Identifying least populated subway stations in NYC

Abstract

NYC subway stations have been consistently making headlines because of rising crime inside the stations. One of the important reasons behind this soaring crime is the depleting rider traffic due to the ongoing pandemic from 2020. The aim of this project is to use MTA turnstile data to locate stations with extremely reduced rider traffic. Matplotlib and seaborn were used to recognize any outliers and patterns on specific days and times across different stations.

Design

The project is an attempt to find solutions regarding unsafe subway stations by utilizing the turnstile data. In depth analysis will identify sparsely populated stations during specific times of the day and days of the week which 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

Data was imported from the website into pandas dataframe using pd.to_csv. Each row represents recorded entries and exits across 378 stations over 4 hour time periods. Entry and exit columns represent cumulative values where each 4 hour values are added to the previous ones. Other columns include C/A, Unit, SCP, Station ,Linename, Division, Date, Time, Desc, C/A, Unit, SCP and Station represent a unique turnstile. As per meta data on the website, RECOVR AUD under Desc refers to system resetting. Data was cleaned by dropping duplicate rows and absolution of negative count of entry and exit values. Outliers were identified throughout the analysis and were dropped.

Algorithms

Key metrics used for analysis -

- Rider Traffic : Sum of Entry and Exit turnstile count
- Unsafe: less than 25% of average rider traffic for each day

Data was analyzed using pandas operations such as groupby, sort_values and loc. Cleaned data was visualized using histograms and line plots.

Tools

- Pandas,SQLite for querying data
- Matplotlib, seaborn for visualizations
- Jupyter notebook, google presentation for data display

Communication

- Jupyter notebook and Google presentation with key findings is available. Here's an example-