Green Taxi:
Competitor Market Analysis

Assessed By Barbara Mikulášová

Report Generated By User
Written By James Wright

Report Generated On
Friday 09 June 2023

Date First Published
Friday 09 June 2023

Contents

1 Data Quality Report **3**

1.1 Metadata Quality Issues . . . 3
1.2 Duplication Issues . . . 3
1.3 Unexpected Data Value Issues . . . 4
1.4 Missing Data Issues . . . 5
1.4.1 Count of Columns with Missing Values . . . 5
1.5 Extreme Data Issues . . . 6

2 Data Analysis Report **8**

2.1 Assumptions . . . 8
2.2 Average Weekday Expenditure . . . 8
2.3 Feature Effect on the Tip Amount for Green Taxi . . . 9
2.3.1 Pick-up hour . . . 9
2.3.2 Payment type . . . 10
2.3.3 Passenger count . . . 11
2.3.4 Pick-up Borough . . . 12
2.3.5 Trip distance . . . 13
2.4 Accumulated Profits in 2021 . . . 15
2.5 Total Number of Rides . . . 15
2.6 Profit Change Over 2021 . . . 16

2.7 Trip Type Revenues 16
2.8 Most Popular Pick-Up Borough 17
2.9 Peak Hours 18
2.10 Disputed Trips 19

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

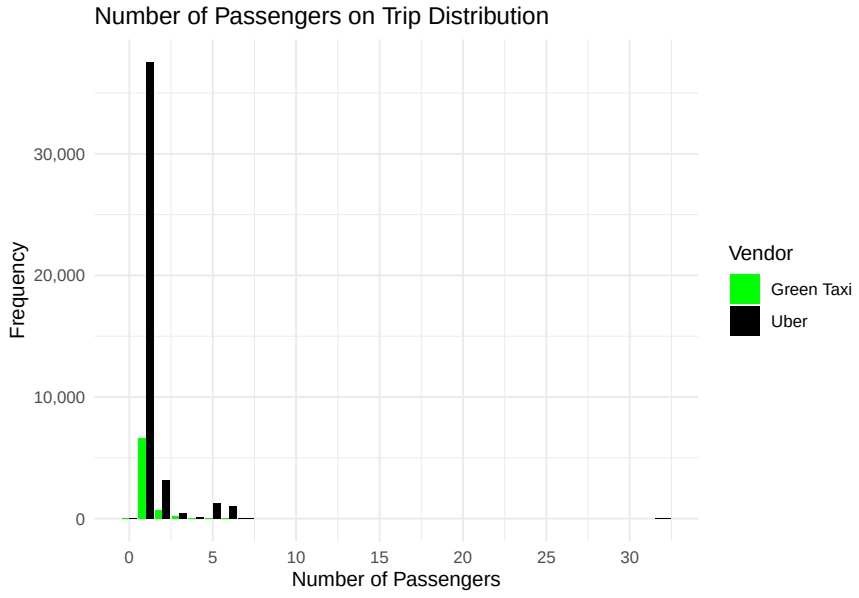
Summary of problems and corrections of duplicates:

- 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.

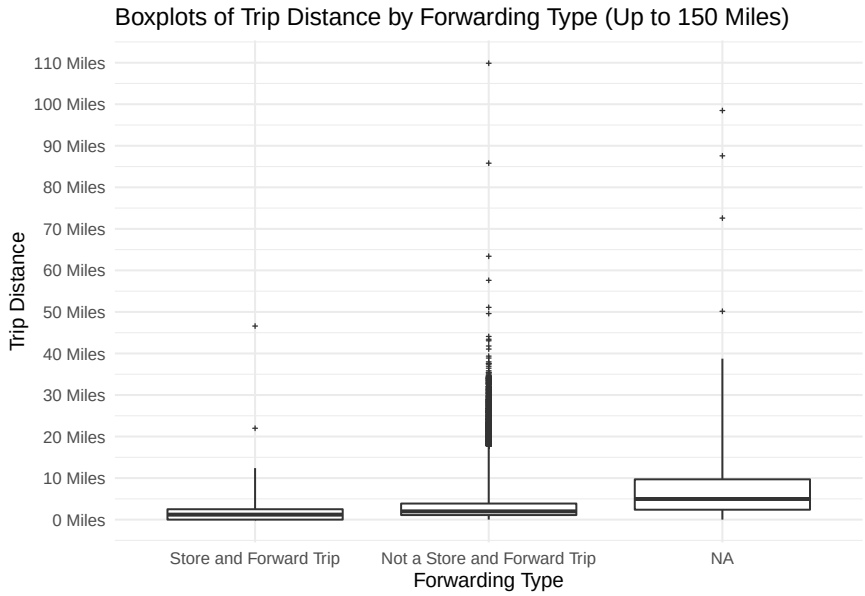
1.4.1 Count of Columns with Missing Values

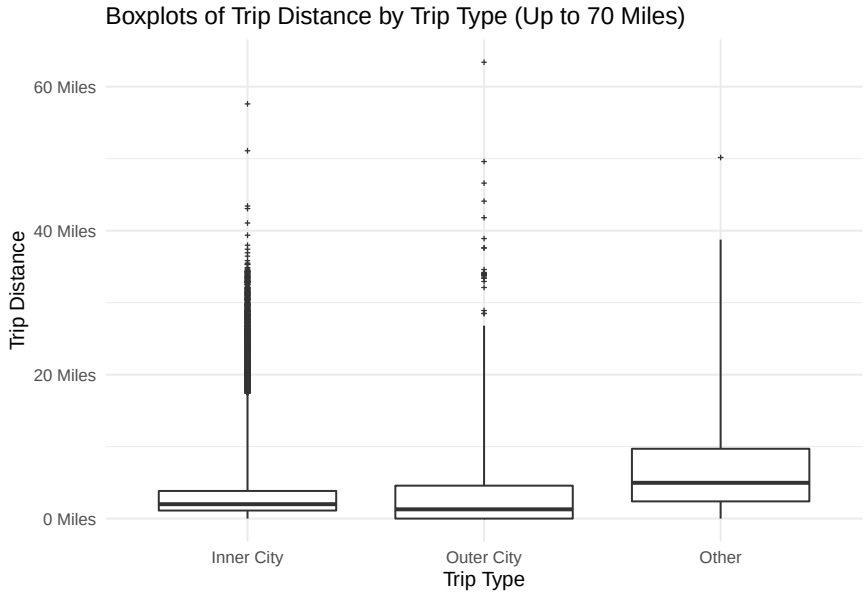
Table 1: Count of Columns with Missing Values

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

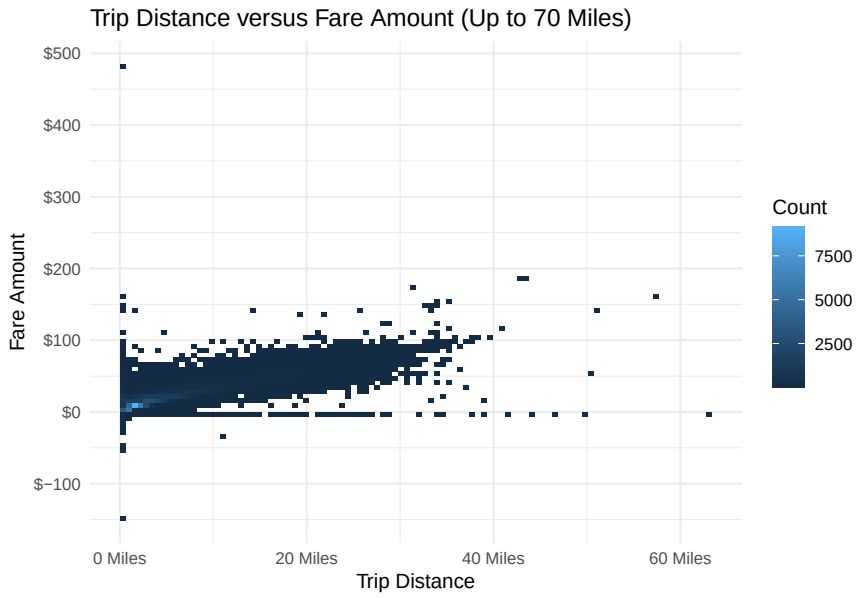
All of the values 32518 come from the other vendor, so data needs to be better collected on that vendor.

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

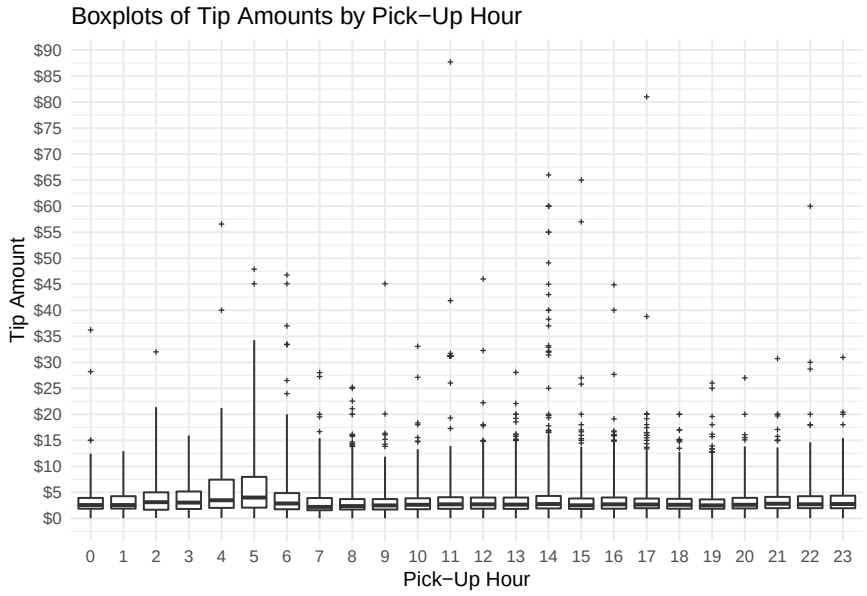
We see the more was typically spent on Sunday on average, with a value of \$18.41.

[1] Formally, this is called an ergodic regime.

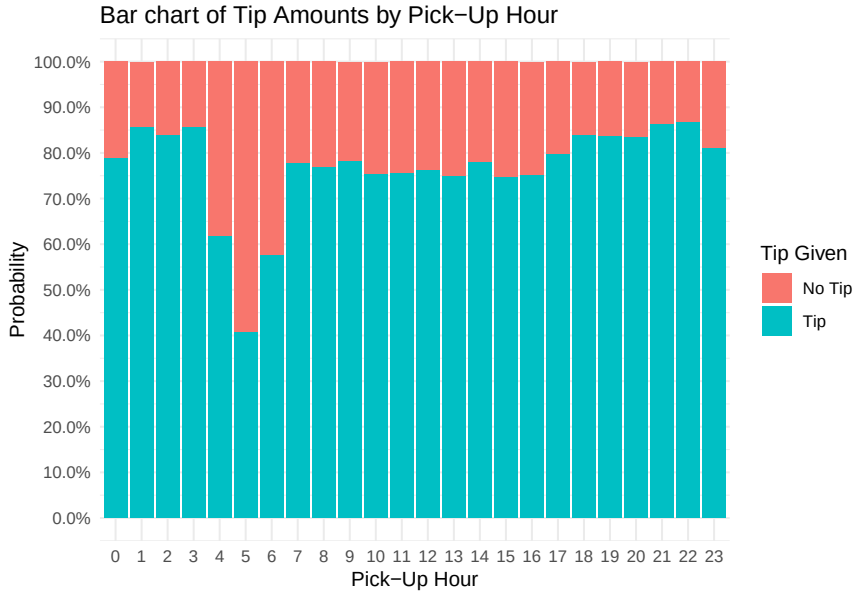
2.3 Feature Effect on the Tip Amount for Green Taxi

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 Pick-up hour



We see from the above boxplot that the early hours of the morning will have higher tips on average, and these tips have greater variability - you are more likely to receive a larger tip in the early hours of the morning. In general, there is a lot of variability in possible tip values for a given pick-up hour.



From 4:00 to 7:00 there is a lower probability to receive a tip than normal - from 40% to 60%, even though the tips given are larger if they are. Throughout the rest of the day, there is a 70%-80% chance of being given a tip.

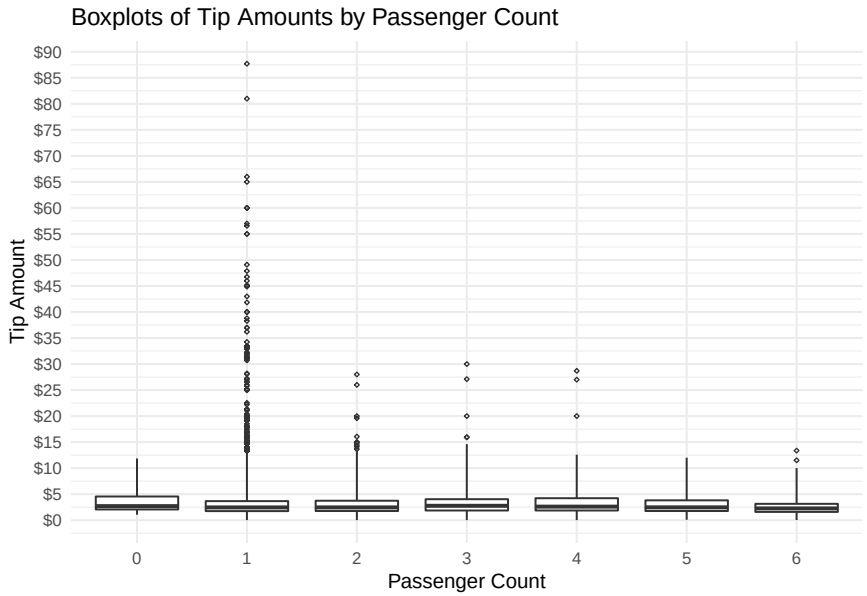
Hence, it follows that pick-up hour is a poor predictor of tip amount for the early morning due to high variability of the amount for a given hour. In addition, the roughly 50%-60% chance of a tip in the early morning also means there is little information gained about the chance of receiving a tip relative to just flipping with a coin in those hours. The rest of the day, this predictor is good for deciding that a tip will occur - telling yout that there will be an 80% chance. For the rest of the day, it is a somewhat poor predictor of the value of a given tip, due to the high variability at the extremes, however 75% of those values are in a small range around $2.50 for those hours - so for a given non-early hour you can be fairly sure 75% of customers that give a tip will give a tip somewhere between $2 and $4.

2.3.2 Payment type

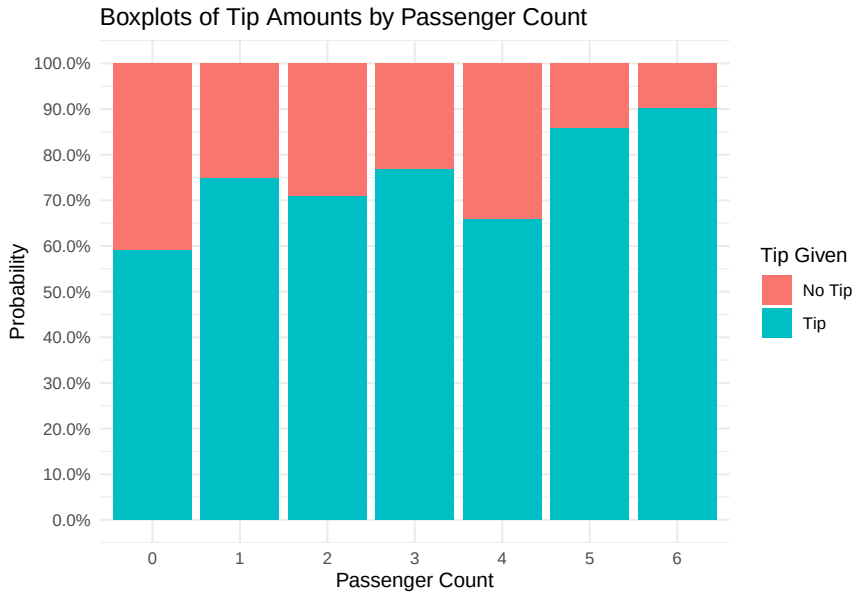
It is impossible to know if payment type has any impact on the tip amount given that the tip amount data only tracks tips made on credit card. More data must be collected on cash tips.

2.3.3 Passenger count

We excluded data from other vendor in the analysis as their passenger accounts are entirely unknown.



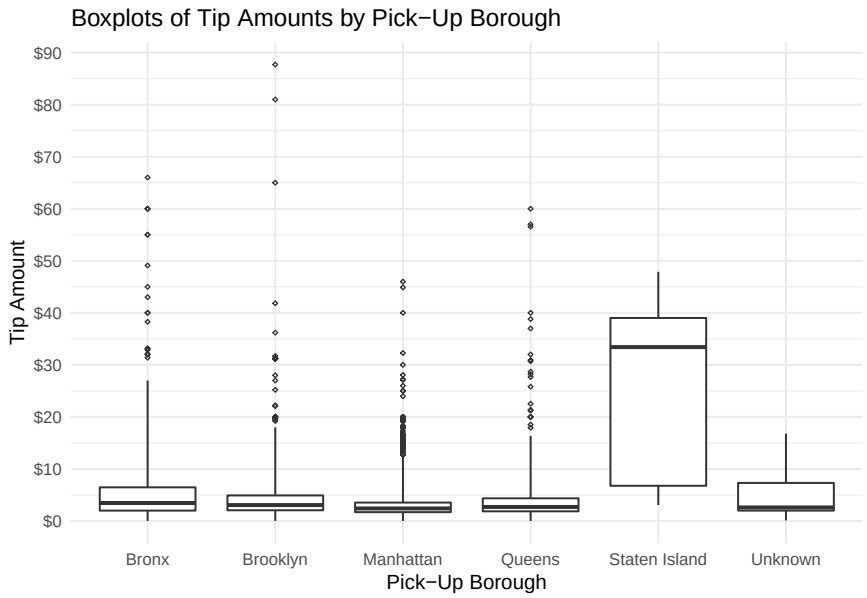
The boxplot above shows that the average tip is likely to be around $2.50 regardless of the number of passengers, and similarly the typical range of tips is similar - roughly 75% of the tips are between $1.25 and $3.75 for 1-6 passengers. For zero passengers (say, deliveries), there is a 75% chance the tip is roughly between $2 and $5. Lower number of passengers are more likely to give larger extreme tips, with a relatively dense 'cluster' of extreme values even around $50 for 1 passenger.



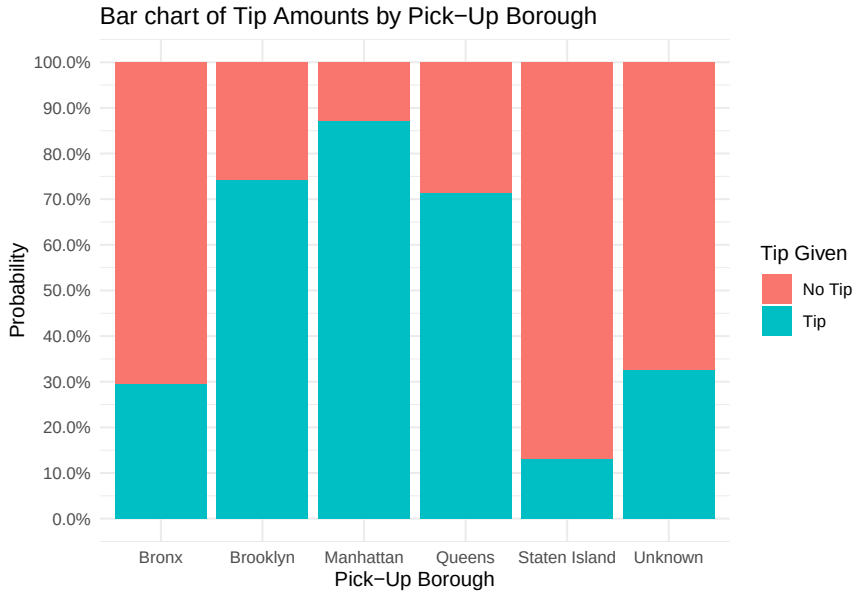
The above graph indicates that 5 or 6 passengers are almost surely going to give any kind of tip with a probability of 85% to 90%. For zero passengers, the probability of receiving a trip falls to 60%. Otherwise, the probability of receiving a tip is around 70%.

Overall, passenger count is only a weak predictor of tip amount for low numbers of passengers, as while there is little variability in the amount given for 75% of tips given, this predictor only improves the chance of receiving a tip for 0-4 passengers slightly above a coin-toss. It is more useful to determine the chance of any tip occurring if you have a 5 or 6 customers, where you would expect a roughly $2-$5 tip with roughly 70% certainty.

2.3.4 Pick-up Borough



We see that the average tip varies by pick-up borough, with Staten Island giving the highest tip of around $35. Staten Island also has extreme variability of the tip amount with 75% of tips laying somewhere between $7.50 and $40, and more generally the tip amount variability changes a lot between borough.

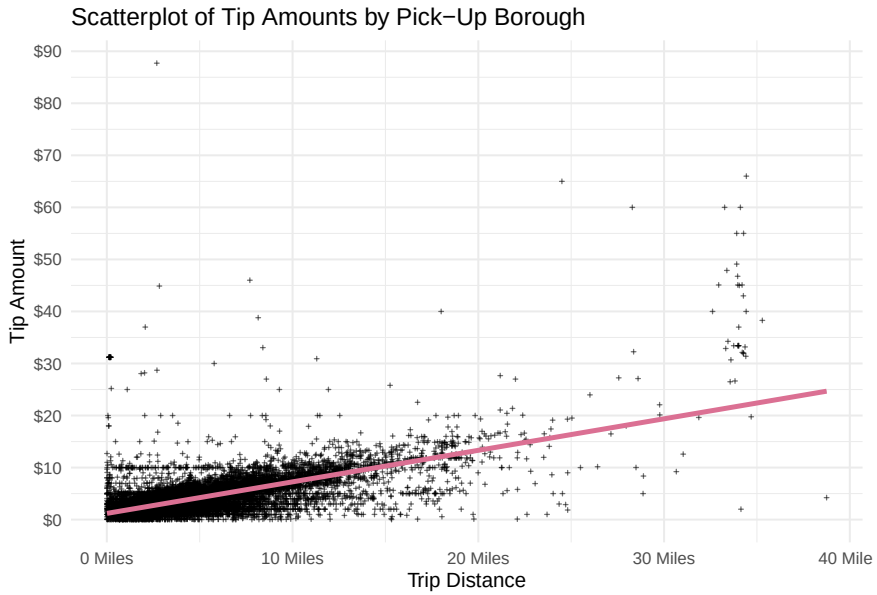


We see that the probability of receiving a tip is highly dependent on the borough, with Staten Island and the Bronx only tipping 12% and 30% of the time respectively, while Manhattan tips 85% of the time and Brooklyn and Queens tip over 70% of the time. As the probabilities are either all far above or below 50%, this makes borough a good predictor of whether there is no tip or a tip - picking up from a given borough you can be fairly sure you will or won't get a tip for that trip. On the other hand, the variability in tip amount, particularly for Staten Island, means that knowledge of the expected tip amount given - if it is given - is poor. The situation is better for Brooklyn, Manhatten, and Queens, where this predictor is capable of both telling you that you would receive a tip *and* 75% of the time that tip will be between $2 and $5.

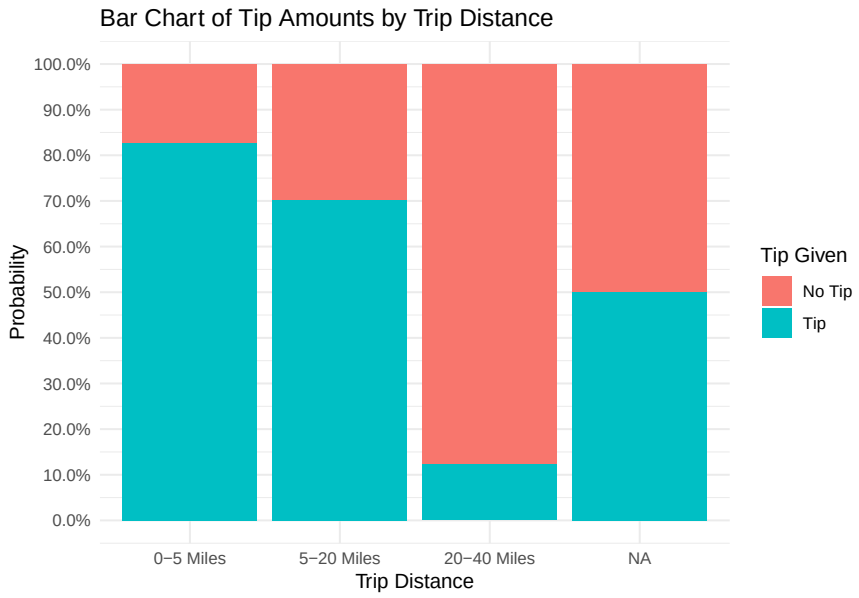
Overall, Borough is useful for predicting if a tip will occur in a given borough, but only useful for predicting the tip in 3 boroughs.

2.3.5 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.



Trip distance appears highly correlated with tip amount, a 1 mile increase in trip distance corresponding to around a $0.75 increase in tip.



We see that for 1-5 mile trips there is a roughly 80% chance of a tip being given up to 20 miles, making the short trip distance a good predictor to use. There is a 20% chance of a tip over 15 miles, which itself is useful as you expect longer trips to therefore yield no tip. Trip distance is therefore an excellent predictor of tip amount, as it both predicts the value of the tip accurately and the chance of the tip from 0-20 miles well, which is a majority of the trips.

2.4 Accumulated Profits in 2021

Table 3: 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.

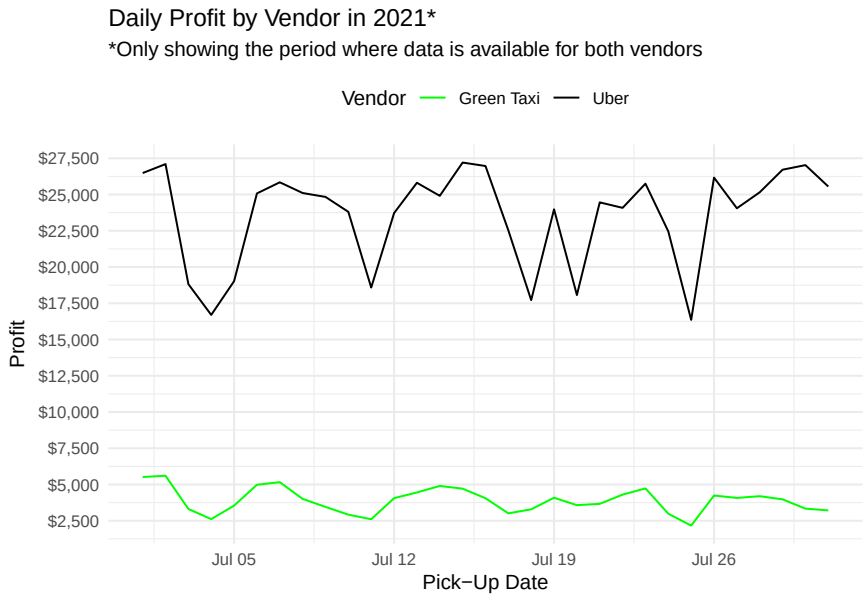
2.5 Total Number of Rides

Table 4: 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

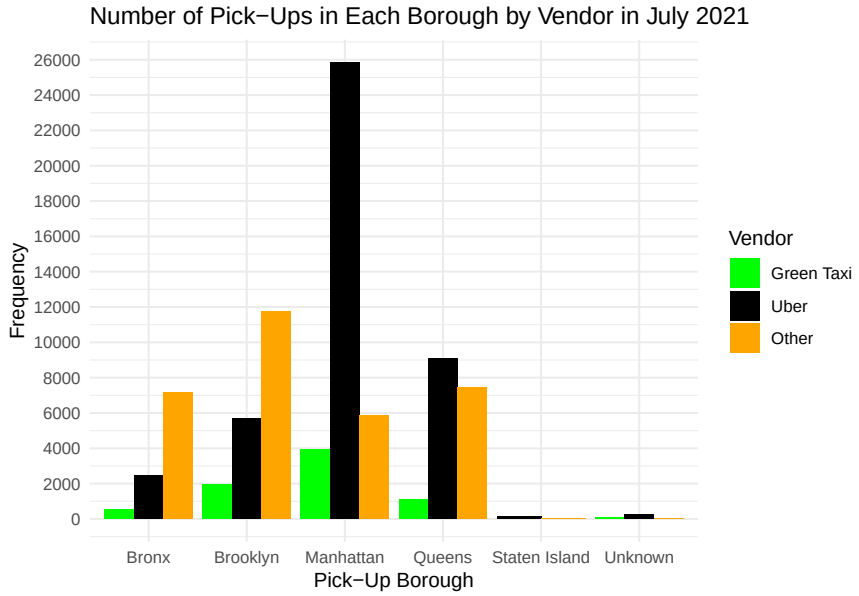


We see that the profit amount for Uber is more variable than Green Taxi, but both Green taxi and Uber both have cyclical daily profits that are seemingly correlated with one another - both Green Taxi and Uber see their profits rise and fall simultaneously with what is likely overall taxi demand in New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

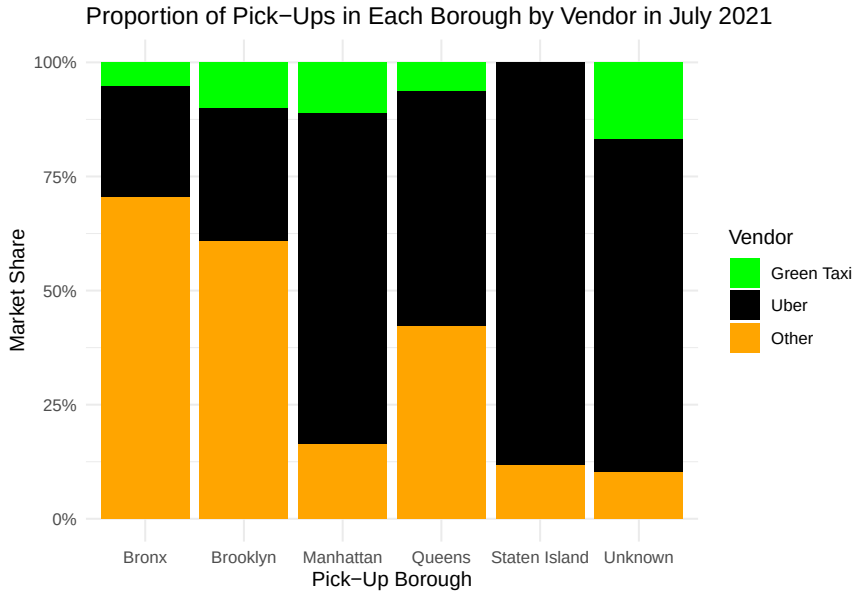
2.7 Trip Type Revenues

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

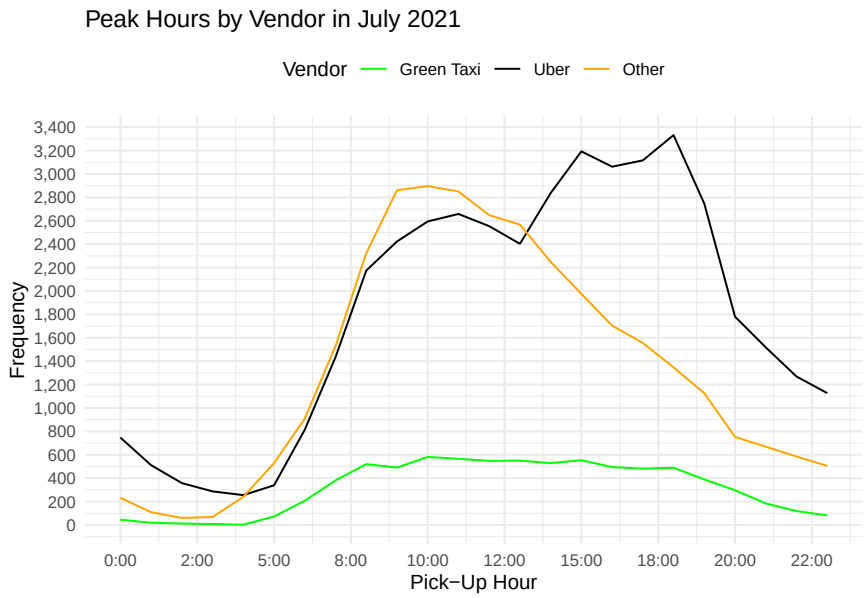


Manhattan is the most popular pick-up borough for both Green Taxi and Uber, while Brooklyn is the most popular pick-up borough for the other competitor.

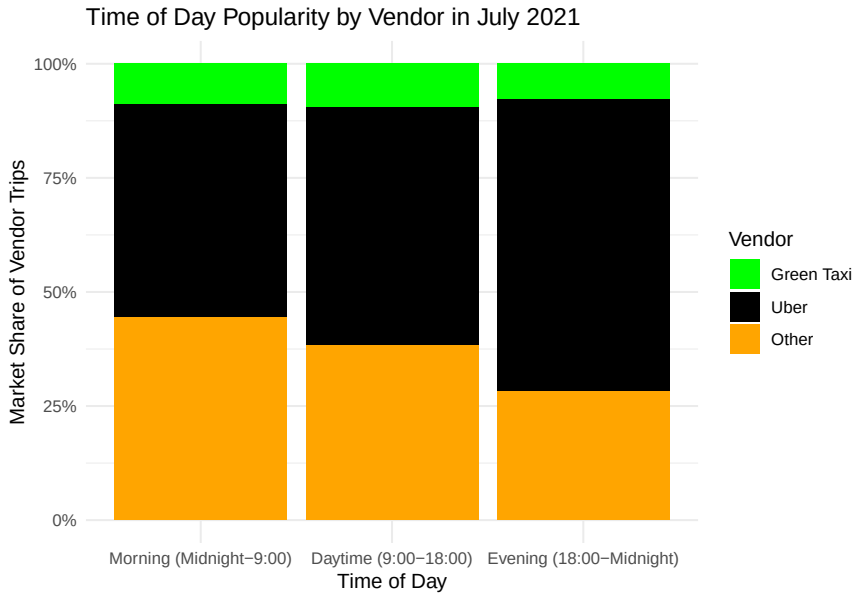


In the Bronx and Queens, Green Taxi has a much smaller market share than its competitors, with the other competitor taking most of the market share in the Bronx. Green taxi also does not compete on Staten Island where Uber takes most of the market share.

2.9 Peak Hours

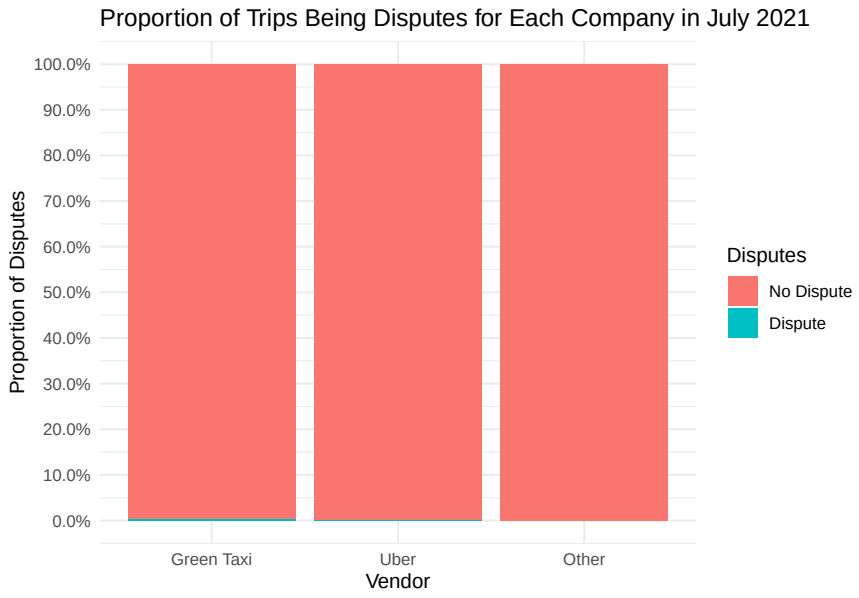


Green taxi has its peak hours ranging between , while the other competitor has its peak hours around 10:00, and Uber has its peak hours from 15:00 to 19:00.



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%.