Comparing Uber and Yellow Taxi in NYC

Khanh Nam Nguyen
Student ID: 1367184
https://github.com/MAST30034-AppliedDataScience/project-1-individual-knam2609

August 28, 2024

1 Introduction

In New York City, the competition between Yellow Taxis and Uber reflects a broader shift in urban transportation. This project, conducted from June to November 2023, compares the performance of these services based on data from this period. By examining metrics like pricing, availability, and efficiency, the study aims to highlight trends in consumer behavior and service dynamics. As ride-hailing apps continue to influence the transportation industry, this analysis offers valuable insights for various stakeholders interested in the evolving landscape of urban mobility in one of the world's busiest cities.

2 Preprocessing

1. **Yellow Taxi Data**

   - **Column Renaming and Feature Addition:** During data preprocessing, it became evident that some column names, like `tpep_pickup_datetime` and `tpep_dropoff_datetime`, were unnecessarily complex and hindered the dataset's clarity. To enhance readability and facilitate analysis, these columns were renamed with simpler labels. Additionally, new features were introduced to enrich the dataset: `trip_time` was calculated as the difference between `pickup_datetime` and `dropoff_datetime`, giving a clear measure of trip duration. The `fare_per_miles` feature was created by dividing `total_amount` by `trip_distance`, offering insights into trip cost efficiency. These adjustments not only streamlined the dataset but also added valuable metrics for deeper analysis.

   - **Data Integrity and Outlier Management:** During data cleaning, it was clear that the dataset had numerous out-of-range and incorrect negative values across various columns. For example, negative values appeared in the `fare_amount`, `tip_amount`, and `trip_time` columns, and some records reported implausible trip distances over 90,000 miles, trip times near 6,000 minutes, and `fare_per_miles` values ranging from 33,000$to - 17,000$.

     Initially, I tried using quantiles to remove outliers. While this method worked for filtering extreme `fare_per_miles` values, the quantiles for other columns were too small to apply effectively at this dataset's scale. Given the complexity, I decided to defer final data cleaning decisions until I could compare these anomalies with the Uber data. This comparison would provide context and help avoid distorting the dataset or obscuring meaningful trends.

   - **Missing Data:** In the yellow taxi dataset, a significant number of records had null values in multiple columns, including `passenger_count`, `RatecodeID`, `store_and_forward_flag`, `congestion_surcharge`, and `airport_fee`. These missing values were clustered within the same rows, indicating that the data was missing completely at random and not due to any external factors or data collection issues. Given the dataset's large size, the impact of these null entries on the overall analysis was minimal. Therefore, the most effective approach was to remove all records containing these null values, ensuring the dataset's integrity and usability while maintaining a robust foundation for analysis.

2. **Uber Data**

   - **Column Renaming and Feature Addition:** To ensure consistency and facilitate comparison between the Uber and yellow taxi datasets, several Uber columns were renamed to match the yellow taxi naming conventions.

     New features were also added: `trip_time` was converted from seconds to minutes, and `total_amount` was created by summing fare-related columns. `waiting_time` was calculated as the difference between `request_datetime` and `on_scene_datetime`, and `fare_per_miles` was introduced to assess trip cost efficiency.

   - **Data Anomalies:** Upon inspecting the Uber data, it was clear that the issues identified in the yellow taxi data were less severe here. While negative values persisted in fare-related columns, there were fewer extreme outliers, making the dataset more manageable.

     The `fare_per_miles` column still showed some unusually high values, but quantiles remained useful for outlier removal. However, quantiles for `trip_distance` and `trip_time` were too low to be practical, similar to the yellow taxi data. Despite being more robust, the Uber data still required careful handling to address anomalies.

     Interestingly, the `waiting_time` column had some extremely high values, which is understandable due to Uber's advance booking system. However, negative values were incorrect and needed to be removed.

   - **Missing Data:** During preprocessing, it was noted that missing values were concentrated in the `originating_base_num` and `on_scene_datetime` columns, exclusively in Lyft trips. This posed a challenge, as `waiting_time`, a key metric for service efficiency, couldn't be calculated without `on_scene_datetime`. This limitation led to the decision to focus solely on analyzing Uber data.

     Additionally, missing values in the `fare_per_mile` column were linked to trips with zero `trip_distance`, leading to undefined fare calculations. These zero-distance trips require careful handling to avoid distorting the analysis.

     These issues underscore the importance of thorough data cleaning to ensure accurate and reliable analysis, making it essential to address missing and anomalous data properly.

3. **Decisions**

   - **Final Data Cleaning and Preprocessing Decisions:** After thoroughly inspecting the Uber and Yellow Taxi datasets, several key decisions were made to refine the data for analysis.

     Firstly, all records with null values were removed, as these incomplete entries were deemed unfit for meaningful analysis, ensuring the remaining dataset is robust.

To address negative or zero values, a targeted approach was used to filter out logically impossible figures, particularly in fare-related columns, to maintain data integrity.

For outlier removal, a quantile-based method was applied, especially in the `fare_per_miles` column, to preserve the overall data distribution while eliminating skewed values. Maximum values from Uber data served as benchmarks to filter out unreasonable entries in the Yellow Taxi dataset.

Additionally, flag indicator columns were removed as they were irrelevant to the analysis, helping to streamline the dataset.

Lastly, records with invalid location IDs and timestamps were dropped to maintain accuracy in geospatial and temporal analyses.

4. **Weather Data**

   - In regards to the weather data, I chose to focus on the three most important aspects of weather which are temperature, humidity and precipitation. Besides, I will join the weather data to the taxis data based on date.

3 Analysis and Geospatial Visualisation

1. **Yellow Taxi Data**

   - **Features Correlation:**

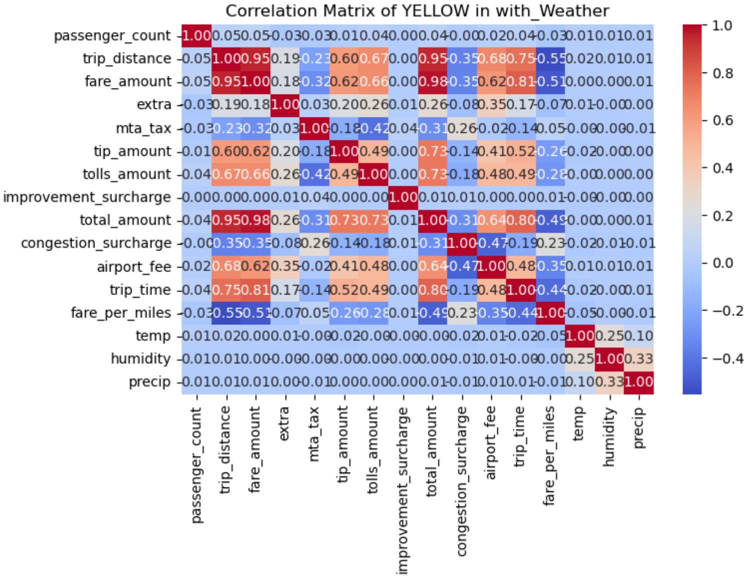

Figure 1: Features Correlation in Yellow Taxi Data

- As expected, from Figure 1, `trip_distance`, `trip_time`, and `total_amount` are highly correlated, confirming the intuitive relationship that longer trips, which naturally take more time, also result in higher fares.
- `tolls_amount` and `tip_amount` are also highly correlated with both `trip_distance` and `total_amount`. This suggests that longer trips are more likely to include highway

segments, where tolls are incurred, leading to higher tolls for drivers and passengers. The correlation between tips and distance, as well as total fare, is understandable, as tips are often a percentage of the fare. Moreover, the moderate correlation between tips and tolls could indicate that passengers may be more inclined to tip generously when tolls are higher, perhaps to support drivers in covering these additional costs.

- `airport_fee` is also positively correlated with `trip_distance` and `trip_time`, indicating that longer trips are often associated with requests made at the airport. This correlation makes sense, as airport trips typically involve greater distances and time due to the location of airports relative to other parts of the city.
- `fare_per_miles` is negatively correlated with both `trip_distance` and `total_amount`. This is expected, as longer trips tend to have a lower fare per mile, and higher total fares, which are indicative of longer trips, also correlate with lower fare per mile. This relationship suggests that the pricing structure reduces the fare per mile as the trip distance increases, providing a discount on longer trips.
- The weather attributes (`temp`, `humidity`, `precip`) show very weak or no correlation with most of the taxi-related metrics. This suggests that weather conditions during this period might not significantly affect taxi trip distances, fares, or other metrics in this dataset.

- **Features Distribution**

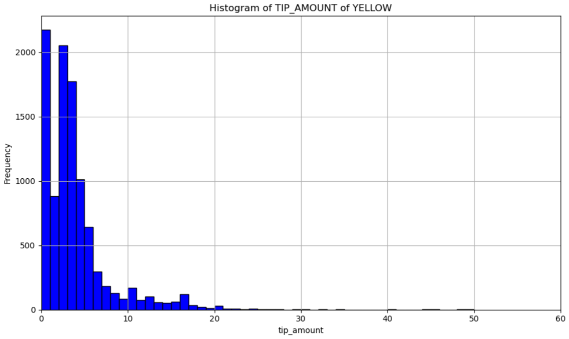

Figure 2: Tips Histogram

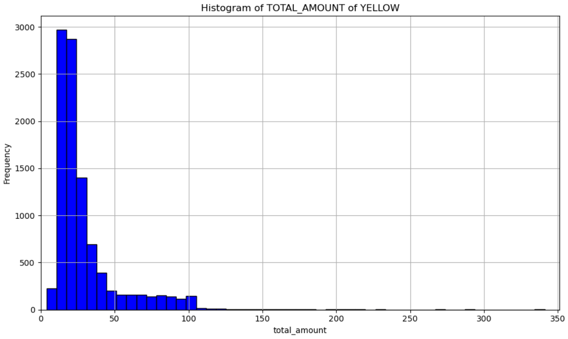

Figure 3: Total Fare Histogram

  - From figure 2 and 3, the distribution of numerical features in the dataset is left-skewed, which is expected given the nature of the data. Most trips are relatively short, and the tips given are generally modest, which naturally results in a left-skewed distribution. This skewness does not indicate any abnormal patterns but rather reflects the typical behavior of taxi trip data, where extreme values are rare, and the majority of data points cluster around smaller values. Therefore, the dataset remains suitable for analysis, with the skewness being a natural characteristic of the data.

- **Time Series Analysis**
  - The daily data from figure 4 and 5 do not provide any significant insights, the only significant detail to mention here is that the tips and total fare increase suddenly in late September maybe due to this is the best time in fall for tourists to come visit.
  - The hourly data from figure 6 and 7 show that the tips and total fare is the highest around 5am and lowest around 2-3am for tips and 2-3am and 10am for total fare. The

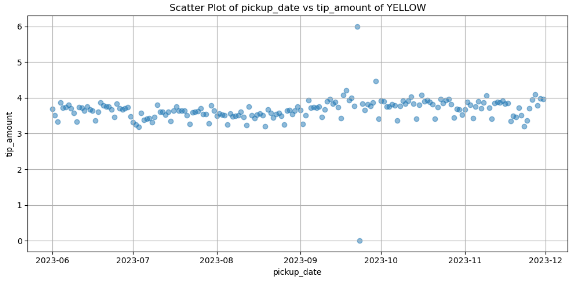

Figure 4: Daily Tips

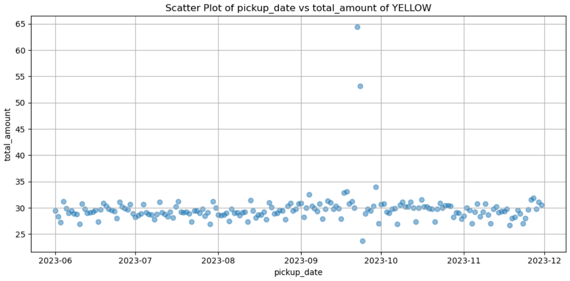

Figure 5: Daily Total Fare

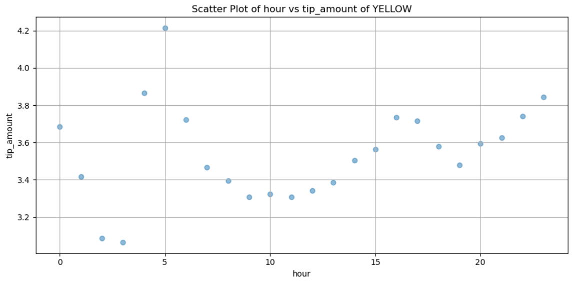

Figure 6: Hourly Tips

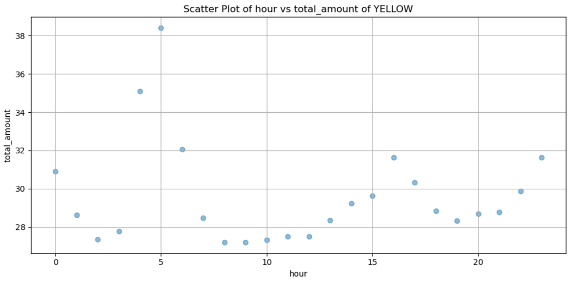

Figure 7: Hourly Total Fare

reasons for these are because passengers are more prone to tip drivers in early morning and late night trips as we can see the tips is also higher from 8pm to 12am. The total fare is high possibly due to the trips are from the airport whose passengers are people who landed in early morning or at night. Both tips and fare are the lowest during 2-3am, this is understandable because there are not many people using taxis around that time. The total fare is also low during the day might be the result of not being in the rush hour. In the afternoon when the rush hour approaches around 5-7pm, both the tips and total fare increase.

- From figure 8, 9, 10, congestion surcharge, trip time and fare/mile all rise as rush hour approaches which is very natural. The values are always the lowest at around 2-3pm as no ones is using taxi around that time.

2. **Uber Data**

- **Features Correlation**

  - From Figure 11, we observe that, similar to the Yellow Taxi data, `total_amount`, `trip_distance`, and `trip_time` are highly positively correlated. Additionally, `driver_pay`, `tolls`, `bcf`, and `sales_tax` also show strong positive correlations with these trip-related attributes. This makes sense as `bcf` and `sales_tax` are surcharges added to trips, and therefore, they naturally increase with higher trip distances and total amounts. Similarly, `driver_pay` tends to be higher for longer trips with higher fares, which further strengthens the positive correlation with these trip attributes.

  - Everthing else looks mostly similar to Yellow taxi, but there is a difference in tips. The tips in Uber data does not correlate with trip attributes as highly as in Yellow taxi, it can be explained by Uber drivers seldomly get tipped.

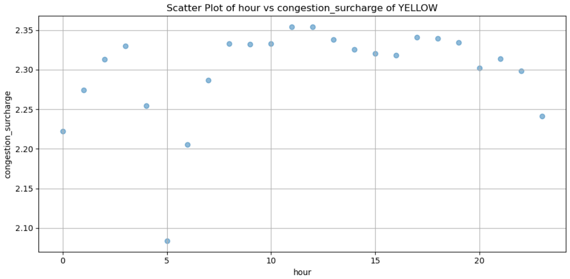

Figure 8: Hourly Congestion Surcharge

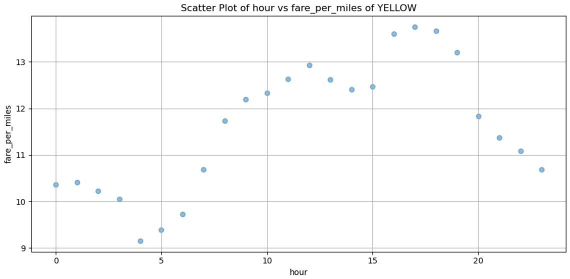

Figure 9: Hourly Fare/Mile

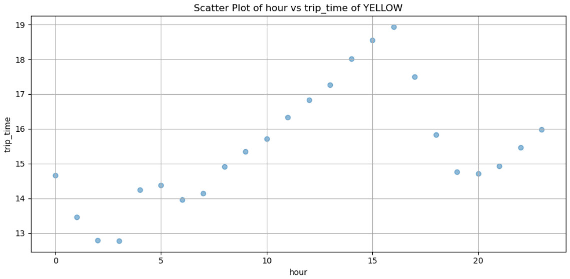

Figure 10: Hourly Trip Time

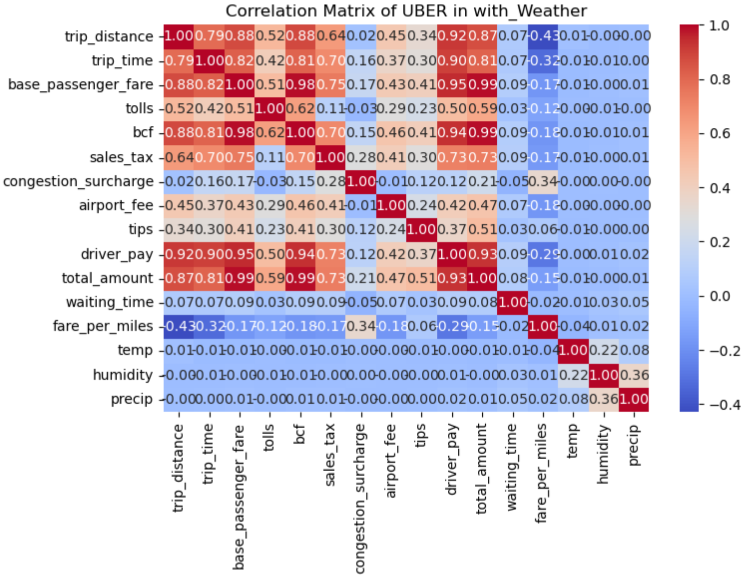

Figure 11: Features Correlation in Uber Data

- **Features Distribution**

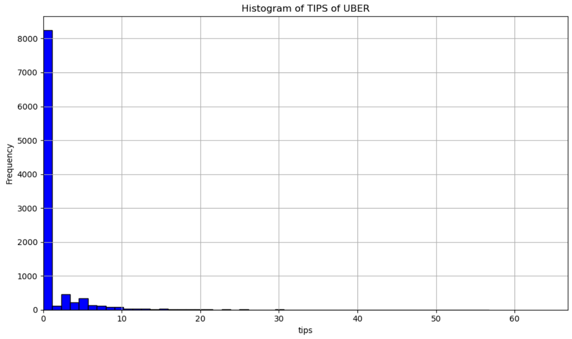

Figure 12: Tips Histogram

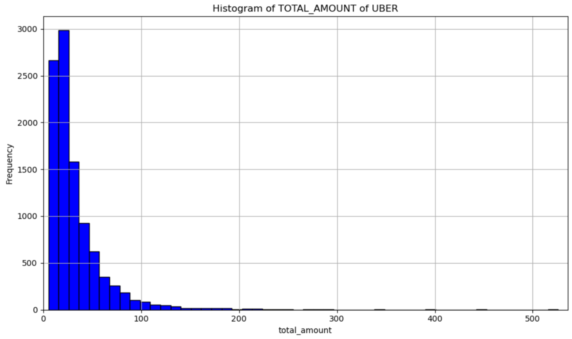

Figure 13: Total Fare Histogram

  - From figure 12 and 13, we can see the tips and total fare are also naturally left-skewed but the tips is much more left-skewed as Uber drivers do not get much tips. The numerical features from Uber data are all left-skewed, naturally of course, just like those in Yellow taxi. But we should have a look at `waiting_time` because it is very unique for Uber.

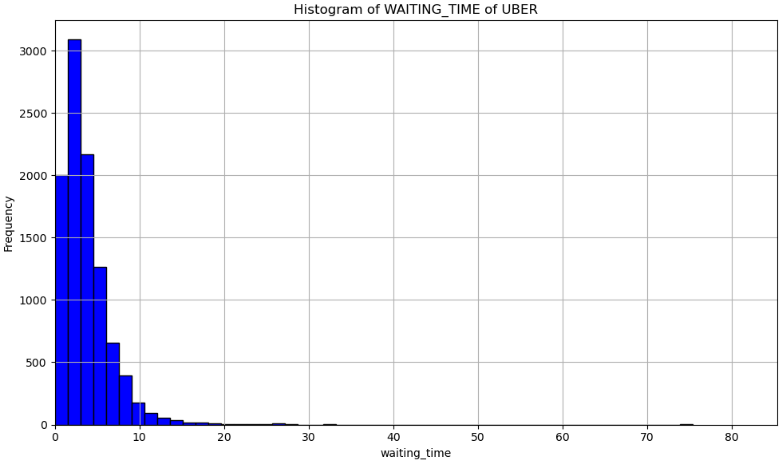

Figure 14: Waiting Time Histogram

  - We can see from figure 14 that the waiting time is also left-skewed, indicating that Uber is doing a good job keeping the waiting time of customers as low as possible. This allows Uber to compete with street-hail taxi like Yellow taxi in terms of arrival speed.

- **Time Series Analysis**
  - I found the daily analysis is not very interesting after analyzing Yellow taxi as it does not show any trends, Uber data is no different, the only difference is from figure 15, the points are more sparse compared to Yellow taxi. This can be explained by how Uber fare fluctuates more because it it calculated by the app's algorithm, unlike fixed fare from Yellow taxi.
  - We can see from figure 16 that the tips does not peak at early morning at 5pm anymore

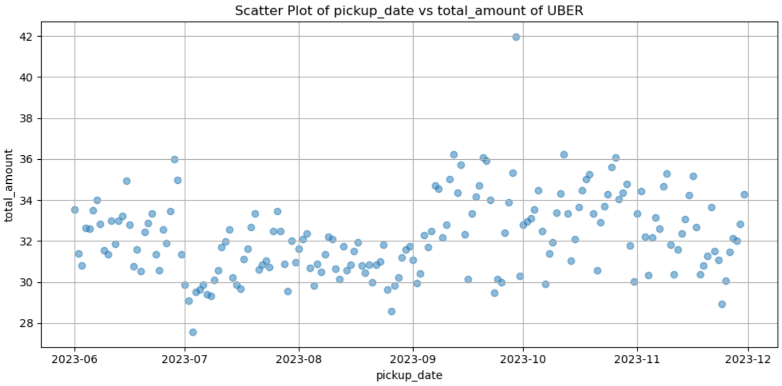

Figure 15: Daily Total Fare

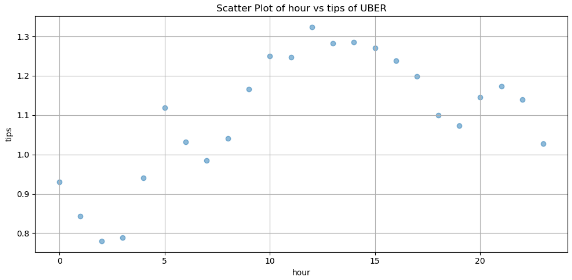

Figure 16: Hourly Tips

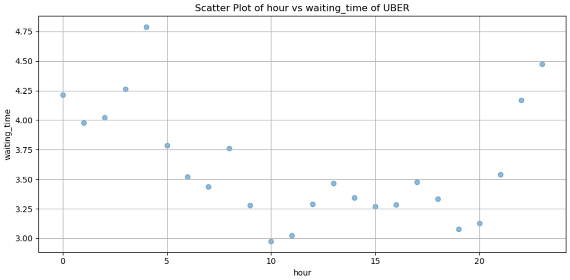

Figure 17: Hourly Waiting Time

as passengers do not usually tip Uber drivers. However, as noon and early approaches, the tips gradually increases even though it is not rush hour, showing that Uber drivers are more likely to get tipped when the road is less busy, possibly because they are nicer to passengers during those times.

  - From figure 17, I found it surprising that the waiting time is very short during rush hour, which maybe due to more drivers go to work at that time hoping to get more money. This once more confirms that Uber has done a phenomenal job at providing fast service to its customers. The longest waiting times are around late night to 2-3pm, the obvious reason is there are not many drivers working at that time.

- **Spatial Visualisation and Comparisons**
  - From figure 18 and 19, we can see that Uber covers all 5 boroughs, the only place with no trips is Newark Airport because of the ban. With Yellow taxi, although it covers Newark Airport but does not cover most of Staten Island and some other places, this is maybe due to Yellow taxi drivers do not want to take trip in outer borough like Staten Island. In the inner boroughs like Brooklyn, Queens and Manhattan, there is a huge competition between the two. However, Uber with its huge number of drivers, are still able to cover the whole area more effectively while Yellow taxi can only focus on JFK airport, LaGuardia Airport, East Manhattan and its central area. This illustrates how Uber dominates Yellow taxi in terms of coverage. Next we will compare drivers' tips, my expectation is Yellow taxi drivers will get way more tips than Uber drivers as all the data shown above are proving this idea.

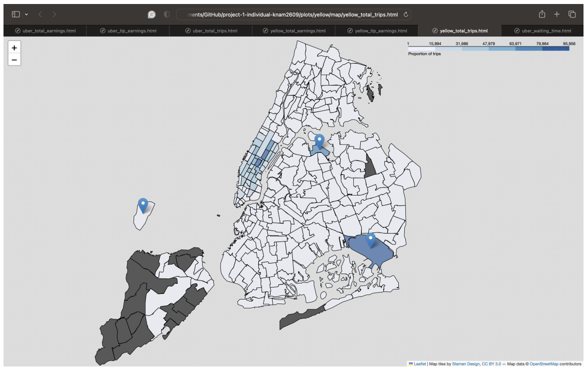
ments/GitHub/project-1-individual-knam2609/plots/yellow/map/yellow_total_trips.html
uber_total_earnings.html
uber_tip_earnings.html
uber_total_trips.html
yellow_total_earnings.html
yellow_tip_earnings.html
yellow_total_trips.html
uber_waiting_time.html
1
15,994
31,986
47,979
63,971
79,964
95,956
Proportion of trips
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 18: Yellow Total Trips

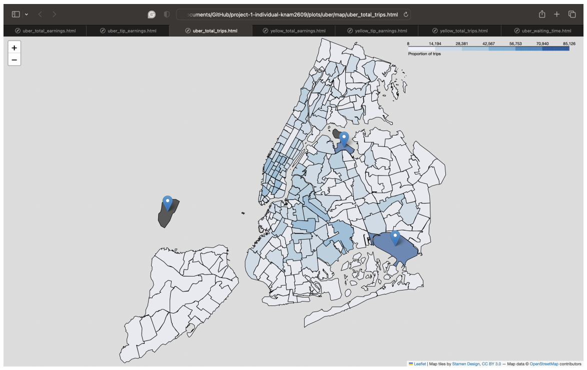
cuments/GitHub/project-1-individual-knam2609/plots/uber/map/uber_total_trips.html
uber_total_earnings.html
uber_tip_earnings.html
uber_total_trips.html
yellow_total_earnings.html
yellow_tip_earnings.html
yellow_total_trips.html
uber_waiting_time.html
8
14,194
28,381
42,567
56,753
70,940
85,126
Proportion of trips
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 19: Uber Total Trips

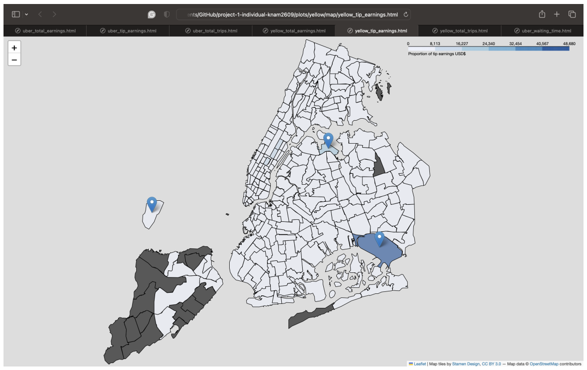
nts/GitHub/project-1-individual-knam2609/plots/yellow/map/yellow_tip_earnings.html
uber_total_earnings.html
uber_tip_earnings.html
uber_total_trips.html
yellow_total_earnings.html
yellow_tip_earnings.html
yellow_total_trips.html
uber_waiting_time.html
0
8,113
16,227
24,340
32,454
40,567
48,680
Proportion of tip earnings USD$
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 20: Yellow Tips Earnings

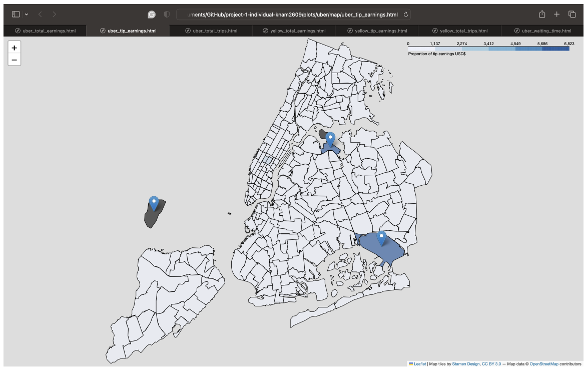
uments/GitHub/project-1-individual-knam2609/plots/uber/map/uber_tip_earnings.html
uber_total_earnings.html
uber_tip_earnings.html
uber_total_trips.html
yellow_total_earnings.html
yellow_tip_earnings.html
yellow_total_trips.html
uber_waiting_time.html
0
1,137
2,274
3,412
4,549
5,686
6,823
Proportion of tip earnings USD$
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 21: Uber Tips Earnings

- As expected, figure 20 and 21 confirm that although Uber' coverage is far greater, its drivers barely get tipped while Yellow taxi drivers get about 8 times more tips despite having fewer drivers. Now let's have a look at Uber's average waiting time to see if it is really that good like how the data above has shown us.
- As we can see from Figure 22, most average waiting time is 4-5 minutes which is more than good. So the idea of Uber being very fast with its service is confirmed and we can now understand why Uber is so dominant agaisnt Yellow taxi.

Feature	Yellow	Uber
fare/mile	8.22	6.24
fare/trip	29.48	32.17
tips/trip	3.67	1.16
tips/mile	1.02	0.22
miles/trip	3.58	5.15
avg wait	NONE	3.67

Table 1: Comparison Table

- Table 1 confirms everything I have analyzed above. Furthermore, we can see that Uber is actually cheaper than Yellow taxi in terms of fare per mile. This might be due to discounts Uber gives its customer and its algorithm also lower the fare during non-rush hours. We can see the average waiting time of Uber is only 3.67 including those booked-in-advance trips, showing us how fast customers can get their trip after

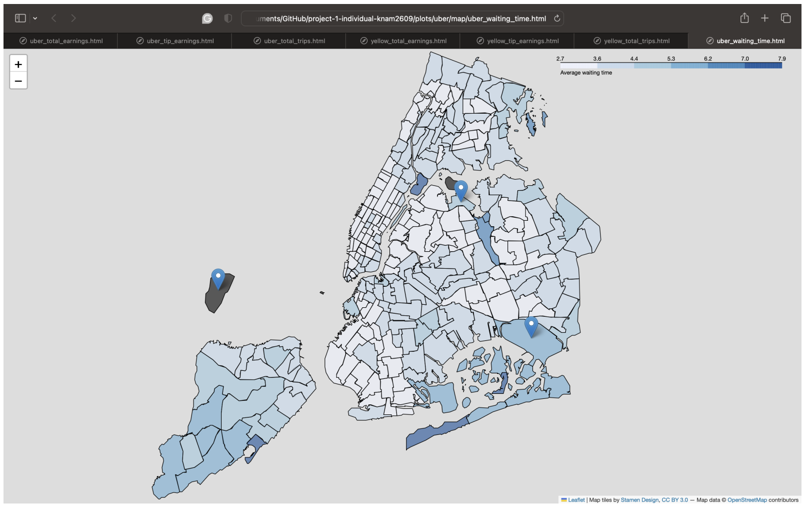
ments/GitHub/project-1-individual-knam2609/plots/uber/map/uber_waiting_time.html
uber_total_earnings.html
uber_tip_earnings.html
uber_total_trips.html
yellow_total_earnings.html
yellow_tip_earnings.html
yellow_total_trips.html
uber_waiting_time.html
+
−
2.7
3.6
4.4
5.3
6.2
7.0
7.9
Average waiting time
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Figure 22: Uber Average Waiting Time

requesting. To confirm Uber is actually cheaper than Yellow taxi, we can look back at their hourly average fare per mile.

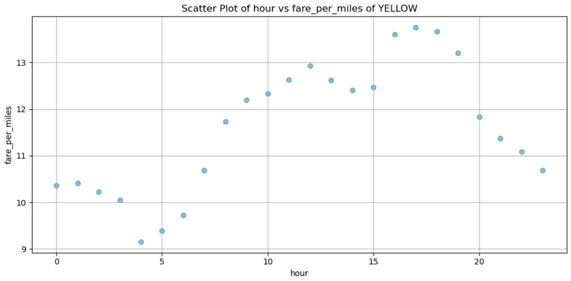

Figure 23: Yellow Hourly Fare Per Mile

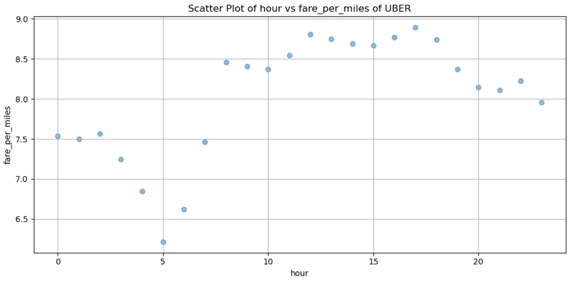

Figure 24: Uber Hourly Fare Per Mile

– From figure 23 and 24, yes, Uber is indeed cheaper than Yellow taxi!

4 Statistical Modelling

- **Goal and Preprocessing:** My goal is using Linear Regression to predict the `total_amount` in November 2023 of Yellow taxi and Uber using data from June to October 2023, I will predict then compare Yellow's model and Uber's model. For preprocessing, I will get rid of fare-related features which directly contribute to total fare, encode categorical features and turn timestamp into year, month, hour and minute features. I will not normalize or standardize trip distance or trip time because I want them to contribute mostly to the result.

- **Model Result Comparison**

Metric	Yellow	Uber
R^2	0.94	0.85
RMSE	5.36	11.39

Table 2: Model Evaluation Table

- From Table 2, we can observe that Yellow model performs better than Uber model. This is within my expectation because Uber has an algorithm to calculate their fare based on the mean time situation. Besides, Uber also has a lot of different discounts for its customers so it will of course be harder to predict its trip fare. Yellow taxi on the other hand, has fixed fare and does not offer official discounts so its fare can be easily predicted by trip distance and time.

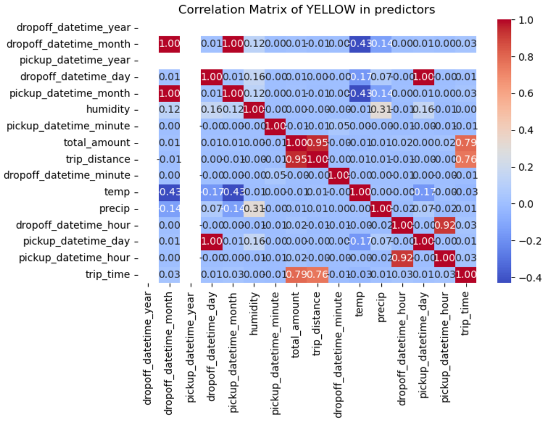

Figure 25: Yellow Correlation

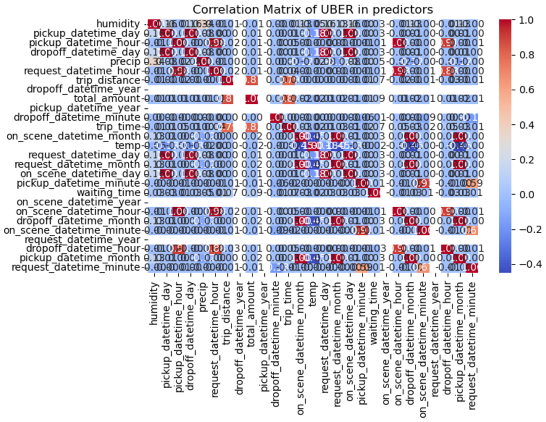

Figure 26: Uber Correlation

- We can tell from figure 25 and 26 that my decision to not normalizing or standardizing trip distance and time was reasonable as they are highly correlated with total fare while other predictor are not. The other reason for why Uber model performs worse is it has much more predictors that are not correlated with total fare, if we want to improve the model, we will need to do some feature selection.

5 Recommendations

- **For The Two Companies:**
  - Uber is already dominating Yellow taxi so I do not have much to recommend them but if I have to, I would suggest them including more bonuses for drivers when taking trips at the airport, at late night or in early morning to attract more employees. It is because Uber drivers rarely get tipped so bonuses would encourage them to work more during those extreme hours and get more customers for the company.
  - Yellow taxi should really consider giving its customers more discounts as its fare are not as competitive as Uber. Most of its trips are at the airport so a lot of its customers can be foreigners or outsiders that do not know the price and too afraid to use Uber on their first day to a new city. However, in most of the other places, I do not see the reason to use Yellow taxi as Uber is way cheaper and more convenient. To improve even more, Yellow taxi should send more drivers to the outer boroughs such as Staten Island to increase the coverage, it can also station their drivers at specific points or advertise using the yellow color more to make sure passengers can always see their iconic yellow taxi color. This might be a very innovative idea because people say ”Out of Sight Out of Mind” so if people see yellow everywhere, they are more likely to use Yellow taxi.
- **For passengers:** If you are new to the city or have just arrived at the airport, you should

use Yellow taxi as they are more convenient because you can see them everywhere and directly communicate with the drivers immediately. However, I am still a fan of Uber as they are more convenient and arguably cheaper so I would still recommend using Uber in general.

- **For drivers:** I would suggest you work for both companies part time. You can work your morning trip for Yellow taxi at the airports and in Manhattan to earn some big trips and big tips, then in the afternoon work as an Uber driver in inner boroughs to get more trips and bonuses during rush hours. If you find it hard to get trips because of high competition with other drivers, I suggest going to Staten Island to find trips because there is almost no yellow taxis here so you can compete better. This will maximise your income and make sure you do have trips the whole days.