

EDA Project Proposal

Question/need:

- To see if there is a relationship between traffic and criminal activities
- Will heavy traffic lead to more criminal activities?
- The Police department and the Office of public safety from New York will be benefit from exploring this question.

Data Description:

- Primary data source is the MTA turnstile data set which is given by the instructor
- The MTA turnstile data set can provide the traffic information for the whole New York state
- Secondary data source is the NYPD Complaint_Data Historic which is found on Kaggle.com
- The NYPD Complaint Data Historic set can provide the time and location of the crime committed in New York
- The evaluation period is twelve weeks from 2014/12/27 to 2015/04/03
- Mainly focusing on the traffic from each area, and the number of criminal activities for each area

Tools:

- Create a SQL database and the table with the labels for the MTA turnstile data set using sqlite3
- Use SQLAlchemy to extract the MTA turnstile data set into Python
- Explore the MTA turnstile data set and the NYPD Complaint Data Historic set in Pandas
- Filter the useful columns in Pandas for both data sets
- Use Matplotlib or Seaborn to graph out the number of traffic per area and the number of criminal activities per area and per day

MVP:

- To find out if there is a relationship between heavy traffic and the number criminal activities