|| Background:

We are tourism Authority in New York City .One of the tour guides has experience a noticeable traffic at some Subway stations particularly (E 149 ST,59 ST COLUMBU ,VERNON-JACKSON ),which leads to unsatisfied impression from the tour's members .The guide seeks to rescue the authority from the money loss and bad reputation .So he has investigated his passion about Data Science to warn MTA (Metropolitan Transit Authority) about most traffic stations , in order to help them facilitate the transportation via subway stations . At the end there will be valuable, actionable insight into their transportation performance – Collect data that leads to better strategic decisions in the MTA organization .

# II. Setup

# Dataset Description:

# We used historical data from [MTA’s turnstiles](http://web.mta.info/developers/turnstile.html), deciding to focus on data from January and April 2017.

* The months of **January to April**seem to be enough lead time to monitor the traffic in NY.

Let's start the journey

*Importing Required Libraries*

# import libraries

import pandas as pd

import datetime

import numpy as np

%matplotlib inline

import matplotlib.pyplot as plt

## Loading the dataset into Python -

def get\_data(week\_nums):

url = "http://web.mta.info/developers/data/nyct/turnstile/turnstile\_{}.txt"

dfs= []

for week\_num in week\_nums:

file\_url = url.format(week\_num)

dfs.append(pd.read\_csv(file\_url))

return pd.concat(dfs)

week\_nums = [170107, 170114, 170128, 170204, 170211, 170218, 17022

170304, 170304, 170311, 170311, 170318, 170325, 170401,170408, 170415, 170422, 170429]

turnstiles\_df = get\_data(week\_nums)

## Show first five rows of the data

turnstiles\_df.head(data)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

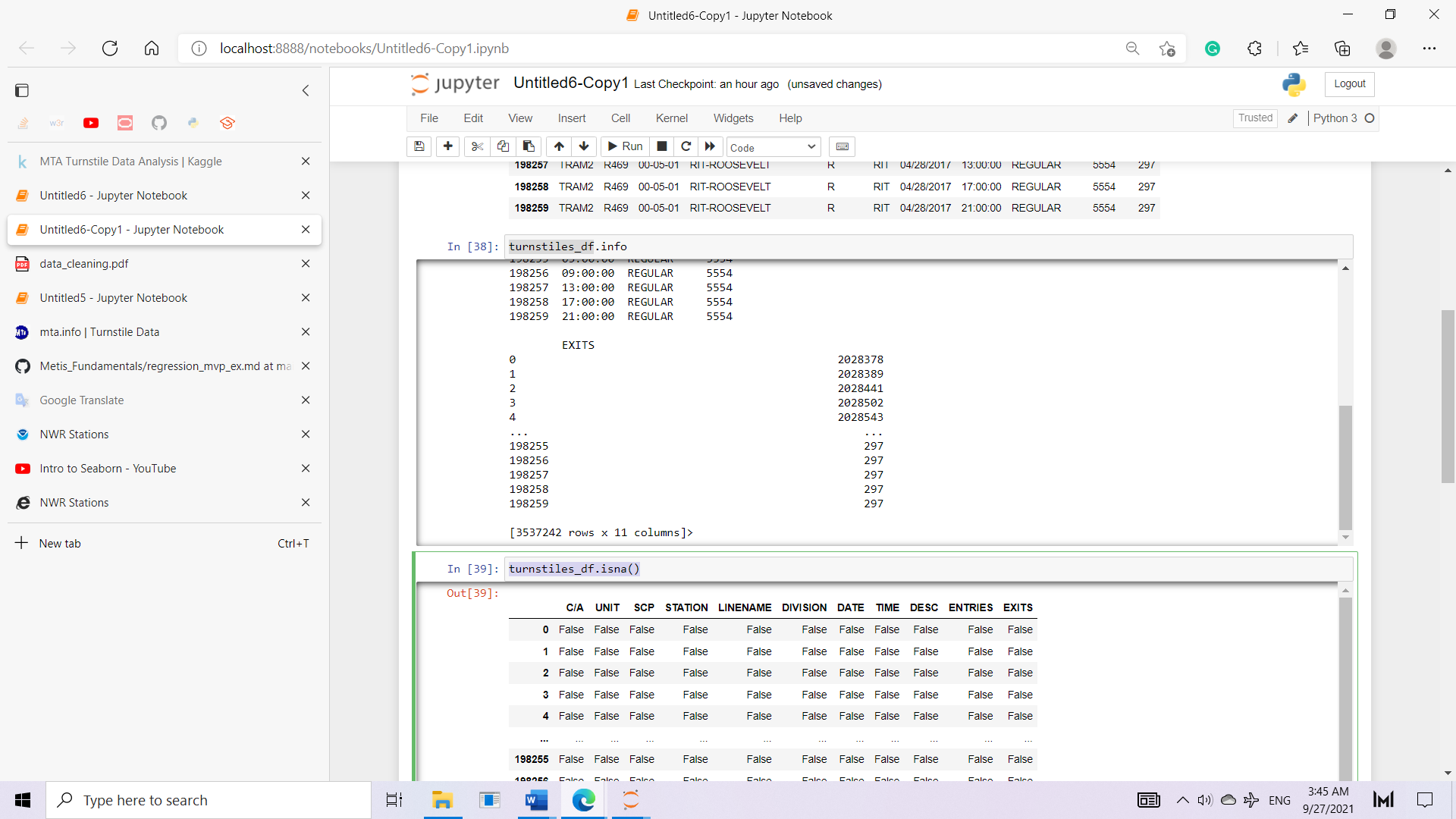
III. Preparing the data

# Exploratory Data Analysis:

missing values: I have checked if the data contains null values or not by

using isnull()method , and it returned false which indicates to the data frame doesn't contain missing values , and for making sure I tried ~.isna() function to check each specific column and it returned false value for null values.

turnstiles\_df.isna()



* -we want to make sure the error in spelling avoided by removing spaces via strip method

As :

turnstiles\_df.columns = [column.strip() for column in turnstiles\_df.columns]

* turnstiles\_df.columns

- focusing on the repeated values:

(I found repeated values in entries and exists according to various dates, for march month, I will display only unique values via unique()method

* inconsistent values: there is no inconsistent values. also, to simplify the analysis process, I checked the column names and I found it is understood.
* In addition, to ensure that we are able to make any calculation or statistics on data, I have checked the data types for the desired column, and I found the data is numeric -int64, which means it is possible to make calculation.

……………………………………

Mask:

portions of data mask

mask = ((turnstiles\_df["C/A"] == "A002") &

(turnstiles\_df["UNIT"] == "R051") &

(turnstiles\_df["SCP"] == "02-00-00") &

(turnstiles\_df["STATION"] == "59 ST"))

turnstiles\_df[mask].head()

# and when I find exact daily entry instead of cumulative, we need to specify the date to show the daily.

# shift () enables us to shift line and then find the next entry, while by difference of current and next entries ,the daily entry obtained !

# '' space to continue the remaining work'

# IV. Insights from the data

We found the top 5 stations by aggregate traffic were as follows:

1. E 149 ST
2. 59 ST COLUMBUS
3. VERNON-JACKSON