Exploratory data analysis on MTA turnstile data

Abstract

The goal of this project was to use eda on mta turnstile data to find the most crowded times of people who use the MTA service in the days before Independence Day. I worked on public data authority site where I choose the (June - July) periods of (2016 - 2019).

Design

This project originates from the SDAIA Academy Data Science Bootcamp. The data provided by MTA site. Exploring data via Python visualization such as seaborn library for best results in identifying the busiest times for Independence Day booth setup.

Data

The data is about Metropolitan Transportation Authority (MTA) Turnstile, and I obtained the data from MTA site.
The period of data I analyzed was 3 weeks of (June - July) from (2016-2019) to get best results. The data contains 2,383,351 row and 11 columns. Also, I use an external data which is about stations, the data contains the names of nyc stations as well as there boroughs', division, and linename. The data is about 748 row and 5 columns.

Algorithms

Feature Engineering

- Check for the duplicates by grouping C/A, time, date, unit, scp, and station and then remove them.
- Remove the spaces of columns by using strip.
- Add the nedded columns, such as:

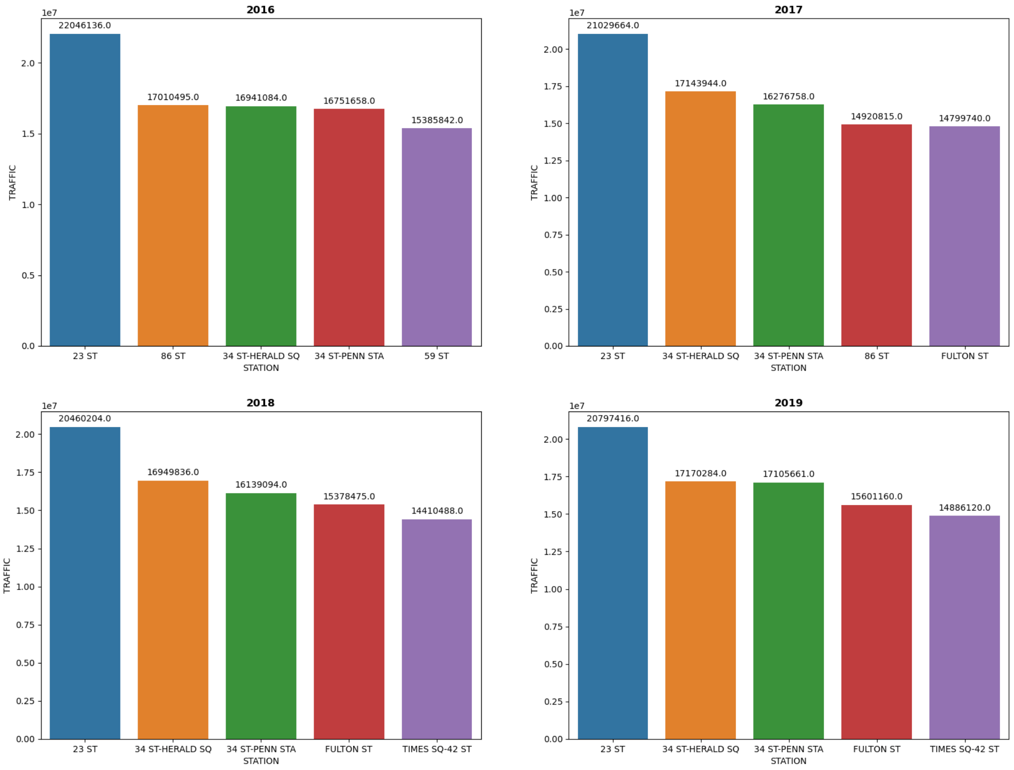
  Day name: day of the week.
  Turnstile:by grouping C/A, time, date, unit, scp, and station.
  Datetime: by concat the the date with time and then convert it to datetime object.
  Entries num: daily number of entries.
  Exits num: daily number of Exits.
  Traffic: by collect the daily entries and daily exits.

Tools

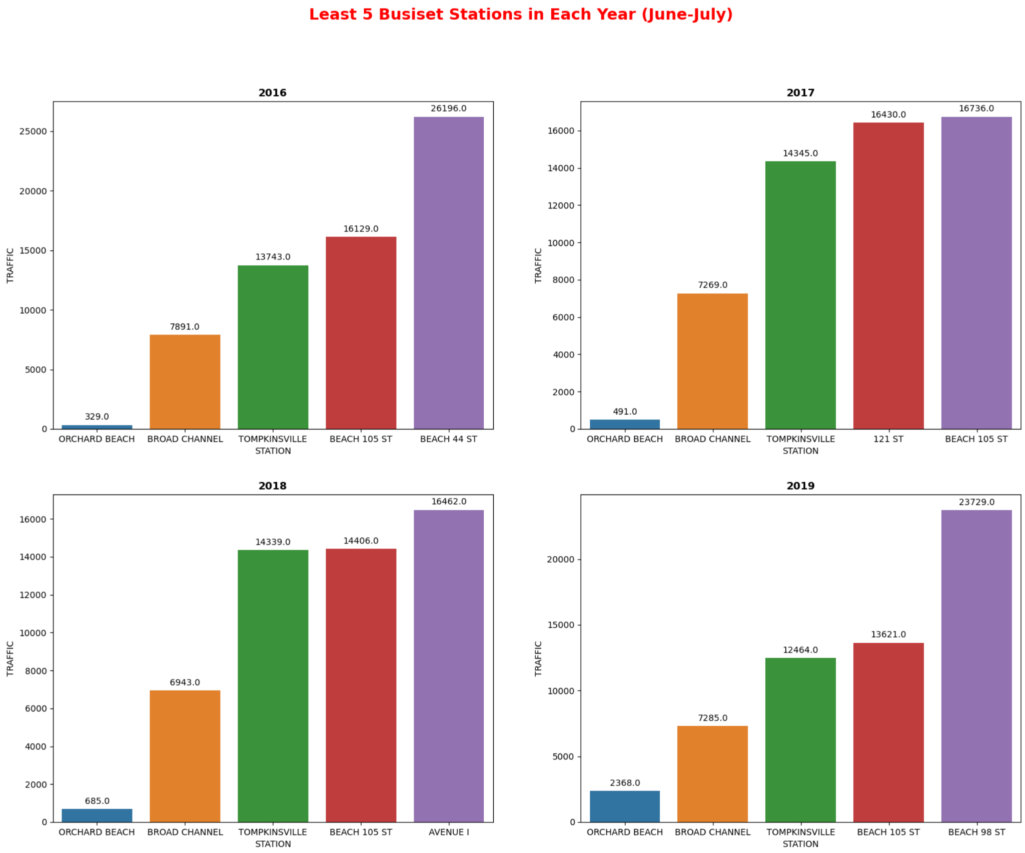
I used python, Jupyter nootbook and SQLite as technologies, where I use pandas, numpy, matplotlib, seaborn, and SQLAlchemy as libraries

Communication

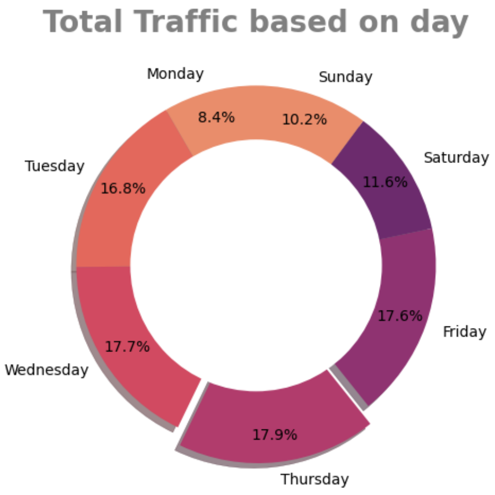
Top 5 Busiset Stations in (June-July)



The graph shows the top 5 busiest stations in each year by calculating Traffic, and we can see that 23 ST is the most busiest station for all years.

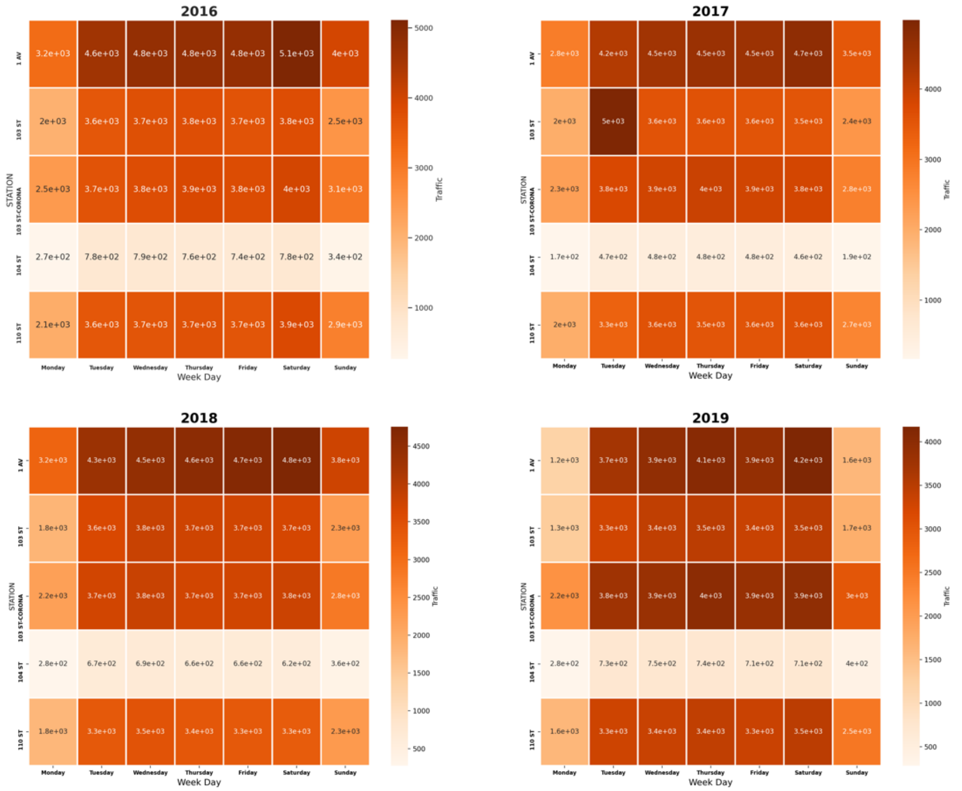


The graph shows the least 5 busiest stations in each year by calculating Traffic, and we can see that ORCHARD BEACH is the least busy station for all years.



The graph shows the total traffic for each day, which proves that the Thursday is the most trafficed day.

Top 5 Busiest Stations / Day



The graph shows the busiest days based on the top 5 busiest stations each year, proving that from Tuesday to Saturday, the stations are very crowded.