EDA Project:

The Relationship between MTA Traffic Data and NYPD Crime Rate Data

Abstract

According to NYPD recent report the crime rate in the subway has increased by 93% compared to last year. Public safety is a concern to transit agencies and the NYPD and making
an action to solve this issue is required at this critical time. Subway crimes were never this high
and there are specific times and stations that have a higher crime rate than others. In order to help tracing the issue, this project used correlation model to find the relationship between the most crowded stations and the stations with highs crime rate. Finding out about the relationship between those variables can help in preventing more crimes or developing better solutions.
Both MTA and NYPD data were part of the data analysis process, and they were not the perfect datasets, and they need so much cleaning and filtering. After all those steps, there was no enough data to draw the conclusion that the busiest station has the highest crime rate.

Design

MTA data was provided along with NYPD crime reports, both data have the situations in common and sorting those stations by the highest in crime rate and the busiest during the period of 3 months (the last 3 months in 2020). Correlation model was created to find the relationship between those variables to understand the underlying issue of transport crimes.

Data

The MTA dataset contains 2738673 rows x 11 columns only two of them were numerical, while the NYPD dataset has 9167 entries with 35 features that vary in type between categorical and numerical. The data that was used from MTA database was weekly based and the one from NYPD was a daily report. Getting data such as the total number

of crimes per station and information about victims and criminals all were collected from the daily crimes' reports. However, information about the busy stations and total numbers of ridership on each of those stations was obtained from the MTA database. Analyzed both data sets help in formulating a better idea about when was the time that most crimes occur and who was the most targeted victims. The sample was all the people who use NYC subway in a regular basis.

Algorithms

1. Upload both datasets to the Jupyter Notebook.
2. Convert the daily report to weekly in the NYPD.
3. Cleaning the data, dropping duplicates and imputing any missing value.
4. Finding the sum of entries were taken for each station, considering the uniqueness of each of the "C/A, UNIT, and SCP" for each station.
5. Finding the busiest station was calculated from the sum of all the entries for each day of the period between Oct to Dec in each station.
6. At the end sorted them and took the largest 10 station in regard of traffic.
7. Mapping the result from both data.

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication