MTA and Citi Bike project

By: Juan Herrera

Abstract

The goal of this project was to have a better understanding on the specific time at which some Citi Bike stations have a high demand of customers in order to reduce stockout. The MTA data contains many important features that along with the Citi Bike data helped me achieve this goal. I had to work on cleaning the data, creating new features and successfully filtering the data in order to achieve the results, then I used Matplotlib to visualize my results.

Design

These are datasets that many people have used before to solve different problems. In this case I was intrigued about the fact of whether there was a correlation between the high volume of people at the subway and the amount of Citi Bikes rented.

Data

I focused on the summer (June, July, August) of 2021 for this project. The subway data contains 2722610 entries where each represents a specific station, unit, time and date. The subway data contains twelve features, but some of the features that were most important to my project were Exits which specifies the number of exits recorded at that time, Station which is just the station name, Date, Time and the C/A, Unit and SCP which these last three helped me identify the distinct turnstiles in the dataset.

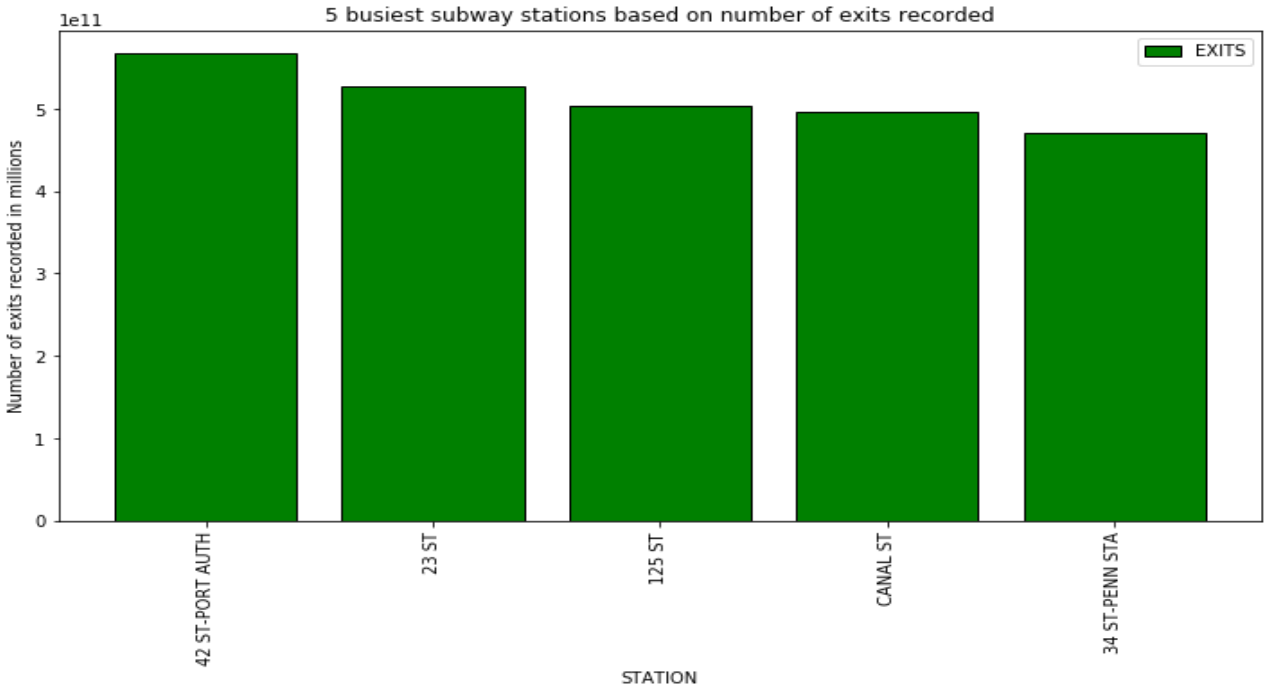
For the Citi Bike data, I worked with 9334532 entries where each represented a different ride. This dataset contained 13 features. A few features highlights include “start_lat” and “start_lng” which are the latitude and longitude of the station where the ride started and it was helpful to find the distance between the 5 busiest subway stations and each Citi Bike station.

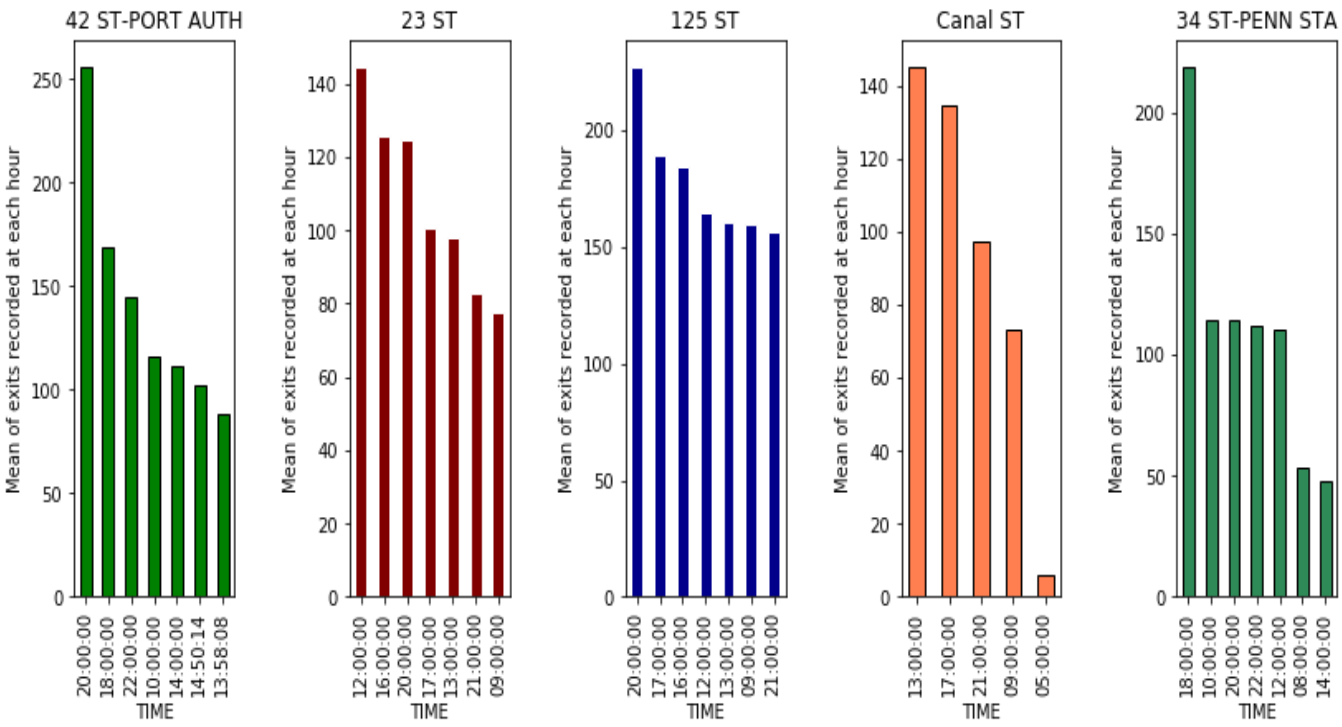
Algorithm

1. Created a new column on the subway data to find the difference in the number of exits among hours
2. Used haversine function to find the distance in km between the 5 busiest subway stations and all the Citi Bike stations included in the dataset
3. Created a new column on the Citi Bike dataset that specified the specific time in which the ride had started, if it started at 17:23, the new column would have it as 17
4. Implemented a lot of filtering and loops to find out what Citi Bike stations were in proximity to the 5 busiest subway stations

Communication

These are some of the visualizations that I used which include the 5 busiest stations and the times in which they are the busiest. However, the slides show how I used haversine and what conclusions I came up with.





Tools:

Python, SQL, Pandas, Matplotlib, Haversine