# Exploratory Data Analysis For MTA Turnstile Data

Prepared by: Asma Althakafi

## Abstract

The goal of this project was to analyze the Metropolitan Transportation Authority (MTA) Turnstile Data to see which stations have a lot of people, to help the Café chain to determine the places where they will open new branches in New York City. I worked with MTA turnstile data which contains three months from 2021 which are June, July, and August. I Stored the data in the database, cleaned it, added more features, then analyzed and visualized the data. Finally, I built an interactive dashboard to visualize and communicate my results using Tableau.

## Design

This project originates from the SDAIA academy. The data is provided by MTA and presents turnstile data in Metro New York City. Finding the stations with the most people would enable the Café chain to determine where they will open new branches in New York City.

## Data

The dataset is three months in 2021 which are June, July, and August. It contains 2,753,240 turnstile data entries with 11 features for each. The main features I used are:

- STATION: Represents the station name the device is located at.
- DATE, TIME: Represents the date in (MM-DD-YY) format, and time in (hh:mm: ss) format.
- ENTRIES, EXITS: Represents the cumulative entry and exist register value for a turnstile.

## Algorithms

Feature Engineering

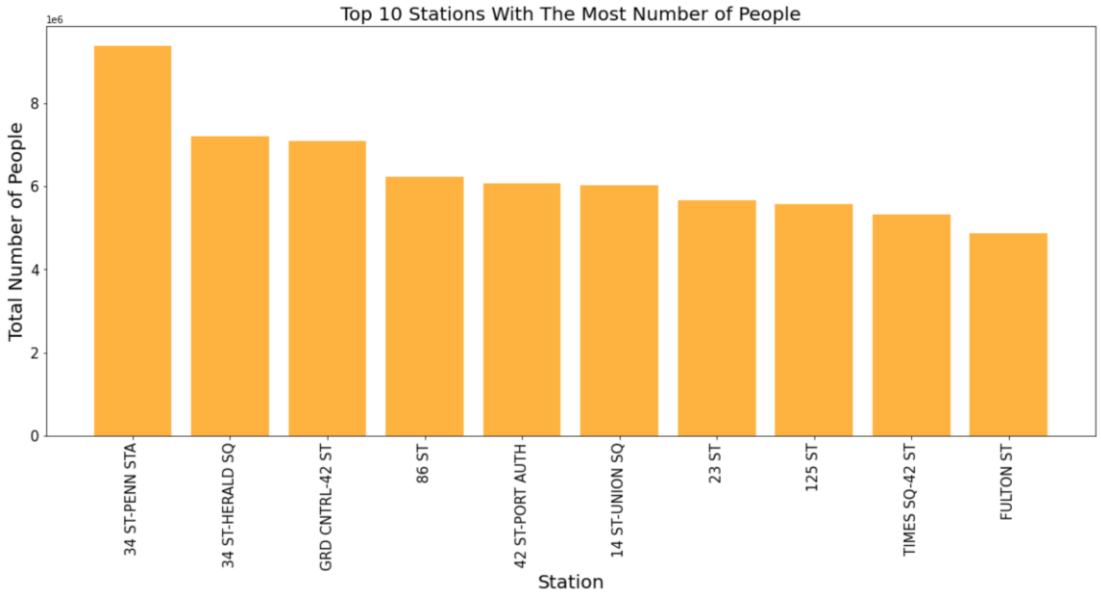
1. Adding a feature that concatenating the DATE and TIME, then convert it to 'data time' type.
2. Count the number of daily entries and exits from the cumulative entries and exits.
3. Adding a feature that sums the daily entries and exits.

## Tools

- Pandas for data manipulation.
- SQLAlchemy and SQLite to create the database.
- Numpy, Matplotlib, and Seaborn for plotting.
- Tableau for interactive visualizations.

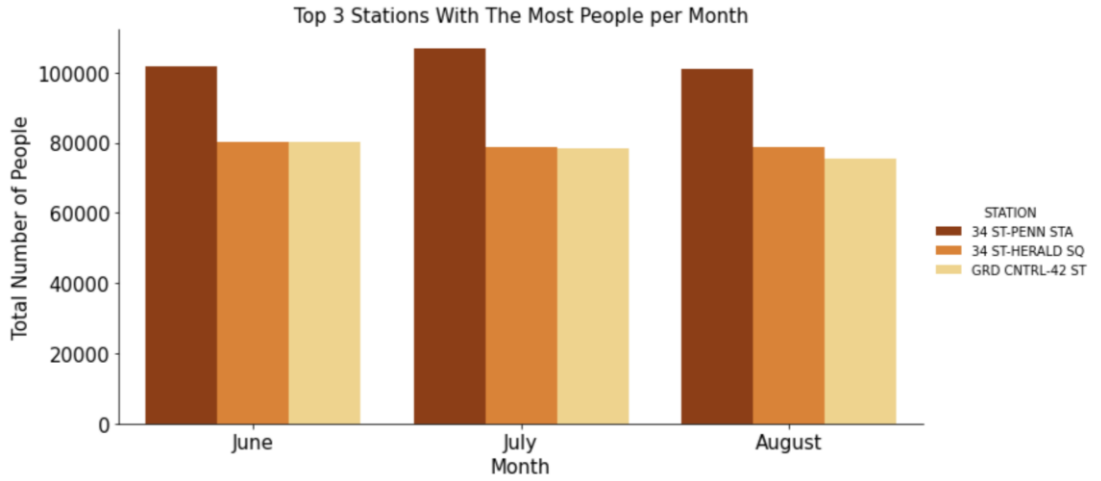
## Communication

In addition to the slides and visuals presented, here I show the plots:

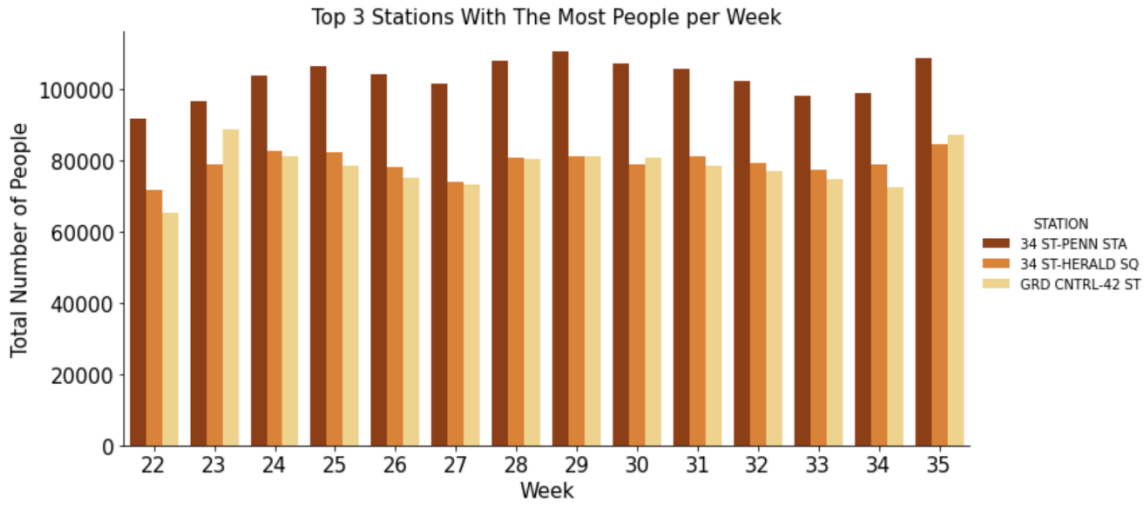



As we can see the top 3 stations with the most number of people are: 34 ST-PENN STA, 34 ST-HERALD SQ, GRD CNTRL-42 ST.

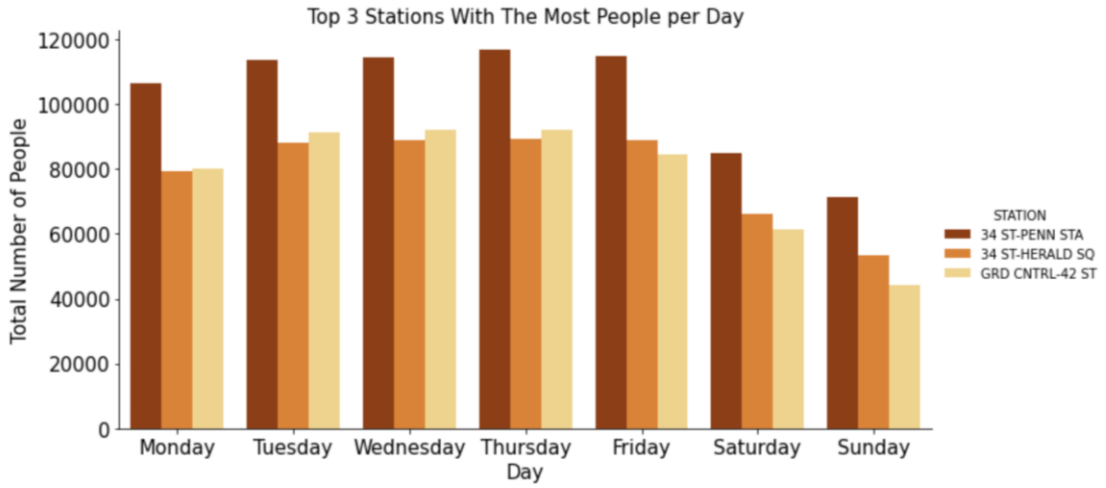
**More analysis for the top 3 stations:**



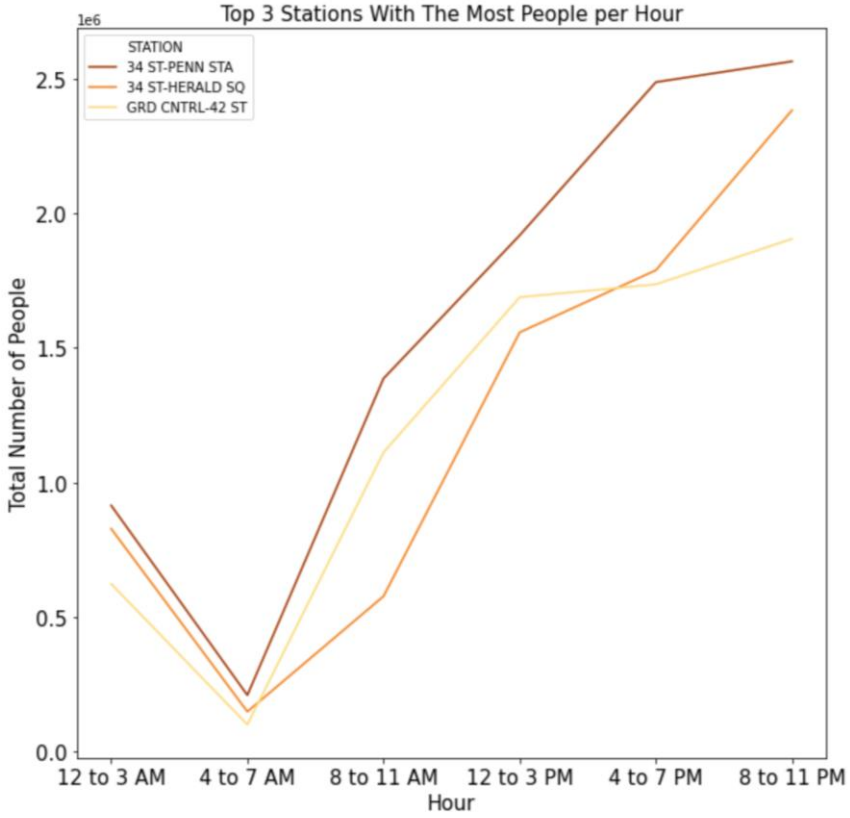

34 ST-PENN STA has the most number of people in June, July, and August.




34 ST-PENN STA always has the most number of people in each week.

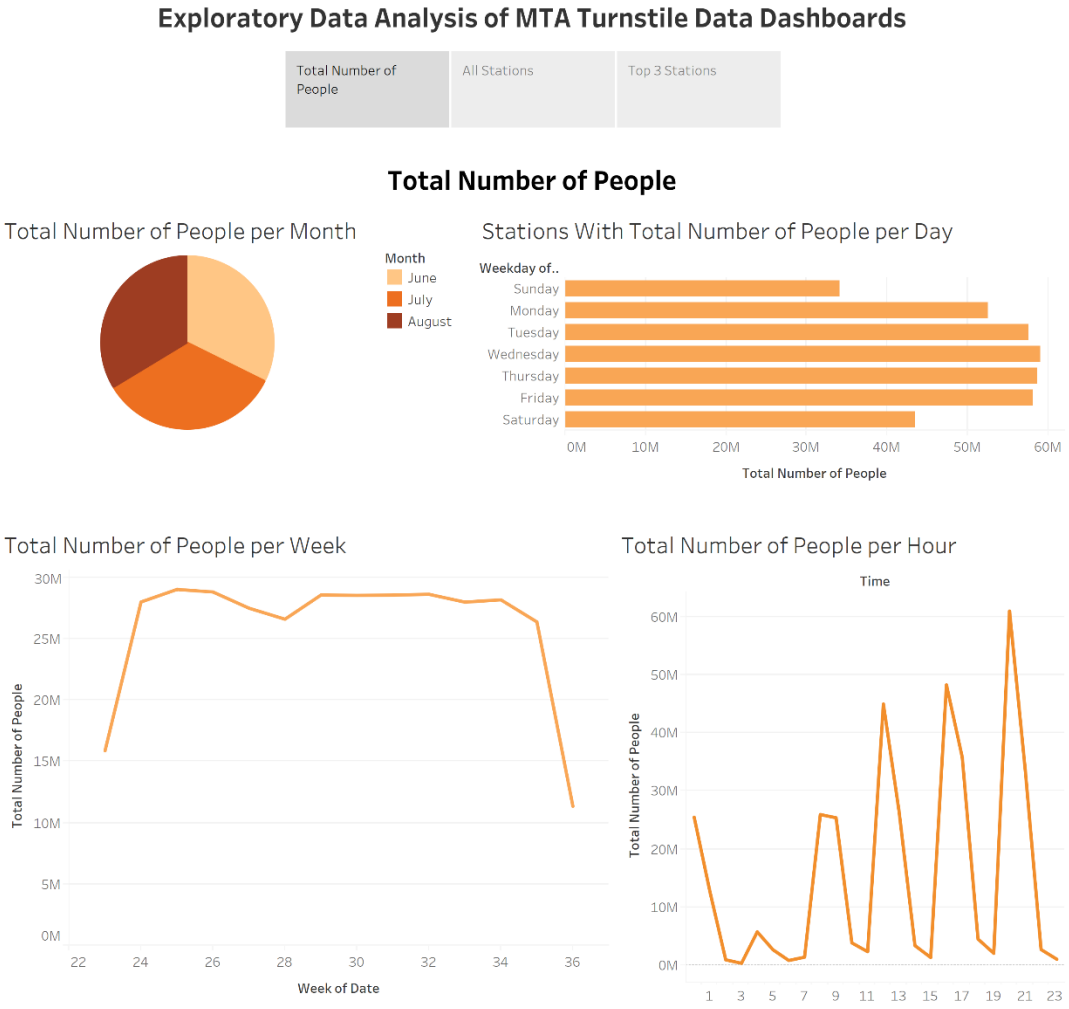



34 ST-PENN STA has the most number of people on Thursday, 34 ST-HERALD SQ on Thursday, and GRD CNTRL-42 ST on Wednesday.

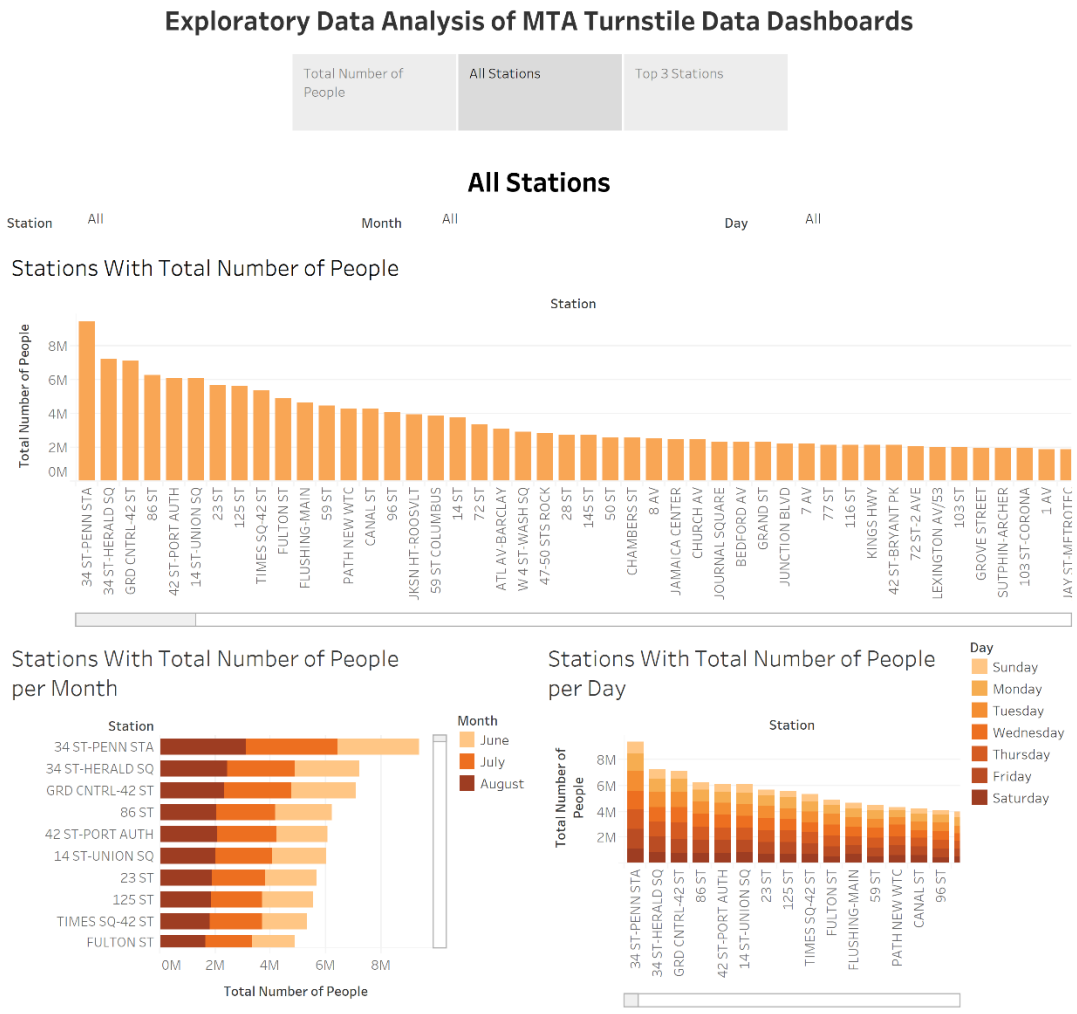



34 ST-PENN STA has the most number of people between 4 to 11 PM, 34 ST-HERALD SQ between 8 to 11 PM, and GRD CNTRL-42 ST between 8 to 11 PM.

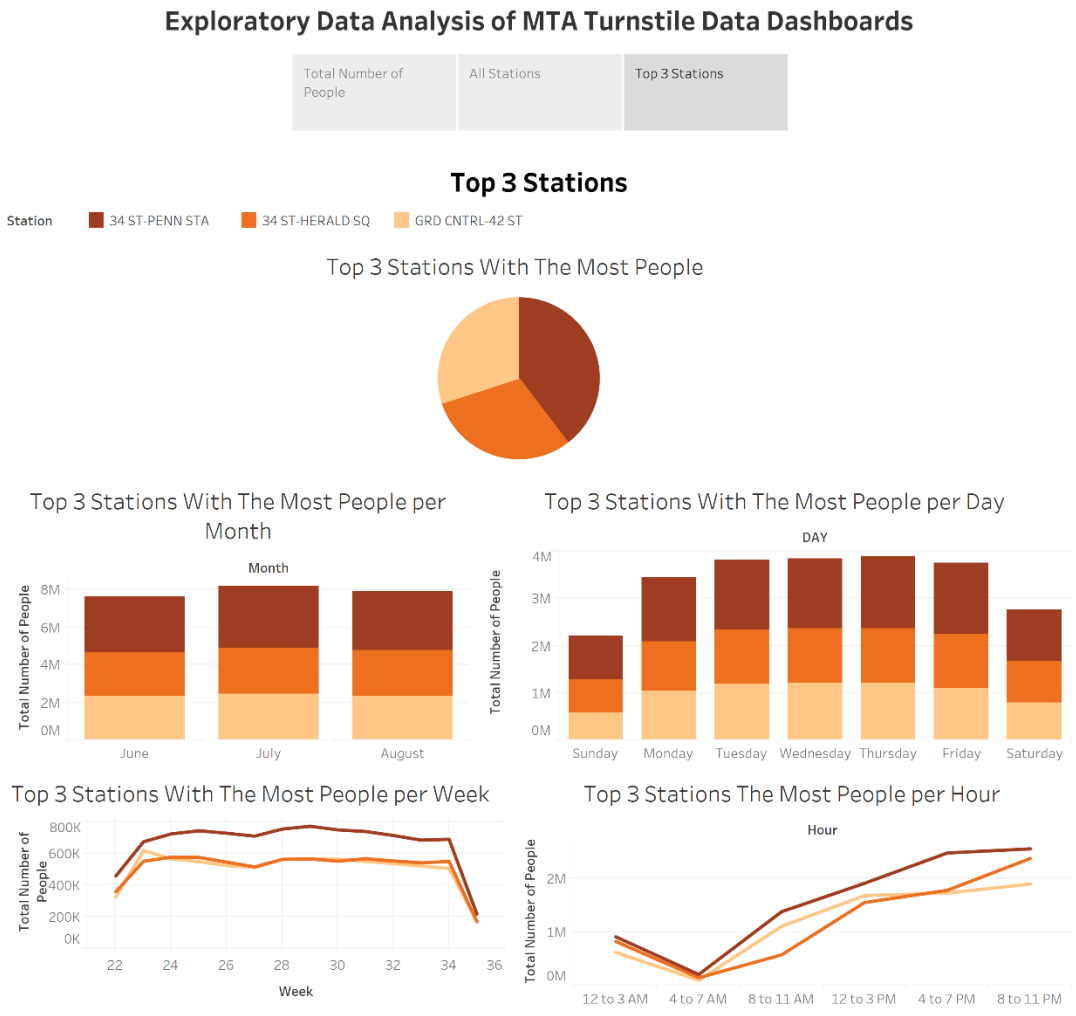
## Dashboards:




This dashboard shows the total number of people per month, week, day, and hour.




This dashboard shows the total number of people of each station per month and day.




This dashboard shows the total number of people of the top 3 stations per month, week, day, and hour.