Analysis of Yellow Taxi Drivers' Hourly Wage in NYC based on Data Visualisation and Linear Model

Jiacheng Zhang
Student ID: 1056574

August 15, 2021

1 Introduction

The New York City Taxi and Limousine Commission (TLC) collects the trip records of the yellow taxis, green taxis, for-hire vehicles (FHV), and high volume for-hire vehicles (HVFHV) in recent years. Taxi service, as one of the most universal pick-up services in NYC, has more than 200,000 TLC licensees that can complete 1,000,000 trips each day in total [1]. However, under such high demand, the income levels of taxi drivers are still varying a lot. This report aims to analyse how the factors such as weather, pick-up locations, start hour, and trip distance (maybe more) affect the hourly wage and provide some suggestions to the drives on how to earn more money in a fixed time interval.

2 Data Selection

The data sets chosen were from June 2019 to August 2019. It is assumed that the data in 2020 has lost generality because of the impact brought by COVID-19, which indicated that 2019 was the latest year that can be chosen to perform analysis. Three consecutive months were chosen as they encompass the data of the same season, which is more representative if the weather is considered as a factor to analyse. The data sets selected contain 18 features and more than 19 million instances in total.

2.1 Yellow Trip Data

The yellow trip data was selected mainly because of two reasons. Firstly, they are the only vehicles licensed to pick up street-hailing passengers anywhere in NYC, and secondly, they have much more instances than the green taxi and the for-hire vehicles [2][3]. As a result, yellow taxi is regarded as the most representative type of licensed taxi in NYC to perform analysis.

2.2 Weather Data

The weather data chosen were from the same period as the yellow trip data. It contains more than 20 weather conditions to describe the weather in NYC [4]. To simplify the analysis, weather conditions were divided into two categories (good or bad) according to the description. Additionally, the data is clean, which means no extra preprocessing is required.

2.3 Additional Data

The TLC website published the taxi zone lookup table and taxi zone shape file which provided information that can be used for geospatial visualisation [3].

3 Preprocessing of Yellow Trip Data

3.1 Data Cleaning

- Removed any instances that have missing values (i.e. NA).
- Filtered the passenger counts to a minimum of 1 and a maximum of 6, as stated in the TLC website [5].
- Filtered the pick-up and drop-off date time to be strictly between June 2019 and August 2019, as the analysis required such a timeline.
- Filtered the MTA tax to \$0.5, as stated in the TLC website [6].
- Filtered the congestion surcharge to either \$0.0 or \$2.5, as stated in the TLC website [6]. \$0.75 was filtered because the sample size was too small which would not affect the overall data.
- Filtered the improvement surcharge to \$0.3, as specified in the data dictionary.
- Filtered the payment type to 1, as the total amount did not include the cash tips and other payment types were not relevant to analysis.
- Filtered the rate codes to 1, 2, and 3, as trips from the airports would also be included in the analysis.
- Filtered the location ID to a range from 1 to 263, as other IDs were not valid in the taxi zone lookup table.
- Filtered rate codes labelled as standard taxi fares to start at \$2.5, as stated in the TLC website [6].
- Filtered the trip distance to be non-negative, as negative trip distances were invalid.
- Removed other outliers of continuous attributes such as tip amount, fare amount, total amount, and trip distance by the interquartile range (IQR).

3.2 Feature Engineering

- Added a duration (mins) by subtracting the pick-up time from the drop-off time. Filtered the duration with a lower bound of 1 minute, and an upper bound of 54.43 minutes (decided by IQR) since extremely short/long trips are not representative which have no analysis power.
- Added an hourly wage (USD/hour) using (tip amount+fare amount)/duration*60, which is a measure of the hourly income of a driver. Filtered the hourly wage by IQR to prevent the impact of the extreme values.
- Extracted the start hour and the start date from the pick-up time, which were used to merge the weather data.
- Merged the weather data into the yellow trip data as a new attribute.

3.3 Attribute Selection

- Removed VendorID and Store_and_fwd_flag, as they are irrelevant to the task and cannot produce any useful information.
- Removed passenger_count since there is no extra charge for extra passengers [6], and hence has no analysis power in this task.
- Removed tpep_pickup_datetime and tpep_dropoff_datetime, as more useful attributes (start hour and start date) have been extracted.
- Removed congestion_surcharge, extra, mta_tax, improvement_surcharge, and tolls_amount, because they have almost no effect on a driver's income.
- Removed total_amount, as it can be directly calculated by other attributes such as tip_amount and fare_amount.
- Removed payment_type, as it is always consistent which cannot provide any useful information.
- Removed RatecodeID, as PUlocationID and DOlocationID can provide more detailed and useful information.

After the preprocessing, the dataset contains 9 attributes (trip_distance, PULocationID, DOLocationID, fare_amount, tip_amount, duration, start_hour, start_date, and hourly_wage) and more than 7.9 million instances.

4 Analysis and Visualisation

4.1 Preliminary Analysis

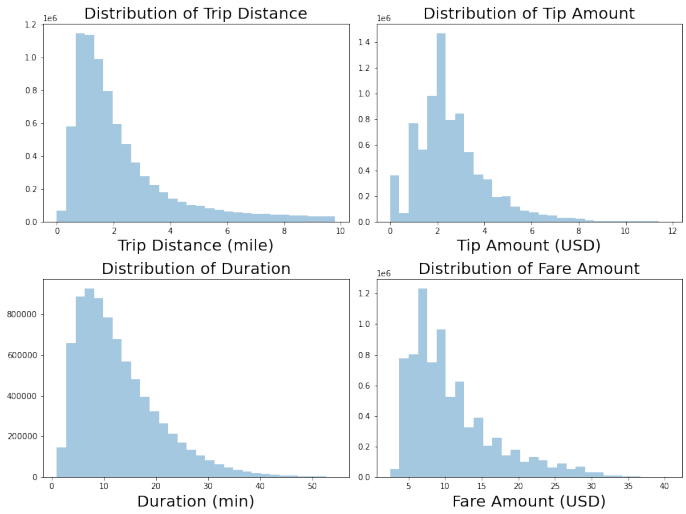

Figure 1: Distributions of Continuous Predictors

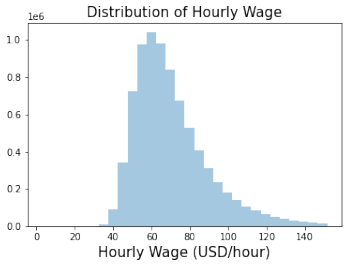

Figure 2: Distribution of Response Variable

Figure 1 illustrates the distributions of the continuous predictors and Figure 2 shows the distribution of the response variable after removing outliers. One noticeable point is that there are many 0s in the distribution of the tip amount, which is reasonable since some passengers may not be willing to give tips in the real life. In a linear model, it is assumed that all the variables will follow a normal distribution.

However, the distributions are now right-skewed which indicates that a log transformation can be applied.

4.2 Correlation Analysis

From Figure 3 shown below, it can be found that the fare amount is highly correlated to the trip distance and the duration. It is reasonable because the fare amount is measured based on the trip distance and the duration [6]. However, a linear model requires all the predictor variables to be independent of each other. Therefore, the fare amount will be removed in future analysis.

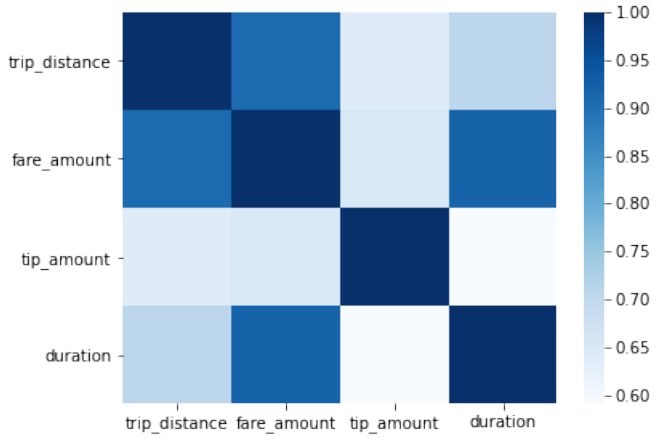

Figure 3: Pearson Correlation of the Continuous Predictors

4.3 Transformation

After a log transformation, it can be seen from Figure 4 that most of the sample quantiles fit the theoretical values. As a result, these attributes will be treated as normally distributed and all the remaining predictor variables will be considered as independent from each other.

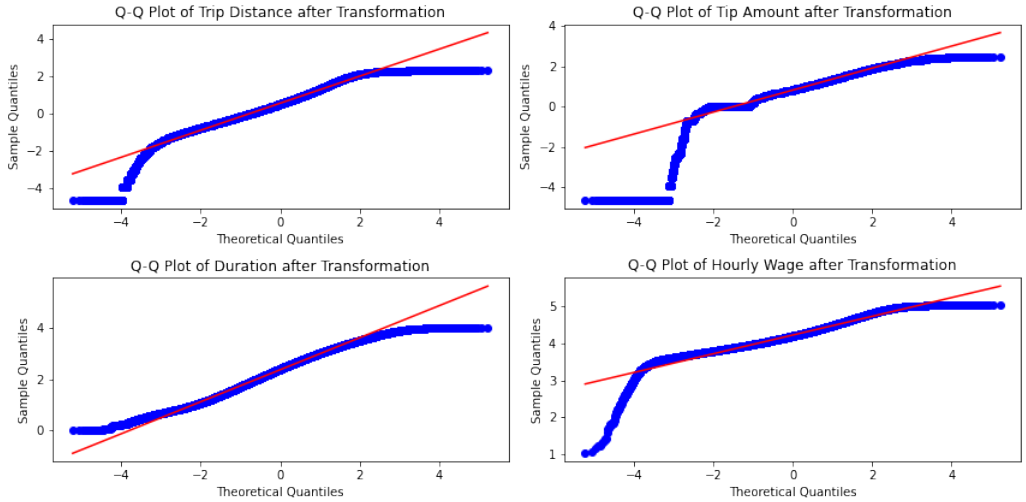

Figure 4: Q-Q Plots of Continuous Attributes after Transformation

4.4 Geospatial Visualisation

Figure 5 illustrates the average hourly wage of different zones in NYC and Figure 6 shows the average trip counts per day in NYC, both from June 2019 to August 2019 grouped by pick-up locations. It can be seen from Figure 5 that picking up passengers around the Broad Channel, including the JFK airport, had the highest hourly wage (over 80 USD/hour), followed by several zones of Staten Island (approximately 80 USD/hour). However, it is arbitrary to conclude that picking up passengers in these areas can earn more money. Because as shown in Figure 6, these areas had low taxi demand which indicated that drivers may have difficulties finding a passenger. Therefore, combining two figures, the LaGuardia airport and central Manhattan were the places with both high demand and considerably high profitability.

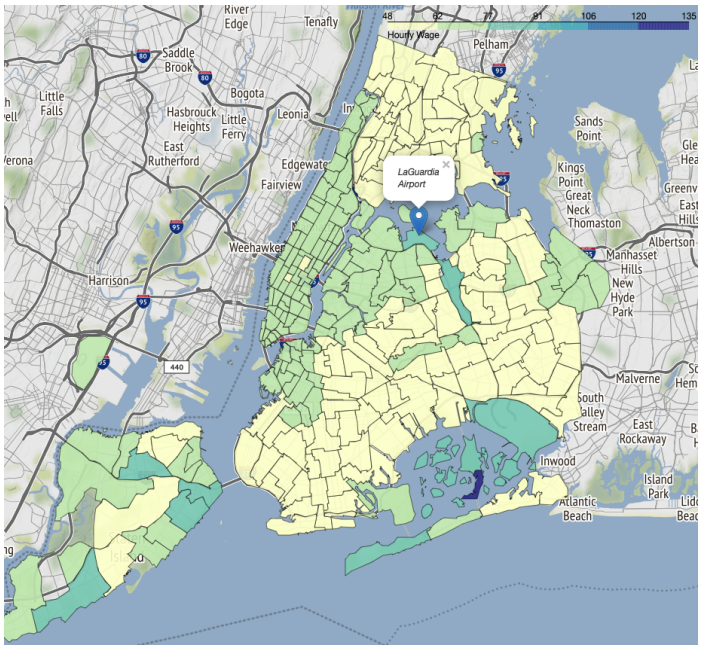

Figure 5: Average Hourly Wage for Pick-ups

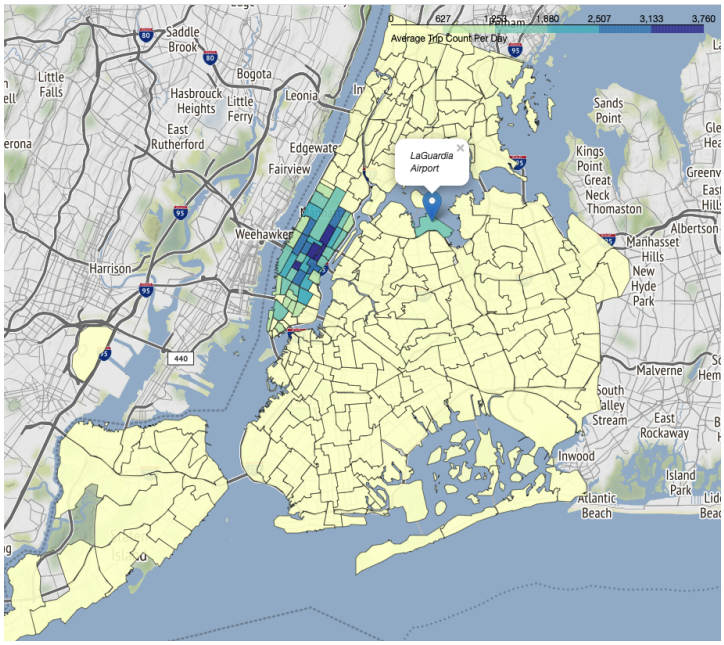

Figure 6: Average Trips per Day for Pick-ups

4.5 Exploration on Attributes of Interest

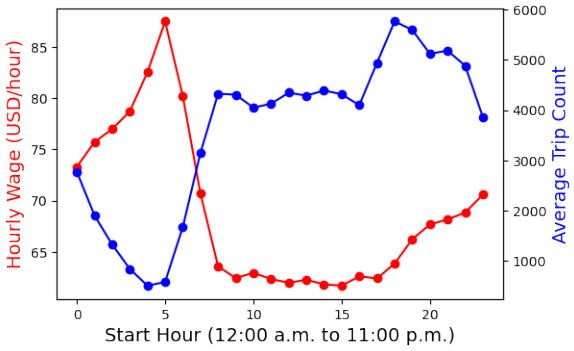

Figure 7: Comparison in Average Trip Counts and Average Hourly Wage by Time

Figure 7 compares the average trip counts and the average hourly wage in NYC at different periods

in a day. It is noticeable that the hourly wage achieved the climax while the average trip counts dropped to the trough at 5 a.m., which again supported that the insufficiency of trips at certain areas or certain periods might be the main reason that led to an over-estimated hourly wage. One possible explanation of this is that there might be no traffic congestion in the early morning so that the drivers could finish the trip shortly which resulted in a high hourly wage. However, it was not representative as the drivers might spend much time finding a passenger at that time. Therefore, It might be better to pick up passengers from 5 p.m. to 11 p.m. since the hourly wage was increasing during this period with a high taxi demand.

4.6 Discussion for future analysis

Previous sections have captured and discussed some useful information by exploring the relationships between the hourly wage and some attributes such as pick-up locations, start hours, and the trip frequency. Some conclusions can already be derived from these plots. As a result, to obtain a more comprehensive analysis, the modelling section will mainly focus on exploring how other attributes such as weather, trip distance, and interactions between them will affect the hourly wage.

5 Statistical Modelling

5.1 Model Choice

Linear statistical model is chosen to perform general interactions between attributes for two reasons. Firstly, the remaining attributes are less likely to have complex relationships by common sense, and secondly, the linear model has good interpretability and low time complexity.

5.2 Attribute Choice

Trip distance, weather, start hour, and an interaction between the first two attributes were chosen as the predictor variables, where the response variable was the hourly wage. Tip amount and the duration were removed as they will lead to an over-estimated performance.

5.3 Model Assumption

1. Response and predictors have a linear and additive relationship
2. Predictors are normally distributed and independent from each other
3. Errors of the model is normally distributed and independent from the predictors

Distributions and independence have been checked in previous sections. Therefore, the normality of the errors and the heteroskedasticity will be checked through Figure 8:

- The 'Residuals vs Fitted' plot has a mean of 0 and a constant variance which is expected to see.
- Most of the standardized values are close to the theoretical quantiles in the 'Normal Q-Q' plot which is also expected to see.
- There is no increasing/decreasing trend in the size of the residuals in the 'Scale-Location' plot which means the model has an equal variance.
- The 'Residuals vs Leverage' plot indicated that no point has a large distorting effect on the model.

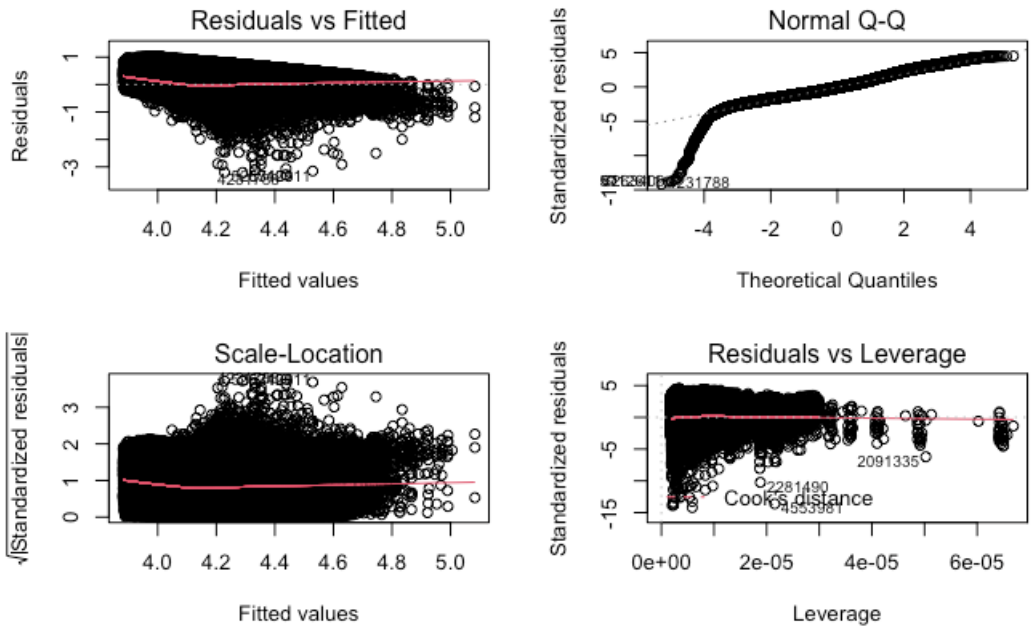

Figure 8: Diagnostic Plots of the Model

5.4 Feature Selection

Stepwise selection is applied for feature selection using Bayesian information criterion (BIC) as the goodness-of-fit measure with the following equation:

$$\begin{aligned} BIC &= -2\ln(likelihood) + p\ln n \\ &= n\ln(\frac{SSres}{n}) + p\ln n + const \end{aligned}$$

BIC is chosen because the penalty term includes the sample size which highlighted that the penalty term will not be ignored as the sample size increases. Hence, it is suitable for a large-scale dataset.

5.5 Model Performance

From Table 1, it can be seen that the mean squared error (MSE) of the model is approximately equal to 0, which means the data fits well using the linear model. However, the R square of the linear model is only 0.1741, which indicated that although the model captured some information between the response variable and the predictor variables, the correlations were at a low level. Furthermore, according to the result of Figure 9, the weather has almost no interaction with the trip distance. Such a result is reasonable because some continuous features were removed at the early stage due to the violation of model assumptions, and some other attributes such as location IDs were excluded from the model since they may lead to an over-complicated analysis. As a result, it might be difficult to discover any significant findings.

Model	R Square	Mean Squared Error
Linear Model	0.1741	0.05359

Table 1: Performance of the Linear Model

```
Analysis of Variance Table

Model 1: hourly_wage ~ 1
Model 2: hourly_wage ~ trip_distance + start_hour + weather
Model 3: hourly_wage ~ trip_distance + start_hour + +weather + trip_distance *
    weather
   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1 7947771 515689
2 7947746 425998 25     89691 66946.5 < 2.2e-16 ***
3 7947745 425918  1        80  1495.9 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 9: Analysis of Variance on Interaction

6 Recommendations and Discussion

All of the coefficients of the predictor variables in the linear model are approaching 0, which indicated that factors such as weather and trip distance have almost no effect on the hourly wage. Therefore, yellow taxi drivers can pick up passengers without considering the weather and the trip distance, which means long trip distances would not decrease the income of the driver. Furthermore, as shown by Figures 5-7, the yellow taxi drivers are recommended to pick up passengers around the LaGuardia airport and central Manhattan from 5 p.m. to 11 p.m., since these areas and periods have considerably high hourly wages and high taxi demand.

7 Limitations and Future Work

Due to the limited processing ability of the local computer, and the limited memory of the server, only three months of the data are chosen. Sampling trip records from a longer period of time may have a better analysis power. Another improvement can be made by choosing more statistical models as baselines to compare with the linear model and choosing more metrics to obtain a more comprehensive evaluation.

References

[1] The New York City Taxi and Limousine Commission 2021, *About TLC*, accessed 9th August 2021, https://www1.nyc.gov/site/tlc/about/about-tlc.page.

[2] The New York City Taxi and Limousine Commission 2021, *Your Ride*, accessed 9th August 2021, https://www1.nyc.gov/site/tlc/passengers/your-ride.page.

[3] The New York City Taxi and Limousine Commission 2021, *TLC Trip Record Data*, accessed 9th August 2021, https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[4] Weather Underground 2019, *New York City, NY Weather History*, accessed 5th August 2021, https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA

[5] The New York City Taxi and Limousine Commission 2021, *Passenger Frequently Asked Questions - How Many People Can Fit Into A Yellow Taxicab?*, accessed 4th August 2021, https://www1.nyc.gov/site/tlc/passengers/passenger-frequently-asked-questions.page.

[6] The New York City Taxi and Limousine Commission 2021, *Taxi Fare - Standard Metered Fare*, accessed 4th August 2021, https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.