MTA Exploratory Data Analysis Project Proposal

By

Sohaib Albakri

Data Science Bootcamp

SDAIA Academy

Riyadh – Saudi Arabia

[September 2021]

Problem

Most of the people use the train as their main transportation but when they arrive at the destination, they need to look for a taxi or use transportation apps, but sometimes the station is crowded so there are not enough drivers which makes them walk the rest of the way or spend more waiting for a driver and because of the traffic the people needs to wait in a long line to buy something to eat or drink which is annoying. On the other side, the shops in the less crowded stations are losing which makes theme close their business.

Solution

In this project, I will be doing Exploratory data analysis for the Metropolitan Transportation Authority datasets in New York. I plan to study the crowded stations and when the traffic occurs, and we can use this information to add more food trucks and booths to these stations and help the people to find without waiting for long time. Also, we can provide transportation facilities like bikes and more taxis in traffic times. To help the shops in less crowded stations, we can plan for events and carnivals to increase the number of visitors in these stations which will increase the sales.

Dataset

The dataset we will explore are the entry/exit statistics for the turnstiles at the control area at a specific date and time starting from May 2010 to August 2021 the data are taken from the Metropolitan Transportation Authority website. The following table explains the dataset in detail:

Column	Description	Data type
C/A	Control Area	String
UNIT	Remote Unit for a station	String
SCP	Subunit Channel Position represents a specific address for a device	String
STATION	Represents the station name the device is located at	String
LINE NAME	Represents all train lines that can be boarded at this station	String

DIVISION	Represents the Line originally the station belonged to	String
DATE	Represents the date in (MM-DD-YY) format	String
TIME	Represents the time (hh:mm:ss) for a scheduled audit event	String
DESC	Represent the "REGULAR" scheduled audit event	String
ENTRIES	The cumulative entry register value for a device	Integer
EXITS	The cumulative exit register value for a device	Integer

Features

First, I will remove the duplicated data to get clear results, also I will add delivered columns from the current columns which are as follows:

Column	Description	Data type
NUM_ENTRIES	Subtracting the cumulative entries from the previous row for the same turnstile	Integer
NUM_EXITS	Subtracting the cumulative exits from the previous row for the same turnstile	Integer
TRAFFIC	This column is the result of adding the NUM_ENTRIES and NUM_EXITS	Integer

By grouping the data by the station name I will use these new columns to sort the data to find the most crowded stations and when does the crowd happen.

Tools

To explore and analyze the data, I will use the following tools:

Tool	Description
Jupyter notebook	Contains cells of Python code and human-readable text
pandas	Library is written in Python for data manipulation and analysis
NumPy	Library for the Python, adding support for multi-dimensional arrays and matrices, along with a large collection of mathematical functions
matplotlib	Matplotlib is a plotting library for Python

Conclusion

I expect after analyzing and exploring the data by finding the most and less crowded stations and the traffic time in each station we can put a plan to make the people transportation easy and reduce their waiting time. Also, we can help the business owners to increase their sales and income by making the less crowded stations attract people to visit.