Jonathan Hernandez

DATA698 Fall 2019 Final Project

NYC Traffic Fare Levels and Volume

Table of Contents

[Introduction](#_hlfybb675uef)

[Data Sources](#_1e1ibj7a2itu)

[Literature Review](#_6cf6dt45ljkx)

[Hypothesis and Model Simulation](#_j9ggb4ysihyz)

[Getting and Cleaning the Data](#_q0kurw2sc3ul)

[Exploratory Data Analysis](#_r79buvw1pc5v)

[Gathering Historical Fares](#_ftjlvx8ba9go)

[Training Machine Learning Models For Predicting Vehicle Volume](#_3pfx8oo8wtur)

[Creating Simulated Fares](#_q20te2a50q95)

[References](#_is6vgsnzz6sr)

### Introduction

New York City’s Metropolitan Transit Authority (MTA) controls all public transportation of NYC. Over the years traffic congestion has been a common problem for NYC bridges and tunnels and many ideas have been proposed by the MTA. For this project, I’ve decided to look into data that contains bridge and tunnel volume for NYC bridges and tunnels. The reason for looking at this data is to create simulation fare rates and find a suitable fare rate that maximizes profit and revenue, but also minimizes traffic congestion and in turn, reducing carbon emissions and accidents.

### Data Sources

Examining Traffic among NYC bridges and tunnels:

<https://www.kaggle.com/new-york-state/nys-metropolitan-transport-authority-mta-data>

This is a csv file of NYC bridge and tunnel tolls from 2010 is a dataset that contains hourly traffic on the MTA tunnels and bridges. The CSV file contains over 1.5 million observations and 6 columns. Details of the features are as follows:

* Plaza ID: numerical ID that represents each toll plaza.
* Date: Date of the measurement taken of traffic
* Hour: The hour associated with the data
* Direction: Direction of traffic (Inbound or Outbound)
* # Vehicles – E-Zpass: Number of vehicles that pass through each bridge or tunnel
* # Vehicles – Cash/Vtoll – Number of vehicles that paid in cash

### Literature Review

According to a nyc.sstreetsblog.org blog post back in 2007, variable pricing at MTA bridges and tunnels would ease traffic by for example, encourage some drivers to shift their trips to off-peak times. Doing so would decrease traffic by about 4.9 to 11.8 percent. Other options mentioned in the article would charge more during peak hours and less during off-peak times. This article shows how altering prices can help to reduce traffic which is part of the purpose of this project. This also shows that for our simulated fares, traffic should be less in the off-peak hours and more otherwise.

Another article from silive.com shows that implementing nyc congestion pricing can help to also reduce congestion and improve travel speeds. The idea is to create a two-way congestion pricing fee and doing so would make drivers travel less and may increase revenue. Various pricing models are mentioned like flat all day price, low/moderate/high peak period price. The site gothamist.com also reports that MTA and the city have decided to apply congestion pricing in the year 2020 or 2021.

### Hypothesis and Model Simulation

The hypothesis will be finding an optimization model to simulate bridge and tunnel toll rates that maximize revenue and mass transit volume while minimizing traffic congestion, accidents and/or carbon emissions. This model can motivate NYC residents and Metro area commuters to use mass transit versus using vehicles. Doing this can help to reduce the city’s use of carbon emissions and footprint and reduce traffic congestion.

The optimization approach is to append random fares to the dataset, compute revenue for that given day or year and see which price yields the highest revenue. Examples that will be used are assuming the prices are fixed throughout the years and slowly increasing fare hikes. Each model can be examined and visualized one by one to see how the revenues change over time and fares. The models can also assume all vehicles pay the same price or look into other data that shows how many and which types of vehicles enter the bridges and tunnels.

### Getting and Cleaning the Data

The language of choice will be in Python. Various libraries such as Numpy, Scipy, Matplotlib, Pandas, Seaborn will most likely be used. Machine Learning algorithms will be used from the sklearn module.

For the approach of how to create simulations and model, different algorithms such as simple linear regression, elastic net and random forest models will be used. These models will be a way to predict vehicle volume based on simulated fares.

To get an idea of how much vehicle volume for the EZ-pass or total of vehicles paid in cash, simulated fares ranging from $4.00 to $20.00 will be used in one dollar increments. For each simulated price, a model will be fit and the result will be a model for predicting vehicle count.

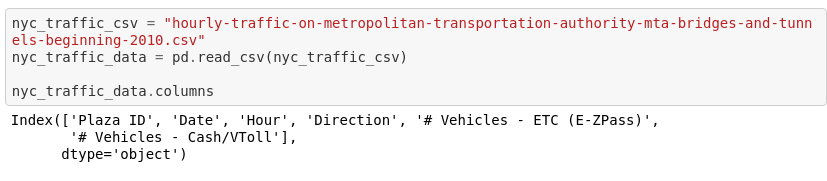
After acquiring the dataset, we see that the dataset contains 1467219 rows and 6 observations. Each observation contains the number of vehicles that passed through a particular bridge/tunnel via cash or EZ-pass within a specified hour. The 6 columns are as follows:

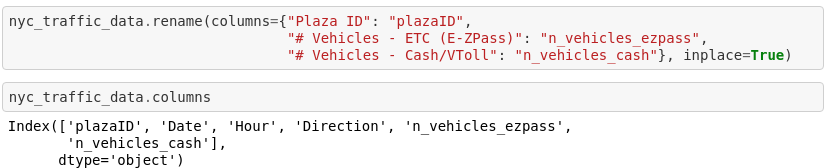
To better read and understand which plaza ID goes to which one, a mapping scheme had to be implemented. In NYC, there was a tolling switch over date and values 1 to 11 are the original historical values and values 21 to 30 were the values after the switch over date. Also to avoid having spaces in variables, the features ‘Plaza ID’, ‘# Vehicles - E-Zpass’ and, ‘# Vehicles - Cash/Vtoll’ were changed to ‘plazaID’, ‘n\_vehicles\_ezpass’ and, ‘n\_vehicles\_cash’ respectively. Finally, the time after the date in the ‘Date’ column was removed as the hourly total was already included in the ‘Hour’ column.

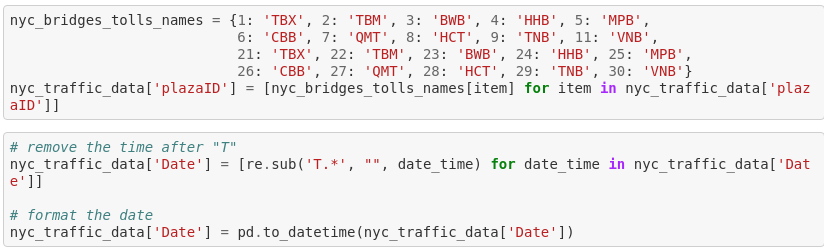
In addition, the abbreviations of the bridges and tunnels are as follows:

* TBX - Robert F. Kennedy Bridge Queens/Bronx Plaza
* TBM - Robert F. Kennedy Bridge Manhattan Plaza
* BWB - Bronx-Whitestone Bridge
* HHB - Henry Hudson Bridge
* MPB - Marine Parkway-Gil Hodges Memorial Bridge
* CBB - Cross Bay Veterans Memorial Bridge
* QMT - Queens Midtown Tunnel
* HCT - Hugh L. Carey Tunnel (was formerly known as the Brooklyn-Battery Tunnel (BBT)
* TNB - Throgs Neck Bridge
* VNB - Verrazano-Narrows Bridge

The below snippet shows the column renaming, plazaID mapping and, removing the time from the Date column.





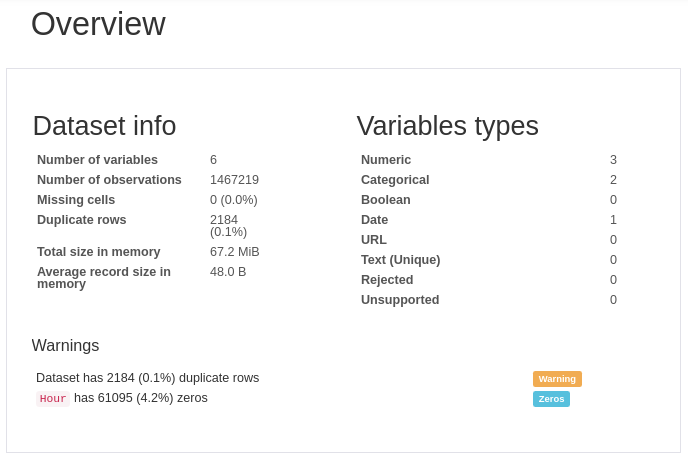


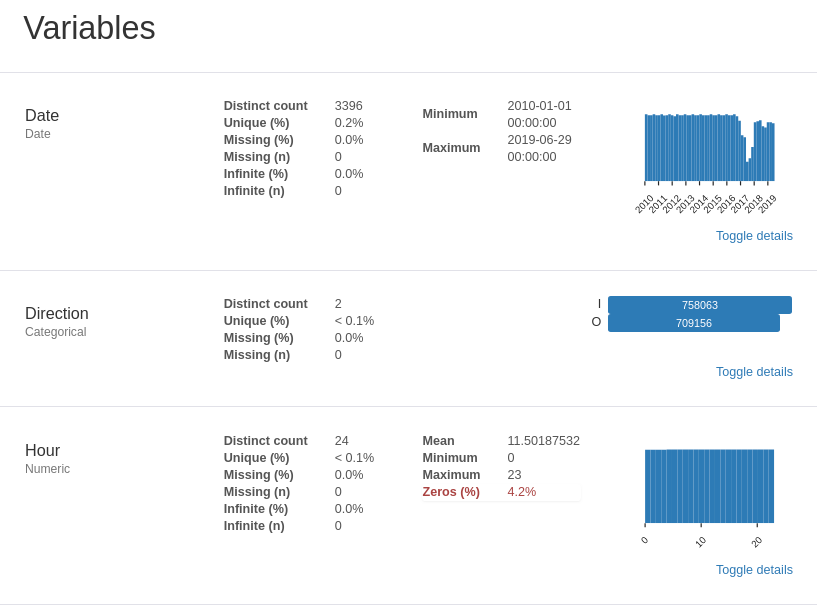
With the data cleaned, the next step was to do exploratory data analysis on the dataset see what the data shows us in detail and check for any missing values and/or outliers.

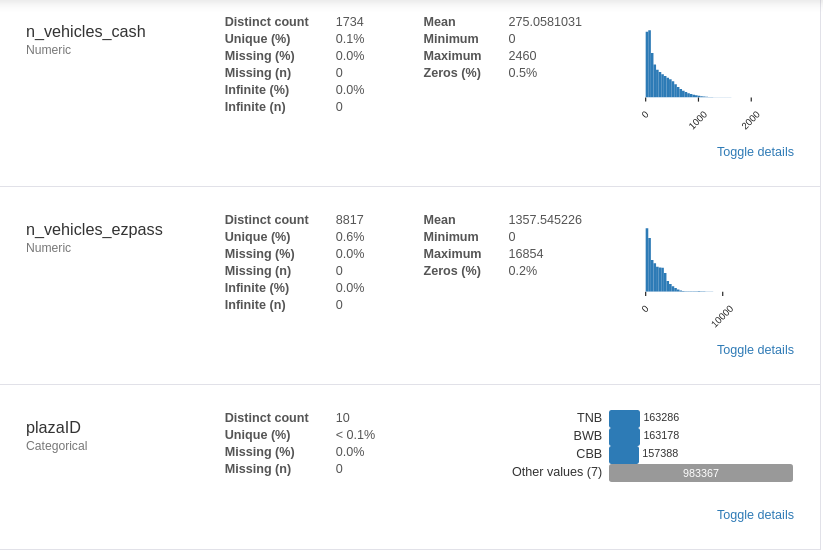
### Exploratory Data Analysis

Using a single line of python code and using the ‘pandas\_profiling’ module, we can see many things such as correlation amongst the features, missing values, outliers, previews of the data and so on. What makes this a good function to use is that in a single line of code, one can gain insight on their data and is a time-saver when working on large datasets.

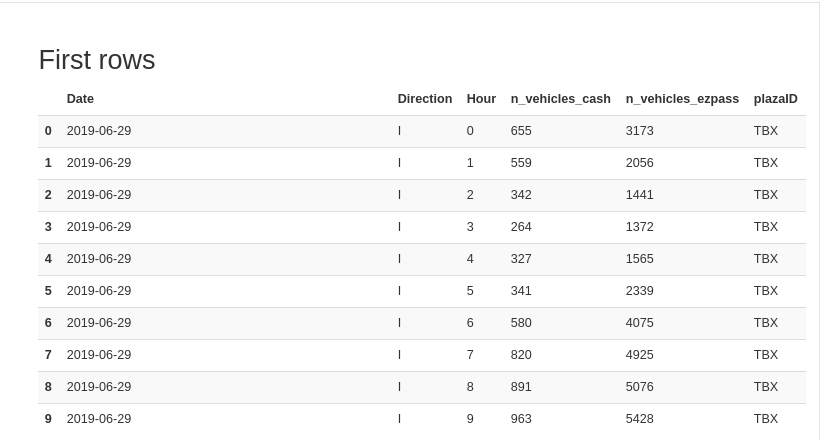
Below is an overview of the data such as the dataset information and the variables and their type. The duplicate rows is the fact that some rows had some columns repeated (for example the plazaID and Date columns. In addition, there were no rows removed from the dataset.

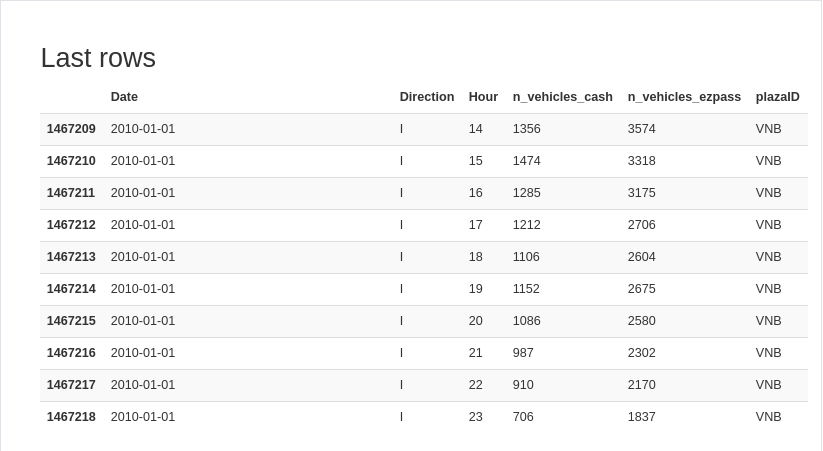




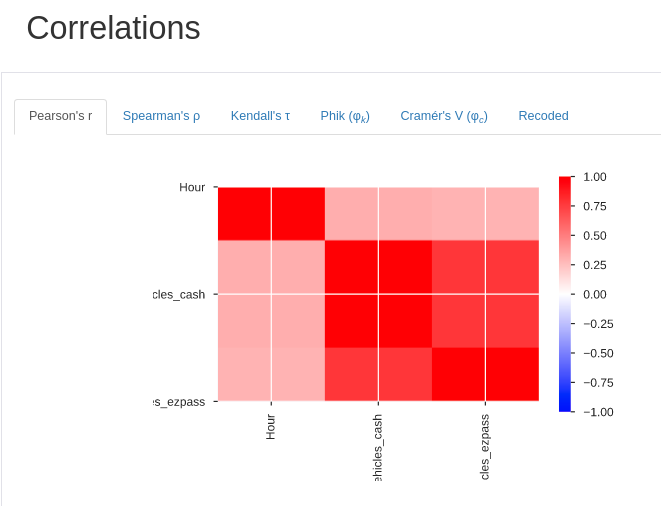


Preview of the data:





Using Pearson’s correlation amongst the variables, we can see that there are no negatively skewed correlations and that there are moderately positive correlation amongst Hour and the n\_vehicles\_cash and n\_vehicles\_ezpass variables. Also there are high positive correlations amongst vehicle counts for both cash and EZ-pass.



Based on this EDA, we see the following:

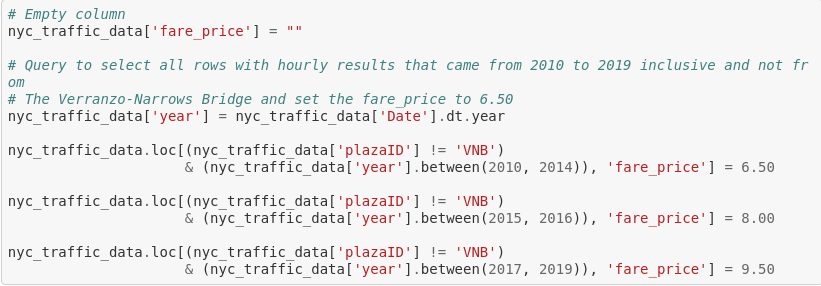
* There are no missing data which saves time having to decide what to decide on handling missing items.
* Around 2017, there are much smaller measurements around mid 2017 meaning that something may have occurred with the NYC DOT that caused less measurements to be taken during that time.

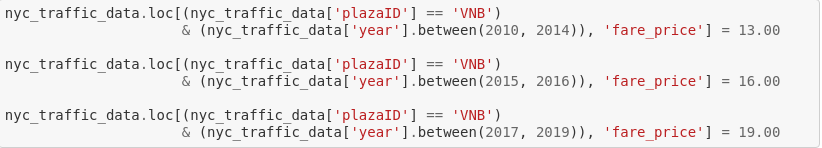
### Gathering Historical Fares

In order to train the regression algorithms that will be used for estimated vehicle traffic, we will need to find out how how much were the fares for several bridges and tunnels. According to several wikipedia pages regarding NYC bridges and tolls, majority of them charged $6.50 during 2010-2015, $8.00 from 2015-2017 and $9.50 from April 2019 afterwards $9.50. The Verrazano-Narrows bridge however, charged double for years mentioned.

For this project, we will assume all bridges and tolls were the prices mentioned above except for the Verrazano-Narrows bridge.

The next step was to create a new feature called ‘fare\_price’ and append the fare\_prices for each bridge and year.





Once that was done, the next step was to begin training machine learning regression models and seeing which one has the smallest RMSE and MAE values.

### Training Machine Learning Models For Predicting Vehicle Volume

With all the data ready to go, three different types of regression algorithms were to be used and compare them based on metrics to see which one would be ultimately be used. The explanatory variables were the plazaID, Hour, year, Direction and, fare\_price. The response variable was the n\_vehicles\_ezpass. Either the number of vehicles in cash or in EZ-pass could have been used but the idea was to see how many vehicles would go through each bridge and tunnel via EZ-pass.

To prepare the models to be trained, the variables plazaID and Direction had to be encoded into numerical values so this way the regression models would work and that there would be no errors if the data were strings and characters. A method called one-hot encoding was used to do a quick mapping scheme to map the plazaID’s and Directions to integers.

Next, the dataset was then split into a 70/30 training/test set meaning 70% of the data was chosen randomly to train the models and the remaining 30% was used to make predictions and evaluate the model.

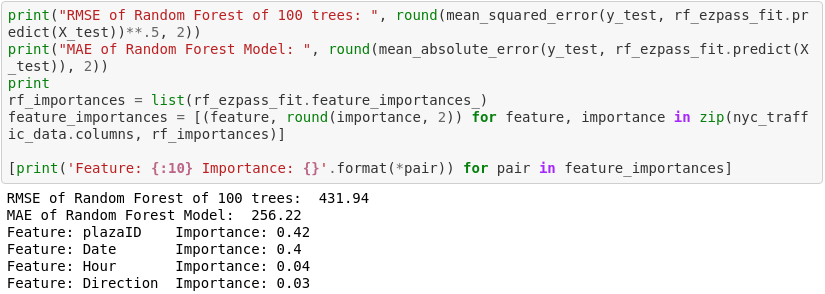
For multiple linear regression, the approach was to use all the explanatory variables in order to train a linear regression model and see the results. For this, once a best-fit model was used for predicting EZ-pass usage, the root mean squared error and mean absolute error were calculated where the RMSE was approximately 1030 and the mean absolute error was about 790.

Next was the elastic net model. This is a regularization model that prevents overfitting and can also penalize features and thus do feature extraction. Using the same training/test set, the results of the RMSE and MAE were about the same as the multiple linear regression model.

The last model that was deployed was the random forest model. The idea was to use the default of 100 trees and once the model was run, the RMSE and MAE were smaller around 430 and 256 respectively. As this model had the smallest value in terms of these metrics, it was going to be the algorithm of choice.

Looking at the most important features from the random forest model, it was shown that the plazaID and Date were the most important features contributing to about 42% and 40% respectively. Other variables like Hour and Direction only contributed about 4%. This should be reasonable as the traffic volume depends mostly on location and date and time.

Random Forest code and feature importance:



### Creating Simulated Fares

Since the random forest model worked well when predicting traffic volume, to see whether increasing the price would decrease the traffic congestion, simulated fare prices were created. These simulated prices ranged from 4,5,6…,20 dollars and the approach was to create a new feature ‘vehicle\_ezpass\_<fare price>’ where <fare price> is the simulated fare price. The data in each of these new columns would be the estimated ezpass vehicle count based off a random forest model.

To achieve this, a function was created that performed the following:

1. Appended the simulated fares using the same query as when appending the historical fares. Also, by doubling the simulated fare price for the Verrazano-Narrows bridge, the simulation also captures the historical fare price behavior.
2. With the simulated price, train a random forest model using the same 70/30 training/test split and also use the same features but only use the simulated price and not the original one.
3. Predict on the whole dataset with our model what the traffic would look like under the simulated fare price.
4. Repeat steps 1 through 3 for each simulated fare price.
5. The result will be an updated dataset that contains estimated traffic for each simulated price.

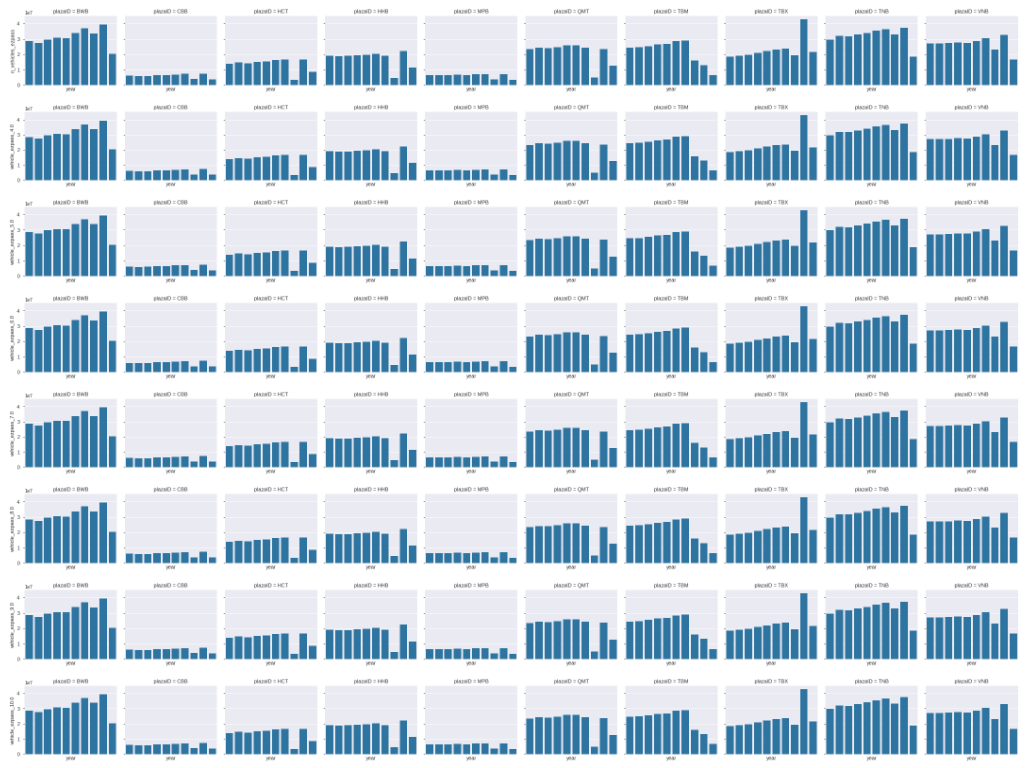
### Results and Conclusion

Once all the predictions were done for each simulated fare price, it was time to plot the traffic congestion over time for each simulated price. For each of the 10 toll plazas, a barplot was created to see how the behavior looked if fare prices were a given amount.

The below plots show the ez pass traffic congestion over time from 2010 to June 2019. Each column in the plot is based on a particular Location. Each row shows the result of each simulated fare price starting from the original to up to $20.00. Surprisingly, the traffic volume remained mostly the same.

From left to right the bridge/tunnels are BWB, CBB, HCT, HHB, MPB, QMT, TBM, TBX, TNB and, VNB.

From top to bottom are the traffic volumes of the original fare price, simulated price of $4.00, $5.00 up to $20.00.



### 

Based on these plots, the results are shocking in the sense that despite raising the fare price over time a fixed amount, the traffic volume remained about the same. This can lead one to believe that no matter the price increase or fixed high price, commuters will still use the bridges and tunnels regardless. This can also show that even when maximizing fare prices, traffic congestion might not be easily reduced.

In conclusion, we saw that for applying fixed fares for several prices, not much as changed. Some items to include for future research that may see change in traffic volume would be

* Including more features such as CO2 emissions, weather/temperature, vehicle type/make and model.
* Using a different regression model.
* Gather more data or even use a different dataset.
* Rather than fixed price, use similar simulated fares but with increasing pricing over time.

### References

Varone, Jason. “Variable Pricing at MTA Bridges and Tunnels Would Ease Traffic” StreetsBlog NYC <https://nyc.streetsblog.org/2007/09/26/variable-pricing-at-mta-bridges-and-tunnels-would-ease-traffic/> 24 Sept, 2007

Nessen, Stephen. “MTA And City Move Forward On Congestion Pricing” Gothamist <https://gothamist.com/news/mta-and-city-move-forward-on-congestion-pricing> 19 June, 2019

Bascome, Erik “NYC Congestion Pricing New Report Recommends Variable 2-way Fees” <https://www.silive.com/news/2019/09/nyc-congestion-pricing-new-report-recommends-variable-2-way-fees.html> 18 Sep, 2019