Exploratory Data Analysis of the MTA Turnstile Project proposal

Rawan Al-Ahmadi

SDAIA Academy

Mohamed Al Moghalis

September 29,2021

Project proposal:

In this project the task is to use the MTA turnstile Data to help a non-profit organization called 'The kids are the future', hand out an event flyer to the summer charity event that they are throwing for children with cancer. The organization expect to achieve these goals:

- Getting enough exposure to advertise for the event and the organization work.
- Target the high-to-medium-income commuters of the subway in New York city.

The Approach:

My plan in the project is to find:

- What the stations that have a lot of commuters during the morning/noon period. To determine what are the best places that volunteers can be assigned at to distribute the event flyer and advertise for the charity work.
- The stations where people are more likely to donate or attend the event due to economic status.

Data Description:

- The primary data I am going to use in this model is the MTA Turnstile data obtained from the official website:
  - "mta.info | Turnstile Data".
- Also, I will be using the census data to determine what are the wealthiest neighbourhoods (high household income), in New York city.
  - "Census - Geography Profile".
- I would be using the data of summer 2019 which is the months from late May through early-mid October, so the months I will collect data from are:
  - June, July, August, September of the year 2019.
- In this project I am expecting to analyse the number of entries and exists from the metro stations along other characteristics such as time, date, most commuted stations, trends in transfer stations, and understand what the relationships between all these characteristics using statistical methods as:
  - Correlation, covariance, mean, median, mode, standard deviation, and regression.

An example of the data that I will explore during the analysis:

- I will use the number of entries plus the exits to find out the traffic number for a specific station during a day, week, and then eventually, during the whole duration of the data.
- I will use the date and the time data with the station name to find out the best times the volunteers at the charity can best reach commuters.
- I will join the census data using the location of houses column with the location column of the busiest metro stations that I will get from the MTA data.

That is some of the data that I thought of exploring first to better understand the data, before starting my analysis.

- After accruing the busiest station's locations, I will look at the surrounding neighbourhoods to assess the economic status of the subway commuters.
- My targets are subway commuters with high-to-medium income, traveling through the city of New York during the morning/noon period.

Tools:

- In this project I will be using:
  - Jupyter
  - Git
  - Matplotlib
  - Seaborn
  - Sql
  - SQLite
  - Potentially I will use other visualization tools like tableau, pandas profiling.