Data Science Bootcamp

Project 1_Report

Exploratory Data Analysis (EDA)

For Metropolitan Transportation Authority (MTA)

Presented by: Ashjan Basri

9 September, 2021

I. Project overview

The public transportation crimes are increasing during the last couple of years in New York City (NYC). There are 27 crimes per 100,000 trips approcemilty reported in New yourk times news [2]. Therefore, Metropolitan Transportation Authority (MTA) is planning to make NYC subway stations safer for riders. Our objectives of this project are to reduce the crime rates at NYC stations, and find the optimal subway stations and time to place and distribute NYC security officer on the platforms who have the task of observing and securing the station areas and facilities. This project will analyse the MTA an open-source dataset for three months starting from April, May and June. The rest of the report is organized as the following: Section two presents data description. Section three presents analysis mythology. Section four presents data analysis tools. Section five presents experiment and analysis results and report is concluded with conclusion in section six.

II. Study Methodology

The methodology of this project as follow, firs phase is web scarping from MTA website [1] using Urllib to import the data. We used three months from 10/04/2021 to 26/06/2021. The total number of rows is 2511481 and the total number of columns is 11. We drop some of the columns that are not need in our analysis such as 'LINENAME', 'DIVISION', 'DESC'. Second phase is dataset cleaning such as drop duplicate values, missing values, Nan values and outlier values. Remove the irregular events in the data set. Third phase, apply the analysis methods includes calculating the total number traffic on the all stations but we want to look at the least crowded station that would have high number of crimes. So we calculate the previous entry subtracted from previous exits then the total traffic = sum entry difference and exit difference.

III. Data Description

Our project uses an opensource dataset MTA turnstile [2], we are aiming to scrape a total of three months of the data. The description of the dataset as follow:

Variables	Discerptions
C/A	Control Area (A002).
UNIT	Remote Unit for a station.
SCP	Subunit Channel Position represents a specific address for a device (02-00-00).
STATION	Represents the station name the device is located at.

LINENAME	Represents all train lines that can be boarded at this station. Normally lines are represented by one character. LINENAME 456NQR represents train server for 4, 5, 6, N, Q, and R trains.
DIVISION	Represents the Line originally the station belonged to BMT, IRT, or IND.
DATE	Represents the date (MM-DD-YY).
TIME	Represents the time (hh:mm:ss) for a scheduled audit event.
DESC	Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours) 1. Audits may occur more than 4 hours due to planning, or troubleshooting activities. 2. Additionally, there may be a "RECOVR AUD" entry: This refers to a missed audit that was recovered.
ENTRIES	The cumulative entry register value for a device.
EXIST	The cumulative exit register value for a device.

IV. Tools

There are some tools that we used for the analysis such as matplot, seaborn, pandas and NumPy.

V. Data analysis

After we prepared the data, we used SQL methods and aggregations techniques to analyse the data. Then we used some of visualisation tools such as seaborn and matplotlib to look at the distribution across stations and show the least crowded stations of the total traffic as figure 1 shows. The distribution is heavily right-skewed. So, it means that the distribution of the

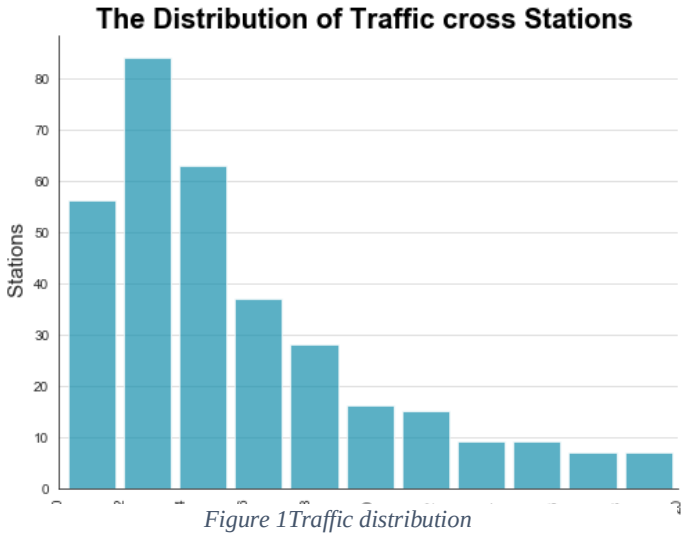


Figure 1Traffic distribution

traffic is not normally distribute. Then we calculated the average total traffic of the least station and we found Broad Channel is the least station that is not crowded with riders as figure 2 shows.

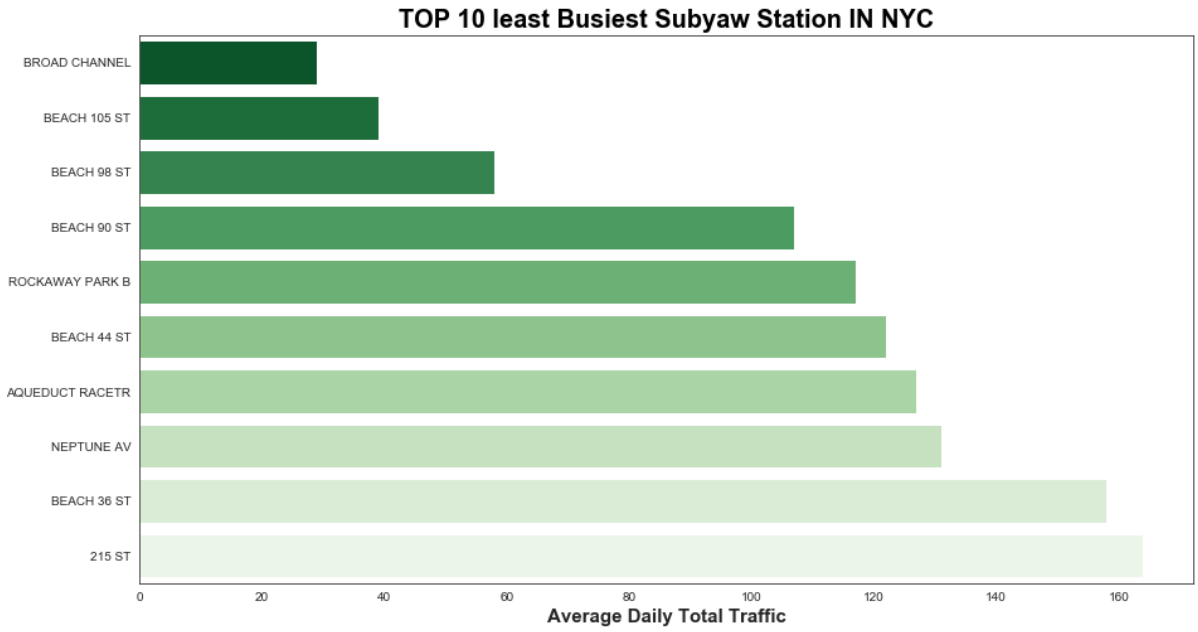
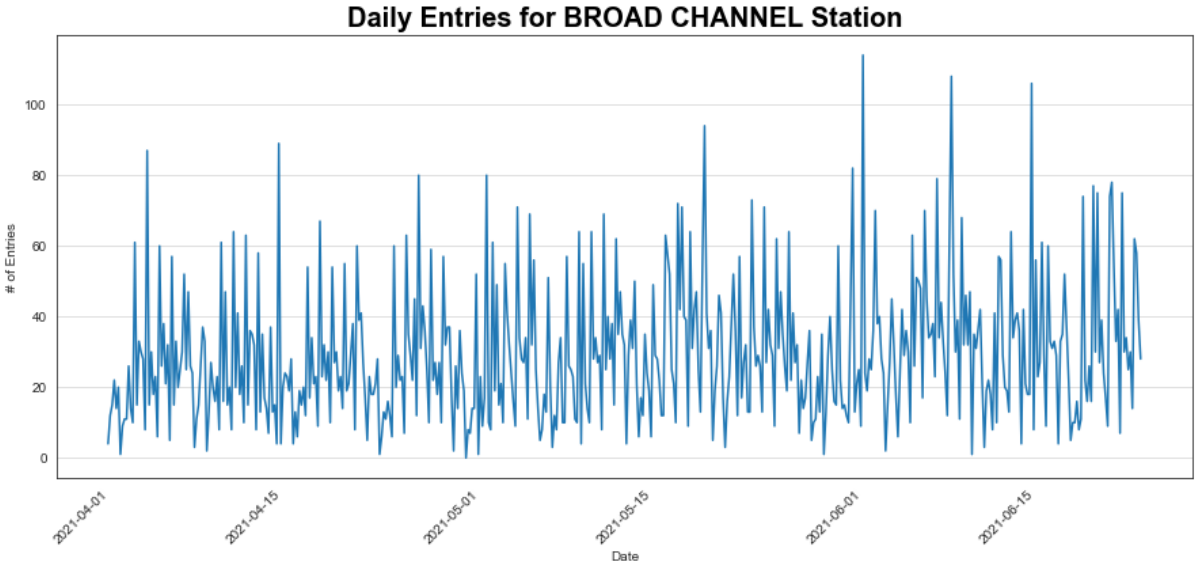
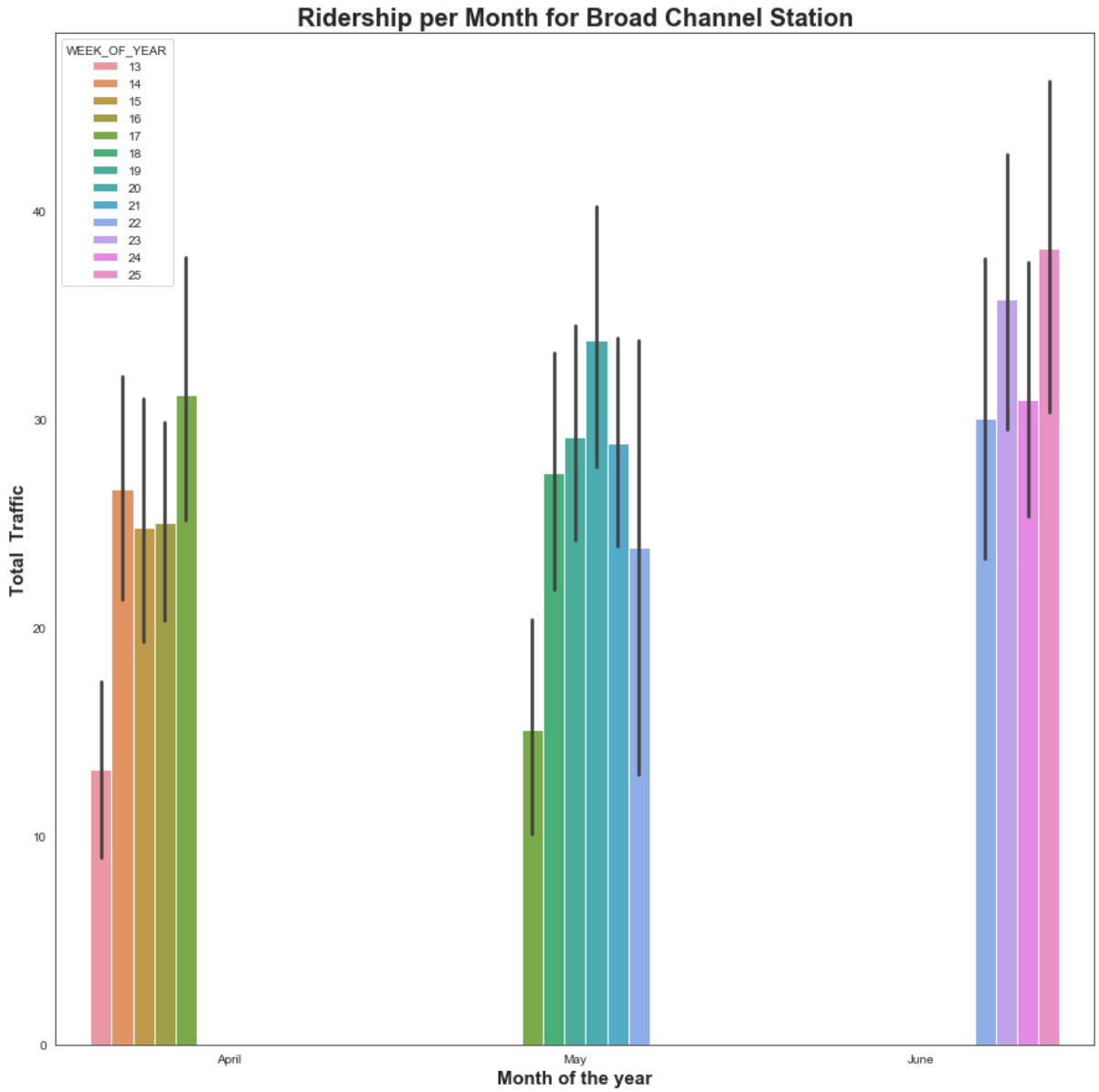


Figure 2 Top 10 least stations

We investigated more on Broad Channel station and saw the daily traffics and we found that the weekend has the least traffic comparing with the weekdays traffics as figure 3 shows.



Also, we analysis the total traffic on Broad Channel station on monthly and we found June has the highest total traffic which represent the summer time in New York as figure 4 shows.



VI. Conclusion

Broad Channel station is placed in the suburb far from the central city of NYC. Education level and low income are common in these areas therefore, there are a high rate of crimes For these reasons, we recommend the MTA office to focuses on the least crowded station such as Broad Channel station and distribute security officer to monitor the subway facilities. In our future work we would analyses the traffic pattern for riders during the day's morning and night to fiend the correlation between them.

Reference

[1] Eyewitness,abc,MTA survey suggests it's not just COVID keeping riders away from subway accessed on 1, august,2021, http://web.mta.info/developers/turnstile.html.

[2] Metropolitan Transportation Authority, accessed on 1, august,2021 http://web.mta.info/developers/turnstile.html.