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

When we talk about the symbols of the New York, you might think the Statue of Liberty. But one thing that also is the symbols of the New York city is taxi. There are two varieties taxicabs yellow and green in New York City. They have same fare structure. The difference is that the yellow one can pick up passengers anywhere in the five boroughs, while the green one can pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens and Staten Island. The New York City Taxi and Limousine Commission (TLC) a private companies that license the taxi operate.[1]. The only vehicles that can pick up passengers anywhere in New York City is the taxi. Each taxi must have a medallion affixed to it by law.[2]. Taxis are comfy, they are available all the time, It is important for us to find out the how weather effect the New York City taxi business.

2 EDA

2.1 Load Data

2.1.1 Yellow Taxi Trip Records, weather and Taxi Zone data

The Yellow Taxi Trip Records data we use is in the year 2019 January April, July, and October yellow taxi trip data in New York City. Since these 4 months represent the four seasons of the New York City. Which can use for weather analysis. The data is provided in New York City Taxi and Limousine Commission.

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. The data collects in 2019 all days weather data, include each day the minimum temperature(measured in Fahrenheit), maximum temperature(measured in Fahrenheit), average temperature(measured in Fahrenheit), precipitation(measured in inches), new snow fall(measured in inches), and current snow depth.[3]

The Taxi zone data [4]. By using this data we can get where the taxi in New York City, combine with the Taxi Zone Look Up Table file we can draw a map in New York City.

2.2 Preprocess Data

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 Clean Data

First thing we check is the missing data. Yellow taxi data set missing value is list below: 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 1 and in figure 3,we plot the data distribution.

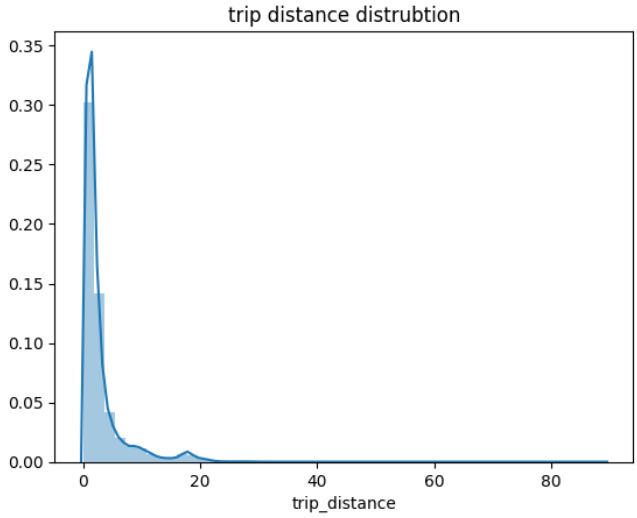

Figure 1: trip distance distribution

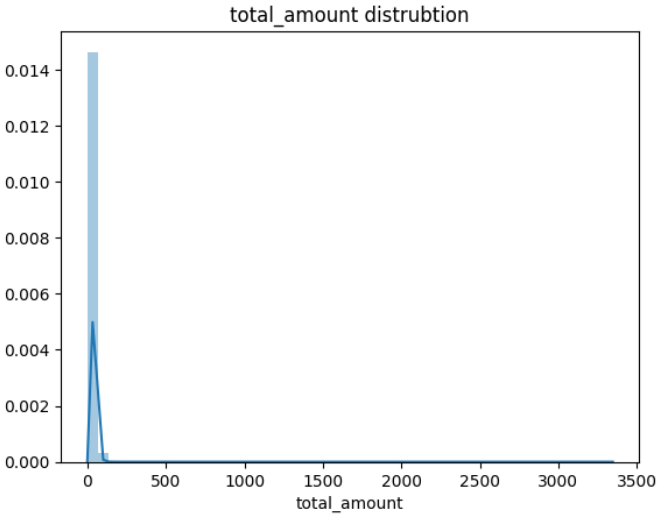

Figure 2: total amount distribution

2.2.2 Create Feature

The difference between pickup and drop-off time is a useful information for us to know the period of each trip. The features: tpep_pickup_datetime, tpep_dropoff_datetime, store_and_fwd_flag that are no longer the features we needed. Finally, I select the date in the 4 months of both taxi and weather dataset.

2.3 Visualisation

2.3.1 Map

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.

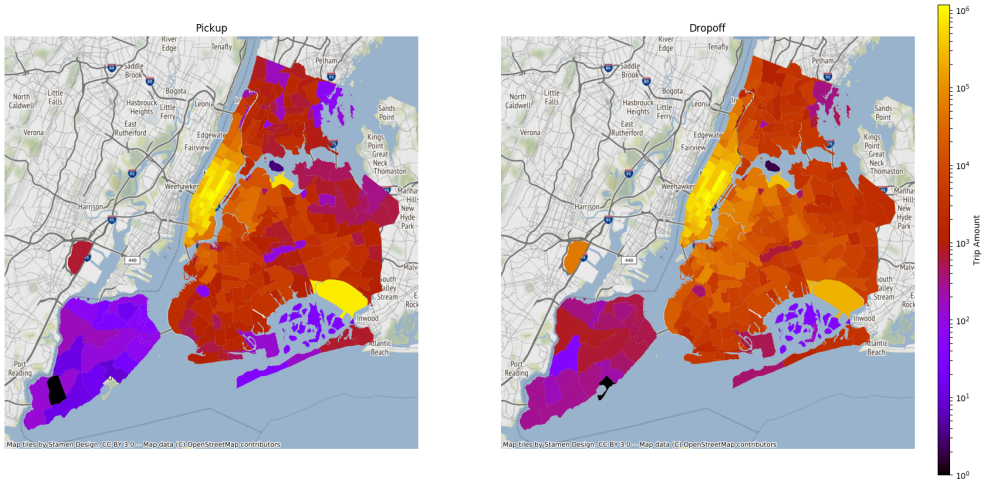

Figure 3: Taxi Pickup Frequency

2.3.2 Weather Impact

In this section we investigate how weather affects the taxi business. The following scatter plots shows the weather Average temperature of the day in F versus different data features like:

'passenger_count','trip_distance','tip_amount','total_amount' 'period'.

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.

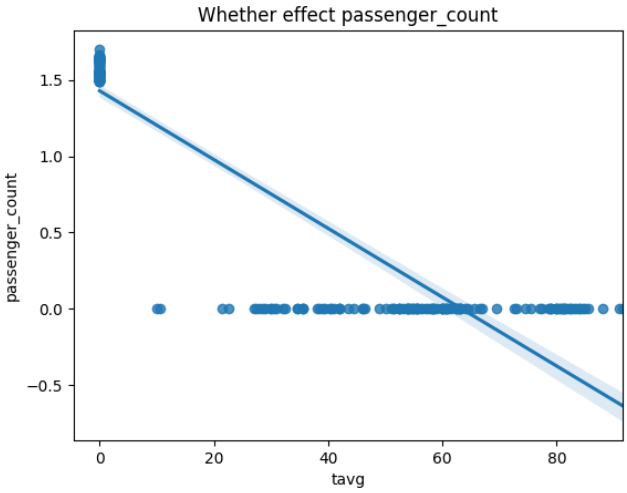

Figure 4: Weather Impact passenger count

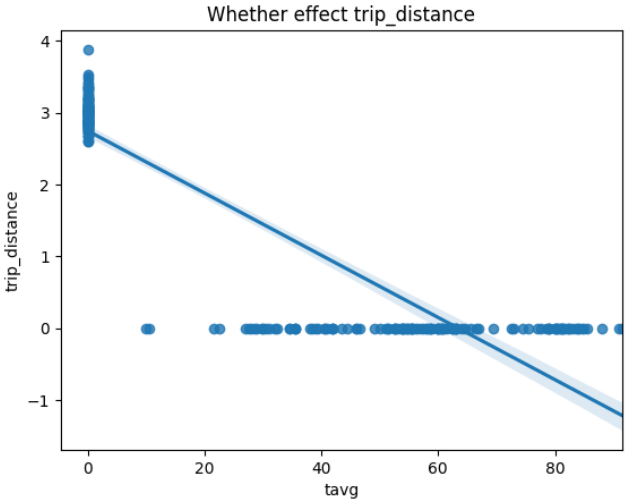

Figure 5: Weather Impact trip distance

3.1 Correlation

The correlation heat map about data features 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.

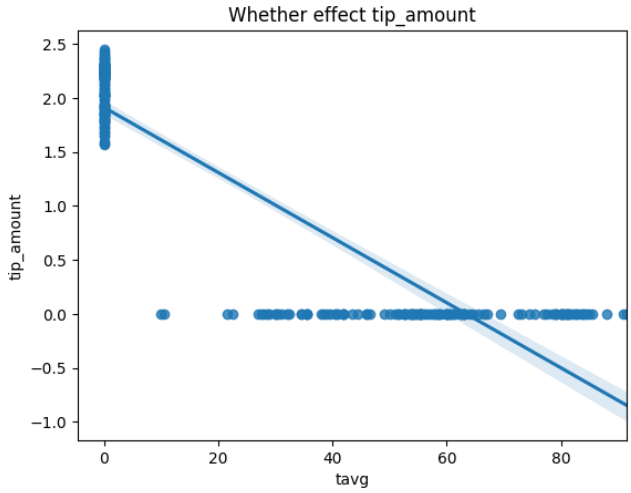

Figure 6: Weather Impact tip amount

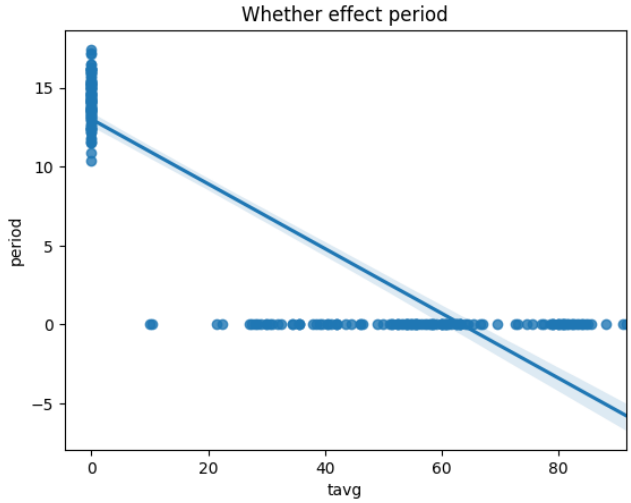

Figure 7: Weather Impact period

3.3 Evaluation Results

When we create model, we need to know how well it perform. The evaluation step is for us to know how well the model perform. By using Root Mean Square Error and R2 score. We get a general idea about how well the model is perform. In the table1, we can see that all model perform the same.

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.

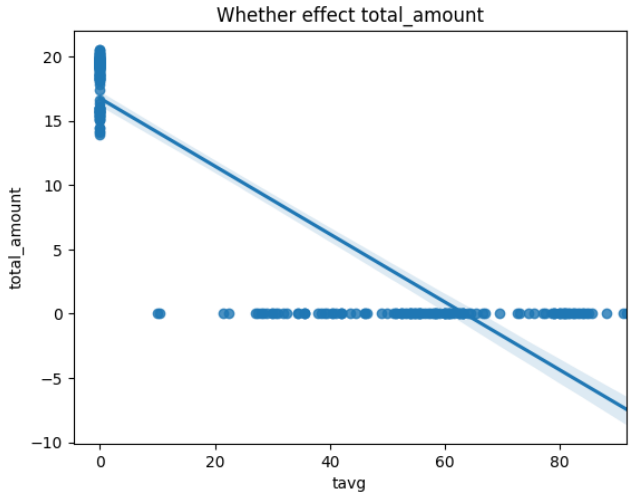

Figure 8: Weather Impact total amount

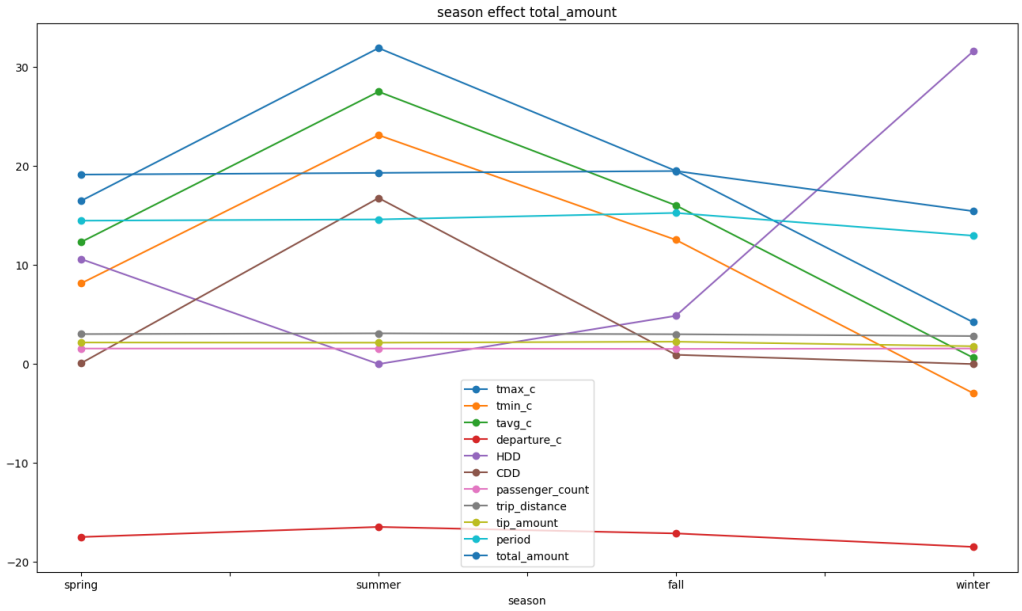

Figure 9: Season affect

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

All in all, by looking at the 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

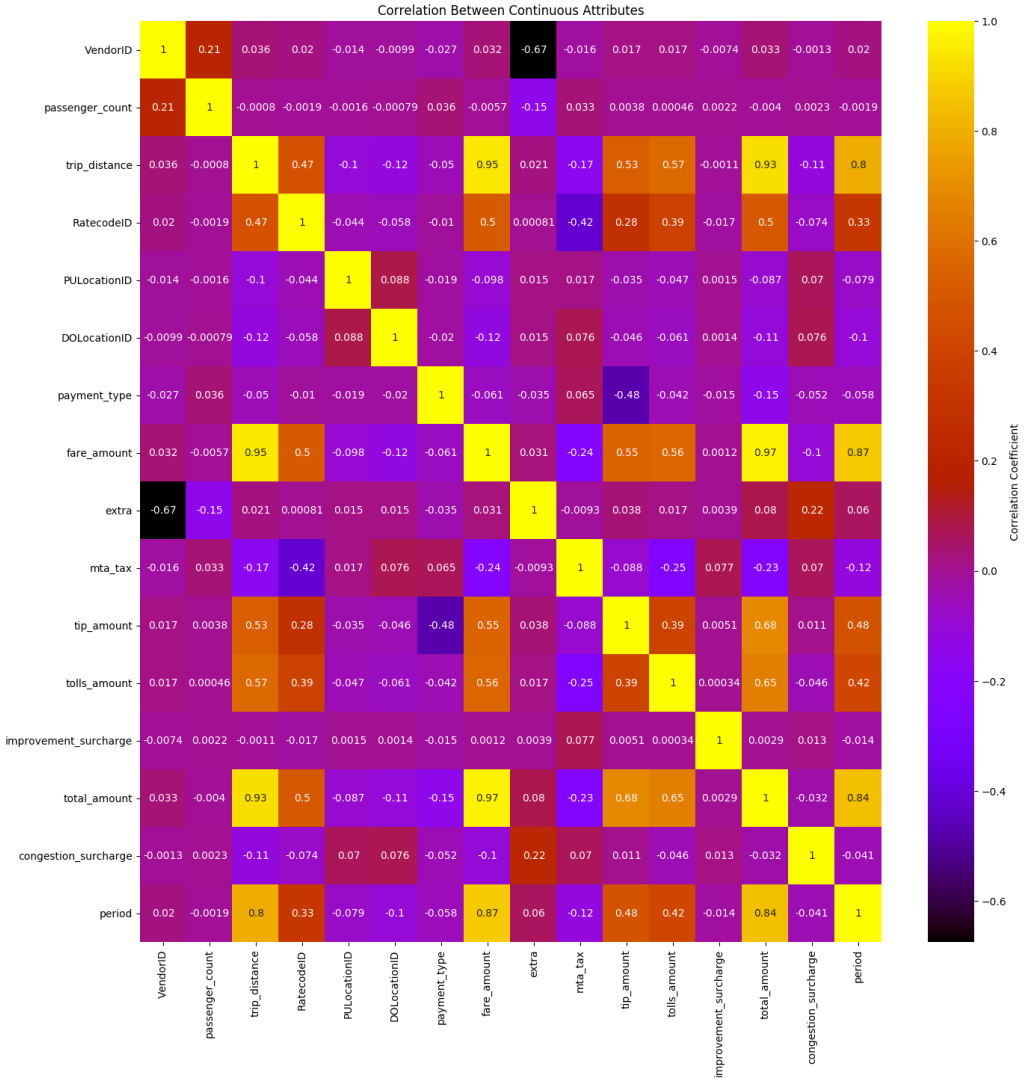

Figure 10: Correlation Coefficient

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

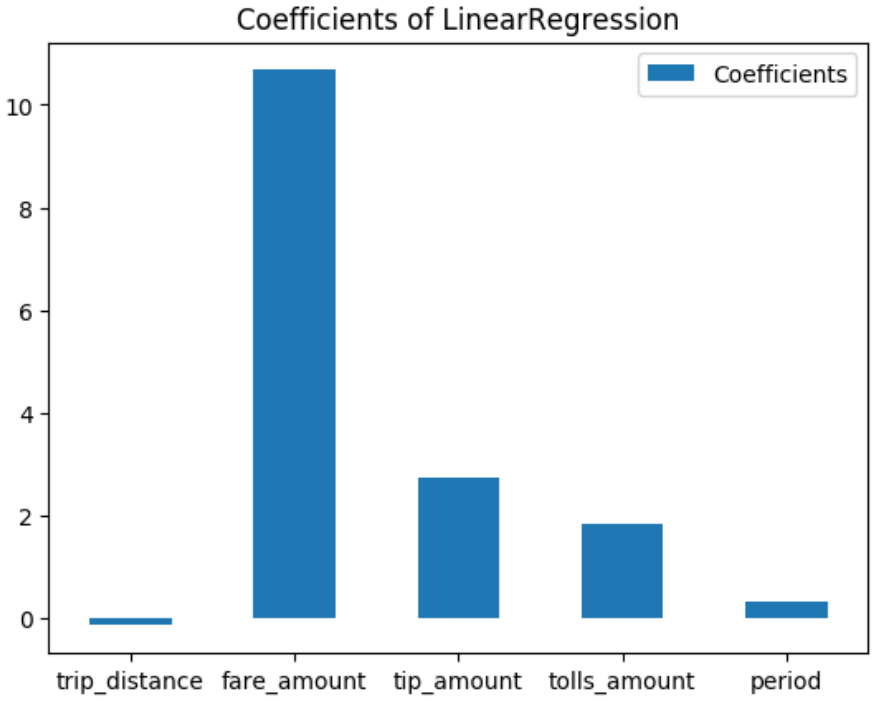

Figure 11: Linear

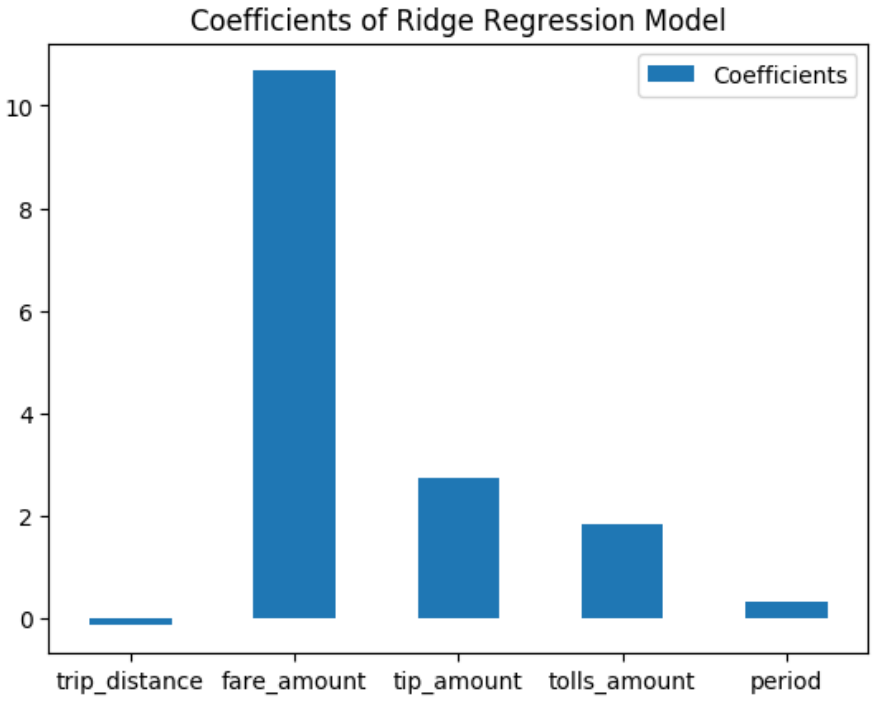

Figure 12: Ridge

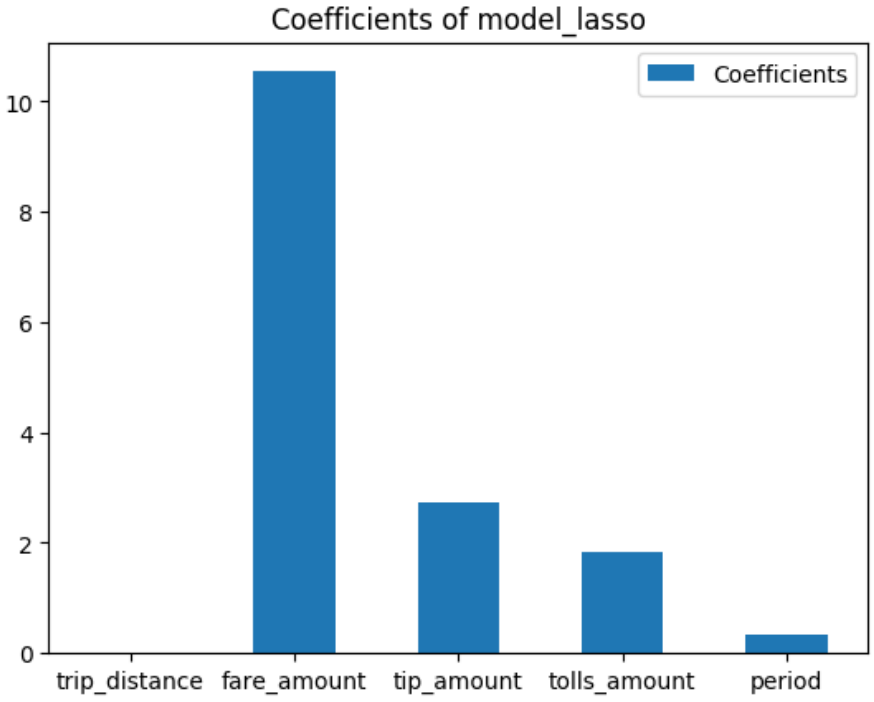

Figure 13: Lasso

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).