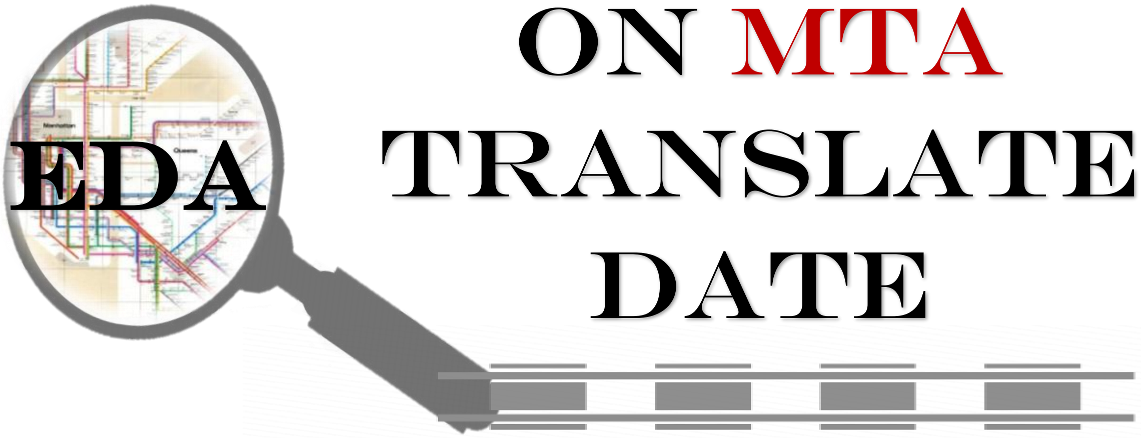
SDAIA T5_bootcamp
Raghad Rashed Al-Saqer
Oct.02.2021



EDA ON MTA TRANSLATE DATE

After the end of the 1st phase of the internship at SDAIA Academy camp and in order to fulfil the first requirement as a data science and AI intern, I had to work on my 1st project of this kind which is Exploratory Data Analysis (EDA) for the publicly available MTA dataset. I aim by that to provide insight into the fictional entity (RSTABUS) and meet its needs.

Background

RSTABUS is an imaginary bus company covering several regions in the United States of America. It is concerned with carrying passengers from several specific centres to the destination stations throughout the day. RSTABUS wants to cover more areas of high demand and there are no buses provided for the customers to connect them to the destination station, so that would attract more customers to benefit from its services and also to expand the field of work and cover a larger space for the geographical area at the level of the United States of America.

It needs to know the (locations) of the stations with the specifications mentioned, in order to study the case and find out the appropriate number of needed buses, establish bus centres for each station, and identify drivers and so on. In addition, RSTABUS needs the daily schedule of rides times for that station to connect it with those buses after providing them. My task is to help RSTABUS by using the MTA dataset to provide it with the information needed; so that it could then draw up a business plan for an optimal service.

Question/ Need

1. What are the stations that I am looking to target in my analysis? They are the busiest stations of demographic or location and do not have bus service. I used this aim to work on most of my analysis.

2. What are the set of data parts that will be targeted to be analysed? I assumed that the date of the event would be on the period form (Jul to Oct 2021), to focus on collecting only the last data from the MTA and to get the latest results.
3. After knowing the locations of the highest stations traffic and not having bus service, RSTABUS will be able to start drawing up a business plan to provide the service to customers.
4. Knowing the traffic throughout the day will be sufficient to determine the number of buses that will be provided and to determine their operating times.
5. Bus centres will be built based on the most traffic stations only.

Data Description

- I used MTA Turnstile Data, deciding to focus on data (from 3 Jul to 25 Sep 2021) to recommend a list of the top 5 stations with the highest amount of ridership and which does not have a transportation service.
- I will also take care of the same MTA dataset after calculating the traffic by extracting the departure times of the buses for the service.
- RSTABUS will get a model of ten high-demand stations, with their geographical location with the riding times of each station.

Tools

- SQLite
- Python (sqlalchemy, Numpy and Pandas, etc…)
- Google Map to ensure that the stations serve a good geographic area and a bus centre may serve a remote line station.