Jonathan Hernandez

DATA98 Project Proposal Draft

Examining Various Trends of NYC MTA Activity

Introduction:

New York City’s Metropolitan Transit Authority (MTA) has been around for many years and heavily controls the majority of transportation of NYC. Anyone who has lived in NYC or even visited the city that never sleeps knows that besides taxis, MTA transportation is always around if needed to move around. It is also perhaps the most popular and busiest transportation system in the country. 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 about how passengers and drivers move 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.

Research questions and Hypothesis:

Question 1) Are drivers more likely to drive through MTA operated tunnels and bridges in the AM more than the PM?

Question 2) Can we predict the total amount of MTA Full fares for future weeks?

Question 3) Has number of full fare uses decreased while subway fares increased?

These questions will enable us to see how traffic activity is in NYC as well as examine how fare increases have affected toll payments. One can use this data to better predict how people will use MTA as fares increase and traffic flow for better monitoring.

Data Acquisition:

Data will be acquired from Kaggle (website for data science competitions and datasets)

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

Python will be used to analyze these datasets and various python libraries for data analytics and using regression techniques and/or time series methods will be used to answer questions.

For question 1, we can use hypothesis testing where the null hypothesis can be if drivers are more active in the PM than in the AM and the alternative hypothesis can be the opposite.

Question 2 can be solved or predicted using machine learning techniques and as data is updated, we can test our predictions so using a training/test set might be a good approach.

Question 3 can be worked on using possibly a Time Series approach (moving average, ARIMA model).

Data Visualization (Exploratory Data Analysis)

Using Python’s libraries like matplotlib and seaborn to create bar or time series data and look into other libraries to perhaps plot geo-spatial data to see visually where in NYC are there metro card swipes.

GENERAL COMMENTS

Jonathan,

While the analysis of traffic patterns is a viable subject matter for a research project, you need to develop a more ambitious scope than simply comparing and visualizing different amounts of drivers and fares.

I don’t mean to be overly critical, but what you’re suggesting in your draft amounts to descriptive statistics. So what would represent a sufficiently sophisticated model? As a suggestion, you could develop a simulation of a graduated/extrapolated pricing levels of fares (starting with the previous pricing and including the new surge pricing) to maximize revenue while minimizing traffic congestion at peak (rush hour) times of the day.

The choice is completely yours; you just need to choose/propose a more complex/challenging task for your senior research project in this program. Data Science, after all, is a discipline of designing and developing models that some the relationship of different variables in order to better understand phenomena.

As I’ve done with the proposal drafts of your other classmates, I’ve posted my feedback via track changes in the “Feedback on Draft Proposals” folder in the “Weekly Assignments and Feedback on Drafts” section on the BB course site. I’ve also attached it to this e-mail.

As always, please let me know if you have any questions.

A