Green Taxi:
Competitor Market Analysis

Assessed By Barbara Mikulášová

Report Generated By User
Written By James Wright

Report Generated On
Saturday, 10 June 2023

Date First Published
Friday, 09 June 2023

Contents

1 Data Quality Report

1.1 Metadata Quality Issues

Summary of problems and corrections with the metadata on import:

- All variable names should be made upper-case by convention for formatting clarity;
- There should be a consistent variable name word-separation convention across table variables;
  - For example, variable PAYMENTTYPE in taxi_trips renamed to PAYMENT_TYPE.
- There should be a consistent variable name conventions across all tables;
  - For example, variable LOCATION_CODEID in taxi_time_location was renamed to LOCATIONID for a consistent ID format across tables and common name across tables.
- Missing variables
  - Tip amount only includes credit card tips, data should be collected on cash tips to help future analysis.
  - The entire variable LOCATION_DETAILS column in taxi_time_location was missing but is in appendix;
    - Added containing NA values, for completeness, on load.
- Ill-defined groups in appendix - ‘no charge’, ‘dispute’, ‘voided’ need clear definitions of their categorisation for potential prediction of missing value purposes.

1.2 Duplication Issues

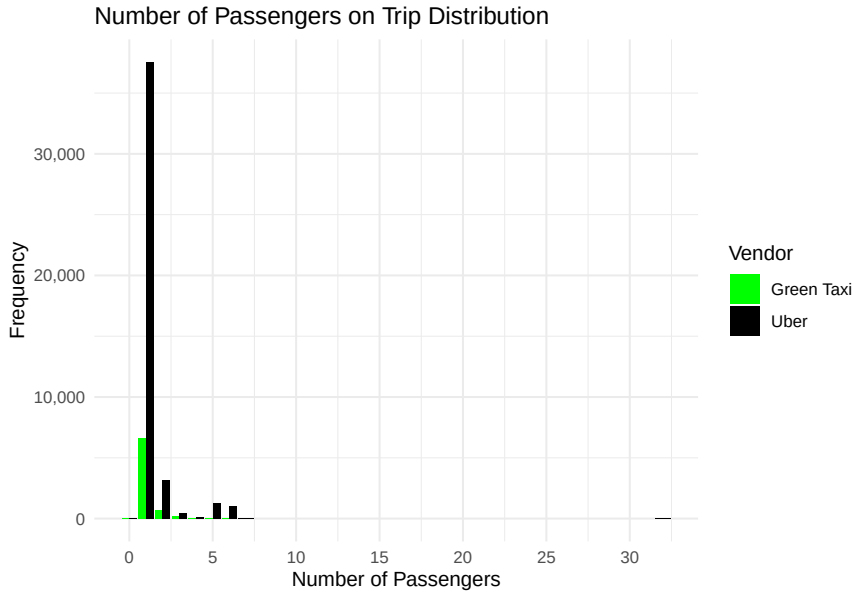
Summary of problems and corrections of duplicates:

- There should not be duplicates stored in the data.
- There were 67 duplicates in the taxi_trips data, 0 duplicates in the taxi_zone_lookup data, and 0 duplicates in the taxi_time_location data.
- The duplicates have been saved as an Excel file in the ‘Reports’ folder to help review which data gets duplicated in the collection process.

1.3 Unexpected Data Value Issues

Summary of problems and corrections of the data values:

- Unexpected values;
  - There are negative and positive fare amounts. This might make sense as they could be refunds for voided and free trips so they have been kept in the cleaned data, but for clarity data should be collected or flags assigned to point out where a $\pm$ sign comes from.
  - In rows of taxi_zone_lookup, there is a zone 'NV' which is inconsistent with the naming of other zones. Care should be taken inputting values to be consistent.
    - This was treated as typo for NA based on the surrounding context of other NAs in the row, and made 'Unknown' for reporting purposes.
  - Number of passengers has values up to 30. The typical capacity of a taxi is only going to be 6 people, so data is possibly being collected on the wrong kinds of vehicles or entered wrongly.
- Unexpected categories - all data category values must be accounted for / explained;
  - Payment type '7' and rate code '8' are not defined in the appendix but exist in the data.
    - Redefined '7' values as Unknown '5' in the cleaned data for reporting purposes.
  - Rate code '8' is not defined in the appendix but exists in the data.
    - Leaving as it is, as all correspond to vendor 3 who might have their own rate code like New York cabs.
- Value distribution
  - There is a disproportionate amount of Other versus Uber data versus Green taxis, this may reflect the proportion of the true market share by each vendor, but this must be checked to make sure the data collection is fair.



1.4 Missing Data Issues

- Inconsistency with naming convention for missing values - e.g. Unknown, N/A, NA, NV.
  - All not available data should follow a common naming convention.
- The entire E-Hail fee column is missing. Data must be collected for all stored variables or stop storing the variable.
  - For analysis purposes, the column was dropped as it was not needed.
- Other vendors have entirely missing columns of values for forward flag, rate code, e-hail fee, passenger count, and congestion surcharge.
  - This data needs to be collected or other data collected that enables you to infer these values using known values.
- Data was only available in 2021 for late June to early August. More data should be collected for more comprehensive analysis of 2021.

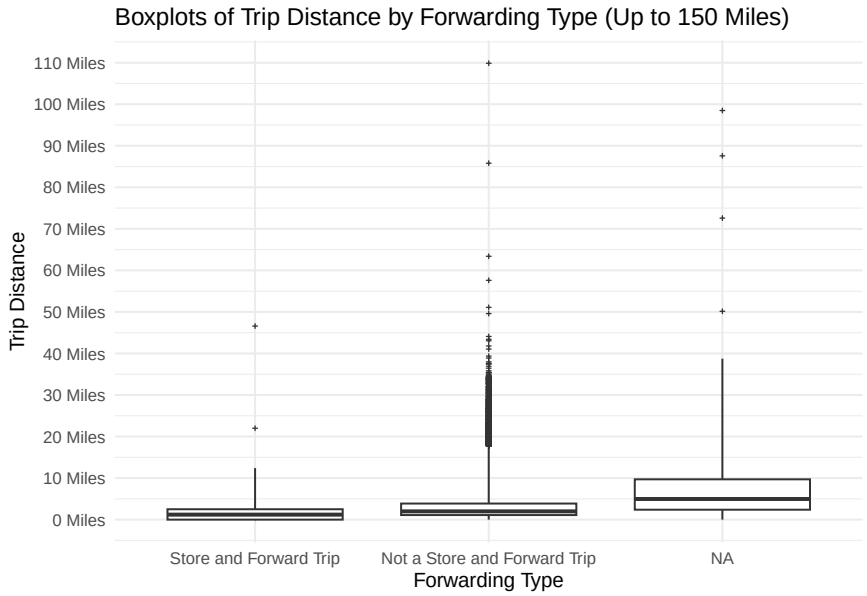
1.4.1 Count of Columns with Missing Values

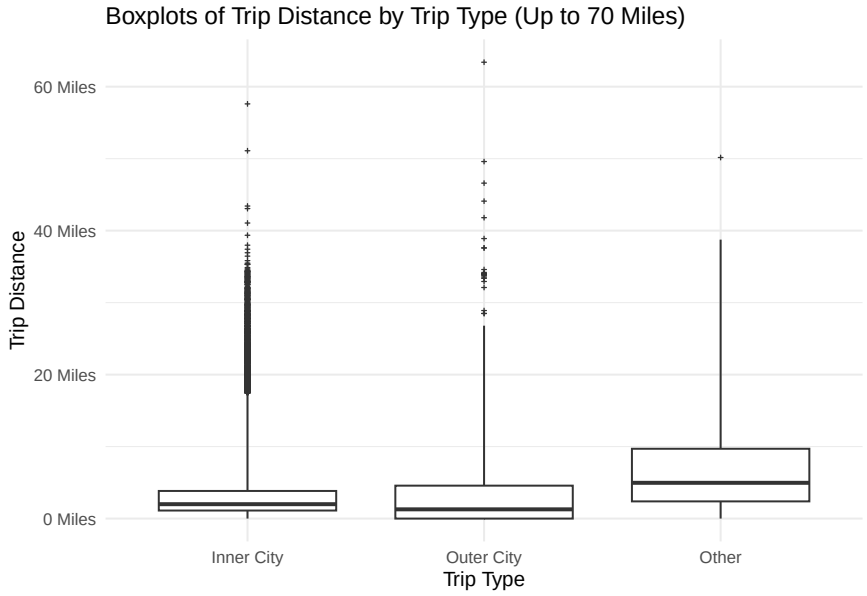
Table 1: Count of Columns with Missing Values

Forward Flag	Passenger Count	E-Hail Fee	Congestion Surcharge
32518	32518	83691	32518

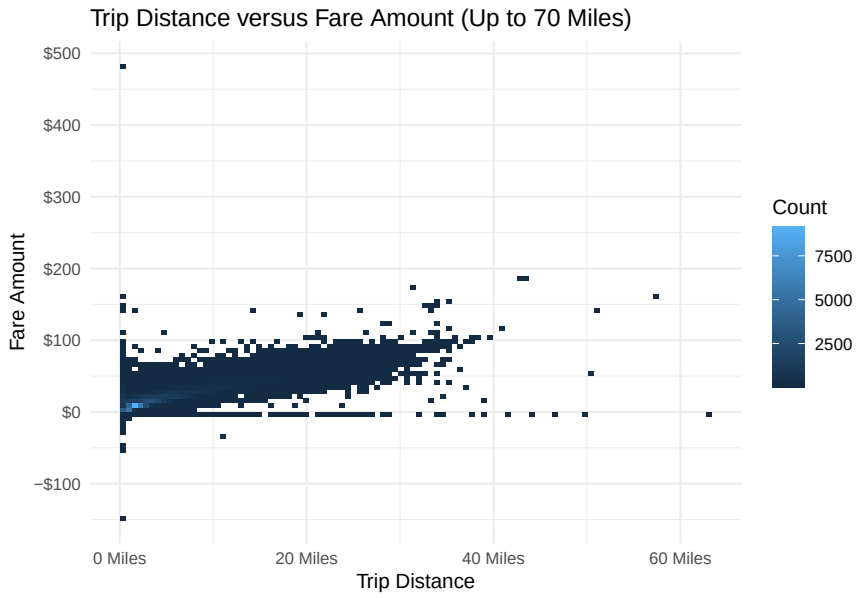
All of the values 32518 come from the other vendor, so data needs to be better collected on that vendor.

1.5 Extreme Data Issues





- There are distances travelled as large as 270,000 miles - this is likely poorly entered data as most of the extreme values come from the other category, which has an unknown automatic forwarding flag.
  - Data only really needs to be collected up to around 40 miles as this is the maximum normal distance journey you can expect a New York taxi or its competitotrs to travel, and this is reflected by the scatter density of outliers for the pink taxi and uber categories once filtered to 40 miles.
  - We restrict the data to up to 40 miles in the cleaned data for analysis.



- There is a single outlier fairs of over $200 for values of around 0 Miles, while also unexpected fairs of $50 or lower for distances over 10,000 miles - which is clearly wrong intuitively. There must be mistakes in how the data for fare amounts and distances are collected, or the wrong data is being collected to include non-standard ‘taxis’.
  - Most fairs in the 70 mile range are between -$200 and $200, with a few outliers above and below, so restricted the data to fares between -$200 and $200 for analysis. We retain 0 mile and negative fairs as these may have been disputes or voided trips.

2 Data Analysis Report

Important Note: This analysis is conducted using only July 2021 data, as only data from late June to early August was available. We chose to consider only July 2021 data so as to present the data over a well-defined non-arbitrary time frame. A natural way to extend the estimate to the analogous yearly results would be to multiply the presented July values by 12, however this estimates accuracy would depend on July's similarity to the other 11 months, such as March or December. To do so reliably, it must be the case that 'on average' a year is balanced out[1] so that July is a typical representative of a typical month in a year. We have no intuitive reason to believe July could be representative of the rest of the year on average so cannot make this assumption - in fact, you would expect months such as July and December to be atypically busy in New York due to tourism and extreme weather, that is, external factors will be influencing demand, traffic flow, and so on for the data we have available. Hence, we leave the analysis results for the month of July only.

We conduct the analysis on data from distances of up to 70 miles as this is taken to be the typical competition radius of interest for a New York taxi provider.

2.1 Assumptions

We will take an estimate cost-per-mile for our non-Green taxi data to be that of a typical diesel car, roughly \$0.20 - less than 5% of Ubers are electric, and we will assume the other competitor is also not green. We will also take an estimate cost-per-mile for Green taxis to be that of a typical electric car, roughly \$0.03.

Hence, profit per trip is computed as, where $\alpha_{i=1}$ would be \$0.03 for Green taxi, or $\alpha_{i=2}$ and $\alpha_{i=3}$ would be \$0.20 for Uber or other,

$$\begin{aligned}\text{Profit} = &\underbrace{\text{Total Amount}}_{\text{Revenue-Per-Trip}} \\ &- \underbrace{\left(\alpha_i \text{Miles} + \text{Congestion Surcharge} + \text{Tolls Amount} + \text{MTA Tax} + \text{Improvement Surcharge}\right)}_{\text{Cost-Per-Trip}}.\end{aligned}$$

2.2 Average Weekday Expenditure

Table 2: Average Amount Spent on Taxi Trips with Green Taxi On A Weekday in July 2021

Weekday	Average Spend
Sunday	\$18.41
Monday	\$17.76
Tuesday	\$17.77
Wednesday	\$17.78
Thursday	\$17.99
Friday	\$17.39
Saturday	\$16.28

We see the more was typically spent on Sunday on average, with a value of \$18.41.

[1] Aside: For reference, making this is type of estimate is formally called an ergodic regime..

2.3 Feature Effect on the Tip Amount for Green Taxi

We only consider non-zero dollar tips to study the influence of features on the tip amount. We also independently study the effect of the features on tip or no tip. As the tip data only includes the payments with credit card, we only consider the credit card payment type.

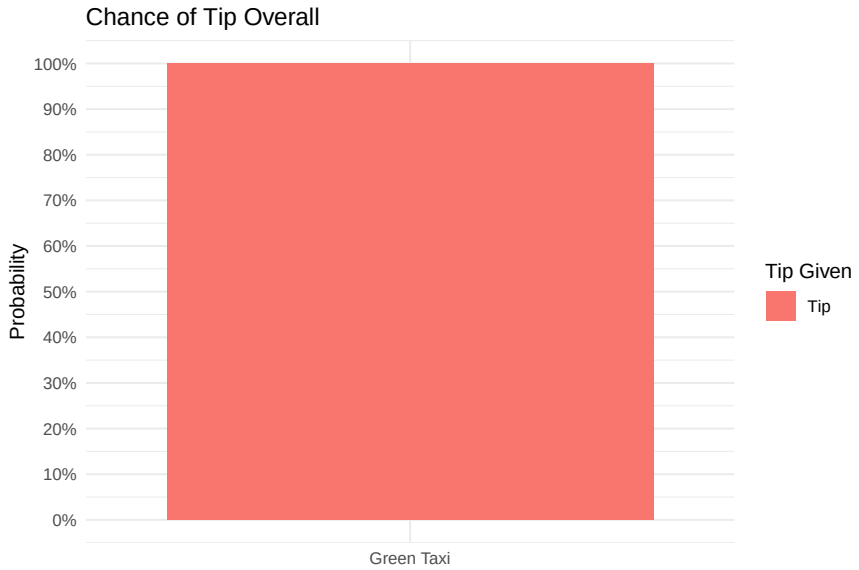
2.3.1 No Feature Prediction

Below we present how we would make a prediction of the tip amount without use of any features, as a benchmark of the relative performance of our features.

Table 3: Mean Tip Amount in July 2021

Mean Tip Amount ($)
3.16

Without using a feature we would guess a tip amount of $3.16 for any given journey.



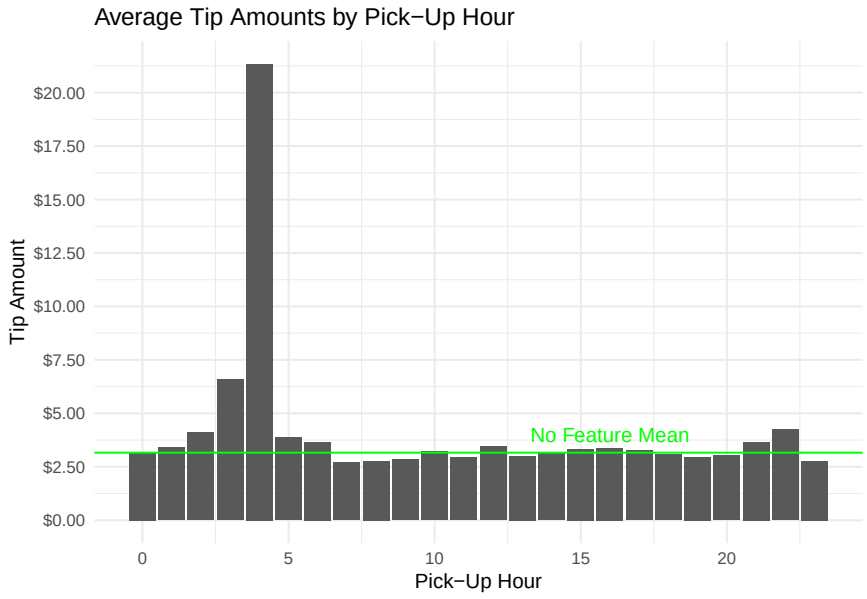
We see there is around a 62% chance of being given a tip without using a feature to support our guess, which means predicting a tip is only slightly more likely than a coin-toss without using a feature to support the prediction.

2.3.2 Pick-Up Hour

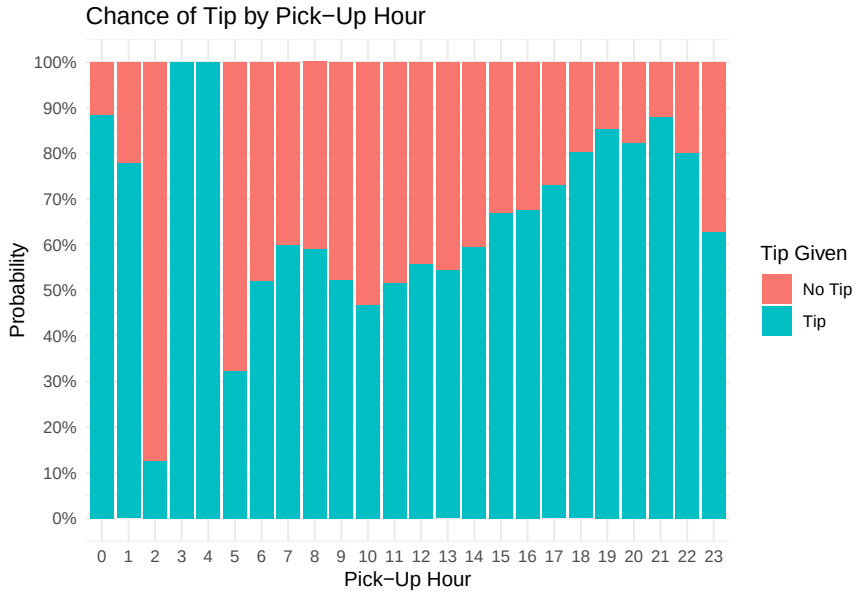
Table 4: Mean Tip Amount in July 2021 by Pick-Up Hour

Pick-Up Hour	Mean Tip Amount ($)	Group Size
0	3.20	23
1	3.39	7
2	4.10	1
3	6.58	2
4	21.35	2
5	3.87	10
6	3.65	67
7	2.69	152
8	2.78	206
9	2.84	167
10	3.21	165
11	2.94	171
12	3.44	180
13	3.01	170
14	3.17	169
15	3.34	210
16	3.37	173
17	3.26	202
18	3.07	226
19	2.96	185
20	3.03	148
21	3.65	88
22	4.23	56
23	2.77	27

We cannot make conclusions on the hours of 1:00 to 6:00 as the sample size is too small.



We see from the above bar plot that the early hours of the morning 2:00-6:00 will have higher tips on average than our featureless guess $3.16, however the sample sizes is too small for this to be considered more than random variation and so we cannot make a conclusion from this - more data must be collected. Outside of these early hours, the tips across all hours are not too distant from our featureless average.



For the hours of midnight to 6:00 we cannot make a conclusion about the predictive power of a tip or no tip, as the sample sizes are far too small in these hours and more data needs to be collected. In the evening, there is an over 70% chance of receiving a tip versus a 62% chance estimate if we did not use this feature. From 7:00 to 17:00 this feature does not give us much more predictive power on whether or not a tip will be given than our featureless guess 62% probability as these values are only between 58% and 70%.

Hence, it follows that during daytime hours, a feature of pick-up hour would be a poor predictor of tip amount in general as it is neither good at predicting whether a tip will be given or not, and does not predict a value much different from a featureless average tip anyway. From 17:00 to 23:00 this feature is good for predicting whether or not a tip itself will occur, however the tip amount predicted is no better than a average estimate without the feature. In the early hours of the morning we don't have large enough sample data to make a conclusion.

In summary, this feature is not a useful predictor of tip amount - as during the daytime hours it is poor at predicting whether or not a tip will happen or improve the estimate beyond the overall tip average itself. Outside of daytime hours, it will predict whether or not a tip will occur but not give an estimate better than the overall average.

2.3.3 Payment Type

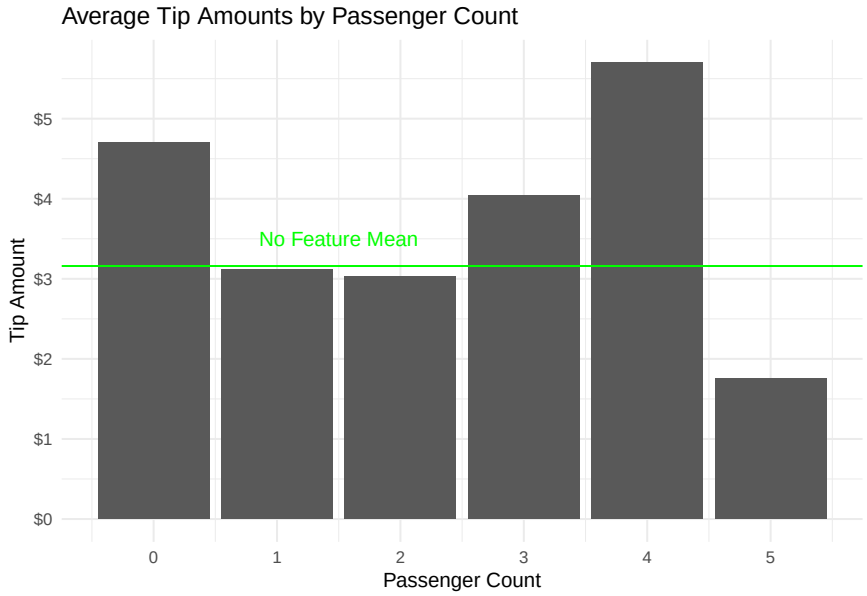
It is impossible to know if payment type has any impact on the tip amount given that the tip amount data only tracks tips made on credit card. More data must be collected on cash tips.

2.3.4 Passenger Count

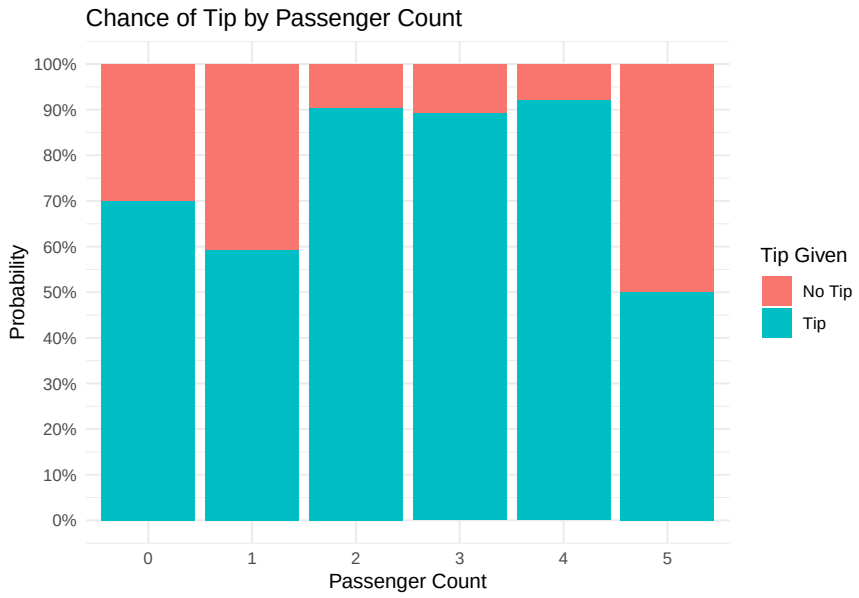
Table 5: Mean Tip Amount in July 2021 by Passenger Count

Passenger Count	Mean Tip Amount ($)	Group Size
0	4.70	7
1	3.12	2339
2	3.03	363
3	4.04	74
4	5.70	23
5	1.75	1

We cannot make conclusions for 0 and 5 passengers as the sample size is too small.



The bar plot above shows that the average tip by passenger count is noticeably different than for our featureless mean for 3 and 4, making it a useful predictor of the tip amount given. On the other hand, for 1 and 2 passengers it predicts similarly to the featureless mean. As a majority of trips are 1 or 2 passengers, this makes it a poor predictor of tip amount for the typical trip. While 0 and 5 passengers vary significantly from the featureless mean, the sample size for these is too small to conclude anything other than random variation.



The above graph indicates that 2 to 4 passengers are almost surely going to give any kind of tip with a probability of around 90%. For 1 passengers, the probability of receiving a tip is between 60% and 70% which is not much different from our featureless estimate of 62%. For 0 passengers (say, delivery) and 5 passengers we don't have enough data to draw a conclusion.

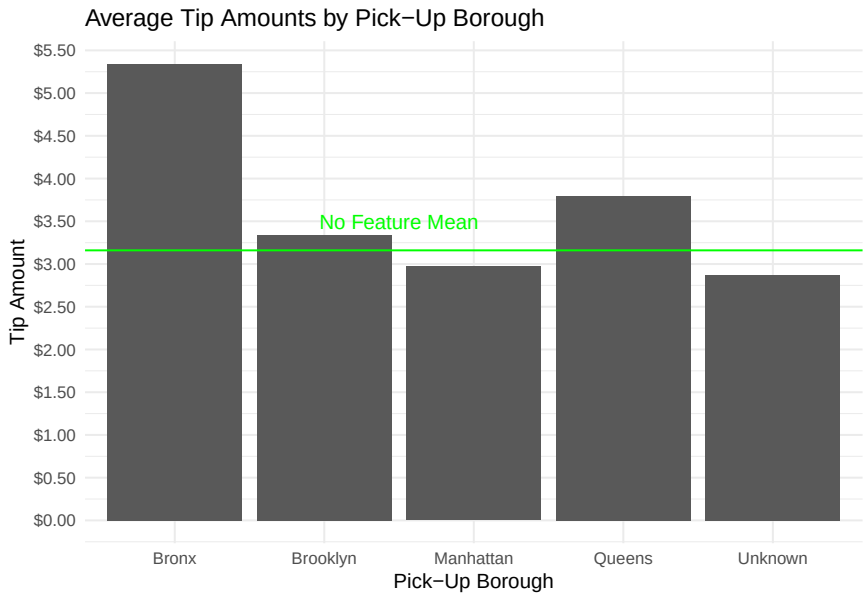
Overall, passenger count is a useless predictor of tip amount for single passenger journeys, as it provides almost the same information as a featureless estimate. For 2-4 passenger journeys it is a useful feature as it tells you to expect a tip, and although for 2 passengers it provides the same estimate - for 3 and 4 passengers provides a better estimate than a featureless estimate. Hence, it is a good feature to predict tip amount taken as a whole.

2.3.5 Pick-Up Borough

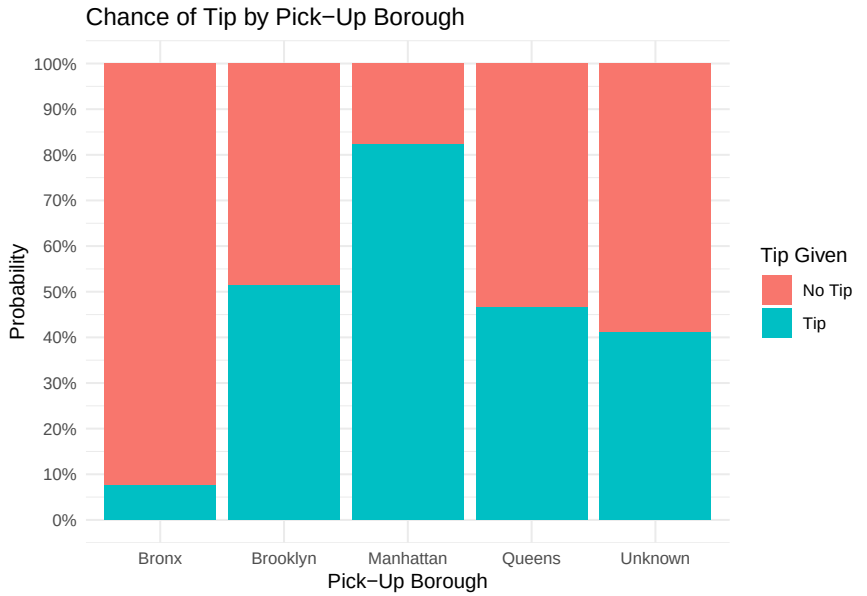
Table 6: Mean Tip Amount in July 2021 by Pick-Up Borough

Pick-Up Borough	Mean Tip Amount ($)	Group Size
Bronx	5.34	28
Brooklyn	3.34	656
Manhattan	2.97	1862
Queens	3.80	254
Unknown	2.87	7

We cannot make conclusions for unknown borough as the sample size is too small, however as we do not know the borough this would not help anyway.



We see that the average tip does not vary significantly by pick-up borough from the featureless tip amount estimate $3.16, with the exception of the Bronx which estimates $5.34.

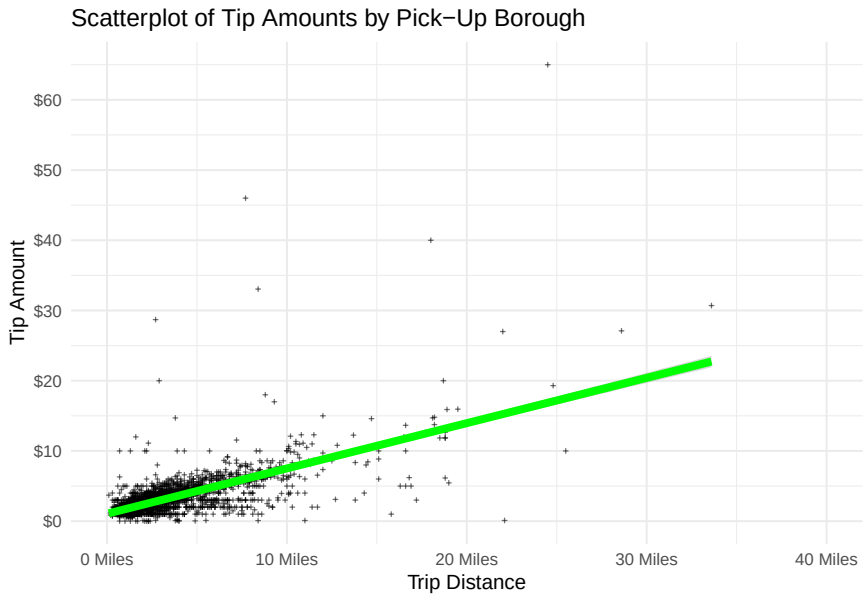


Moreover, we see that the probability of the pick-up borough Bronx leading to a tip is less than 10%, so this different in estimate discussed above is irrelevant, but this is useful to determine that a $0 will likely be given. We see aside from Manhattan and the Bronx, this feature provides no more information in predicting whether or not a tip will occur than a coin toss. In Manhattan, it tells you there is an 80% chance a tip will occur, even though we know from above this tip amount, $5.34 will be in the region of our featureless estimate $2.97.

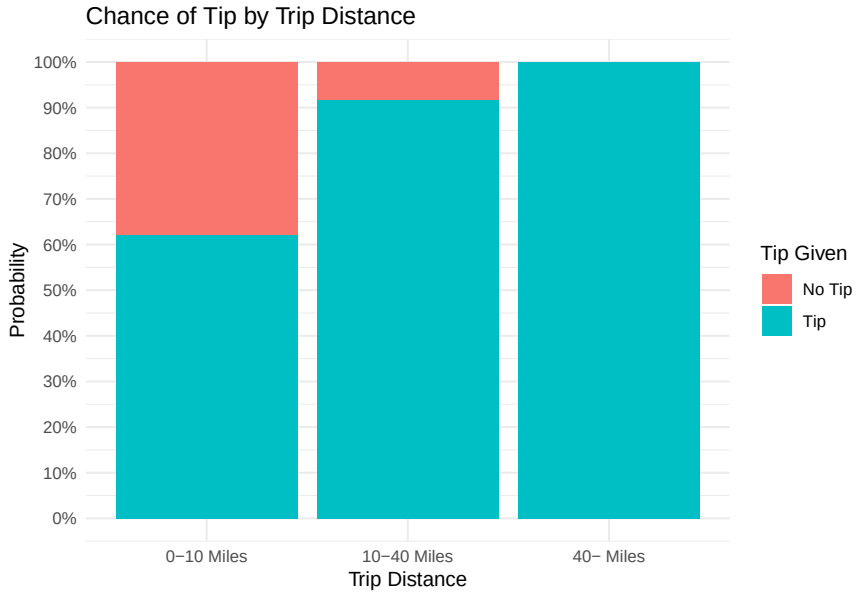
Overall, Borough is only useful for providing tip amount information on if a tip will occur in Manhattan and the Bronx boroughs.

2.3.6 Trip distance

We include only non-zero trip distances to avoid cancellation charges and so on, and restrict the visualisation to 40 miles as this is where a majority of the data plotted lies.



For less than 10 miles, we see that trip distance appears to follow a linearly increasing relationship with tip amount, a 1 mile increase in trip distance corresponding to around a $0.75 increase in tip. As distances increase beyond 10 miles, our tip amounts are increasingly distant from the line of best fit, making it a poor tip amount predictor here.



We see that for 0-10 mile trips there is a roughly 62% chance of a tip being given, the same as guessing based on no features, which makes sense as this is where most of our data lies so should reflect the overall

probability. We see from 10+ miles there is a 90%+ probability of being given a tip, however as stated above, the amount predicted will likely be wrong for these values. Given the discussion, trip distance is a good predictor for a majority of trips (up to 10 miles) if paired with another feature that more accurately predicts whether or not a tip will occur.

2.4 Accumulated Profits in 2021

Table 7: Total Profits Per Company in July 2021

Company	Total Profit
Green Taxi	$120,790
Uber	$730,046
Other	$987,781

We see that the other vendor has accumulated the most profit, with a value of $987,781, followed closely by Uber with $730,046. Green taxi had significantly less profit with $120,790 for July 2021.

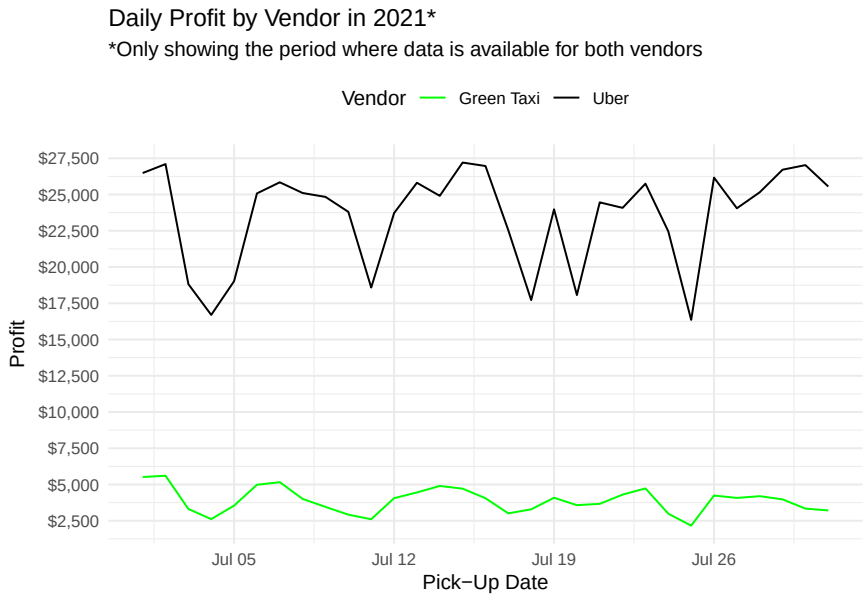
2.5 Total Number of Rides

Table 8: Total Number of Rides Per Company in July 2021 (Non-Voided Trips)

Company	Number of Rides
Green Taxi	7,619
Uber	43,530
Other	32,288

Uber and the other vendor have significantly more non-voided rides with values of 43,530 and 32,288 respectively in July 2021, than Green Taxi which has 7,619.

2.6 Profit Change Over 2021

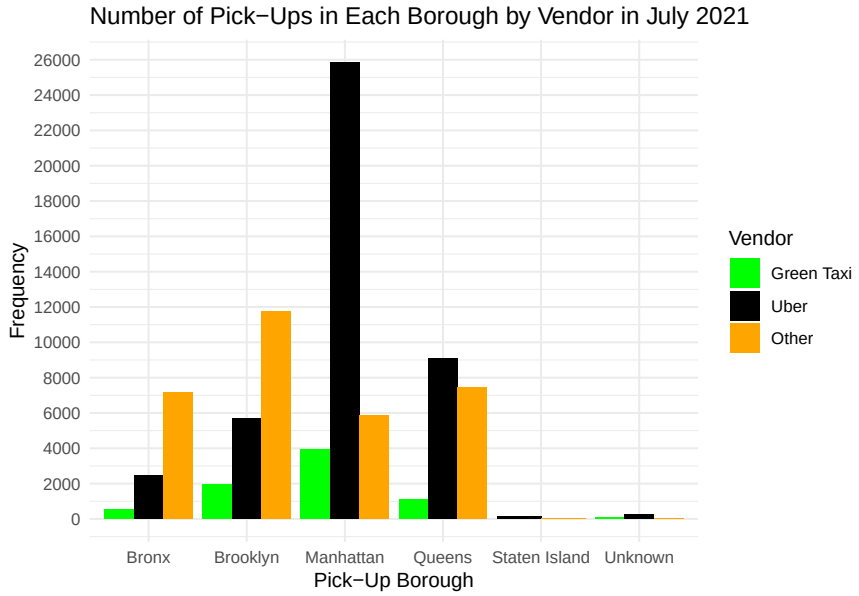


We see that the profit amount for Uber is more variable than Green Taxi, but both Green taxi and Uber both have cyclical daily profits that are seemingly correlated with one another - both Green Taxi and Uber see their profits rise and fall simultaneously with what is likely overall taxi demand in New York City. We also see that Green taxi has significantly lower daily profits than Uber, which makes sense as it makes less trips.

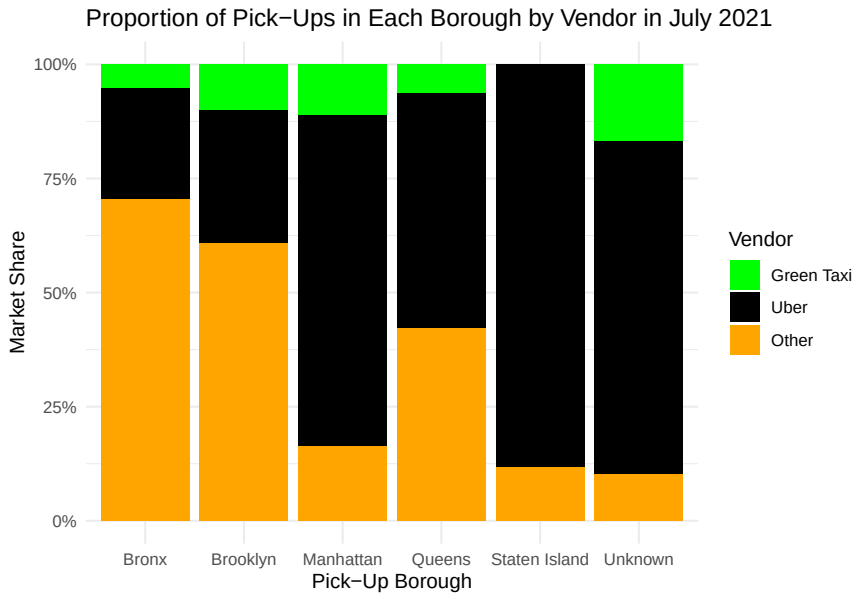
2.7 Trip Type Revenues

We see that inner city trips brought significantly more revenue to Green Taxi than outer city ones, and make up a majority of the revenue.

2.8 Most Popular Pick-Up Borough

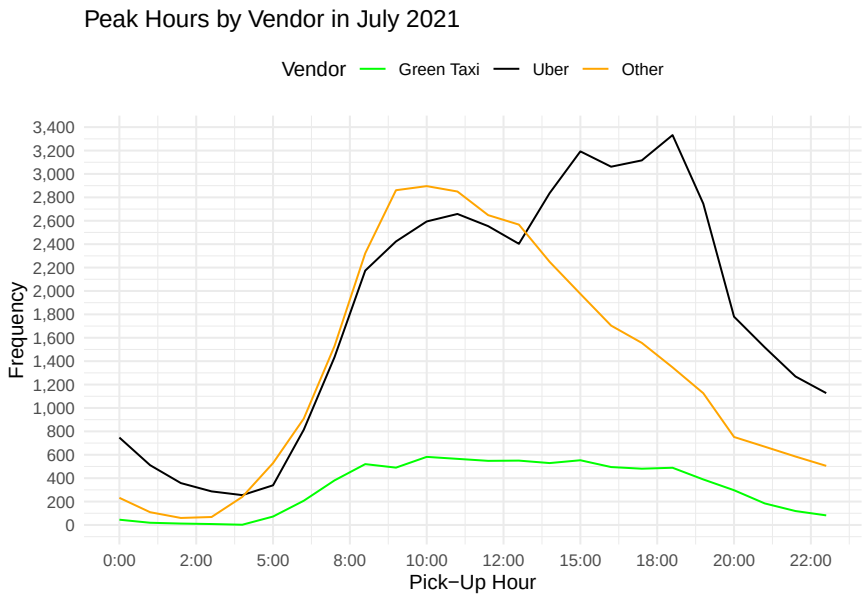


Manhattan is the most popular pick-up borough for both Green Taxi and Uber, while Brooklyn is the most popular pick-up borough for the other competitor.

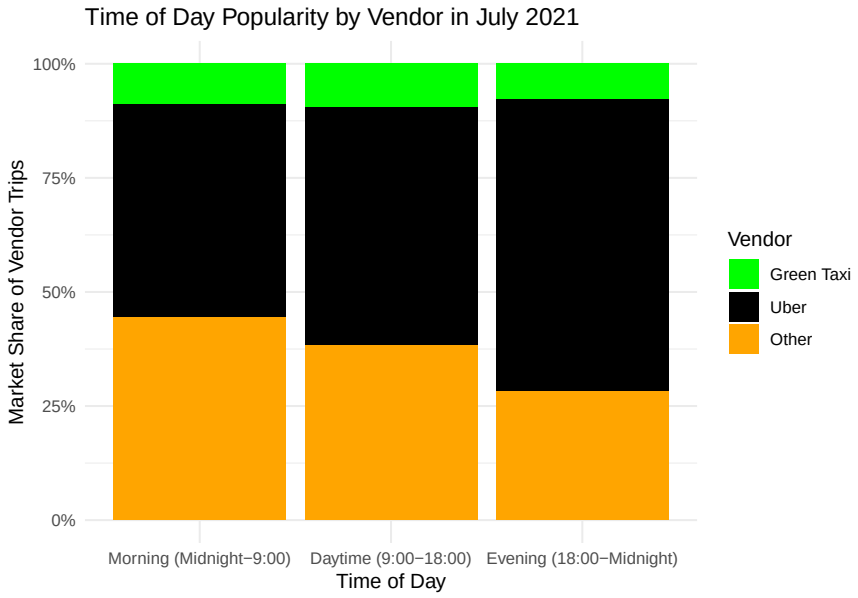


In the Bronx and Queens, Green Taxi has a much smaller market share than its competitors, with the other competitor taking most of the market share in the Bronx. Green taxi also does not compete on Staten Island where Uber takes most of the market share.

2.9 Peak Hours

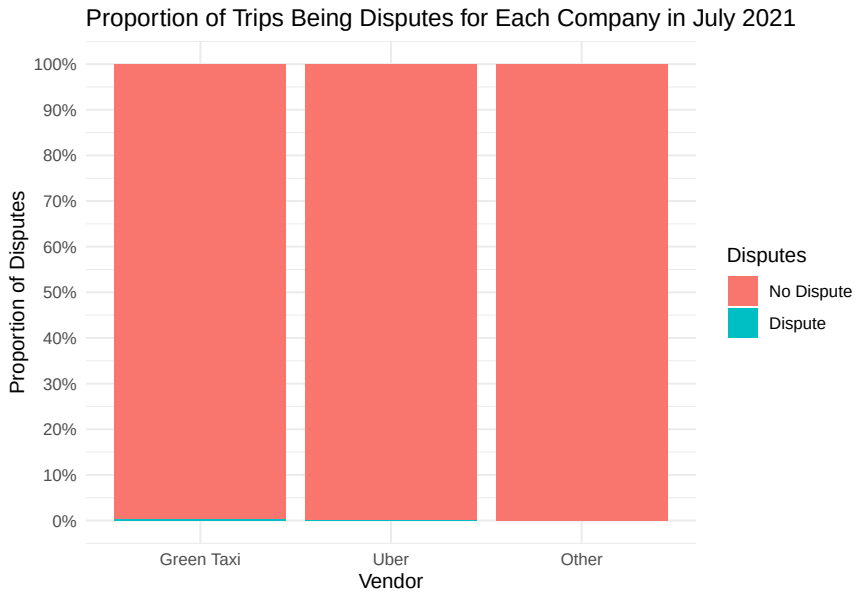


Green taxi has its peak hours ranging between , while the other competitor has its peak hours around 10:00, and Uber has its peak hours from 15:00 to 19:00.



Green Taxi is equally active throughout the day, while Uber is most active in the evening hours and the other competitor is most active in the morning hours.

2.10 Disputed Trips



Green Taxi has the most disputed trips in a 70 mile radius of New York City, however it is still a tiny proportion of the total trips at less than 1%.