Business Opportunities- Extra Services Plan for New York City Taxi Passengers

Taxi Trip Record Dataset Analysis

Haofei Xiao
Student ID: 1072038

August 8, 2021

1 Introduction

In New York City, there are various types of taxi licenses: yellow, green and for hire vehicles. In this project, the focus is on yellow and green taxis. Yellow cabs are the official and iconic taxis in NYC mainly served in Manhattan. Green taxis are allowed to pick up passengers in northern Manhattan (north of West 110th street and East 96th street), and anywhere in the Bronx, Brooklyn, Staten Island, and Queens (excluding the airports).

Nowadays, driving people to their destination should not be the only focus of the taxi company, rather, some appropriate offers of extra services could help boost company revenue and reputation. For example, passengers might be willing to pay more for a luxury car model. Thus, The aim of the project is to provide **advice to cab companies about the passengers that might be willing to pay more money for special offers**, based on the useful information extracted from the dataset during 2016.

1.1 Data

The Trip Record Data of **yellow and green** licenses during **April to June in 2016** are selected for analysis. April to June is the relatively busy work period during the year, thus reflecting a reliable background statistic. Moreover, the selected timeline is prior to the Covid dynamic, so it shows more of a normal economic environment, **assuming** the client is asking for the plan during the regular period.

Attributes of both yellow and green are similar, including pick up and drop off datetime, longitude and latitude, passenger count, trip distance, and the total fee amount which includes extra payment, tax, tips, tolls, and improvement surcharge. The payment type only includes cash and credit card as the method because the majority of the cases are contained in these two types.

1.2 Other external data

The extra dataset is supposed to help determine **consumer behavior** based on the trip record data because tourists often travel through taxis and live in hotels. The extra dataset lists most of the hotels in New York City, and the attributes engineered are the longitude and latitude, the star rating, high

and low rate of the hotel. The **assumption** made in this project is that there is an effective amount of passengers taking taxis as transportation are tourists, and tourists usually take taxis to hotels.

2 Preprocessing

The trip record dataset downloaded from TLC and hotel record from Kaggle are arranged in csv format with neat data type. However, applying data preprocessing is necessary to gain a typical dataset for the project analysis. The data cleaning process is done using Python; datasets are broken into chunks during implementation to reduce memory usage, and processed data are saved for further access.

2.1 Cleaning process

Because the yellow and green dataset are having similar attributes, the cleaning steps for both dataset are as following:

- Remove Vendor ID, store and fwd flag because of the irrelevance.
- Remove Fare amount after reading through the statistic, the fare amount of a trip is usually fixed and not useful for analysis.
- Remove Ehail fee because this column is all 0s in the statistical overview which has no contribution to the analysis.
- Remove the rows that have the same pick up and drop off time because these instances usually result in **outliers assuming fault enters**.
- Remove the instances in pick up, drop off in the respect of longitude and latitude that have values of 0 because they are often the **outliers** or **missing values**.
- Filtered passenger count, trip distance and all other fields involving charges (e.g. tip amount) that are greater than 0 because the instances with 0 or less make no sense and are noisy data.
- Filtered payment type of only cash and credit card because these two contain the majority of the instances.

For external dataset containing NYC hotel information, the cleaning process work as follow:

- Remove the id, name, address, city, state and postcode because they are irrelevant to the analysis.
- Remove the latitude and longitude, star rating and hotel rating that contain 0 because of missing value or noise.

2.2 Dataset shape

Before the preprocessing, the data is very poorly fitted in normal distribution and have many missing values and outliers on both sides of the distribution. After being preprocessed and taking logarithm, the overall shape of data is satisfactory. There are a few outliers as shown in figure 1, but after the filtering, those outliers should contain useful information and should not be removed.

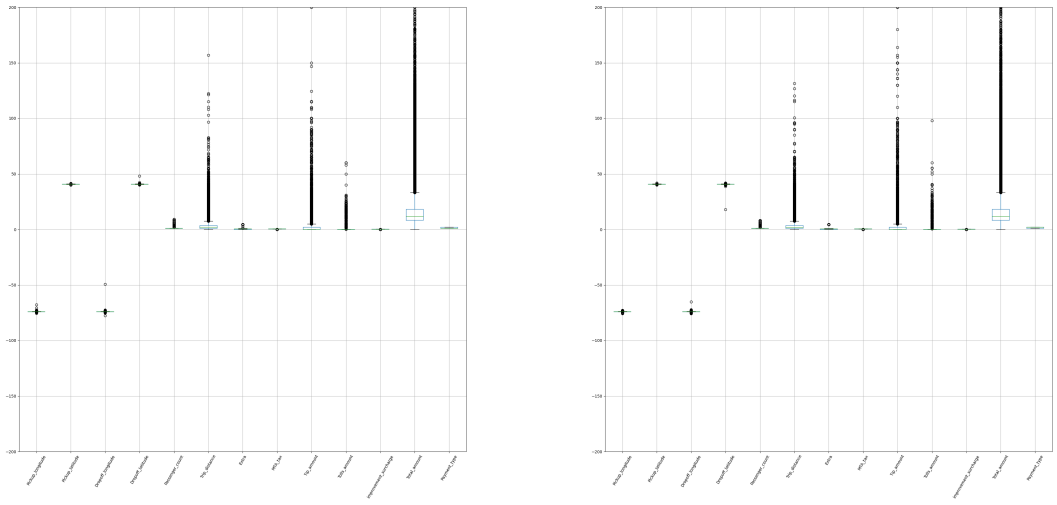

Figure 1: shape of dataset green(left) vs yellow taxi(right) after processed (for the attributes that have negative value filtered out, so there is no negative value presented)

3 Data Visualisation

The dataset visualisation takes three aspects into account: the general correlation and distribution, the correlation of interested fields, and the overall geolocation of the taxi trip record. The purpose of visualising the data is to have direct examination of important variables in both macro and micro scope.

3.1 Heatmap

The figure 2 shows the correlation between the fields of interest. The pickup and dropoff datetime is converted into travel time for visualization. The feature engineering plot below shows that for both types of license, pick up and drop off latitude and longitude are relatively well correlated with trip distance, tip and tolls amount, and travel times, which make sense when considering common sense that the location of destination is truly relevant to these elements.

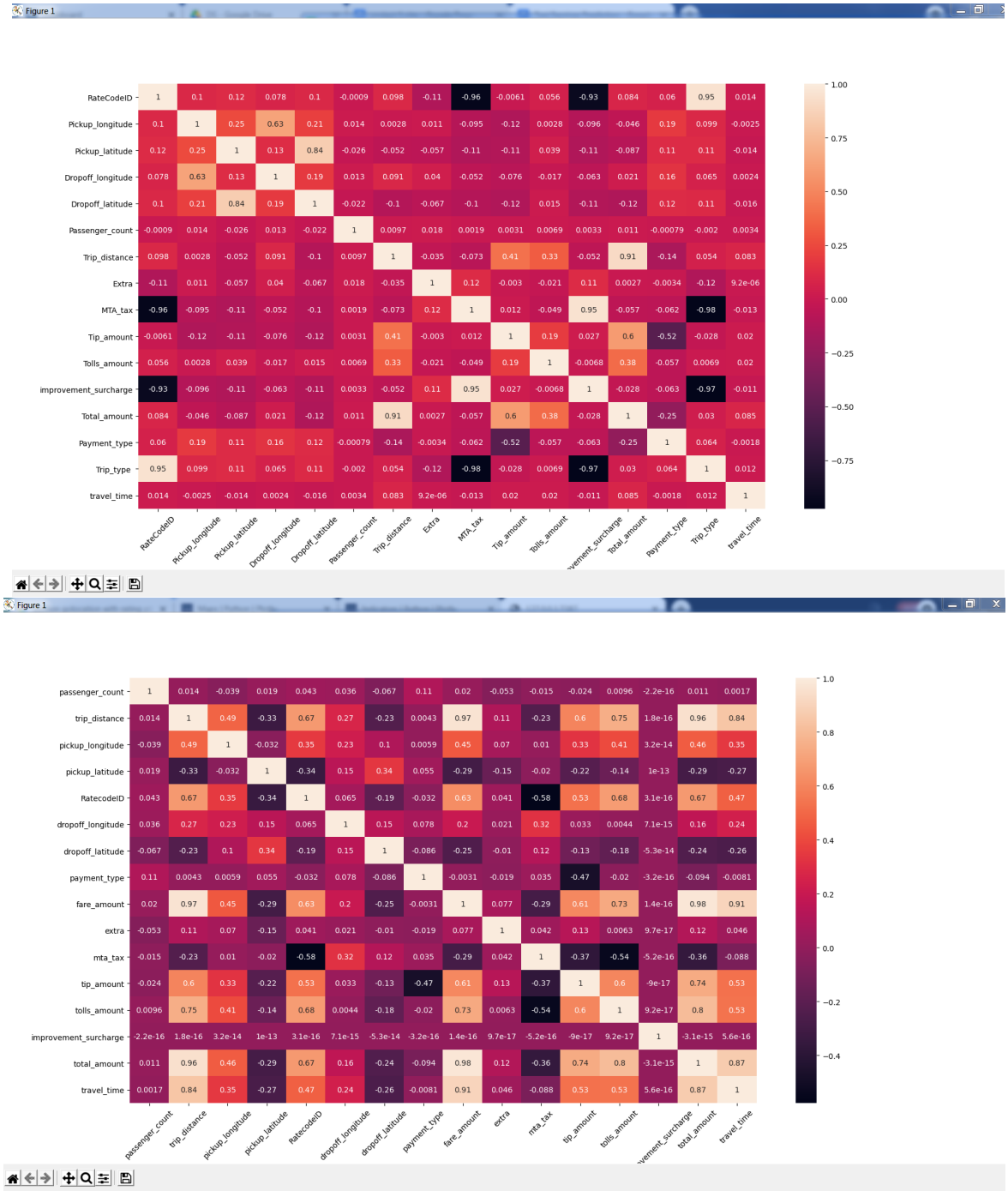

Figure 2: heatmap of green (up) vs yellow (down) feature correlation

3.2 Pairwise relationship

To add on to the information revealed in the heatmap, a pairwise plot is constructed to zoom in the scope, reflecting the stronger correlated features. **Kernel Density Estimation** is used to produce the plot for more direct visualization, and the distribution is aggregated by **payment method** to provide a contrast between credit card and cash.

The pairwise KDE distribution plots show strong comparison between the two licenses. Green taxis have passenger flow, consumption and time spent that are generally less than yellow taxis. Moreover,

it is clear that yellow cabs receive mostly credit cards as payment methods, while the green cabs receive more cash payment. This might be because people who live in the city center have different spending power than those that live in the suburbs, corresponding to the service region of the taxis. **From this insight, the extra services provided should start with yellow taxi passengers for their higher spending power.**

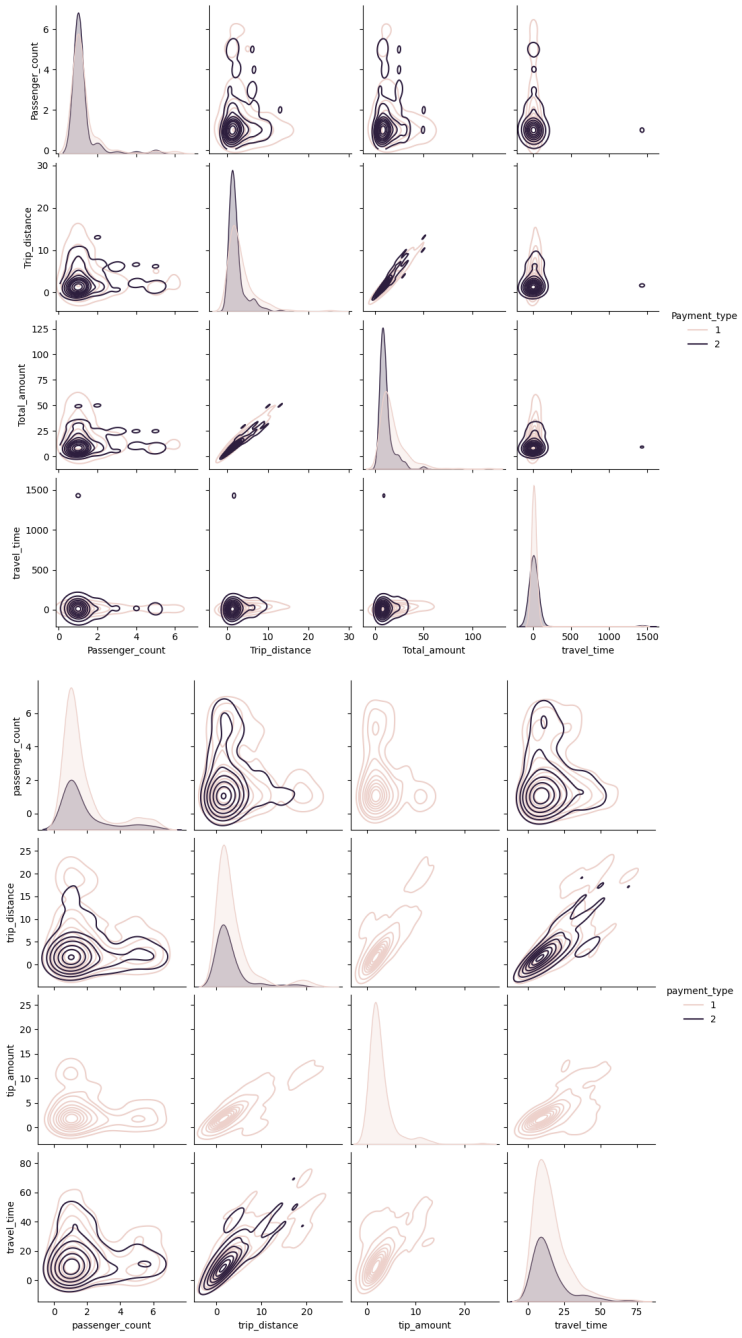

Figure 3: Pairwise relationship between well correlated features in Kernel Density Estimation, Green (up) vs Yellow (down)

3.3 Geolocation and Mapping

It is very important to see the travelling route of the records because this provides the regional economy background, and helps the service company determine the location of focus.

From the density graph below, the pick up map correctly shows the general service area of green taxis, which excludes Manhattan, the region of central city. However, the most popular spots are still around Manhattan along the coast, as well as the drop off area.

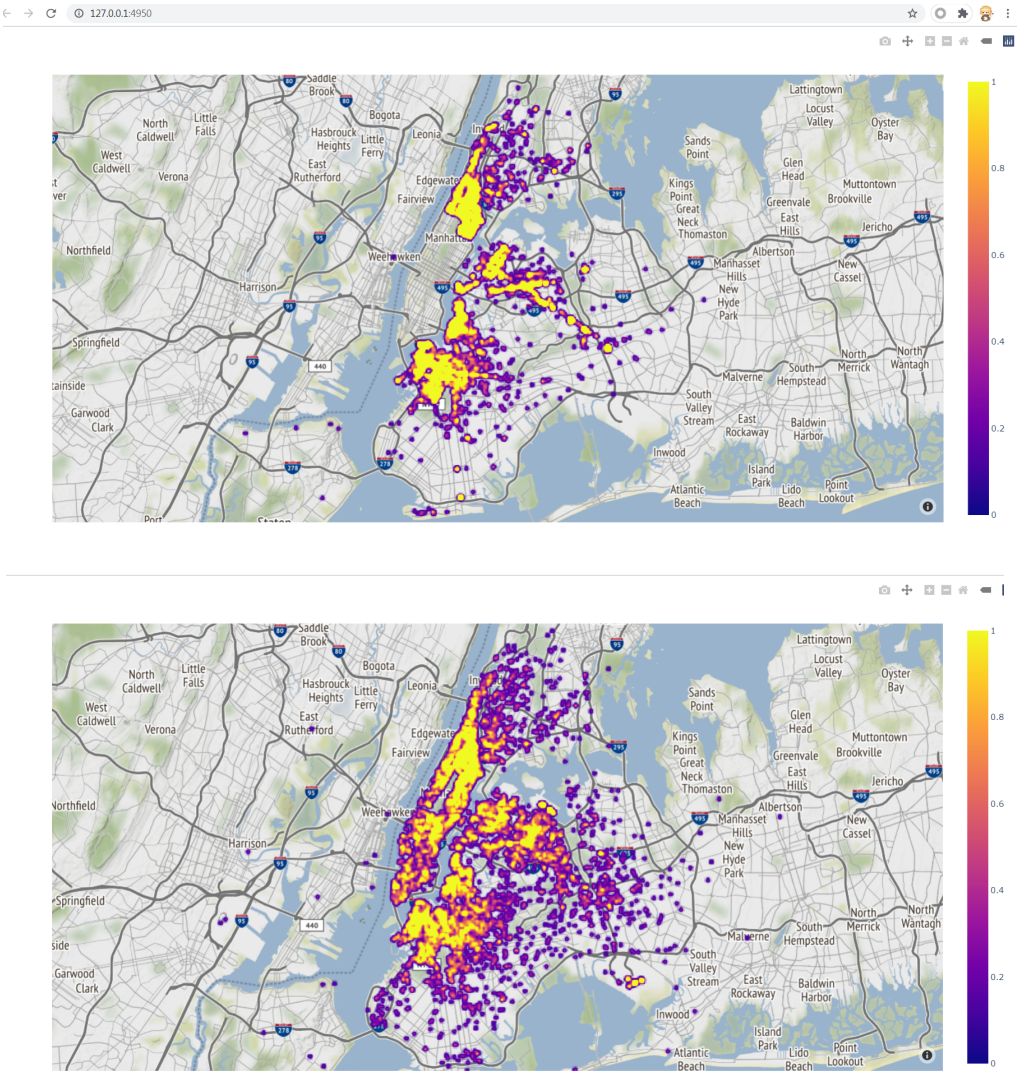

Figure 4: Green taxi pick up (up) vs drop off (down) location

On the other hand, yellow taxis have mostly all the pick up spots from the central city, and radiate to Brooklyn.

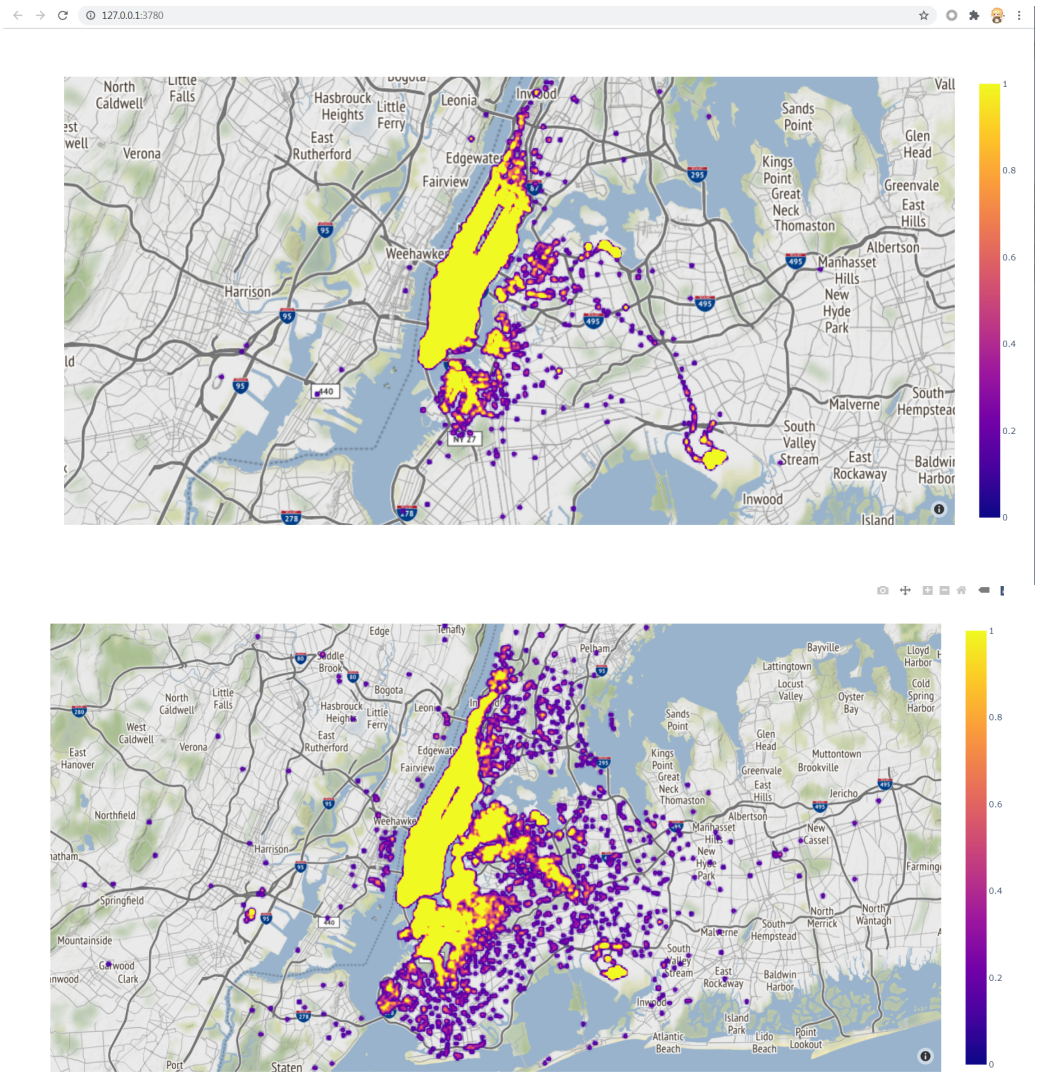

Figure 5: Yellow taxi pick up (up) vs drop off (down) location

From the information gained above, the most active region regarding transportation should be within Manhattan and Brooklyn, and the area along the coast. **Thus, when attempting extra services, these regions should have several pins during the beta test**.

3.4 Additional Information

When researching the destination of the trip records, it turns out that many of the drop off locations are matched with the building location of hotels.

This is an interesting discovery that might reflect some customer behavior. The heatmap below shows the rate of hotels is well correlated with the location and star rating.

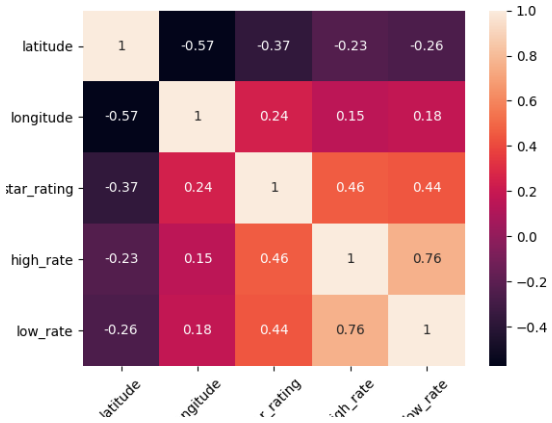

Figure 6: Heatmap of external dataset (hotel information) showing engineered feature correlation

From the information gained above, considering the project is aiming for people that are willing to spend extra money, and learn that the higher star rating a hotel has, the higher the rate (as shown in figure left below), which are located in Manhattan near the central park(right in the figure below). Despite the number of five star rating hotels being small (which make sense because rich people are not the majority), their locations are more concentrated within the yellow taxi serving area, meaning the yellow taxi passengers are still in the spotlight. The service company could take cooperation with the five star hotels into account. For example, a special service could be tourists take the yellow taxi from the airport straight to the cooperated hotels.

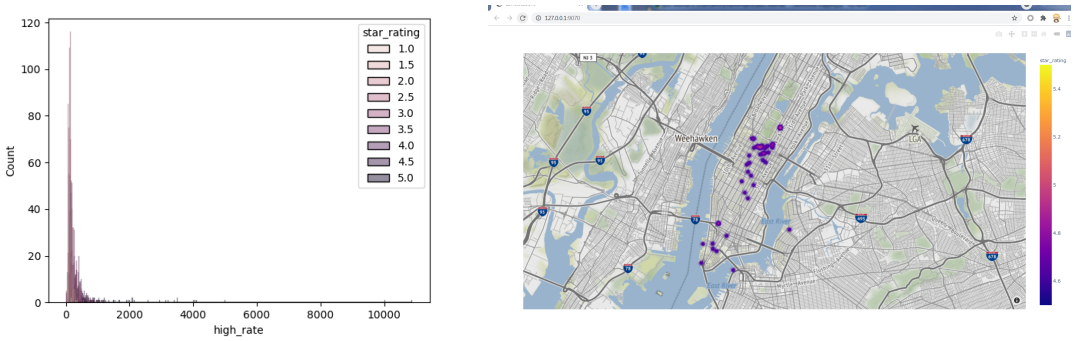

Figure 7: hotel rate respect to star rating (left) and Location of five star hotels (right)

4 Prediction and analysis

In the previous steps, the project has explored the feature correlation, and explained the target people and main region of attempting taxi extra service. It is important to discover the amount of money that passengers are possibly willing to spend for this launch for determining the service cooperation and content. Therefore, predicting the tip amount that **yellow taxi passengers** are willing to pay under a regular economic background is useful. In this prediction, **only** credit cards will be considered because it is the **main payment method** in the Manhattan area as discovered in the previous steps.

4.1 Model

The model used is the **Support Vector Machines**. It is effective in this project- the predicting outcome is the tip amount which is a continuous value, so there should be a regression model. It is a supervised machine learning model that works relatively memory sufficient. Moreover, since it has an advantage on a dataset that has less noise, which in this project the dataset is being cleaned, it should be a satisfactory choice.

4.2 Feature Selection

To obtain a precise result, the **least absolute shrinkage and selection operator** is used to regulate and select the variables. It also helps the accuracy and interpretability of the SVM model. The plot below shows that the trip distance and variables related to charges are well correlated with the amount of tip paid. Therefore, when performing prediction, these are variables that need to be taken into accountability.

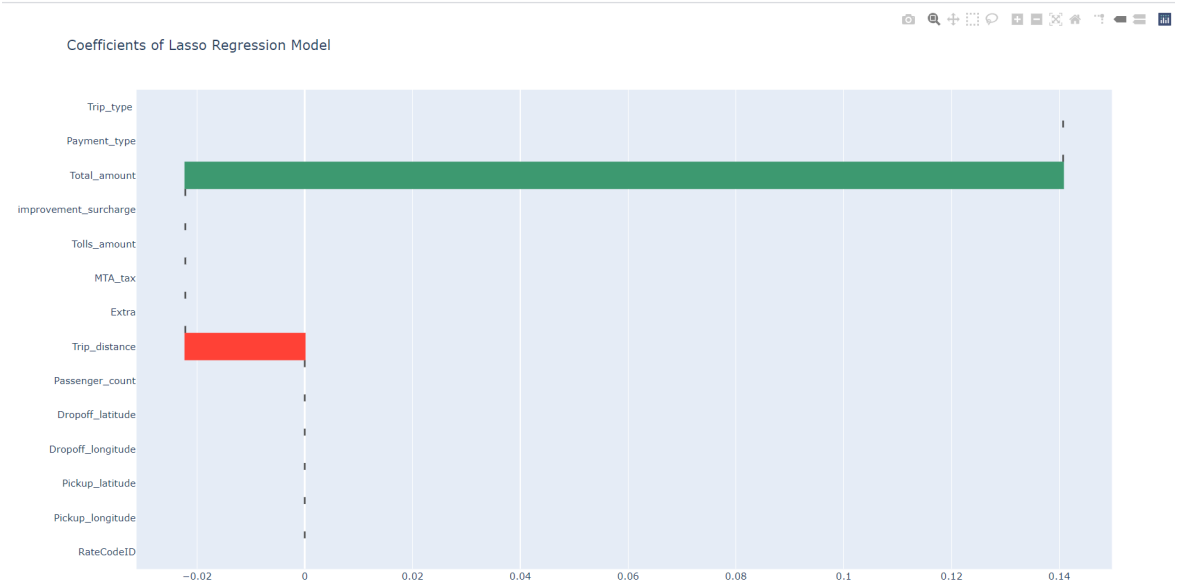

Figure 8: coefficients of lasso feature selection regarding tip amount

4.3 Model Performance and Error Analysis

The result of the prediction is satisfactory. The **coefficient of determination** is about 0.77, meaning the predictive values are well fitted with the actual value. The mean absolute and **mean squared error** are both very low. Thus, the overall model performance is outstanding.

```
R square: 0.7692804117272423
Mean Absolute error: 0.527288745528214
Mean Squared Error:: 1.5309385197254284
```

Figure 9: statistic of model performance

The plot below shows the fitted and actual value of tip amount that yellow taxi passengers might pay

in future. The shape of the prediction is generally normal distributed with a little right skewness. This makes sense because the people that are willing to spend more money will not be the majority in the statistics, but are the focus of the project to provide extra services.

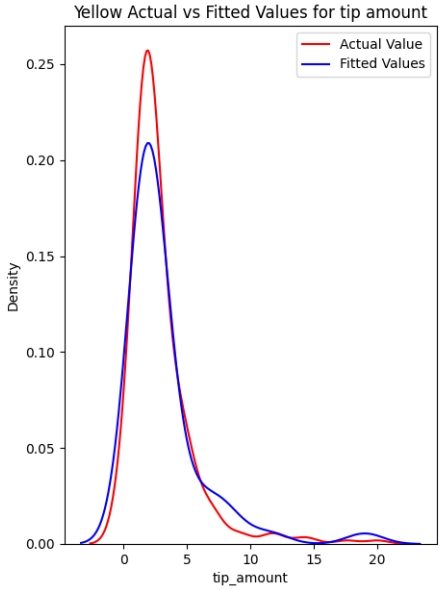

Figure 10: Yellow taxi actual vs fitted value over tip amount prediction

5 Discussion, Recommendation and Conclusion

It is reasonable to advise that in New York City, if the taxi company wants to plan extra services for more revenue and reputation, start with the passengers taking yellow taxis in the Manhattan area who pay tips more than 10 dollars. There are several considerations.

First of all, when comparing the statistics between green and yellow cabs, the yellow taxis show a surpass in passenger flow, money consumption and travel time over green taxis, which contains capability for extra service. Secondly, the most active and busy region is within central Manhattan, and has almost all the five star rating hotels which attracts the higher class people or tourists- and this is the yellow taxi serving area. Last but not least, as a new plan, the taxi service company should minimise the cost in the testing stage, so it is essential to target the consumers.
There are some limitations in this project as well. For example, the assumption about tourists made up most of the passengers is not rigorous, but this will need further research about the population that is made up of the taxi passengers..

In conclusion, this project is the initial analysis for the extra service plan, and has suggested about getting started on the studied field, but further determination will need more specific analysis.

(word count: 1794)

References

[1] Pickhardt, Stephen. "How to Get a Taxi in NYC." Free Tours by Foot, 11 June 2021, freetoursbyfoot.com/how-to-get-a-taxi-in-nyc/

[2] "TLC Trip Record Data - TLC." Nyc.Gov, 2021, www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[3] "New York Hotels." Kaggle, 16 Nov. 2017, www.kaggle.com/gdberrio/new-york-hotels.