City of Boston: Bus Performance

Team F
Jialu Li (jli0318@bu.edu, Class of 2024)
Junyi Li (lijunyi@bu.edu, Class of 2024)
Qinfeng Li (ql2016@bu.edu Class of 2024 (Grad school))
Yifei Zhou (joeyifei@bu.edu, Class of 2024 (Grad school))
Laksanawist Mutiraj, (mutiraj@bu.edu, Class of 2026)

Introduction

The Massachusetts Bay Transportation Authority (MBTA) is not just a transit system, but a crucial lifeline for over a million daily commuters in the Boston area, significantly contributing to the region's economy with an estimated annual value of $11.5 billion. Yet, the quality of bus service and its performance varies across different neighborhoods, raising concerns about equitable access to transportation. This disparity has implications for economic opportunities, environmental sustainability, and social equity. To address this, there is a need for a comprehensive, data-driven analysis of MBTA's bus service performance trends, with a focus on geographic and demographic disparities. This project, in collaboration with BU Spark!, aims to uncover these trends, highlight potential inequities, and inform decision-making to enhance transit accessibility for all Boston residents.
In general, we collected data from various sources including MBTA open data portal, Boston Analyze website, National Weather Service, etc. to help construct a multi-view analysis. Besides doing data analysis using multiple data science tools, we decided to move forward and incorporate the usage of Machine Learning model from the lecture to create a prediction system for potentially benefiting the passengers of MBTA bus.
The project's big picture revolves around leveraging technology and data science to improve public transportation, ultimately striving to make Boston a more connected and equitable city. By analyzing and addressing the disparities in MBTA bus service, this initiative not only enhances daily commuting experiences but also contributes to broader goals of social justice, economic growth, and environmental sustainability in the region.

Base Analysis

In order to address base questions including:

1. What are the end-to-end travel times for different bus routes
2. Are there disparities in the service levels of different routes? (which lines are late more often than others)
3. What are the population sizes and characteristics of the communities serviced by different bus routes (e.g. race, ethnicity, age, people with disabilities/ vulnerabilities)?
4. If there are service level disparities, are there differences in the characteristics of the people most impacted?
5. Which neighborhoods are served better/worse by the MTBA bus system, which routes are better/worse, differences in quality of service by class/race

We mainly collected data from the MBTA Open Data Portal and Analyze Boston website. The datasets that we processed are Boston Neighborhoods Boundaries data from 2020, Bus Network Redesign Draft Bus Routes, Rapid Transit and Bus Prediction Accuracy, Commuter Rail Reliability, MBTA Bus Arrival Departure time, and Bus Ridership. For certain questions, we analyzed certain dataset solely. And for certain questions such as the third one, we used combinations of several datasets in order to present a specific result.

Through **EDA**, we got:

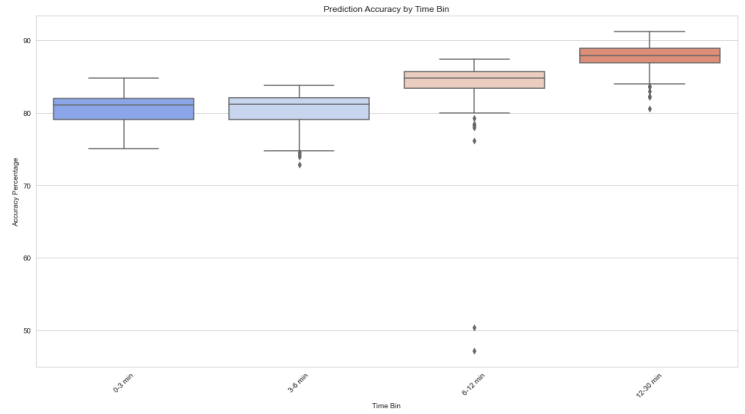

fig1. Prediction Accuracy by Time Bin

The median accuracy is consistently high across all time bins, hovering around the 80-90% range. The shorter duration predictions (e.g., "0-3 min" and "3-6 min") have slightly tighter interquartile ranges, indicating more consistent prediction accuracies. Longer duration predictions (e.g., "12-30 min") have wider interquartile ranges, suggesting more variability in their accuracies.

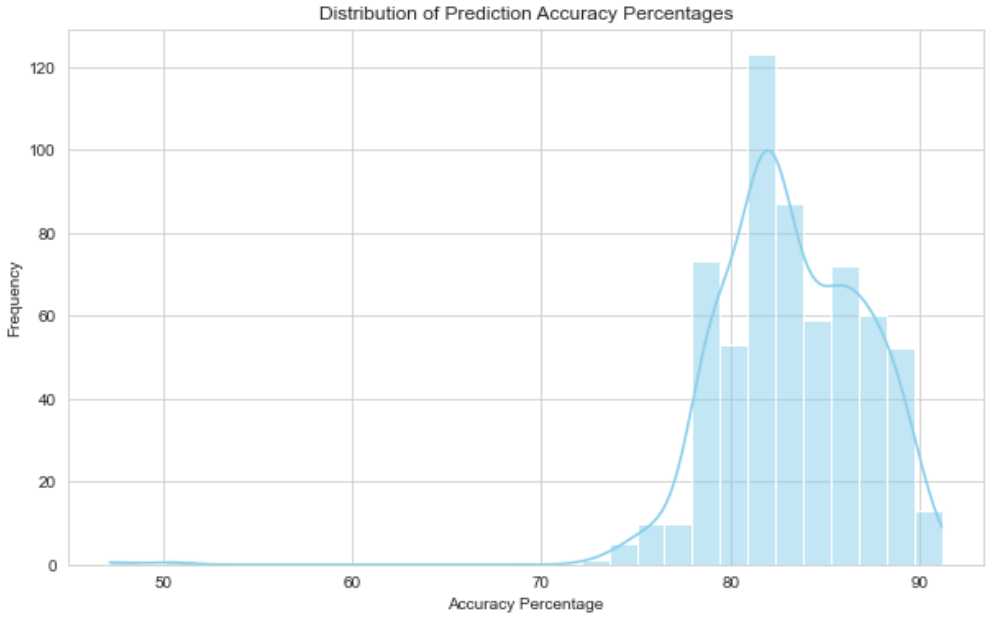

fig2. Distribution of Prediction Accuracy Percentages

From fig2, most of the predictions seem to be clustered around 75% to 90% accuracy, indicating that a significant portion of the bus predictions are fairly accurate.

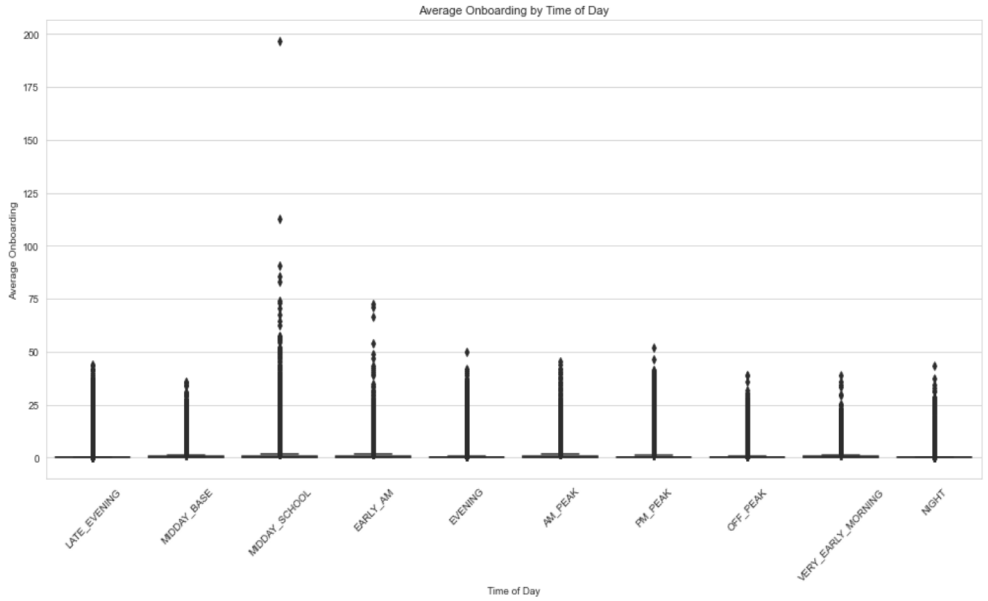

fig3. Average Onboarding by Time of Day

In fig3, the "AM_PEAK" time period shows the highest median onboarding, indicating that morning commutes have a high influx of passengers. The "MIDDAY_BASE" time period also has a relatively high median onboarding, suggesting steady ridership during midday hours. Time periods like "EARLY_AM" and "LATE_EVENING" have lower medians and narrower interquartile ranges, implying lesser and more consistent ridership during these hours.

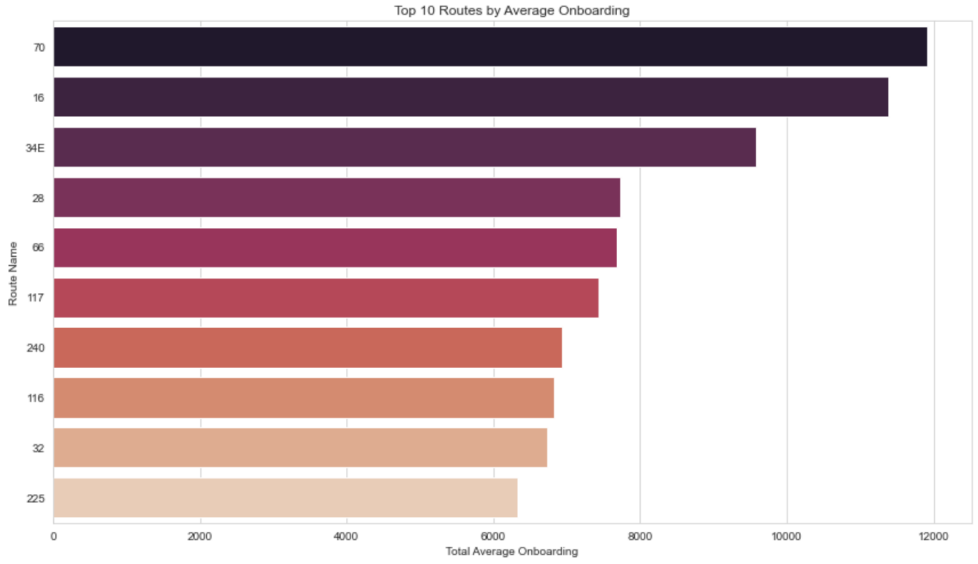

fig4. top 10 Routes by Average Onboarding

And fig 4 lists the top 10 routes grouped by average onboarding. As the graph shows, the bus route with highest average onboarding is route 70 which has nearly 12000 passengers onboarding each day and followed by route 16 and 34E.

Analyze Data

To address base question 1:

We process the data to filter out the transit type which are not bus.

```
# Filtering the data to consider only bus routes
df_bus_reliability = df_reliability[df_reliability['mode_type'] == 'Bus']
```

We processed the data of bus arrival and departure time to calculate the end-to-end travel time.
In order to get the end-to-end travel times for different bus routes represent the average duration it takes for a bus to travel from the starting point to the endpoint of a specific route. Follow the following procedure:
·Filtering the data to consider only bus routes
·Extracting the earliest and latest time points for each route and direction
·Merging the two dataframes to get both min and max times in one dataframe
·Calculating the end-to-end travel time for each route and direction
·Displaying the end-to-end travel times for different bus routes
We got:

	route_id	direction_id	travel_time_seconds	average_travel_time
0	01	Inbound	2186.178082191781	0 days 00:36:26.178082192
1	01	Outbound	2036.2826523777628	0 days 00:33:56.282652378
2	04	Inbound	1362.3664122137404	0 days 00:22:42.366412214
3	04	Outbound	1295.1330798479087	0 days 00:21:35.133079848
4	07	Inbound	983.1116687578419	0 days 00:16:23.111668758
5	07	Outbound	876.9397590361446	0 days 00:14:36.939759036
6	08	Inbound	2973.036211699164	0 days 00:49:33.036211699
7	08	Outbound	3279.767441860465	0 days 00:54:39.767441860

fig 5. CSV file contains the travel time for each bus route

This is the final output we got to answer question 1 and the details will be present later in the report.

To address base question 2:

weekly	mode	route_id	bin	arrival_departure	num_predictions	num_accurate_predictions	ObjectId	accuracy_percentage
2021-01-22 05:00:00+00:00	bus	Unknown	12-30 min	departure	1530075	1394979	400	91.17062889074064
2021-03-05 05:00:00+00:00	bus	Unknown	12-30 min	departure	1612256	1463544	224	90.7761546553401
2020-10-30 04:00:00+00:00	bus	Unknown	12-30 min	departure	1652403	1499831	352	90.76665922296195
2021-01-29 05:00:00+00:00	bus	Unknown	12-30 min	departure	1619321	1469203	204	90.72957122151816
2020-12-11 05:00:00+00:00	bus	Unknown	12-30 min	departure	1528106	1382793	376	90.49064659127049
2022-01-28 05:00:00+00:00	bus	Unknown	12-30 min	departure	1566549	1410309	296	90.02648496791355

gtfs_route_id	reliability_score
9703	32.00941915227626
449	40.25524468576182
448	40.630198757585696
459	42.997043635764754
747	45.480942004577706
195	49.19917090635013
19	49.205904165525475
41	49.24615399524695
70A	49.418184535977765
wad	51.08695652173913
14	51.1622605126052 86
8	51.44105297509801
701	51.50904158891645

fig 6&7 CSV files containing calculated accuracy/reliability score for each bus route

We calculated the accuracy percentage using the Rapid Transit and Bus Prediction Accuracy dataset. We also used another dataset of Commuter Rail Reliability to calculate the reliability score for specific bus routes. We generate these two files. Fig 6 ranks the accuracy score

descendingly and fig 7 ranks the reliability score ascendingly. They can straightforwardly help us answer the second base question.

To address base question 3:
It's about the different characteristics for different bus routes,
we combined the usage of 4 different data files.

```
# Load datasets
neighborhoods_path = "/Users/lijunyi/Downloads/Census2020_BG_Neighborhoods/Census2020_BG_Neighborhoods.shp"
bus_routes_path = "/Users/lijunyi/Downloads/cs506/Bus_Network_Redesign_Draft_Bus_Routes/Bus_Network_Redesign_Draf
neighborhood_data_path = "/Users/lijunyi/Downloads/Boston_Neighborhood_Boundaries_approximated_by_2020_Census_Blo
coordinates_map = pd.read_csv('/Users/lijunyi/Downloads/coordinates_map.txt', sep="\t", header=None, names=["Long
```

Then follow the procedure:
·Convert the CRS of bus_routes to match that of neighborhoods; Determine intersecting neighborhoods;
·Extract relevant columns from neighborhood_data for merging and merge;
·Determine the predominant racial group for each bus route;
·Merge color data back to bus_routes;
·Plot each racial group
Eventually got a plot:

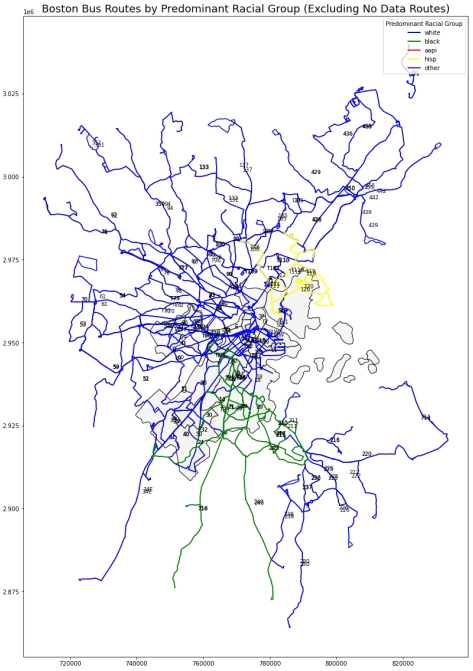

fig 8. Boston Bus Routes by Predominant Racial Group

This visualizes the predominant racial groups with respect to specific bus routes. For example, for bus routes that are mapped to green, the predominant race within the area is Black. Hence, we can easily capture the relationship between racial groups and bus routes through the graph.

To address base question 4:
To address the question about service level disparities and the characteristics of the most impacted people, we need to merge these datasets based on the bus route IDs and then analyze the relationship between the reliability scores and the demographic data. And then generate this heatmap based on the merged dataset:

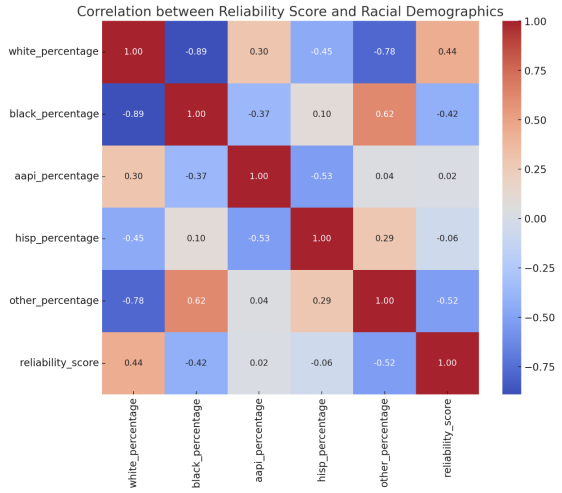

fig 9. Correlation Matrix of Reliability Score and Racial Demographics

In a correlation matrix, a value close to 1 indicates a strong positive correlation. A value close to -1 indicates a strong negative correlation. A value around 0 suggests little to no correlation. Thus, through this correlation matrix with value labeled out, we can answer the fourth question.

To address base question 5:

For this question, we used *Boston_Neighborhood_Boundaries_Approximated_by_2020_Census_Tracts*, which contains the Boston neighborhood boundaries, estimated by the 2020 census data, in GeoJson format. We also dived into *MBTA Bus, Commuter Rail, & Rapid Transit Reliability* dataset provided by MBTA. This dataset provides the historical reliability data under different metrics for the bus lines from 2015 to 2023. We only extracted the reliability data in 2023 for better reference of the current service quality of MBTA bus. For further study, we plan to compare the service between seasons and across the years.

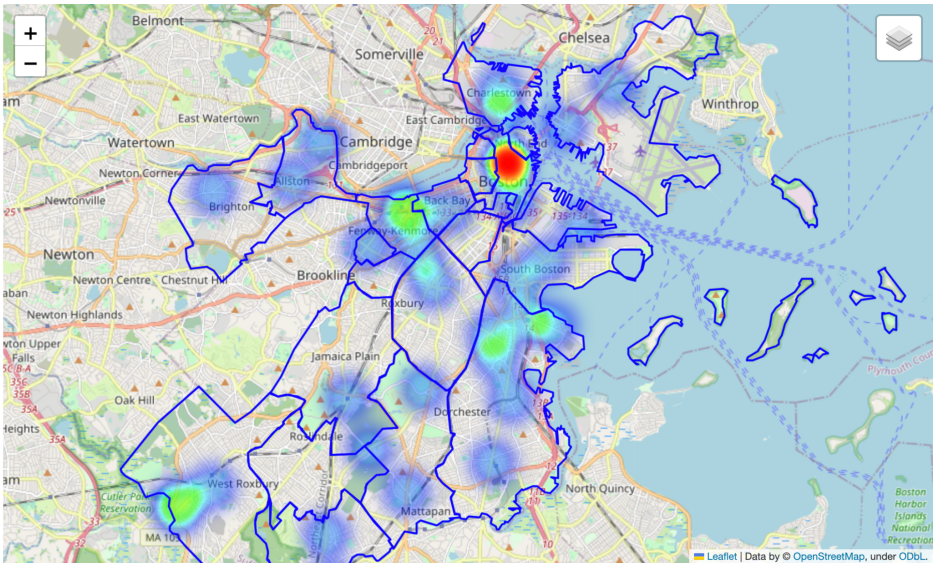

fig 10.Heatmap for Boston neighborhood according to service density

This map visualizes the service level of each neighborhood in the greater Boston area. The blue lines indicate the boundaries of different Boston neighborhoods. To better visualize the

service level across the neighborhoods, we used a heat map by the number of bus lines serving a specific area instead of marking the bus terminals on the map (what we did for deliverable 1). As we can see from the first sight, Downtown area has the most dense service level, followed by Allston, Jamaica Plain, Mattapan, and Dorchester.

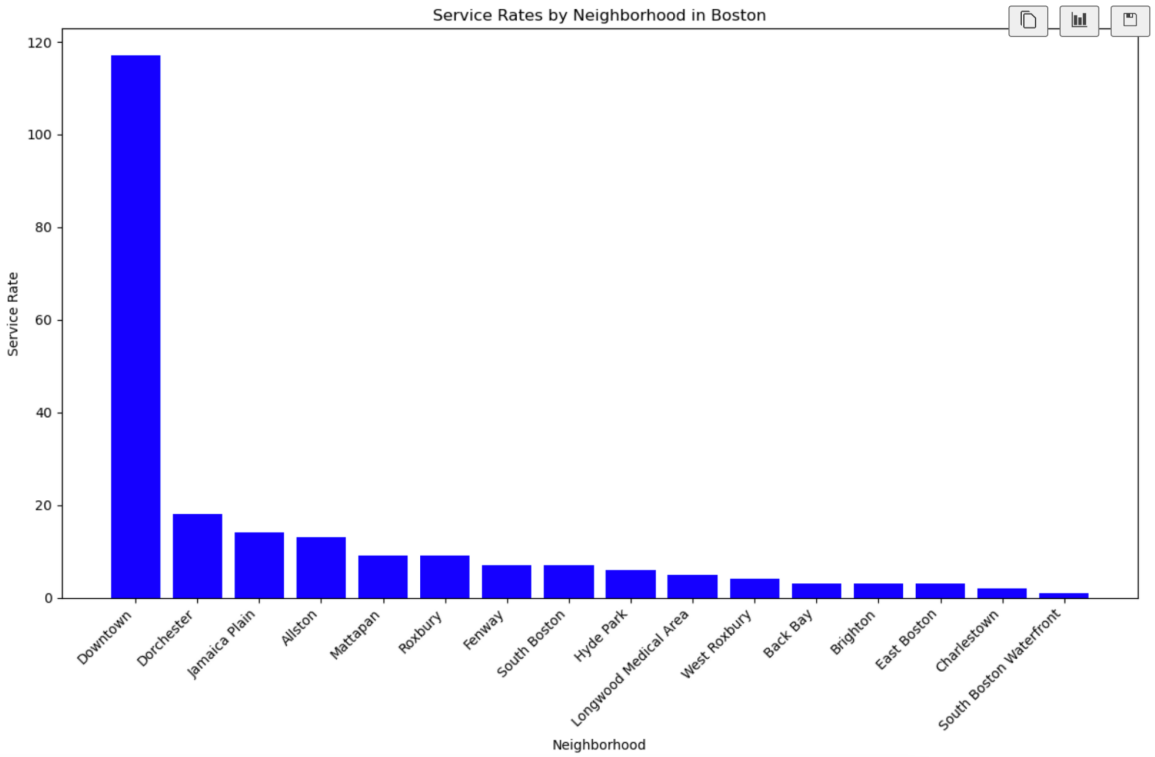

fig 11. Bar plot for service rates across neighborhoods

This figure shows the level of service (number of bus lines) serving different Boston neighborhoods. Downtown, Dorchester, Jamaica Plain, Allston, and Mattapan are the top 5 neighborhood with the highest service level.

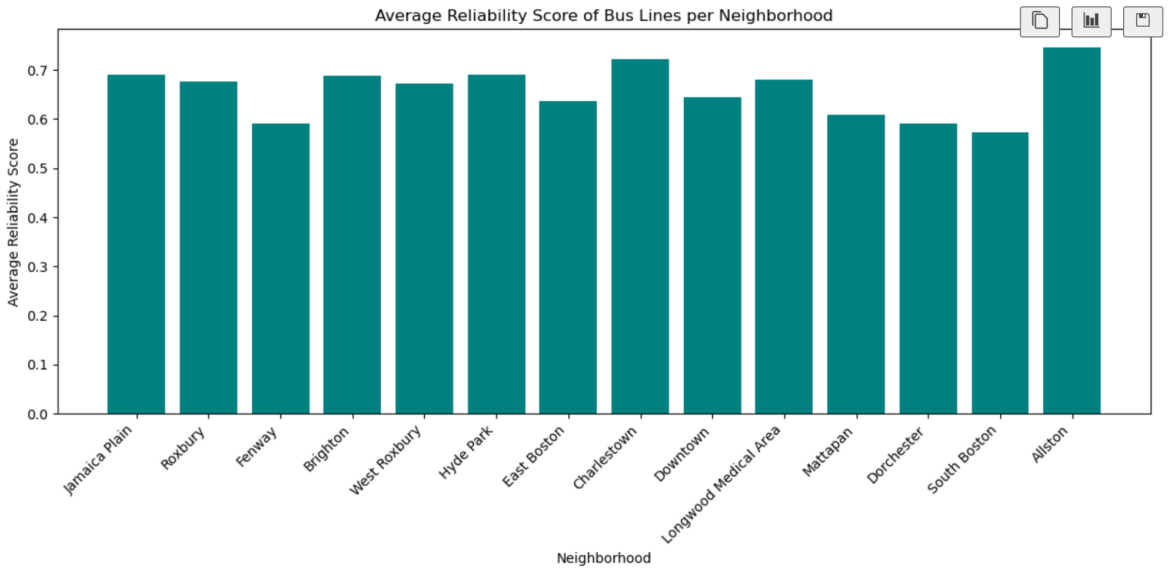

fig 12. Average Reliability Score of Bus Line per Neighborhood

The bar plot gives us insight into the service reliability of different neighborhoods. Allston has the best reliability score of over 0.7, with a top 5 service rate. For other well-serviced areas like Dorchester and Mattapan, their reliability is significantly lower than other areas. Therefore, we can conclude that there are service disparities between different neighborhoods with different service levels.

	gtfs_route_id_modified	metric_type	aggregated_otp_numerator	aggregated_otp_denominator	final_result
0	10	Headway / Schedule Adherence	79832.0	144490.0	0.552509
1	100	Headway / Schedule Adherence	76673.0	94523.0	0.811157
2	105	Headway / Schedule Adherence	23640.0	34222.0	0.690784
3	106	Headway / Schedule Adherence	74699.0	102093.0	0.731676
4	108	Headway / Schedule Adherence	64667.0	93414.0	0.692262

fig 13. Service reliability for individual bus route

This figure shows the average reliability scores of bus lines for each neighborhood. The metric type of the reliability score is “Headway/Schedule Adherence”, which is the on-time rate across the origin station, midway checkpoint station, and end station. According to MBTA, if the bus arrives within 3 min of the scheduled departure time, the bus is considered on time. We processed the *MBTA Bus, Commuter Rail, & Rapid Transit Reliability* dataset and grouped each bus line’s reliability score with the same metric type. The image shown below is the aggregated data.

Results:

- For question 1:

The average_trip_durations.csv file in our repo and *fig 5* provides specific end-to-end average travel times for bus routes in seconds and a more readable duration format. For instance, route 1 inbound has an average travel time of 36 minutes and 26 seconds, while route 1 outbound averages at 33 minutes and 56 seconds. Similarly, route 4 inbound averages at 22 minutes and 42 seconds, and outbound at 21 minutes and 35 seconds. Most of them have end-to-end travel time within 40 min.

- For question 2:

The conclusion is that there exist great disparities in the service levels between different bus routes, *fig6 & 7*

The disparity in service levels between routes is pronounced when comparing the extremes:

- Route 9703 has the lowest reliability score at 32.01, which is significantly lower than the highest reliability score of 92.59 for route CR-Shuttle003.
- Route 449 with a score of 40.26 and route 448 with 40.63 are far below route CR-Shuttle002 and CR-Shuttle001, both of which have a score of 85.82.
- Even between the fifth lowest and fifth highest, route 747 scores 45.48 in contrast to route 73, which scores 81.98.

There are more details in the csv file in our repo.

- For question 3:

The bus_routes_race_data.csv file contains demographic data related to bus routes, including total population and the number of people from different racial and ethnic groups, along with their respective percentages of the total population for each bus route.
For example, the bus route labeled as "10" services a community with a total population of 539,832 people. Among them, 210,242 identify as White (38.95%), 145,108 as Black (26.88%), 66,182 as Asian American and Pacific Islander (AAPI) (12.26%), 107,284 as Hispanic (19.87%), and 11,016 as Other (2.04%).
Some routes, like route "100", have zero population recorded for all racial groups, which may indicate missing data or an error in the dataset.
This data can be used to characterize the populations serviced by different bus routes, highlighting the diversity and potential vulnerabilities within communities. It is essential to analyze this data further to understand the impact of bus service levels on these diverse groups.

- For question 4:
Here is the conclusion for differences in the characteristics of the people most impacted, *fig 9*

-White Percentage: There is a positive correlation (0.44) between the percentage of white individuals in an area and the reliability score. This suggests that areas with a higher percentage of white individuals may experience better bus service reliability.

-Black Percentage: There is a negative correlation (-0.42) between the percentage of Black individuals in an area and the reliability score. This indicates that areas with a higher percentage of Black individuals may be more likely to experience less reliable bus service.

-AAPI Percentage: The correlation here is negligible (0.02), implying that the percentage of AAPI individuals in an area does not significantly correlate with bus service reliability.

-Hispanic Percentage: The correlation is slightly negative (-0.06), but close to zero, suggesting that the percentage of Hispanic individuals in an area does not have a strong correlation with the reliability of bus services.

-Other Percentage: There is a negative correlation (-0.52) between the percentage of individuals of other races in an area and the reliability score. This may suggest that areas with a higher percentage of individuals from other races might experience lower reliability in bus services.

- For question 5:
Our conclusion Boston's Downtown, Allston, Dorchester, Jamaica Plain, Mattapan, and Roxbury are the neighborhoods served best by the MBTA bus, *fig 10,11,12,13*. However, service level disparities exist within these five areas. Allston has the best reliability score, whereas Dorchester and Mattapan experience the most delay compared to others. Now we cannot conclude which racial group is most impacted by these desparities. General census data could be inaccurate while depicting the demographic group experiencing the service level disparities since the generalized population data is different from the actual ridership data (many of the neighborhoods are dominated by whites, but it doesn't mean that white people ride the buses more). We have put the analysis of racial data in the extension project by analyzing the average income of these areas and demographic data of frequent riders.

Extension Analysis

We propose to extend our current geospatial analysis of Boston's transit systems by integrating additional datasets that consider factors like service disruptions and accessibility features. This effort aims to provide a more inclusive story of the city's transit dynamics, particularly focusing on the distribution and development of resources for disabled accessibility around bus stops. We also propose to delve deeper into the intersection of public transit data, specifically focusing on Blue Bike and bus data in Boston. The goal is to uncover insights into how different modes of transit interplay and affect urban mobility. In addition to analysis, we decided to move further by doing a practical move, which is a delay prediction system based on route ID and temperature inputs. With the use of this system, the passengers could gain real benefits.

The reason for doing this extension to the base project is that we aim to extend the geospatial analysis of Boston's transit systems, integrating data on service disruptions and accessibility, especially focusing on disabled access at bus stops. This approach seeks to provide a comprehensive view of the city's transit dynamics and improve urban mobility by analyzing how different modes, like Blue Bikes and buses, interact. Additionally, we plan to develop a practical delay prediction system based on route ID and temperature, offering direct benefits to passengers by enhancing their commuting experience.

Questions for Analysis

- What area do bus stops and blue bike stations cover, and how could that make the blue bike a possible alternate for bus?
- What are the key user patterns observed in the Bluebikes bike-sharing system in Boston?
- What is the distribution of income situation within each neighborhood the bus routes cover?
- What are the accessibility gaps in the current transit network, and how might they affect riders with disabilities?
- How to build a delay minutes prediction system based on temperature and route id?

Datasets & Sources

We will use the following datasets: Blue Bike data: Details of bike usage, station locations, and trip information; Bus data: Bus routes, stop locations, and ridership statistics; Boston income data: To analyze the socio-economic factors affecting transit use; Boston Weather data: Contains daily temperature of Boston from 2013-2023.

Data Visualizations

- Heatmaps showing the density of Blue Bike stations and bus stops.
- Correlation plots between bike and bus usage.
- Socio-economic overlays on transit maps to identify service gaps.
- Charts correlating service disruptions with changes in Blue Bike usage.
- Accessibility heat maps indicating potential areas for infrastructure improvement.
- GeoJSON map with a combination of bus and blue bike stops

Visualization for Extention Question 1: What area do bus stops and blue bike stations cover, and how could that make the blue bike a possible alternative for the bus?

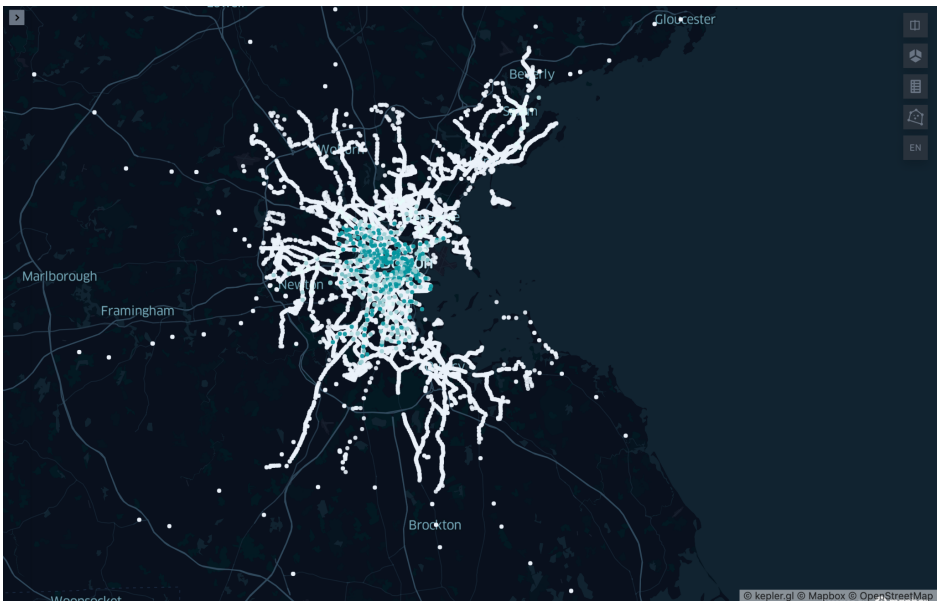

fig.14. Visualization of Blue Bikes Stations and Bus Stops

The map visualizing bus stops and Blue Bike stations in Boston was produced by first examining and parsing the geospatial data of bus stops from a CSV file into GeoJSON format. The blue dots represent the Blue Bikes stations and the white dots represent MBTA bus stops. This data was then merged with an existing GeoJSON file containing the locations of Blue Bike stations. The combined dataset was saved as an updated GeoJSON file, which was likely visualized using a mapping platform such as Kepler. gl, as seen in the watermark of the provided map.

It displays a dense network of bus stops, and Blue Bike stations concentrated in the central city areas, signifying robust multimodal transit options likely catering to higher population densities and commercial activities. As one moves towards the outskirts, a notable thinning of this network indicates potential transit service gaps. The proximity of bike stations to bus stops in the center suggests good integration for efficient transfers, but this integration appears to diminish outwardly. This spatial data could inform transit authorities about potential areas for infrastructure expansion, aid in urban planning for better service coverage, and support environmental goals by encouraging reduced car usage through accessible public transit options.

Visualization for Extension Question 2: What are the key user patterns observed in the Bluebikes bike-sharing system in Boston?

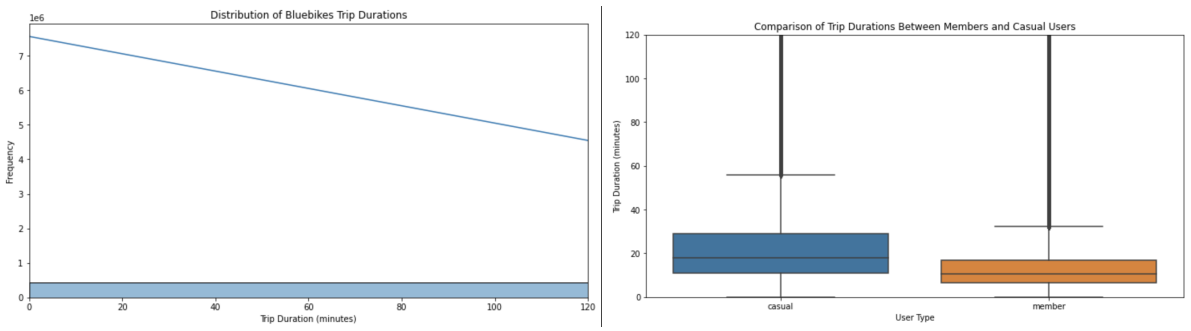

fig.15, 16 Distributions and Comparisons of Blue Bikes Trip Durations

Our analysis of the Bluebikes data reveals distinct user patterns in the bike-sharing system. The majority of the trips are concise, predominantly under 30 minutes, signifying that users primarily opt for Bluebikes for quick, efficient travel within the city. This trend is a testament to the convenience and accessibility of the bike-sharing system for short urban commutes.

Furthermore, when we delve into the differences between members and casual users, a striking pattern emerges. Casual users generally engage in longer trips than members, suggesting a varied utilization of the service. Members appear to use the service for routine, shorter commutes, whereas casual users might be leveraging the system for leisurely rides or exploratory journeys around the city. This distinction in usage patterns underscores the flexibility and appeal of Bluebikes to a diverse range of users, catering to both regular commuters and occasional riders seeking to navigate Boston in an eco-friendly and healthy way.

Visualization for Extension Question 3: What are the accessibility gaps in the current transit network, and how might they affect riders with disabilities?

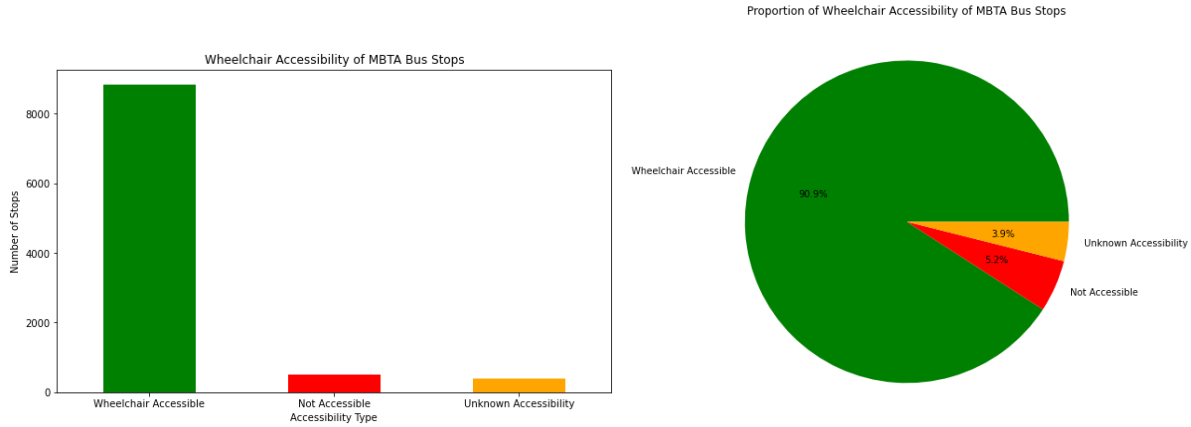

fig.17, 18. Wheelchair Accessibility Analysis

A significant majority of stops are wheelchair accessible, as indicated by the green segments in both charts. A small fraction of stops are not accessible (red), and a relatively minor portion has an unknown status (orange).

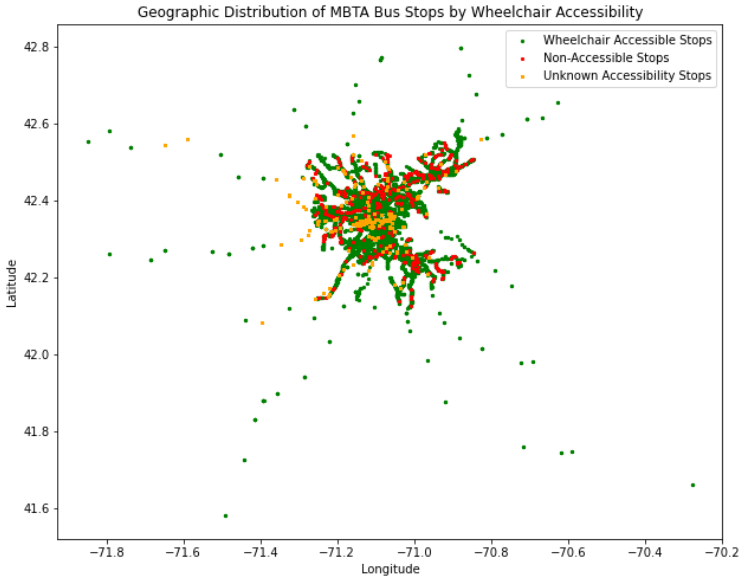

fig.19. Wheelchair Accessible Stops on Map

This visualization provides a clear geographic perspective on the accessibility of the MBTA bus network, emphasizing the focus on making the majority of stops wheelchair accessible.

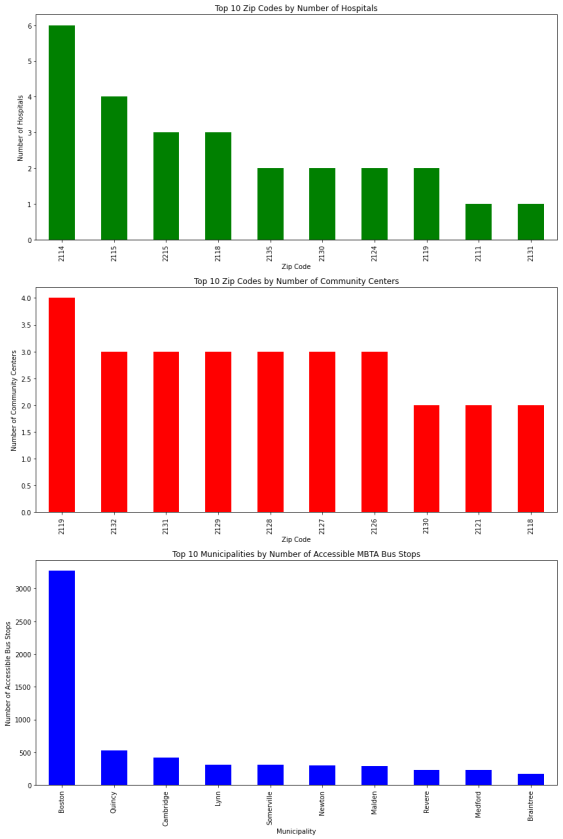

fig.20. Concentration of Hospitals& Availability of Community Centers

The visualizations depict a comprehensive overview of wheelchair accessibility at MBTA bus stops. The bar chart and pie chart clearly show that most bus stops are wheelchair accessible, indicating that the MBTA has made considerable efforts to accommodate passengers with disabilities. Specifically, the pie chart highlights that 90.9% of the bus stops are wheelchair accessible, with only a small fraction being inaccessible or of unknown accessibility status.

The geographic distribution map further provides insights into the spatial spread of accessible and non-accessible bus stops across the Boston area. It appears that wheelchair-accessible stops are widely distributed, suggesting that people with disabilities have substantial coverage and can access the bus system in numerous locations. However, the non-accessible and unknown accessibility stops, though few, might indicate potential areas for improvement where the MBTA can focus its efforts to ensure complete inclusivity.

The bar charts showing the number of hospitals and community centers by zip code suggest an additional layer of analysis. They imply that certain areas, especially those with a higher number of hospitals and community centers, may have a greater demand for accessible transportation due to a higher likelihood of mobility-impaired individuals visiting these facilities. Therefore, ensuring that bus stops in these zip codes are fully accessible is crucial for providing equitable access to healthcare and community services.

Lastly, the bar chart displaying the top 10 municipalities by the number of accessible MBTA bus stops indicates that Boston has the highest number of accessible stops. This is encouraging as it suggests that the city's core, likely to have the highest demand for public transportation, is largely accommodating to passengers with disabilities. However, there is room for improvement in other municipalities to ensure equitable access throughout the Greater Boston area.

Overall, these visualizations show a positive trend towards accessibility but also underscore the importance of continued efforts to improve and ensure that the MBTA bus system is truly inclusive for all riders, regardless of physical ability.

Visualization for Extension Question 4: What is the distribution of income situation within each neighborhood the bus routes cover?

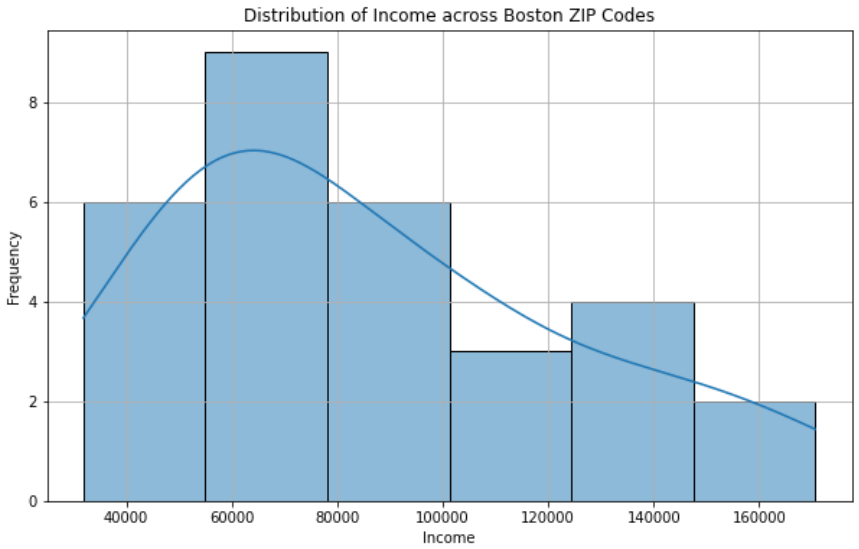

fig.21. Distribution of Income Across Boston Zipcodes

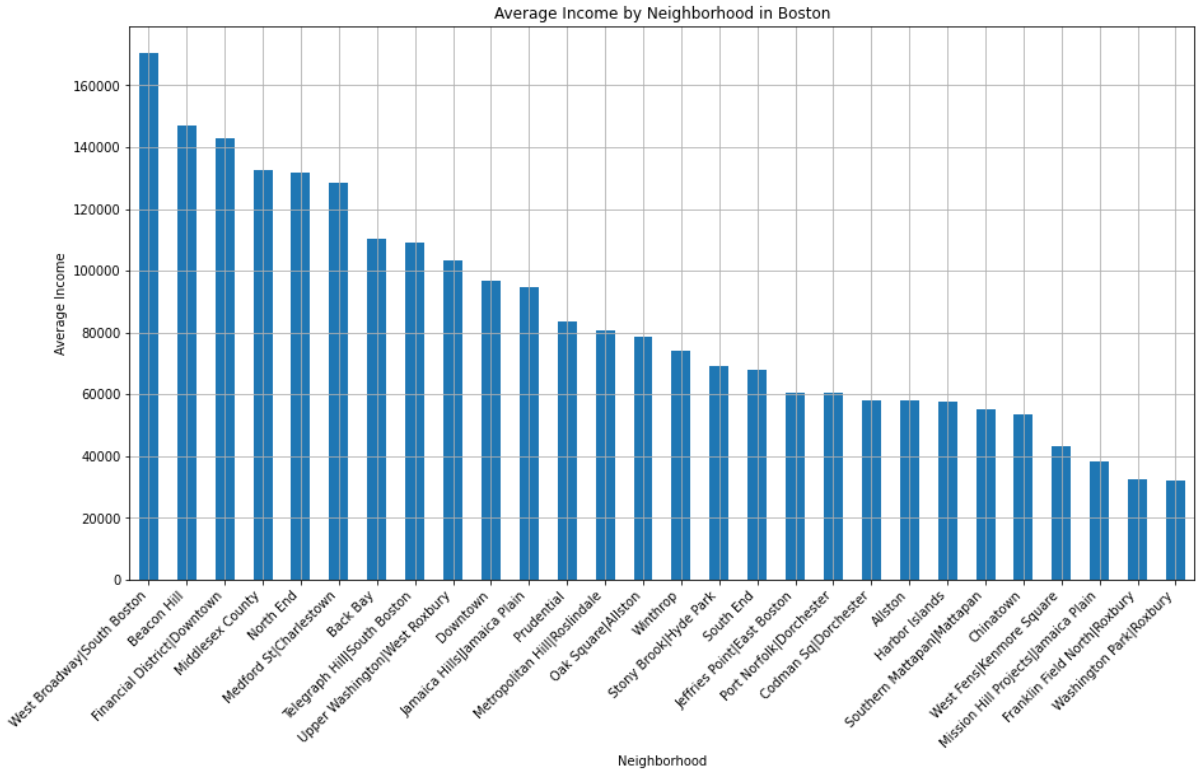

fig.22. Income Analysis on Boston Neighborhoods

The histograms and summary statistics depict a right-skewed income distribution across Boston neighborhoods, with an average income of approximately $85,316 and a wide range from $31,900 to $170,588. This suggests significant economic diversity, with a concentration of wealth in neighborhoods like West Broadway/South Boston and Beacon Hill, contrasted by lower-income areas such as Chinatown and Washington Park/Roxbury. The observed

income disparity highlights the varying economic challenges and affluence across the city, which could inform policy decisions related to urban development, social services, and resource allocation to address the needs of different communities.

Our analysis of the Boston neighborhood income and the reliability of the bus service is shown below. We can see that Dorchester, East Boston, and Mattapan have low-reliability scores and low incomes, whereas some of the higher-income areas like Charlestown have the best reliability scores. The map shows the correlation of reliability scores and income level for each neighborhood.

	Neighborhood	Income	Average Reliability Score
0	Allston	64748.0	0.746070
1	Charlestown	128403.0	0.722110
2	Dorchester	59209.0	0.591057
3	Downtown	119716.0	0.644411
4	East Boston	60579.0	0.636689
5	Hyde Park	69262.0	0.690540
6	Jamaica Plain	66516.5	0.690790
7	Mattapan	55024.0	0.608524
8	Roxbury	32200.0	0.676490
9	South Boston	139868.5	0.572159
10	West Roxbury	103337.0	0.671909

fig.23. Relationship Between Neighborhood Income and Reliability

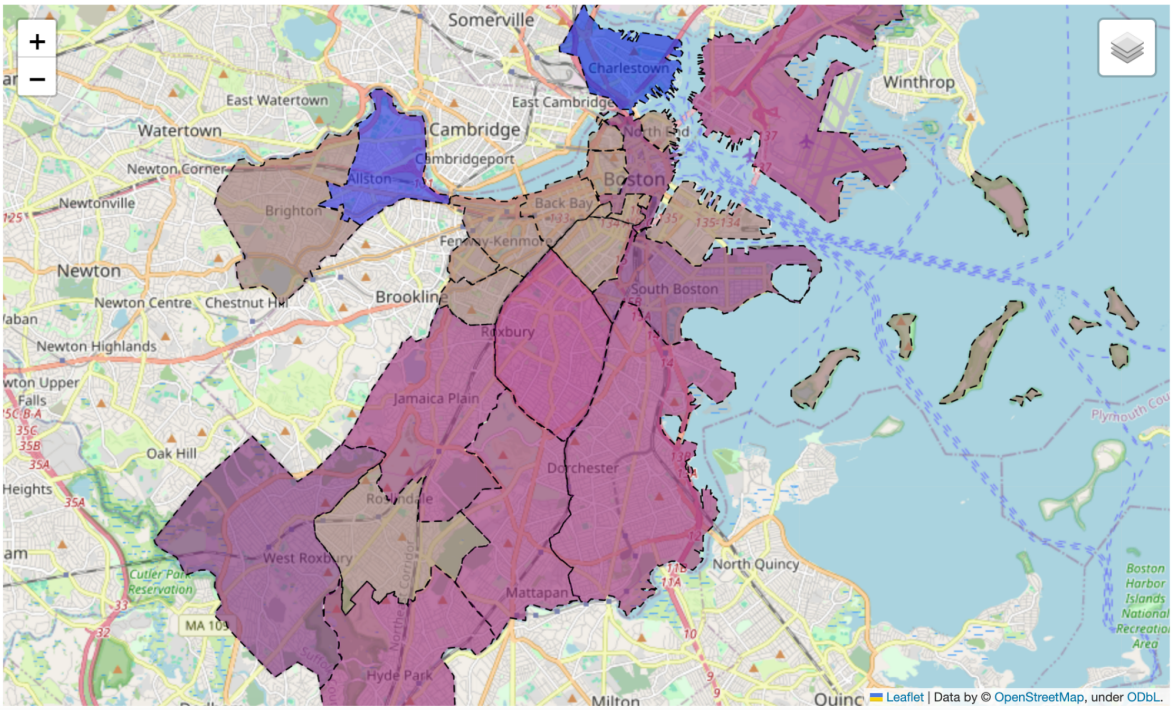

fig.24. Map Visualizations of Neighborhood Income and Service Reliability

Visualization for Extension Question 5: How to build a delay minutes prediction system based on temperature and route id?

Our 'Delay Prediction System' utilizes advanced data processing and machine learning techniques to forecast potential delays in MBTA bus services. We meticulously merged

comprehensive datasets from the MBTA open data portal and weather information, focusing on the intricate relationship between route-specific factors and temperature variations. By employing a Linear Regression model, trained on one-hot encoded route IDs and temperature data, we achieved a nuanced understanding of delay patterns. The system's capability is encapsulated in the predict_delay function, which predicts the delay time for any given bus route under variable temperature conditions. This predictive model is not only a technical feat but also a crucial tool for enhancing the reliability and efficiency of public transport in Boston

Here is the example usage:

```
# Example usage:
route_id_input = '57'  # example route_id as string
temperature_input = -10  # example temperature in Celsius
```

with the above inputs from the user, our prediction will give the following prediction outputs:

```
The predicted delay for route 57 at -10°C is approximately 5.46 minutes.
```

Results

1. “What area does bus stops and blue bike stations covered, and how could that make blue bike a possible alternate for bus?”

The bus stops and Blue Bike stations are concentrated in the central city areas of Boston, indicating a robust multimodal transit network that caters to high population densities and commercial activities. As one moves towards the outskirts, the network becomes sparser, suggesting potential gaps in transit service. The close proximity of bike stations to bus stops in central areas suggests good integration, making Blue Bike a viable alternative for efficient transfers, especially where bus frequency is lower or in case of service disruptions

2. “What are the key user patterns observed in the Bluebikes bike-sharing system in Boston?”

Bluebikes user patterns indicate that most trips are under 30 minutes, implying that the system is used for quick and efficient travel within the city. The service caters to two main user groups: members, who use it for routine, shorter commutes, and casual users, who tend to engage in longer trips, possibly for leisure or exploration

3. “What is the distribution of income situation within each neighborhood the bus routes cover?”

The income distribution across Boston neighborhoods is right-skewed, with significant economic diversity. There is a concentration of wealth in areas like West Broadway/South Boston and Beacon Hill, while lower-income areas include Chinatown and Washington Park/Roxbury. This disparity highlights the economic challenges and affluence across the city and could inform resource allocation to address community needs

To further address the minor limitation we encountered during answering base question 5, now this economic and service-level disparity necessitates a deeper investigation into the demographics of actual bus ridership, which may differ from general census data. While some areas predominantly consist of white residents, it doesn't inherently imply that they are the primary users of the bus system. By incorporating a detailed analysis of average income and the demographics of frequent riders, a more accurate picture of who is most affected by transit disparities can emerge. Such targeted analysis is crucial for developing equitable transit policies and ensuring that service improvements reach those most in need, promoting inclusivity across Boston's diverse neighborhoods

4. “What are the accessibility gaps in the current transit network, and how might they affect riders with disabilities?”

Most MBTA bus stops are wheelchair accessible, with 90.9% of stops accommodating passengers with disabilities. However, there are still some stops that are not accessible or have unknown accessibility status. These gaps suggest areas where the MBTA can improve to ensure complete inclusivity. Ensuring that bus stops near hospitals and community centers are accessible is crucial for providing equitable access to healthcare and community services, particularly for mobility-impaired individuals

5. “How to build a delay minutes prediction system based on temperature and route id?”

The delay prediction system for MBTA bus services is built using advanced data processing and machine learning techniques. The model employs a Linear Regression approach, trained on one-hot encoded route IDs and temperature data, to forecast potential delays. The predictive function, predict_delay, can then estimate the delay time for any bus route given the temperature conditions, enhancing the reliability and efficiency of public transport in Boston

Conclusion

This comprehensive study on the Massachusetts Bay Transportation Authority (MBTA) bus service in Boston reveals significant disparities in service levels across different routes and neighborhoods. Our data-driven analysis, utilizing sources like MBTA's open data portal and Boston Analyze website, highlighted variations in prediction accuracy, onboarding trends, and service reliability, correlating these with demographic disparities.

Key findings include:

- Prediction Accuracy: Most bus predictions show a high accuracy rate, with more variability in longer duration predictions.
- Onboarding Patterns: Peak periods like morning commutes show the highest onboarding, with notable differences across routes.
- Service Disparities: There is a pronounced disparity in service reliability among different bus routes, with some neighborhoods receiving better service than others.
- Demographic Correlation: Our analysis also suggests a correlation between service reliability and racial demographics, with areas having a higher percentage of white individuals experiencing better service reliability.

Suggestions for Improvement

1. Enhanced Data Analysis: Further analysis focusing on the demographics of actual bus ridership can offer deeper insights into service disparities.

2. Service Level Optimization: Prioritize improving service reliability in neighborhoods with historically lower service levels, especially those with higher percentages of minority groups.

3. Accessibility Focus: Ensure all bus stops, particularly near essential services like hospitals and community centers, are wheelchair accessible.

4. Technology Integration: Develop predictive systems to enhance service efficiency, such as delay prediction models based on route ID and environmental factors.

5. Policy Initiatives: Address economic and social disparities by aligning transit policies with the needs of diverse communities, ensuring equitable access to public transportation.

By addressing these suggestions, the MBTA can enhance its service quality, promote social equity, and ensure that the bus system serves as a reliable and accessible transit option for all Boston residents.

Personal Contribution

Junyi Li:
Focus on solving base questions 1,2,3. Forming extension proposal and doing analysis for extension questions related to blue bike as an alternate, disabling access, and building a prediction system. Collaborated in drafting the report.

Jialu Li:
Answered base question 4,5. Geographical data visualization/representation and service quality analysis with historical data. Processed demographic and ridership data to provide detailed insights into the impact on different neighborhoods. Collaborated in drafting the report.

Qinfeng Li, Yifei Zhou, Laksanawisit Mutiraj: Data collection and presentation slides.