**IEOR 4523 Project Report**

Team Member:

Boyuan Xia, Maokang Lin, Joshua Ye

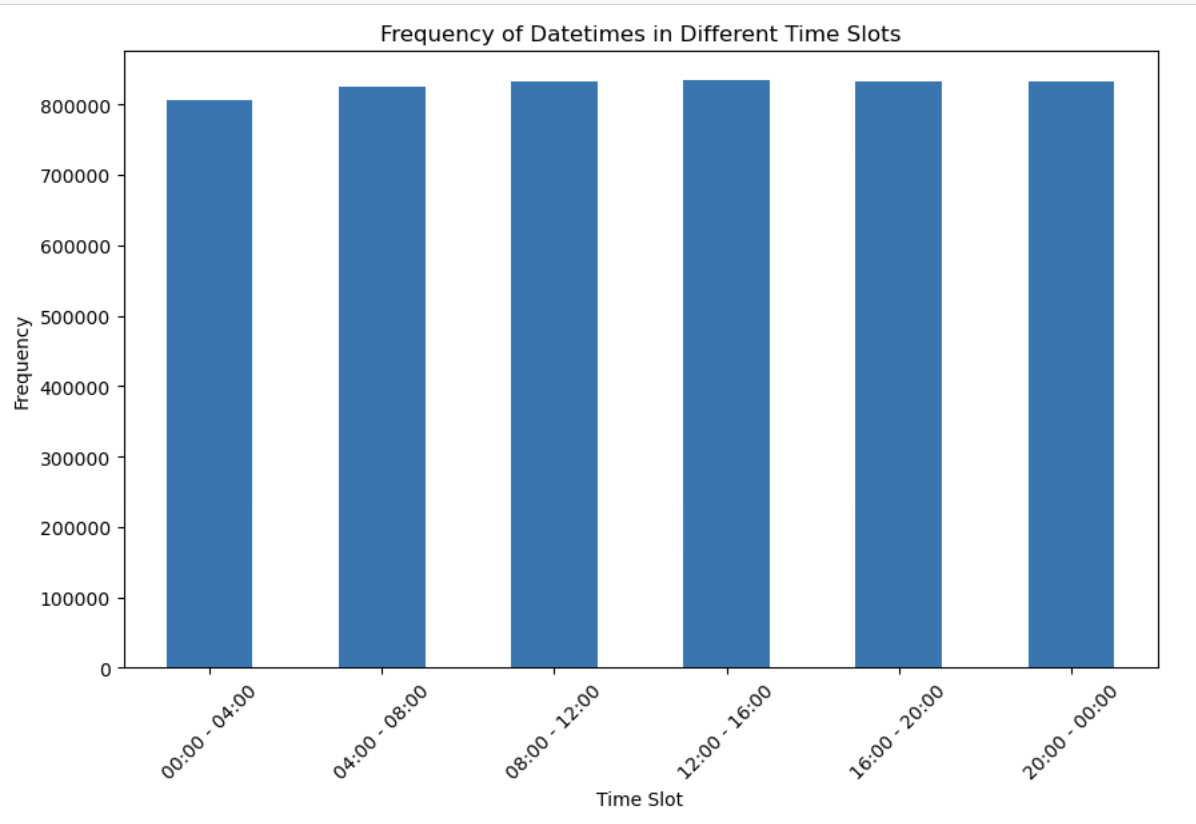
Cining Liu, Jui-Jia Chen

Fall 2023

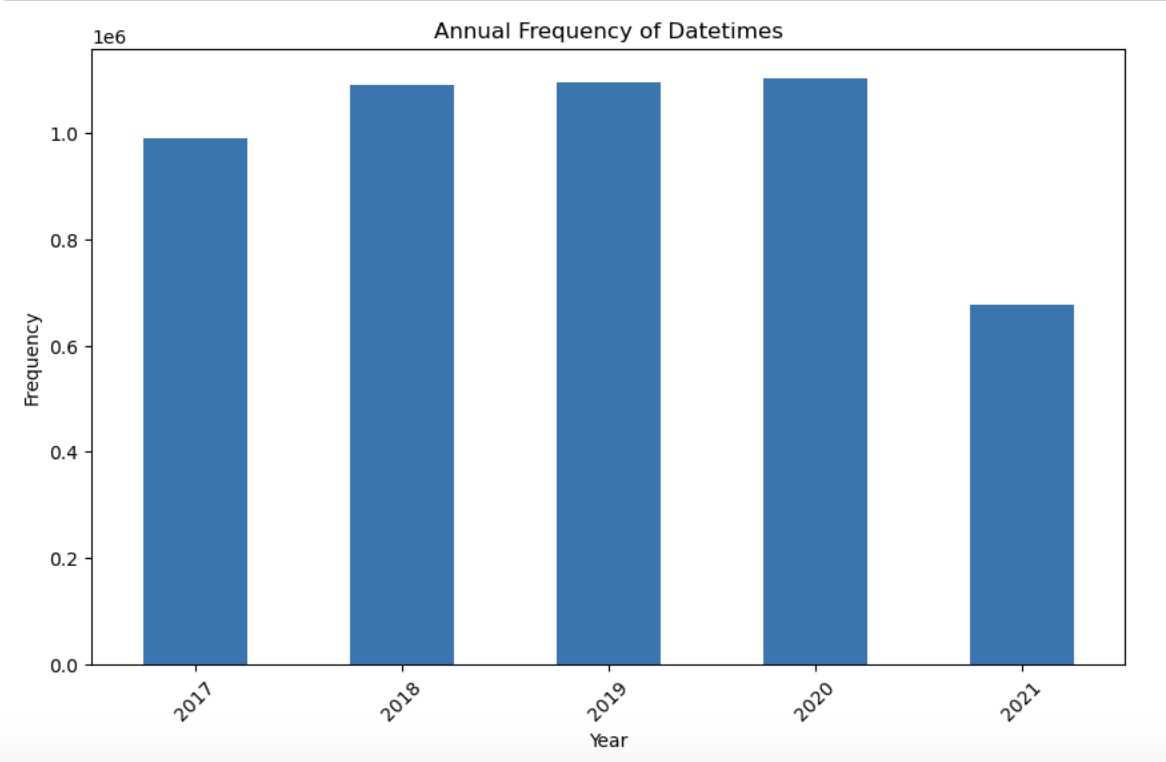
1. Introduction:

The subway system plays a crucial role in the transportation network of New York City (NYC). As part of our daily commute and exploration of the city, we delved into a comprehensive examination of passenger flows at individual subway stations across NYC. By integrating demographic factors associated with various neighborhoods in the city, we conducted an analysis and prediction of subway traffic patterns. Our goal is to provide insights into the intricate relationship between subway usage and the demographic characteristics of NYC neighborhoods.

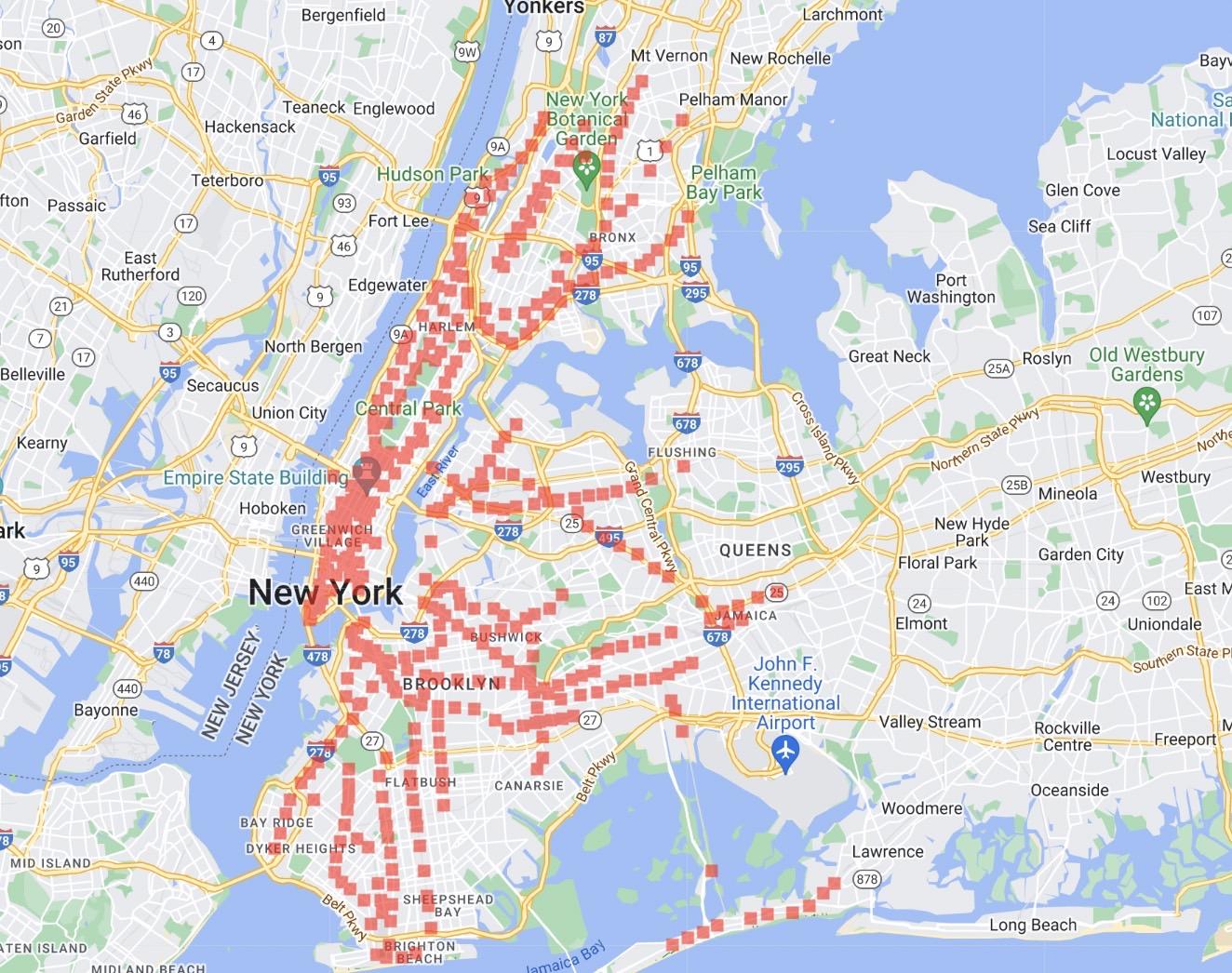
Moreover, considering the impact of the pandemic on subway traffic, leading to a decrease in passenger flow, we extended our analysis to include data from both prior to and following January 2020. This allows us to assess the recovery rate and its correlation with neighborhood factors, providing valuable insights into the complex dynamics of subway usage in the post-pandemic context.

1. Exploratory Data Analysis :
   1. Since “Datetime” is divided into six time slots a day, which means four hours per timeslot, we want to know the passenger flow of each timeslot. So we use a bar chart to visualize it. From the plot, we know that there is no significant difference among each time slot.

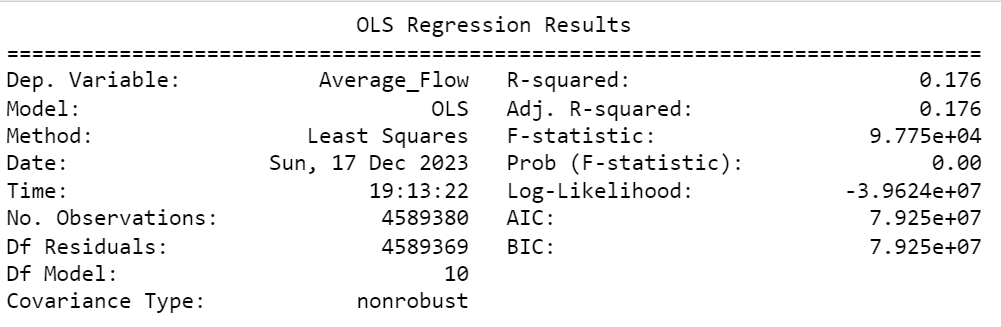
* 1. We are also interested in different years’ passenger flow so we visualize it. Counting each year’s total passenger flow. Since the data in 2021 only contains until August so the number was less that year.



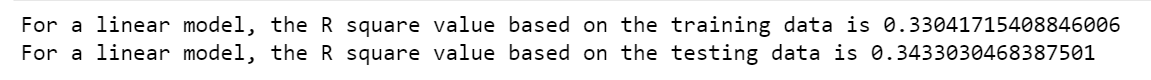
* 1. And we also calculate the top ten neighborhoods that have the most subway stations, which can be the reference for people who take the subway frequently as an indicator when choosing where to live. We visualize the outcome using Google Map heatmap to know the distribution of the subway stations.



1. Data preprocessing
   1. Data Cleaning: We obtained two datasets from Kaggle: one focused on subway traffic and the other on the census data of New York City neighborhoods. The subway traffic data spans from February 4th, 2017, to August 13th, 2021, capturing entries and exits at 4-hour intervals across 469 stations. Comprising 4,589,380 entries and 17 columns, this dataset forms the backbone of our analysis. Additionally, we incorporate NYC neighborhood data, associating each station with one of 51 neighborhoods and augmenting our insights with 87 financial and demographic variables.
   2. Variables Selection:
      1. **Dependent Variable**: We used “Average Flow" as our dependent variable, which represents the average passenger flow at a given subway station, accounting for variations in subway lines and different times throughout the day.
      2. **Data Selecting:** From the 17 columns in the dataset, we narrowed down our focus to 5 key columns: 'Datetime,' 'Stop Name', 'Line,' 'Neighborhood,' and 'Entries' from the NYC subway traffic dataset. Afterward, we sorted the entries values for each stop name and time. Leveraging the 'Entries' column, which signifies passenger flow, we computed the average flow for each subway line at every station during different times.
      3. **Independent variable:** Considering the large number of variables in the NYC neighborhood dataset, we utilized Random Forest to pinpoint variables with an importance index equal to or greater than 0.005.
      4. Furthermore, to investigate potential multicollinearity among the chosen variables (14 in total), we examined their correlation and calculated the Variance Inflation Factor (VIF). To address concerns related to high correlation and VIF, we excluded specific variables: "Students performing at grade level in math, 4th grade," "Moderately rent-burdened households," "Severely rent-burdened households," and "Median household income, homeowners (2018$)." Variables were dropped based on a correlation value surpassing 0.7 and a high VIF exceeding 10.
      5. Our final selection comprised 10 columns: ‘Housing units,’ ‘Percent white,’ ‘Moderately rent-burdened households, moderate income,’ ‘Pre-foreclosure notice rate (per 1,000 1-4 family and condo properties),’ ‘Median sales price per unit, condominium,’ ‘Home purchase loans to LMI borrowers,’ ‘Refinance loan rate (per 1,000 properties),’ ‘Single-person households,’ ‘Residential units within 12 miles of a subway station,’ and ‘Racial diversity index’ as our independent variables.
2. Regression and Modeling
   1. Linear Regression: The first model we used in our analysis was linear regression. We initially applied Multiple Linear Regression using simple Ordinary Least Squares on the data to regress the ten selected features, with a constant added, on average subway flow. Despite obtaining a low R-squared score of 17.6%, all of our feature variables were statistically significant as they had p-values of zero. This is indicative of something important missing - that average flow changes every four hours but features do not. As a matter of fact, we found that there are only six values for the whole year for each station.



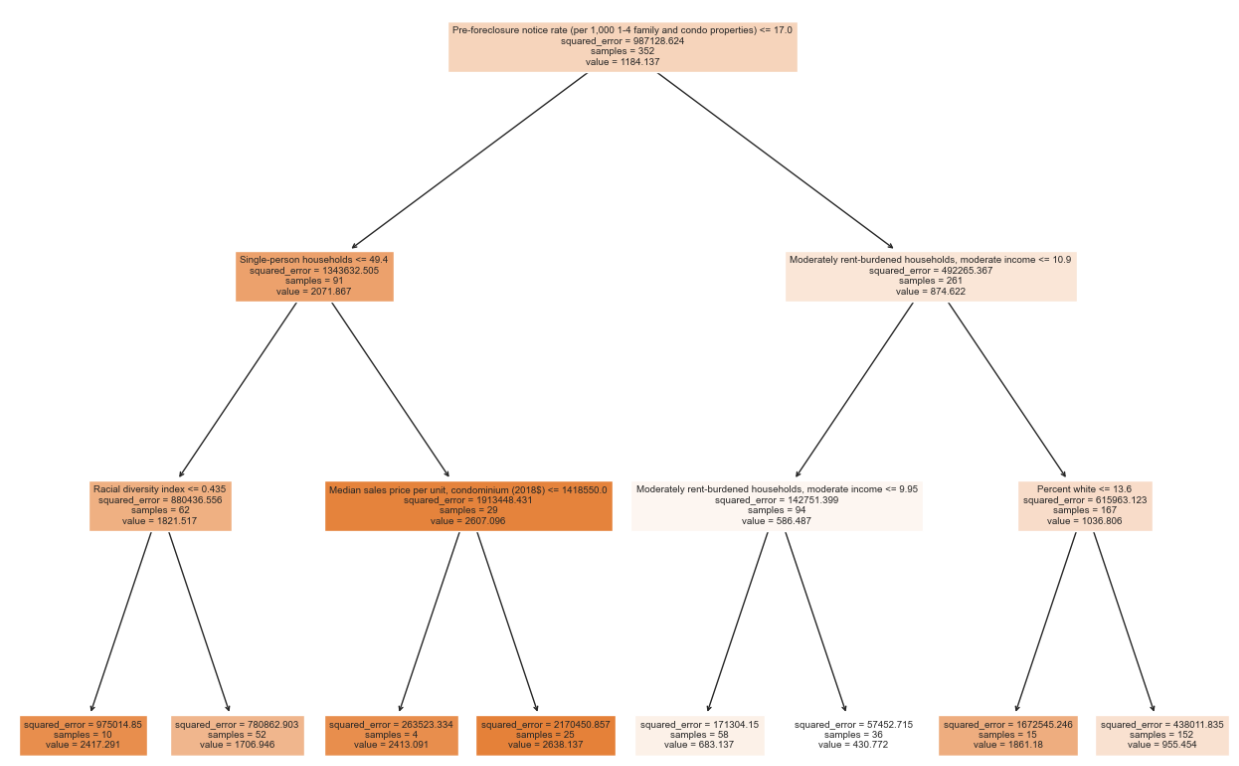
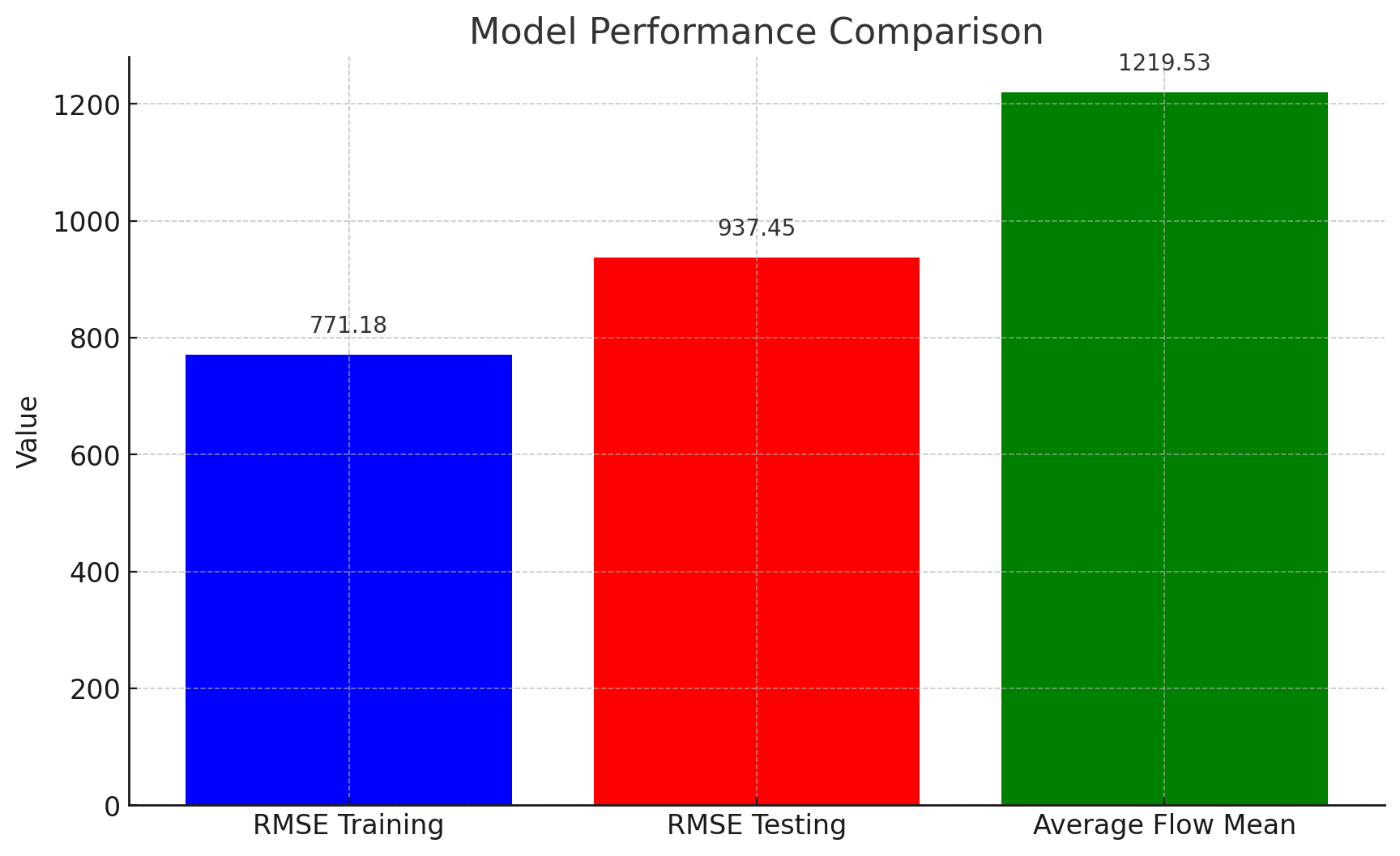
We realized the oversight of not including time as a variable and reanalyzed by splitting the dataset into its individual lines and stations. The mean average flow value across four years for each station became the new dependent variable, which should be identical to the mean of the six values of average flow per day. Through a for loop, we iterated through the lines and stations and added each station’s information on the dependent variables and their means of average flows to lists, build our modified dataset. So we want to know if the mean average flow (of 6 values) of each station can be predicted by the station specific independent variables.

We used 80% of the data as training data to build a linear model and the rest as testing. We fit the model into both on our training and testing data, getting R-squared scores of 33% and 34.3% respectively. As the test score is double that of 17.6%, we know we have successfully improved our linear model. 

* 1. Decision Tree: We utilized the decision tree regression model as our second model. Through meticulous data preprocessing, we extracted key features such as the proportion of single-person households, economic status indicators, and demographic and housing characteristics, which are thought to significantly affect subway ridership.

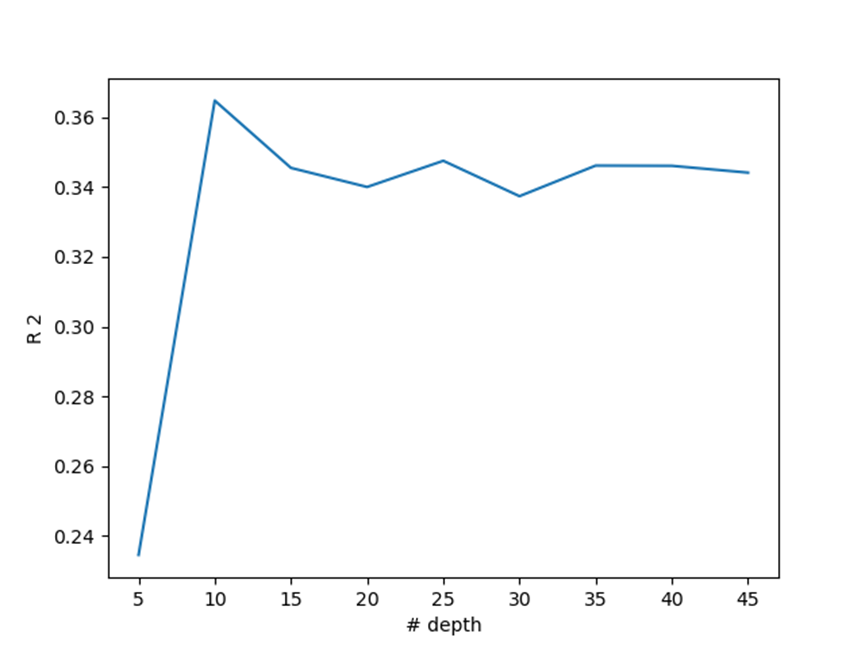
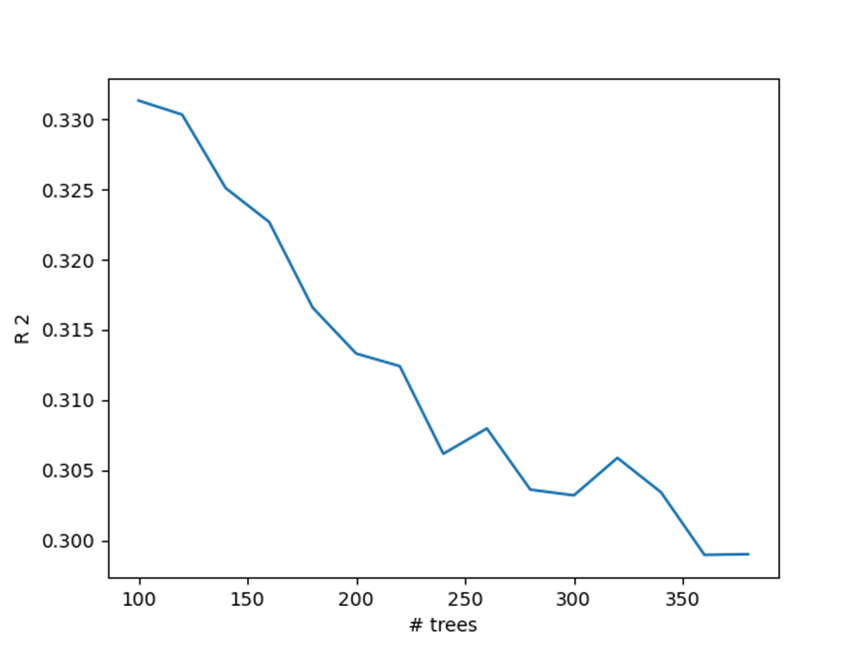
The training of the model took place after splitting the dataset into an 80% training set and a 20% testing set. The decision tree model was chosen for its simplicity and ability to handle nonlinear data. To prevent overfitting, the model depth was capped at three levels. The model's performance was evaluated using RMSE and R² values, with the training set achieving an RMSE of 771.18 and the testing set 937.45, and R² values of 0.398 and 0.229, respectively, reflecting the model's shortcomings in data fitting.

The visualization of the decision tree revealed several insights. For instance, the tree's initial split based on the proportion of single-person households may indicate that living patterns have a direct impact on subway usage. Other economic and housing factors, such as rent burden and property prices, also occupy prominent positions in the tree, suggesting that economic conditions might have predictive value for passenger flow.

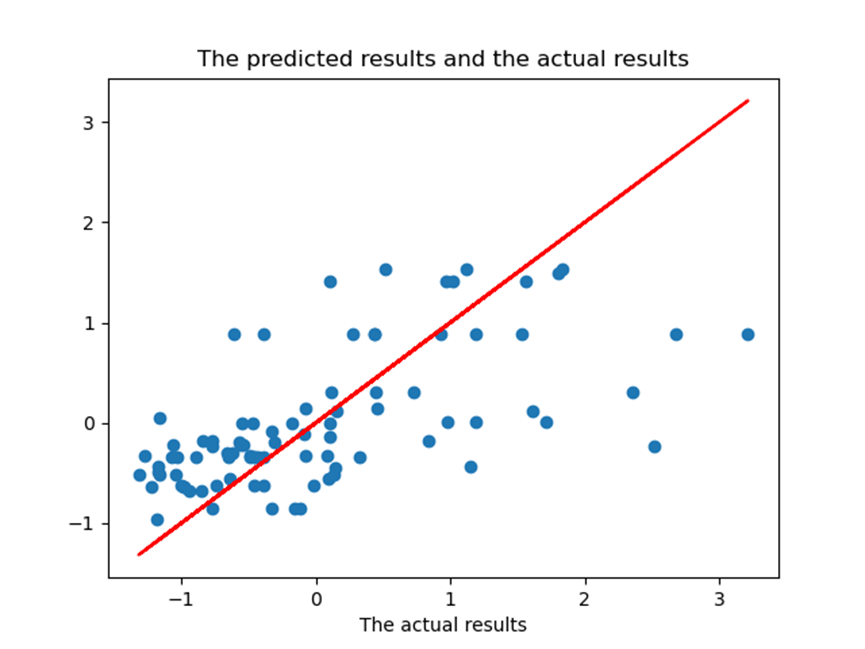
 

However, there are limitations to the model's accuracy. Although the RMSE is below the average flow value, the model's inadequacy in explaining data variance points to a deficiency in capturing linear relationships and temporal dependencies. Notably, the model failed to effectively account for temporal factors, such as the peaks and troughs in flow throughout the day or differences between weekdays and weekends. These shortcomings likely led to the model's diminished performance on the test set, indicating limited predictive ability for unseen data.

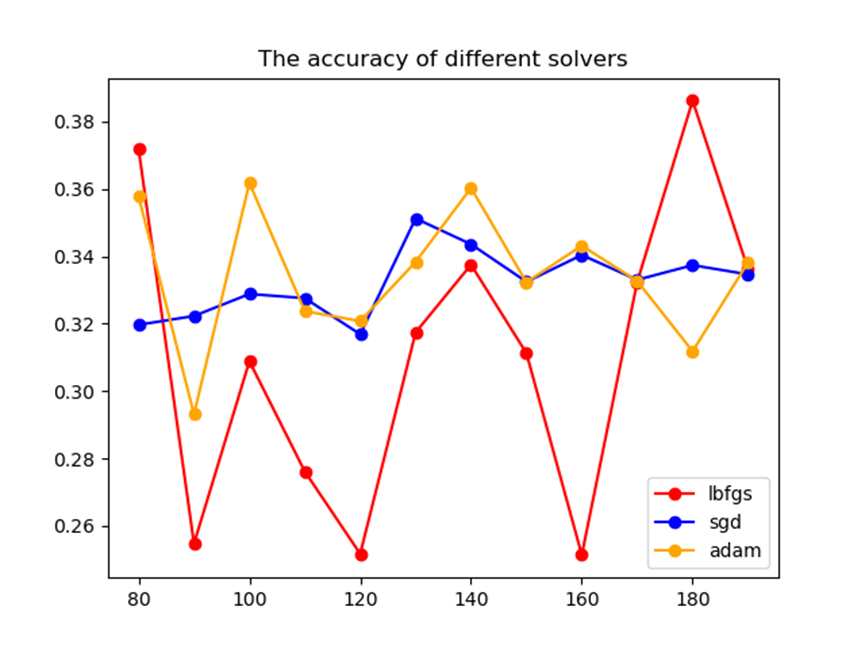
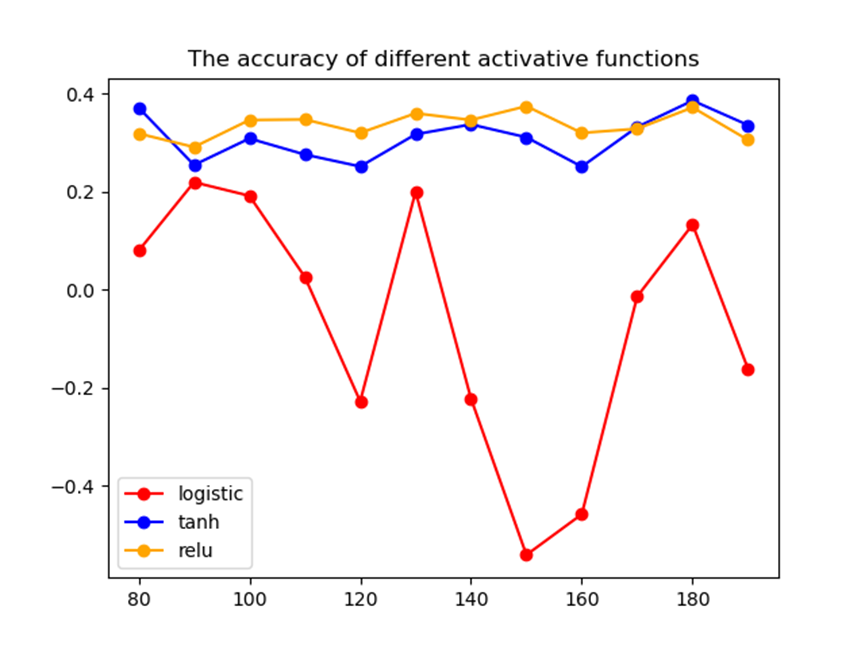
* 1. Random Forests: The third model we use is Random Forest, which has higher accuracy than a single Decision Tree. And for this part, it is worthwhile to pay attention to the parameters: number of trees and max depth of each tree. So here we use RandomForestRegressor to run our model, and the range of number of trees is from 100 to 300 and max depth is from 5 to 45. For the number of trees, we find that the model of 100 trees performs best with accuracy > 0.330. It means more trees, lower accuracy, which may be caused by overfitting. And for max depth, The model with max\_depth around 10 performs best with accuracy > 0.360, and there is no obvious difference when max\_depth > 10. This implies there's a non-linear relationship in our data.



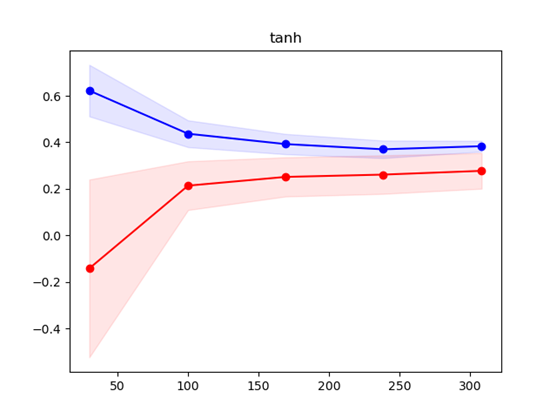
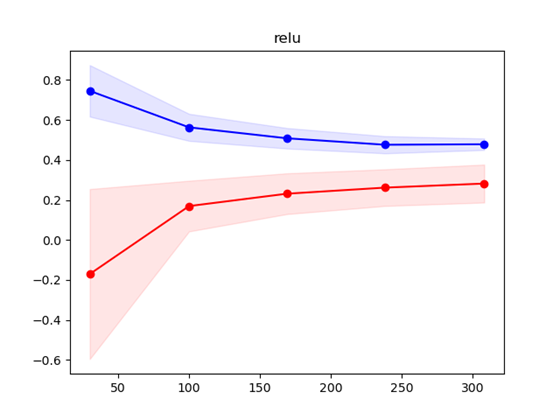
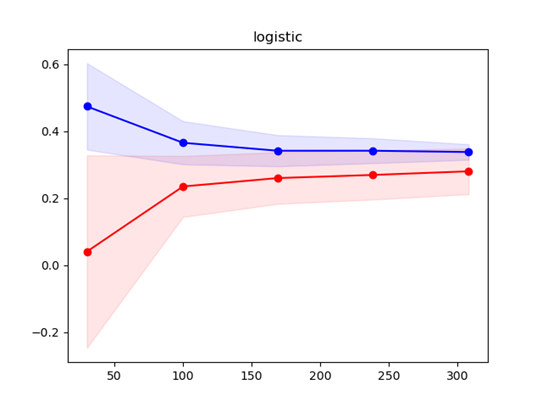
* 1. Neural Network: With all kinds of parameters and hidden layers, Neural Network is a complex model which has the best result for our project. In this section, we still focus on the influence of parameters: solvers, activation functions and hidden layers.



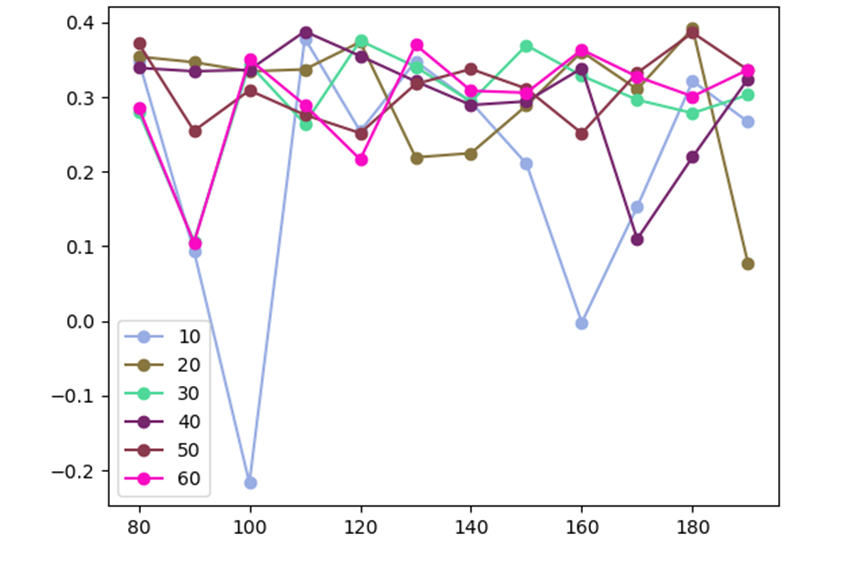
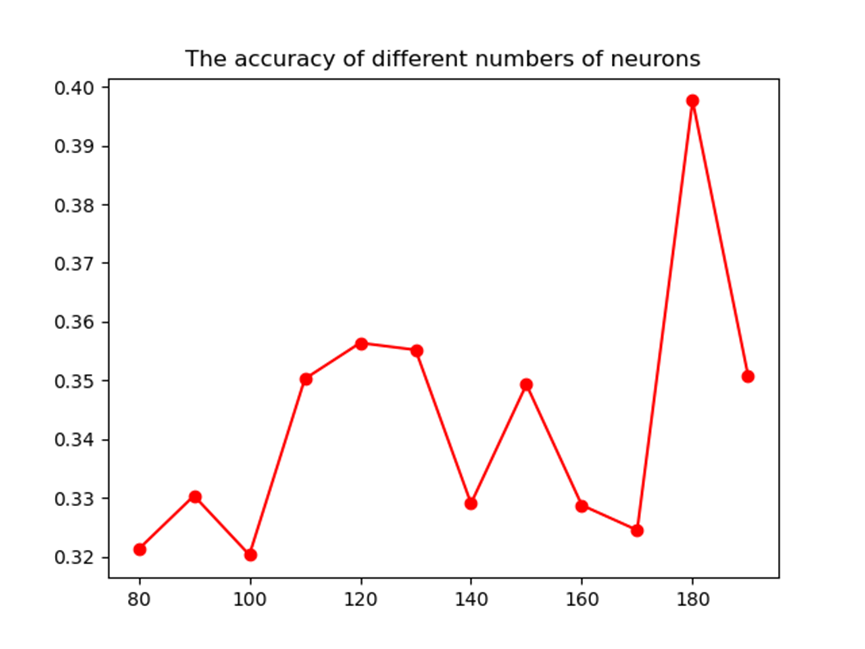
We examined the accuracy of different solvers under different numbers of neurons, and found that both Relu and Tanh activation functions have relevant high accuracy, which means this kind of activation mechanism may be more useful than Logistic. And for different solvers, we found ‘lbfgs’ works best.



To select an activation function from Relu and Tanh, we also test the learning curve of the functions, which can help identify the fitting condition of our models. Logistic and Tanh have relevant better fitting ability, while Relu is quite underfitting. It may be caused by a gradient vanishing problem in negative value regions.

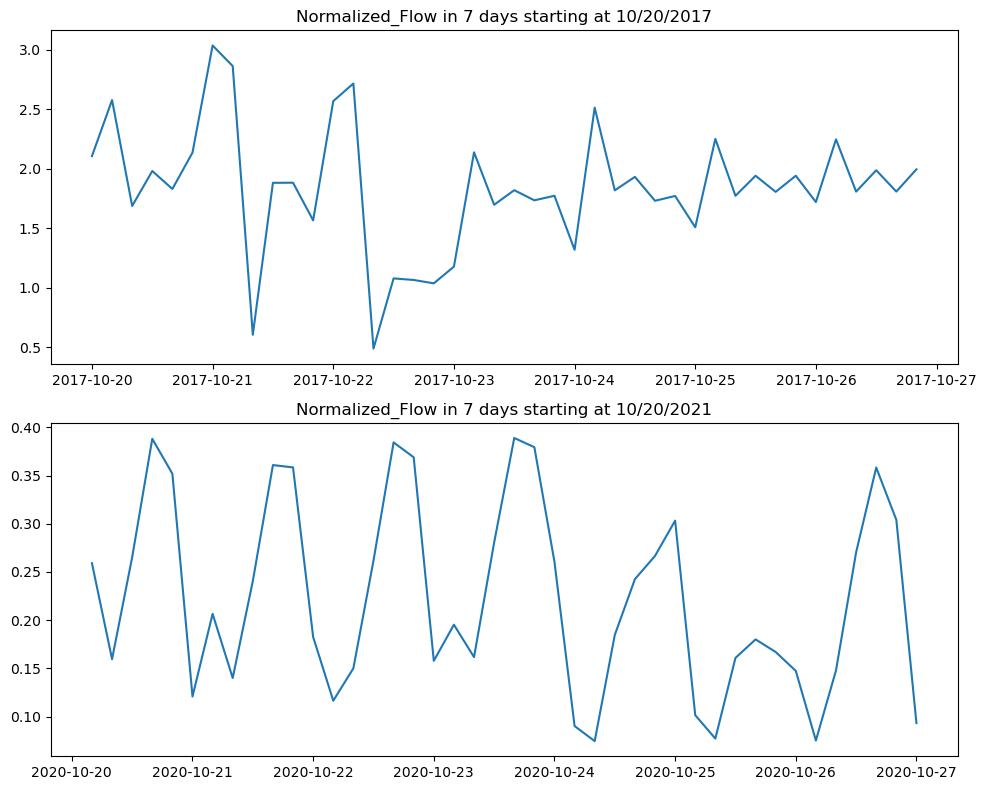
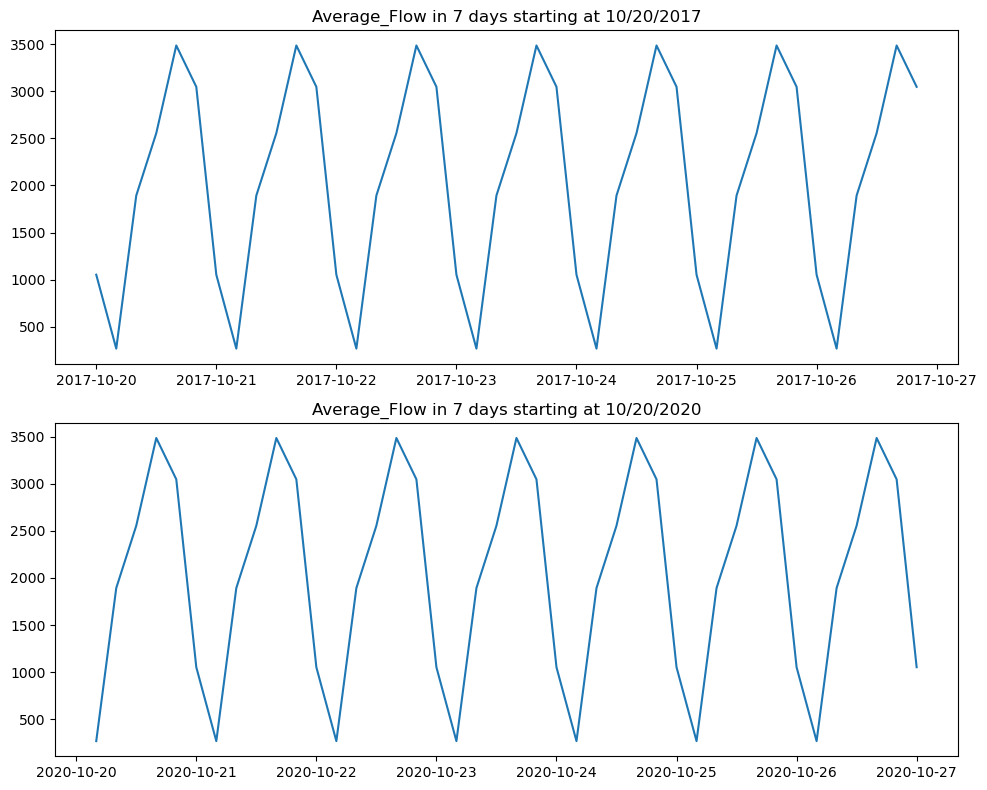


In order to discover the impact of both the number of neurons and hidden layers, the examination on neurons is divided into two groups: one hidden layer and two hidden layers. From the result of one hidden layer, we find around 180 neurons the model works best. After adding another hidden layer, the second layer with 50 neurons stands out with an accuracy around 0.38.

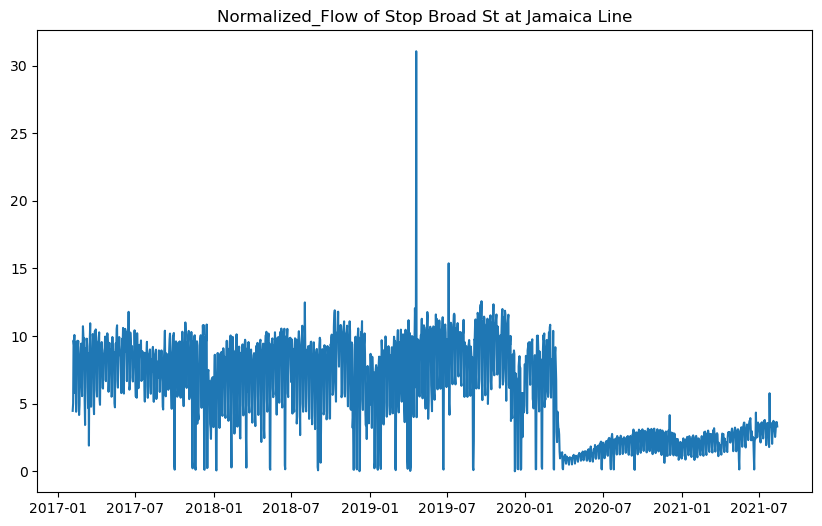
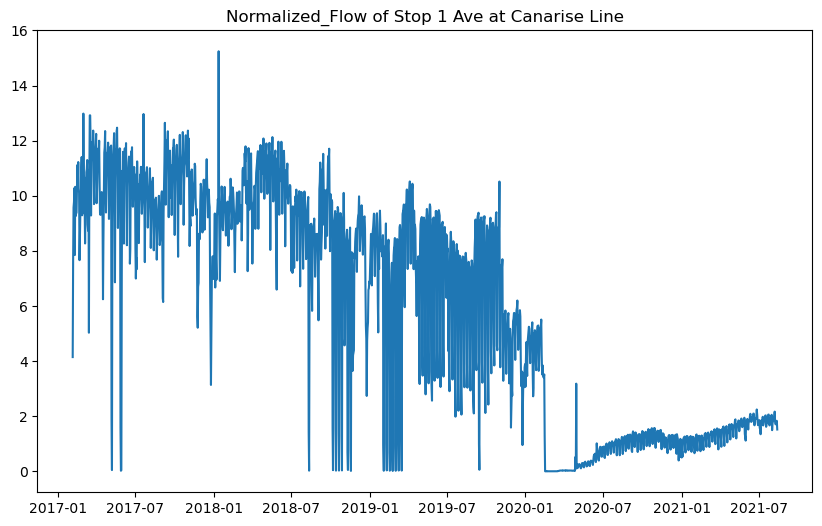


* 1. Time Series Model: The final model we will present is the times series model, which is quite different from the rest. Each station's subway flow forms individual time series of their own in the entire dataset. So we found that instead of performing time series analysis on it collectively, it would be more practical to examine the invidual stations’ time series separately. Two specific stations, 1st Avenue Station (Canarsie Line) and Broad Street Station (Jamaica Line), were selected for analysis. Besides, we transformed the data from hourly to daily because it would be more useful.

Attempts to create a time series of Average Flow revealed consistent values daily from 2017 to 2021, making predictions meaningless. Therefore, we chose Normalized flow as a more suitable dependent variable for this model. The differences between average flow and normalized flow for 1st Avene station are displayed clearly in our diagrams.

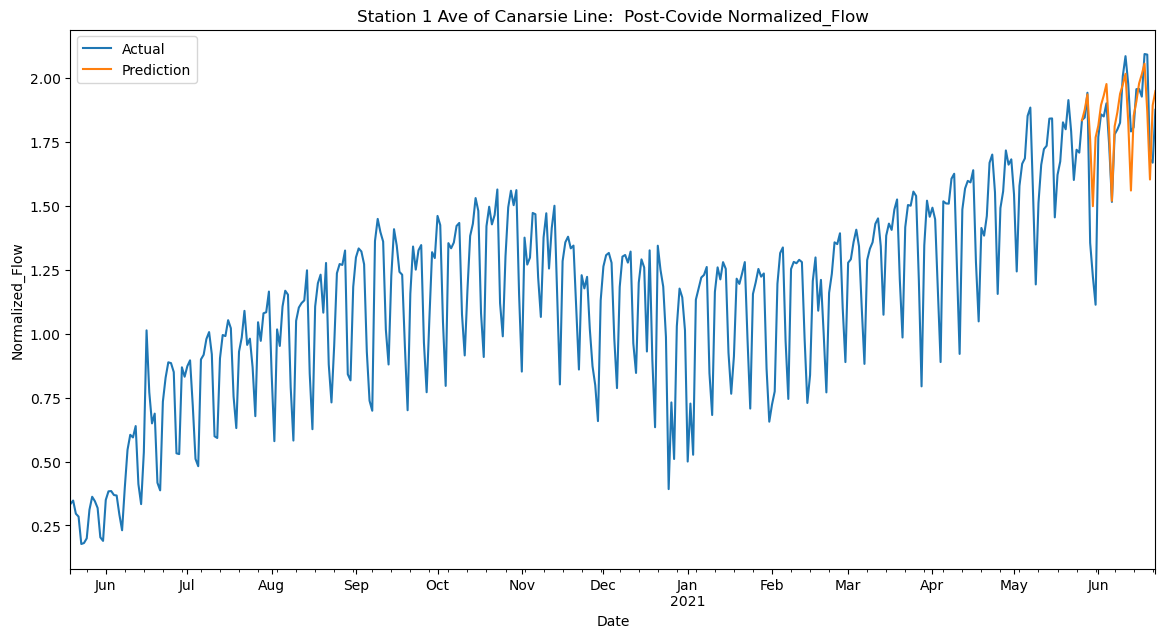
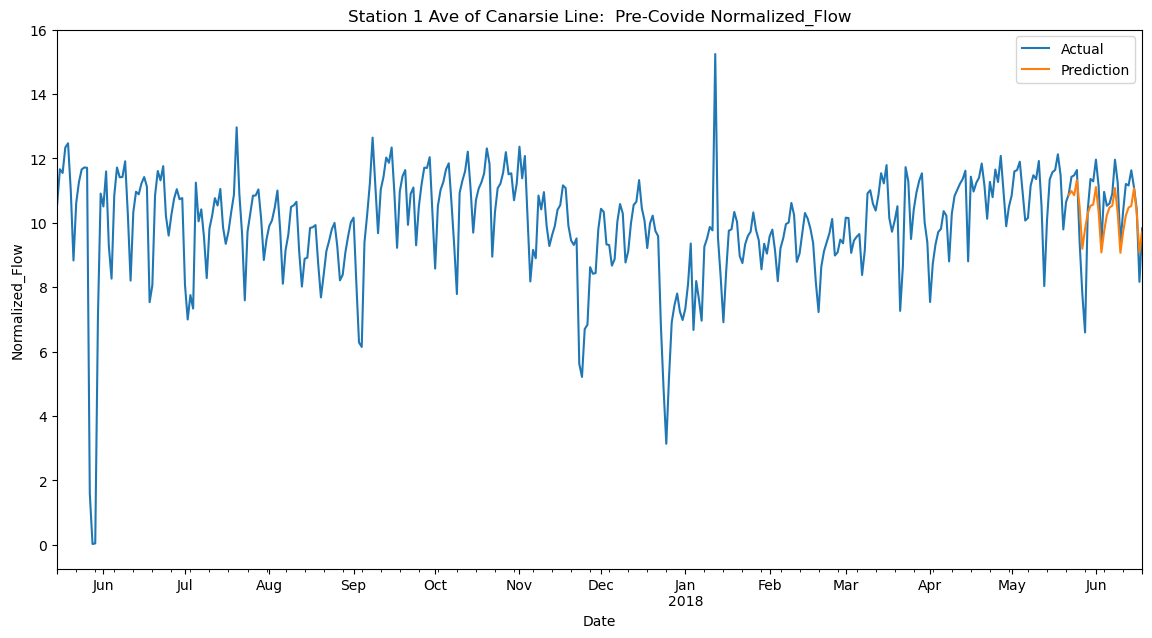


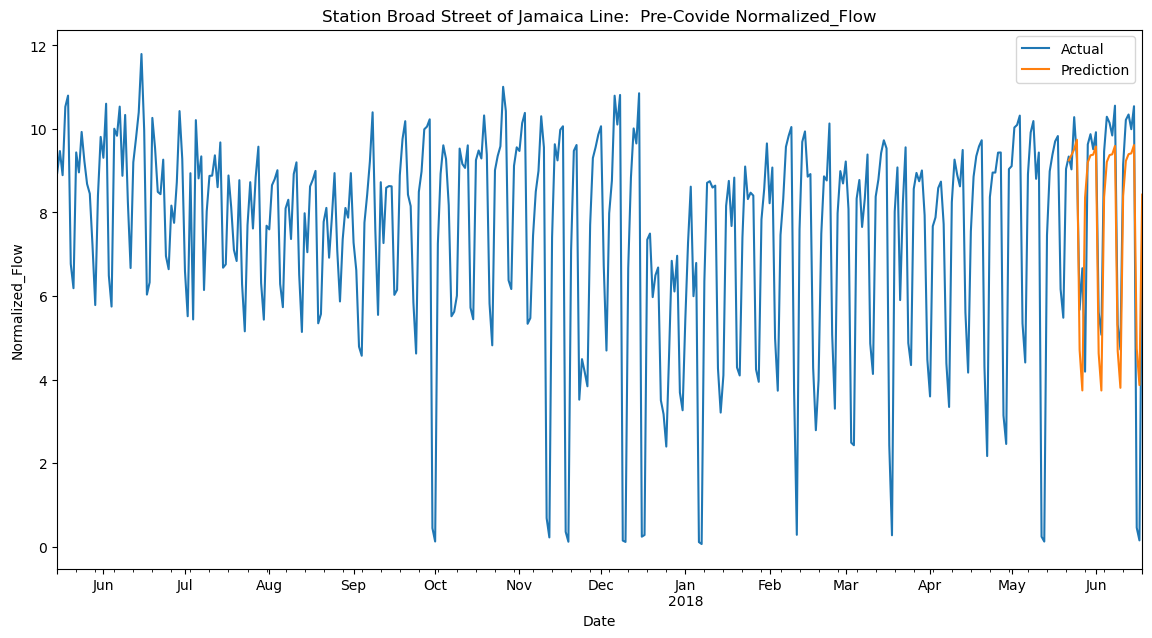
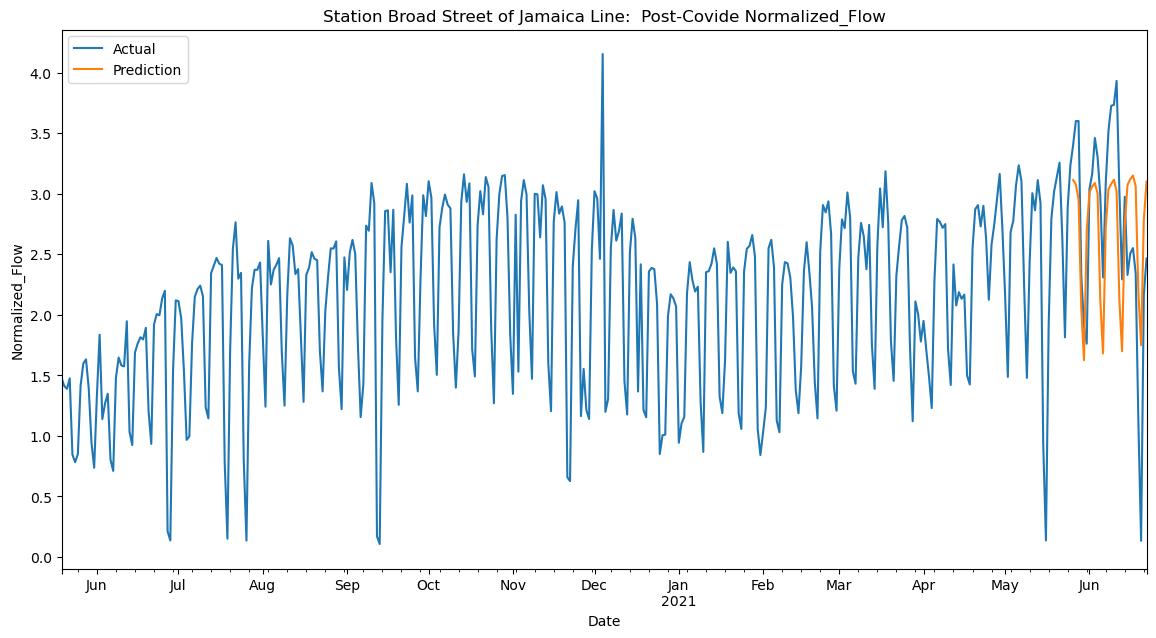
We first attemped to perform time series data analysis on each station's Average\_Flow but only found it to be meaningless as the Average\_Flow only cycles through the same 6 values every day through out the whole period, for 6 specific hours (e.g, 0:00, 4:00, 8:00, 12:00, 16:00, 20:00). Obviously this data has been processed based on some averaging method. Instead, we found that ridership data from another column, Normalize\_Flow, makes a lot of more sense. For example, it clearly shows both the short-term (e.g., over different hours during a day and over different days over a week), and long-term varations (e.g., the drastic disruption caused by Covid-19 and and slow recovery after the peak of the pandemic). Further investigations into the Normalized\_Flow ridership data reveals two facts: (1) the ridership of the same station has significant differences over different period of time such as pre-Covid and post-Covid period; and (2) different stations also have quite different ridership data during the same period.



As expected, each station's Normalized\_Flow has shown two strong seasonalities: seasonality cross the 24-hour period and the seasonality cross the 7-day period. In this analysis we decided to only study the daily ridership data instead of the original per reading in every 4 hours, because we believed predicting daily ridership is of more interest in pratical. To do so, we summed the Normalized\_Flow cross each day to obtain the daily data. A simple analysis on the daily data shows strong seasonality cross 7 days, as expected. Auto ARIMA model with seasonality is used to predict the daily ridership of the next 28 days after the model is trained with roughtly 365 days of data, which means roughly ~13-month of data (400 days of data) is used for the analysis. In addition, two different 13-month windows ( are chosen for each station, one before Covid and one after, to show the drastic change of the ridership caused by Covid. The data from mulitple stations were analyzed and we will use two randomly selected stations (Stop 1 Ave at Canarise Line, and Stop Broad St. at Jamaica Line) to demonstrate our work.

And so we applied seasonal ARIMA models to the four time series we have, First Avenue before and after Covid, and Broad Street before and after COVID. In both cases, we used the final 28 days as testing data and 1 year data before that as training data. And the four fitted ARIMA models produced different results, whether it be the number of autoregressive terms, moving average terms, or integrated terms, which is seen in the printed SARIMAX results.





Even though the ARIMA models with seasonality can predict the ridership data into the next 28 days with good accuracy, there are still ways to improve the prediction. For example, if we could include information such as if a day is a holiday, and the daily weather related information, we shall expect better results.

5 Analysis and Conclusion

1. **Analysis on regression**

Through Regression, we find that some factors, including Housing units and Single-person households, are significant and can influence the flow in the subway effectively.

1. Single-person households (positive): Single-person is more likely to take the subway rather than driving, which may be the reason that this factor can improve the subway flow in a positive way.
2. Housing units (positive): The number of houses in that neighborhood. It is reasonable that more residents can bring more subway passengers.
3. Residential units within 12 miles of a subway station (positive): The same as Housing units.
4. Pre-foreclosure notice rate (positive): For people who cannot afford their mortgage, it is most possible that they are facing financial difficulties, and this may encourage them to take the subway and result in a positive impact on the subway flow.
5. Moderately rent-burdened households, moderate income (positive): The rent also becomes a heavy burden for these populations, and this can be explained by financial difficulties as well.
6. **Analysis on Models**

For the four models we use for modelling, the neural network works best with an accuracy of 0.386. However, the R Square is not satisfactory, the prediction ability of our model still needs more improvement. To solve this problem, it is expected to observe a non-pandemic period after COVID-19 to remove the impact of pandemic. Besides, performing more feature engineering on our data and taking deep learning into models are also promising approaches.

Our time series analysis concludes that ARIMA models vary within individual stations before and after COVID, and differ among different stations in the same time period. This is due to the fact that we produced different SARIMAX results for the four fitted ARIMA models - 1st Avenue Station and Broad Street Station, before and after COVID.

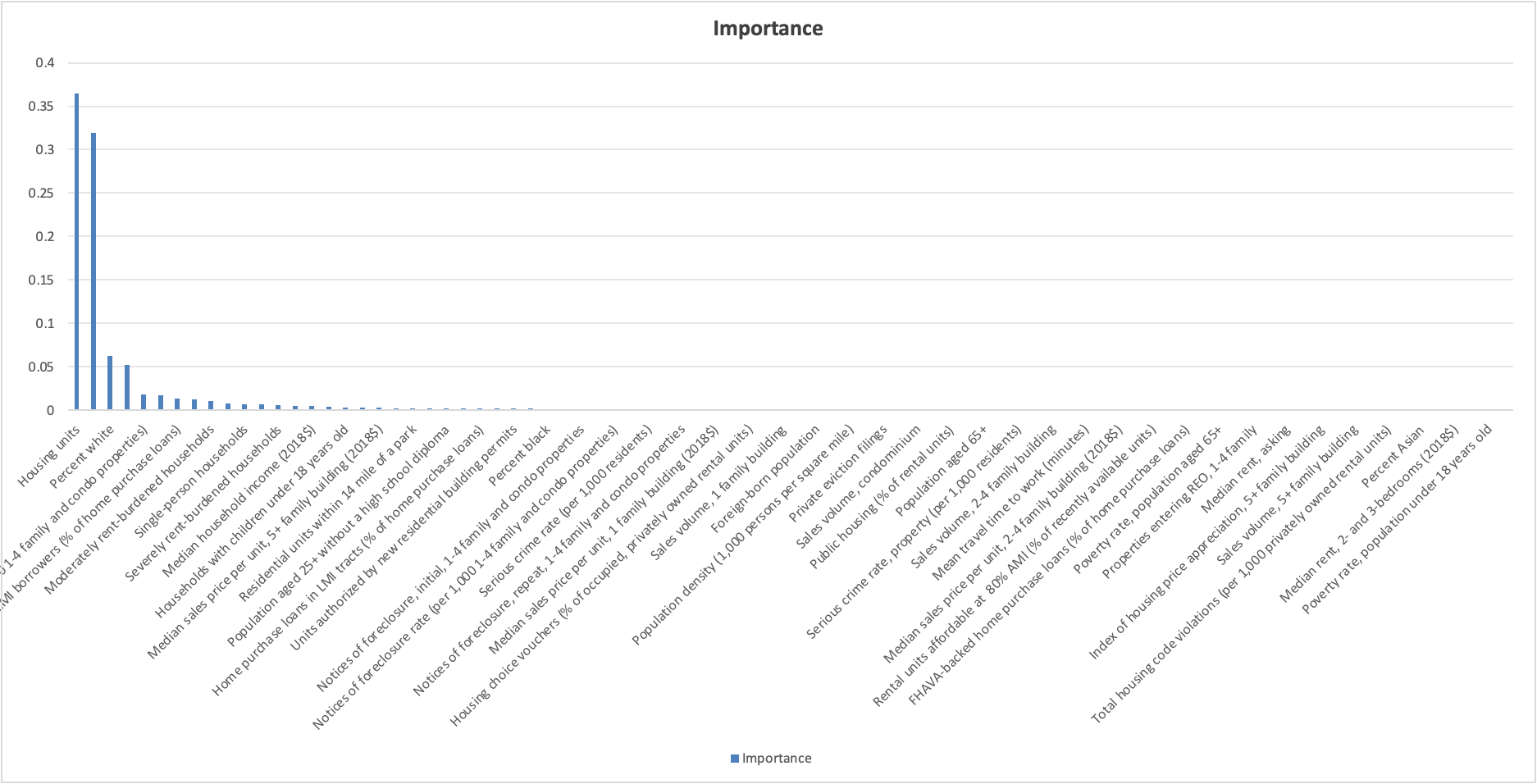
| **Models** | **Linear Regression** | **Random Forest** | **Neural Network** | **Decision Tree** |
| --- | --- | --- | --- | --- |
| **R Square** | **0.343** | **0.339** | **0.386** | **0.229** |

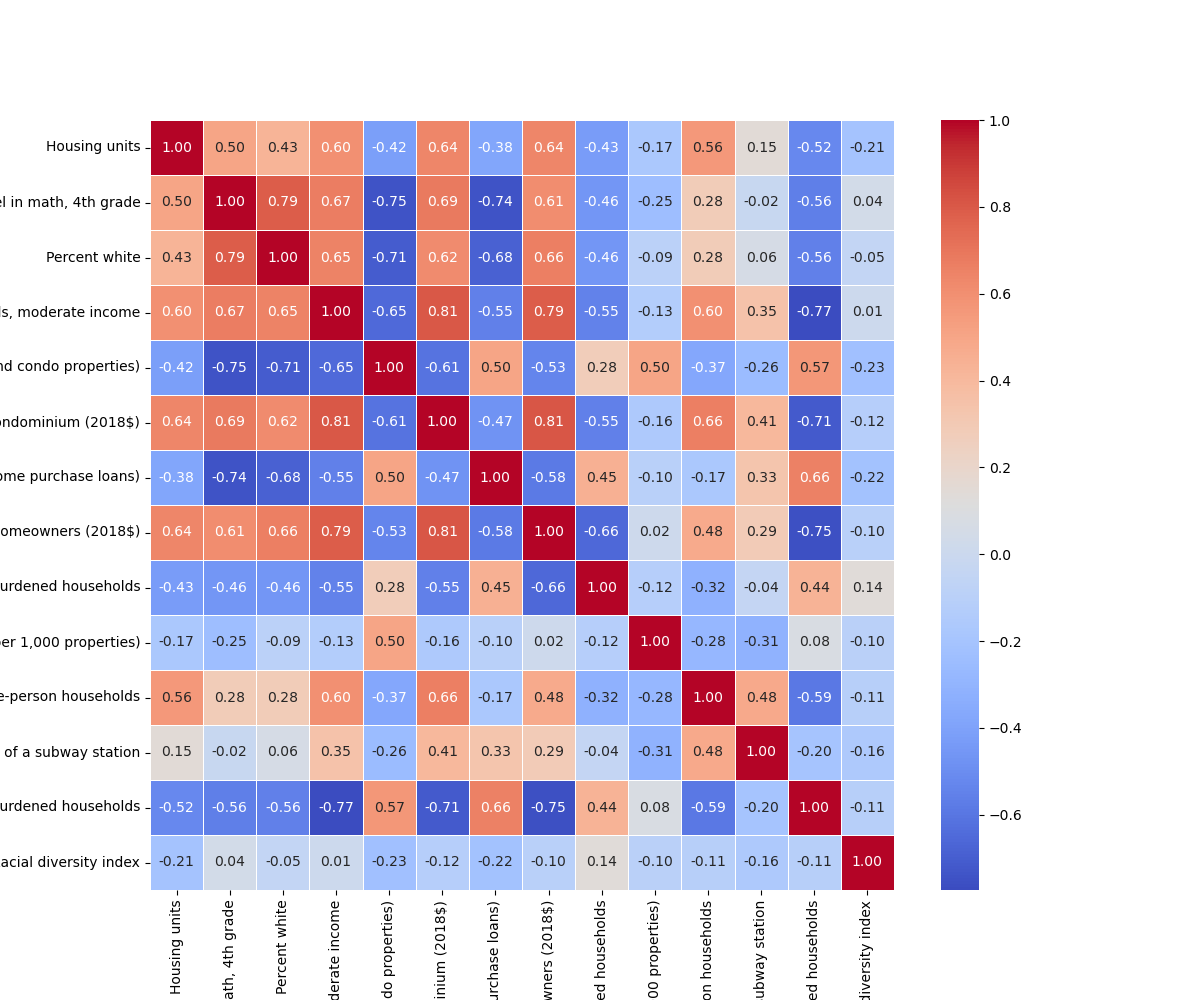
1. **Overall Conclusion and future improvement**
   1. Future Improvement:
      1. Implementing real-time data streams, such as current weather conditions, special events, and traffic updates, could significantly enhance the predictive power of our models. This would allow for more dynamic and responsive forecasting.
      2. For the linear regression model, we can consider implementing techniques like Ridge or Lasso regression to prevent overfitting and handle multicollinearity. As for the decision tree model, we can tune different tree depths to balance between underfitting and overfitting. Moreover, in the random forest part, we need to ensure that the trees in the forest are not highly correlated to maintain the diversity in the model and increase the accuracy.
   2. Overall Conclusion: Our findings highlight the importance of the subway's proximity to residential areas. Therefore, investing in more accessible subway stations and improving connectivity in underserved neighborhoods could boost usage and convenience. Furthermore, we can provide some suggestions for subway station infrastructure development. For example, expanding capacity in overcrowded stations or optimizing train schedules to meet fluctuating demand. As for the neighborhoods identified as having lower subway usage, the government can launch targeted campaigns to encourage public transportation, which also makes New York City more sustainable.

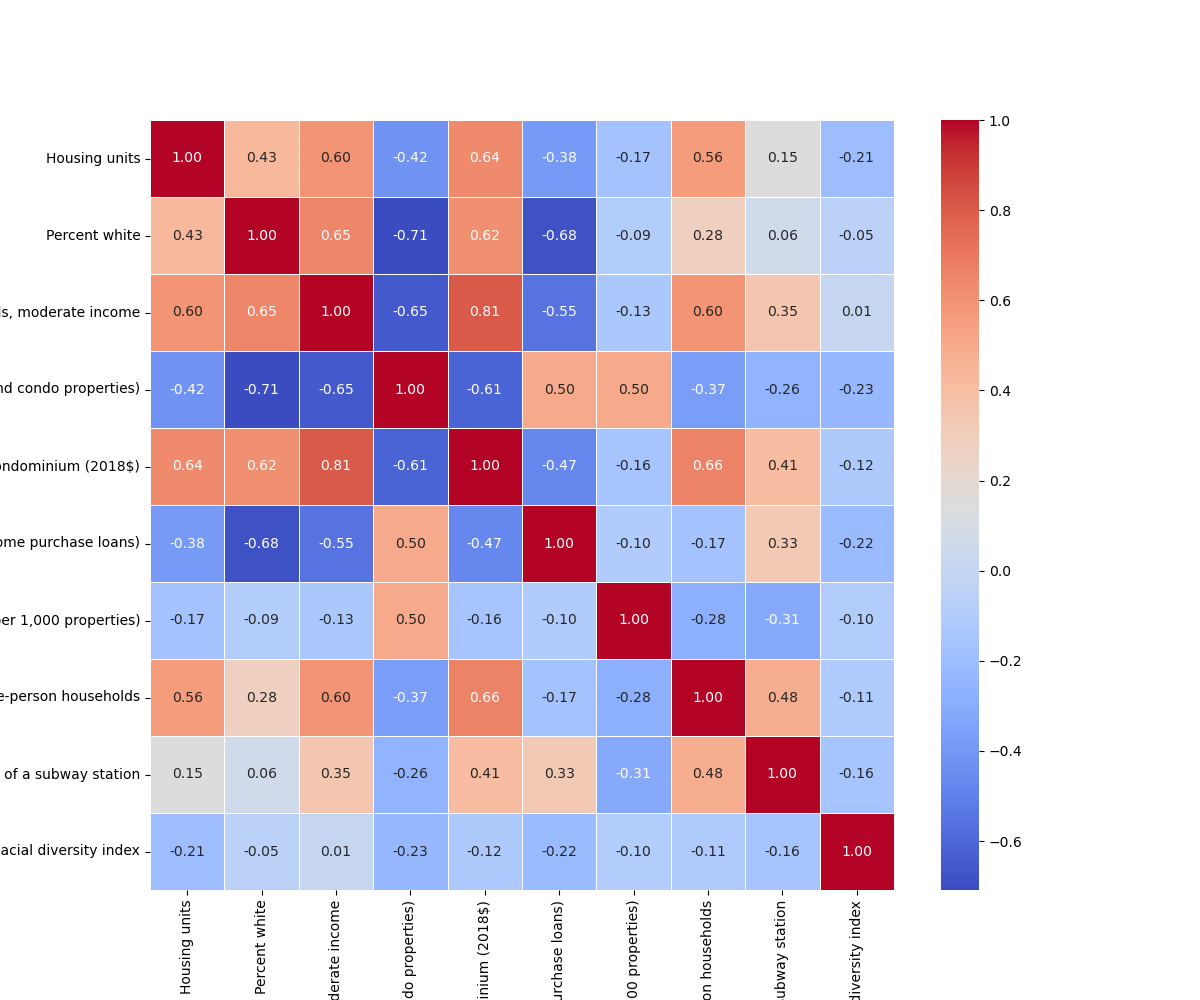
APPENDIX Ⅰ: Variables Definition

| Independent variables | Definition |
| --- | --- |
| Housing units | The number of houses in that neighborhood |
| Percent white | The percentage of whites in that neighborhood |
| Moderately rent-burdened households, moderate income | For moderate-income households, the percentage of them who pay 30-50% of their income on rent |
| Pre-foreclosure notice rate (per 1,000 1-4 family and condo properties) | The number of families who failed to make payments on their mortgage for every thousand families |
| Median sales price per unit, condominium | The median price of a condominium |
| Home purchase loans to LMI borrowers | For residents who have income no more than 120% of median income, the percentage of them who have home purchase loans |
| Refinance loan rate (per 1,000 properties) | The number of properties that has a refinance loan per 1000 properties |
| Single-person households | The percentage of residents who live in a single-person household |
| Residential units within 12 miles of a subway station | The percentage of residential units within 12 miles of a subway station |
| Racial diversity index | How likely two people chosen at random will be from different race and ethnicity groups |

APPENDIX Ⅱ: Variables Importance and Correlation Matrix







APPENDIX IⅡ: Linear Regression Output