Exploratory Data Analysis (EDA) Proposal for Outfront on the MTA turnstiles

Leena AlQasem
Leenabdulh@gmail.com

Banan Alhethlool
banan.alhethlool@gmail.com

- ## Abstract

The first project of Data science Bootcamp T5 is called the Exploratory Data Analysis (EDA) for the MTA dataset turnstile using python libraries such as Pandas, NumPy, and extra. Below, we will shed light on the company, Outfront media, that we cooperate with to assist them with their issue and help them with the dataset; After that, we have the dataset description and the scope and the methodology of the project, and analysis will be described with results. Finally, we have a recommendation for the Outfront to solve their issue.

- ## Business objective

Outfront is one of the largest media and outdoor displaying advertising companies. It provides an integrated and target platform. Nowadays, commercial advertising is necessary, and due to the nature of the subway audience, ads are seen by millions of people multiple times a day. Therefore, knowing rush hours, idle times, and whether Covid-19 has an effect or not is essential because it is a cost-effective option for both Outfront and businesses planning to work with them.

- ## Methodology

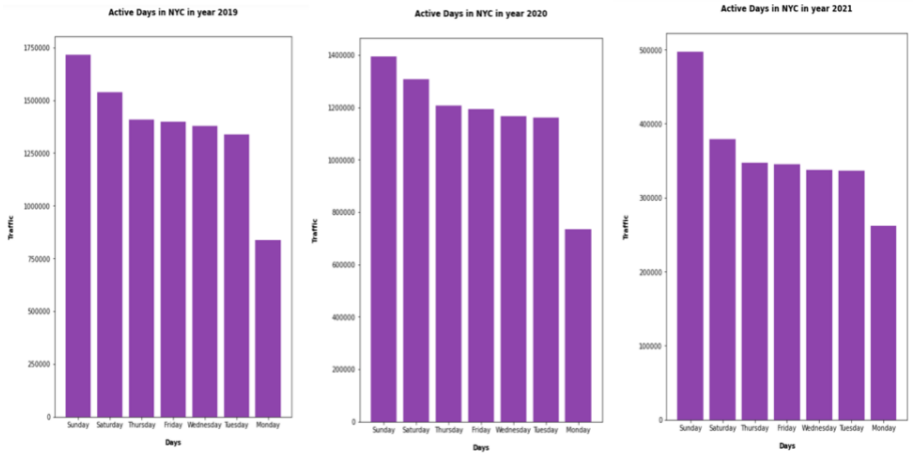
In this section, we first understood the company's problem by asking what the goal of the dataset is? Then, we gathered the data from the MTA turnstile dataset, a public source obtained from (http://web.mta.info/developers/turnstile.html). Furthermore, we started working on the dataset by cleaning and filtering, such as removing whitespaces, formatting the date and time, handling missing values, removing zero values, handling negative numbers, duplicates values, and outliers. Finally, we found the traffic by finding the daily entries and analyzing it with the station, days, and period time to help Outfront finding the rush hours for displaying the ads. So, to succeed in this project we used some technologies such as Jupyter Notebook, Python, SQL, and SQLlite. Also, we used Libraries in python for such as Pandas, NumPy, Matplotlib, and Seaborn.

- ## Analysis and Results

In the analysis and result section, we divide our analysis into three parts for each year. Hence, we first visualize Active Days, then Active periods, And finally, the Top 10 Crowded Stations. Before explaining each part, every year from each analysis has slight differences.

- **Active Days**

In the active day, we observed that weekends are more active than weekdays, especially on the first day of the week. In addition, after Covid-19, the number of people using public transportation has decreased year by year.

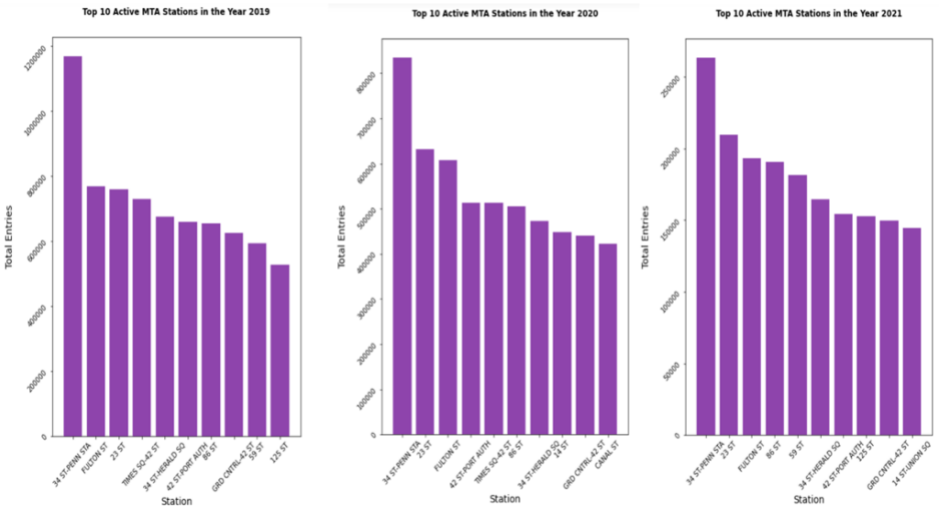


- **Active Period Times**

In the active period time, it is same as the describe previously about covid-19, so as shown in the heatmap, darker color means more traffic. Therefore, afternoon and night in the march were less traffic than in February in the morning and afternoon.

- **Top Crowded Stations**

Finally, we have the top 10 crowded stations, and in all three years, the crowded top station is 34 Street Penn Station.



- **Recommendations**

After exploring and analyzing the dataset, we recommend Outfront put their advertisement more on weekends cause ads are not displaying forever. Maybe it takes only one day, two days, or maybe hours. So, if its hourly ads its suitable to choose periods time that have rush hours, which is in afternoons. And finally, for more vistors and views we recommend to choose at least the top 5 stations such as 34 Street Penn Station.