EDA_Project_Proposal

Question/need:

- What is the framing question of your analysis, or the purpose of the model/system you plan to build?

Crime rate is soaring in the NYC subway system as a result of fewer riders during the pandemic. Specifically, an 8% increase in assaults was seen from the past year despite less riders. Riders in isolated stations and trains are becoming easy targets for a criminal. Understanding the rider traffic data from the past few months might locate areas and times of the day where caution is required.

- Who benefits from exploring this question or building this model/system?

Analysis on rider data might help authorities to come up with a prevention plan by adding extra security or warning riders travelling through risk identified routes to practice caution. Crime rate can be controlled eventually making subways safe for the public to ride again.

Data Description:

- What dataset(s) do you plan to use, and how will you obtain the data?

MTA Turnstile data from January 2021: http://web.mta.info/developers/turnstile.html is being used and imported into jupyter notebook utilizing pandas read function.

- What is an individual sample/unit of analysis in this project? What characteristics/features do you expect to work with?

Each row in the data which represents entry and exit of a rider is my unit of analysis. I aim to utilize features such as station, date, time, entry and exit.

- If modeling, what will you predict as your target?

None

Tools:

- How do you intend to meet the tools requirement of the project?

Data is downloaded and added to a database. SQLAlchemy will be used for querying on Python. Data cleaning and visualization will be performed utilizing pandas libraries in a jupyter notebook.

Pandas libraries such as numpy, seaborn and matplotlib will be used.

- Are you planning in advance to need or use additional tools beyond those required?

Maybe Tableau for additional NYC map visualizations

MVP Goal:

- What would a minimum viable product (MVP) look like for this project?

MVP will contain data imported from the website, ready to be queried using SQL on python. Using SQL queries on python specific chunks of data will be extracted.

Initial analysis is performed to learn about any nulls and data type inconsistencies utilizing pandas functions. Visualizations with matplotlib and seaborn will be used to recognize any outliers and patterns on specific days, times or stations in the data.

Source:

https://www.nydailynews.com/new-york/nyc-crime/ny-nypd-assaults-rise-subway-cops-20210218-gmn7tw5hlnhefnoagpnsu376wq-story.html

https://nypost.com/2021/04/18/subway-crime-still-outpacing-ridership-despite-drop-in-march/