Applied Data Science (MAST30034) Project 1: Quantitative Analysis

Quantitative Analysis Of The New York City Taxi Data

Yuhao Zhai
Student ID: 1067899
Github repo with commit

October 27, 2022

1 Introduction

In New York City, taxicabs come in two varieties: yellow and green; they are widely recognizable symbols of the city. Taxis painted yellow (medallion taxis) are able to pick up passengers anywhere in the five boroughs. Those painted apple green (street hail livery vehicles, commonly known as "boro taxis"), which began to appear in August 2013, are allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island. Both types have the same fare structure. Taxicabs are operated by private companies and licensed by the New York City Taxi and Limousine Commission (TLC). It also oversees over 40,000 other for-hire vehicles, including "black cars", commuter vans, and ambulettes.[1]. Taxicabs are the only vehicles that have the right to pick up street-hailing and prearranged passengers anywhere in New York City. By law, there are 13,587 taxis in New York City and each taxi must have a medallion affixed to it. Medallions are auctioned by the City and are transferrable on the open market by licensed brokers.[2]. It is important for us to find out the how weather effect the New York City taxi business.

2 Preprocessing, Analysis, and Geospatial Visualisation

2.1 Data Selection

2.1.1 Yellow Taxi Trip Records

In this project, we use yellow taxi trip data for the year 2019 January April, July, and October. Since these 4 months represent the four seasons of the New York City. Which can use for weather analysis. The data set can be downloaded from from the New York City Taxi and Limousine Commission. The data set has 18 columns: such as pick-up and drop-off dates/times, pick-up and drop-off locations, journey distances, all kinds of fares, fares types, payment types, and driver-reported the number of passengers.

2.1.2 weather

The weather data I use is from kaggle New York City Weather Data 2019.This dataset was gathered to be used in the Big Data Derby 2022 competition. However, it can obviously be used for any other

purposes. Horses are affected by weather conditions, so knowing the day's temperature, snowfall, precipitation, etc will definitely be useful.Weather data collected from the National Weather Service. It contains daily data from all days in 2019. It contains for each day the minimum temperature, maximum temperature, average temperature, precipitation, new snow fall, and current snow depth. The temperature is measured in Fahrenheit and the depth is measured in inches. T means that there is a trace of precipitation.[3]

2.1.3 Taxi Zone data

The taxi zone data also downloaded [4]. This data got each taxi zones geometric information. Also with the Taxi Zone Look Up Table file contains TLC taxi zone location IDs, location names, and each zone corresponding boroughs.

2.2 Preprocessing

It approximately got 28696876 trip records for 2019 year 4 months data set with 19 columns of attributes. The data takes 4.1G memory. By changing the datatypes we reduce the data memory usage.

2.2.1 Data Cleaning

After we reduce the memory usage, we check the missing value in the yellow taxi data set. 151810 values are missing for passenger_count, RatecodeID, store_and_fwd_flag features. 5008025 missing values for the congestion_surcharge feature. According to the feature context, I am able to filter out some error like the distance can't be negative. Finally, the PULocationID and DOLocationID should within NYC taxi zone.[4]

Next, in figure 2 and in figure 3,we plot the data distribution.

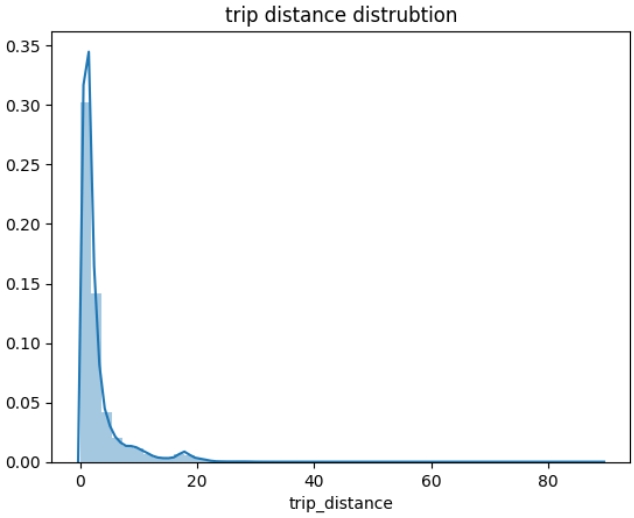

Figure 1: trip distance distribution

2.2.2 Feature Engineering

The difference between pickup and drop-off time is a useful information for us to know the period of each trip. The features like tpep_pickup_datetime, tpep_dropoff_datetime, store_and_fwd_flag that are

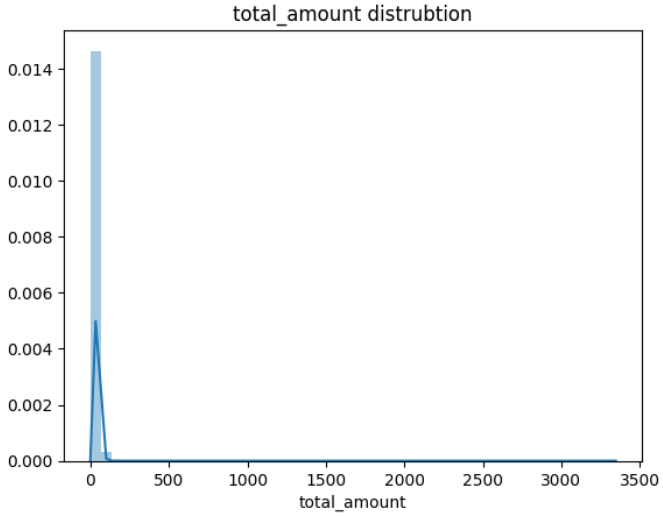

Figure 2: total amount distribution

no longer the features we mainly focus on, these features are dropped. Finally, I select the date in the 4 months of both taxi and weather dataset.

2.3 Geospatial Visualisation

2.3.1 Geospatial and Time Analysis

These 4 months the total amount of activity in New York City is show. In figure 8, The brighter and yellow regions indicate that there are exits more taxi activity. In Manhattan it got most pickups and drop-off activities.

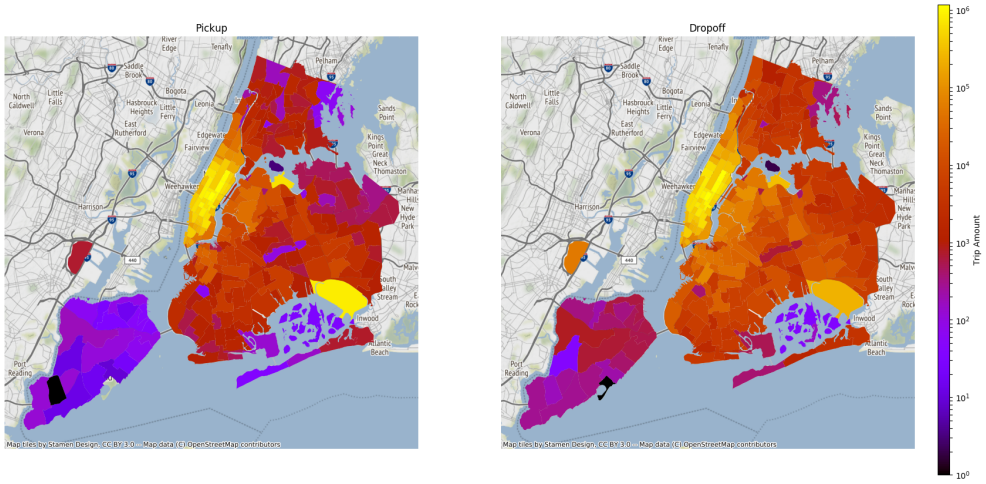

Figure 3: Taxi Pickup Frequency

2.3.2 Weather Impact

In this section we investigate how weather affects the taxi bussiness. The following scatter plots shows the weather Average temperature of the day in F versus different data features like 'passenger_count','trip_distance','ti

'total_amount'.

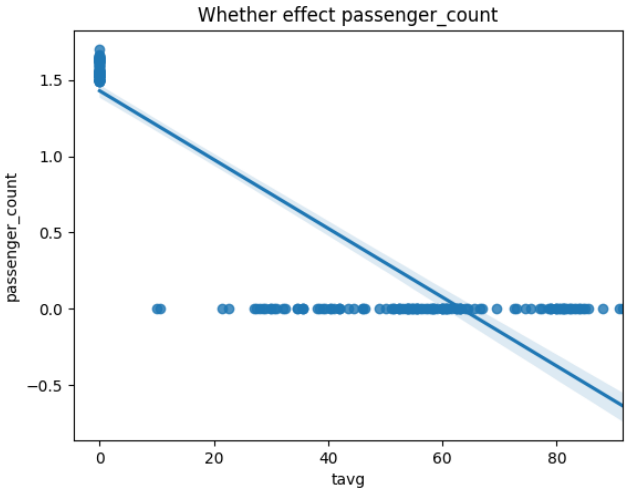

Figure 4: Weather Impact passenger count

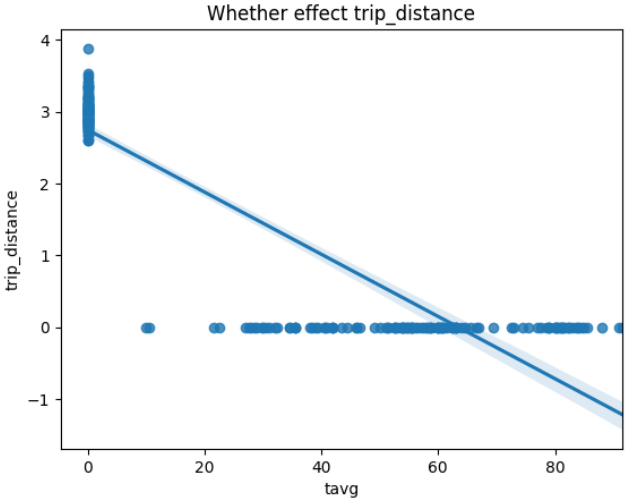

Figure 5: Weather Impact trip distance

From figure 8 , We can draw the conclusion that the taxi bussiness get more tip and trip distance when the weather gets cold.

2.3.3 Season affect

In this section we shows that how the seasons affect the taxi bussiness. As shown in figure 13, The trip distance, tip amount doesn't change much through the seasons. The period and total amount get decrease in winter compare to the other seasons.

3 Modelling

In Modeling section we want to see which features contribute to the total_amount feature. Since the total_amount is the most important for taxi driver.

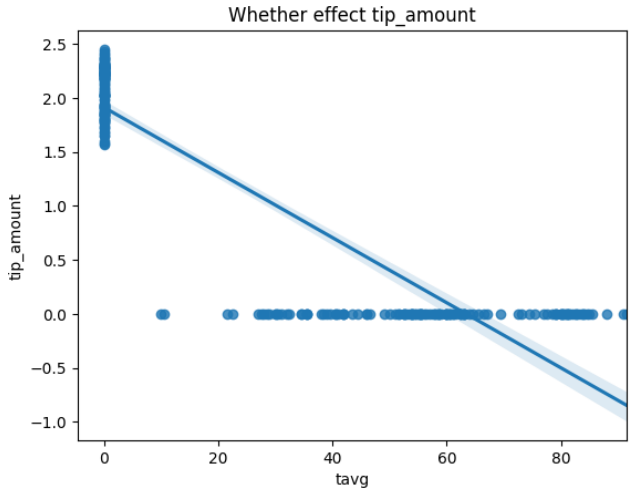

Figure 6: Weather Impact tip amount

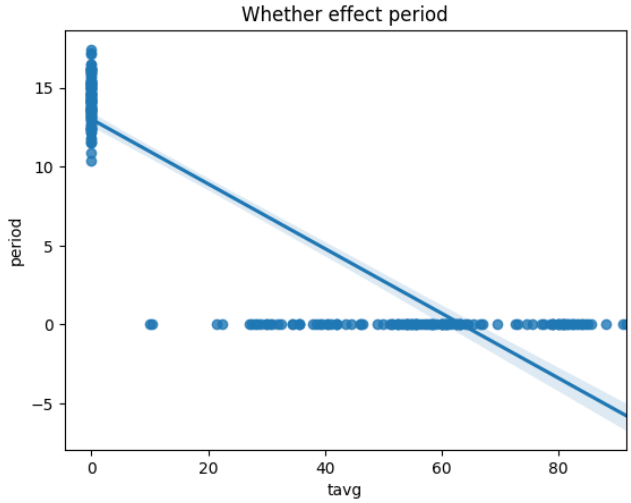

Figure 7: Weather Impact period

3.1 Correlation Coefficient

The correlation between attributes in the yellow taxi 2019year 4 months shown in figure 10, the features like 'trip_distance', 'fare_amount', 'tip_amount', 'tolls_amount', 'period' get high correlation scores of the total_amount.

3.2 regression model

The total_amount data feature will be predict using three regression models. They are linear regression model, Ridge regression model and Lasso regression model. According to the correlation heat map we created. We use feature 'trip_distance', 'fare_amount', 'tip_amount', 'tolls_amount', 'period' as our X feature. The y is of course the total_amount. The train-test split method is used to randomly split sampled data into 70% training set and 30% testing set.

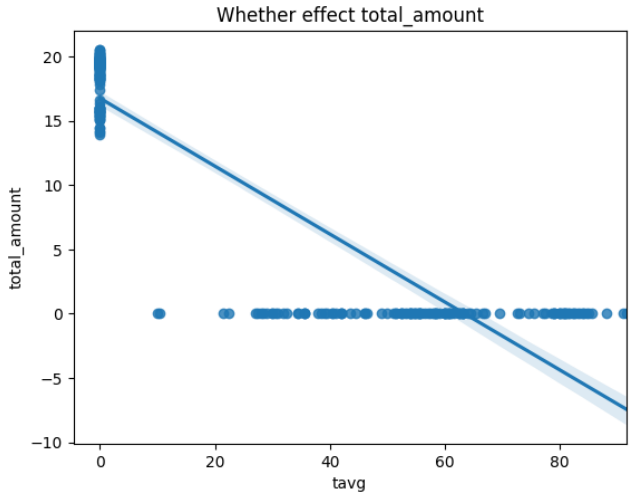

Figure 8: Weather Impact total amount

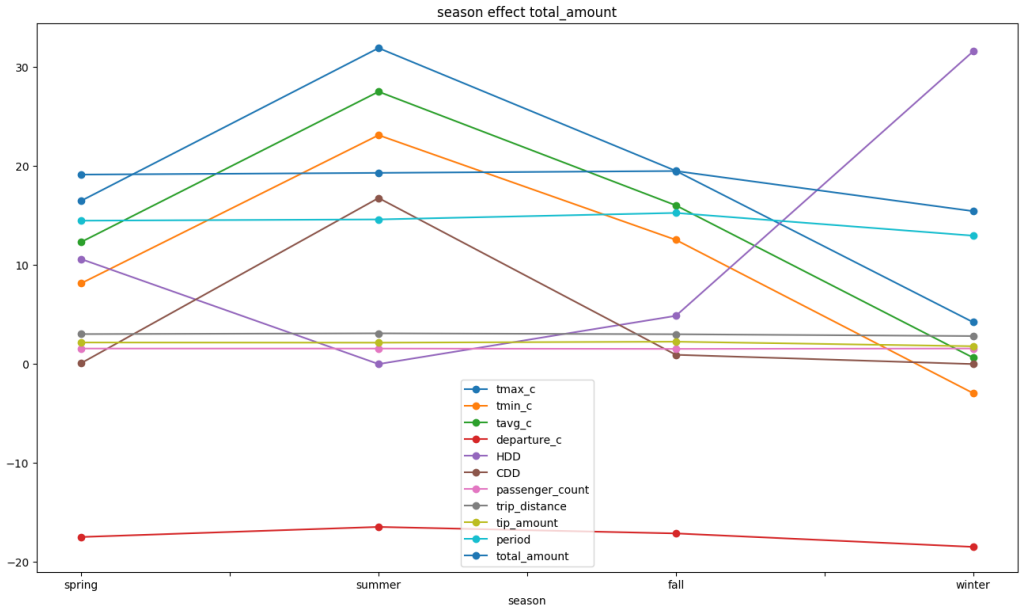

Figure 9: Season affect

3.3 Evaluation and Results

We evaluate the model by using Root Mean Square Error R2 score. The result is show in Table1.

The coefficient of the three models are show:

3.4 Discussion

As we can see of the three regression models, the fare_amount features contribute to the final total_amount most, since its coefficient is the biggest. We can indicate that the more fare_amount the more total_amount.

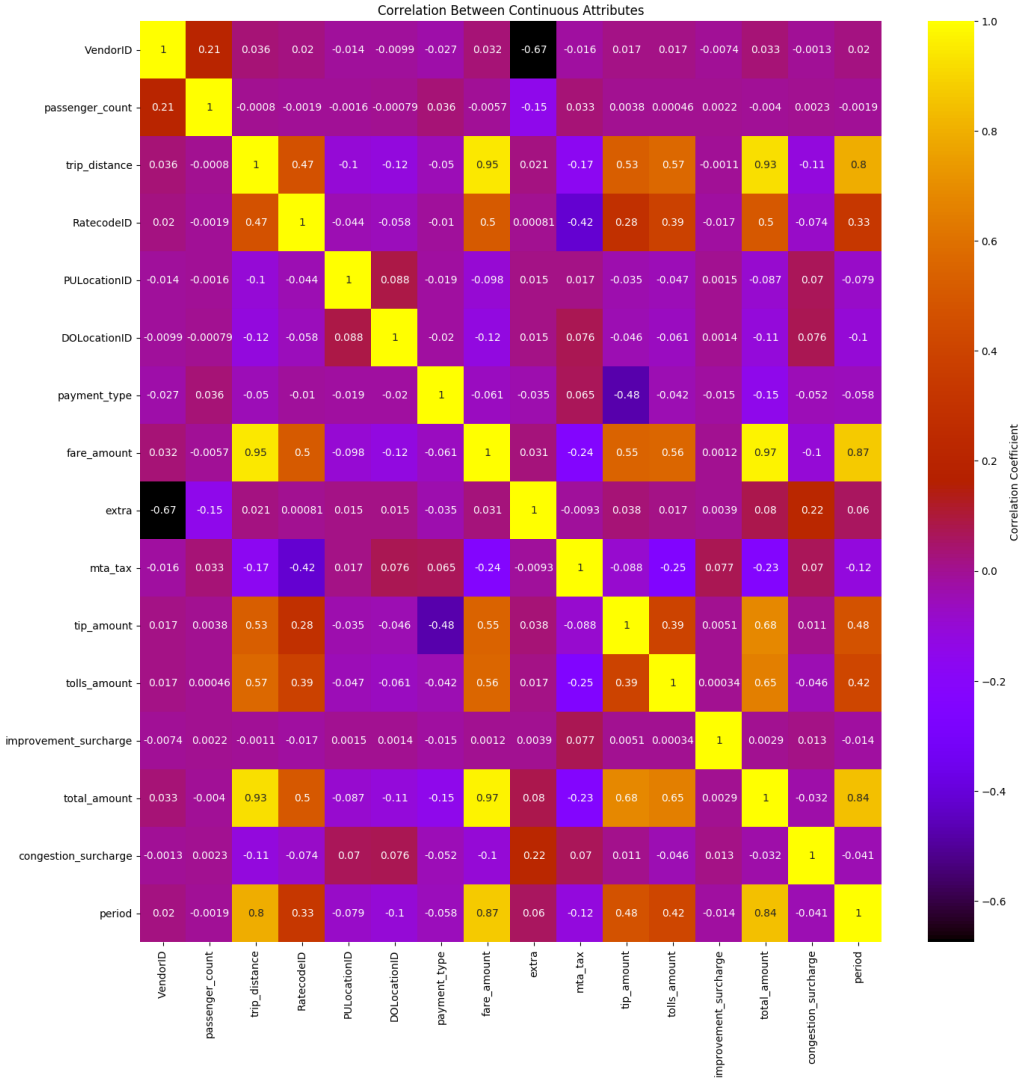

Figure 10: Correlation Coefficient

4 Recommendations

After all the analysis, we can learn that the weather affect the New York City taxi industry greatly. The colder the weather is the more people want to take taxi. But the cold day taxi trip distance is less than other season. In order for the taxi driver make more money the fare amount can take higher.

5 Conclusion

In conclusion, after doing the exploratory analysis of yellow taxi trip records in 2019 4 months and combine with New York City Weather Data 2019 data. We learn much more about the New York City taxi industry. The visualisation help us taxi usage on different weather. We also build model to predict the total amount. The fare_amount affect the total_amount the most.

Table 1: result table

Model	Train RMSE	Test RMSE	Train R2	Test R2
Linear	1.25	1.25	1.0	1.0
Ridge	1.25	1.25	1.0	1.0
Lasso	1.25	1.25	1.0	1.0

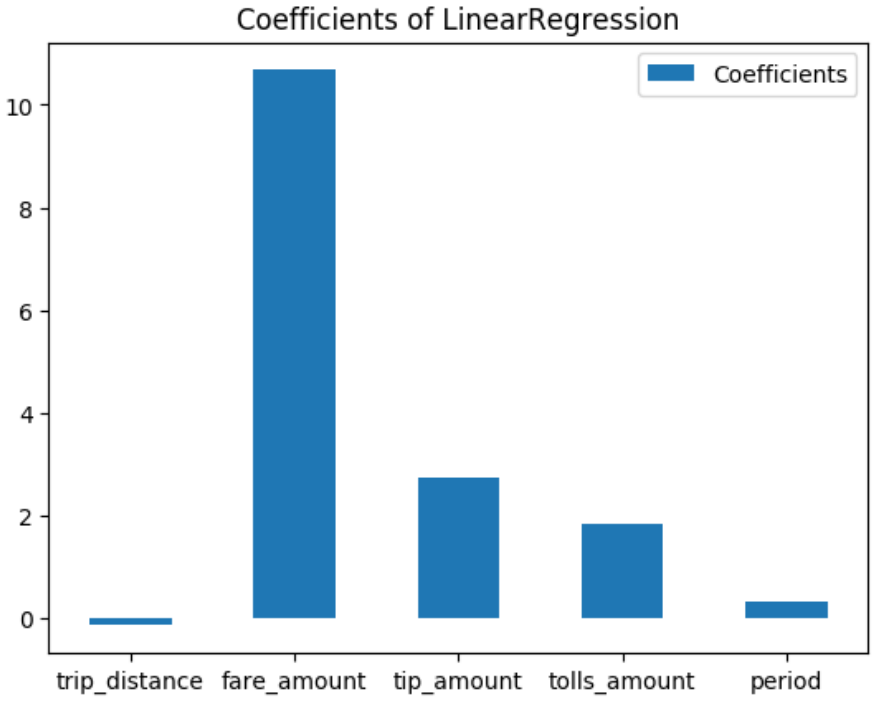

Figure 11: Linear

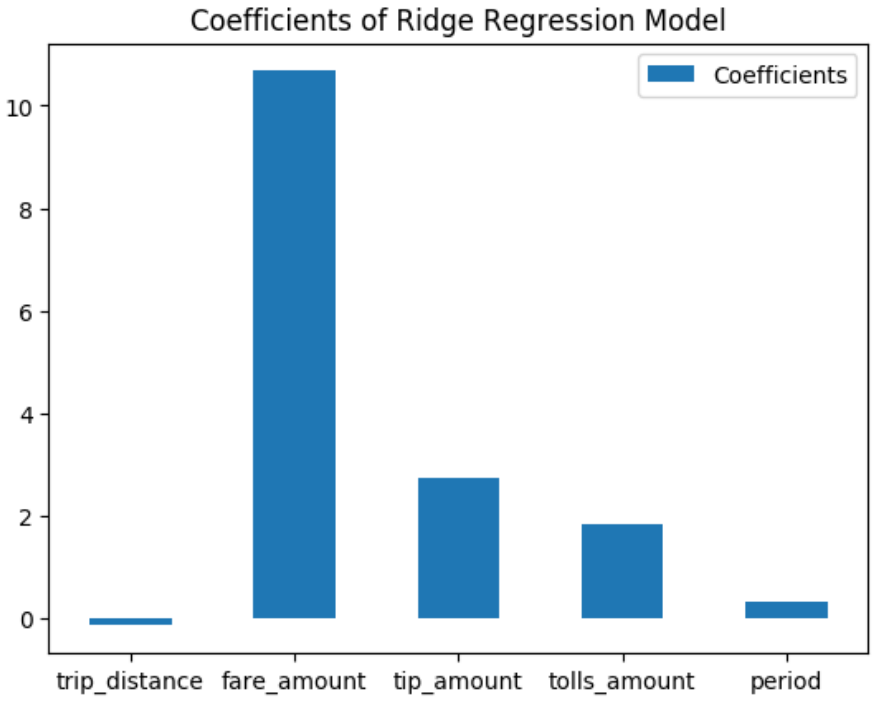

Figure 12: Ridge

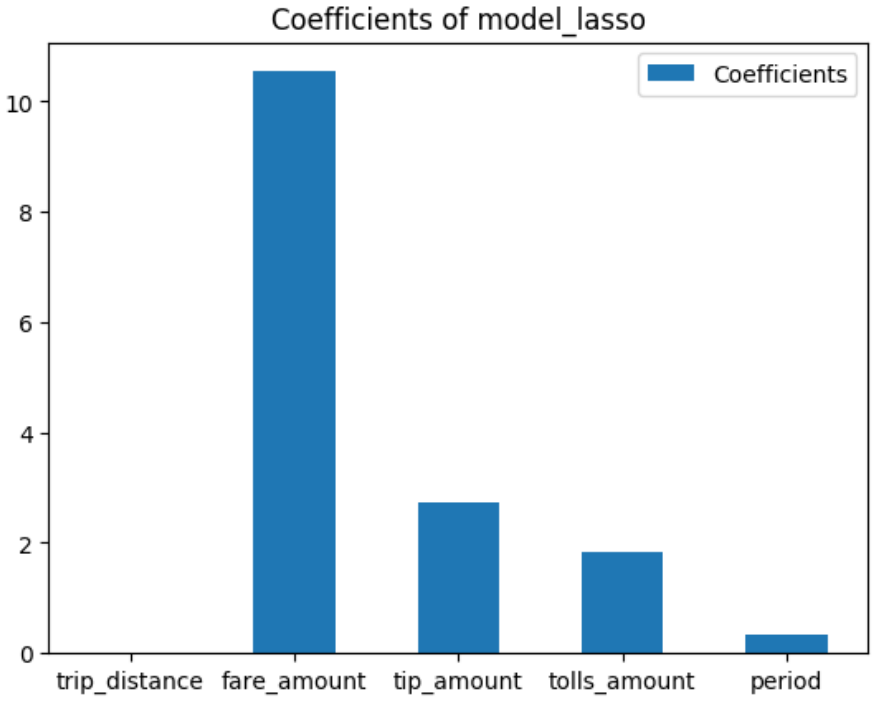

Figure 13: Lasso

References

[1] Wikipedia. *Taxis of New York City*. 2022. URL: https://en.wikipedia.org/wiki/Taxis_of_New_York_City (visited on 08/14/2022).

[2] NYC government. *Yellow Cab*. 2022. URL: https://www1.nyc.gov/site/tlc/businesses/yellow-cab.page (visited on 08/14/2022).

[3] Kaggle. *New York City Weather Data 2019*. 2022. URL: https://www.kaggle.com/datasets/alejopaullier/new-york-city-weather-data-2019 (visited on 08/14/2022).

[4] NYC government. *Taxi Zone Lookup Table*. 2022. URL: https://d37ci6vzurychx.cloudfront.net/misc/taxi+_zone_lookup.csv (visited on 08/14/2022).