# Exploratory Data Analysis on MTA Turnstile Data

**Prepared by:** Sara Al Abdulsalam

**Abstract:**

The goal of this project was to explore the data of New York MTA data to help give insights for a metro stations reconstruction purpose. I worked with a data publicly provided by Metropolitan Transportation Authority site where I chose to work with the first five months of the year 2021, visualize the data using python matplotlib library. After visualization, I extracted the valuable information and reflect it on the problem.

**Design:**

This project originates from SDAIA Academy data science bootcamp -T5- EDA model project. The data is provided by Metropolitan Transportation Authority site. Exploring the data via python visualization libraries such as matplotlib library which would enable the New York metro station to take actions of reconstructions to improve the riders experience.

**Data:**

The data that will help us taking insights about the nature of metro stations congestions days, hours, and stations is a from MTA website that provides a series of data files containing numbers of cumulative entries and exits by stations, turnstile, with their dates and time specified. The metro data records are weekly produced and mostly collected every 4 hours.

In this project and in order to carry out the insights I will use the first five months of the current year 2021 data.

Features provided in the dataset are:

- C/A = Control Area (e.g., A002) which is a string
- UNIT = Remote Unit for a station (e.g., R051) which is a string
- SCP = Subunit Channel Position represents an specific address for a device (e.g., 02-00-00) which is a string
- STATION = Represents the station name the device is located at which is a string
- DATE = Represents the date (MM-DD-YY) which is a data type
- TIME = Represents the time (hh:mm:ss) for a scheduled audit event which is a time type
- ENTRIES = The cumulative entry register value for a device which is integer

- EXIST = The cumulative exit register value for a device which is integer

I will add more features which are the following:

- Turnstile = which is a combination of C/A + unit + SCP can be used to locate the near by places around the turnstile on google map
- entries_num = which is the number of entries for the station timestamp observed by taking the difference of the cumulative entries and the previous one.
- exits_num = which is the number of entries for the station timestamp observed by taking the difference of the cumulative exits and the previous one.
- Weekday = the weekday name to distinguish between weekdays and weekends.
- Day_type = is the day weekend or a weekday.
- Congestion = which is the number of entries and exists added up to know how busy the station is.

**Algorithms:**

Feature Engineering:

- Extract the exact number of entries and exists by taking the difference between the records grouped by Turnstile identifier sorted by date and time.
- Extract the day name based on the date of the record.
- Classify the Date to weekday or weekend, since it's an American data so Saturday and Sunday are weekends.
- Extract the total traffic which is the summation of entries and exists.

**Tools:**

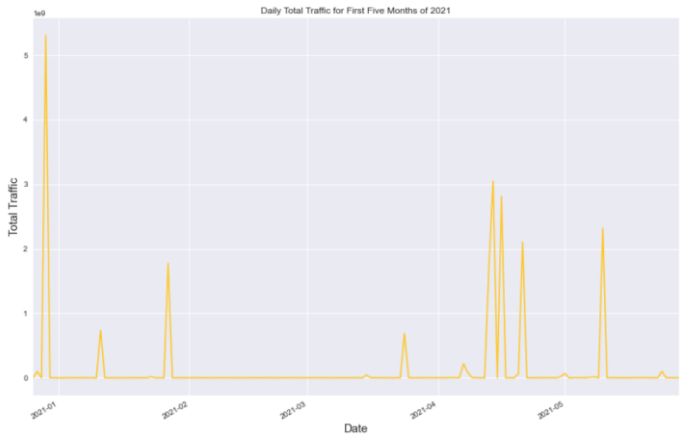
To carry out the project and explore the data, I will be using Jupyter lab to use python language. In addition to Python library which are:

Matplotlib, and Seaborn for data visualization.

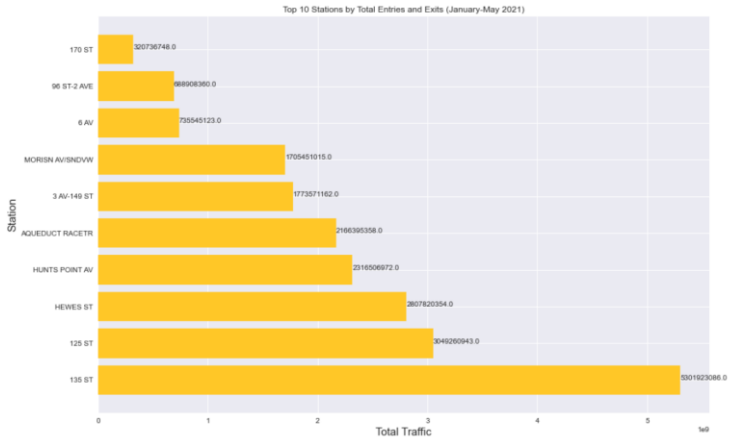
Numby, and Panda for data read and write operations.

**Communication:**

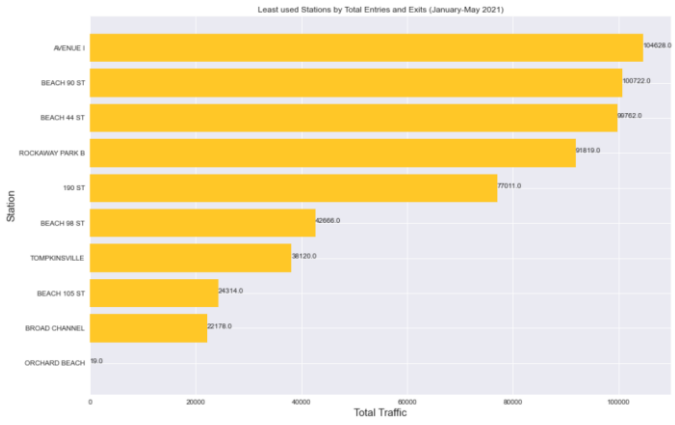
In addition to the slides and visuals presented, here I show the charts



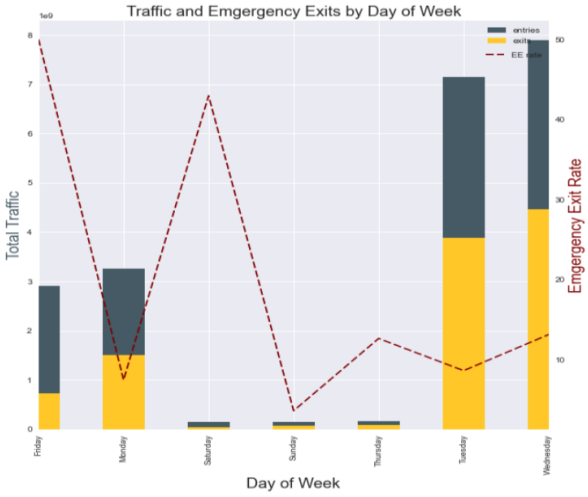

As we can see the total traffic in the first month of 2021 wan unnormally large.




Here are the top 10 stations in terms of traffic and we can see that *135 ST* is the most trafficked station.




Here are the bottom 10 stations and we observe that *ORCHARD BEACH* is the least used. Least trafficked stations can be closed and enlarge the most trafficked stations.




As shown in the figure some riders use Emergency Exists for non-emergence events to get out faster. We also observe that the stations are not used on the weekend as much as weekdays. However, we cant be assured due to the nature of our dataset which is a skewed data (More weekdays than weekends) as shown on the pie chart below.