MAST30034 Project 1

Kennedy Ker Huan Guok
Student ID: 1039169

December 11, 2022

1 Introduction

Accompanied by the rise of the Internet and generalisation of smartphones, businesses have gone digital, and vehicle-hire is no exception. The existence of e-hail companies like Uber and Lyft have caused taxi drivers to lose customers. Because of some bad eggs among taxi drivers, they have a bad reputation for not using the meter, taking long routes or tricking visitors to pay more, which became a stereotype causing passengers to choose e-hail over taxis.

This study aims to help passengers understand what factors affect taxi fares and whether passengers are overcharged. E-hailing took over the vehicle-hire industry sometime around 2016. Since green taxis are allowed to accept dispatches like e-hailing services in addition to street hailing, we will look into factors that affect green taxi fares in this study. Hence, green taxi data from 2016 will be the main dataset for this study.

2 Data Selection

The dataset of this study is made up of 6 files. The data in these files are raw data which may not be complete and requires preprocessing and analysis.

2.1 New York City TLC Dataset

The raw taxi trips data are collected and downloaded from the NYC Taxi & Limousine Commission [1], which is available on the Internet. These data are the main data required in this study. Green taxi data of January, February and March 2016 is used. There are a total of 21 columns, these columns consist of features including date-time, pickup and dropoff coordinates, fare amount, trip distance. At the meantime, the rows are made up of rides, which consists of more than 4.5 million instances before preprocessing. Two other files of the NYC taxi zone are also required to visualise the zones across NYC.

2.2 External Dataset

External dataset is essential to find out whether other factors can affect the price of a green taxi ride. External data is any data that is not obtained from the TLC website.

2.2.1 New York City Traffic Collisions Dataset

A dataset of all police reported motor vehicle collisions in New York City is obtained from the NYC Open Data [2], which is publicly available. The data collected in this dataset ranges across January

2012 to mid of July 2021. It contains information such as date-time, type of vehicle, coordinates, number of people involved and etc. It contains more than 1.8 million traffic collisions data that needs to be cleaned and preprocessed.

3 Preprocessing

3.1 NYC TLC Dataset

The preprocessing step starts off by cleaning the NYC TLC Dataset. In this stage, the same steps are applied to both the training data and testing data. This stage is to remove any possible outliers and mistakes in the dataset. The following instances are removed:

- contain coordinates outside of NYC [3].
- have negative numbers in features involving trip distance, money paid and number of passengers.
- fare is paid but trip distance is zero.
- fare is paid but dropoff and pickup time is the same.
- fare per mile is above $50.00.
- fare per mile is below $0.10.

Instances mentioned above are probably caused by faulty taxi meters or mistakes made when recording the data. The average fare per mile is $4.24, hence any instance that differs too much and illogical is removed. A plot that shows the difference in instances before and after cleaning is shown in (*Figure 1*).

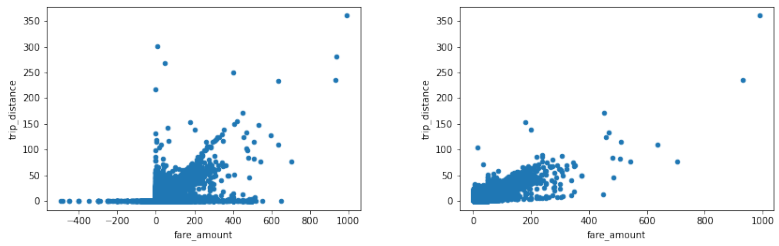


Figure 1: Scatter plot of **trip_distance** against **fare_amount** before and after cleaning instances

Apart from the conditions mentioned above, instances that did not adhere to the rules by the NYC TLC fare rates [4] are removed:

- fare per minute is less than $0.50.
- fare amount is less than $2.50.

In the process of calculating the fare per minute, a new feature: **time_elapsed** is engineered by using the dropoff and pickup date-time.
To reduce memory consumption, features that do not contribute to the study or are replaced with new features are removed:

- **extra**, miscellaneous and extra surcharges.

- **mta_tax**, tax triggered automatically.
- **store_and_fwd_flag**, indicates whether trip record was held in memory.
- **vendorid**, LPEP provider.
- **improvement_surcharge**, improvement surcharge assessed.
- **lpep_pickup_datetime**, pickup date-time.
- **lpep_dropoff_datetime**, dropoff date-time.

In order to merge the external dataset to the main dataset, new features are added:

- **p_weekday**, day of the week when the pickup happened.
- **d_weekday**, day of the week when the dropoff happened.
- **p_hour**, the hour when the pickup happened.
- **d_hour**, the hour when the dropoff happened.
- **pickupX**, pickup latitude in mercator format.
- **pickupY**, pickup longitude in mercator format.
- **dropoffX**, pickup latitude in mercator format.
- **dropoffY**, pickup longitude in mercator format.
- **pickup_zone**.
- **dropoff_zone**.

The **pickup_zone** and **dropoff_zone** are boroughs where pickups and dropoffs happens. There are seven different boroughs:

- **JFK**, JFK International Airport.
- **EWR**, Newark Liberty International Airport.
- **Manhattan**
- **Staten Island**
- **Queens**
- **Brooklyn**
- **Bronx**

The coordinates of the boundaries of the boroughs are modified from the shape files taken from the TLC website. However, JFK International Airport is not included in the shape files, hence the coordinates are taken from Google Maps [5]. The exact coordinate between the dropoff and pickup points is taken, then an approximate 10 metre boundary is calculated. Since the area is small, the curvature of the earth can be ignored.

3.2 NYC Traffic Collisions Dataset

In this external dataset, preprocessing starts off by removing irrelevant instances:

- Longitude or latitude is absent.
- Outside the date range of January to March 2016.

A few features are engineered in the process of preprocessing the traffic collisions dataset to allow merging the taxi dataset:

- **BOROUGH**, the area where the collision happened.
- **ENDTIME**, one hour after the collision happened, which is assumed that the traffic flow will be affected within this period.

It is observed that all collisions that happened during a ride did not involve any injuries or deaths. Hence, only the following features are required:

- **DATETIME**, date-time when the collision happened.
- **ENDTIME**, date-time when the effect of collision ended.
- **BOROUGH**, the area where the collision happened.

Using the above features, two new features are engineered in the main dataset:

- **pickup_affected_by_collision**
- **dropoff_affected_by_collision**

Whenever a pickup or dropoff is affected by a collision, the above feature will show **True**, and **False** otherwise.

4 Data Analysis

(*Figure 2*) shows the pickup and dropoff points in NYC respectively. It is obvious that pickups in Downtown Manhattan is lacking. This is because green taxis are only allowed to pickup in Upper Manhattan, Brooklyn, the Bronx, Staten Island and Queens [6]. Hence, green taxi pickups are concentrated in the upper part of Midtown Manhattan and Brooklyn. In contrast, the dropoffs concentrate all across Manhattan including Downtown Manhattan, but the hottest dropoff zone is still Upper Manhattan and Brooklyn.

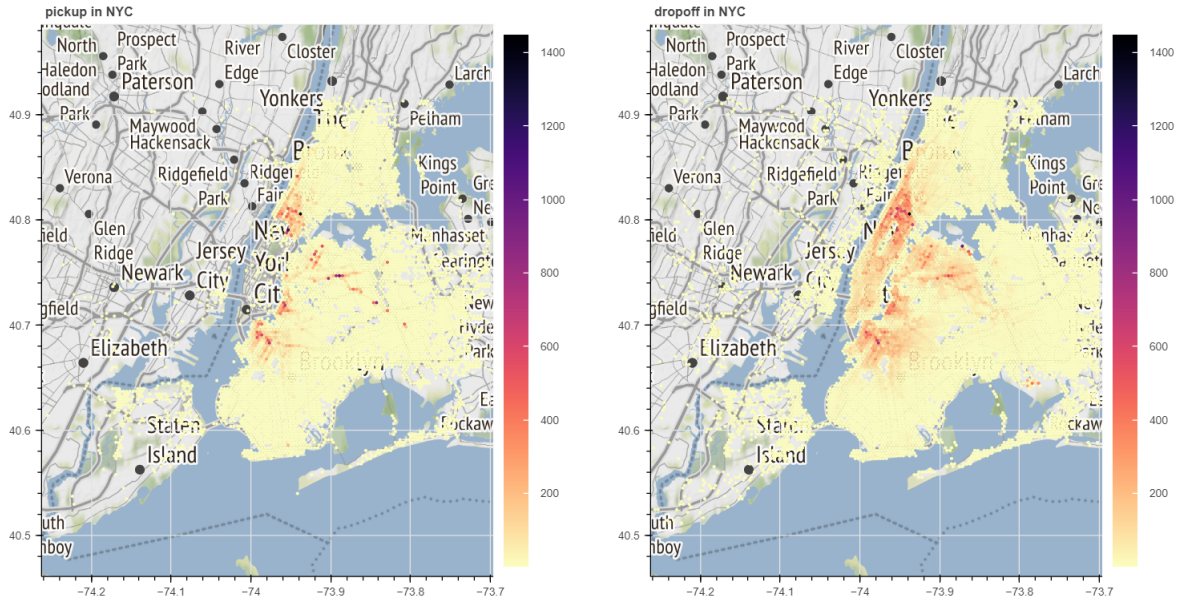


Figure 2: Map of Pickup and Dropoff Points

4.1 Main Attribute Analysis: Taxi Fare

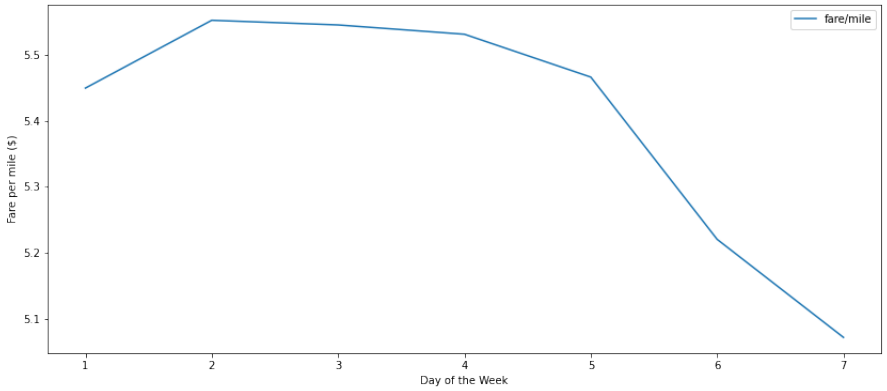


Figure 3: Graph of Fare per Mile against Day of the Week

(*Figure 3*) shows that the fare per mile on weekdays is at least \$5.40/mile, and is lower during the weekends around \$5.10/mile to \$5.20/mile. This is likely due to the \$4.50 rush hour surcharge [4] at 4pm to 8pm on weekdays. The increase in red and orange colours in (*Figure 4*) shows an increase in number of dropoffs on Sundays compared to Mondays, but the fare per mile on Sundays is still lower despite the high demand for taxi rides.

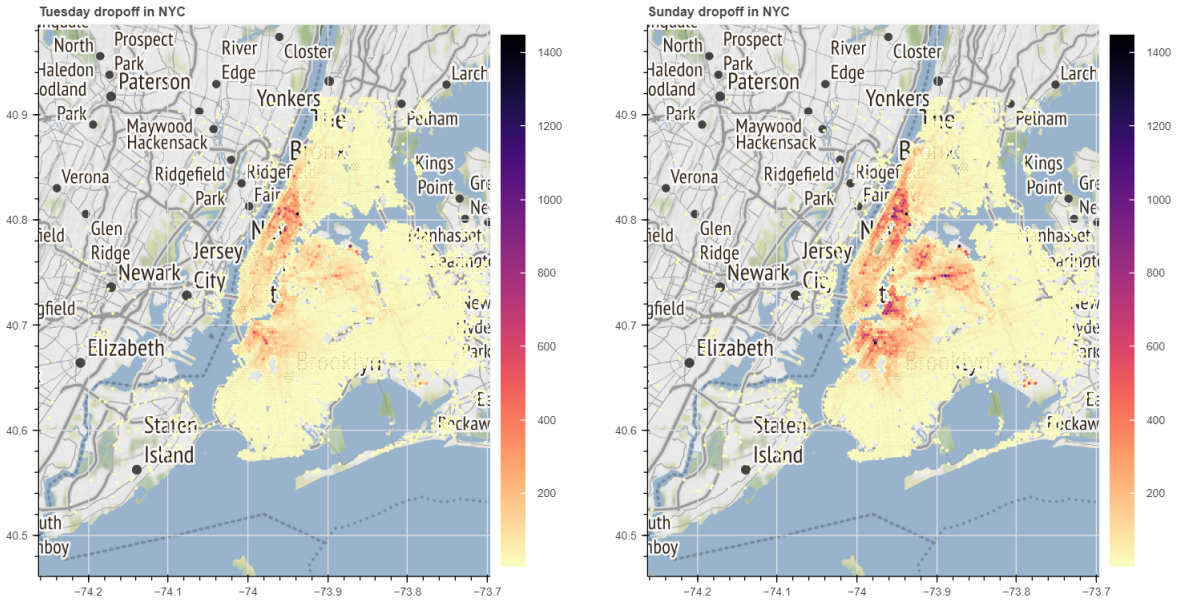


Figure 4: Map of Dropoffs on Tuedays and Sundays

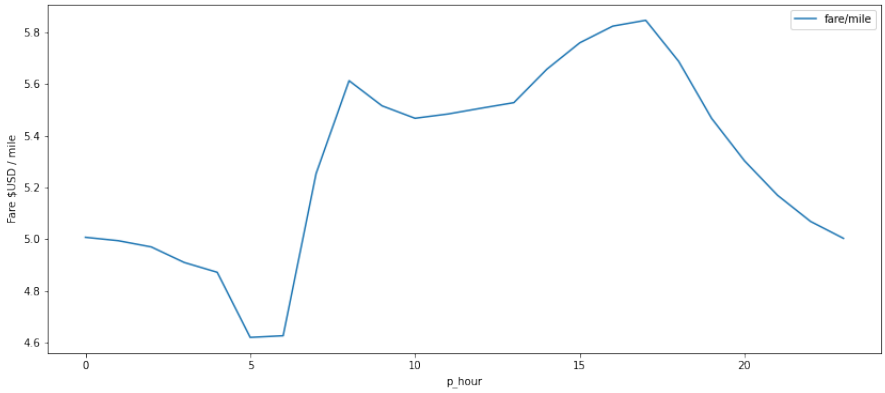


Figure 5: Graph of Fare per Mile against Hour of the Day

According to (*Figure 5*), the fare per mile is lowest at \$4.60/mile from 5 to 6 in the morning and peaks at \$5.80/mile around 5 to 6 in the evening. This is also due to the rush hour surcharge plus the congestion surcharge of \$2.75 [4] during peak hours.

In (*Figure 6*), the fare per mile for all zones are identical except for EWR. This is interesting because according to the TLC Taxi Fare Rules [4], trips to EWR have a \$17.50 surcharge, which means no surcharge for pickups at EWR. Upon further analysis shown in (*Figure 7*), it appears that there are rides where pickup and dropoff zones are both in EWR, and trip distance is less than a mile. This might be a result of disagreements between passengers and drivers, causing the trip to end prematurely.

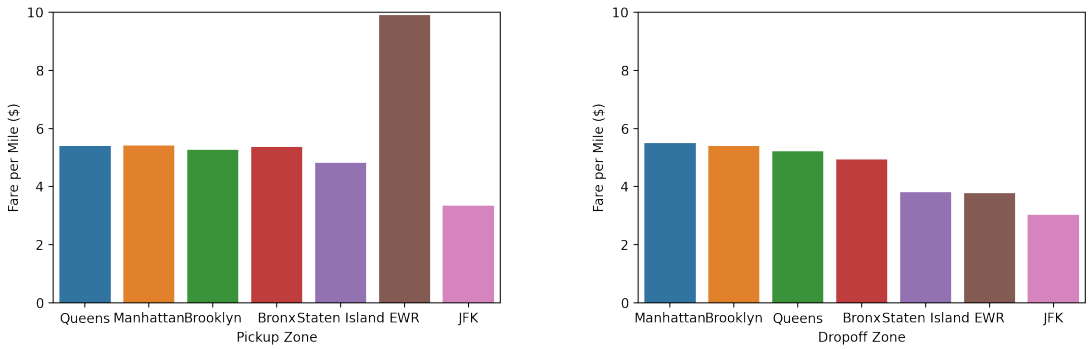


Figure 6: Barplot of Fare per Mile against Pickup Zone and Dropoff Zone

total_amount	trip_distance	p_weekday	p_hour	d_weekday	d_hour	time_elapsed	pickup_zone	dropoff_zone
78.30	23.77	5	15	5	16	74.683333	EWR	Manhattan
96.50	0.76	2	7	2	6	1.650000	EWR	EWR
36.30	6.49	3	11	3	10	8.366667	EWR	NaN
3.30	19.30	6	17	6	16	0.100000	EWR	EWR
63.00	15.40	2	7	2	6	0.933333	EWR	EWR
99.66	18.79	2	18	2	17	43.000000	EWR	Manhattan
17.50	32.10	6	16	6	15	0.433333	EWR	EWR
82.88	20.00	2	15	2	15	33.233333	EWR	Brooklyn
129.09	35.63	5	7	5	6	53.466667	EWR	Queens
93.50	0.80	6	6	6	5	0.016667	EWR	EWR
99.30	1.00	6	8	6	7	0.050000	EWR	EWR
96.00	2.29	7	15	7	14	3.616667	EWR	EWR
3.30	0.11	7	15	7	14	0.516667	EWR	EWR
53.80	0.07	7	19	7	18	1.583333	EWR	EWR
91.30	20.71	7	23	7	22	39.216667	EWR	Manhattan
33.30	15.50	1	7	1	6	0.300000	EWR	EWR
118.30	0.50	2	8	2	7	0.050000	EWR	EWR
5.15	0.50	2	8	2	7	1.100000	EWR	EWR
4.80	0.10	5	14	5	13	3.133333	EWR	EWR
16.30	4.53	6	10	6	9	13.800000	EWR	NaN

Figure 7: Table showing Pickup and Dropoff Zones of EWR

4.2 External Dataset Attribute Analysis: Traffic Collisions

(*Figure 8*) shows rides affected by collisions. It is obvious that traffic collisions happen mostly across places Manhattan and Brooklyn, where traffic volume is higher. One thing to note is that rides affected by collisions happen only on Sundays around 2pm to 3 pm as shown in (*Figure 9*). However, (*Figure 10*) shows that collisions happen regularly throughout the week regardless of whether it affects a ride. It is likely that rides affected by collisions on other days are probably removed during data cleaning.

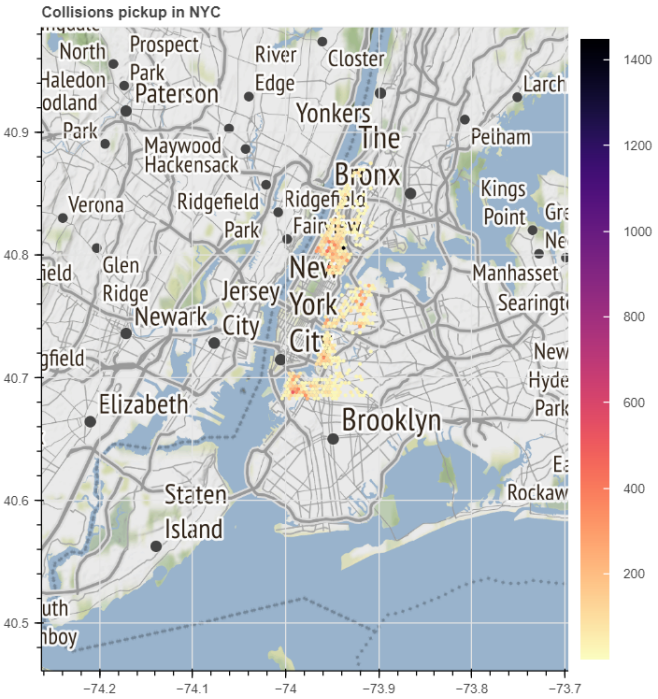


Figure 8: Map of Rides Affected by Collisions

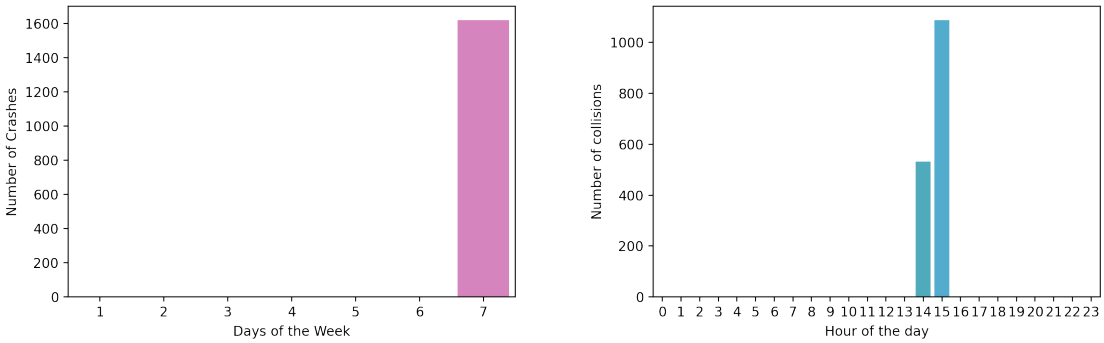


Figure 9: Number of Rides affected by Collisions against day of the week and hour of the day

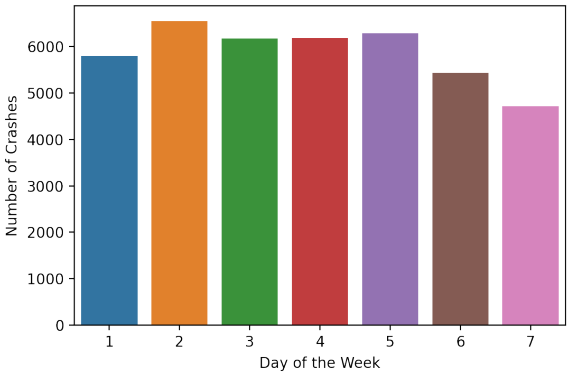


Figure 10: Bar plots of All Collisions vs day of the week and time of the day.

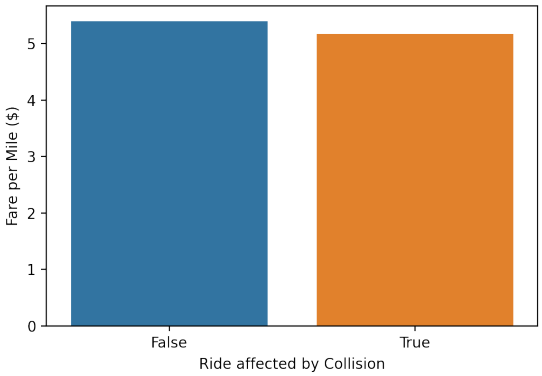
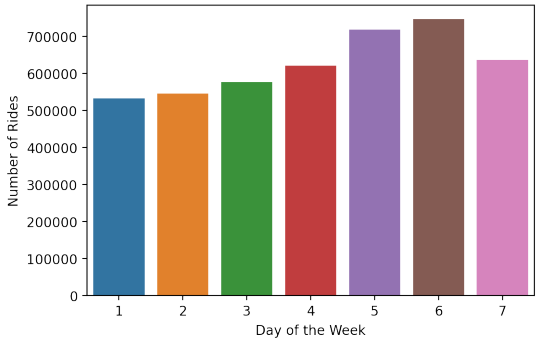


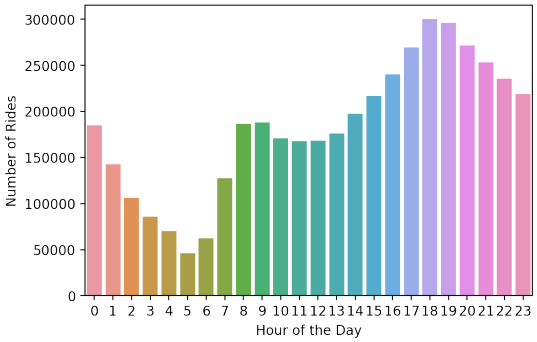
Figure 11: Fare per mile when a ride is affected by a collision.

It is expected that the fare is higher due to congestion surcharges, traffic overflow and so on when a ride is affected by a traffic collision. However, this expectation is defied because the fare per mile is actually similar when the ride is affected by a collision, as shown in (*Figure 11*). However, out of the 4.38 million rides, only 1620 were affected by collisions, therefore the sample size is too small to deduce that the fare per mile between them is similar.

4.3 Other Attributes Analysis



(a) Number of Rides against Day of the Week



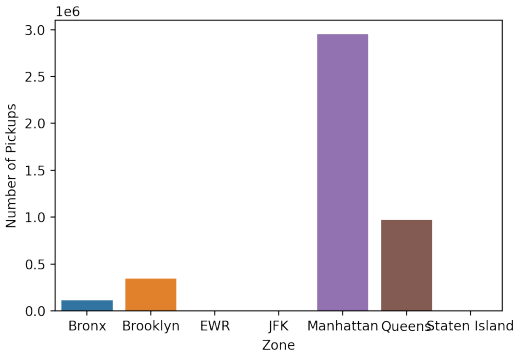
(b) Number of Rides Hour of the Day

Figure 12

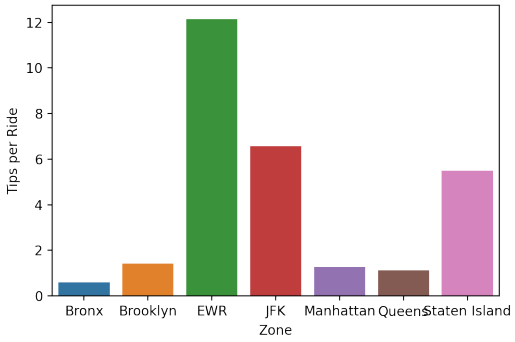
From (*Figure 12(a)*), there is a gradual increase in number of rides from Monday to Saturday, then a slight decrease on Sunday. This is because people usually go out after work on Fridays and weekends. Meanwhile, (*Figure 12(b)*) shows the lowest number of rides at 5am, which is not surprising. The number of rides increases as working hours start, and reaches the highest amount at 6pm when happy hour starts.

Shown in (*Figure 13(a)*), there are more pickups in Manhattan compared to the other boroughs combined, despite being the third most populated borough in New York City [7]. This is likely because Manhattan is so densely populated that people rather take cabs than drive and get stuck in traffic.

As seen in (*Figure 13(b)*), trips to EWR earned the most tips, followed by JFK and Staten Island. These three boroughs are relatively far from New York City Centre, hence better tips.



(a) Number of Pickups against Zone



(b) Tips per Ride against Zone

Figure 13

5 Statistical Modelling

5.1 Feature Selection

For feature selection in this model, *LASSO Regularisation* is used. This is because the dataset is considered small. *LASSO Regularisation* shrinks data values, making it suitable for models with high multicollinearity [8]. *LASSO Regularisation* uses the following mathematical equation:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} |\beta_j| \tag{1}$$

where λ is the amount of shrinkage. The equation can be interpreted as *Residual Sum of Squares + λ * Sum of the absolute value of the magnitude of coefficients* [8].
Using LASSO Regularisation, the following features are selected:

- **trip_distance**
- **time_elapsed**
- **pickupX**
- **pickupY**
- **dropoffX**
- **dropoffY**

Upon investigation, **time_elapsed** and **trip_distance** have a correlation of 0.798, which is improper to have both features in the model. **Time_elapsed** is removed from the model because **time_elapsed** is affected by **trip_distance**.

5.2 Linear Regression

The dataset is fitted to a **Linear Regression** model , which is covered in Machine Learning by Dr.Kris Ehinger [9]. This is because there is a strong linear relationship between fare amount and other features, as seen in (*Figure 1*). Besides, with most outliers removed, linear regression can perform better than other models.

5.3 Evaluation

The model is fitted with January, February and March 2016 datasets, and tested with fare amounts from April, May and July 2016.
As seen in (*Figure 14*), the residuals are close to 0 and distributed horizontally in a straight line, apart from an outlier on the top right corner. This proves that linear regression is a suitable model for this dataset.

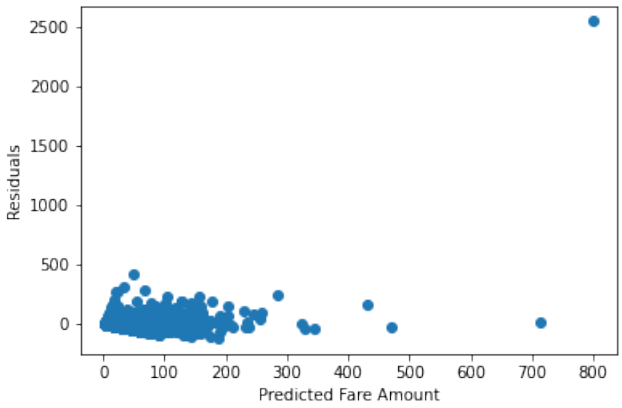


Figure 14: Plot of Residuals against Predicted Fare Amount.

Comparing the predicted fare amount from the linear regression model to the actual fare amount taken from the test dataset, a *mean squared error* of **8.81** and **0.89** *coefficient of determination* are obtained. The mean squared error shows that the difference between predicted value and actual value is acceptable, while the high coefficient of determination shows that there is a strong relationship between the features used and the fare amount.

6 Discussion

Taxi and e-hail have similar fare metrics, but e-hail actually displays an estimated fare before booking, while the final taxi fare is only revealed upon reaching the destination, unless negotiated before riding. Besides, e-hail doesn't charge based on the congestion of traffic, only during peak hours whereas taxi rides charges on both situations.
However, based on the findings above, the overall trend of taxi fare only increases following an increase in trip distance, which means most taxi drivers charge according to the TLC Fare Rules. As mentioned in **Section 4.1**, high demands for taxi rides won't affect the fare amount, whereas e-hailing apps often show "prices may be higher to due high demand".
Hence, if you're in a rush, street-hailing for a taxi is much quicker than using e-hail apps as there is a waiting time for the driver to arrive. Otherwise, e-hailing and taxi don't have much difference. It is a misconception that taxi fares are overall more expensive than e-hailing.

7 Conclusion

It is no secret that the number of e-hail drivers have rocketed throughout the years, causing a huge impact on taxi drivers. However, e-hail companies are doing their best to help taxi drivers by including taxi services in their apps. It is believed that taxi services will not become obsolete as they are considered essential in major cities like New York where street hailing is much more convenient.

References

[1] "Taxi Fare." Taxi Fare - TLC. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[2] "Traffic Collision Dataset." NYC Open Data by New York Police Department (NYPD). https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

[3] "NYC boundaries coordinates." BoundingBox by Klokan Technologies: https://boundingbox.klokantech.com/

[4] "Taxi Fare Rules." Taxi Fare - TLC. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

[5] "Coordinates of dropoff and pickup points of JFK." John F. Kennedy International Airport, Google Maps, 2021. https://www.google.com/maps/place/John+F.+Kennedy+International+Airport/@40.6446662,-73

[6] "Green Cab Pickup Zones." Green Cab - TLC. https://www1.nyc.gov/site/tlc/businesses/green-cab.page

[7] "Population of Boroughs in New York City." Boroughs of New York City by Wikipedia. https://en.wikipedia.org/wiki/Boroughs_of_New_York_City

[8] "Lasso Regression." A Complete Understanding of LASSO Regression by Great Learning Team, contributed by Dinesh Kumar, 4 September 2020.

[9] "Linear Regression." Machine Learning COMP30027 by Dr. Kris Ehinger, University of Melbourne.