Exploratory Data Analysis (EDA) on the MTA turnstiles Data

Randa Almohammadi Randa1414@gmail.com

Modhi Alhbrdi Modhi.alhbrdi@outlook.com

Abstract:

Data services team looking for the best solutions for the clients, Metropolitan Transportation Authority in the most important cases in the use of metro stations for the largest cities in the United States from the business and community aspects. And reach the required satisfaction levels serve the community as well as business needs.

Business Objectives:

Most of New York subway stations are very crowded, and a need for the New Yorkers day life. The project here takes the benefit of time spending at the stations, for the MTA and the stations users experience. Since the station visitor time spending is the key, for MTA financial satisfaction needs, we provide the solution by interactive, safe and entertain environment for the station visitors.

Therefore, the model we are planning to build using this data is, designing a place where entertaining shows can be performed at the most crowded stations and at the busiest hours. This model could include many aspects, increase visitors' awareness of the local places near the station, such as museums, restaurants, and other tourist sites, the social responsibility to support local talents and community campaigns.

Approach and methodology:

The data used in this project is taken from the MTA website MTA turnstile data. It presents the complete status of the metro stations in New York City, with eleven features. Three months of the MTA has been selected to be used in this project which are November, December of 2018, and January of 2019. Those three months were selected due to the new year holiday and before the pandemic of covid-19.

Tools: Python with Jupiter Notebook

- Numpy, Pandas and datetime libraries for data manipulation.

- Matplot, and seaborn libraries for data visualization.

- SQLite, and Sqlalchemy library on Python to connect with database.

Methods:

The methods used for this project are as following. Data cleaning, data analysis, and data visualization. On data cleaning, null values were checked for, and there were none. Then, duplicated values were dropped. Records have been selected only from the regular time audit. After that, specific columns were selected and other were dropped for the purpose of obtaining a good analysis. Next, data were sorted and analyzed through calculating the

difference between entries and exits which will provide accurate number of people entering and exiting the stations for every specific time frame. Specific day of the week were added to the data frame in order to know the total of people entering or exiting the stations in a specific day of the week. Outliers were dropped in order to make the analysis more accurate. Finally, data visualization was used for displaying data analysis and results on charts. It provides a clear and an easy way to read such big data.

Analysis:

The data of MTA turnstile were analyzed through finding the top five busiest stations in NY city ranked by average daily traffic. TIMES SEQUARE-42 station appears to be the top one busiest station. Then it was followed by 14 St. UNION SQUARE, 42 St. PPRT AUTH, Fulton St, and 59 ST Columbus respectively. Those stations were determined based on the number of entries. The analysis was narrowed down to find the average daily traffic on the top five busiest stations. It shows that weekdays are busier than weekends, and the traffic will be at its highest mid-week. The traffic was calculated by using the number of people entering the stations. In addition, the traffic on the top five busiest stations were recalculated using the number of people exiting the stations, and the result of that was the same top five stations. Therefore, both entries analysis and exits analysis shows the same top five stations to be the busiest.

Results and Recommendations:

The results of MTA turnstile data analysis show that the traffic on the top five busiest stations is very close to each other. Therefore, any station from the top five can be used to implement the idea of preparing a place for entertaining shows. Using the top one which is the times square station is recommended since it has the most traffic on NYC. The other 4 station can be used for activating campaigns. We found out that the times square is busiest on the morning at 11 am. This time can be used for targeting students and work-related business. The other busiest time is at 7 pm which we recommend to be used for local bands and other community related shows. We are able to determine specific busiest hour of the day for every station in the top five. This time can be used to offer it to companies who would like to come and perform advertising entertaining shows.