Hayat Aldhahri

Data Scientist

Project Proposal

EDA on MTA Turnstile Data

# Introduction

I was assigned by the New York Orphanage Organization to assess them in their upcoming fundraising campaign in November 2021 by analyzing the data available from Metropolitan Transportation Authority. I will be performing an Exploratory Data Analysis using the freely available data from the MTA website. Our goal is to identify the location or areas with the most traffic to place the fundraisers at those locations. Since the campaign will last for one week, from 15th to 21st of November, I will analyze and explore data of 3 months. By the end, I will provide recommendations and insights to improve the fundraising campaign to help the organization achieve its target. Below shows a timeline for the project.

# Data Description

The primary data set I will be utilizing is sourced from ([MTA Turnstile Data](http://web.mta.info/developers/turnstile.html)) available at the Metropolitan Transportation Authority. The dataset available from year May 2010 to current data (September 2021) on weekly basis. The below table highlights the field description of the dataset. I will be using 3 months data from year 2019 (September to November). Since COVID 19 Pandemic affected the transportation in year 2020, data from 2020 were not included in the analysis.

|  |  |
| --- | --- |
| Field Name | Description |
| C/A | Control Area (A002) |
| UNIT | Remote Unit for a station (R051) |
| SCP | Subunit Channel Position represents an 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 |
| 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) |
| ENTRIES | The comulative entry register value for a device |
| EXITS | The cumulative exit register value for a device |

# Algorithm

Diagram

Description automatically generated

Following the above chart and workflow with data, I will start the process by first cleaning the data. The below are major activities:

* Removing duplicates.
* Remove null values, if present.
* Check if there are any inconsistency in texts and typos and correct them.
* Remove outliers that could affect the analysis.

To ensure working correctly with the date and time, both columns will be combined in a new column and converted into datetime format.

Since ENTRIES and EXITS fields are cumulative, I will get the difference in turnstile counts using panda built in functions. This will ensure that the total traffic (entries/exits) are determined for each specific period.

# Tools

To achieve the desired results and perform EDA on the dataset, I will be using different tools including, SQLite, Python, SQLAlchemy and Jupyter Notebook. EDA through Python will be done using pandas, matplotib and seaborn.

# MVP Goal

The MVP Goal would be looking at one station as an example and explore the traffic for the period specified and visualize the results. MVP will be provided on Monday 27th September to get the feedback and progress on the analysis.

# Conclusion

In conclusion, the outcomes from the analysis will support New York Orphanage Organization to locate the fundraisers on most busiest areas and achieve the campaign target.