Citibike New York City

October 2015 - October 2020 statistical data analysis

Project Summary

Group 16

Wang, Su, s3999wang
Mabbayad, Mabi, mmabbaya
Xue, Kira, w27xue
Jajcanin, Igor, ijajcani
Cebula, Daniel Adam, dacebula
Shi, Wenwen, w65shi

This project is analyzing data gathered by New York City Citi Bike Program. We have collected and analyzed data from Oct 2015 to Oct 2020.

Collected data was based on start and end time of each bike trip, duration of the trip and bike model based on model ID. Data does no include other factors of the trip, related to weather conditions, traffic congestions nor actual trip path.

The project includes data gathering, cleaning, grouping and data analysis based on the bike ID (type) , gander and trip duration, origin and destination station information that allowed to calculate trip time and distance duration.

Data source is https://www.citibikenyc.com/system-data. This data is originally NYC Citi Bike Rental Program from Oct 2015 to Oct 2020. These are fairly large data collection with total of 5GB of data.

Here are the attributes information:

TRIP_DURATION': Duration of the trip (seconds)
'START_TIME': Trip Start date and time
'STOP_TIME': Trip End date and time
'START_STATION_ID': Trip Start Station ID
'START_STATION_NAME': Trip Start Station Name
'START_STATION_LATITUDE': Trip Start Station Latitude
'START_STATION_LONGITUDE': Trip Start Station Longitude
'STOP_STATION_ID': Trip End Station ID
'STOP_STATION_NAME': Trip End Station Name
'STOP_STATION_LATITUDE': Trip End Station Latitude
'STOP_STATION_LONGITUDE': Trip End Station Longitude
'BIKE_ID': Bike ID
'USER_TYPE': User_type (Customer – 24 hrs or 3-day pass user, Subscriber – Annual Member)
'BIRTH_YEAR': Year of Birth
'GENDER': Gender (0 – unknown, 1 – male, 2 – female)

After data cleaning, we did the analysis to better understand dataset. Some of the available features included geographical coordinates of the stations that we used to calculate distance between start and end stations using Haversine formula. Also, by using categorical function we managed to significantly reduce size the dataframe to more than 1/3 of the original 4.9GB and that dataset was later used for data analysis.

Upon the analysis on relationship between bike trip duration and distance between start and end station, we found there is no linear relationship between then, even though common sense might conclude that the longer the duration with further the distance would be.

Based on the other analysis that we made we found top 5 performing bikes (based on bike ID). Plots made out of it revealed to be typical Poisson distribution. Furthermore, the histograms of Bike Trip Duration stand and end station indicated to combined Gaussian and exponential distribution.

Based on that we made 3 Bayesian models

1. Total count of bike trips
2. Total duration
3. Total distance.

We created SARIMA time series analysis, since the data provided contain timestamps for each trip, and we tried to used them to predict how many trips could be expected in the future. After cleaning startTime columns we aggregated columns based on hour and summarized trips on hourly bases. That allow us to create time series on number of bike for each hour.

We completed analysis to prove seasonality of the data. Though it was hard to determine if the trend exists there was indication of the seasonality.

Based on ACF and PACT plots we used SARIMA mode and ensured that the model is valid by analyzing the residuals. The model validation plots show that:

- The residuals plot shows no trend in the residuals
- The histogram shows that residuals are normally distributed and centered on zero
- The normal Q-Q plot is somewhat linear
- The correlogram plot shows that some values are outside of the 95% confidence bounds

November data was used for forecast, while October being used as training data). We found that the model was able to capture general trends but as not very accurate at predicting exact values.

We might need to adjust the model to account exogenous factors and also there ridership was significantly different due to COVID situation.

Lastly, data required in OLS analysis were users birth year, distance, trip duration and speed. Birthyear was provided in the dataset, whereas distance and speed were calculated and grouped for each birth year. The analysis was focused on users who were born after 1994.

We wanted to prove relationship between birth year and duration, distance and speed (in m/s). We found that there is strong liner correlation between birth year and distance as well as duration. The results show that speed is much higher for younger bike riders compared to older people.

Categorical function proved to be very efficient in reducing the size of the dataframe. Overall dataset was fairly scattered and it appear to be no linear relationship between indicating that longer duration would mean traveling longer destinations. Analysis indicate of combined Gaussian and exponential distribution. Based on Bayesian models it appears the exponential distribution is a good for fitting posterior randomly sampled values. On time series analysis there was showing of a seasonality trend. Using SARIMA model was good to capture general trends. However, the it was lacking in predicting exact values. That was most likely due to the differences in test data due to COVID lockdowns and downturns. T-test to prove relationship in distance traveled and gender showed larger T values. Therefore, the conclusion is that the data for duration time periods and distances between stations have quite different distributions for women and men. Finally, OLS method used to show dependences of age group (i.e., birth year) and distance and speed show that younger population were traveling on higher speed.