Checkpoint A Report

**Task Division**

Data Collection: Yifei Zhou, Qinfeng Li, Laksanawisit Mutiraj

Exploratory Data Analysis: Jialu Li, Junyi Li

**Problem Statement:**

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.

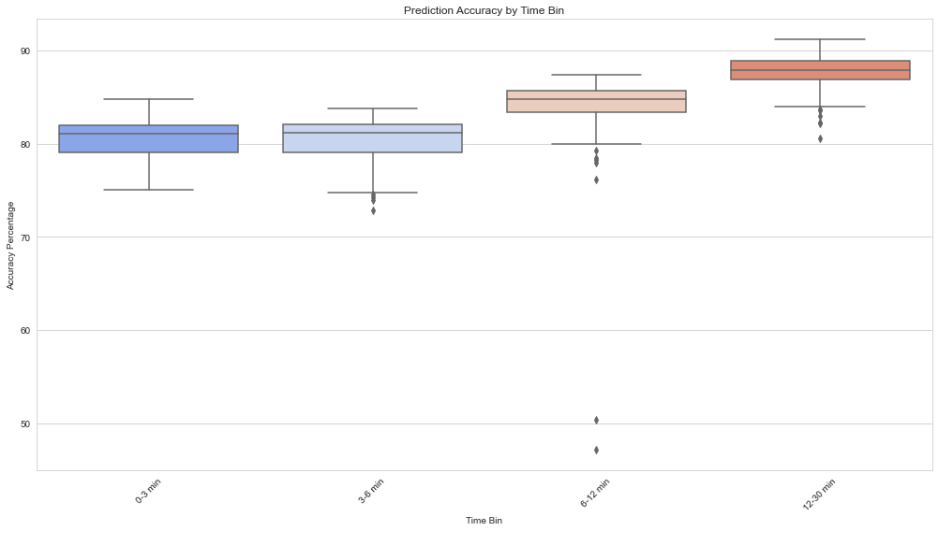
**Data Collection and Preprocessing**

In order to address 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. This can include questions about traffic information, which neighborhoods are served better/worse by the MTBA bus system, which routes are better/worse, differences in quality of service by class/race, contributing variables, ect.

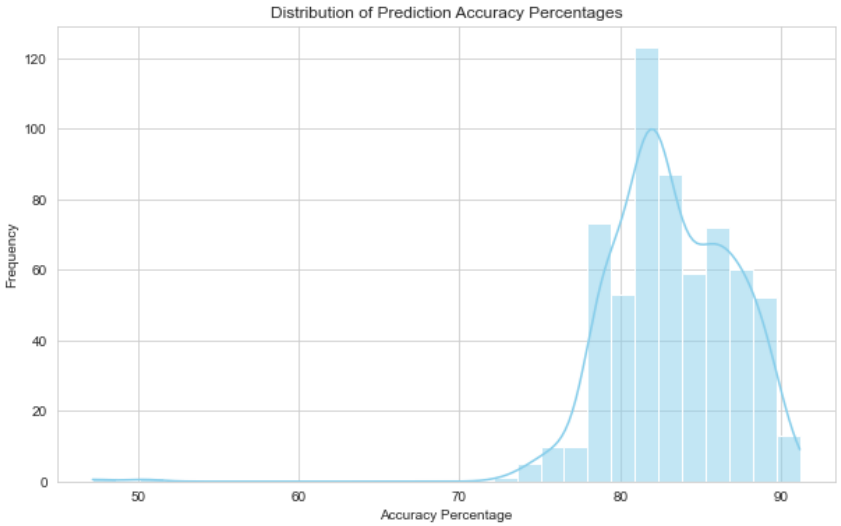
We mainly collected data from MBTA Open Data Portal and Analyze Boston website. The datasets that we processed are Boston Neighborhoods Boundaries data from 2020, Bus Network Redesign Draft Bus Routes, Rapid Transit and Bus Prediction Accuracy, Commuter Rail Reliability, MBTA Bus Arrival Departure time, and Bus Ridership. For certain question, 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:

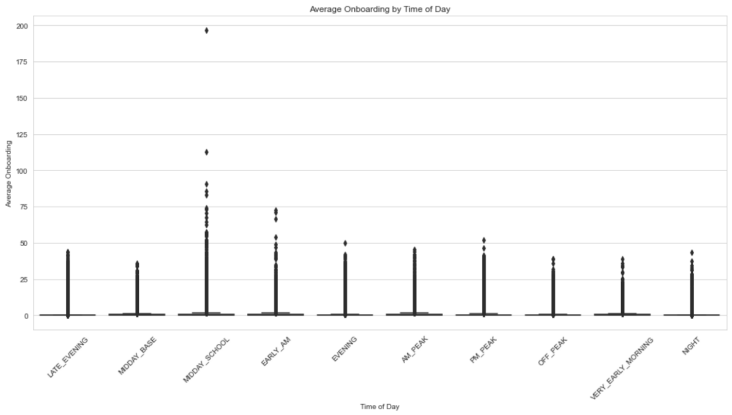



The median accuracy is consistently high across all time bins, hovering around the 80-90% range. The shorter duration predictions (e.g., "0-3 min" and "3-6 min") have slightly tighter interquartile ranges,

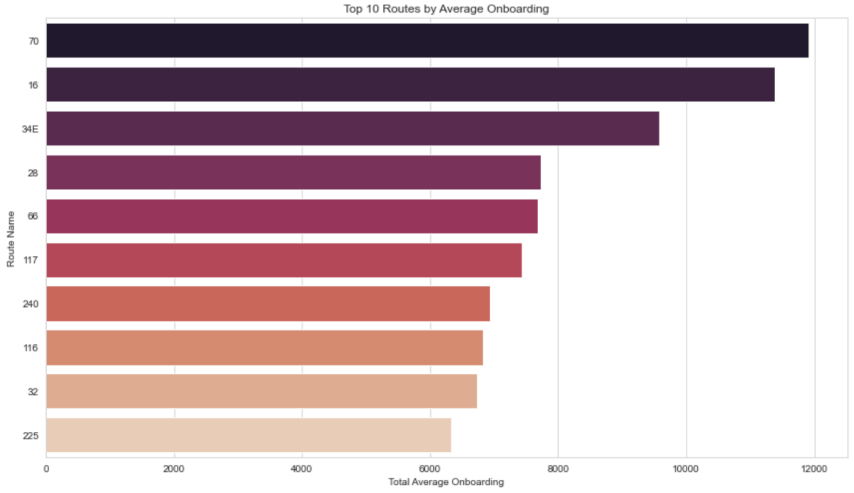
indicating more consistent prediction accuracies. Longer duration predictions (e.g., "12-30 min") have wider interquartile ranges, suggesting more variability in their accuracies.




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.




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.



And here lists the top 10 routes grouped by average onboarding.

# 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 |

## 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 |

Then we calculated the accuracy percentage using the Rapid Transit and Bus Prediction Accuracy dataset.

| | gtfs_route_id | reliability_score |
|---|---|---|
| 1 | 9703 | 32.00941915227626 |
| 2 | 449 | 40.25524468576182 |
| 3 | 448 | 40.630198757585696 |
| 4 | 459 | 42.997043635764754 |
| 5 | 747 | 45.480942004577706 |
| 6 | 195 | 49.19917090635013 |
| 7 | 19 | 49.205904165525475 |
| 8 | 41 | 49.24615399524695 |
| 9 | 70A | 49.418184535977765 |
| 10 | wad | 51.08695652173913 |
| 11 | 14 | 51.162260512605286 |
| 12 | 8 | 51.44105297509801 |
| 13 | 701 | 51.50904158891645 |

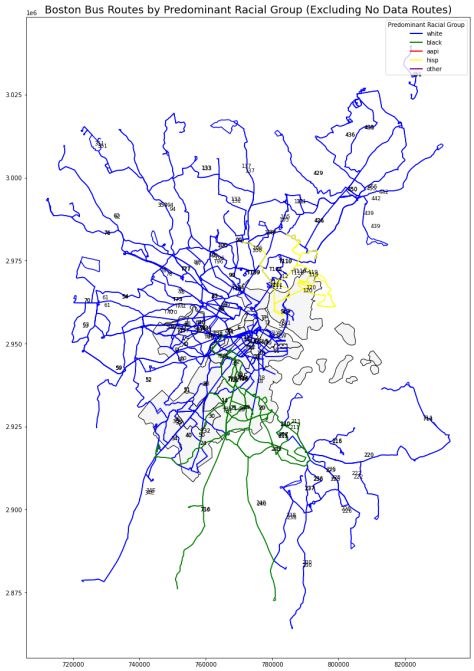
We also used another dataset of Commuter Rail Reliability to calculate the reliability score for specific bus routes.

## 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
```

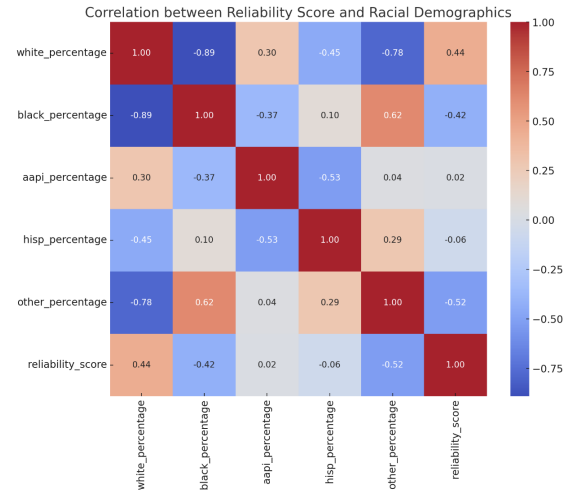
Then follow the procedure:
Convert the CRS of bus_routes to match that of neighborhoods
Determine intersecting neighborhoods
Extract relevant columns from neighborhood_data for merging and merge
Determine the predominant racial group for each bus route
Merge color data back to bus_routes
Plot each racial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




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

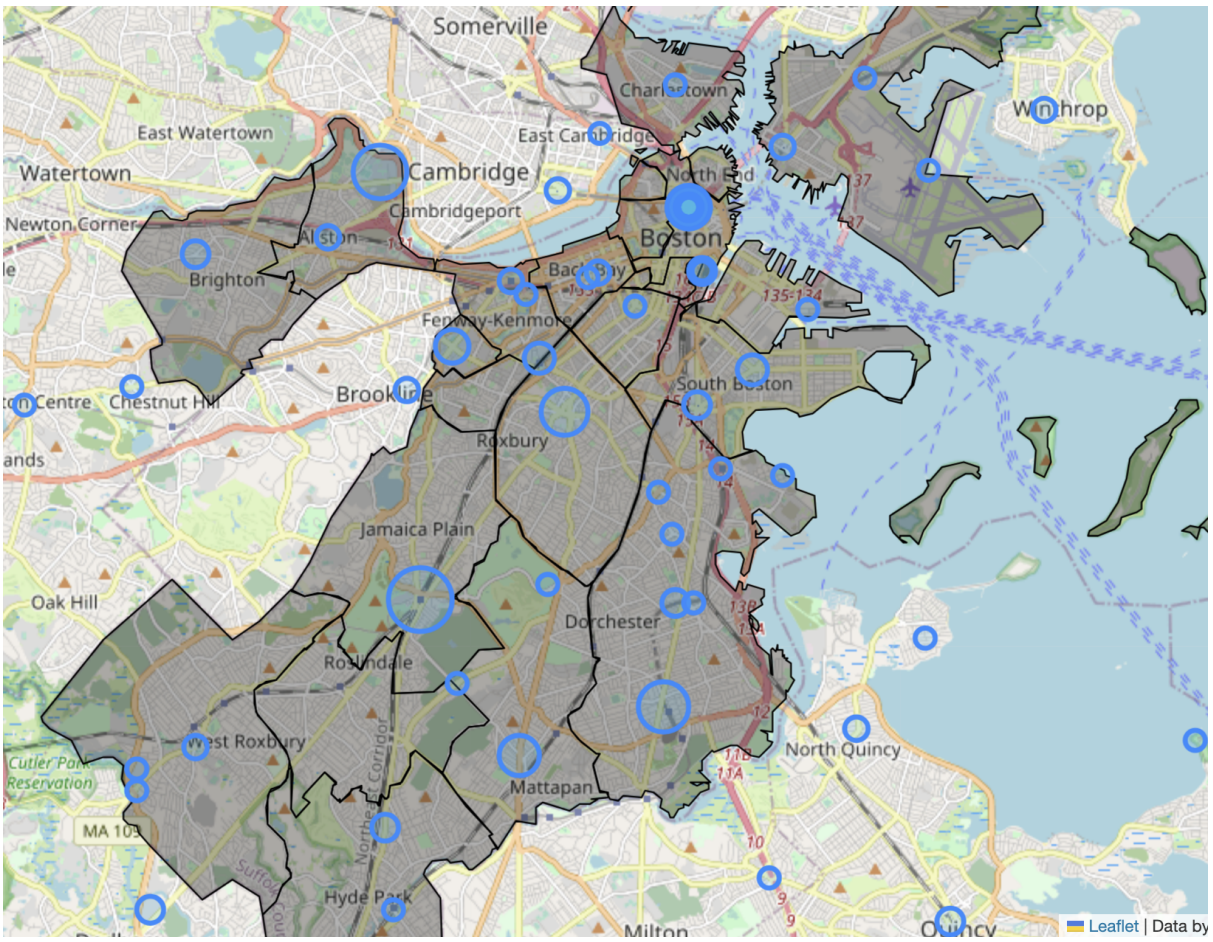
## To address base question 4:

To address your 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:




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.

## To address base question 5:




This map visualizes the service level of each neighborhood in the greater Boston area. The blue circles designate the bus terminals that service multiple bus lines. The size of the circle indicates the number of bus lines served by a specific terminal. As we can see, The bigger terminal resides in Boston's downtown area, Allston, Dorchester, Jamaica Plain, Mattapan, and Roxbury. These areas are the most populated neighborhoods in the greater Boston area.

## Answer to Base Questions:

For question 1:

The average_trip_durations.csv file in our repo 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

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

-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:
We can partially answer this question with the generated map of the bus terminals and their service levels. Boston’s downtown, Allston, Dorchester, Jamaica Plain, Mattapan, and Roxbury are the neighborhoods served best by the MBTA bus. However, further research has to be done on the service level disparities in these neighborhoods compared to others as well as the ridership demographics. 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).

## Challenges & Limitations:

1. The layer of neighborhood boundaries is only limited to the Boston Area
2. Lack of detailed census data for now
3. The biggest challenge is matching data from the different dataset; e.g some bus id of the certain dataset is missing but it contains other useful information
4. Also, the exact locations for smaller bus stops (non-terminal) are lacking. For now, we're only able to use the data fetched from Google Maps API, which, for some of the bus terminals, is not accurate.

## Assumptions:

Given the outlined challenges and limitations, we must assume that our analysis is constrained by the geographic scope of the Boston Area and may not account for nuanced socio-demographic dynamics outside this region. The lack of detailed census data necessitates a cautious interpretation of demographic impacts, acknowledging the potential for missing or underrepresented variables that could influence service level disparities. The difficulty in matching data from different datasets is likely to introduce gaps in our understanding of service reliability across all bus routes; thus, we proceed with the available data while recognizing the incomplete picture it may present. The absence of exact locations for smaller bus stops suggests that our analysis may better reflect service reliability at larger terminals rather than along entire routes.

## Next Steps:

The next steps would include seeking more comprehensive datasets, enhancing the precision of stop locations, and potentially using advanced data matching techniques to bridge the information gaps between different sources. We will search for more accurate and broader neighborhood boundaries and bus stop location data to refine our model. Modify the Data Preprocessing Section to better handle missing data, to match the pattern across different datasets, and eliminate the shortcoming of missing bus id.

## Project Completion Plan:

To address the challenges and limitations identified in the preliminary stages of our analysis, our project completion plan involves several strategic steps. We will intensify efforts to acquire more comprehensive datasets, including detailed census data and more accurate neighborhood boundary delineations. We recognize the need to refine the precision of bus stop locations and will look to utilize alternative sources beyond the Google Maps API to rectify inaccuracies. Find detailed census data within each neighborhood and ridership data. Pair it with the bus-stop coverage/reliability data to see which demographic group is more impacted. Add alternative mobility data around the bus stops/lines with the least reliability score. Additionally, we plan to develop advanced data matching techniques to reconcile discrepancies between different datasets, especially where bus IDs are missing or inconsistent. This will improve the robustness of our data and the validity of our findings. Our data preprocessing protocols will be adjusted to mitigate the issues stemming from the lack of granular data at smaller bus stop levels, aiming to create a more representative and accurate depiction of service reliability across the Boston Area.

We are trying to complete all the steps mentioned above in 1-2 weeks.

## Extension Proposal:

In the proposed extension, we aim to delve deeper into the relationship between public transit reliability and socioeconomic factors within the Boston Area. By integrating additional data sets, such as income levels, employment statistics, and educational attainment, alongside our current demographic and service reliability figures, we believe we can uncover more nuanced insights into how and why certain communities are underserved by public transit. The rationale for this extension is twofold: firstly, it will allow us to understand the broader social implications of transit disparities and secondly, it aligns with our team's commitment to promoting equitable transportation policy. We hypothesize that areas with lower income and educational levels will correlate with poorer service reliability, reflecting broader social inequities. To substantiate our analysis, we will seek out datasets from local government databases and educational institutions. The visualizations will include layered heatmaps and scatter plots, with axes representing socioeconomic indicators against reliability metrics. This additional information will enrich our understanding of the transit ecosystem and provide valuable insights for stakeholders aiming to address service disparities.