Abstract :

In this project, my tasked on MTA turnstile data to help new company to advertise their brand, business and other purposes. Newuture is new tourism company. It will create experiential travel opportunities for tourists coming to US. Newuture want to know the average number of the passengers in each station and find out the importance of this station to start advertising (promotional distribution, survey, …etc). also, they try to catch the peak time so they will send their employees for advertising and Their trips will be planed based on the number of trains with few numbers of passengers.

Design:
1.Get data: MTA subway turnstile data
2.Process-analyze data: total traffic, traffic density(traffic per station), commuter index(peak rush hours - nonpeak hours):We calculated traffic density of people (population per turnstiles), and created original indicator, Then we used the product of these parameters to choose Top 10 MTA station.
3.Visualize analysis
4.Presentation

Data:
 I used historical data from MTA's turnstiles, deciding to focus on data from Jul to Sep 2021. Th e dataset contains 2724418 rows × 12 columns ('C/A', 'UNIT', 'SCP', 'STATION', 'LINENAME', 'DIVISION', 'DATE', 'TIME', 'DESC', 'ENTRIES','EXITS)
Algorithms
   1- Read the data from the website
   2- Checking the Data Frame
   3- Clean the data
        a. Get rid of the duplicate entry
        b. Get rid of the null value
        c. Remove outlier value
   4- Data analysis:
        a. Creating new columns:
             i. DATETIME - TIME_INTERVAL- PREV_DATETIME- PREV_EXITS
        b. Find the value of Dailytraffic = ENTRIES + EXITS
        c. Create 2 table one for entreis and the other for exists
        d. Merge the table
        e. Grouping the same data above by stations to find of a sum of daily entires and exists for each station
   5- Visualization
Tools:
   1- Python(sqlalchemy Numpy and Pandas) for data manipulation
   2- Matplotlib and Seaborn for plotting
   3- Google Map for choosing station have valuable places near to customers (museum, peach, Disneyland, …etc).

Communication:

Top Most Trafficed Stations

Station	Total Traffic
34 ST-PENN STA	
34 ST-HERALD SQ	
GRD CNTRL-42 ST	
86 ST	
14 ST-UNION SQ	
42 ST-PORT AUTH	
23 ST	
125 ST	
TIMES SQ-42 ST	
FULTON ST	

(X-axis: Total Traffic, 0.0 to 1.0 ×10⁷)

Less Trafficed Stations

Station	Total Traffic
ORCHARD BEACH	
BROAD CHANNEL	
TOMPKINSVILLE	
BEACH 105 ST	
BEACH 98 ST	
BEACH 44 ST	
AQUEDUCT RACETR	
NEPTUNE AV	
BEACH 90 ST	
ROCKAWAY PARK B	

(X-axis: Total Traffic, 0 to 80000)

Station Traffic in the Week

Top 10 Stations	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
34 ST-PENN STA	1391.8k	1623.5k	1545.3k	1512.6k	1533.7k	1097.0k	943.0k
34 ST-HERALD SQ	1036.7k	1165.7k	1174.8k	1151.1k	1183.8k	864.1k	719.1k
GRD CNTRL-42 ST	1045.1k	1227.7k	1251.7k	1217.8k	1139.3k	695.4k	557.1k
86 ST	926.7k	1002.9k	1011.4k	1021.0k	1026.6k	733.0k	583.9k
14 ST-UNION SQ	861.8k	958.7k	1003.3k	946.7k	1009.7k	845.8k	666.7k
42 ST-PORT AUTH	891.1k	977.6k	992.6k	986.3k	1039.4k	744.0k	614.4k
23 ST	817.7k	964.0k	987.8k	952.8k	948.7k	625.2k	485.6k
125 ST	813.5k	889.0k	896.6k	871.7k	919.1k	681.1k	561.8k
TIMES SQ-42 ST	783.0k	859.2k	866.9k	848.0k	825.1k	746.3k	619.2k
FULTON ST	739.9k	845.5k	866.4k	845.6k	829.4k	540.3k	425.5k

(Colorbar scale: 0.6 to 1.6 ×10⁶)

Day of the Week

Station Traffic in the day

Top 10 Stations	0:00 to 8:00	9:00 to 16:00	17:00 to 24:00
34 ST-PENN STA	1914.9k	4225.3k	3506.6k
34 ST-HERALD SQ	1563.7k	3347.0k	2384.7k
GRD CNTRL-42 ST	1163.6k	3238.4k	2732.1k
86 ST	1235.2k	2872.4k	2198.1k
14 ST-UNION SQ	667.4k	2107.2k	3518.1k
42 ST-PORT AUTH	1478.8k	2448.6k	2318.0k
23 ST	940.1k	2545.9k	2295.8k
125 ST	1061.1k	2365.8k	2205.9k
TIMES SQ-42 ST	1494.0k	2393.8k	1659.9k
FULTON ST	516.6k	2087.9k	2488.1k

TIME_INTERVAL