MTA Turnstile Exploratory Data Analysis

Abstract

The purpose of this project is to make an exploratory data analysis for the New York subway MTA turnstile data to find the busiest stations and also to get the most traffic station and the busiest day and hour related to a specific station. My analysis was done on the data given by the website which is (http://web.mta.info/developers/turnstile.html). I did some of data analysis to achieve this purpose and plot some graphs to make a better data visualization.

Design

First of all I scrapped the data from the website of MTA data (http://web.mta.info/developers/turnstile.html). Starting from scrapping three months of data from 30 June to 28 of August and putting them in sql file in a table called test then applying some data cleaning , adding some columns help us in visualization, finally plot the graphs and get more intuition about the data and total traffic.

Data

The New York subway MTA turnstile data is a series of data files containing cumulative number of entries and exits by station, turnstile, date and time. Data files are produced weekly, data records are collected typically every 4 hours with some exceptions. the data which I will use now to analyze is from June to August for 2020 year.

Structure of the data

>> C/A = Control Area (e.g., A002)

>> unit = Remote Unit for a station (e.g., R051)

>> SCP = Subunit Channel Position represents an specific address for a device (e.g., 02-00-00)

>> station code = C/A + unit, locating a station

>> turnstile = C/A + unit + SCP, locating a turnstile

>> Station = Represents the station name the device is located at

>> 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

Analysis Steps

First of all , I scrapped the data and put them in a sql file consisting of one table containing the whole information of the Subway MTA data the loading the data in a python data frame . after that I performed some data cleaning, I added some additional columns which helps me get more visualization and understanding the data clearly. These columns are the difference between entries and the difference between exits with the previous record the a added a column called total traffic which is the sum of the two columns Entries difference and exits difference to work on it as it is the most important column which helps me to get the most traffic station and the most traffic day and hour related to a specific station. Also I dropped some unnecessary columns such as line name desc,… etc. as explained in the python file .I used some

powerful libraries in python which help me to do the analysis of the data such as pandas ,matplotlib ,seaborn …etc. finally I plotted some graphs explaining the most busiest station and the most busiest hour and day related to a specific station and get the results as shown I the python file.

Conclusion and summary

1.from the above Exploratory Data analysis we can conclude that the most traffic station is the "world trade CTR" station as we see this from the above plots. we also can deduce that the top 10 stations which have most traffics are ['WORLD TRADE CTR', '191 ST', 'BAY RIDGE AV', 'HUNTS POINT AV', '4AV-9 ST', '96 ST-2 AVE', 'TREMONT AV', 'CROWN HTS-UTICA', '207 ST', 'BOWERY']. Also we can see that Monday is the most traffic day for the "world trade CTR " and at time 12 AM is the busiest hour traffic in the day.
2.the most busiest day for "191 ST " Station is Thursday
3.4 AM is the most traffic hour at Wednesday At "BAY BRIDGE AV" station.
4.the most busiest day for "HUNTS POINT AV " Station is Monday
5.the most busiest day for "4AV-9 ST " Station is Thursday
6.the most busiest day for "96 ST-2 AVE " Station is Monday
7.6.the most busiest day for "TREMONT AV " Station is Tuesday