Going Places with Data: A Study of the MBTA

Tim Demling, Will Hanvey, Dan McCusker-Alvarez, Emma Sommers, Nick Perrotta, and Daniel Xu

Northeastern University, Boston, MA, USA

Introduction

In the city of Boston, public transit and the MBTA are services that over 34% of residents rely on every day to get to work, go to school, and visit neighboring areas. With the recent election of Boston's new mayor, talks of revising and improving Boston's MBTA system have rekindled, bringing into question how the MBTA currently functions to serve Boston's residents' needs. Our project was aimed at gaining a better understanding of how the MBTA's Orange Line performed as a whole last year and, more specifically, how timely it was on a daily basis. Along with this analysis, we hoped to further discern how the MBTA's performance could be affecting the various groups of people that reside in Boston and whether or not certain groups of people were being advantaged or disadvantaged by its service—if so, being able to understand which groups and which demographics need better service, would bring about suggestions on how the MBTA can make smarter and more inclusive improvements in the future.

Modeling

We pulled data from a large variety of sources. Our primary location was the MBTA's 2020 Rapid Transit Headways data, which contained data on every train leaving every station in 2020 [1]. We also used weather data from Weather Underground, general data from MBTA General Transit Feed Specification, and GeoJson and demographics data from Analyze Boston [2,3,4,5].

To prepare our data, we first started with general EDA and cleaning that involved fixing issues with some columns having NaNs and some columns having incorrect direction IDs. From there, we merged the dataframe with the GFTS data to make columns for the current stop, next stop, and prior stop [3]. Using this data, we were able to calculate the travel time for each train from its current stop to its next stop. We did this by creating smaller data frames based on date, then finding the train arriving at the next stop that departs the earliest after 45 seconds have passed from the next stop. This is long enough to ensure that we're not counting trains close to each other since trains are required to maintain a certain distance between them. Additionally, we added a cut-off point of about an hour, in case trains were taken out of service.

Now, with this data, we were able to create a pivot table to give us specific information about how trains departing each stop did in each direction on each day. We used the median instead of the mean for this, though, to avoid the influence of outliers. From here, we added a new column,

called ‘slow days,’ that we determined was any day (based on station and direction) where the mean was over 5% greater than the mean. Slow days would, then, give us a metric of if trains faced a high probability of delays on any specific day, departing any specific station, heading towards either Forest Hills or Oak Grove. Finally, we joined this dataframe with weather data from Weather Underground [2].

From here, our goal was to develop machine learning algorithms to predict the time between stations and whether delays were likely, given the departure station, arrival station, month, day of the week, temperature, and precipitation. We decided on a KNN model for the latter, which was able to predict the slow day flag with 99.5% accuracy in our test set. For the former, we used a Random Forest Regression model, which was able to predict the travel time between stations with an average error of only seven seconds in our test set. Finally, we implemented both of these models into an interactive python code that would allow a user to input the features and give an output based on them.

The purpose of the demographics code was to determine if the demographics of a neighborhood were correlated to the quality of service that the neighborhood received as indicated by the delay times experienced in a neighborhood. We decided the best way to implement working with the data was by creating a class for each sheet of the excel file we wanted to examine. The excel sheet was manually reduced to include only the statistics that could have genuinely meaningful relationships to train service. Additionally, sheets without uniform formatting were deleted. From there, we included methods to isolate a specific statistic and visualize it on a map of Boston neighborhoods. To create the map of Boston, we used the plot method of geopandas with a geojson file [4] that had the coordinates of each Boston neighborhood. Each neighborhood was colored based on the value of the statistic in the excel sheet. We then used contextly to add a base map of the city of Boston beneath the neighborhoods. In addition to these visualizations, we used the NumPy corrcoef method to quantify the correlation between these variables. This method uses the equation corrcoef(x, y) = cov(x, y) / (stddv(x) * stddv(y)). We then examined the most positively and most negatively correlated variables with delay times and visualized them as both a bar chart and line plot.

Use Cases

Our model, which measures how likely there is to be a delay and what that delay time will be, could be used for MBTA travelers. For instance, a traveler could indicate the desired destination stop and a departure stop, and our code could return whether a delay is expected and, if so, what the delay time is, as well as the average travel time. Ideally, this could be implemented into google maps or the MBTA’s own website to predict arrival times more accurately. There are, of course, limitations of the code at this time. One of those is the lack of live data, as our project uses only historical data from 2020. A solution could be using data from the last five years but increasing the weight of data from the past year when computing averages. While this is unfortunately only

available through the Python console right now, it would be interesting to turn this into a website in the future. And although the rest of our project may not be particularly application-oriented, our data and analysis pose important questions and considerations on decisions regarding the MBTA. Our visualizations about the delay times each neighborhood experiences and the demographic groups within those neighborhoods have the potential to, hopefully, enable future modifications and improvements along the orange line regarding those most disadvantaged by its current issues.

Analysis and Results

As can be seen in Appendix 1, “Daily Average Time Between Stops and 5 Coldest Days”, the top coldest days are often correlated with high delay times; this is true in January because of a snowstorm that happened. Further, the scatter plot shows a decrease in delay times in March that slowly picks up around August. We can assume that this is due to COVID-19. In March, the country went into lockdown, and many people started working virtually and thus did not use public transportation. This also confirms that when fewer people take the subway, it will run faster.

Our group also wanted to look at how direction affects delay times for each station because the data shows different numbers for the average delay time for each day at a given station for each direction. Our group was able to add a new column getting the difference between Forest Hills’ delay time and Oak Grove’s delay time. If the number was negative, that meant that going towards Oak Grove had a greater delay time. Our group then averaged out all the differences throughout the year for each station. This bar graph, seen in Appendix 2, shows how more stations expected a greater delay time going towards Oak Grove. Additionally, what also really sparked the group's interest was how there was over a 200 second greater delay time for the station Green Street for the Oak Grove direction. Finally, the other bar chart, seen in Appendix 3, shows how there are many more stations that have a greater delay time going towards Oak Grove, thirteen, while Forest Hills only has three.

Similarly, Appendix 4 shows that the train from Chinatown to Downtown Crossing has a higher spread of delay times. The number of instances of delays is higher than for the train from Ruggle to Roxbury Crossing, excluding outliers. Indeed, for Chinatown to Downtown Crossing, the number of instances is mainly between delay time of 80 and 95 seconds, with a few more instances between 100 and 110 seconds. The trips from Ruggles to Roxbury Crossing, on the other hand, are almost entirely between 90 and 100 seconds. This visualization reveals that the train is more unreliable when going from Chinatown to Downtown Crossing than when going from Ruggles to Roxbury Crossing. This also isn’t related to the actual distance between the stations, as both seem to take around the same amount of time.

Our group additionally wanted to see how both temperature and precipitation affected delay times. First, looking at the temperature and delay time graph, seen in Appendix 5, to our surprise, we saw that temperature had little effect on the delay time. Although some of the highest delay times were when the temperature was below 30, the regression line shows how there is only a small

correlation between temperature and delay time. Next, looking at precipitation and delay time which is shown in Appendix 6, the amount of precipitation had little to no effect on delay time as seen by the regression line as well. Only when there is a combination of low temperature and precipitation is present delay times could be greater. Looking at both Appendix 5 and 6, the two times where the delay time was greater than 600 was when the temperature was below 30 and precipitation was present.

For demographics correlation, we examined some of the most positively and negatively correlated variables. A positive correlation with delay times would indicate that an increase in % of that demographic in a neighborhood correlates with a higher average delay time. One of the most positively correlated variables was the percent of the neighborhood that had an average income over $150,000. This makes intuitive sense, as we've seen higher delays closer to downtown Boston, which houses wealthier Bostonians. More variables with high positive correlations (as shown in Appendix 7) include % with one vehicle, % white, and % who have a commute of 30-59 minutes. One variable with a large negative correlation was % income $25,000 to $34,999, which also follows from seeing fewer delays further from downtown Boston, where less wealthy Bostonians reside. Other variables with high negative correlation (shown in Appendix 8) include % no access to a vehicle, % Black/African-American, and % with a commute of 60+ minutes.

Conclusion

Overall, we're very pleased with the results of our project. The models we made were very accurate, and we were able to derive some very interesting results. However, we did have some initial goals that we were not able to achieve. On the modeling and statistical side, for example, we originally wanted to use all the MBTA lines instead of just Orange, but the scale of working with the Orange Line data was already very large, and the Red and Green line's branching meant that our calculated delay times for them were flimsy, at best. As a result, we felt that focusing our efforts on one line and receiving a very good result would be better than receiving less accurate results with more lines. Because of focusing on just one line, we were able to create an accurate model that was able to predict MBTA travel times, and we feel quite proud of our results.

For the demographics side of the project, we were successfully able to visualize and compute the correlation between the average orange line delays in a neighborhood and specific demographics of that neighborhood. There were generally weak correlations between the delays and demographic data, but this was to be expected since almost none of the demographics examined would have any causal relationship with MBTA delays.

Author Contributions

Since we had a group of six, we found it best fit to divide our group into two sections. Tim Demiling, Emma Sommers, and Daniel Xu focused on demographics, while William Hanvey, Dan McCusker-Alvarez, and Nick Perrotta focused on general analysis, statistics, and modeling of the MBTA data.

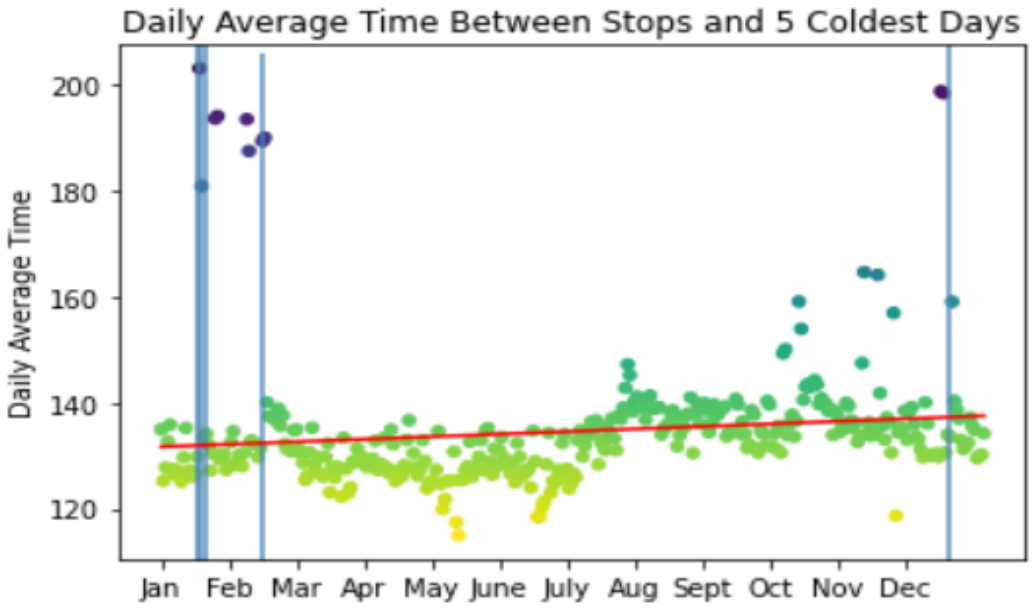
Tim Demling: Wrote correlation analysis methods, contributed to demographics analysis writeup
William Hanvey: EDA, building data frames, preparing data, building models, wrote large parts of write-up.
Dan McCusker-Alvarez: Prepared visualizations, interpreted visualizations for analysis of the data.
Nick Perrotta: Helped clean the data, dealt with weather data, created visualizations and the code for it and contributed to the project and report.
Emma Sommers: Wrote statistics class with Daniel, initial geopandas visualizations, and bar chart code. Helped with presentation and report
Daniel Xu: Did a lot of presentation work regarding demographics slides and overall aesthetic features. Wrote code to visualize the demographic relationships using geopandas and contextily and code that manipulated calculated delay times into usable data—also contributed to the write-up.

References

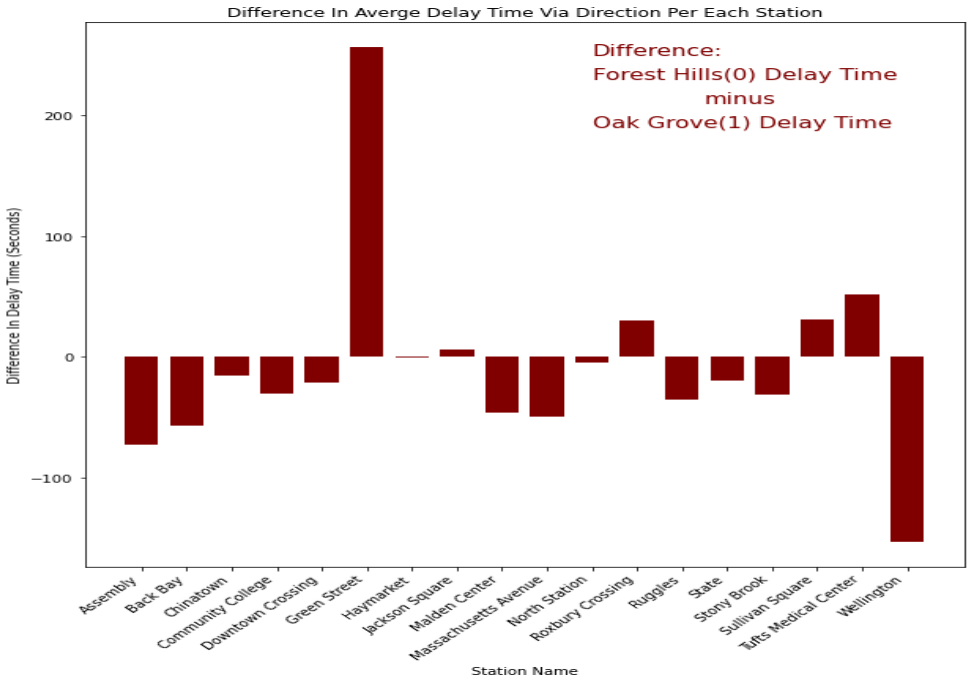
1. “MBTA Rapid Transit Headways 2020.” *MBTA Blue Book Open Data Portal*, 2021, https://mbta-massdot.opendata.arcgis.com/datasets/mbta-rapid-transit-headways-2020/explore.
2. “Boston, MA Weather History” Weather Underground, 2020, https://www.wunderground.com/history/daily/us/ma/boston/KBOS.
3. “GTFS.” *MBTA*, MBTA, https://www.mbta.com/developers/gtfs.
4. Analyze Boston. *Boston Neighborhoods* (GeoJson). Boston, MA. Boston.gov. 7 Dec 2020. https://data.boston.gov/dataset/boston-neighborhoods/resource/cc535415-d3a9-408d-aec4-4844ba3cc63b.
5. Analyze Boston. *Boston Neighborhood Demographics 2015-2019* (XLSX).. Boston, MA. Boston.gov. 23 Feb 2021. https://data.boston.gov/dataset/neighborhood-demographics/resource/d8c23c6a-b868-4ba4-8a3b-b9615a21be07.

Appendixes

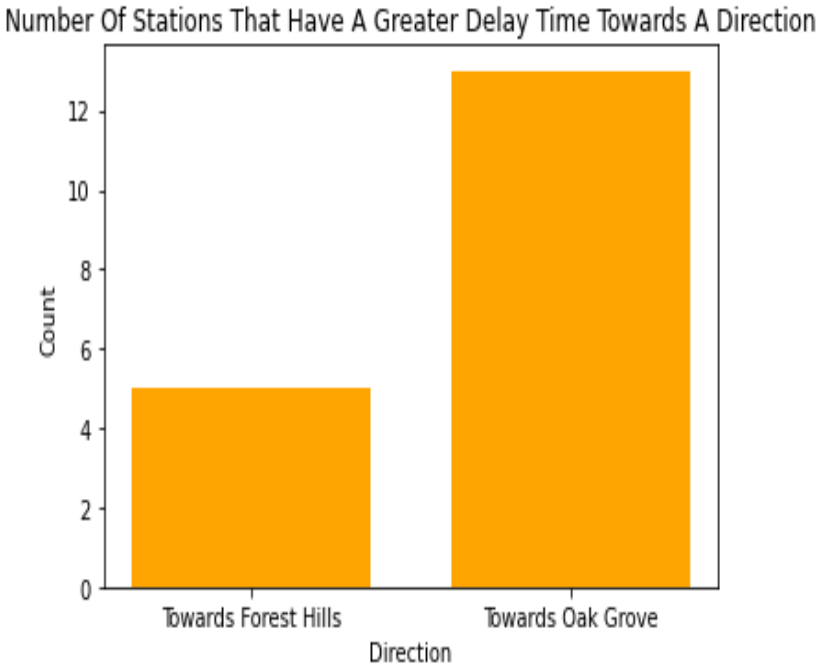
Appendix 1 - Daily Delays and Temperature



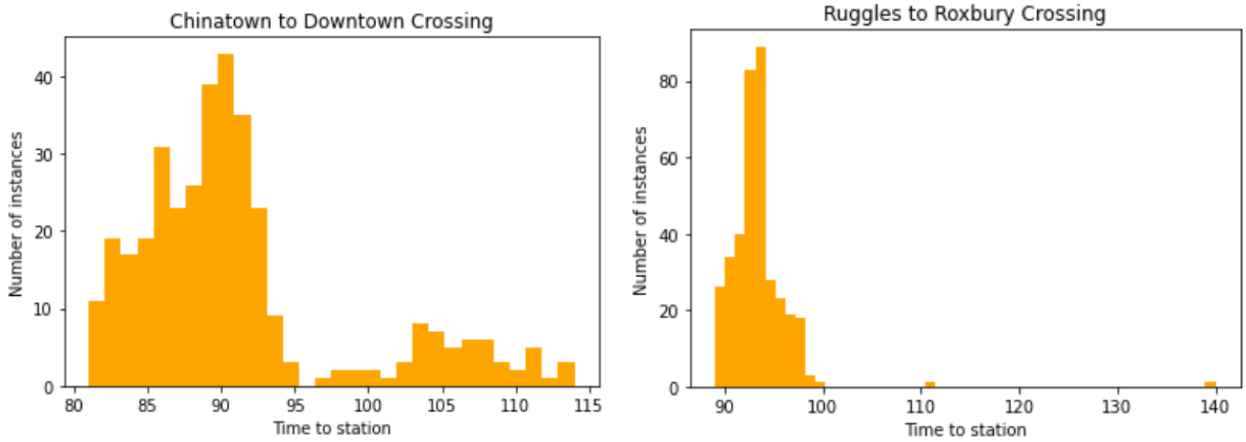
Appendix 2 - Difference in Delay Time between Directions



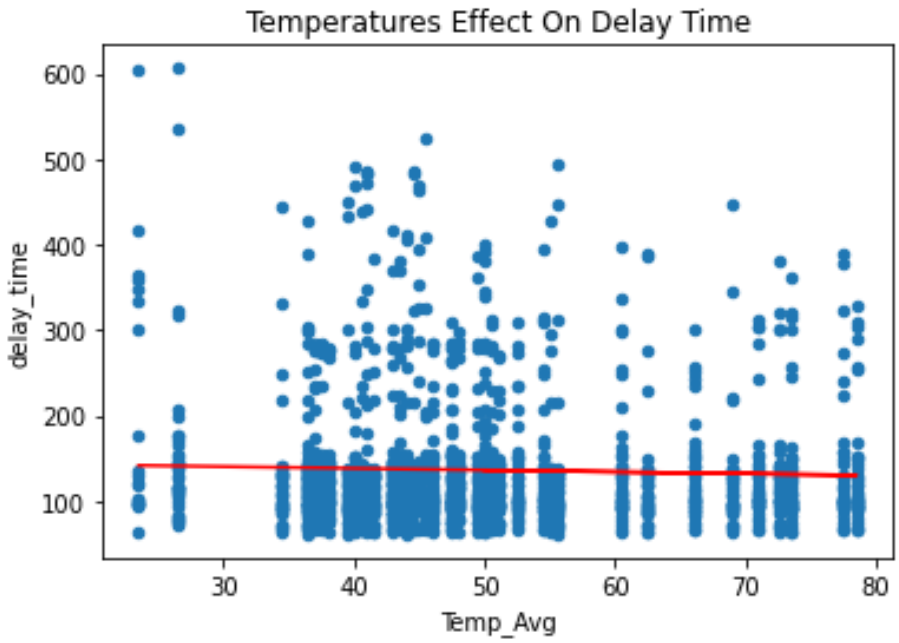
Appendix 3 - Directional Delays



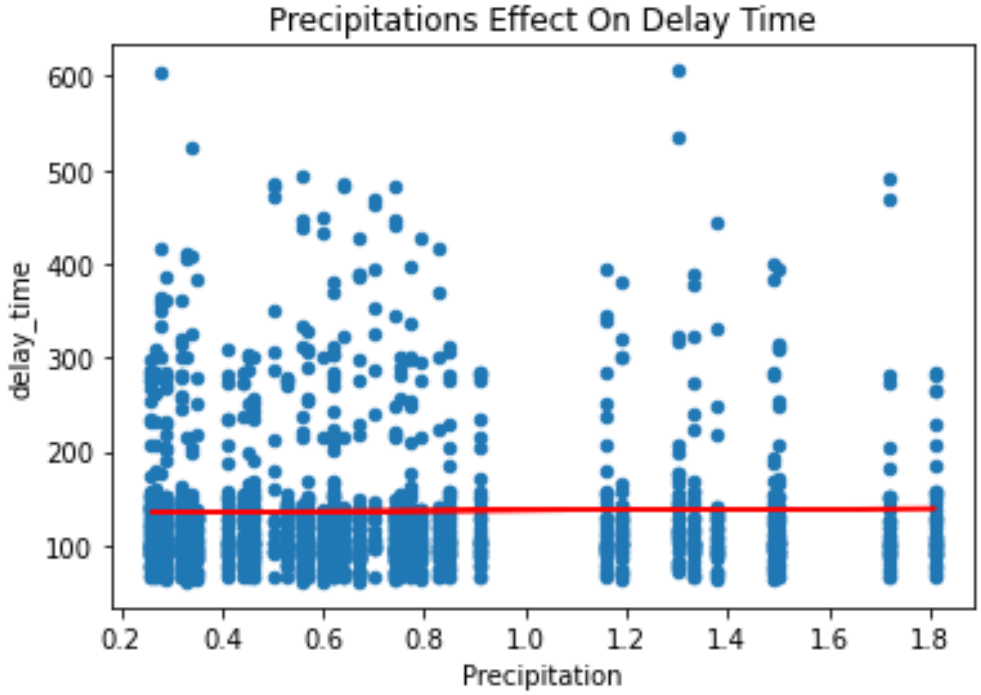
Appendix 4 - Spread of Delay Time Between Stations



Appendix 5 - Temperature's Effect on Delay Time



Appendix 6 - Precipitation's Effect on Delays



Appendix 7 - Demographics with the Highest Positive Correlation

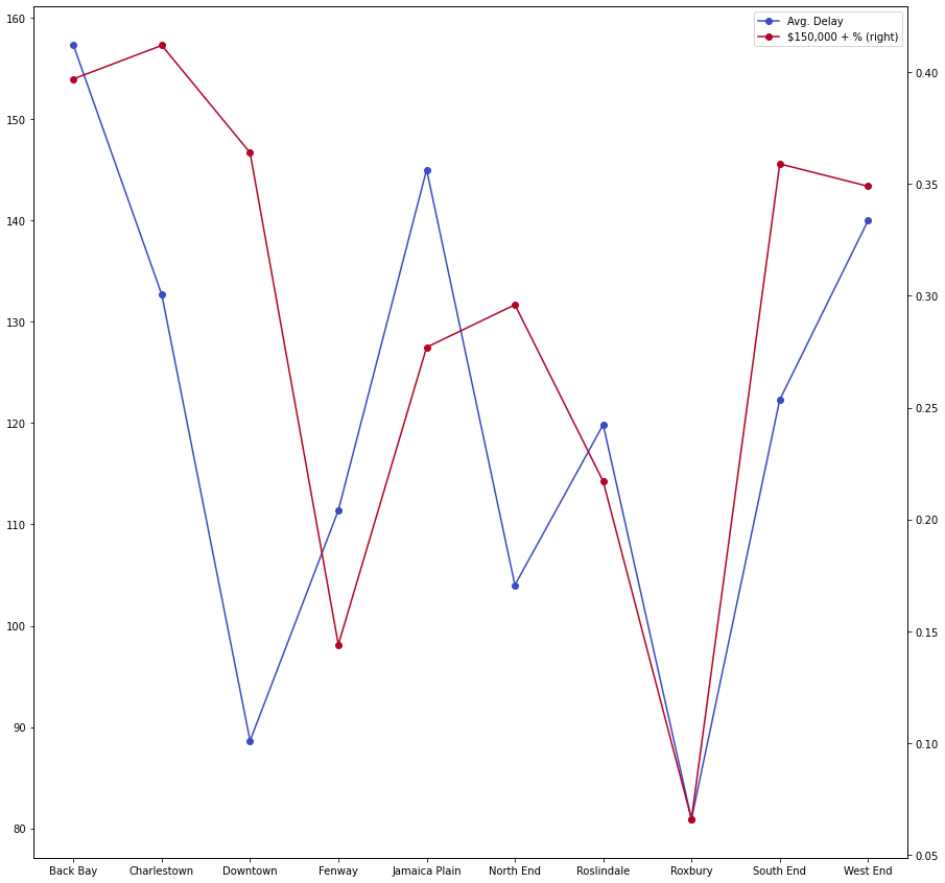


Fig. 1 Line Plot for Median Income $150,000 +

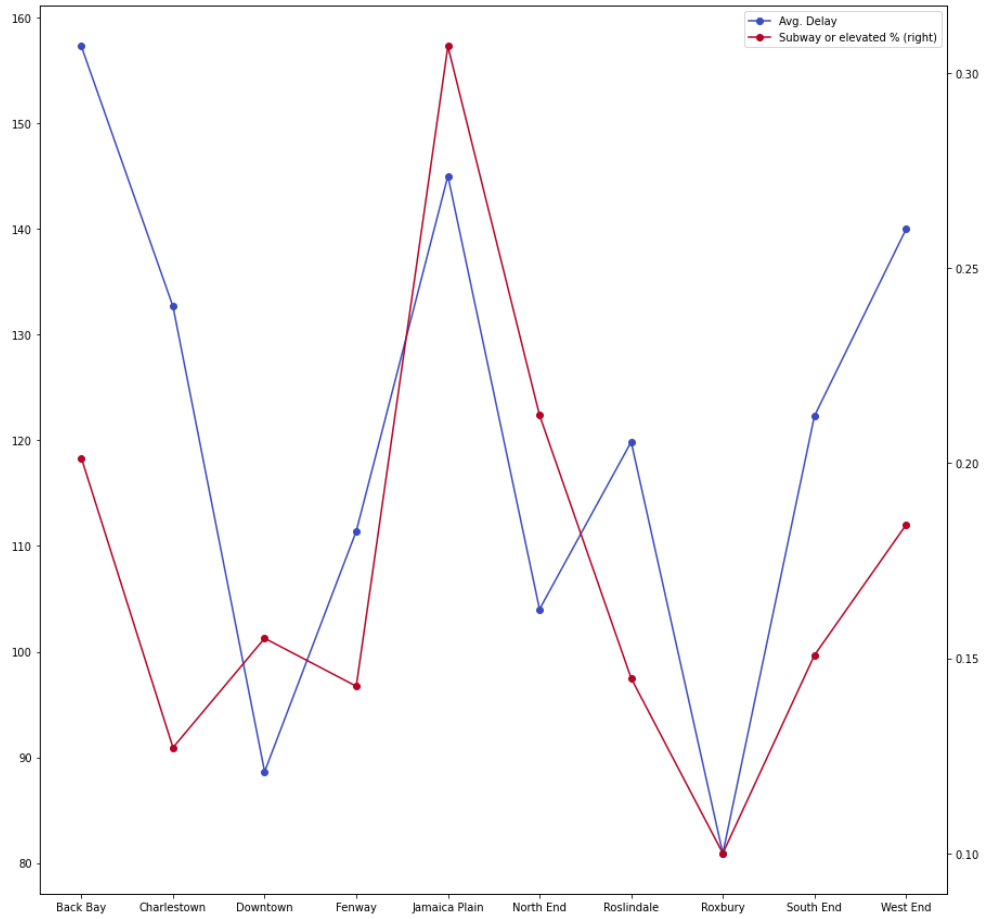


Fig. 2 Line Chart for Percentage of People Commuting by Subway

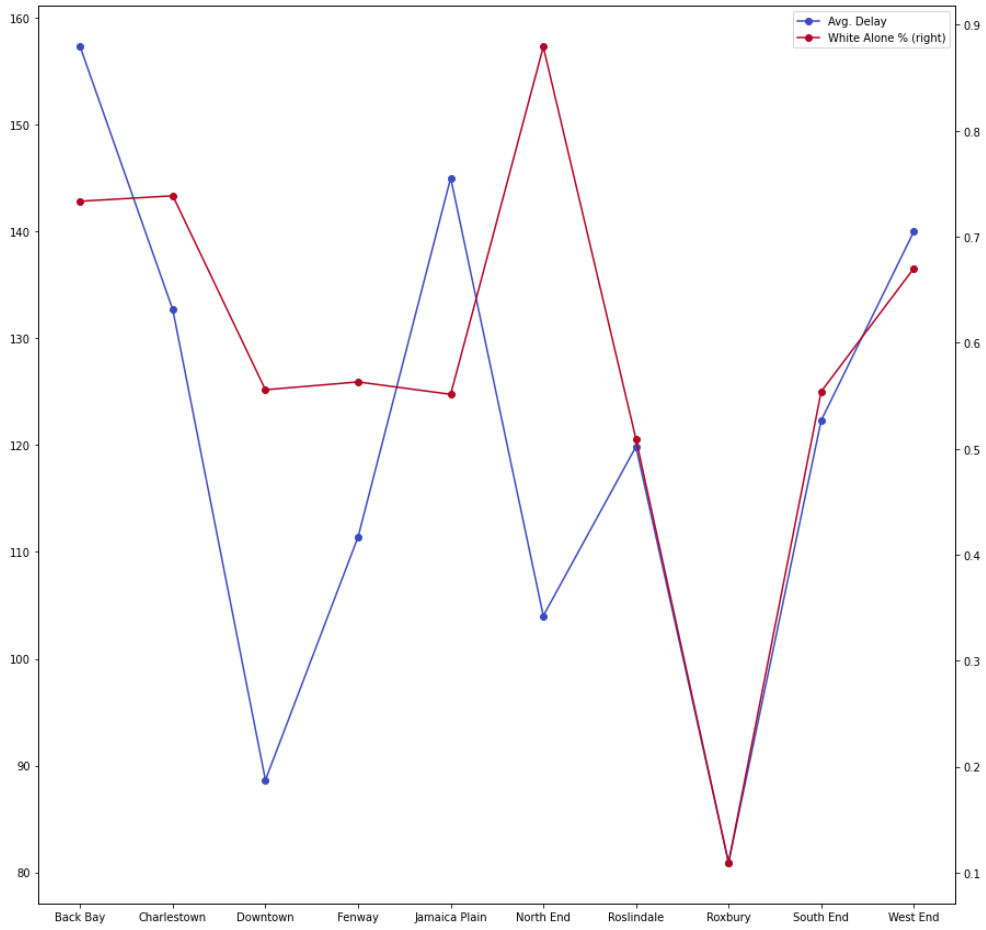


Fig. 3 Line Plot for Percentage of White Residents

Appendix 8 - Demographics with the Highest Negative Correlation

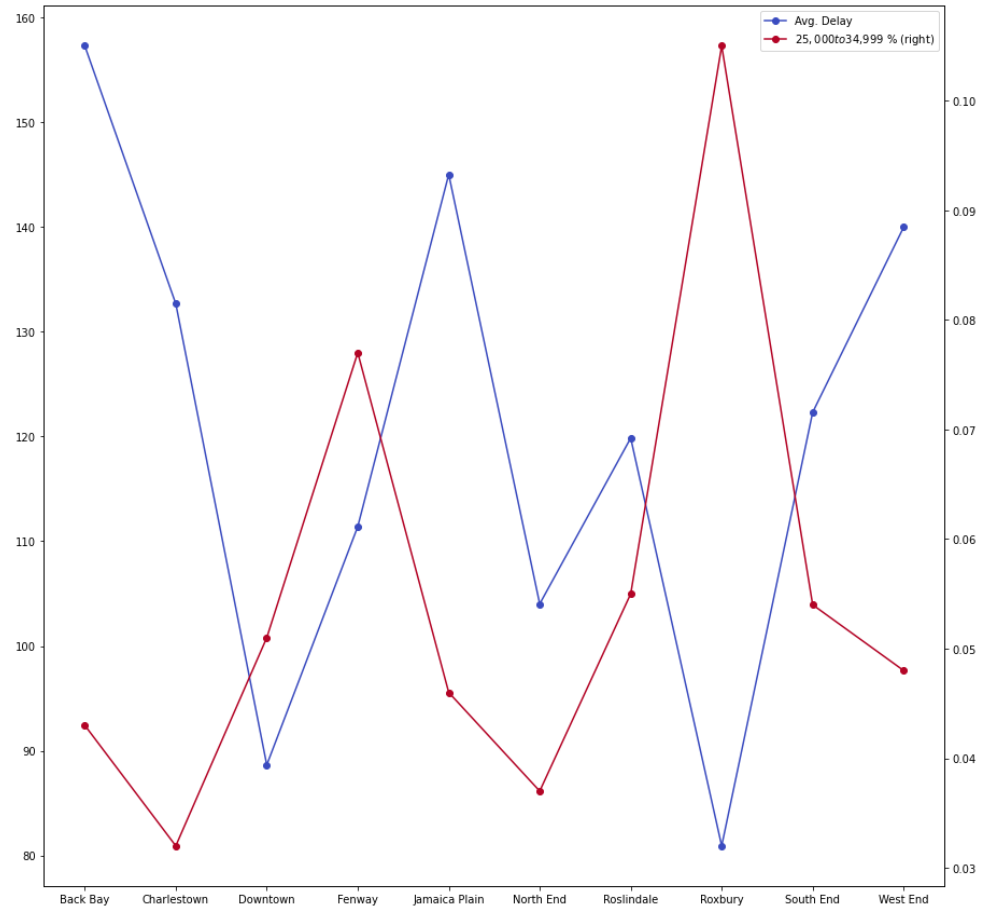


Fig. 1 Line Plot for Median Income $25,000 - 34,999

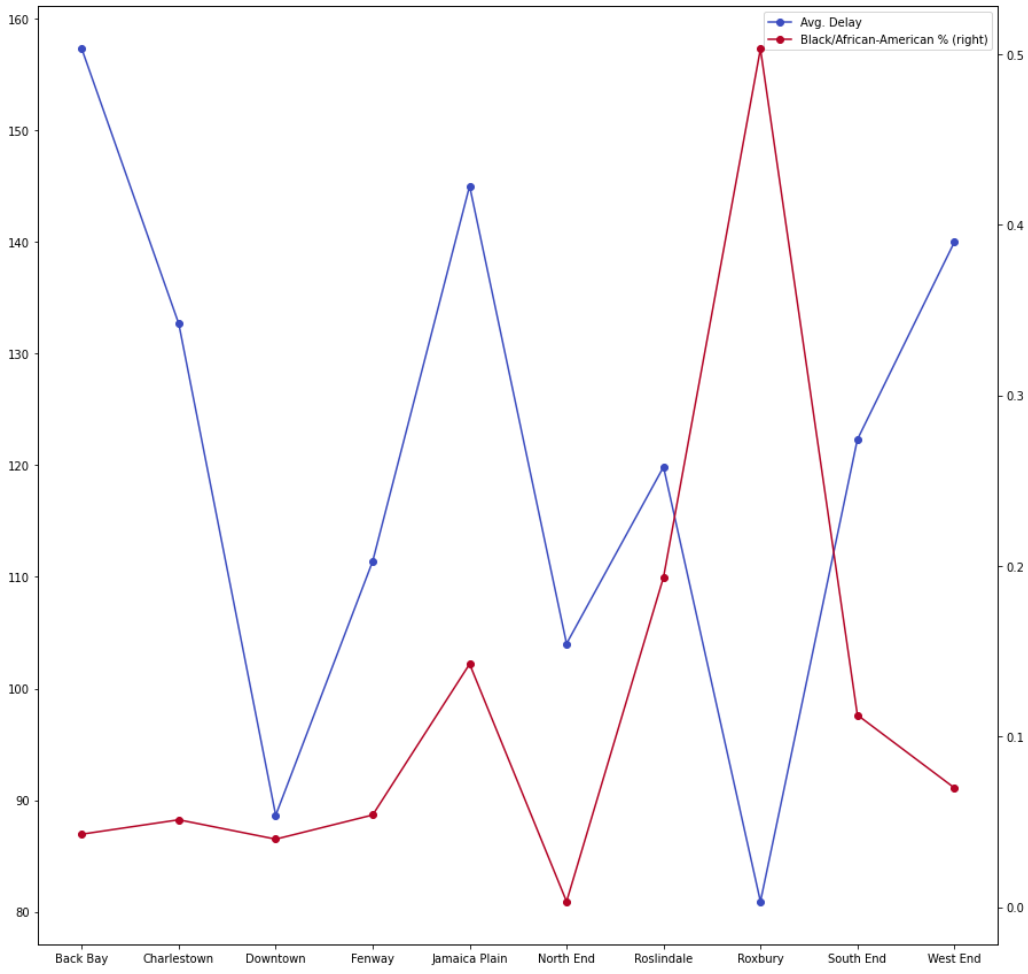


Fig. 2 Line Plot for Percentage of Black/African American Residents

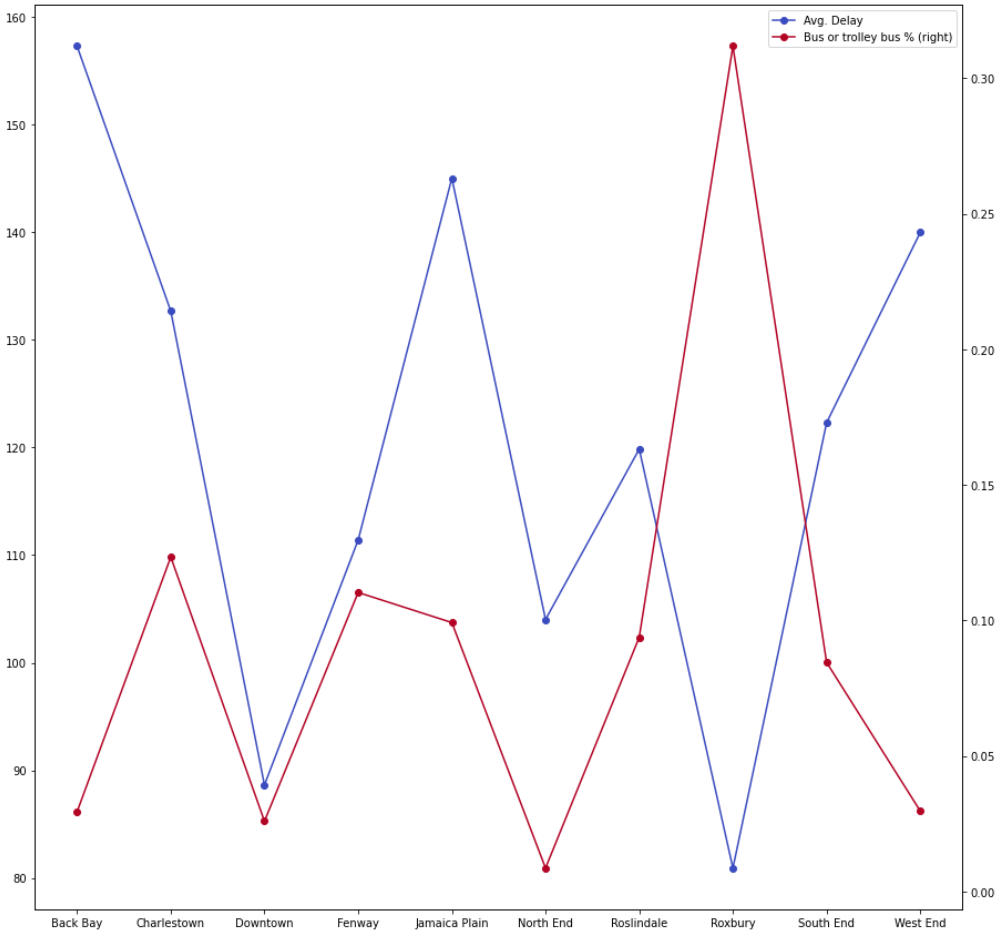


Fig. 3 Line Plot for Percentage of Residents Commuting by Bus or Trolley