Citi Bike Project – Update 08.30.2024

Introduction

As in a city heavily troubled by congestion, lots of New Yorkers choose to ride a bike to commute or do an in-city travel. Although riding through a city is a great way to feel the city, more likely to be shocked by the awful ground transportation in this metropolitan, visitors accept riding bikes as an awesome option to travel from one location to another in New York. As a major bike sharing provider, Citi Bike is likely to be a perfect representative to depict the bike sharing environment in New York City. As a cycling lover, I would like to investigate in the topic of bike sharing in New York, though the subject itself is interesting.

Data Preparation

Before we even start to prepare the data, we need to know what is the population of interest. To understand the landscape of bike sharing in New York City, we may consider all the people in the New York City as the population. However, that can be difficult. Thankfully, we can take a look at the population who ride Citi bikes and extend the findings to the larger population with some logical reasoning. Hence, the population of interest in the major part of this project is the people who ride Citi bikes.

Thanks to Citi Bike, it is easy to access the bike sharing data on its website and thus allowing me to start such a project more easily. The raw data all come from Citi Bike's System Data. You can overview the data they provide there and download trip history data for your own investigation.

The data you can download includes:

- Ride ID
- Rideable type - classic_bike or electric_bike
- Started at - when the rider start to ride
- Ended at - when the rider stop riding
- Start station name
- Start station ID
- End station name
- End station ID
- Start latitude
- Start longitude
- End latitude
- End longitude
- Member or casual ride

Considering the huge size of the datasets and the data types provided, it is far from plausible to use all these data to carry out the analysis. Statistically, it is good to sample these data with logically correct methods and use sampled data which is of a much smaller size to do analysis. Such an analysis is expected to be appliable to the whole population. Here, we are going to use Simple Random Sampling (SRS). By using SRS, each individual can be selected with equal probability. Understanding SRS may be vulnerable to sampling error, we are not denying other sampling methods as moving on the project.

The Citi bike ride history data are divided into months. Data of each month may be divided into multiple datasets. To help simplify and automate the process of data retrieving, sampling, and

early-stage feature engineering, we are going to write some functions. The functions are intended for the datasets having the data as introduced earlier.

- The first function is to retrieve monthly data and sample it. It takes the year, month, file path, and sample fraction in and output a randomly sampled dataset.
- The second function is to repeat the first function for a whole year or a list of years.
- The third function is to do some basic feature engineering. It can make sure the data type of each column is good to use. Some features are added as well. For example, the duration of the ride is added by calculating the difference between "started at" and "ended at".

To handle missing data, we need to check the context each time. As we inspect the missing values more carefully, we find some rides are missing end data like end station name, end latitude, and end longitude. It might be interesting to do a little more research on that topic. Namely, does that mean some sensors/parts are broken, or that bike is stolen? Anyway, for this project, considering the large size of the dataset, we would like to drop all rows of data having any missing values.

Stage 1 - Analysis of July 2024

In this stage, we are going to investigate in a recent month (July 2024) to have a basic understanding of the recent bike sharing environment in New York.

According to what Citi Bike provides, there are over 4,700,000 rides in the month of July. We set the sample fraction to be 0.01 which is resulting in a random sample of about 47,000 rides. A dataset of such size is good for us to wrangle with on a personal computer. Since we sampled with a fraction of 0.01, all the number in this part is supposed to be multiplied by 100.

When we carry out Exploratory Data Analysis, plotting is a good means to help us understand the data. Tableau is a powerful tool available to handle the task in an elegant way and we are going to use it as a main force to help plot. Another strength of Tableau is that we can make dynamic dashboards by combining multiple sub-plots we created to obtain more insights and interesting and meaningful findings.

We plot the ride count by day and divide into member and casual riders.We find some interesting things:

- Overall, the ride count fluctuates like a wave of a cycle of a week.
- Overall, more people ride from Monday to Friday.
- However, more casual riders are during the weekends than working days.
- Members' riding count waves more intensely, compared with casual riders.

A logical explanation could be the majority of Citi Bike users are using Citi bikes as a transportation to commute daily for their jobs. Note that we are not saying all those members are using bikes to commute. During weekends, a little bit more casual take a Citi bike to travel around. This could because of either more tourists or other public transportation's special operation during the weekends like fewer subway shifts/services.

What are some top popular Citi Bike stations in New York? This could be an interesting question, and we can easily find out the answer with this dataset. As we find out, "Broadway & E 14 St" is the most popular station where people pick up a bike and "W 21 St & 6 Ave" is the most popular station where people drop off their bikes. Both two stations are near Flatiron District and Union Square where both locals and tourists gather together.