CITY OF BOSTON: TRANSIT AND PERFORMANCE
TEAM C

FINAL REPORT

Team Members

1. Chandrahas Aroori, Class of 2025 - Team Representative
   (Email: charoori@bu.edu, GitHub: Exorust)
2. Manushi Munshi, Class of 2025
   (Email: manushi@bu.edu, GitHub: manushimunshi14)
3. Munir Siddiqui, Class of 2024
   (Email: munirsid@bu.edu, GitHub: munirsidd)
4. Patrick Browne, Class of 2025
   (Email: pbrowne2@bu.edu, GitHub: pbrowne011)
5. Haocheng Liu, Class of 2025
   (Email: easonlhc@bu.edu, GitHub: HaochengL)

1. Introduction

Public transportation plays a vital role in cities like Boston by providing mobility and access to jobs, services, healthcare, education, and amenities. However, historically not all communities have been served equally by public transit, often due to inequitable urban planning and infrastructure investments. This report aims to analyze the performance of MBTA bus routes across Boston in relation to the socioeconomic demographics of the communities they serve. The goal is to identify any disparities in service levels or quality and assess whether they may disproportionately impact marginalized groups.

Specifically, this analysis will focus on examining MBTA bus operational data from January 2022, alongside Census demographic data aggregated at the neighborhood level in Boston. Key questions to be explored include: What are the end-to-end travel times for different bus routes across the city? Are there noticeable disparities in on-time performance and reliability between routes? What are the racial, economic, and age demographics of the communities primarily served by each bus route? If clear service disparities emerge from the data, do they appear to adversely and disproportionately impact disadvantaged communities?

To answer these questions, the report will utilize data science techniques including data collection APIs, cleaning and joining of datasets, exploratory data analysis, statistical summaries, data visualization, and geographic mapping. Findings will provide critical insights into potential service inequities along demographic lines. Recommendations will also be made on how to focus future transit investments to improve equity in service quality and accessibility for marginalized groups.

Public transit equity has broad implications for social justice, economic mobility, and environmental sustainability in diverse cities like Boston. This data-driven analysis aims to provide an objective, comprehensive understanding of where service gaps exist across bus routes, who they impact, and how they can be addressed. The goal is to promote discussion and progress towards an affordable, reliable transportation network that connects and serves all Boston communities.

2. Data Collection and Cleaning

The first step in the data cleaning process was to join and filter the MBTA data to focus only on January 2022. This involved accessing the MBTA API to get real-time location, route, and schedule data for all bus trips during that month. The raw data was then filtered and aggregated to create a consolidated dataset for analysis.

Next, key metrics like route travel times and on-time percentages were calculated. The scheduled and actual departure/arrival times in the MBTA data were used to compute total travel times for each route and trip. These travel times were averaged to get the mean time for each route. The scheduled versus actual times were also compared to calculate on-time performance.

When answering future questions, we will use spatial joins to associate the bus routes and stops with demographic data from the 2020 Census. We will first identify the census tracts that each stop is located

in, and then we will link the tract-level population statistics to the stops and routes. This will allow us to append useful demographic data like income levels and race/ethnicity percentages to the MBTA data.

Our MBTA dataset was checked for quality and missing values were handled prior to analysis.

3. Exploratory Data Analysis

We worked with two datasets, the MBTA & Census dataset. Here are the results of the EDA for both of the datasets.

MBTA Dataset

Fig. 1 the proportions of measurements that are classified as either starting point, endpoints, or neither (midpoints). From this, we can see that over 28% of our collected data comes from either a starting or ending location. This possibly varies depending on the route (i.e., shorter routes may have less midpoints); however, it is possible that we have the same number of midpoints (5) for each route, with 1 start and 1 end point.

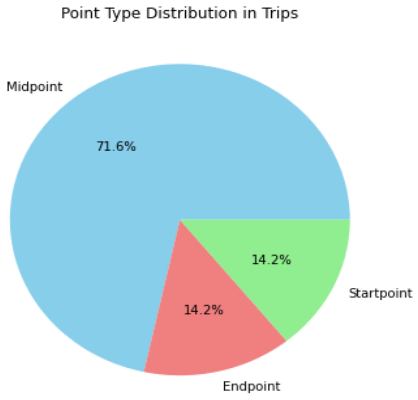

Fig. 1: Point distribution in MBTA dataset

Fig. 2 shows the total number of bus routes that were run each day on the MBTA for the month of January. We can see a clear pattern where more bus routes (~90,000) are run during the week than on the weekends, and that Saturday (~60,000) has more bus routes than Sunday (~40,000). We can also see that there is a holiday schedule, as Martin Luther King, Jr. Day had roughly the bus volume of an average Saturday.

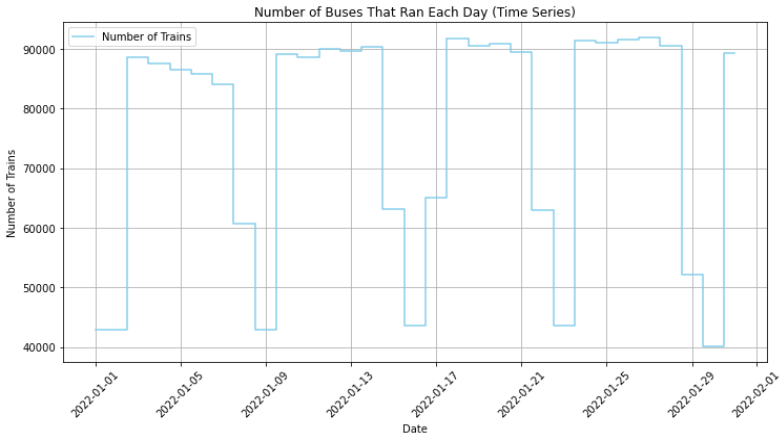

Fig. 2: Daily count of buses that ran in Jan 2022

In Fig. 3 we can see the top 10 most frequent and least frequent lines by count of routes run. We see that the "111" bus route is by far the most frequent, with a roughly 66% increase in total number of routes compared to the "28" route (the second-most frequent). This makes sense; upon looking at the MBTA website, we see that this route has scheduled departures every 5 minutes. This may be because there are less buses running in this area. We observe that the 192, 193, and 194 routes are listed as least frequent, only being run 350-400 times total. Upon searching the MBTA website, these routes are not listed, indicating that they may have been removed for their lack of frequency (and passengers). The "171" route is a special low-service route to Logan Airport.

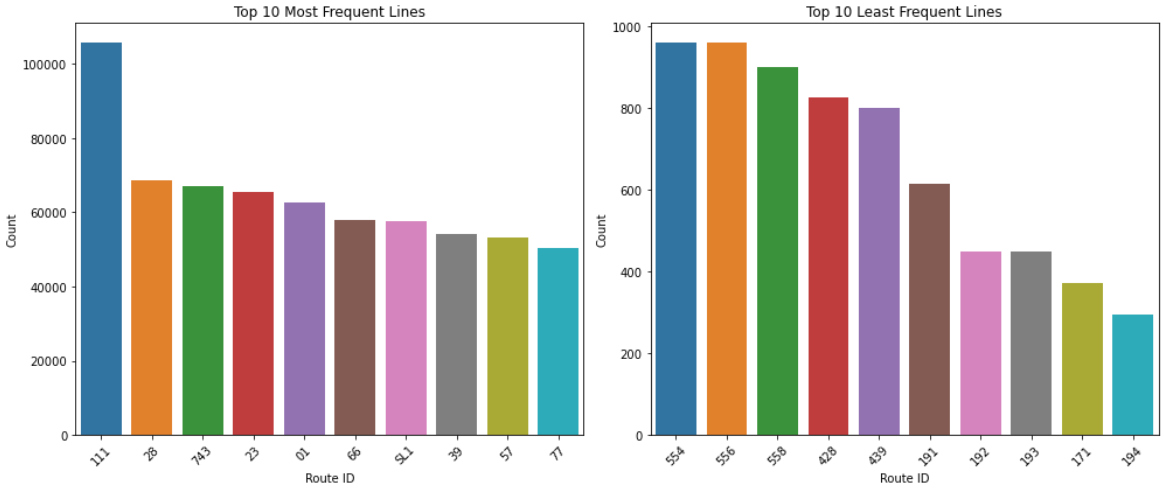

Fig. 3: Top 10 most and least frequent bus lines

Fig. 4 plots "maximum average headway" for the top 10 bus routes. Headway is defined as the "distance or duration between vehicles in a transit system measured in space or time." Thus, we see which routes have the most time between buses when run in this graph. It appears that three routes (the 117, 116, and 746) have far more time between them than all the other routes plotted.

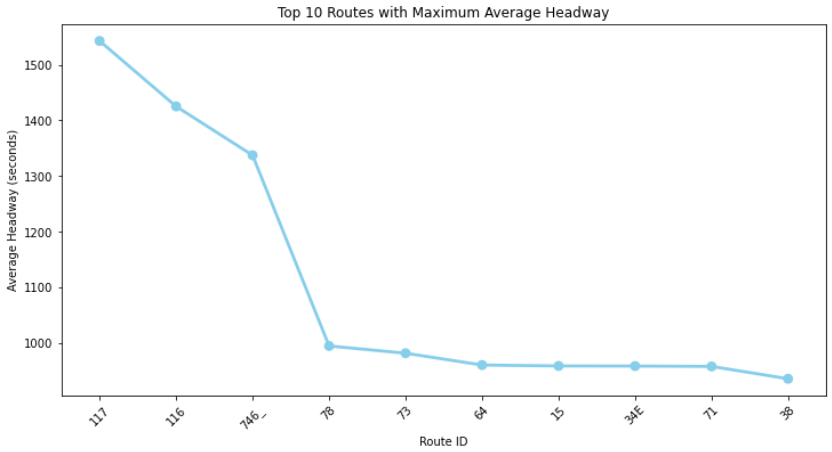

Fig. 4: Top 10 routes having maximum average headway

Census Dataset

Fig. 5 is the population distribution in the 23 neighborhoods of the city of Boston. The darker portions represent a high density of population as compared to the lighter portions which represent lower population density. We can observe that the neighborhoods of Dorchester and Longwood have the highest and lowest population respectively, in the city of Boston.

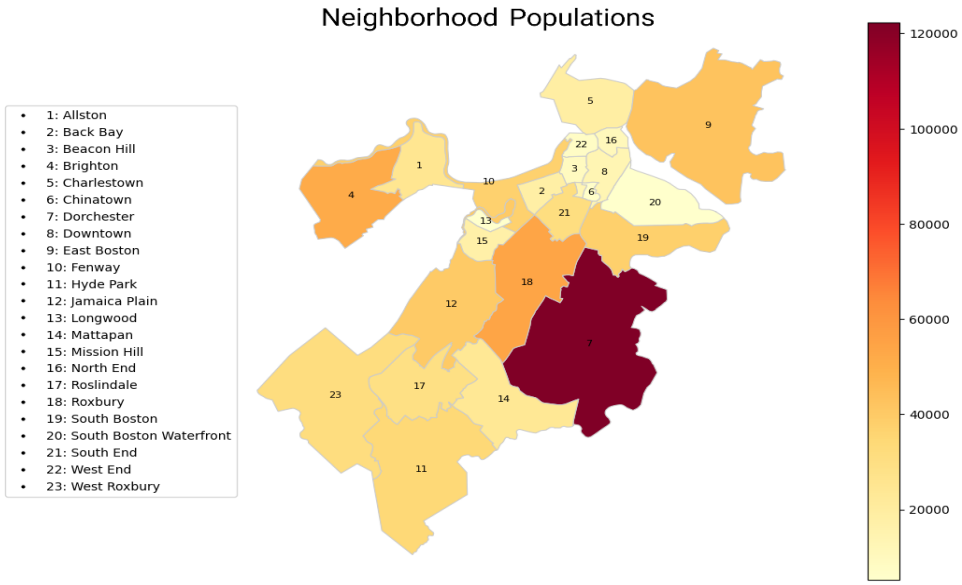

Fig. 5: Population distribution of Boston by neighborhood

The pie charts in Fig. 6 represent the ethnic and age distribution of the most populated neighborhood in Boston - Dorchester. We can infer from the pie chart that over 50% of the population is Black, White, or African American whereas the other 50% is composed of Hispanic, Latino, Asian or other races. Further, over 78% of the population is over the age of 18 in Dorchester.

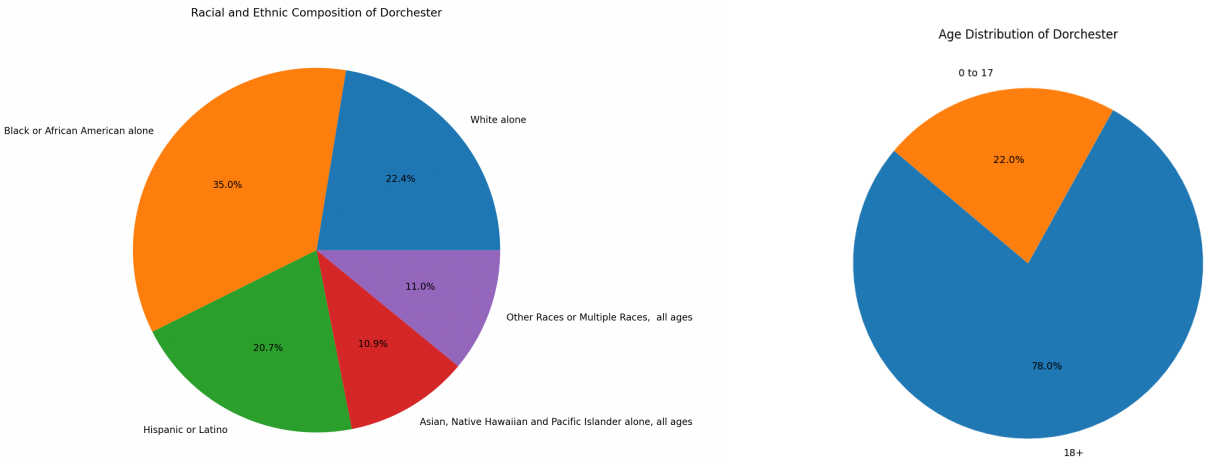

Fig. 6: Racial and age distribution of the most populated neighborhood of Boston

The pie charts below show the ethnic and age distribution of the least populated neighborhood of Longwood. The first pie chart shows that over 50% of the population is white with other races having less population percentage in Longwood. The second pie chart shows that only 0.8% of the people living in Longwood are under the age of 18.

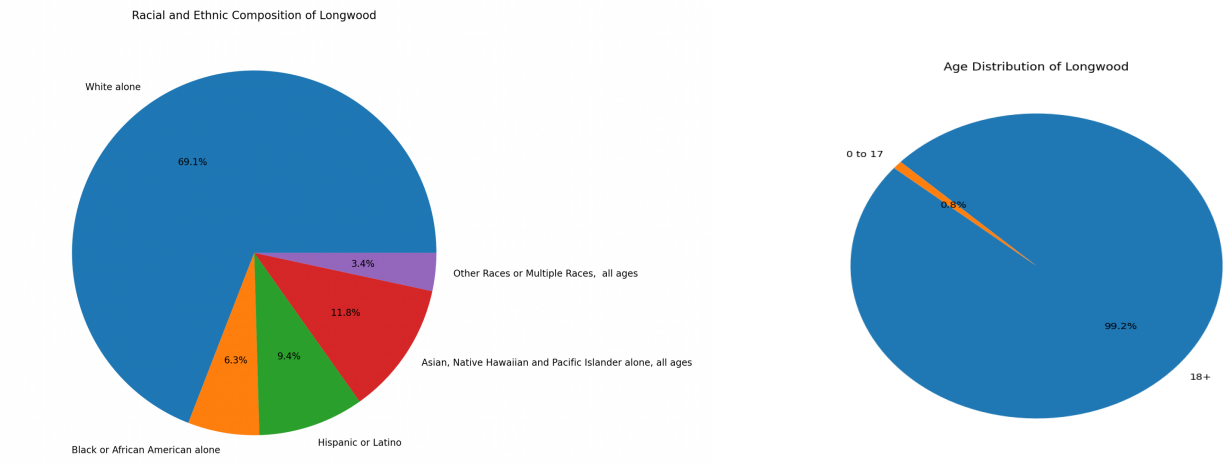

Fig. 7: Racial and age distribution of the least populated neighborhood of Boston

4. Visualizations for Addressing Base Questions

The end to end travel times and average travel times for the bus routes were calculated from the MBTA bus data. Two important factors for service disparities were defined - lateness and punctuality. *Lateness* is measured as the difference between actual total travel time and scheduled total trip time. *Punctuality* here is defined as the difference between actual departure time and scheduled departure time from the start point.

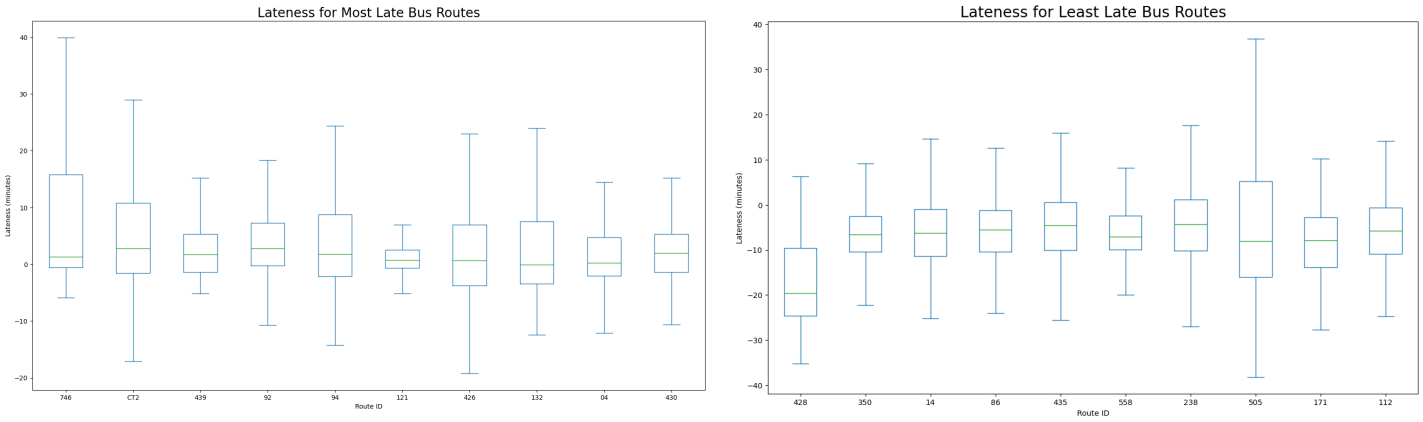

Fig. 8: Lateness for top 10 most and least late bus routes

In Fig. 8, we can see the boxplots of the lateness for the top 10 most and least late bus routes. An interesting thing to note is that the most late bus routes have median lateness close to zero, and apart from a couple, they are roughly symmetrical with the routes sometimes taking less time than scheduled and sometimes taking longer. The maximum lateness goes up to 40 minutes for route 746. On the other hand, the least late routes have negative median lateness. This means that the routes, on average, took less time than scheduled. Again, they are roughly symmetrical with the routes sometimes taking less time than scheduled and sometimes taking longer. The minimum lateness is around -40 minutes for route 505. This route is an interesting one as it has a very large spread. The maximum lateness for this route is around 40 minutes and the minimum is around -40 minutes.

In Fig. 9, we can see the boxplots of the delay for the top 10 least and most punctual routes. The least punctual bus routes have a very large spread with the absolute delay ranging from around 10 to 30 minutes. On the other hand, the most punctual bus routes have a significantly lower spread, with the absolute delay ranging from around 2 to 5 minutes. This means that the most punctual routes were more consistent and departed within 5 minutes of their scheduled time, whereas the least punctual routes had a much higher variability.

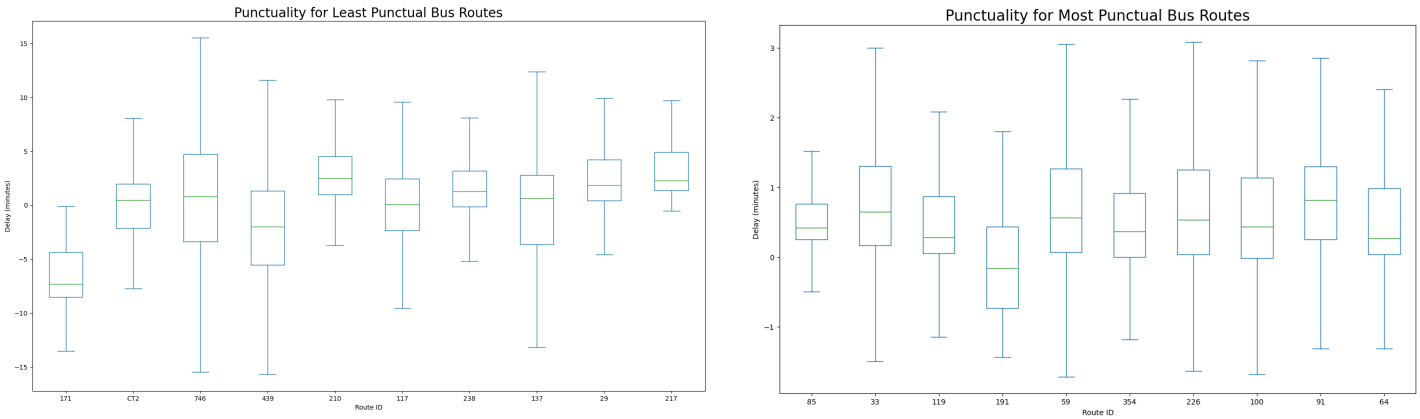

Fig. 9: Punctuality for top 10 least and most punctual bus routes

The average travel times for the top 10 least late and most late are depicted in the bar plots in Fig. 10. From the figure, we can see that the average travel time for the most late bus routes ranges from 10 to 40 minutes whereas that of the least late bus routes ranges from around 25 to 60 minutes.

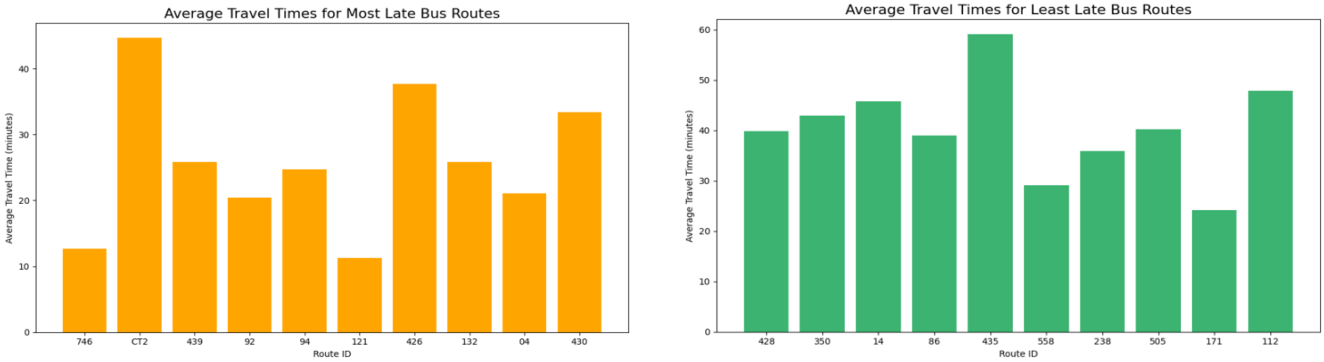

Fig. 10: Average travel times for top 10 most and least late bus routes

The average travel times for the least punctual and most punctual depicted in the bar plots in Fig. 11. From the figure, we can see that the average travel time for the least punctual late bus routes ranges from 10 to 45 minutes whereas that of the most punctual bus routes ranges from around 15 to 40 minutes.

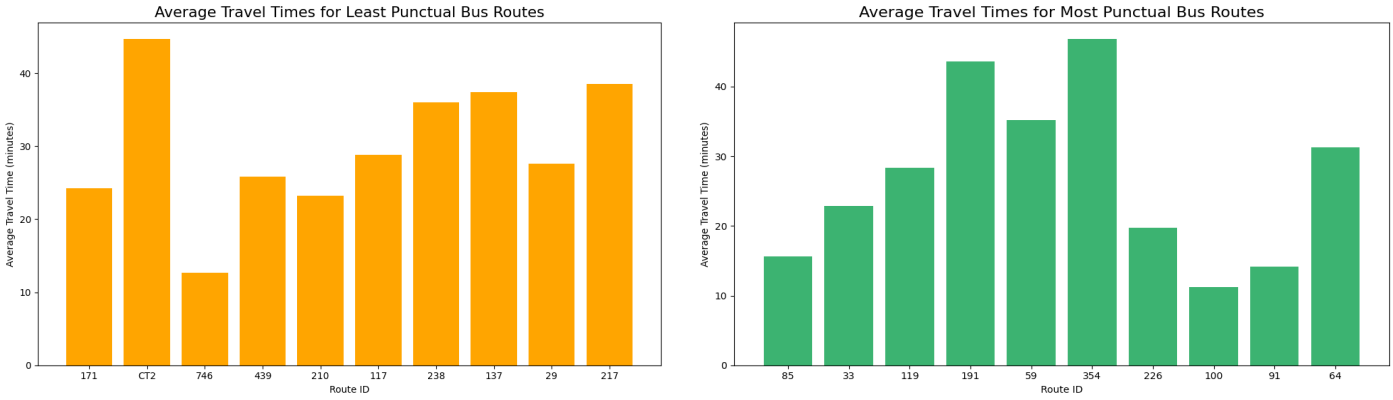

Fig. 11: Average travel times for top 10 least and most punctual bus routes

The service level disparities were further analyzed in terms of average lateness for routes. *Average lateness of routes* was calculated as the average difference between the actual total trip time and the scheduled total trip time. The average lateness for the 10 least late and most late bus routes are plotted Fig. 12. It can be observed that the least late bus routes have negative values of average lateness. A negative value means that on average, the trip took less time than scheduled, which is favorable. The maximum average lateness for the most late bus routes is around 7-8 minutes.

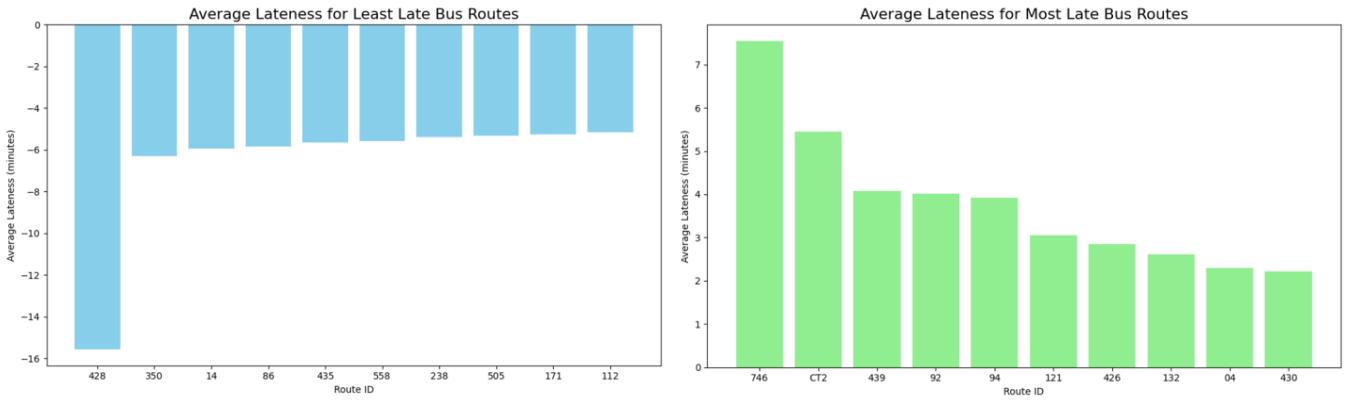

Fig. 12: Average lateness for top 10 least and most late bus routes

Another parameter for analyzing the service level disparaties was the *average absolute delay* of bus routes. It was calculated as the average absolute difference between the actual bus departure time from its start point and the scheduled departure time. We use the absolute difference because a bus leaving earlier than scheduled is not favorable, just like a bus leaving later than scheduled. Therefore, an average absolute delay close to zero is most favorable as it indicates that on average, the bus departed very close to its scheduled time. The average absolute delay for the least punctual and most punctual bus routes are plotted in the below graphs. Fig. 13 plots the average absolute delays for the top 10 most and least punctual routes.

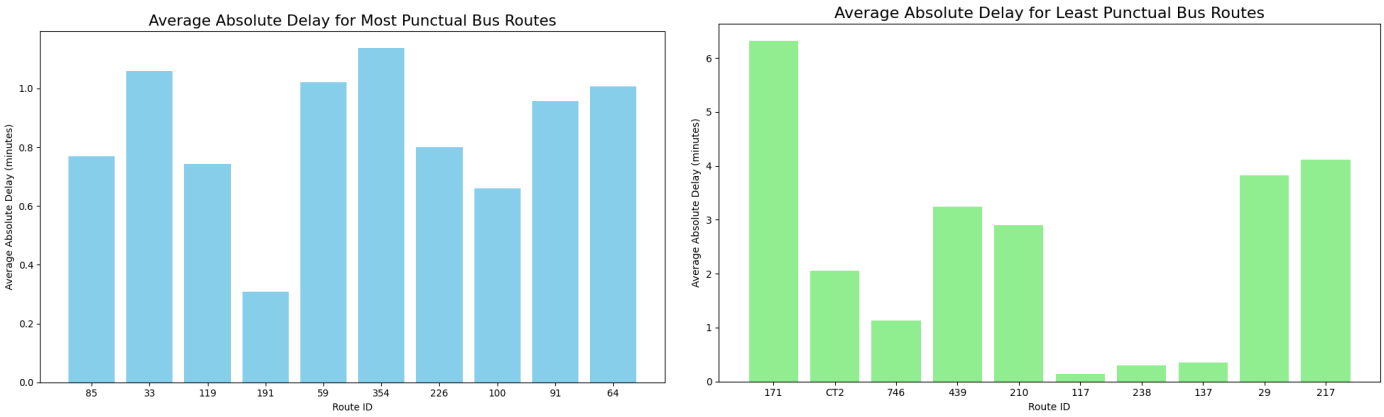

Fig. 13: Average lateness for top 10 least and most punctual bus routes

For addressing the rest of the base questions, the MBTA data was correlated to the census data. The MBTA V3 API was used to get the latitude and longitude coordinates of the stops for all the routes. The stops were then mapped to the neighborhoods of the city of Boston using the City of Boston neighborhoods geojson file. Since the census data had information pertaining to the 23 neighborhoods of Boston city only while the MBTA data had route information for the entire state of Massachusetts, some stops belonging to other cities of Massachusetts could not be mapped to the neighborhoods.

Fig. 14 depicts the total population of the most accessible and least accessible routes. Fig. 15 and 16 plot the racial and ethnic distribution of these routes. *Most accessible routes* here are defined as the routes serving maximum number of people overall and least accessible routes mean the routes serving a small proportion of the population. It can be observed that the most accessible routes serve a wider range of races including white, black, asian, hispanic or latino. The least accessible routes on the other hand serve predominantly white or black populations and their customers do not have a wide race variety.

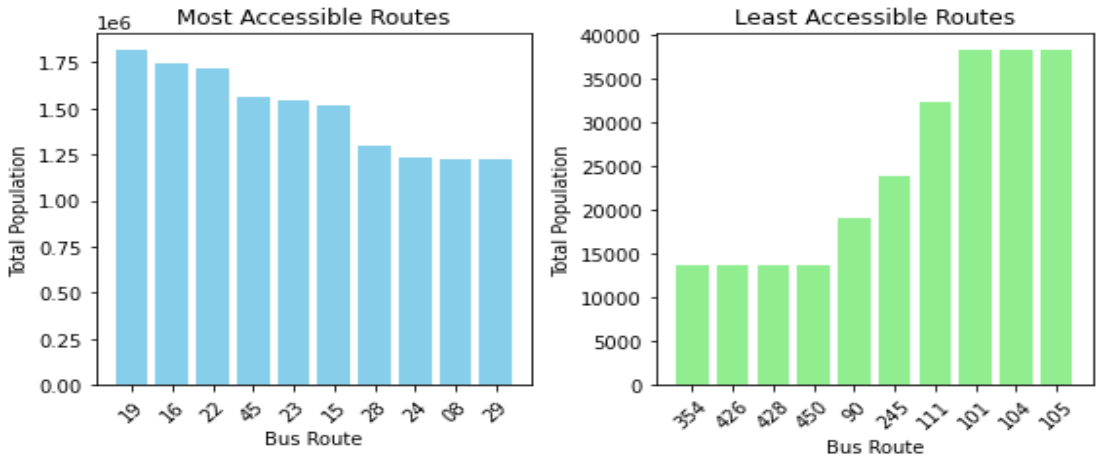

Fig. 14: Total population of top 10 most and least accessible routes

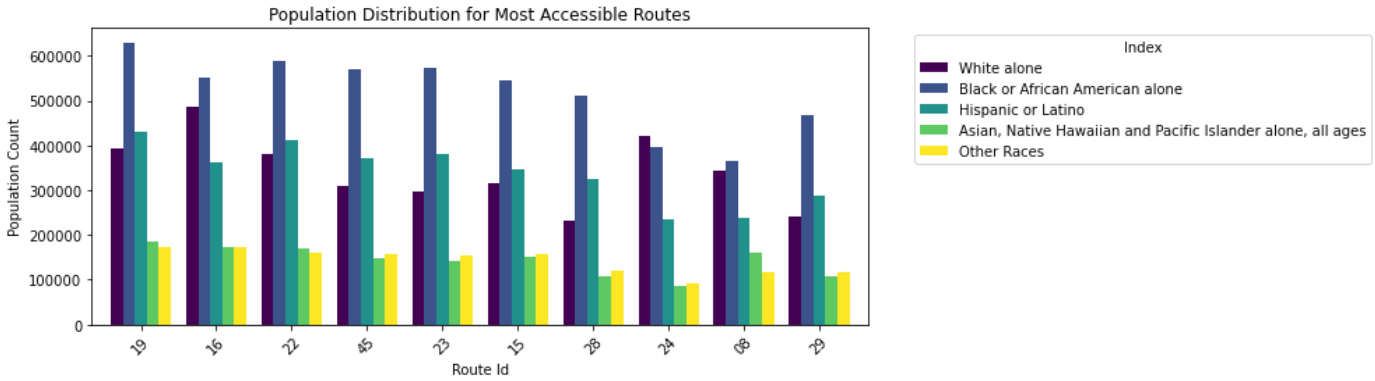

Fig. 15: Racial and ethnic distribution of top 10 most accessible routes

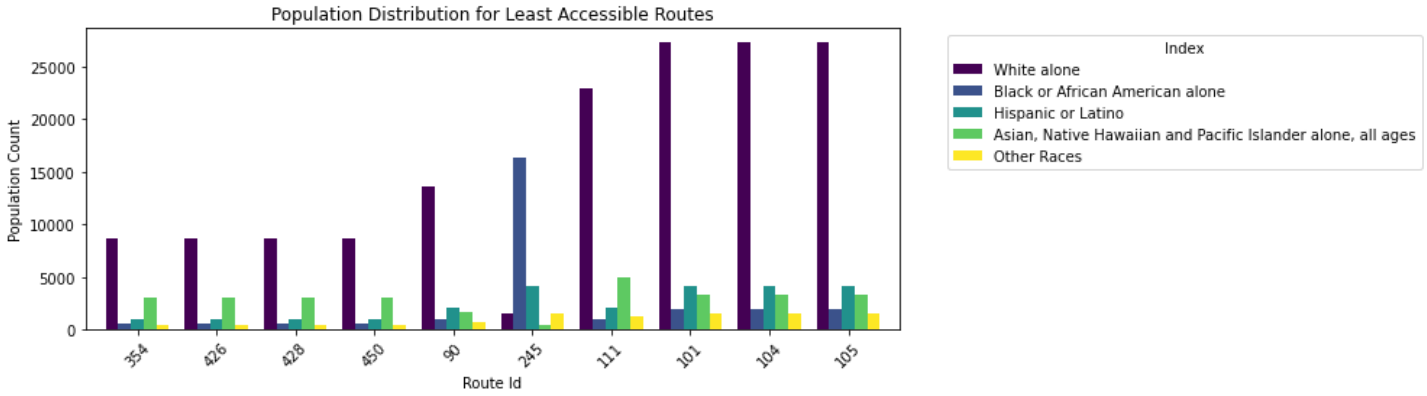

Fig. 16: Racial and ethnic distribution of top 10 most accessible routes

Our next step was to similarly analyze the dependency of Lateness and the ethnic distribution of the people affected by the service level disparities. In Fig. 17 we see that for the most late routes, most of the times, the overall number of people riding the bus is less. And, we also see that for the least late routes, there is a somewhat similar distribution of ethnicity. With the exception of route 86, where we see it serves more of the white populations.

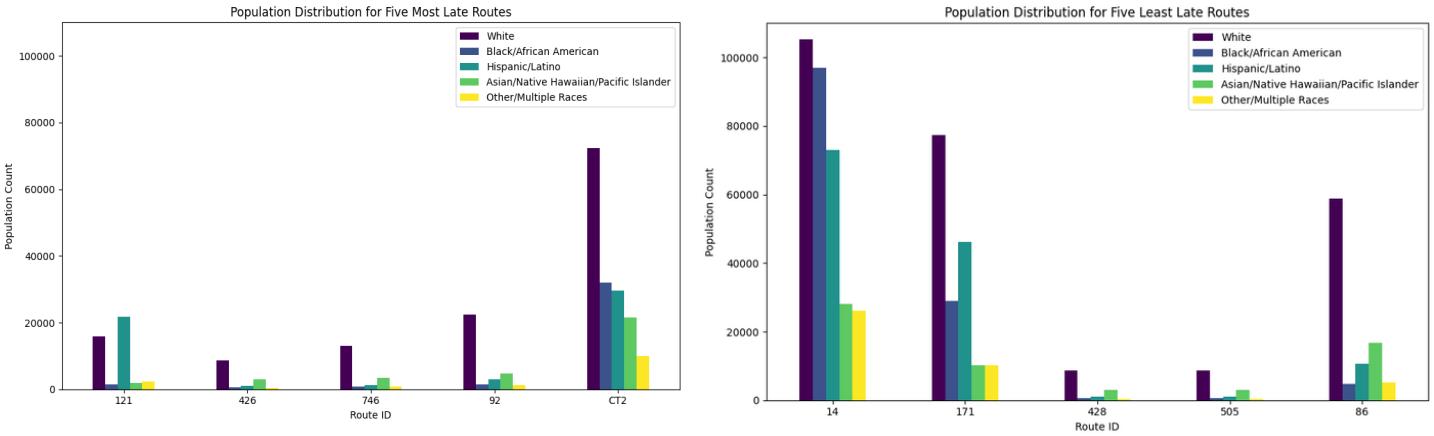

Fig. 17: Population and racial distribution for the top 5 most late and least late routes

We started to perform the same analysis on punctuality. Where we saw quite different results than the last two comparisons. In Fig 18, we see that the least punctual routes serve a wide distribution of communities. When we compare the most punctual routes, we see that most of the routes serve a higher percentage of white neighborhoods with the exception of route 191.

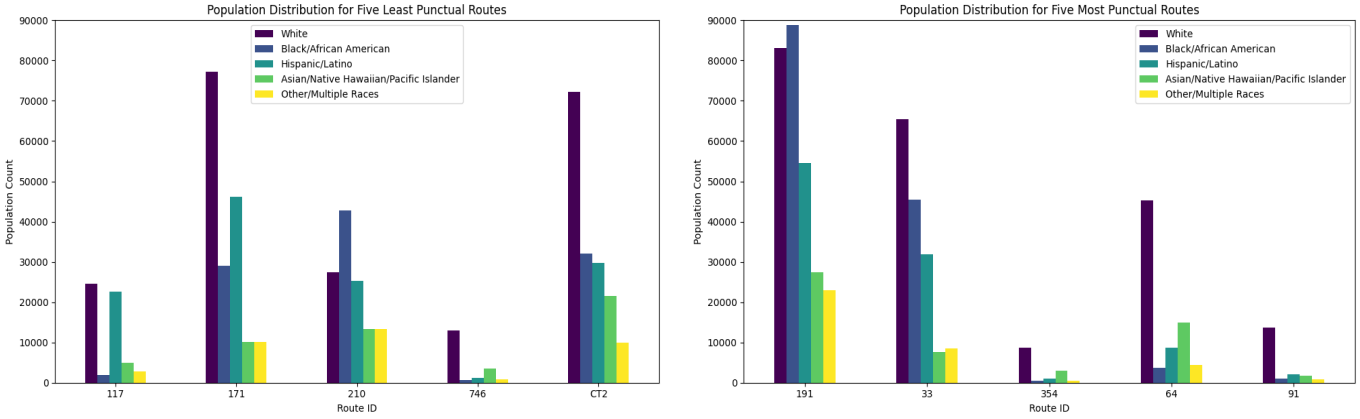

Fig. 18: Population and racial distribution for the top 5 most and least punctual routes

One theory that we had was that there might be some correlation between the college students and the Routes. The MBTA would possibly either add routes which serve college students more or the college students would also overcrowd the system leading to possible delays. To figure this out, we compared College students to Lateness & Punctuality.

The CT2 bus route is used by approximately 23,600 students, which is significantly higher than any other route in both the most late and least punctual categories. Routes 746, 92, and 117 each serve around 3,200 students and are present in both categories, suggesting high student usage regardless of the route's punctuality. Routes 121 and 210 have no students using them. The data highlights the importance of CT2 and could indicate a need to prioritize reliability improvements for routes with high ridership.

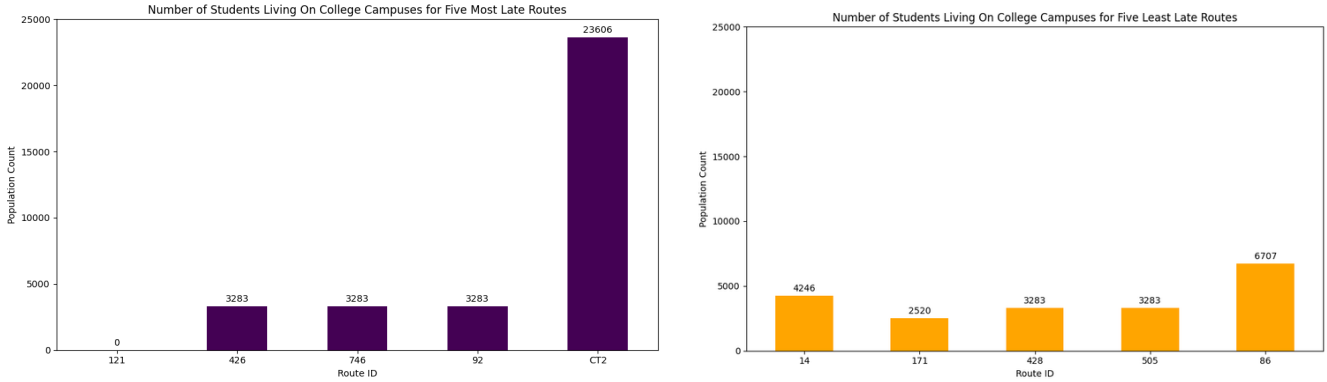

Fig. 19: Number of students living on college campus for the top 5 most and least late routes

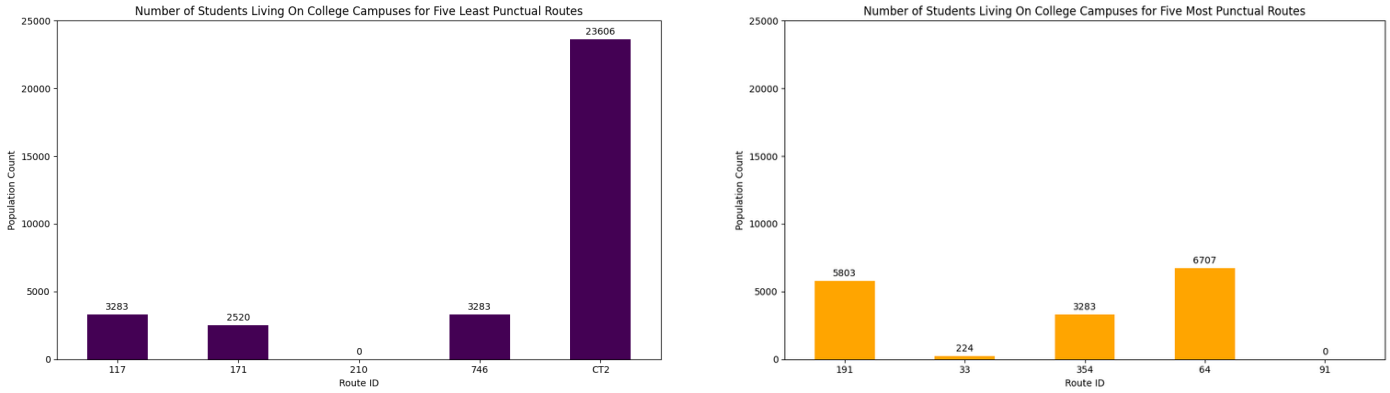

Fig. 20: Number of students living on college campus for the top 5 most and least punctual routes

Fig. 21 shows the age distribution of the neighborhoods served by the top five least and most punctual routes. We suspected that the routes that are less punctual would serve neighborhoods with more children. However, the plot shows that there is no direct correlation between the proportion of children and the punctuality of bus routes.

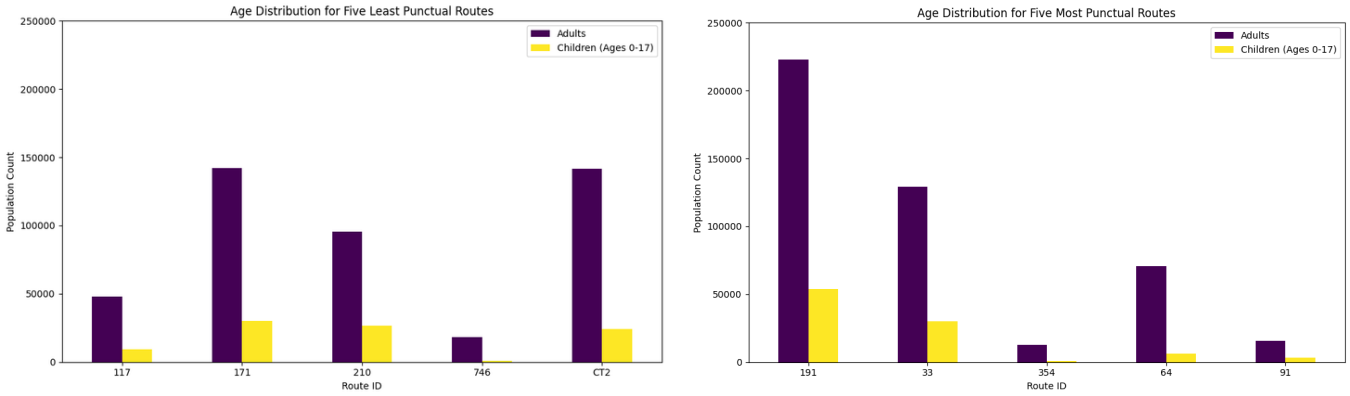

Fig. 21: Age Distribution for the top 5 most late and least punctual routes

5. Extension Project

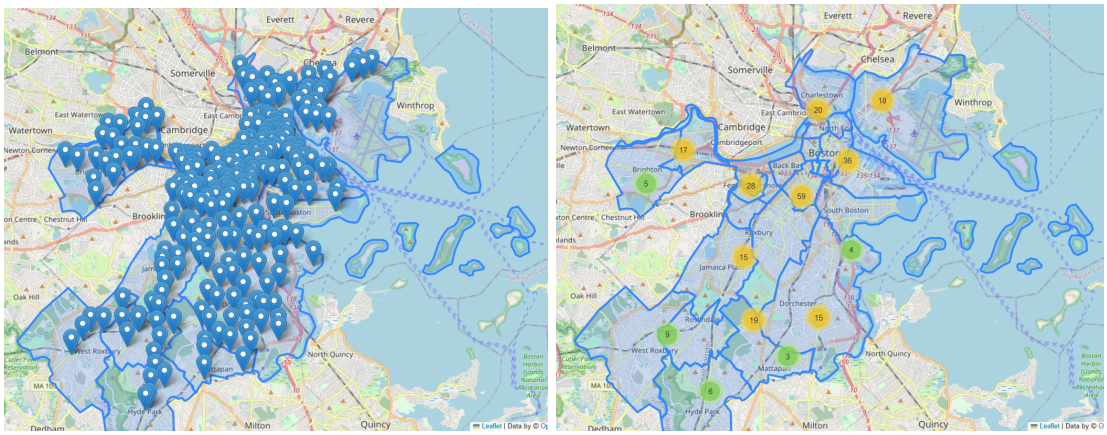

Fig 22: Location of Bluebikes stations across Boston

The first image depicts a map of Boston with markers representing the locations of Bluebikes stations throughout the city. At initial glance, it is apparent that the distribution of stations is not even across all neighborhoods. There is a high concentration of Bluebikes stations clustered in the downtown and central neighborhoods of Boston. Specifically, the area encompassing major attractions like the Boston Common, the waterfront, Fenway Park, and various universities has a very high density of stations. Within this central zone, the stations appear evenly dispersed along major roads and paths, providing convenient access for residents, workers, and tourists in these parts of the city.

However, the more residential outskirts and neighborhoods farther from downtown Boston have significantly fewer Bluebikes stations. Areas like East Boston, Roxbury, Mattapan, and Hyde Park have large swaths that lack bike share access entirely. Huge sections of these neighborhoods do not have a single station located nearby, which could severely limit transportation options for residents.

This imbalance in the location of Bluebikes stations across Boston neighborhoods can help us visualize and analyze potential disparities in availability and access in different communities. The map visualization makes it abundantly clear that large coverage gaps exist between downtown Boston and the outskirts. The distribution raises concerns about equity, as residents in certain areas may be excluded from utilizing the Bluebikes service simply due to lack of access. By displaying the precise geographic distribution, it will allow for clear identification of communities that are underserved by existing stations. This visualization and analysis sets the foundation for investigating why such access gaps exist, and how to best expand coverage to provide more equitable transportation options.

It appears that this map paints a picture of a bike share system highly concentrated in central commercial districts, with surrounding residential communities unable to enjoy the same benefits. This exemplifies why analyzing Bluebikes availability by neighborhood is an important undertaking, to ensure equal access across all of Boston.

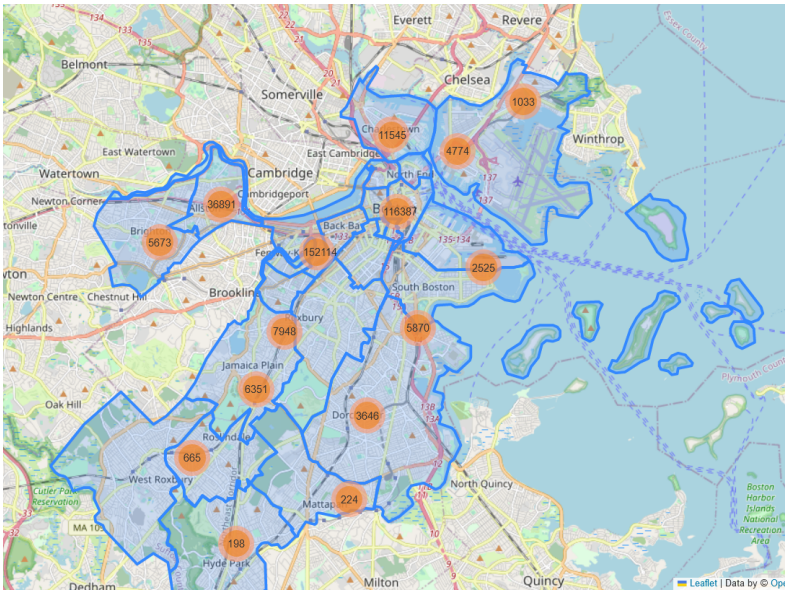

Fig 23: Location of Bluebikes stations aggregated

The map above provides a summary of Bluebikes ridership data for the month of September 2022. It shows that over 200,000 rides were taken on Bluebikes during that period, indicating that it is a widely-used and popular transportation option in Boston. With such high demand, it is important to analyze exactly where these rides are occurring.

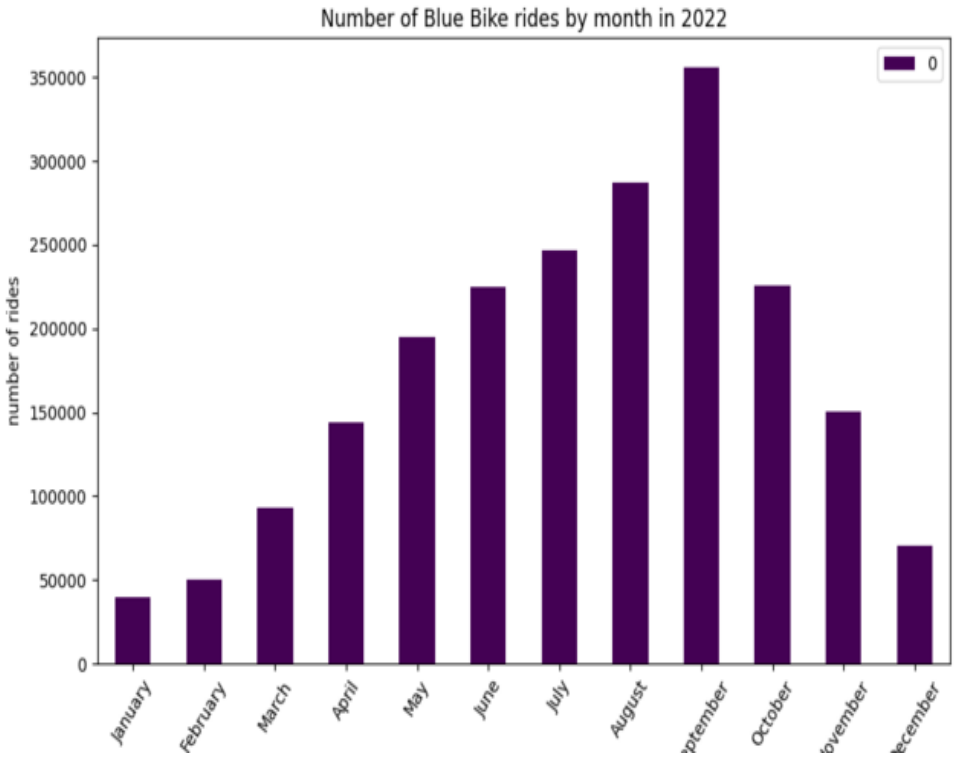

Fig 24: Bluebikes usage by month

Our analysis commences with a month-by-month examination, where a distinct correlation is observed between the usage of Bluebikes and seasonal variations. Notably, usage is at its lowest during the winter and spring months. A progressive increase in usage is evident in the summer months, peaking in September. This peak coincides with the commencement of the academic year, suggesting a potential surge in usage due to the influx of new students.

As done with the MBTA data, we start to perform a neighborhood based analysis to see how the Bluebikes usage correlates with the stations and neighborhoods.

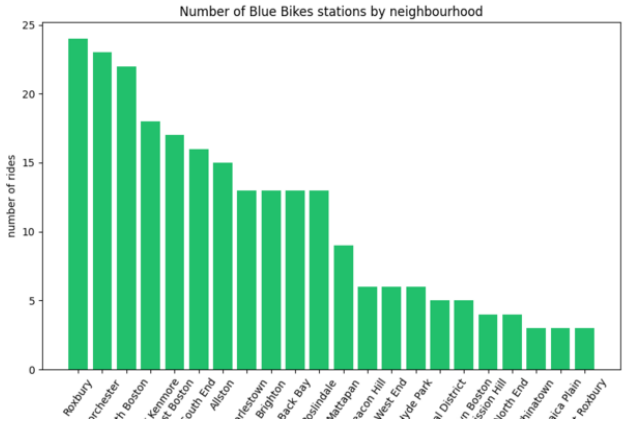

Fig 25: Bluebikes stations by neighborhood

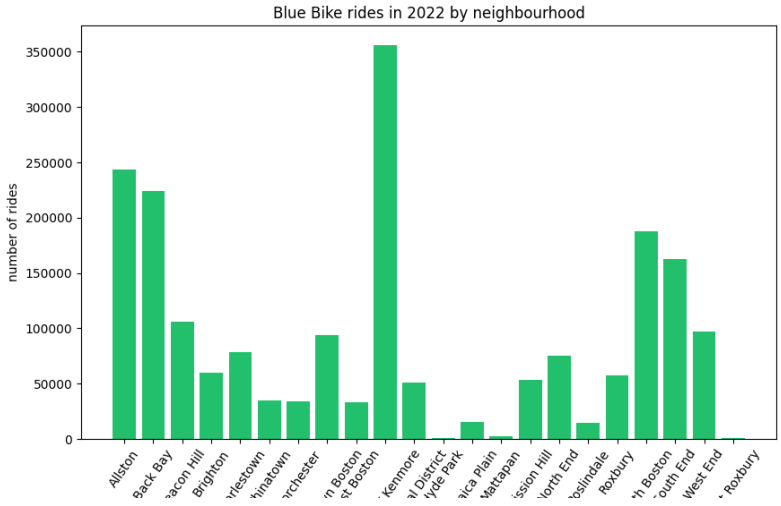

Fig 26: Bluebikes rides by neighborhood

Areas such as Roxbury, Dorchester and North Boston have the highest number of Bluebikes stations (around 20-25). However we can see no correlation between the number of stations and Bluebikes rides.

When compared by Neighborhood, West Boston leads the number of Bluebikes rides by a considerable amount (~350,000). Areas such as Allston & Back Bay also lead in the number of total rides.

Our analysis commences by examining the Bluebikes stations that experienced the highest and lowest usage throughout 2022. Notably, two of the most frequented stations are affiliated with universities: Forsyth St at Huntington Ave, situated in front of Northeastern University, and Commonwealth Ave at Agganis Way, located near Boston University. The remaining three stations are strategically positioned in the vibrant downtown and Back Bay areas, further underscoring their popularity.

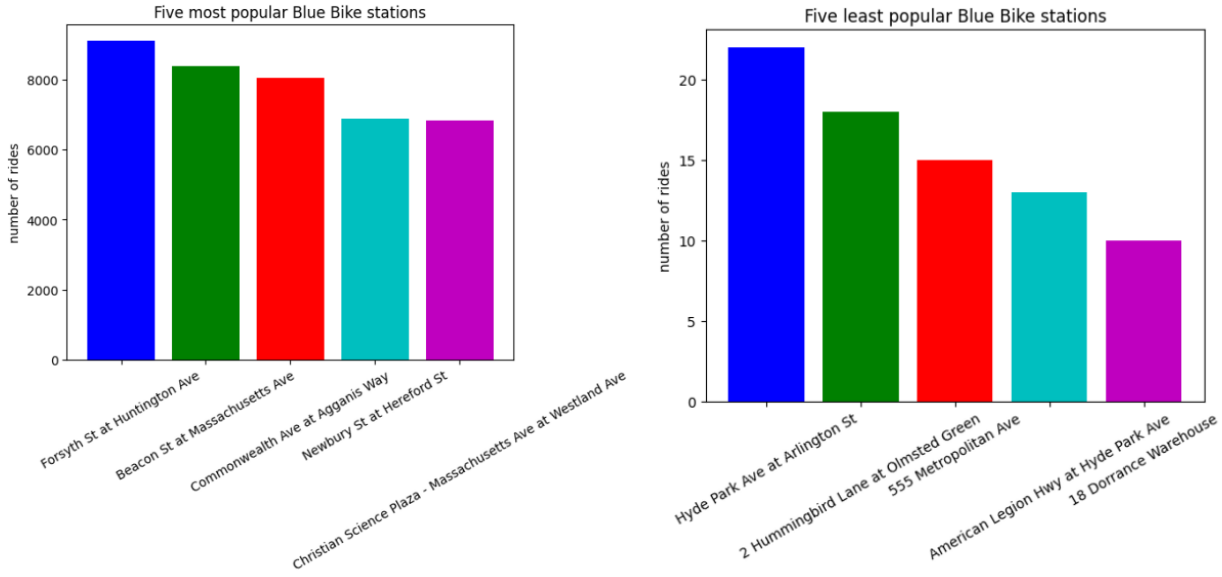

Fig 27: Most popular and least popular Bluebikes stations

Additionally, it becomes apparent that the least utilized stations are predominantly situated in areas characterized by limited connectivity with other Bluebikes stations. An illustrative example is the Hyde Park Ave area, which boasts one of the lowest concentrations of Bluebikes stations, and indeed, two of the least frequented stations are found within this vicinity.

However we will now shift our analysis to two months. We begin by delving into the data for September, which recorded the highest number of rides. Finally, we will conduct a comparison with the month of January, during which we analyzed the MBTA data.

During the months of September, a striking similarity emerges when compared to the annual data. It is evident that the stations with the highest and lowest usage remain virtually unchanged.

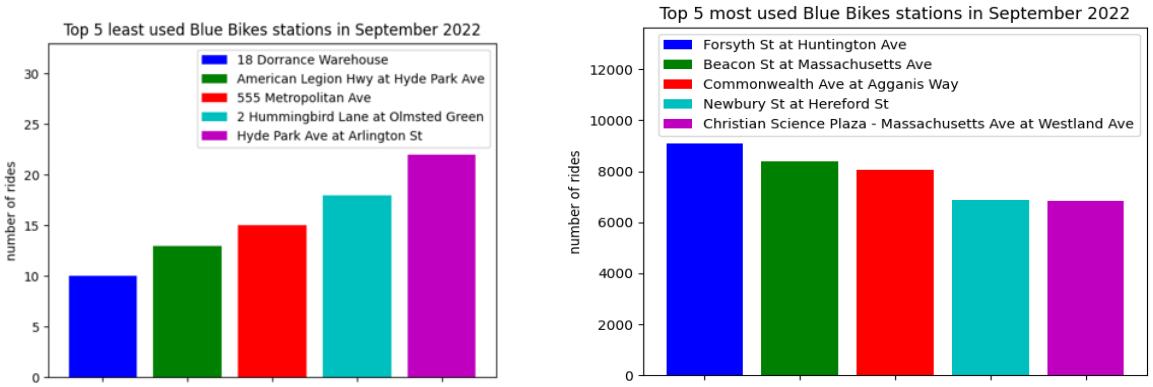

Fig 28: Most popular and least popular Bluebikes stations in September

Nonetheless, in the winter month of January, there is a significant decline in the number of rides, with some of the least used Bluebikes stations registering just a single ride for the entire month. The highest ride count observed in January was approximately 1600. Consequently, it is apparent that the data from the summer months has distorted the annual figures. While the most used bike stations largely mirror those of the summer months, one notable exception is Charles Circle, which is situated adjacent to the commencement point of the Charles River bike track—a plausible explanation for its distinct usage statistics.

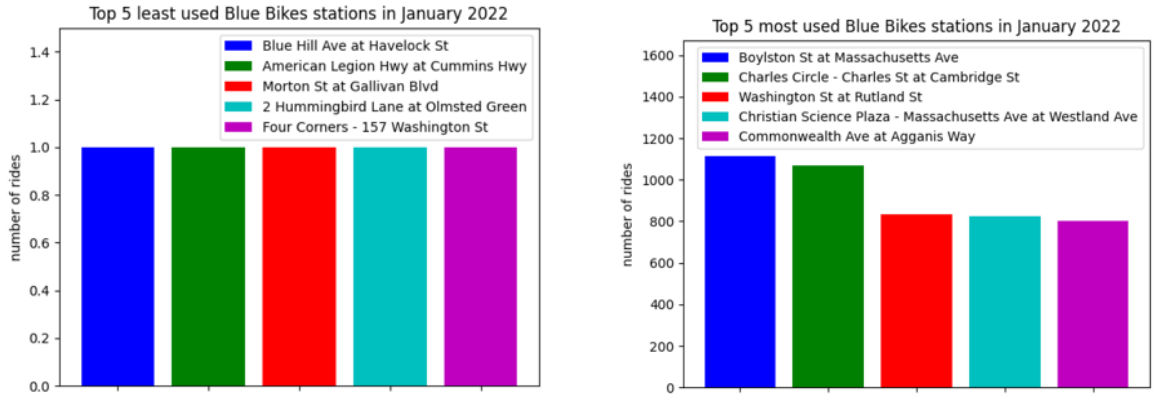

Fig 29: Most popular and least popular Bluebikes stations in January

In a manner similar to the MBTA data, we compare the population distribution of the Bluebikes stations. When we start to compare the usage for the most used stations, we see that they serve a diverse set of communities. Earlier we had also noticed that these stations are used near universities and in common areas in the downtown. So, we can assume that the population distribution for these stations would be normalized.

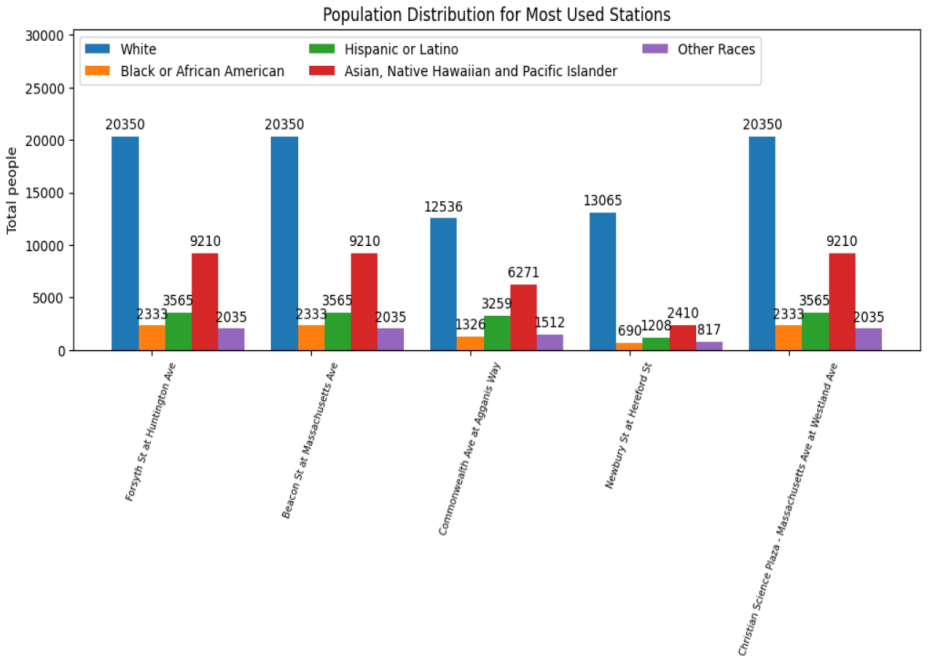

Fig 30: Ethnic demographics of most used stations

When examining the population distribution of the least utilized stations, it becomes evident that they predominantly cater to the African American community. This suggests an imbalance in service provision within these neighborhoods.

Fig 31: Ethnic demographics of least used stations

6. Challenges and Limitations

We faced certain challenges when doing the following:

1. **Merging Datasets:**
   When merging the neighborhood census data with the MBTA dataset, we had to do a large amount of preprocessing. The census data only contained the neighborhoods whereas the MBTA dataset contained latitude and longitude information. We had to use a geojson of the neighborhood boundaries available on the City of Boston website to correlate each stop's coordinates to the neighborhood it belongs to.
2. **Bluebikes datasets:**
   Similar to the MBTA and Neighborhood Census datasets, we had to preprocess to get the correct neighborhoods and stations to analyze. Additionally, the stations and rides datasets were mismatched, both in terms of ID numbers and station names. We similarly used an API to get the correct neighborhood for each station, and had to merge the stations and rides datasets by creating a unique list of stations and defining their attributes.

7. Conclusion & Suggestions

We analyzed the MBTA data initially to discover insights into bus lateness & delays. We found:

- On average bus routes are within the time range for a single route within ± 8 minutes, i.e. a bus route that was supposed to take 30 minutes from Point A to Point B was within the ± 8 minutes range.
- The most punctual bus routes were less than a minute late at their stops.
- The least Accessible Bus routes passing through the smallest neighborhoods and served more predominantly white neighborhoods
- The most punctual routes serve a higher percentage of white neighborhoods as well.
- The CT2 route served a wide majority of college students and is also one of routes which is least punctual and late.

We noticed the Bluebikes are a different mode of transportation and are usually used for last mile transportation and we can see this because of no direct correlation between Bluebikes and MBTA data. We observed:

- Seasonal variations with Bluebikes usage
- High clustering of Bluebikes stations does not necessarily lead to higher usage. While areas like Roxbury, Dorchester, and North Boston have the highest number of Bluebikes stations (approximately 20-25), there is no apparent correlation between station count and Bluebikes rides; West Boston stands out as the leader in Bluebikes rides with a significant number of rides (~350,000).
- Bluebikes usage directly correlates with university locations. Bluebikes stations near Northeastern and Boston University have the highest number of Bluebikes rides.
- The ethnic distribution across neighborhoods had no significant impact on Bluebikes usage.

Some suggestions we came up with:

- Increase bus frequency of line CT2 which serves many college students.

8. Individual Contributions

1. **Chandrahas**
    - EDA of MBTA dataset
    - Correlation of MBTA and census data
    - Addressing base questions 3, 4, 5
    - Final report

2. **Manushi**
    - EDA of MBTA dataset
    - Correlation of MBTA and census data
    - Addressing base questions 3, 4, 5
    - Final report

3. **Munir**
    - Data cleaning and preprocessing
    - EDA of census data
    - Addressing base questions 1, 2, 4, 5
    - Final report

4. **Patrick**
    - Addressing base questions 1, 2
    - Extension proposal
    - Visualizations and EDA for extension project
    - Final report

5. **Eason**
    - EDA of census data
    - Addressing base questions 4, 5
    - Visualizations and EDA for extension project
    - Final report