EDA ON MTA TURNSTILE DATA

Hayat Aldhahri

Data Scientist

Abstract

This report presents an Exploratory Data Analysis to assist New York Orphanage Organization in their upcoming fundraising campaign. The object was to find the most busiest stations within New York stations to locate the fundraisers. I worked with data sourced from (MTA Turnstile Data) available at the Metropolitan Transportation Authority. After performing an EDA on the dataset, plots were generated to visualize and communicate my results. Top five stations were found to be (34 ST-PENN STA, GRD CNTRL-42 ST, 34 ST-HERALD SQ, 23 ST, 14 ST-UNION SQ).

Design

This project provides insights to New York Orphanage Organization to achieve their ambition targets for the November 2021 Fundraising Campaign. The primary dataset was sourced from (MTA), which presents the patterns of transit traffic in New York City. Performing an exploratory data analysis EDA on the data set would assist the client to identify the most busiest stations across New York City and locate the fundraisers for the allocated period and preferred time period as well.

Data

The raw data contains 378 stations and has a shape of (2471446, 11). The data had the necessary information and details to execute the task. A three month dataset was considered appropriate to perform the task, considering that the campaign is set on November. The dataset ranged from September 2019 to November 2019. Datetime were corrected

Algorithm

1. Data Cleaning and sorting:

Data was cleaned were spaces and null values were removed from the dataset. And to ensure working correctly with the date and time, both columns will be combined in a new column and converted into datetime format.

2. Exploratory Data Analysis:

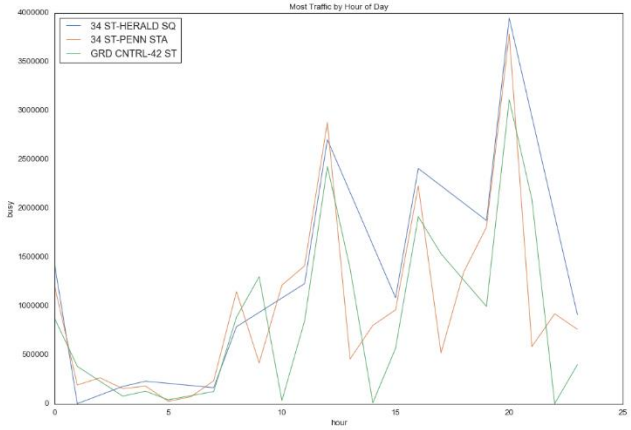
To find the answers for the main question asked, EDA was performed on the data. The data was first sorted by Turnstiles ("C/A", "UNIT", "SCP", "STATION") and ("DATE_TIME") to properly do the aggregation on the stations. Since “ENTRIES” and “EXITS” were cumulative figures, the difference were taken for each turnstile based on the above sorting to find the exact traffic through each turnstile.

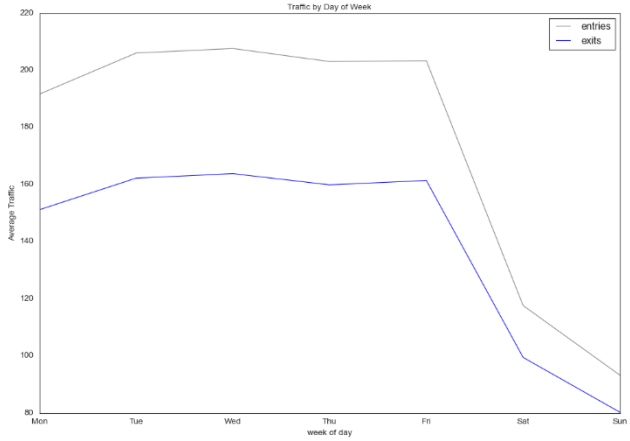
Tools

The following tools were used for the project SQLite, Python, SQLAlchemy and Jupyter Notebook. EDA through Python will be done using pandas, matplotib and seaborn.

Communication

Visuals can be further viewed in my GitHub profile (presentation slides and Jupyter Notebook).


Most Traffic by Hour of Day
34 ST-HERALD SQ
34 ST-PENN STA
GRD CNTRL-42 ST
busy
hour
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
5
10
15
20
25


Traffic by Day of Week
entries
exits
Average Traffic
week of day
220
200
180
160
140
120
100
80
Mon
Tue
Wed
Thu
Fri
Sat
Sun