A STUDY OF THE INCOME OF YELLOW TAXI DRIVERS

MAST30034 PROJECT1

Zixuan XIAO
Student ID: 1132915
Github repo with commit

August 20, 2022

1 Introduction

In New York, the yellow cab has become part of American pop culture, a symbol of the city along with the Empire State Building, the Statue of Liberty, and Times Square. The cab drivers who have shaped all of this have been a lonely presence in the midst of New York's bustle. The current use of yellow cab data to study the income patterns of drivers and make recommendations to yellow cab drivers can help improve the income of yellow cab drivers.

The process of this study consisted of three steps: data pre-processing, visualization, and modeling. The data was collected using TLC trip log data published by the New York Taxi and Limousine Commission (TLC)[1], which includes information such as pickup date/time, pickup location, trip distance, fare details, rate type, payment type, and number of passengers reported by the driver. The data used in the additional dataset is collected by technology providers authorized by the Taxi and Limousine Passenger Enhancement Program (TPEP/LPEP) and provided to the New York City Taxi and Limousine Commission (TLC). Trip data is not created by the TLC and the TLC makes no representations as to the accuracy of such data.

In order to reflect the current daily income of yellow cabs more realistically, the data of January, March, May, August, October, and December, a total of six months of yellow cab trip records in 2021, were selected and all data were stored in PARQUET format[2]. For the data pre-processing process, we performed the following operations: remove missing values, remove outliers, remove data fields that are not useful for this study, and also remove outliers using IQR.

Also this time, we studied the revenue of yellow cabs, where the driver's revenue is determined by the demand and the charge for each service, which is assumed to be influenced by the fare and tip paid by the passenger. This time, we additionally used the weather data for New York City in 2021 released by the National Oceanic and Atmospheric Administration, which includes: rainfall, snow thickness, maximum temperature, and minimum temperature. We analyzed the impact of these external weather (temperature, snow, rain) on yellow cab service hours while visualizing them. In addition we also analyzed the impact of weekends and large events on yellow cab service hours. Finally, a linear prediction model was designed to predict the revenue of yellow cabs.

2 Preprocessing

2.1 Preprocessing

This time, the acquired TLC trip record data should first be pre-processed with data, and the process mainly includes the following steps.

The first step was to remove the missing values, for the data that did have more data, we directly removed the data in the corresponding row, and then removed the outliers, which included: data other than January, March, May, August, October, and December of 2021, data with less than 1 greater than 5 passengers, data with average speed less than 1 MPH and greater than 100 PM, data with fare amount less than 2.50 US dollar, and data with payment type not by cash or credit card. The data was then deleted from the fields that were not useful for this study, and the final data size was changed from 15268660 rows × 17 columns to 9591844 rows × 15 columns.

With the cab fare calculation method published by TLC [1], trips between Manhattan and John F. Kennedy Airport (JFK) in any direction would be charged from 52 US dollar. Therefore, the cost of these trips would be significantly higher than the standard, and it was necessary to classify and discuss them to avoid a large amount of useful data being defined as outliers. Therefore, the data was divided into two parts by the features "RateCodeID" 1 and 2. In this case, outliers are handled using the interquartile range detection (IQR) method, which uses the interquartile range (IQR) of the box line diagram to detect outliers. It provides a criterion for identifying outliers: outliers are usually defined as less than Q1-kIQR or Q3+kIQR.

Q1: the lower quartile, indicating that one quarter of all observations take a smaller value than it.

Q3: the upper quartile, indicating that one-fourth of all observations take values larger than it.

IQR: Interquartile spacing, which is the difference between the upper quartile Q1 and the lower quartile Q3, during which half of all observations are included.

The threshold value for normal data is k times the interquartile range, and the data outside the interquartile range is considered as abnormal data, and k=3 is set in this case. Finally, after IQR processing, the data changed from 9591844 rows × 15 columns to 6960508 rows × 15 columns. the box plots before and after processing using IQR are shown in Fig. 1 and Fig. 2 below.

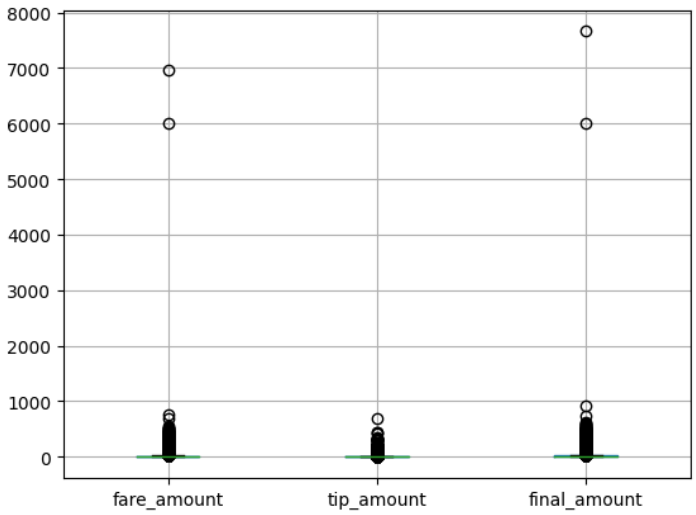

Figure 1: Box plot of IQR before removing outliers

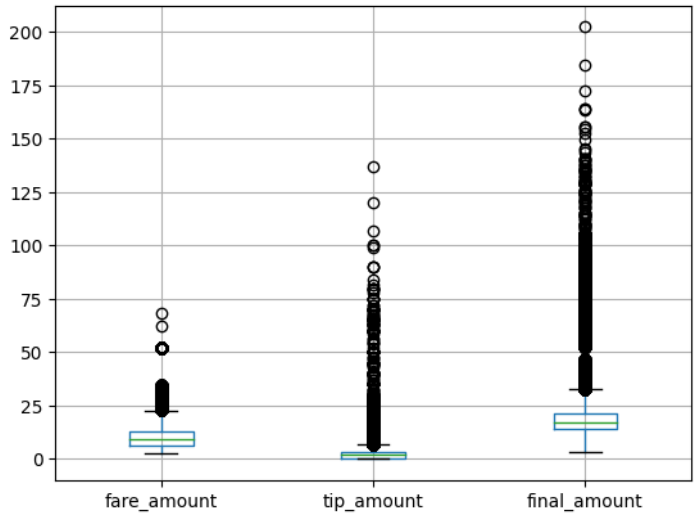

Figure 2: Box plot of IQR before removing outliers

2.2 Weather dataset pre-processing

The weather dataset was also pre-processed, in which the missing values were filled with the average values, and the data were selected for January, March, May, August, October and December of 2021, for a total of 6 months. The data size was changed from 365 rows × 8 columns to 186 rows × 8 columns, and the weather data and cab trip data were merged and saved as taxi_add_weather_21.weather for later visualization and modeling.

3 Analysis and Geospatial Visualisation

3.1 Workday and weekend visualization analysis

The results of this visualization of the service hours of cabs on weekdays and weekends are shown in Figure 3 below.

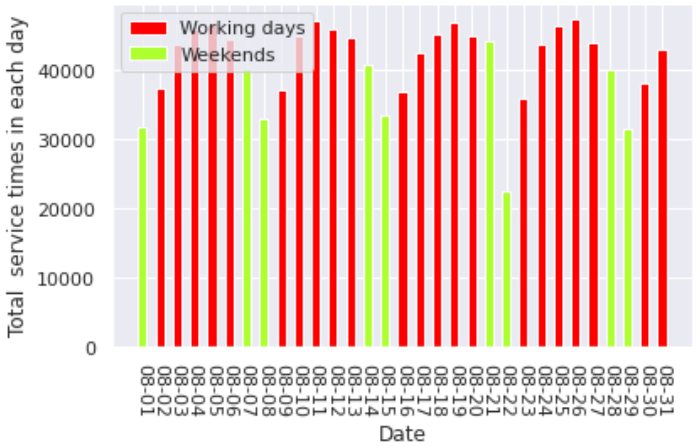

Figure 3: Visualization of service hours on weekdays and weekends

Through Figure 3 we can observe that the yellow bars indicate weekends and the red bars indicate

weekdays. It is obvious to find that the service time of cabs is longer during weekdays, so people prefer to take a taxi to work during weekdays, and during weekends, people rest at home and fewer people go to work, so the service time of yellow cabs becomes shorter.

3.2 Weather Visualization Analysis

3.2.1 Raining

The weather of August 2021 was selected for this visualization, where the service time of rainfall on yellow cabs is shown in Figure 4 below.

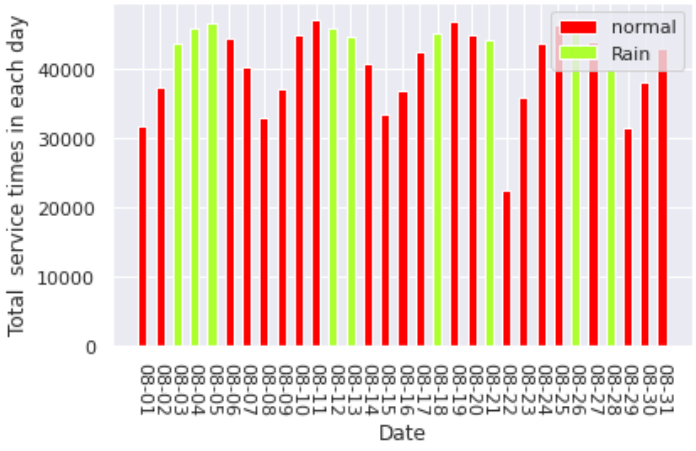

Figure 4: Effect of rain on service hours of yellow cabs

Through Figure 4 we can observe that red indicates normal weather and yellow indicates rainy weather. When it rains, the service time of cabs will increase and passengers are more inclined to travel by cab when it rains, so the rainfall will increase the demand for yellow cabs.

3.2.2 High temperature

The weather in New York in August 2021 was selected, where the red bars indicate normal weather and the yellow bars indicate hot weather (outdoor temperature greater than 85 degrees Celsius).

Through Figure 5 we can observe that high temperature has no significant effect on the service hours of cabs, so we can exclude the demand for yellow cabs in hot weather.

3.2.3 Cold

The weather in New York in January 2021 was selected to represent winter in New York at this moment, where the red bars indicate normal weather and the yellow bars indicate freezing weather (outdoor temperature less than 10 degrees Celsius)

Through Figure 6 we can observe that when the outdoor temperature is less than 10 degrees Celsius, the service hours of yellow cabs are greater and people's demand for cabs is greater, so we can infer that when the outdoor temperature is too low, it will increase people's demand for yellow cabs.

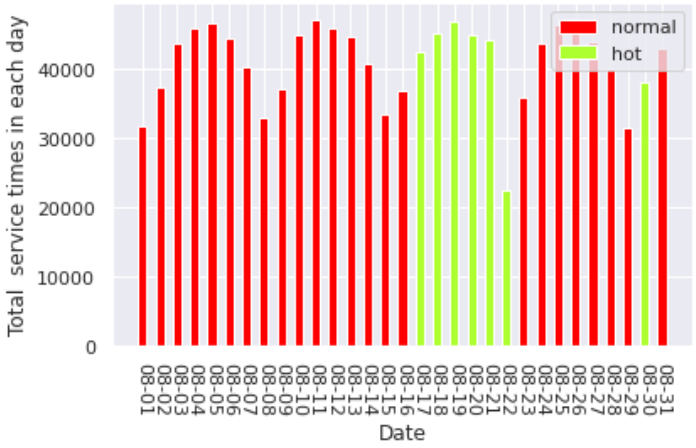

Figure 5: Effect of high temperature on the service time of yellow cabs

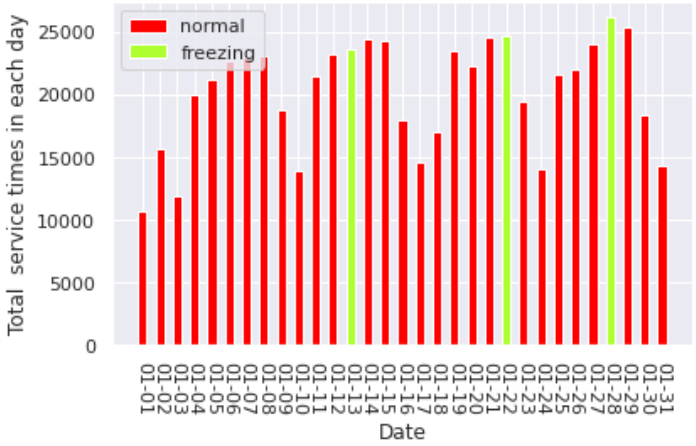

Figure 6: Effect of cold on service hours of yellow cabs

3.2.4 Snowing

This time, the weather of New York in January 2021 was selected to represent winter in New York at this moment, where the red bars indicate normal weather and the yellow bars indicate snowy weather.

Through Figure 7 we can find that when it snows, the service time of cabs will become shorter, the reason is because snow will bring inconvenience to traffic and easily trigger traffic accidents, so when it snows in winter, it will reduce people's demand for yellow cabs

3.3 Visual analysis at different moments of the day

This visualization of the number of cab services count at each moment of the 24-hour day is shown in Figure 8 and Figure 9 below, respectively.

Through Figure 8 we can observe that when the number of yellow cab service is the lowest every day starting from 5:00 am, when there are no pedestrians on the road, the number of yellow cab service gradually increases with time, and when to 19:00 pm, the number of yellow cab service reaches the highest, and then the number of cab service gradually decreases due to late night.

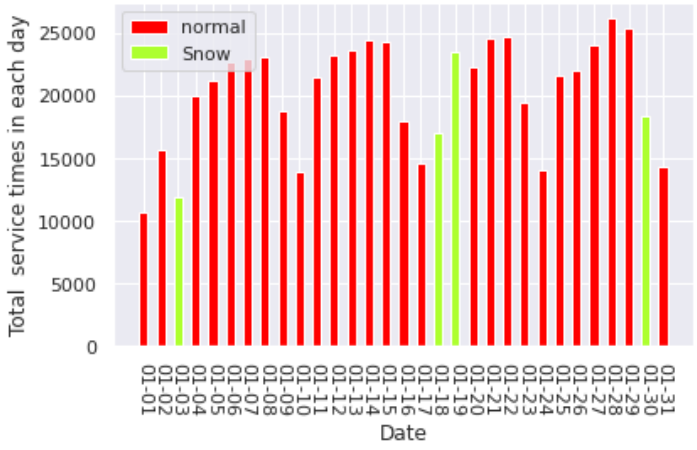

Figure 7: Effect of snow on service hours of yellow cabs

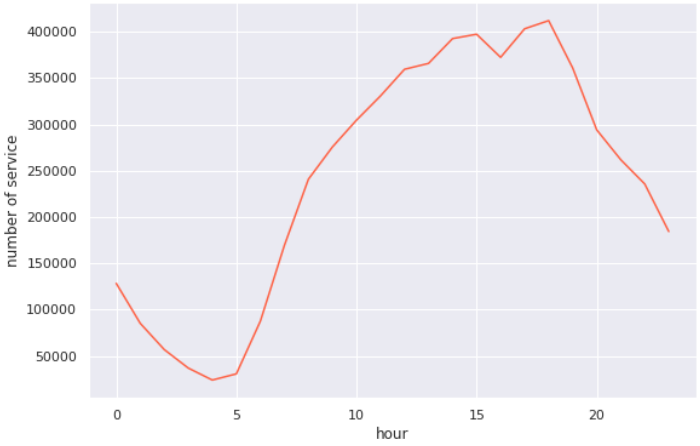

Figure 8: Effect of snow on service hours of yellow cabs

3.4 Geospatial Visualisation

The Charlie Parker Jazz Festival was held in Marcus Garvey Park on August 26th and August 27th, 2021, and the coordinates of Marcus Garvey Park are (40.79162297374497 , -73.92685780906619)[4]. The black area in Figure 9 below circles Marcus Garvey Park, the darker the red means the greater the demand for yellow cabs, as can be observed in Figure 9, the demand in the darker area is significantly higher, so cab drivers are advised to go to the darker area in the figure.

4 Statistical Modelling

4.1 Logistic regression (LR) model

The logistic regression (LR) model is formulated as follows.

$$f\left(x_i, w, b\right) = w \cdot x_i + b \tag{1}$$

In Equation (1), w denotes the weight vector and b denotes the deviation vector. The purpose of this

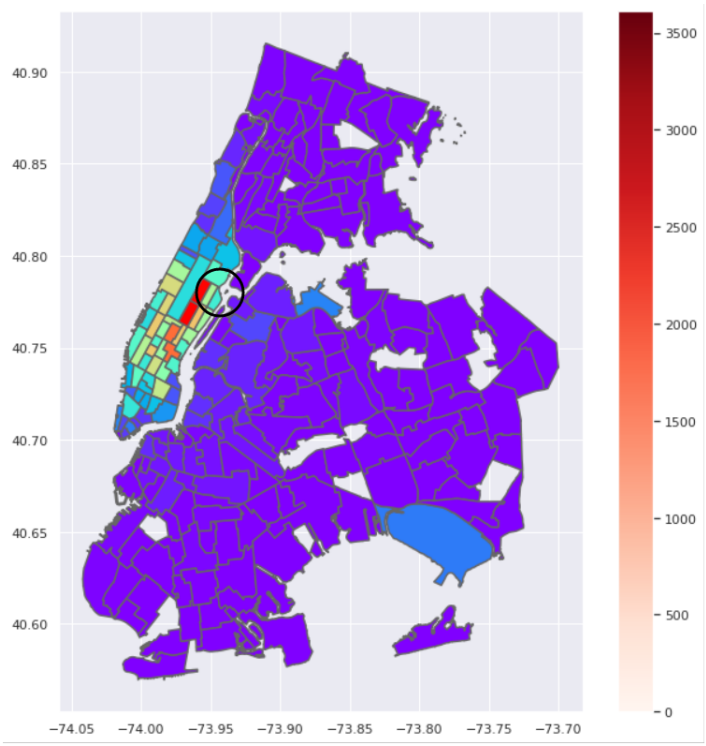

Figure 9: Impact of Charlie Parker Jazz Festival on the number of yellow cab services

model is to predict the consumption that the yellow cab driver may receive. The model was created directly using sklearn.linear model import LinearRegression, and the data set was split into a training set and a test set at the same time. The training set and data are used for training first, and the test set is used for testing. The model was evaluated using sklearn.metrics. The relationship between the true value and the test value is shown in Figure 10 below.

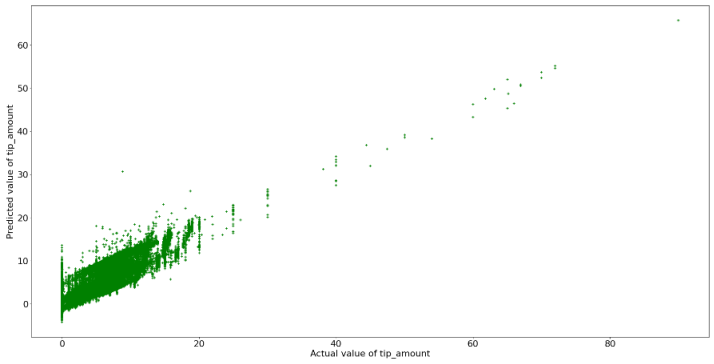

Figure 10: Relationship between true and predicted values

4.2 SVM regression model

The SVM regression model is formulated as follows:

$$\hat{y}_i = w^T x_i + b$$
$$E_\epsilon\left(\hat{y}_i - y_i\right) = \begin{cases} 0, & \text{if } |\hat{y}_i - y_i| < \epsilon; \\ |\hat{y}_i - y_i| - \epsilon, & \text{otherwise} \end{cases} \tag{2}$$

The prediction results were obtained by fitting the test set to the model. By using "r2s core" in sklearn.metrics, the final test accuracy of the logistic regression (LR) model reached 86.6 and the test accuracy of the SVM regression model reached 75.3. The experiments show that the logistic regression (LR) model has the best performance. This also shows the reliability of this model design.

5 Recommendations

The recommendations from this study of yellow cabs are as follows.

cab drivers who work 8 hours a day are recommended to work the following time slots: 8:00 to 12:00, 13:00 to 15:00, and 17:00 to 19:00, because these 8 hours have the highest number of cab services and also the highest revenue.

It is recommended that cab drivers be able to continue working on rainy days, as it is often the case that passengers are more inclined to take cabs and therefore get more revenue.

Cab drivers are also advised to work more hours on Mondays and Fridays, when there is a higher demand for cabs.

6 Conclusion

This study analyzed the trip data of yellow cabs in 2021 provided by TLC and visualized and statically modeled the data. The impact on tips was also analyzed using an external dataset, and a linear logistic regression prediction model was designed with high prediction accuracy.

References

[1] "Www1.nyc.gov. 2022. Taxi Fare - TLC. [online] Available at: ¡https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page¿ [Accessed 15 August 2022].

[2] Www1.nyc.gov. 2022. TLC Trip Record Data - TLC. [online] Available at: ¡https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page¿ [Accessed 15 August 2022].

[3] (NCEI), N., 2022. Daily Summaries Station Details: NY CITY CENTRAL PARK, NY US, GHCND:USW00094728 — Climate Data Online (CDO) — National Climatic Data Center (NCDC). [online] Ncdc.noaa.gov. Available at: ¡https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail¿ [Accessed 15 August 2022].

[4] google.com/maps. 2022. GPS coordinates address of Marcus Garvey Park, New York (NY) United States, Travel to Marcus Garvey Park - google.com/maps. [online] Available at: ¡https://www.google.com/maps/search/Marcus+Garvey+Park/@40.804746,-73.9465183,17z?hl¿ [Accessed 10 August 2021].