Jonathan Hernandez

DATA698 Project Proposal Draft

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)

References

### 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. 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.

The following python code snippet is a primer on how the data is loaded and data transformation is done followed by a facet grid that contains histograms of the number of EZ passes based on each location and direction of traffic:

|  |
| --- |
| **import** re **import** pandas **as** pd **import** seaborn **as** sns **import** matplotlib.pyplot **as** plt  """Importing the dataset of toll and bridge volume in NYC staring from 2010"""  nyc\_traffic\_csv = "hourly-traffic-on-metropolitan-transportation-authority-mta-bridges-and-tunnels-beginning-2010.csv" nyc\_traffic\_data = pd.read\_csv(nyc\_traffic\_csv)  print(nyc\_traffic\_data.shape) print(nyc\_traffic\_data.columns) print(nyc\_traffic\_data.describe()) print(nyc\_traffic\_data.dtypes)  *# Are there any missing values in the dataset* print(nyc\_traffic\_data.isnull().sum())  print(nyc\_traffic\_data.head(n=10))  *### Data cleaning*  *### Replace Plaza Id's with actual bridge/tunnel names* *### Per the MTA\_Hourly Traffic Bridge Tunnel\_Data Dictionary.pdf values*  *# 1 -11 are the original historical values before tolling switchover dates* *# 21 - 30 are the values after the open raad tolling switch over date (around 2017)*  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'}   nyc\_traffic\_data['Plaza ID'] = [nyc\_bridges\_tolls\_names[item] **for** item **in** nyc\_traffic\_data['Plaza ID']]  *# Replace the Dates with simply the date only as the time is not needed given*  *# the Hour column*  *# remove the time after "T"* nyc\_traffic\_data['Date'] = [re.sub('T.\*', "", date\_time) **for** date\_time **in** nyc\_traffic\_data['Date']]  *# format the date* nyc\_traffic\_data['date'] = pd.to\_datetime(nyc\_traffic\_data['Date']) |

The code snippet shows that after the data is acquired, changing the plaza ID’s to their appropriate abbreviations can help in visualization and analysis and can be better understood in terms of the project.

Also by using the isnull() and sum() functions, we can see that there are no missing values in this dataset which will make the cleaning and analysis easier as one does not have to decide on how to deal with the missing data.

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.

### 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