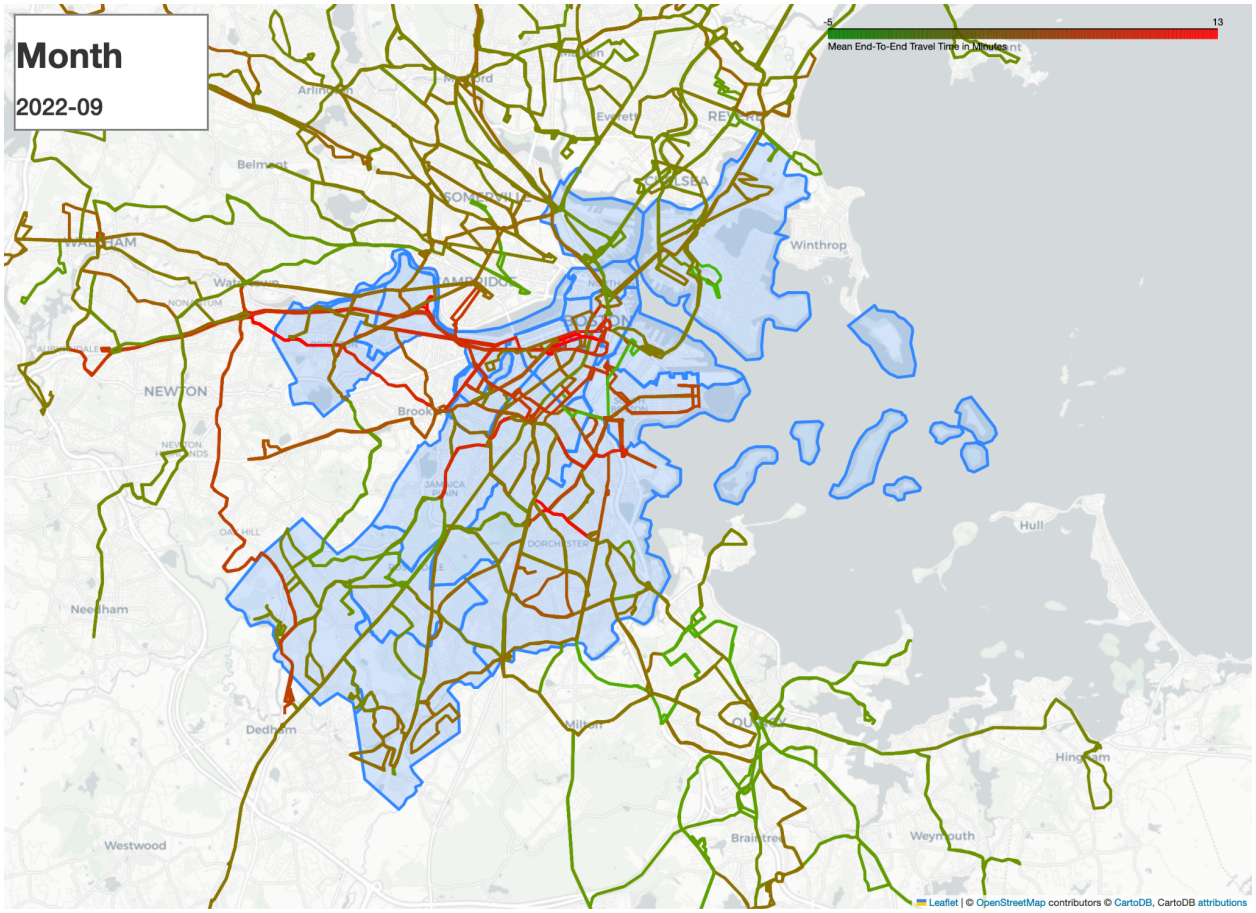
Average Delay Time for Inbound January: 2.54 min

Average Delay Time for Inbound June: 4.23 min

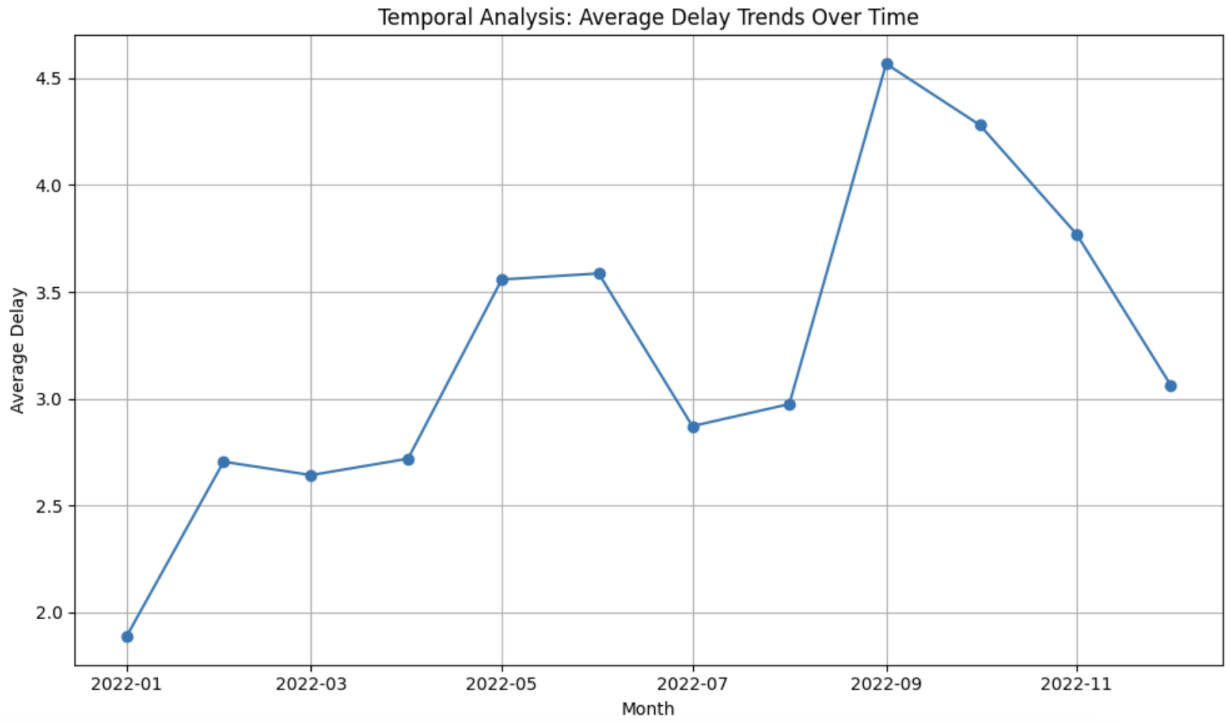
Average Delay Time for Inbound July: 2.78 min

As part of our comprehensive analysis, we incorporated map visualizations to explore potential geographic insights and trends in transportation patterns for each month.

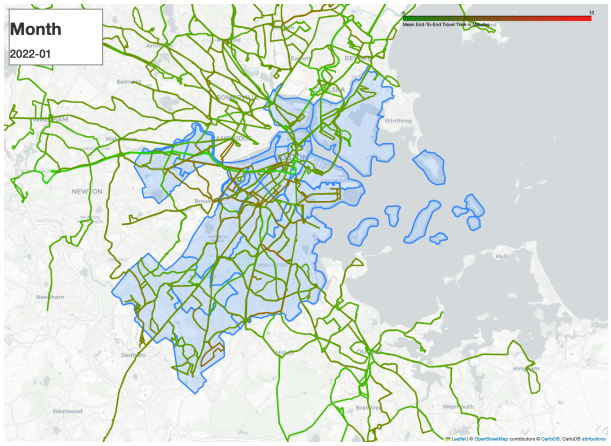
Upon examining the map visualizations, a noteworthy trend emerged in the month of September. This month stood out with a higher incidence of delays compared to others in our dataset. The key observation suggests a probable correlation between the increased delays and the influx of students moving into Boston during this period.

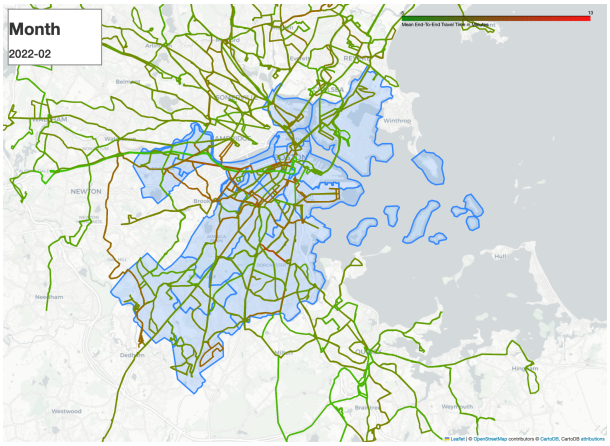


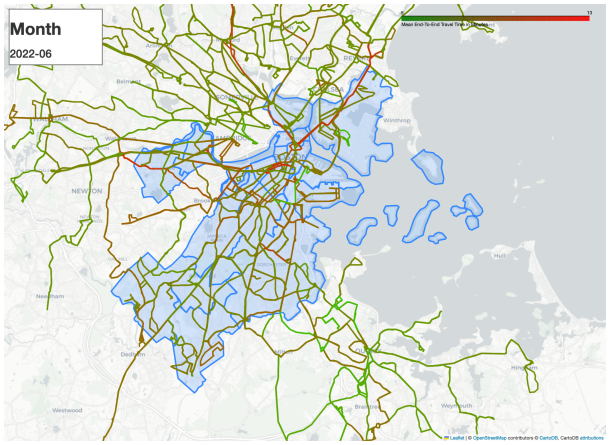
September, being a critical period for academic activities, witnesses a significant surge in students relocating to Boston for the start of the academic year. This influx can lead to heightened demand for transportation services, potentially resulting in delays across various routes.

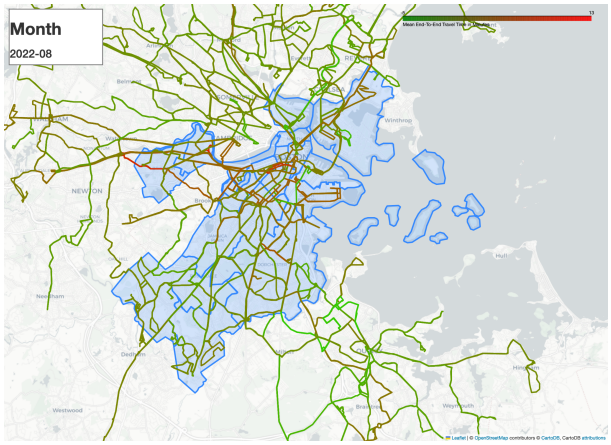


In our exploration of transportation patterns, a notable trend emerged during the winter months of January to March. Contrary to other periods, these months exhibited comparatively lower incidences of delays. This observation raises intriguing questions about the potential influence of seasonal factors on transportation efficiency. The winter season, characterized by reduced outdoor activities and a potential decline in tourism, could contribute to a less congested transportation network. Additionally, weather conditions during these months may play a role in promoting smoother traffic flow. Understanding the seasonal ebb and flow of delays is crucial for urban planning efforts, as it offers valuable insights into how external factors, in this case, the winter season, can contribute to a more resilient and optimized transportation infrastructure. Further investigation into the nuanced interplay between seasons and transportation patterns could provide a deeper understanding of the dynamics at play, facilitating proactive strategies for traffic management and resource allocation.





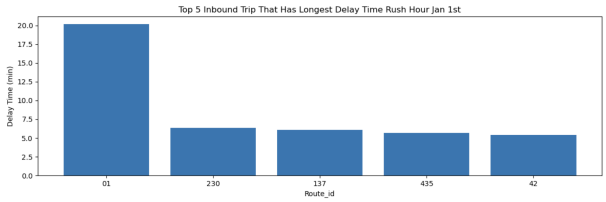


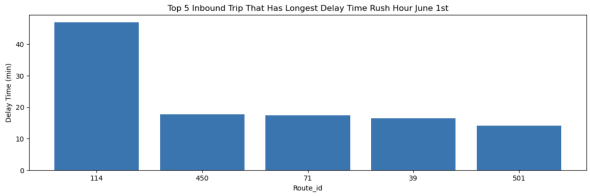


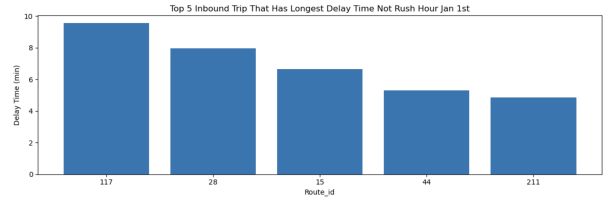
Rush Hour Analysis

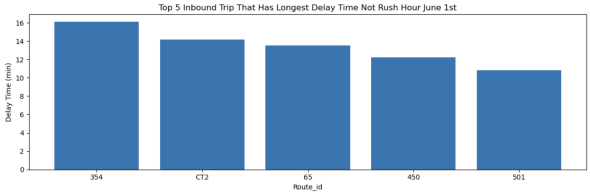
To delve deeper into transportation patterns, we classified entries into rush hour and non-rush hour categories based on timestamp information. The provided code successfully identified rush hours, helping us distinguish between periods of high and low traffic demand.

Using the rush hour classification, we computed the mean delay times for inbound trips during rush and non-rush hours. This information was then visualized in bar charts, showcasing the top 5 inbound trips with the longest delay times for both rush and non-rush hours.





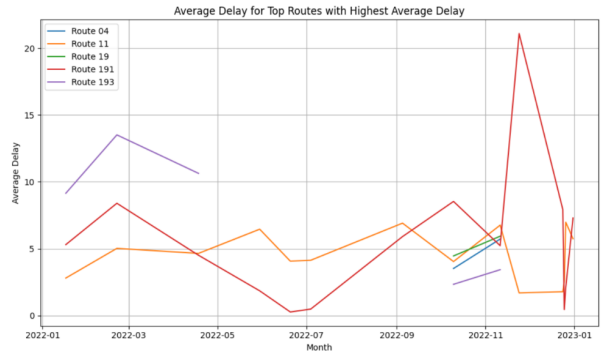


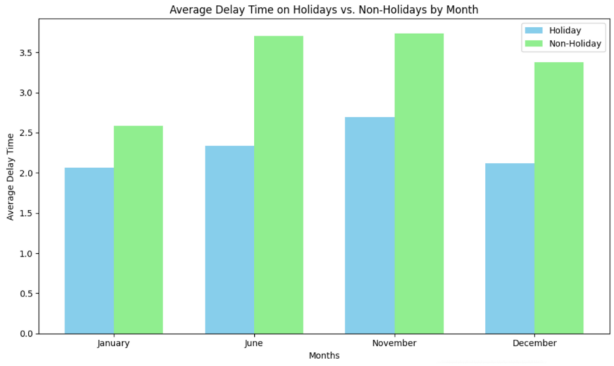


The rush hour analysis uncovered intriguing patterns. For instance, the delay times during rush hours in June were notably higher compared to January, possibly due to factors such as increased tourism and students having more free time to travel.

Holidays vs. Non-Holidays

We also decided to analyze trends and differences between holidays vs non-holidays and were curious to see if it impacted bus delay times.





In our analysis, we utilized a line graph to examine the top routes with the highest average delays during holidays. The general trend indicates that delays seem to be longer during these periods. This could be attributed to increased traffic or other holiday-related factors that affect bus schedules.

However, an interesting observation from the bar chart on the right contradicts this trend. It suggests that non-holidays actually have a longer bus delay time. This could be due to a higher number of routes being operational during non-holidays, leading to increased congestion and subsequently, longer delays.

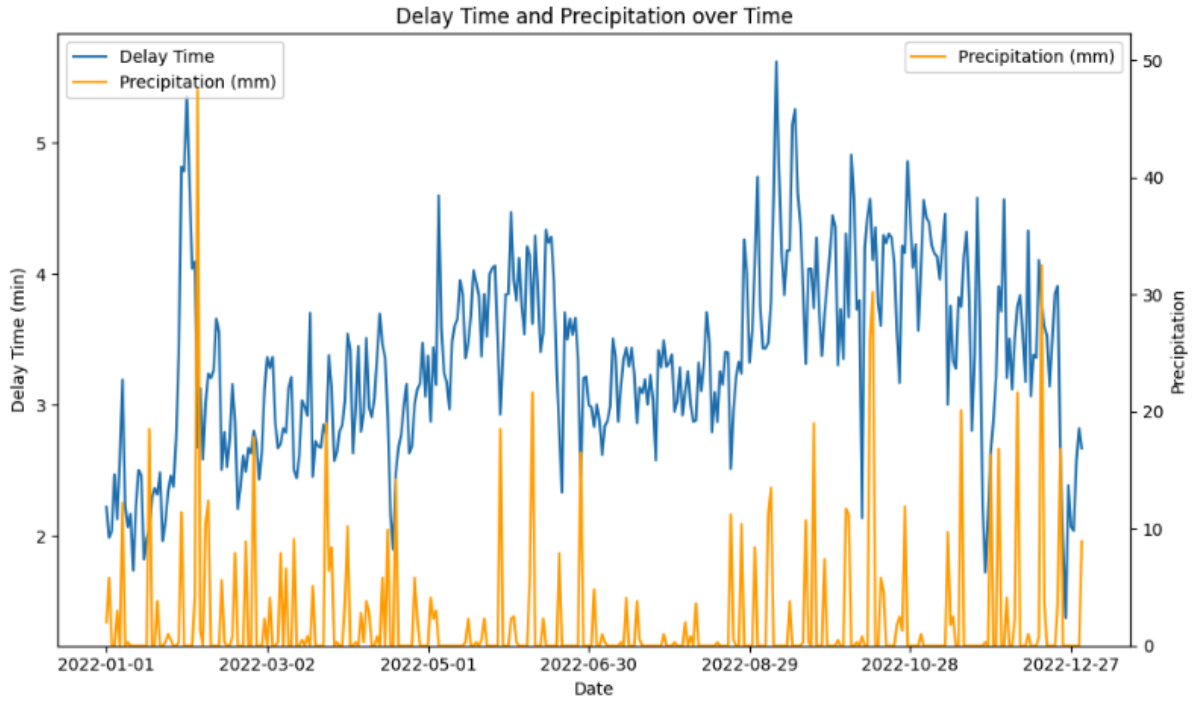
The skew in the delays for specific routes might suggest that certain bus routes are more congested due to their destinations. For instance, routes visiting specific venues or popular

areas might experience higher delays. However, during holidays, there could be less overall delays due to fewer routes being available.

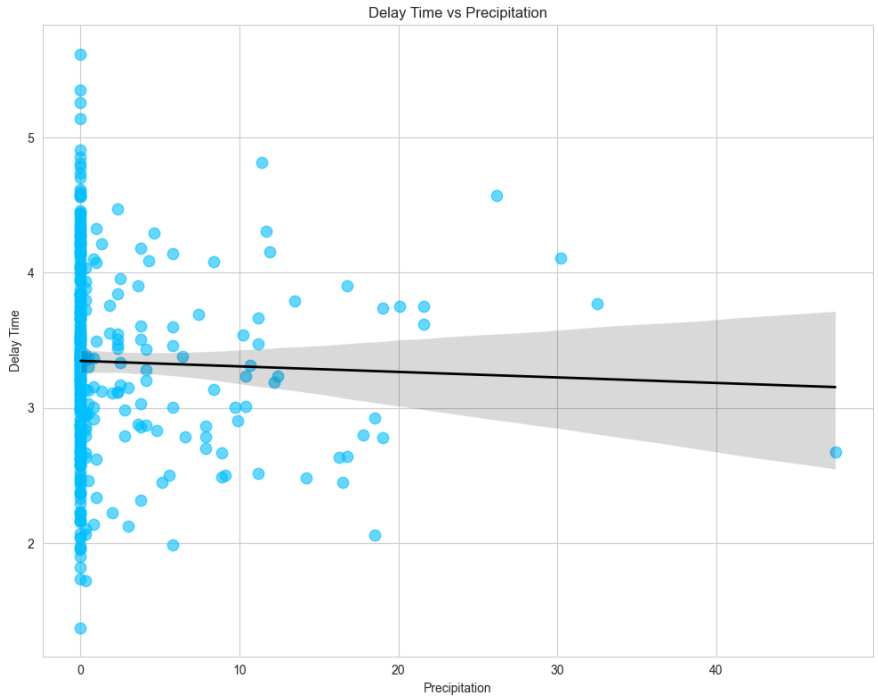
This analysis provides valuable insights into the factors affecting bus delays and can help in devising strategies to improve bus schedules and reduce delays, especially during holidays. Further studies could investigate the specific causes of delays on the most affected routes and propose targeted solutions.

Weather-related Analysis

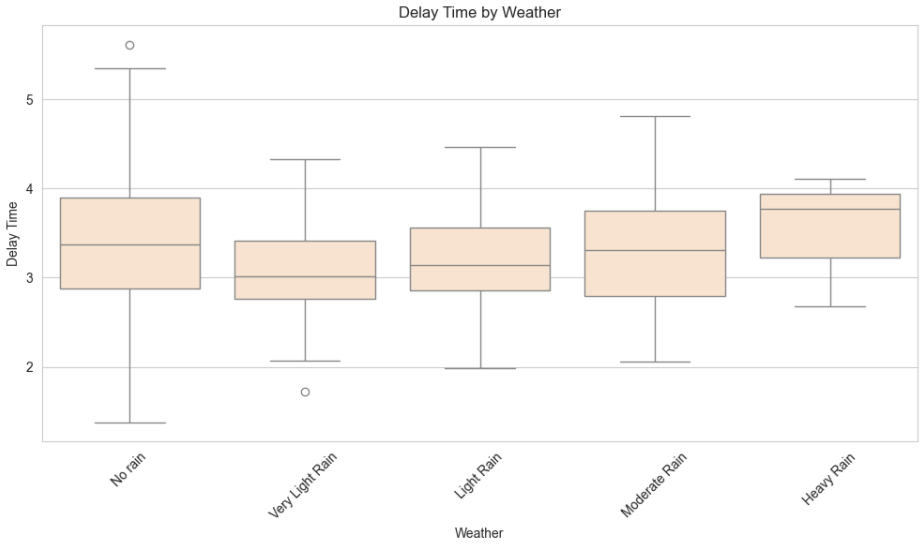
To investigate the impact of weather on bus traffic, we assessed temperature, precipitation, and the monthly mean bus delay times. While our analysis provides valuable insights, conducting a more comprehensive examination, including additional weather conditions and daily precipitation breakdowns, could offer a deeper understanding of weather's influence on bus delays.



The graph illustrating the daily average delay times of buses alongside the average precipitation levels indicates that a strong correlation between the amount of delay and the average precipitation of the day isn't evident. Notably, days with substantial precipitation levels, such as one observed in February from the graph, sometimes exhibit comparatively lower delay times in contrast to other days.

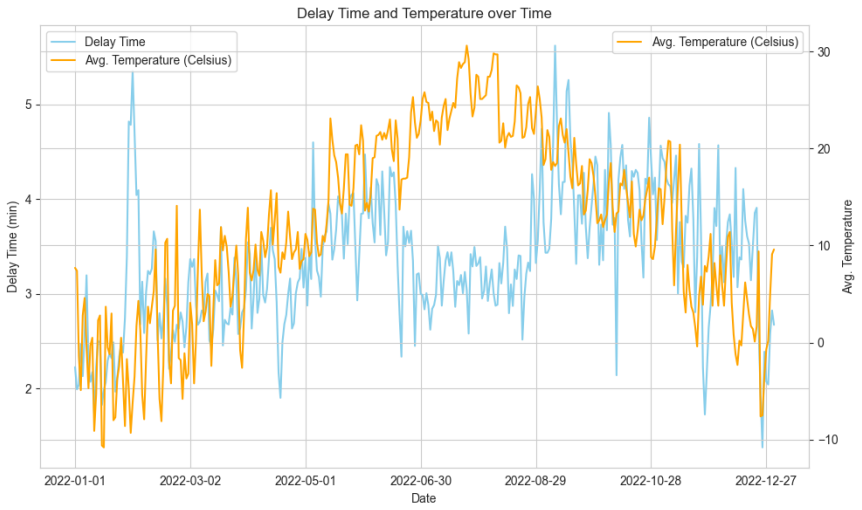


We employed a residual plot to discern any linear relationship between bus delay time and precipitation levels. The clustering of a significant number of data points at a precipitation level of zero indicates that a majority of the days in 2022 were rain-free. Moreover, the variance in delay time diminishes as precipitation levels increase, contradicting the assumption of homoscedasticity in the linear hypothesis. Hence, we conclude that a linear relation between the level of precipitation and delay time isn't supported by our analysis.

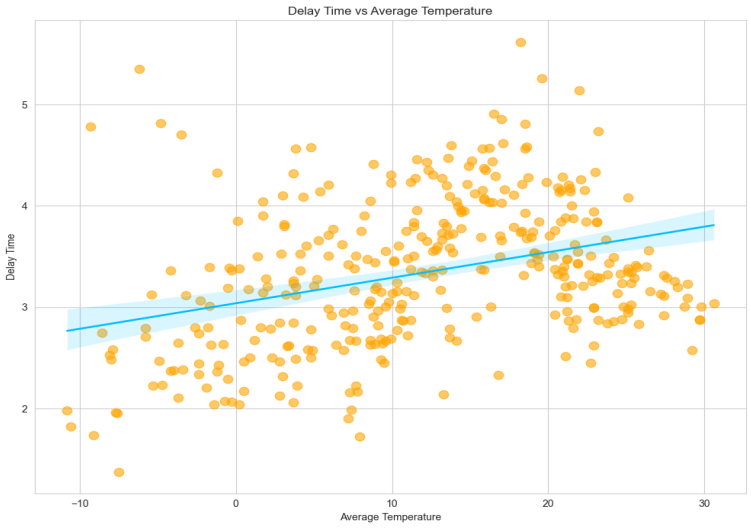


Categorizing precipitation levels into groups such as 'No Rain,' 'Very Light Rain,' 'Light Rain,' etc., based on defined classifications https://www.nchm.gov.bt/attachment/ckfinder/userfiles/files/Rainfall%20intensity%20classification.pdf, revealed

that delay times fluctuate between 3 to 4 minutes across all weather types, with some outliers for 'No Rain' and 'Very Light Rain.' Interestingly, variance in delay time significantly widens when there's no rain, while it notably diminishes during heavy rain episodes.



The graph plotting the average temperature against bus delay time throughout 2022 indicates a somewhat correlated trend, except during the summer months. During summer, while temperatures increase, delay times decrease. Possible reasons include reduced traffic and bus usage due to a large number of students leaving Boston for summer break. Additionally, residents might be less inclined to use buses in the summer due to discomfort caused by higher temperatures at bus stations.



Examining the average daily bus delay time against the daily average temperature of Boston in 2022 reveals a consistent variance in delay distribution with temperature changes. This observation hints at the potential for a linear relationship between delay time and temperature.

Challenges and Limitations

Challenges with Data Sources

Integrating the data proved challenging due to its decentralized nature. The information, scattered across different sources and formats, posed difficulties in merging it effectively. This decentralization made it tough to combine and correlate datasets seamlessly. As a result, extracting cohesive insights became more complicated, limiting the depth of our analysis. The effort to synchronize these decentralized datasets encountered complexities, impacting the efficiency and depth of our analytical endeavors.

Our analysis encountered several constraints that affected the depth and breadth of insights. Accessing historical weather data from external public APIs posed a challenge, limiting our ability to comprehensively assess weather-related influences on bus delays. Additionally, while striving to identify additional predictors or features that might impact delays, our search yielded no significant findings, restricting our ability to explore potentially crucial variables.

Scope Limitation

Furthermore, despite extending our analysis to encompass the entirety of 2022, the constraint of focusing solely on this timeframe implies a potential limitation. The exclusion of other years might restrict the holistic understanding of trends and patterns within the transit system, indicating a scope limitation that could have enriched our findings with broader temporal insights.

Conclusion

The analysis of the MBTA Bus Arrival Departure Times 2022 Dataset has revealed crucial insights into Boston's public transportation system. Our study focused on understanding bus performance and its impact on economic development, environmental sustainability, and equity.

Base Questions:

- We found that travel times varied significantly among different bus routes. Routes closer to Downtown Boston generally had shorter travel times compared to those farther away, indicating potential differences in commuting experiences across the city.
- Our analysis highlighted significant differences in service levels across various bus routes. We identified routes with the longest average delay times, suggesting areas that could benefit from improvements in service efficiency.
- Our study revealed disparities in accessibility and infrastructure, particularly for neighborhoods with a significant Asian population and communities of color. We observed areas lacking wheelchair boarding facilities and identified neighborhoods with fewer bus routes compared to the number of stops, highlighting potential areas for targeted improvement in public transportation.

Extension Project:

- Rush hours exhibited higher delays compared to non-rush hours, indicating increased traffic congestion and potential impacts on commute times.
- We noticed differences in delay patterns between weekdays and weekends, possibly due to altered travel behaviors or traffic densities during these periods.
- We observed varying delay patterns throughout the day, highlighting fluctuations in transportation efficiency during different time brackets.
- While weather conditions were explored, a clear linear relationship between precipitation and bus delays wasn't evident. However, temperature variations seemed to correlate with delay times, particularly during certain months.
- Delays varied across different months, suggesting potential seasonal influences on transportation efficiency. Months with specific events or academic influxes showcased higher delays compared to others.

These conclusions shed light on the nuances of Boston's transit system, highlighting areas for improvement, disparities in service, and potential influences of external factors on transportation patterns.

Suggestions for Potential Improvements

1. Improve Service Equity: Address disparities identified in wheelchair accessibility and route-to-stop ratios in certain neighborhoods. Consider prioritizing these areas for infrastructure improvements to ensure inclusivity and accessibility for all residents.
2. Route Optimization: Focus on routes with longer delays and lower service levels. Consider reallocating resources or adjusting schedules to improve efficiency and reduce delays, especially in areas farther from the city center.
3. Seasonal Planning: Leverage insights into seasonal variations to optimize schedules, staffing, and routes during periods of increased demand or influxes, such as academic seasons or major events.
4. Explore Additional Predictors: Specific to the project itself and despite the challenges, continue the search for additional factors influencing delays. Investigate other potential predictors or features beyond the currently explored variables to gain a more comprehensive understanding of factors impacting bus performance.

Individual Contributions

Taesung Yoon

- Used geospatial mapping to link bus stops to specific Boston neighborhoods
- Incorporated census data to understand population characteristics, including demographics
- Created visualizations (bar charts) to highlight key neighborhood demographics and transportation disparities
- Created map visualizations and animations (GIF) to see trend in bus delays over months;

Jin Young Bang

- Created helper export and data-processing functions for bus delays to be used on other analysis (for Base Questions and Extension Project)
- Performed EDA on processed dataset
- Analyzed processed data to find bus route times; Explored and processed data for Holidays vs. Non-Holidays;
- Created and wrote majority of the reports/deliverables for the team

Minh Le

- Developed functions and classes for utility and quickly retrieve information related to stops, routes, etc. from the MBTA's v3 API
- Visualized data for base question 1 with charts to detect outliers and better understand the average end-to-end travel time for inbound and outbound busses
- Conducted EDA and analysis on weather-related data (temperature, precipitation, etc) to find correlations with bus delay times

Katherine Rimey

- Explore and identified associations between various OOP models on MBTA's v3 API
- Conducted ETL to gather and format data from the MBTA's v3 API into structured tables, facilitating the team's development process by enabling quick and efficient information retrieval
- Conducted EDA and analysis on weather-related data (temperature, precipitation, etc) to find correlations with bus delay times

Duc Minh Nguyen

- Processed the data to calculate delay times for each bus route.
- Analyzed the data to identify the most delayed routes, both inbound and outbound.
- Identified routes with the longest average delay times, indicating potential areas for improvement.
- Explore data for potential project's extension of differences in delay time based on timeframe (rush hour), and season (summer and winter months)

Reproducing Results and How to Use the Codebase

To ensure seamless usage of the codebase and reproduce the results successfully, it's essential to have the following libraries installed in your Python environment:

- Pandas
- Geopandas
- Seaborn
- Numpy
- MatplotLib
- Folium
- Pyproj

- Selenium
- Tqdm (optional)

All the necessary instructions to reproduce the results have been meticulously detailed in the 'main.ipynb' notebook. In case any data or datasets are missing, you can easily access and download them from the shared Google Drive link provided. To ensure a smooth replication of results, please maintain the exact directory names as outlined in the 'ipynb' file. Any deviation might lead to difficulties in reproducing the expected outcomes.