Exploratory Data Analysis on MTA Data

Prepared by: Muneera Alshunaifi

Abstract:

As coffee shop owners who have limited budget, we want first to measure the decision of opening a coffee shop at one of the MTA stations whether is good and profitable idea or not, in order to do that, we need to post some ads at the stations but we do not want to waist and utilize huge amount of money on ads posters. However, we want to use few amounts of ads in which it can be affective and attractive to the targeted customers.

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.

- 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
- LINENAME = Represents all train lines that can be boarded at this station
- Normally lines are represented by one character. LINENAME 456NQR represents train server for 4, 5, 6, N, Q, and R trains. which is a string
- DIVISION = Represents the Line originally the station belonged to BMT, IRT, or IND 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
- DESc = Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours) Audits may occur more than 4 hours due to planning, or troubleshooting activities.

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 and month name based on the date of the record.
- Create new column of the station places name.
- Extract the daily 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 . Numpy, and Panda for data read and write operations.

Communication:

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