MTA TURNSTILE TRAFFIC ANALYSIS

NORAH TAMSAN

ABSTRACT

In this project, the purpose is to find out which stations are the busiest peak times as well as the busiest days to help Variety the Children's Charity of New York run a fundraiser, the Variety the Children's Charity needs to set up 5 booths across New York City. I worked with data provided by The Metropolitan Transportation Authority (MTA), to achieve promising results for this project After cleaning and visualize the data to git the results.

DESIGN

Using the data provided by MTA we must find which stations are the busiest peak times as well as the busiest days to help Variety the Children's Charity of New York run a fundraiser. Assuming that the Variety the Children's Charity is constrained by manpower resources, hence insights from my analysis should identify top stations by traffic, as well as the peak periods in those stations and needs to set up 5 booths across New York City in ether station entry or exit.

DATA

Data from The Metropolitan Transportation Authority (MTA), it shows the entry/exit register values for one turnstile at control area from certain date at certain hours for each subway station in New York City. The data is updated each week and the sample size of this data is 3 months starting from 25 Sep to 9 oct. The entry and exits where used to find the traffic in each station.

ALGORITHM

We started by:

- Derive new columns from the given data PREV_DATE, PREV_ENTRIES, PREV_EXITS.
- Create unique station name for each station
- Finding the daily entries and exits
- Calculate the traffic
- Map the date to a weekday.
- Create time interval

then we start cleaning the data by dropping duplicated rows then shift and create new columns and dropped the NaN values. After that, we check to Find the difference between entries and previous entries to get the actual entries, get rid if any actual entry values greater than 1000000 as they are outliers and reverse the sign/takes abs value of actual entries for turnstiles counting backwards. finally, we start aggregate, and visualize the data to git the results.

TOOLS

- SQL lite: Basic relational database management system (RDBMS) software.
- SQLAlchemy : Python SQL toolkit and Object Relational .
- Pandas library: easy to use open source data analysis and manipulation tool
- Matplotlib : library that allows visualization within Python