Jonathan Hernandez

DATA698 Project Proposal Draft

NYC Traffic Fare Levels and Volume

Introduction:

New York City’s Metropolitan Transit Authority (MTA) controls all public transportation of NYC. While many of New Yorkers are dependent on the MTA for commuting, I’ve decided to examine some trends regarding the MTA and see how data can help in learning how traffic fares affect volume and revenue around NYC. Perhaps someone at the MTA can use these findings to see how much subway fare hikes will affect commuters or can focus on more activity on the less frequent tunnels and bridges.

Data Sources:

Examining Traffic among NYC bridges and tunnels:

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

Literature Review:

According to a 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.

Another article from the MTA mentions about how in March 2019, the MTA wanted to look into congestion pricing alternative technology. The request is from the

Hypothesis and Model Simulation:

The hypothesis will be finding a 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. Other packages or repositories may be needed for creating visualizations of the data for various simulations.

For the approach of how to create simulations and model, different algorithms like linear/non-linear regression, random forests, neural networks, support vector machines will be looked into.

Looking at data from metro card swipes as well as tunnel and bridge toll transactions we can see how revenue is affected.

The 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. The columns 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 though each bridge or tunnel

- # Vehicles – Cash/Vtoll – Number of vehicles that paid in cash

Once the data has been loaded (via python pandas module using the read\_csv() function), it helps to examine whether there are missing values or not. We see using python’s isnull() and sum() functions that the observations don’t have any missing values which will probably make analysis and simulation easier.

# Check whether any columns have missing values or not

print(nyc\_traffic\_data.isnull().sum())

Plaza ID 0

Date 0

Hour 0

Direction 0

# Vehicles - ETC (E-ZPass) 0

# Vehicles - Cash/VToll 0

date 0

dtype: int64

Next, it helps for better readability and understanding that the plaza ID’s should be mapped based on the abbreviation of the bridge or tunnel. For example, ID = 21 corresponds to the Robert F. Kennedy Bridge Queens/Bronx Plaza (TBX) based on the “MTA\_HourlyTrafficBridgeTunnel\_DataDictionary” PDF file.

nyc\_bridges\_tolls\_names = {1: 'TBX', 2: 'TBM', 3: 'BWB', 4: 'HHB', 5: 'MPB',

6: 'CBB', 7: 'QMT', 8: 'HCT', 9: 'TNB', 11: 'VNB',

21: 'TBX', 22: 'TBM', 23: 'BWB', 24: 'HHB', 25: 'MPB',

26: 'CBB', 27: 'QMT', 28: 'HCT', 29: 'TNB', 30: 'VNB'}

# replace the numbers with their corresponding abbreviation

nyc\_traffic\_data['Plaza ID'] = [nyc\_bridges\_tolls\_names[item] for item in nyc\_traffic\_data['Plaza ID']]

The reason why some of the abbreviated values have multiple keys is that the Plaza ID’s from 1-11 were before Open Road Tolling switch over dates which took place mostly around 2017 and 21-30 afterwards.

Next Step is to then start doing Exploratory Data Analysis and start seeing how not only the volume of traffic but the flow of traffic based on each pricing model. As the number of tolls and bridges in NYC is small, visualizations can be made for each location.

For model building and simulation, different algorithms like linear/non-linear regression, random forests, neural networks, support vector machines will be looked into. The idea is to split the data into training and test sets and see which models can help give the most accurate simulations of volume based on fares.