**Proposal**

**Introduction:**

A client reach out to me to do exploratory data analysis because she want to open a booth that sells breakfast, snacks and coffee to go, she wants to set her booth to have a strategic place since COVID 19 pandemic closed her shop and now she wants to know where is the stations that have the most foot traffic after COVID where people mostly will need to take a coffee and breakfast before getting on to the train or after getting off it in the morning time specifically, so I’m going to use the Metropolitan Transportation Authority Turnstile dataset, I will be doing my exploratory data analysis on the period before COVID 19, from November 2, 2019 to December 28 2019 then compare it to the period from November 7,2020 to December 26,2020 ,Hopefully in the end of my EDA I will determine the best places for my client to choose from to open her booth .

**Datasets:**

In my EDA I will be using the MTA turnstile dataset free to access on <http://web.mta.info/developers/turnstile.html> , each database is taken every Saturday of every month and the information inside the database will show you each turnstile in witch control area with the entry and exit date and time below is the field description:

C/A: Control Area (A002)

UNIT : Remote Unit for a station (R051)

SCP: Subunit Channel Position represents a specific address for a device STATION: Represents the station name the device is located at

LINENAME: Represents all train lines that can be boarded at this station

DIVISION: Represents the Line originally the station belonged to BMT, IRT, IND

DATE: Represents the date (MM-DD-YY)

TIME: Represents the time (hh:mm:ss) for a scheduled audit event

DESC : Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours)

ENTRIES: The cumulative entry register value for a device

EXITS: The cumulative exit register value for a device.

**Algorithms:**

First I will be preforming data cleaning by searching for the NULL and duplicates values and removing them, then I will create a new column date-time and concatenate the date and time, Also I will change the data type into datetime so I can perform time operations, also I will calculate the total entries and exits for each day and each station before COVID and compare it to the total entries and exits for each day and each station after COVID happened, then I will combine the numbers for each C/A-UNIT-STATION, then I will calculate the sum of counts for turnstile from the combination of C/A-UNIT-STATION , Also I will be doing some SQL queries for the data .

**Tools:**

I will be using different python libraries with jubyter notebook such as pandas, matblotlib and seaborn,numpy and stat also, SQLite and SQLAlchemy.

**Conclusion:**

In conclusion in the end of my exploratory data analysis I will know what are the busiest stations and that will help my client to choose to open her booth in.