MBTA Bus Performance Analysis Final Report

Fall 2023 CS506

Team A

AlHasan Bahaidarah, Salma Alali, Jasmine Fanchu Zhou, Isaac Chan,

Tiansui Gu

Table of Content

MBTA is a public transportation system serving more than 1 million people a day in the Massachusetts Bay Area. As bus service is one of the most important public transportation, and it is crucial to Boston's economic development and quality of life improvement. Our project looks into the performance of the MBTA system of the Greater Boston Area for the buses specifically. Our goal is to identify the impact of bus performance on the local population and distinguish any disparity among the performance and reliability of the MBTA bus routes. We mainly collected our data from the MBTA database and the City of Boston. We break down the overarching goal into five base questions with a focus on the number of end-to-end operation times of bus lines, on-time performance, and impacted resident demographics across different neighborhoods. Additionally, based on our findings in base questions, we extended our research to focus on the association between Bluebikes usage and bus service along the ten bus routes with the best and the worst-time performance.

1. Methodology and Data Processing

1.1. 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 the 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 neighborhood. 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.

1.2. Data Exploration

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. Our MBTA datasets are from MBTA open data and API. Additionally, we relied on the census data from ANALYZE BOSTON and the 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 neighborhood. 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.

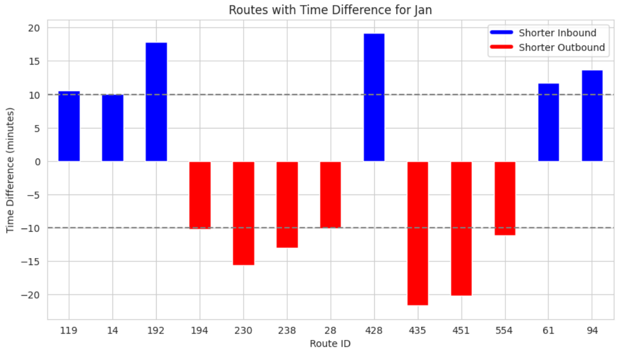
In our analysis, we observed significant population variances across different neighborhoods. To address this, we normalized these figures into percentages, enabling a more equitable comparison. Additionally, we identified and eliminated redundant columns that did not contribute meaningful insights. 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.For example, based on the data we tried to graph the usage of buses based on the population of neighborhoods and the routes. We took the worst-performing routes from the answer to our first question, and we concluded that black people were greatly affected by the 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. This means 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.

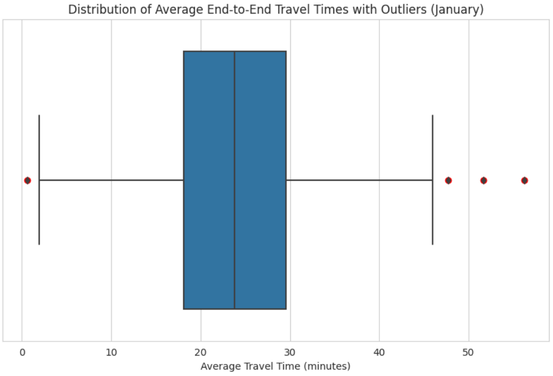
2. Base Question Analysis

2.1. 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, including the times at which they would arrive. From there, we took all routes, 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). In that graph (Figure 1 in Appendix), there were many routes, much too difficult to simply analyze without any further narrowing. As a result, we decided to derive a main idea from such graphs by thinking about what kind of graphs or queries we could ask that would further narrow our graphs into a more substantial finding. As a result, 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 a 10-minute difference in travel time.

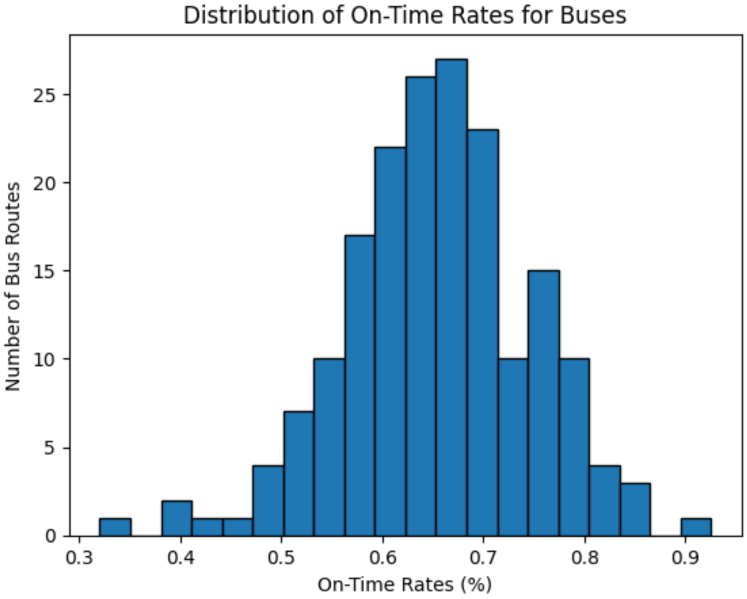
As for our visualizations and findings, we found that trips typically ran between ~18 to ~30 minutes, a max of 45 minutes and outliers going past 50 minutes. We also find that more trips have shorter outbound times than inbound times, and disparities in inbound and outbound times overall reached a max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





2.2. 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. In the data set, we were provided a variable that recorded the count of a bus being on time. Since the count may be affected by the number of stops and number of round trips of bus routes, we created a new variable called ‘ot_rate’ to take the factors mentioned above into consideration. We calculated the OT rate (‘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. The distribution of OT rate for bus routes are roughly follow a normal distribution, where most routes have OT rates between 0.6 to 0.7, indicating that 6 out of 10 measures the bus will be late. There are some extreme cases where the bus routes are almost never late or on time.



We then found out the bus routes with the highest and lowest OT rates. The following tables show the names of the routes with the highest OT rate, the route directions, and their OT rate respectively. Some routes, such as CR-Shuttle003, CR-Shuttle002 and CR-Shuttle002 are recorded in the dataset, but they cannot be located on the MBTA official website. That is, they did not have a route schedule and map. This could imply that these three are just temporary substitute routes for certain lines. Regardless the incomplete information, non of these routes pass through downtown Boston and most of them are North-South oriented.

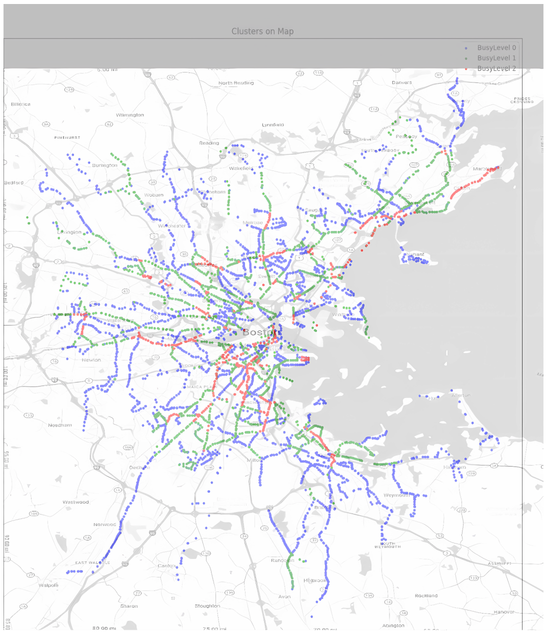
ID for Routes with Best OT Rate	Region	OT Rate
CR-Shuttle003	Unknown	0.925859
CR-Shuttle002	Unknown	0.858203
CR-Shuttle001	Unknown	0.858203
742	Silver Line 2	0.837185
73	(East-West) Harvard -Waverley	0.820220
502	(East-West) Back Bay - Watertown	0.813195
32	(North-South) Jamaica Plain - Readville Manor	0.807782
749	Silver Line 5	0.807251
111	(North-South) Woodlawn - Haymarket	0.803600
751	(Northeast - Southwest) Silver Line 4: Nubian - South Station	0.801902

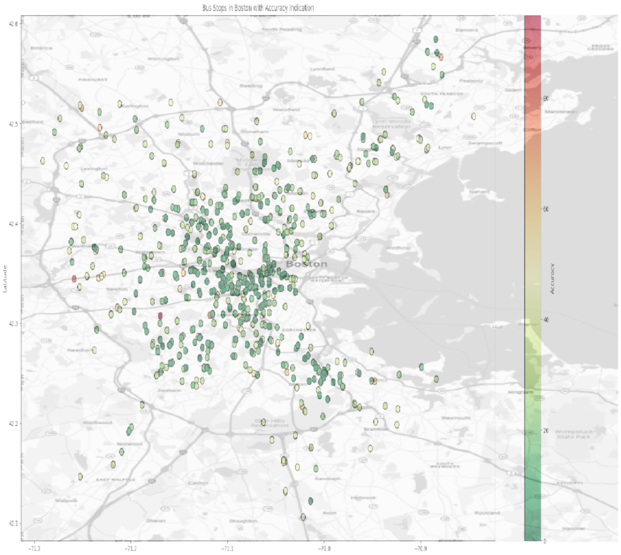
Among routes with the lowest OT rates, many routes are East-West or Northeast-Southwest oriented; four of them (Route 70A, 459, 448, and 449) are revised, eliminated or replaced by either pre-existing or additional routes; and the schedule of one of them (Route 9703) is not found.

ID for Routes with Worst OT Rate	Region	OT Rate
14	(North-South) Health St. - Roslindale Square	0.509825
70A (replaced by Route 61, and added service on Route 70)	(East-West) Cambridge - North Waltham	0.494182
19	(Northwest-Southeast) Kenmore - Fields Corner	0.493452
195	(North-South) Park Street & Tremont Street - Lemuel Shattuck Hospital	0.491992
41	(East-West) JFK/UMass - Centre St & Eliot St	0.488934
747	(North-South) Crosstown 2: Somerville - Fenway	0.458202
459 (replaced by Route 455)	(Northeast-Southwest) Peabody - Revere	0.429970
448 (replaced by Route 441 and 442)	(Northeast-Southwest) Marblehead Historical District - Revere	0.406302
449 (replaced by Route 441 and 442)	(Northeast-Southwest) Marblehead Historical District - Revere	0.402552
9703	(East-West) Jackson Square Station via Mass Pike -Brighton High School	0.320094

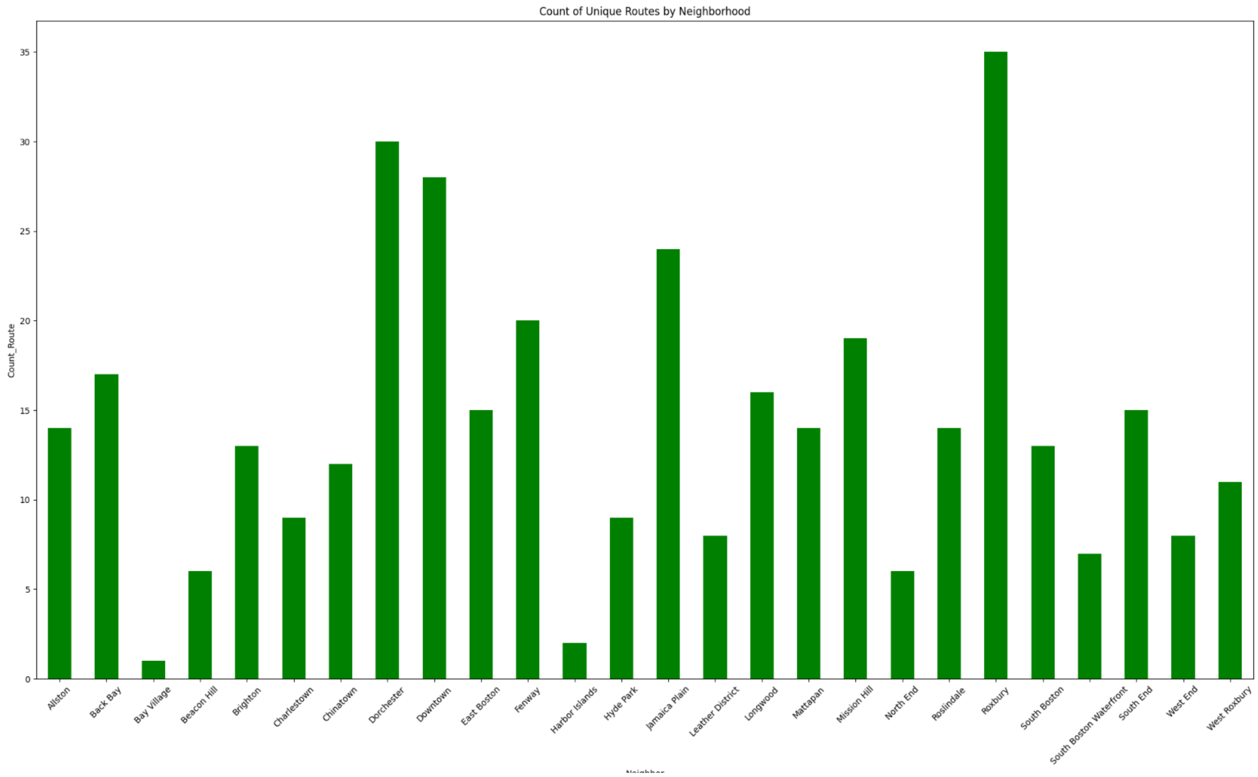
2.3. Question 3: What are the population sizes and characteristics of the communities serviced by different bus routes (e.g. race, ethnicity, age, people with disabilities/ vulnerabilities)?

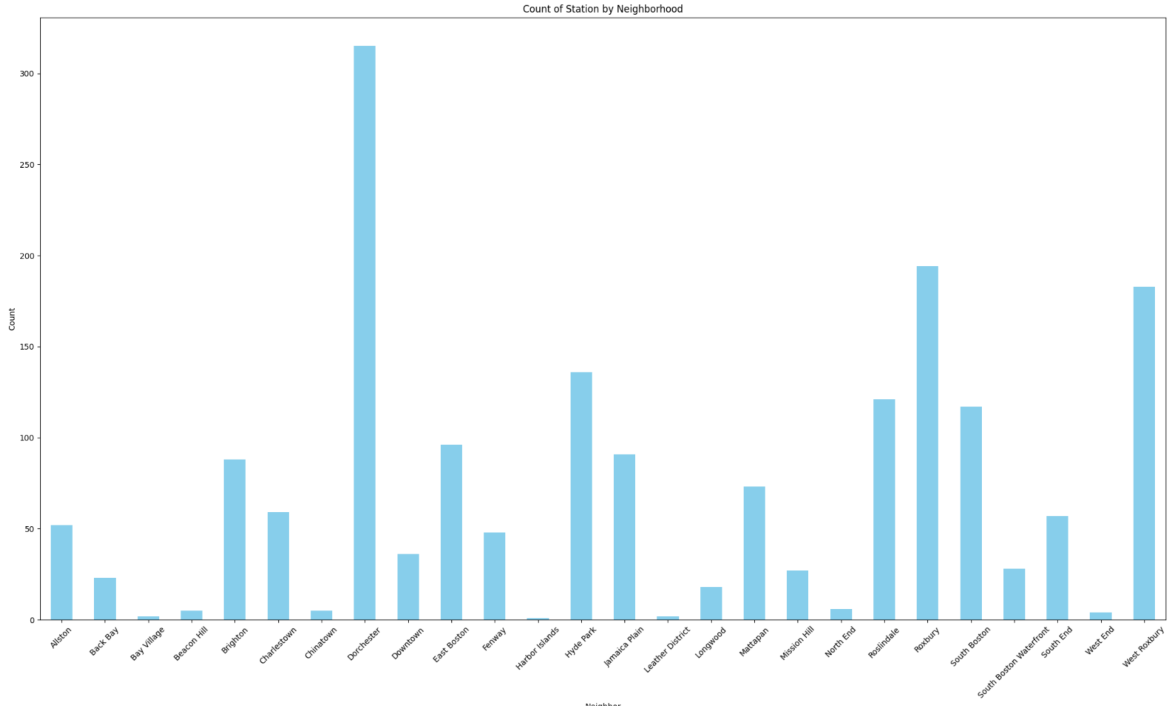
In terms of methodology, it is impractical to find community information for all different bus routes, because there are more than 100 bus routes, and each route can go across various communities. Instead, investigating the on-time performance of bus stops in each community can give a more insightful outcome. For each stop, by plotting the on-time performance on the map of Boston, we see a pattern that almost all stops that are located in central Boston have a very low on-time rate, and for stops that are far from central, the on-time rate is much higher. This finding further attested the answer to Question 2.





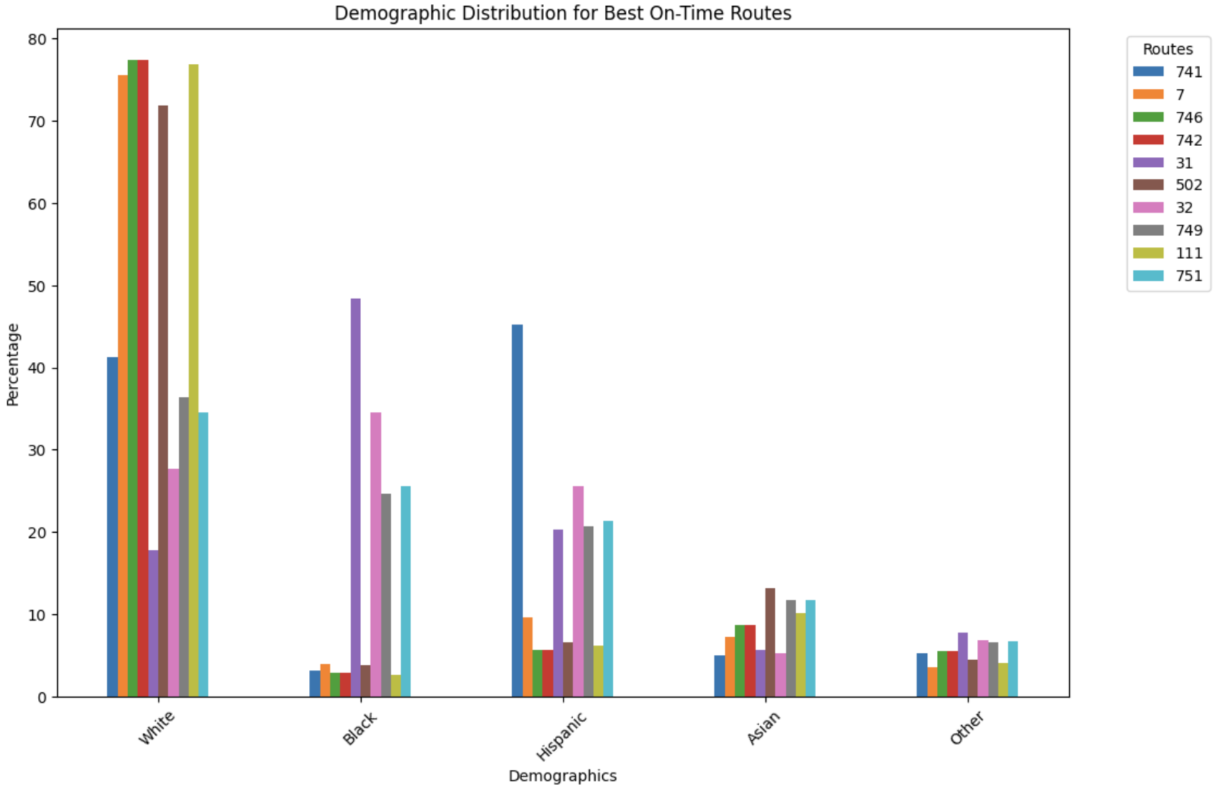
We also generated bar plots on how many routes and stations each community has, so in further analysis, we can check whether each community is evenly treated with bus routes.


Count of Unique Routes by Neighborhood
Count_Route
0
5
10
15
20
25
30
35
Allston
Back Bay
Bay Village
Beacon Hill
Brighton
Charlestown
Chinatown
Dorchester
Downtown
East Boston
Fenway
Harbor Islands
Hyde Park
Jamaica Plain
Leather District
Longwood
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury
Neighbor

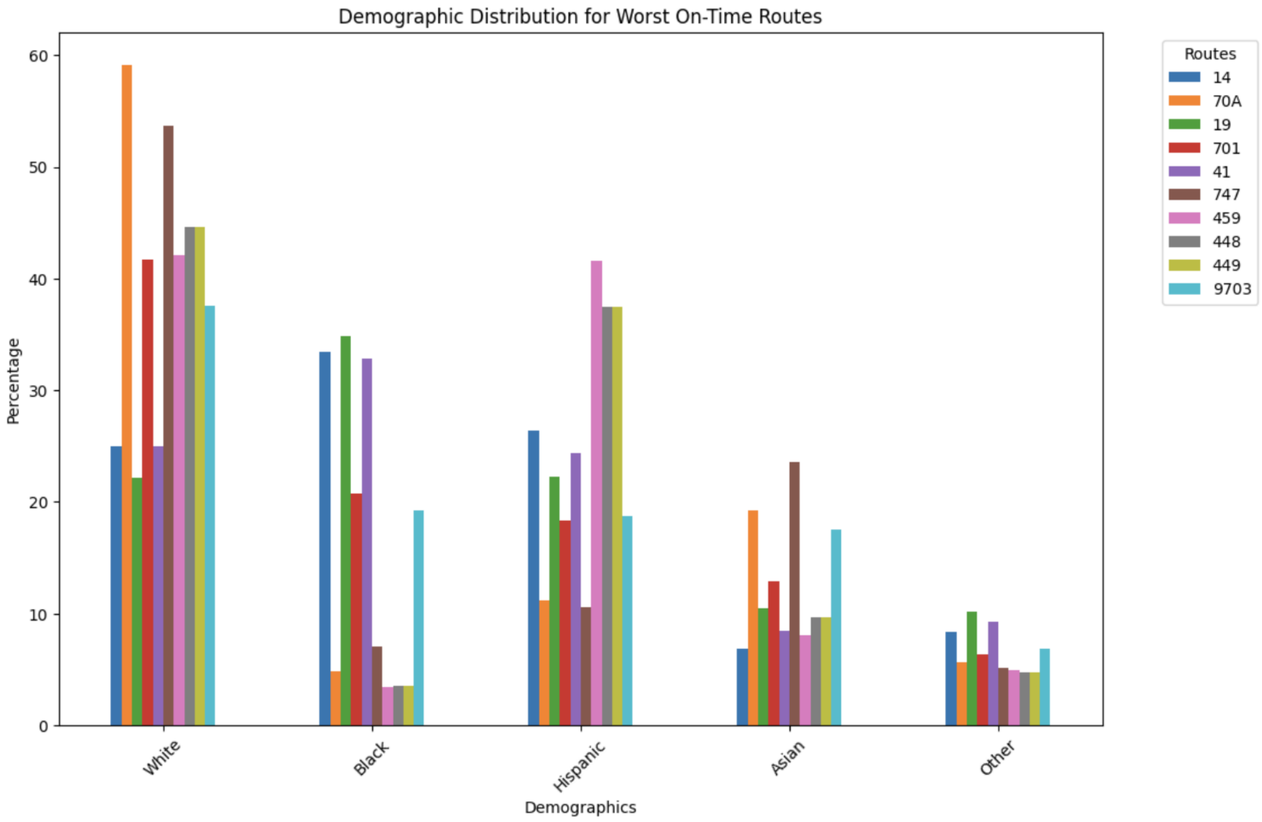

Count of Station by Neighborhood
Count
0
50
100
150
200
250
300
Allston
Back Bay
Bay Village
Beacon Hill
Brighton
Charlestown
Chinatown
Dorchester
Downtown
East Boston
Fenway
Harbor Islands
Hyde Park
Jamaica Plain
Leather District
Longwood
Mattapan
Mission Hill
North End
Roslindale
Roxbury
South Boston
South Boston Waterfront
South End
West End
West Roxbury
Neighbor

2.4. Question 4: Are there differences in the characteristics of the people most impacted by the disparity of bus service?

We used the MBTA’s GTFS map data to identify neighborhoods of each station of each route, using those neighborhoods to identify characteristics on the census data. Using the ten worst routes from the previous base question, we graphed the results below (Figure 1.1).



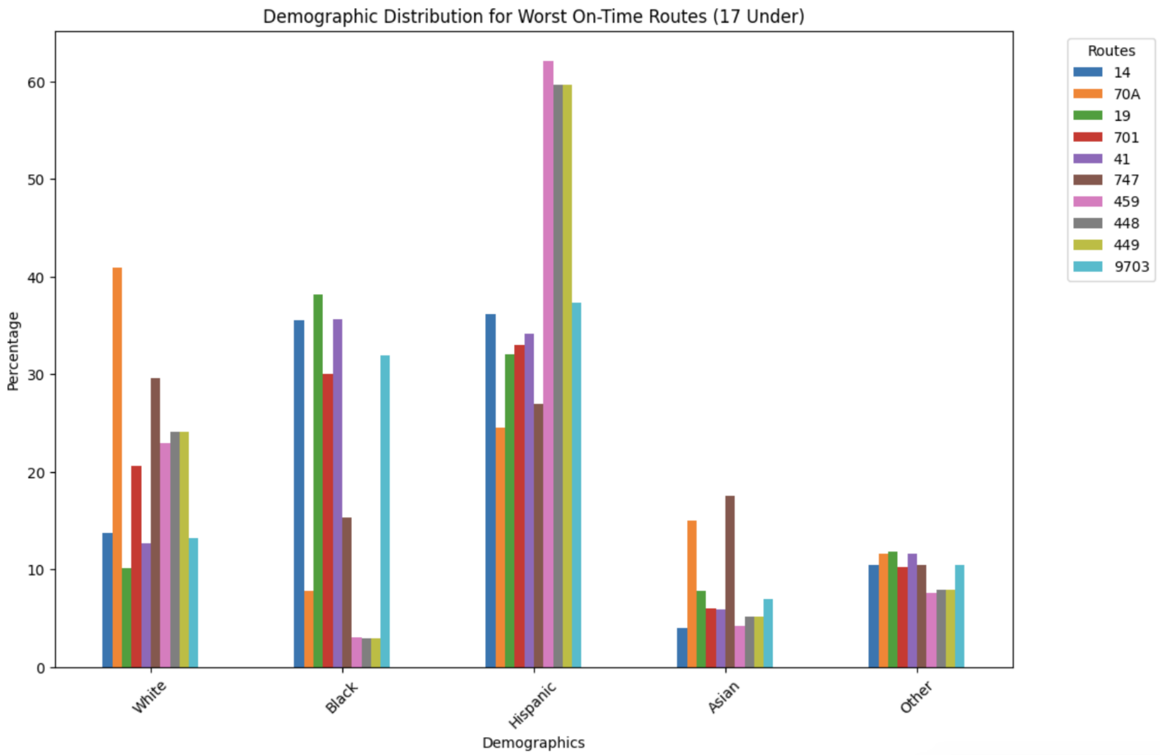
Graph representing the demographic distribution for the top ten worst on-time routes (Figure 1.1)



Graph representing the demographic distribution for the top ten best on-time routes (Figure 1.2)

NOTE: all the graphs under 1.0 have an x-axis that represents the percentage and the y-axis represents the demographics.

We also graphed the demographic distribution for the best on-time routes to provide a contrast (Figure 1.2). As you can observe in both graphs, the demographic distribution is predominantly whites alone. Then the order proceeds as follows: Hispanics, Blacks, Asians, and then Other. It can also be observed that the white population is significantly higher on the best routes compared to the worst.



Graph representing the demographic distribution for the top ten worst on-time routes (17 under) (Figure 1.3)

We extracted and graphed the 17 under-demographic distribution. The Hispanic presence in the 17 and under community seems to be drastically higher, with the White-only population being drastically lower compared to the adult data.

The juvenile facility rate for the top ten best on-time routes was calculated. The average juvenile facility rate among the ten worst on-time routes averaged out to 0.0043, whereas the best routes averaged out to 0.0036, both of which are below 1%, albeit a minor difference where the best routes had a lower juvenile facility rate.

We also proceeded to look at the university rates as it was available in the census data. The average rate for the ten worst routes was calculated to be 0.151, compared to the 0.026 rate for the ten best routes. A potential explanation for this may be higher urban density in neighborhoods with high university rates, where issues like traffic congestion and longer onboarding and offboarding times arise.

2.5. Question 5: Which neighborhoods or groups of people are served better/worse by the MBTA bus system? Which routes are better/worse?

In terms of bus punctuality, people along the bus routes with higher OT rates benefit from the reliability of the riders of the bus routes with lower OT rates. If riders are more sure about the bus schedule, they could plan their time better, minimizing the waste of time waiting for a late bus.

For some neighborhoods, there may be many bus routes passing through, but not many stops are along the route in the region of a certain neighborhood. Further information on the population around a bus stop is needed. As we only have the neighborhood as a geographical unit census, it might be too broad to conclude on the number of population or the specific worst-off or better-off for particular demographics of residents.

3. Extension Analysis

Based on the results of our base questions, we concluded that there is a significant difference between different bus routes. The wide range of bus on-time performance may affect residents' commute choices. For example, shared bikes such as Bluebikes is one of the other kinds of public transportation that is considered sustainable, and the usage of Bluebikes has been increasing over the years. We wonder how the bus's on-time performance and the demand for each bus route are associated with the usage of Bluebikes stations along the bus routes. With a better understanding of how two types of transportation relate to each other and how users would behave concerning the bus's on-time performance, MBTA can better adjust the plan of bus lines and improve the on-time performance to attract more users. We defined three necessary questions to address the correlation between Bluebikes and MBTA bus performances: Where are Bluebikes station hotspots compared to bus stops, what are the average number of trips and trip durations from stations going along the best and worst bus routes, and how does the intensity of Bluebikes usage associate with the accuracy of bus trips?

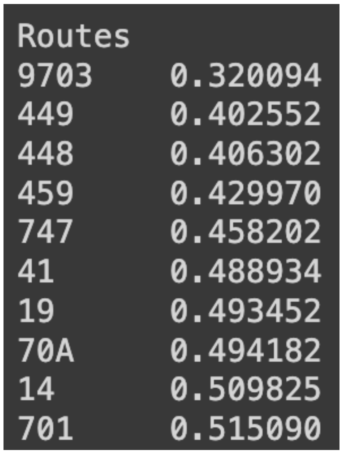
3.1. Data Processing

The demand for buses is defined as the number of people that get on the bus at each stop. The demand for Bluebikess in an area is defined as the number of shared bikes taken out or undocked from a Bluebikes station within that area. The nearby bike stations are defined as those located within a 10-minute walk from any of a given route’s various bus stops. It is imperative to define what datasets we used: Bluebikess’ Tripdata for January 2022, Bluebikess’ Current Stations, Census Boston Neighborhood, MBTA Bus and Rapid Transit Reliability, MBTA Systemwide GTFS Map, and MBTA Bus Prediction Accuracy datasets.

The process was conducted as follows: The Bluebikess’ Tripdata dataset in its raw format contained the trip duration in seconds, but to align it more meaningfully with the rest of our project data exploration, we altered this to be in minutes. The Bluebikess’ current stations dataset had valuable location data in the format of longitude and latitude, but the stations only contained names and not IDs, which may make it difficult to use as primary mode of station identification. Given we had the ID values of the bike stations in the Bluebikes trip dataset, we matched stations with their respective IDs. MBTA GTFS Map dataset initially contained a string of routes per bus stop, which we converted to each row representing a route. MBTA Bus Reliability dataset processing was done in the exact same way conducted in base question 2, calculating the on-time performance (OTP) of the bus routes by taking numerator over the denominator followed by each route’s mean. From there we merged the GTFS and Reliability datasets to have each route’s location data from GTFS and the performance from the Reliability dataset. From there, we took the best and worst OTP routes and used the Haversine formula converting longitude and latitude data to distances in Km to find all Bluebikes stations within 10 minute walking distance from bus stops along each of the buses’ routes. Those were all stored in a mapping of bus route ID to list of Bluebikes station ID. This set up the foundation to conduct our extension proposal.

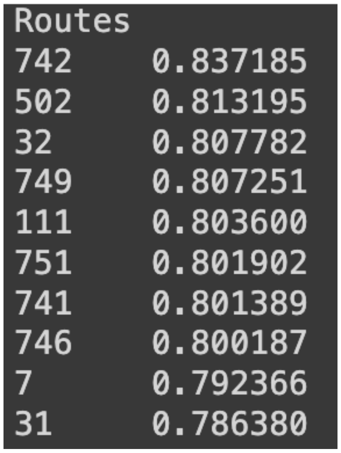
3.2. Results

In finding the relationship between alternative transportation methods (Bluebikess) and MBTA bus performance, we developed a hypothesis that Bluebikes station locations have a positive correlation with bus on-timeness. After preprocessing the data as mentioned above, we were able to derive the ten best and worst MBTA bus routes.



```
Routes
9703    0.320094
449     0.402552
448     0.406302
459     0.429970
747     0.458202
41      0.488934
19      0.493452
70A     0.494182
14      0.509825
701     0.515090
```

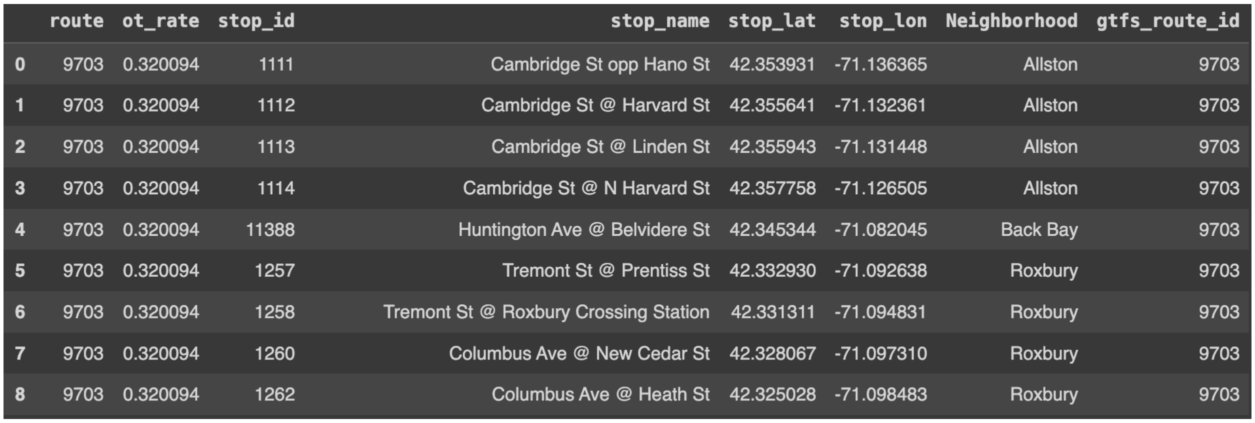
Top ten worst routes and its on-timeness rate (%)



```
Routes
742     0.837185
502     0.813195
32      0.807782
749     0.807251
111     0.803600
751     0.801902
741     0.801389
746     0.800187
7       0.792366
31      0.786380
```

Top ten best routes and their on-timeness rate (%)

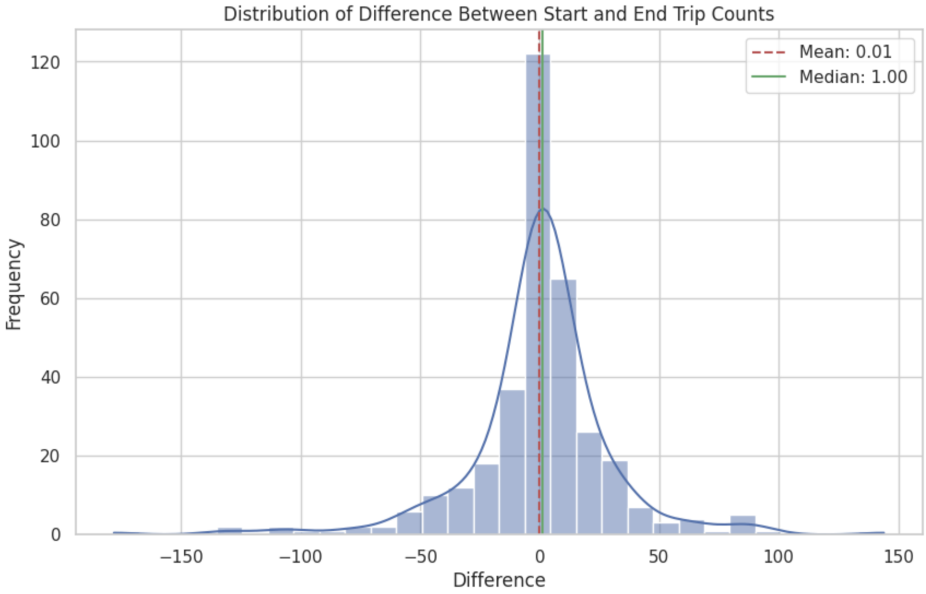
From then on, we merged the Bluebikes dataset in, identifying every bike stop within a ten-minute walk within an MBTA bus station along each of the best and worst routes.



	route	ot_rate	stop_id	stop_name	stop_lat	stop_lon	Neighborhood	gtfs_route_id
0	9703	0.320094	1111	Cambridge St opp Hano St	42.353931	-71.136365	Allston	9703
1	9703	0.320094	1112	Cambridge St @ Harvard St	42.355641	-71.132361	Allston	9703
2	9703	0.320094	1113	Cambridge St @ Linden St	42.355943	-71.131448	Allston	9703
3	9703	0.320094	1114	Cambridge St @ N Harvard St	42.357758	-71.126505	Allston	9703
4	9703	0.320094	11388	Huntington Ave @ Belvidere St	42.345344	-71.082045	Back Bay	9703
5	9703	0.320094	1257	Tremont St @ Prentiss St	42.332930	-71.092638	Roxbury	9703
6	9703	0.320094	1258	Tremont St @ Roxbury Crossing Station	42.331311	-71.094831	Roxbury	9703
7	9703	0.320094	1260	Columbus Ave @ New Cedar St	42.328067	-71.097310	Roxbury	9703
8	9703	0.320094	1262	Columbus Ave @ Heath St	42.325028	-71.098483	Roxbury	9703

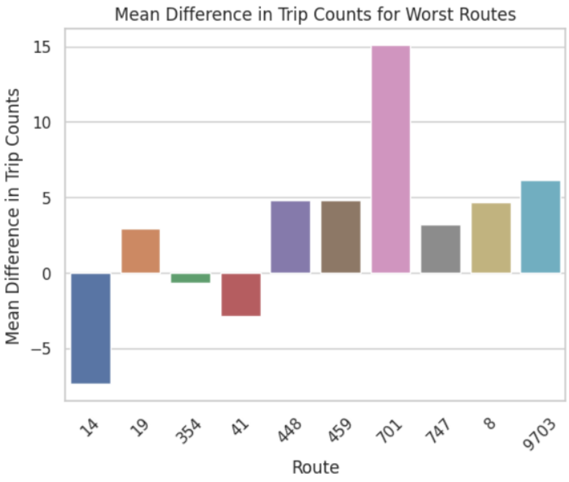
Bluebikes stations along the MBTA bus route "9703"

From the Bluebikes data, we had information on the number of dockings and undockings for each station throughout every period. We were able to create a new column that represented the net dockings for each station throughout the month. The difference between the start and end trip is calculated by count of the taking a bike from the dock and returning a bike to the dock. Below is a graph representing the distribution of docking differences.

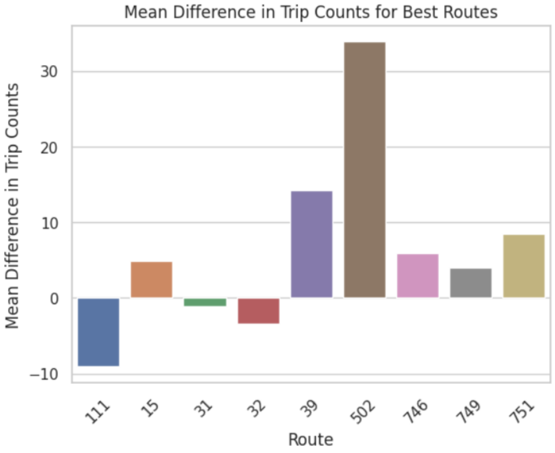


Both the mean and median are close to zero, indicating that they both accurately represent the average among this dataset. Although the average is close to zero, there are many stations where there are net positive and negative values. After observing this, we decided to investigate deeper as to a comparison between the best and worst routes.

Using the derived best and worst routes from the preprocessed data, we calculated the net dockings for the best and worst routes, which are graphed below.



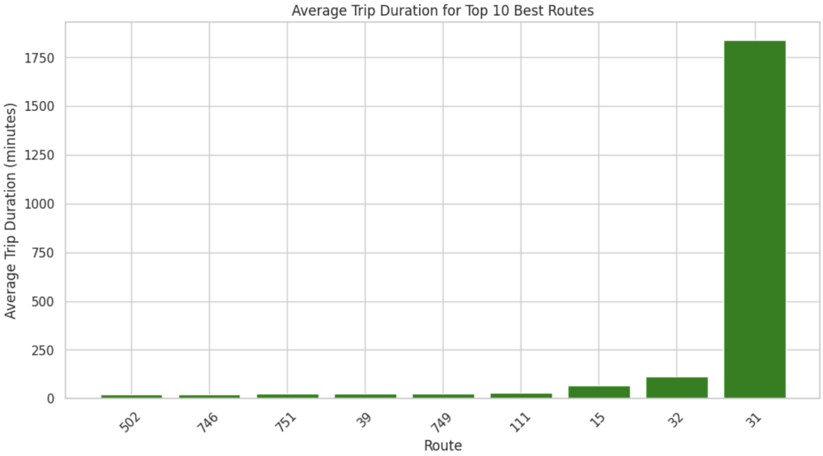
Net dockings for the ten worst routes



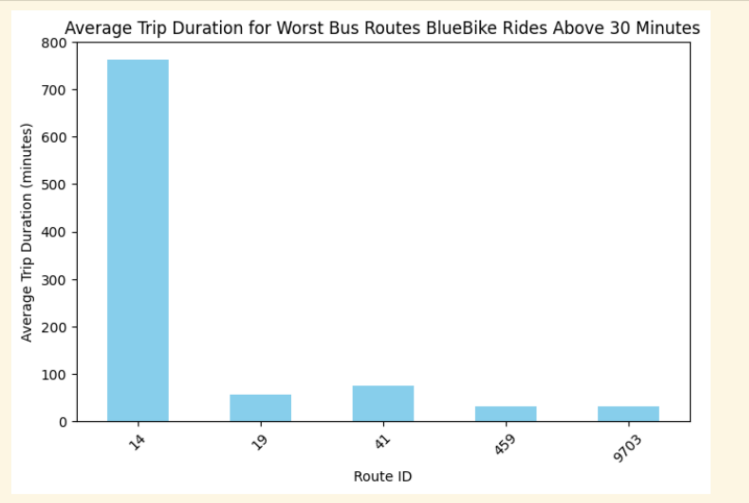
Net dockings for the ten best routes

We can observe that the mean difference in trip counts seems to be larger for the best routes compared to the worst routes. Based on this observation, we can draw a potential correlation between bus route on-timeness and Bluebikes station location, since the best routes seem to have more dockings than undockings.

We also identified another potential factor that might affect the relationship between bus performance and Bluebikess. We were able to extract the end-to-end trip times for Bluebikess, deriving them from the start and end times from the Bluebikes dataset. This was done for each station, where we developed a new dataset that represented the average usage for each Bluebikes station. We then looked at the discrepancy between the best and worst routes in terms of average trip times.

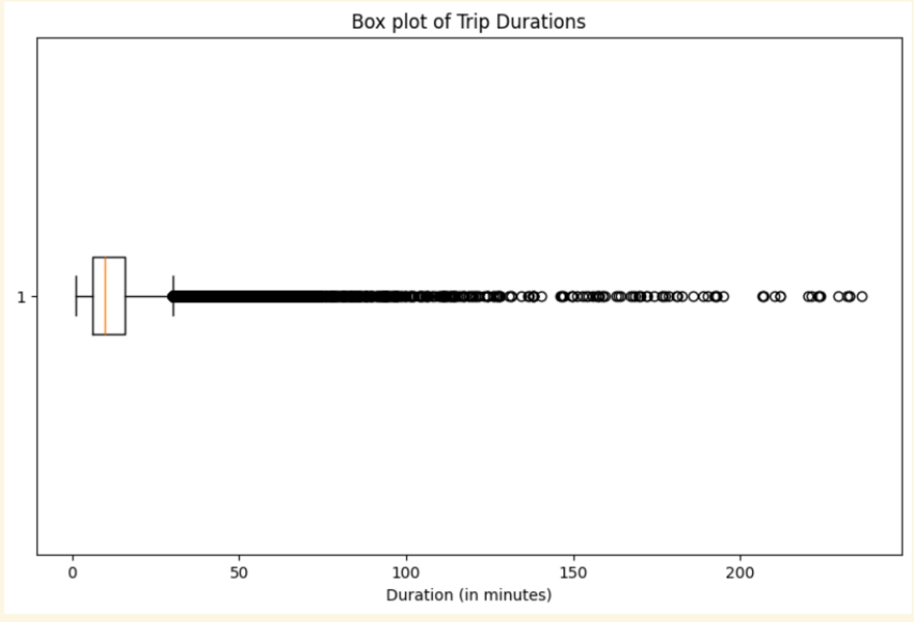


Trip duration for the ten best routes



Trip duration for the ten worst routes

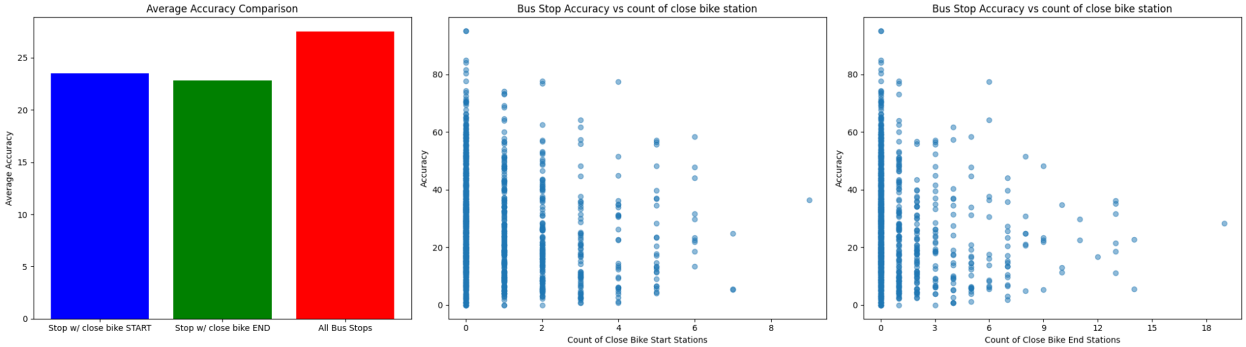
We can observe that the averages for the bus routes are significantly high, which is most likely an account of outliers that represent abnormally long trips. You can see that from the ten worst bus routes, five of them have average trip durations of more than 30 minutes. To investigate this, we graphed a boxplot of trip durations to get a better visualization of the outliers.



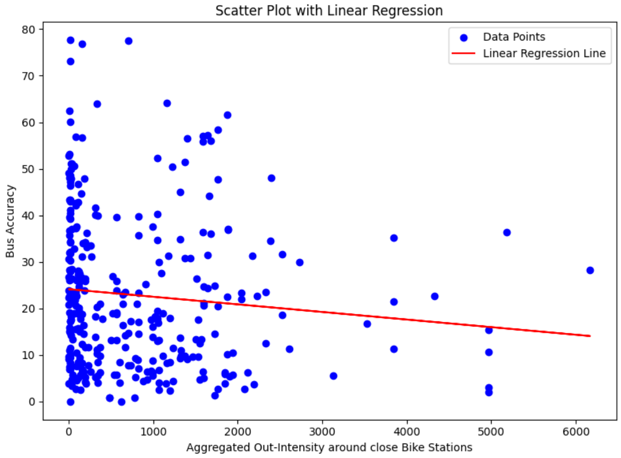
Box plot of trip durations for each Bluebikes station

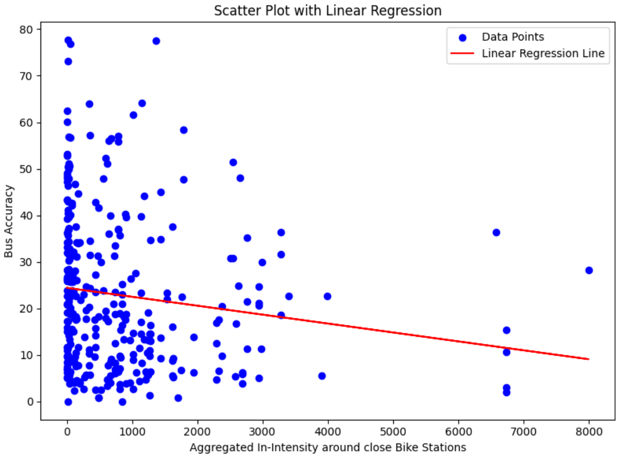
There seems to be a significant number of outliers; however, they are of little significance compared to the median which lies between 10-15 minutes. We did not exclude this from our data because users who use Bluebikess for hours may be people who did not correctly dock their bikes or maybe explorers of the city taking their bikes along with them rather than docking them.

Lastly, we also investigated what is the correlation between the number of Bluebikes stations around a bus stop and the bus stop's on-time rate, and we found the following result:

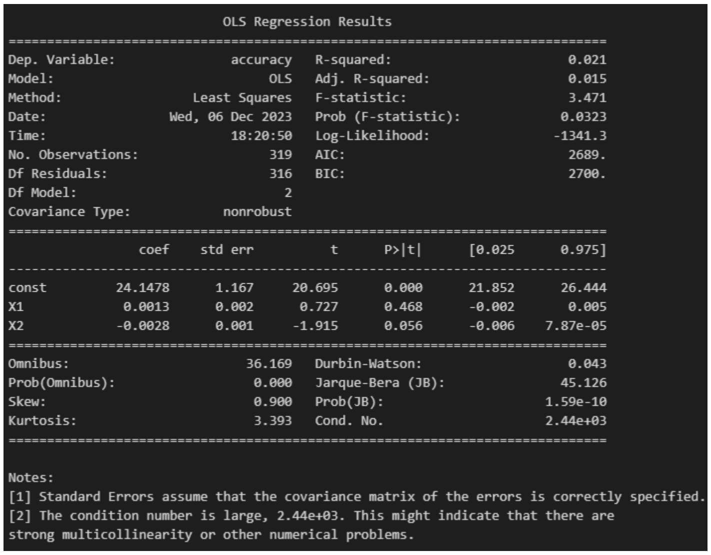


For bus stops that are at least having 1 Bluebikes station within a 5-minute walk (about 500 meters), the stop's average on time rate is slightly lower than the overall bus stops', and it seems that as the bus stop has more Bluebikes stations around it, its on time rate will gradually decrease. To get a better understanding, we define the in/out intensity of a Bluebikes station as the number of rides in/out that station within one month, and we draw a scatter plot with a best-fit line:





We also did an MLR to regress the on-time rate to both aggregated in-intensity and out-intensity of each bus stop. The result shows that the R^2 is a very small number, indicating a concerning fitness for data, because a good model usually has high R^2. Additionally, the 95% CI does not exclude 0 for one of the slopes. This means that the Bluebikes' intensity around a bus stop may have no linear correlation with the bus stop’s on-time performance.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               accuracy   R-squared:                       0.021
Model:                            OLS   Adj. R-squared:                  0.015
Method:                 Least Squares   F-statistic:                     3.471
Date:                Wed, 06 Dec 2023   Prob (F-statistic):             0.0323
Time:                        18:20:50   Log-Likelihood:                -1341.3
No. Observations:                 319   AIC:                             2689.
Df Residuals:                     316   BIC:                             2700.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         24.1478      1.167     20.695      0.000      21.852      26.444
X1             0.0013      0.002      0.727      0.468      -0.002       0.005
X2            -0.0028      0.001     -1.915      0.056      -0.006    7.87e-05
==============================================================================
Omnibus:                       36.169   Durbin-Watson:                   0.043
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.126
Skew:                           0.900   Prob(JB):                     1.59e-10
Kurtosis:                       3.393   Cond. No.                     2.44e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.44e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

4. Conclusion

In conclusion, our analysis of the MBTA arrival departure dataset has provided valuable insights into the performance and characteristics of different bus routes. Through the examination of travel times, on-time performance rates, and demographic distributions, we were able to identify patterns and disparities within the MBTA bus system. The average end-to-end travel times varied between approximately 18 to 30 minutes, with outbound trips generally shorter than inbound trips. Disparities in on-time performance were evident, with most routes exhibiting rates between 0.6 to 0.7, indicating that buses are often late. Routes with the highest and lowest on-time rates were identified, revealing some discrepancies associated with route directions. The analysis of demographic distributions for routes with the best and worst on-time performance highlighted differences in the racial and ethnic demographics along the routes, with predominantly white populations benefiting more from the best routes. Additionally, the examination of juvenile facility rates and university rates showed minor differences between the best and worst routes. Overall, our findings suggest that improvements in bus service reliability could positively impact the communities served by the MBTA, especially in terms of better planning for commuters and minimizing the impact on vulnerable populations.

In the Extension Project, our comprehensive analysis aimed to explore the relationship between alternative transportation methods, specifically Bluebikes, and the on-timeliness of MBTA bus routes. We initially hypothesized a positive correlation between Bluebikes station locations and bus on-timeness. The best routes exhibit larger mean differences in trip counts compared to the worst routes. This observation suggests a potential correlation between bus route on-timeliness and Bluebikes station location, where the best routes tend to have more dockings than undockings. Furthermore, the investigation into average trip durations for Bluebikes stations uncovered outliers, possibly reflecting users who keep bikes for extended periods. Although these outliers did not significantly impact the overall average, they were retained in the data to maintain a comprehensive representation of Bluebikes usage.

Additionally, we explored the correlation between the number of Bluebikes stations around a bus stop and the bus stop's on-time rate. Surprisingly, bus stops with at least one Bluebikes station within a 5-minute walk exhibited a slightly lower average on-time rate than other stops. Further analysis involved defining the in/out intensity of Bluebikes stations and conducting a multiple linear regression, revealing a small R^2 value and a 95% confidence interval that includes 0. This suggests that Bluebikes station intensity may not have a linear correlation with bus stop on-time performance.

In summary, our investigation provides valuable insights into the complex relationship between Bluebikes usage and MBTA bus on-timeness. While patterns suggest a potential connection, the presence of outliers and the nuanced nature of this correlation necessitate further exploration and consideration of various factors influencing public transportation performance.

5. Future Improvements and Limitations

One of the limitations is that the Census dataset only analyses per neighborhood, which is too large a unit to be able to conduct any analyses on a bus stop or block-wide level. This was the primary hurdle preventing us from proceeding with our initial proposal for the extension project. Initially, we wanted to study the usage of each stops, associating them with the neighborhoods they are in. After looking at answering questions 3,4 and 5, it seems that the conclusion drawn are associated with the geographical size of neighborhood. We noticed that the neighborhood, as a geographical unit, is too big to eveluate the impact of bus stops to their surrounding communities. Narrowing down the scope of geo units into smaller units, such as blocks, and analyzing the stops instead of routes would allow us to determine the influence of bus stops better. However, However, we encountered challenges to access the data for stops. There was data on the usage of the underground stations for the trains, but not for bus stops. If similar data is provided for the bus stops in the future it might be possible to answer the question.

Additionally, it was challenging to easily merge or connect Census neighborhood data with any of the MBTA datasets originally given in the project document. This was resolved by finding the MBTA GTFS dataset that helped bridge the gap between MBTA performance datasets and the Census neighborhood dataset.

Last but not least, our analysis on the association of bus on-time performance and demographics are only based on the 10 routes with the highest and lowest on-time performance. Since the the top and bottom performance routes are just 20 out of 600 routes, there might be loss of information when visualizing their relationship. This indicates the need of improved methodology to explore the associations.

6. Appendix

6.1. Additional Materials

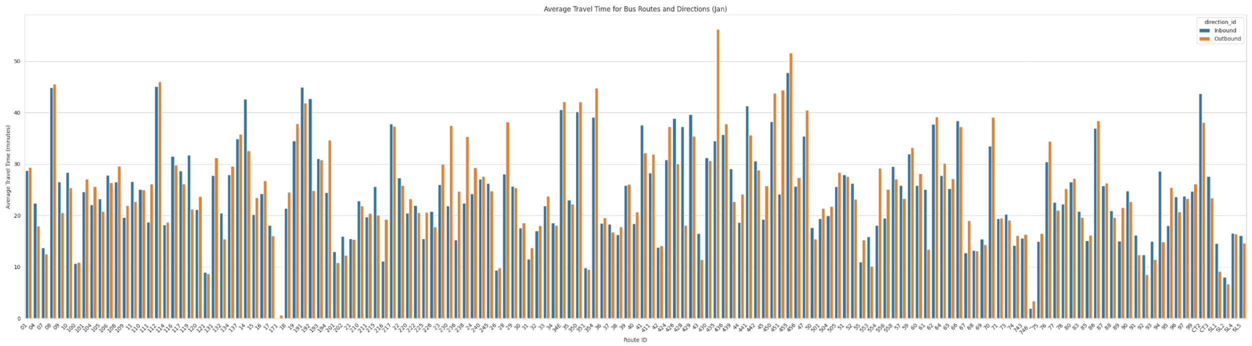


Figure 1

6.2. Individual Contribution

6.2.1. Salma Alali:

- Report: Responsible for writing base question 1's analysis, and the data processing of the extension proposal.
- Worked on base question 1: Helped come up with (along with teammate Isaac) process to answer the question and the idea to further narrow the objective by creating a graph showing inbound-outbound time difference of greater than 10 minutes.
- Worked on base question 4: Helped come up with (along with teammate Isaac) process on how to address the question.
- Contributed to extension project:
  - Helped define questions that will tie back our extension project to the original project.
  - Worked on extension project question 2 (with teammate Isaac): Conducted all data cleaning and feature extraction and quantifying definition of "close bike stations near bus stops".
- Logistical contribution:
  - Made all (excluding one) github repository pull requests involving all team members' work.
  - Organizer of meetings and ensured teammates were up to date and progressing towards each deliverable.
  - Submitter of all GradeScope assignments

6.2.2. Isaac Chan:

- Report: Responsible for base question four as well as extension project results
- Worked on base question 1: Developed the code and answered the question by drawing potential hypotheses so that we can relate data with each other
- Worked on base question 4: Developed (along with teammate Salma) the code to answer this base question. Came up with several hypotheses in terms of university rate and juvenile rate in correlation with on-timeness.
- Contributed to extension project:
  - Helped define questions that will tie back our extension project to the original project.
  - Worked on extension project question 2: Conducted analyses and developed and implemented data extraction (along with teammate Salma)
- Logistical contribution:
  - Updating both the discord and iMessage group chats to make sure the messages are relayed

6.2.3. AlHasan Bahaidarah:

- Report: Responsible for base question five as well as extension project stop location analysis
- Wrote the majority of the report in order to cut down on time needed to write it before the deadline
- Worked on base question 5: Developed the code and answered the question by drawing potential hypotheses so that we can relate data with each other
- Contributed to extension project:
  - Helped define questions that will tie back our extension project to the original project.
  - Worked on extension project question 1: Conducted analyses and developed and implemented data extraction
  - Helped the team graph some of the data that was relevant in coming up with the answers
- Logistical contribution:
  - Found rooms for us to hold meetings in
  - Explained the work we needed to do along the way of the project

6.2.4. Jasmine Fanchu Zhou

- Conducted preliminary research on the proposed business question
- Answered and analyzed base question 2 to get the list of bus on-time rate, and researched each bus route's location.
- Contribution to extension project:
  - Initiated the conversation on rescoping the extension projects, came up with specific research questions, and proposed the idea on the Bluebikes study
  - Rewrote introduction, data, limitation sections in final report from deliverable 2.
  - Wrote conclusion.
- Logistic contribution:
  - Attended all team meetings, and shared ideas with the team
  - Organized Google Folders and files, making sure folders look clean, organize and easy to locate.
  - Reviewed, and formatted slides and reports. Make professional templates and outlines for extension project presentations and final reports. Format slides including revising text, adjusting images, checking content flow and consistency, making table content, etc.

6.2.5. Tiansui Gu:

- Report: Responsible for base question 3 and helped analysis on base question 2, as well as extension project bus stop on-time rate correlation analysis
- Help on base question 2: Processed the data and generated more detailed visualization on bus route/stop on time rate and presented in the presentation
- Worked on base question 3: Generated some basic visualization, and find relation with bus stop on-time rate with the stop's geographic location to get a better understanding of bus performance
- Contribution to extension proposal:
  - Proposed the idea on the Bluebikes study
  - Find possible correlation between bus on-time rate with aggregate count of Bluebikes station/ bike station in/out intensity
- Logistical contribution:

- Attend all group meetings actively and on time
- Share ideas during the group meeting