Date: 11/18/2023 CS506

Team A: AlHasan Bahaidarah, Salma Alali, Jasmine Fanchu Zhou, Isaac Chan, Tiansui Gu

MBTA Bus Performance Analysis

MBTA is a public transportation system serving more than 1 million people a day in the Massachusetts Bay Area. Our project looks into the performance of the MBTA system of Greater Boston Area for the buses specifically. Our goal is to identity the impact of bus performance on local population and distinguish any disparity of bus routes and their influence. We mainly collected our data from MBTA database and the City of Boston. We break down the overarching goal into five base questions towith a focus on differences of bus service disparities, including but not limited to number of bus lines, on-time performance and impacted resident demographic, across different neighborhoods and demographic groups. Additionally, based on our findings in base questions, we extended our research to focus on the impact of bus services on smaller and more precise geographic regions around the bus routes. Though there are challenges and limitation on data collection and analysis, we are highly motivated by the significance of public transportation to Boston economic development and quality of life ramifications.

Data Collection and Cleaning

Our MBTA datasets are from MBTA open data and API. Additionally, we relied on the census data from ANALYZE BOSTON and Census Bureau to provide population and geographical demographics. When comparing service performance across bus routes, we mainly used the MBTA bus dataset. To compare the accessibility of buses in each neighborhood, we combined census data and bus data by neighborhoods. For instance, in addressing question 5, we primarily utilized a census dataset detailing the neighborhoods of Boston. This dataset initially presented some formatting challenges; the column labels were positioned in the first row, causing the data to be stored as strings rather than integers. We fixed this by reformatting the dataset, ensuring correct display and converting numerical values to integers for further manipulation later.

In our analysis, we observed significant population variances across different neighborhoods. To address this, we normalized these figures into percentages, enabling a more equitable comparison. Additionally, we identified and eliminated redundant columns that did not contribute meaningful insights. A notable example was the removal of 'state' and 'city' columns in one of the datasets. Given that the MBTA exclusively operates within Massachusetts, these columns did not add value to our analysis.

We cleaned specific dataset according to our individual needs to work on the respective questions, but we also cleaned up the datasets together if multiple of us need the same dataset. Throughout this process, we continuously evaluated our datasets, actively seeking and discussing potential modifications to enhance clarity and relevance. This approach ensured that our data was not only accurately cleaned but also precisely tailored to address our specific research questions.

Exploratory Data Analysis

Certain trips have significantly longer average travel time outbound trips compared to inbound trips. Any trips that had inconsistencies in where buses last stop, with the expected last stop, were removed as it would have greatly affected the averages.

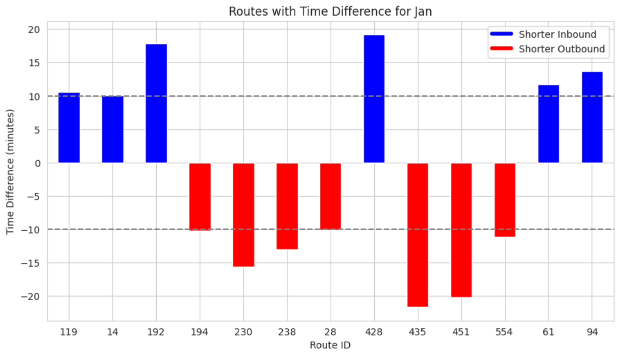
Though some conclusions can be drawn from the results we got, none were conclusive. There were sometimes clear trends in the behavior of some routes that would indicate differential treatment but even with the correlation it is difficult for us to draw casualty simply because the data is often indistinct. For example based on the data we tried to graph the usage of Bus based on population of neighborhoods and the routes. We took the worst performing routes from the answer of our first question, and we came to the conclusion that black people were greatly affected by bad performance of route 19. Upon further examination we found out that a big reason for this is possibly the fact that route 19 passes through the 2 most densely populated neighborhoods: Dorchester and Roxbury. If we were to rank these Neighborhoods based on the percentage of black people in them they are ranked 3rd and 4th respectively. So it might not be based on race specifically but there is a correlation, so we tried to look for more reasons as to why this is the case. We found that although Dorchester is the most densely populated neighborhood and has the most number of stops per neighborhood, it does not have an equally high number of routes. Meaning that many of the stops occur on the same bus, so if one bus trip is performing badly at the beginning of the trip it will most likely continue to perform badly on the entire route, increasingly affecting the neighborhood negatively. So rather than thinking the bus has special treatment, we think it might be an infrastructure problem.

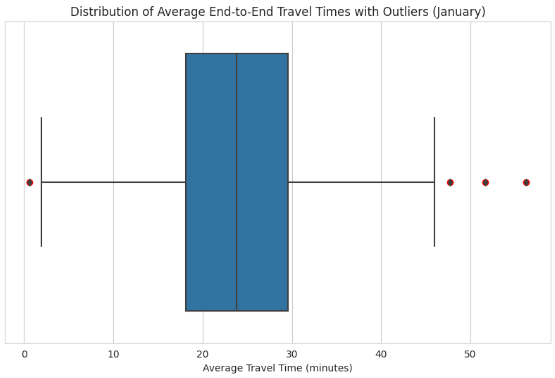
In-Depth Analysis of Base Questions

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

We used the MBTA arrival departure dataset that contained data about the routes and the stops it would run through, containing the times at which they would arrive. From there, we took all routes and calculated the difference in end and start stops' times, and took their averages per route (as each route typically runs more than once a day). For further extraction, we created a boxplot for average end-to-end travel times as well as a graph of routes with a disparity in inbound and outbound trips if either has more than 10 minutes difference in travel time.

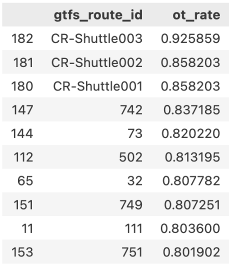
As for our visualizations and findings, we found that trips typically ran between ~18 to ~30 minutes, a max of 45 minutes and outliers going past 50 minutes. We also find that there are more trips that have shorter outbound times than inbound times, and disparities in inbound and outbound times overall reached a max of 20 minutes difference.



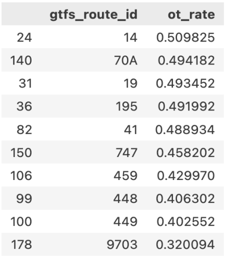


Base Question 2: Are there disparities in the service levels of different routes?

We define disparities as the different service performances. More specifically, we evaluate the rate of being on-time for each bus route. MBTA categorizes the on-time performance (OT) as one of the measurements of bus reliability. We calculated the OT rate by dividing the number of times a bus was on time from the total times of measures. Bus routes with higher OT rates indicate that they are more likely to depart from a stop on time. We then found out the bus routes with the highest and lowest OT rates as following:



	gtfs_route_id	ot_rate
182	CR-Shuttle003	0.925859
181	CR-Shuttle002	0.858203
180	CR-Shuttle001	0.858203
147	742	0.837185
144	73	0.820220
112	502	0.813195
65	32	0.807782
151	749	0.807251
11	111	0.803600
153	751	0.801902



	gtfs_route_id	ot_rate
24	14	0.509825
140	70A	0.494182
31	19	0.493452
36	195	0.491992
82	41	0.488934
150	747	0.458202
106	459	0.429970
99	448	0.406302
100	449	0.402552
178	9703	0.320094

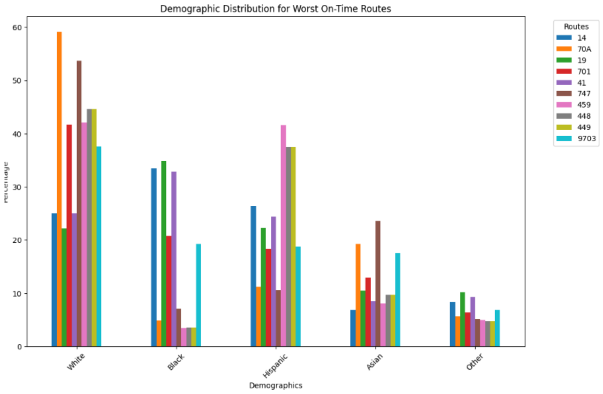
It is note-worthy that most bus routes with higher OT rates are located on the northern side of the Greater Boston area, the lowest OT rates occur mostly around south of Boston, around the Brookline, Jamaica Plain and Dorchester area.

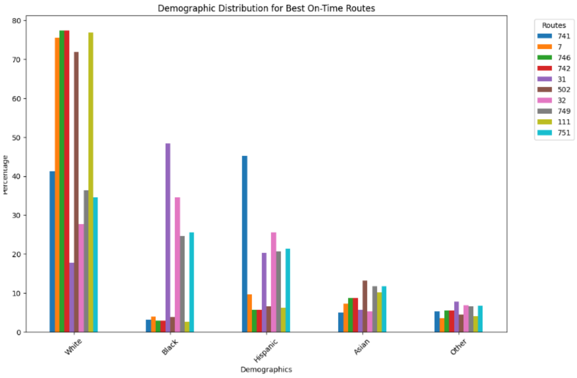
Base Question 3: What are the population sizes and characteristics of the communities serviced by different bus routes (e.g. race, ethnicity, age, people with disabilities/ vulnerabilities)?

In terms of methodology, it is somewhat impractical and meaningless to find community information for all different bus routes, because there are more than 100 bus routes, and each route can go across various communities. Instead, by investigating the on time performance of bus stops in each community can give a more insightful outcome. As for our visualizations and findings, by plotting the on time performance for each stop and show them on the map of Boston, we see a pattern that almost all stops that are located at central Boston has a very low on time rate, and for stops that are far from central, on time rate is much higher; It seems that such performance is independent of race/ethnicity, and almost all stops are friendly to disabled persons.

Base Question 4: If there are service level disparities, are there differences in the characteristics of the people most impacted?

We extracted the top ten worst routes from base question 2, but not all immediately showed in our MBTA-GTFS dataset (which focused on individual stops), so we focused on the top 10 worst routes we were able to locate in GTFS dataset. We joined the GTFS dataset and the census neighborhood datasets on the neighborhood as a common id. From there, we extracted the routes passing through, per stop, and matched them with the worst 10 routes from base question 2. We decided to explore the top 10 routes as well. We managed to find that for the worst 10 routes, the demographic distribution varied per route, but the primary demographics affected overall was in the following order: White (40% average), Black (35% average), Hispanic (25% average), Asian (15% average), and Other (10% average). We attribute this due to Boston having a primarily white demographic overall, according to the census. What was of note was that 5 of the top 10 routes serviced areas primarily white (70%+ of the neighborhood), whereas the other routes do not service just as well neighborhoods that may be primarily any other demographic. The census data included data on university attendance rate. We found that the worst routes based on on-timeness had a higher university attendance rate (0.151 average vs. 0.026 average attendance rate). A potential explanation for this could be that neighborhoods with higher university rates have higher congestion due to higher traffic of passengers. We also used university rates to discover that the top ten routes had much lower university rates compared to the worst ten, indicating university students may be a cause for congestion.

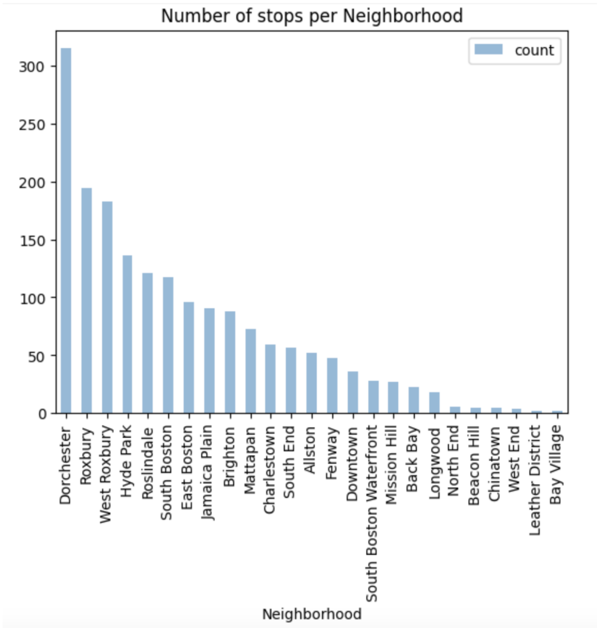


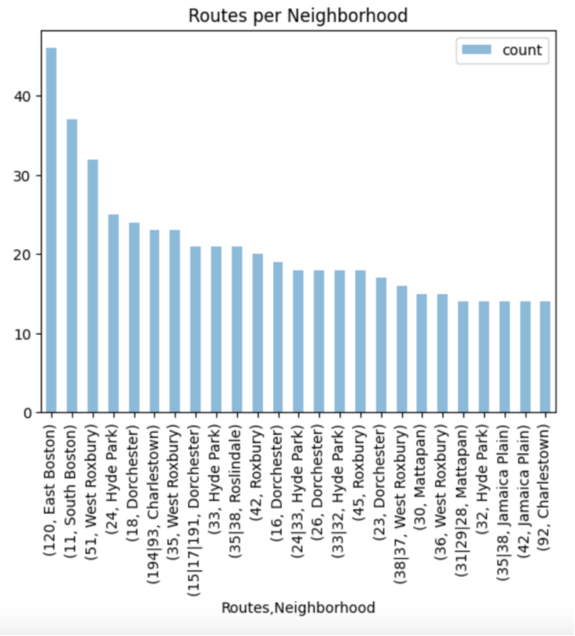


Base Question 5: This can include questions about traffic information, which neighborhoods are served better/worse by the MBTA bus system, which routes are better/worse, differences in quality of service by class/race, contributing variables, ect.

Our process was the following: assuming all races in every neighborhood use the same routes uniformly, finding the percentage of each race living in each neighborhood (giving more weight to the highly dense neighborhoods) and connecting that to the routes going through them. Connecting the performance to the races we found a correlation between them.

Visualizations and Findings: I started by graphing the census data of the populations per neighborhood. Looking at the graph the population density wasn't uniform. Changed that to percentages so that it was most representative. Plotted the number of stops per neighborhood and number of routes per neighborhood. The combination of these 2 showed a disparity in the number of options in public transportation in each neighborhood.



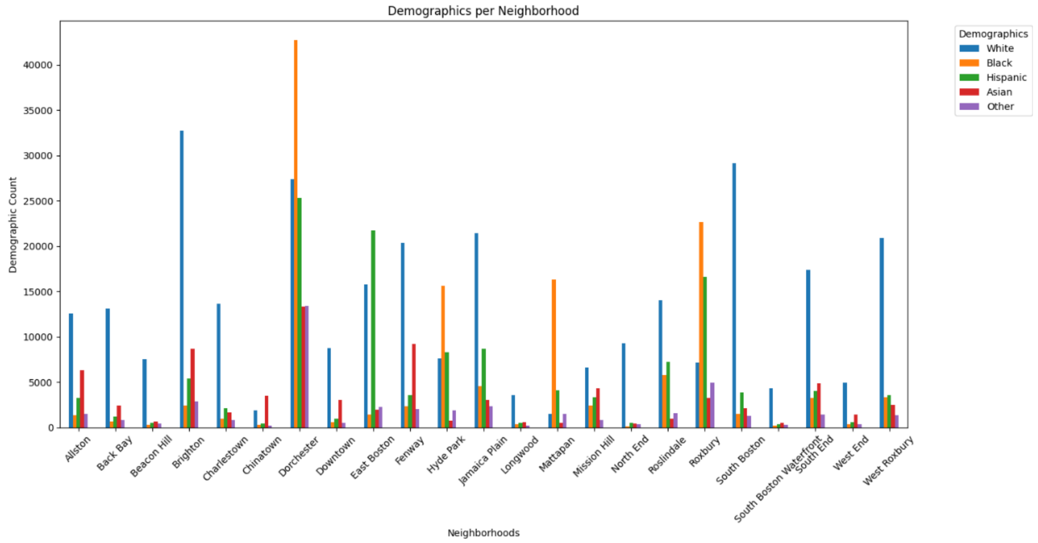


Extension Proposal

Many people could agree that the MBTA routes could use a little more variability. Now to the decision of where to add routes, how to go about making the decision are factors that depend on insights extracted from our implementing concepts learned in class, such as clustering. We thought about clustering the stops based on the density of how many people use those routes/stops in a way that would emphasize the usage/efficacy of each route. We will be making note of where some stops intersect and seeing if they are carrying similar passengers. That will help determine bottleneck stops where there may be large transfers from route to route at one stop, if for example there are few routes in the neighborhood to begin with. Having few routes that span a long distance (with many stops) may be less efficient than multiple routes; it would benefit the MBTA to add routes that would add convenience to the lives of the citizens of Boston and reduce bottleneck stops by allowing for more than one chance to transfer. If we cluster the stops from different routes with each other accounting for the foot traffic we could find the stops where people change buses, therefore accounting to where people might want to add new routes. We could also cluster stops based on foot traffic irrelevant of intersections and this would translate into where we would need to add more frequent buses to the particular route. The way we will be producing these clusters is using the MBTA GTFS dataset that contains longitude, latitude, neighborhood, and route data, per stop. We plan on combining this with Census data to see counts of travelers as the clusters. We expect that bigger clusters will mean more stations in the neighborhoods, when it comes to the counts of routes, and it will mean more users in the neighborhoods when factoring the demographics given by the Census data.

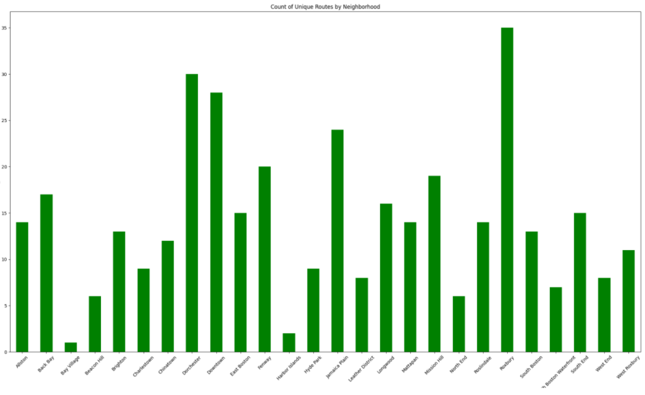
Visualizations and Insights For Extension Project

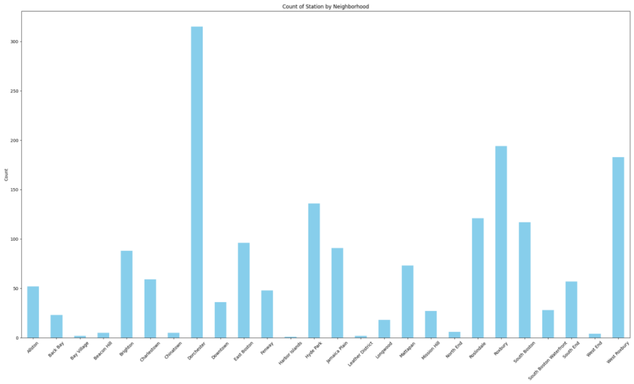
One of the ways we planned on exploring the dataset was by finding the demographics of neighborhoods. The idea behind this is to gain insight into overall percentages of demographics in the respective areas to potentially gain more insight as to what neighborhoods may have more or less clusters of routes. We also use this to take into account the numbers of people per neighborhood.



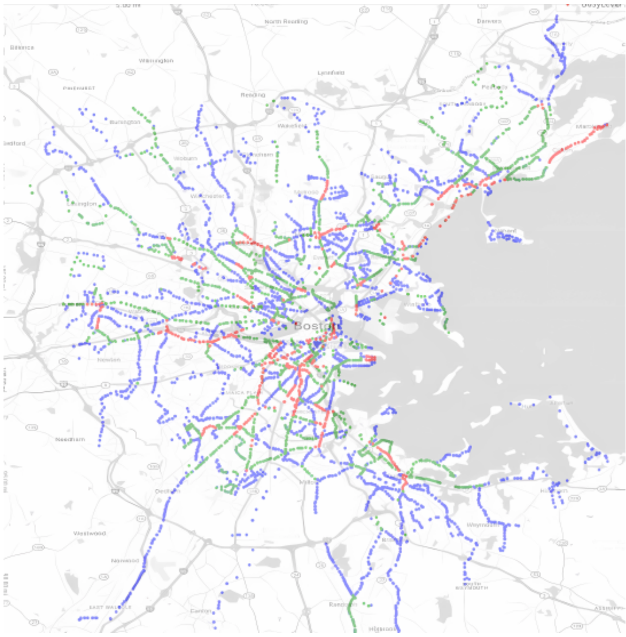
We note here that Dorchester has the most people, and the primary demographic representing the neighborhood is Black. Other than Dorchesters, there are a couple of other neighborhoods we plan on further looking into: Brighton, East Boston, South Boston, Mattapan, Chinatown, and West Roxbury. These neighborhoods are all (but Chinatown) of note due to there being a primary demographic representing the neighborhood, and we believe that it may be of relevance to the number of route clusters. Chinatown is noteworthy as it contains the least number of people living there, but given the economic/business presence in Chinatown, it may be worth looking into how many clusters of routes there may be compared to some neighborhood that is primarily for residence.

Another way we plan on looking into further is the counts of the routes per neighborhood; The idea behind this analysis is to see whether the size of the neighborhood affects the number of routes, whether it affects the number of stops present in the neighborhood, and how the population of the neighborhood comes into play together. In this graph we can see the number of stations per neighborhood. We see that Dorchester has a high number of stations, and it has a high number of unique routes passing through it. Places such as Roxbury are the opposite; there are many unique routes, but much fewer stops in comparison to Dorchester's. With this, we can interpret that there may be frequent stops in Dorchester, whereas in Roxbury, while there are many unique routes, there are less frequent stops, which may mean that it increases foot traffic to switch over from route to route.





The final way we looked at this extension proposal is the number of routes passing through an individual station. Given this graph, we see that there is a predominance of blue stations: These correspond to stations that have only one or two routes passing through them. Those that are green have a larger number, and the ones in red are the ones with the most routes going through them. We primarily see that the further branched out from the Boston city area, the more blue stations there are. This suggests that the further out of the city area we look, there is more divergence in routes, more coverage of area. However, the expectation of more lines overlapping in the city area is not well-realized; this is to say that while there are stops that contain multiple routes passing through them, given that they primarily are in single lines, it may be the case that multiple routes run the same initial path before diverging into paths that are green or blue (less overlapping routes). This is in contrast to what we believe would be more efficient: red stops that are less connected. The reasoning behind why disconnected red stops may be more efficient is that there is not so much overlap in routes going through the same path that trips are made redundant, but at the same time, there exists frequent enough options to change routes.



Individual Contribution

Isaac:

- Preprocessed and completed most of the preliminary analysis for the first base question (end to end times).
- Preprocessed and completed most of the analysis for base question 4 (characteristics among lower serviced routes)
- Wrote report, specifically the base questions section
- Completed and prepared for respective part in presentation
- Attended all meetings and presentations

Salma:

- I worked on both base questions 1 (end-to-end times) and 4 (characteristics in routes with service disparities) with Isaac.
    - I graphed differences in outbound and inbound routes (first base question)
    - Helped merge MBTA GTFS, Census neighborhood data along with 10 best and worst routes (base question 4).
- Organized team meetings
- Keeping the team informed of expected deadlines and following up with progress updates on working towards them.
- Primarily responsible for doing the pull requests to the github repo
- Came up with angles to analyze in the extension proposal

AlHasan

- Was responsible for attending the labs and communicating with the TPM
- Put together most of the slides for the presentations
- Wrote the majority of the report
- Worked mainly on Question 5
- Came up with the idea for the extension proposal

Jasmine Fanchu Zhou

- Worked on Base Question 2, creating visualizations
- Revised Introduction and data collection of the report
- Formatted report

Tiansui Gu

- Worked on Base questions 2 and 3
- Generated geographical visualization based on data we found
- Process data and generate graphs for extension proposal