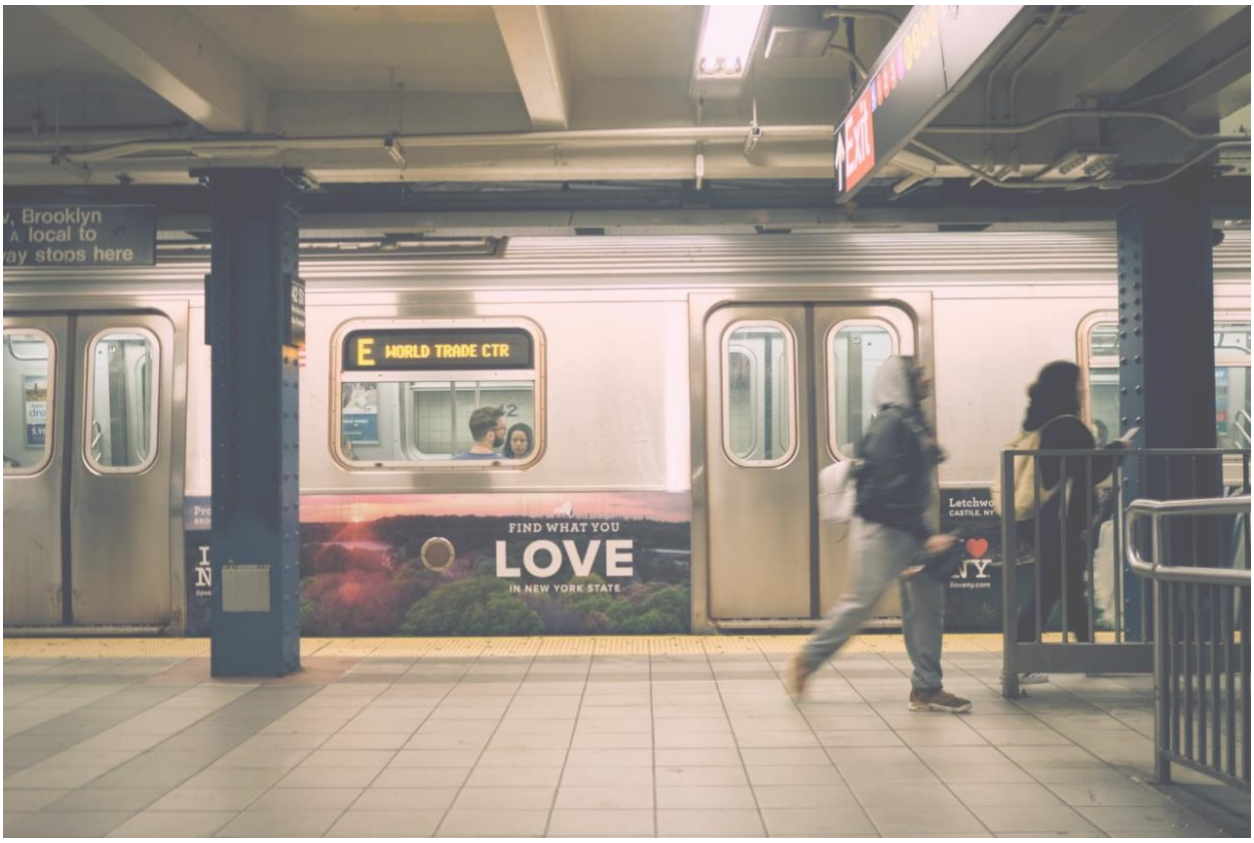
أكاديمية سدايا
SDAIA Academy

MTA Exploratory Data Analysis Project Proposal



Raghad Aloraini
2 September 2021

Introduction

In this project, I will be doing Exploratory data analysis for the Metropolitan Transportation Authority datasets in New York. I plan to find the effect of Coronavirus Pandemic on the Metropolitan Transportation. Then find which Turnstiles are mostly busy and when. At the end of the project, I plan to have some suggestions that may improve the Metropolitan Transportation experience. I will analyze and explore data of 6 months.

Dataset

The datasets I will be using are taken from the Metropolitan Transportation Authority website (http://web.mta.info/developers/turnstile.html). It has many data sets that shows the entry/exit register values for the turnstiles at control area at specific date and time for every week starting from May 2010 to August 2021. Below is the field description:

- C/A = Control Area
- UNIT = Remote Unit for a station
- SCP = Subunit Channel Position represents an specific address for a device
- DATEn = Represents the date (MM-DD-YY)
- TIMEn = Represents the time (hh:mm:ss) for a scheduled audit event
- DEScn = Represent the "REGULAR" scheduled audit event (occurs every 4 hours)
- ENTRIESn = The comulative entry register value for a device
- EXISTn = The cumulative exit register value for a device

Algorithms

For data cleaning and pre-processing, I will start by deleting the duplicate records and check if there are any null values then drop them. Also, I will add new column that will have the date and time and change it is type to datetime to allow me to do time-based operations. Then, I will add new column that will have the exact number of entries at each turnstile at control area by calculating the difference from the previous device. I will visualize the results and try to find any correlations such as a relationship between the date and time and the number of entries.

Tools

To explore ad analyze the data, I will be using different tools such as Jupyter notebook, Excel, SQLite. Also, I will use different libraries with python such as pandas, numpy, matplotlib, stats and math.

Conclusion

To conclude, I expect after analyzing and exploring the data to see the effect of the COVID-19 and if the vaccine has changed the transportation traffic. Also, figure out which Turnstiles and gates are busier. I plan to have some suggestions that may improve the Metropolitan Transportation experience.