EDA-for-MTA-project
Blood Donate

By: nada alqabbani

10/9/2021

❖ Project Background

▪ Idea of Project

Some of common problems in the world is the need for blood donors to help patients. So, I want to take advantage of the places where people gather to solve the problem by introducing the idea.

▪ Solution

Distribute special donation trucks in the most crowded metro stations to reach the largest number of donors and thus help many patients get better.

▪ Assumption

- I assumption when I choice the most station have people and the best time in day, I will collect a large number of donors.
- I assumption in the time of back people from work and the student from university and in the night is most suitable time.

I will distribute many trucks for blood donations in many stations, but I focus to the most station have people and the best time when I implement assumptions I will reach large people and I will help many people in need of blood and hospitals. I will bring to the Ministry of Health a solution to the common shortage problem. So, I took advantage of the crowding with great benefit, spreading and reminding people of the importance of donating.

❖ Data Source

▪ August, September / 2021
▪ December / 2020

- Daily Traffic for each Station
- Daily Traffic for Time Interval
- Daily Traffic for each day

❖ Analysis

I need to analysis the most station crowded, and the time interval have crowd to distribute the trucks blood donors.

- Steps analysis:
    - Load the dataset from MTA Turnstile Data and convert it to a data frame.
    - Loaded Dataset to SQLite and read Database from SQL.
    - Find the distances between column name and removed by **Strip.**
    - Find a nulls values and dropped.
    - Create new column to combine Time and Date to easily analysis.
    - Find unique rows by drop duplicates.
    - Find shape function to know the rows numbers.
    - Create new column to select the time interval.
    - Create three new columns include the date, exits, entries columns values with shift one.
    - Drop a Nan values.
    - Check if have negative values or outlier values.
    - Remove the outlier and negative values and stored the clean column into new column name daily
    - Create new column to combine daily entries and exits name traffic.
    - Use the traffic column to analysis the top station crowded and what time interval by group by function.
    - Then represent the answer into bar and plot charts.

▪ After clean and exploring data to easily analysis and view data I can answer for questions:

    ○ Q1: What the top stations is Crowded?
- 61   34 ST-PENN STA
- 233   GRD CNTRL-42 ST
- 59   34 ST-HERALD SQ
- 110   86 ST
- 14   14 ST-UNION SQ

Top 5 busiest stations
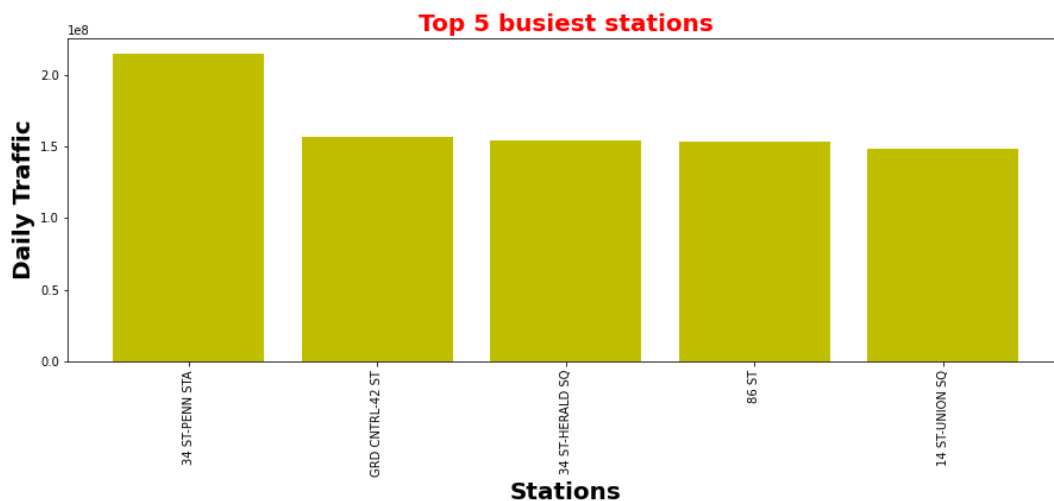
    ○ Q2: What the time interval od day crowded in 34 ST-PENN STA station?
The time crowded is from 0:00 to 20:00 often because I analysis to days:
- Sunday is time crowded is from 0:00 to 20:00 and max number of people is 80000.
- Friday is time crowded is from 0:00 to 20:00 and max number of people is 140000.

The Daily Traffic in Friday & Sunday of (34 ST-PENN STA) Station
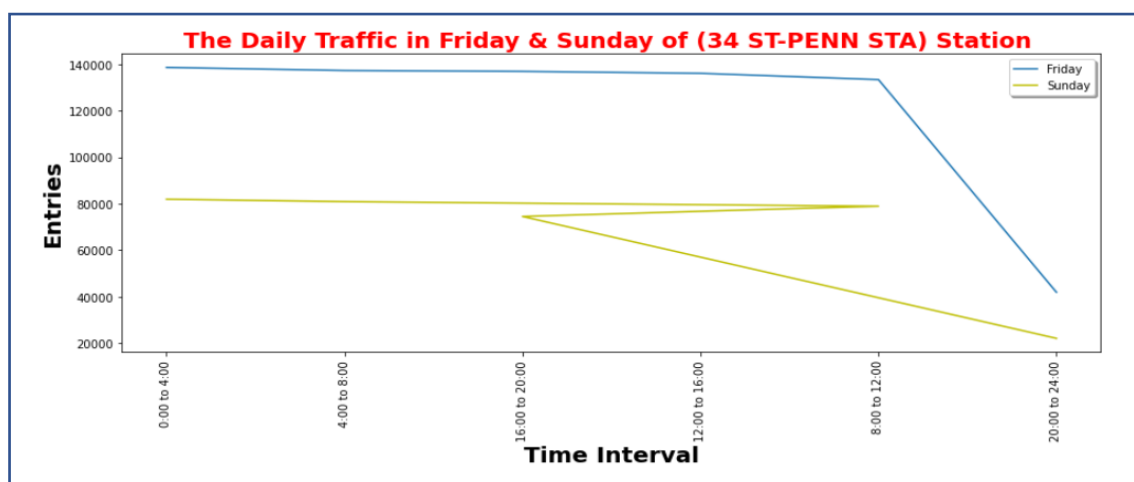
o   Q3: What the Ridership per day for 34 ST-PENN STA station?
- The max day is Saturday and Sunday
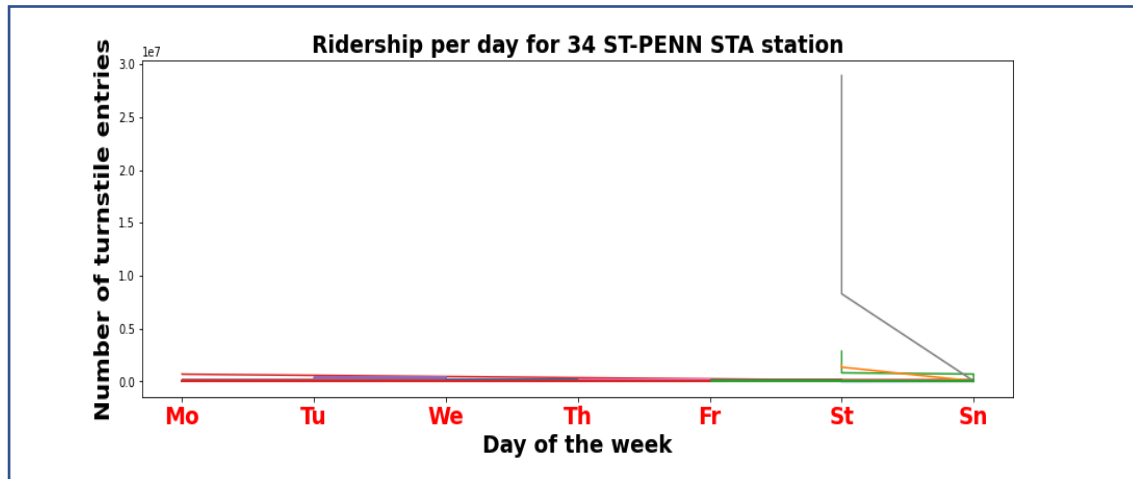
❖ Conclusion:

- I will distribute the trucks of blood donors from 12:00 to 20:00 because is always have daily traffic and is suitable time for people back from work.

- I hope that as many donors as possible will come to help many patients.