NEW YORK TAXI ANALYSIS – YELLOW and GREEN – 2017

Brief intro about the taxi structure in New York :

Yellow Medallion Taxicabs - These are the famous NYC yellow taxis that provide transportation exclusively through street-hails. The number of taxicabs is limited by a finite number of medallions issued by the TLC. You access this mode of transportation by standing in the street and hailing an available taxi with your hand. The pickups are not pre-arranged.

For Hire Vehicles (FHVs) - These are community car services (livery), black car services, or luxury limo services. The number of FHVs is not limited by a finite number of licenses. FHV transportation is accessed by a pre-arrangement with a dispatcher or limo company. These FHVs are *not* permitted to pick up passengers via street hails, as those rides are not considered pre-arranged.

Street Hail Livery (SHL) - This class is a hybrid between yellow taxicabs and For Hire Vehicles that are permitted to additionally accept street-hails above 110th Street in Manhattan and in the outer-boroughs of New York City. Historically, these areas have had sparse yellow taxicab coverage as compared to Manhattan and livery cabs were often operating as "gypsy cabs", illegally accepting street hails. The SHL program will allow livery vehicle owners to license and outfit their vehicles with green borough taxi branding, meters, credit card machines, and ultimately the right to accept street hails in addition to pre-arranged rides. This plan was originally proposed by Mayor Bloomberg in 2011 and finally approved in June of 2013.

Aim of the analysis:

I have tried to do an exploratory analysis using SQL and Tableau to best summarize the data. In this analysis, I have tried to explore the dataset for the year 2017 for both yellow and green cabs. The focal point of the analysis is to explore and summarize the data and to give a holistic picture of what both these dataset holds. I have performed an independent analysis to see how the yellow and green cab markets are and the comparison between the two types cannot be complete as tehey don't serve the same population always and hence I opted to analyse independently for both types.

The analysis components consist on two parts:

1) SQL analysis using Google Bigquery for both Yellow and Green cabs for the year 2017. The queries used to perform analysis are summarized and drafted in the document attached.
2) Build a dashboard using Tableau to show the key metrics identified during EDA and also adhoc during the dashboarding which can provide a better picture of the data to the business users.

YELLOW TAXI ANALYSIS:

These are the most famous and the oldest of taxi services in New York.

Yearly summary Statistics for the Yellow Taxi:

Total number of trips:
113 M

Total number of trips:
11 M

As expected, we can see that number of trips made by yellow cab is 10 times higher than green cabs because of the population it serves and the foot print it holds.

How many of these rides were in morning and how many were during the later part of the day ?

This is an important question to answer because, this helps the vendors to manage the demand and supply throughout the day. For this analysis, I have classified all trips between 6AM-6PM as day trips and others as eve/night trips.

Total number of trips:
Day: 70M
Night: 40 M

Total number of trips:
Day: 6.8M
Night: 4.8M

In both the cabs, the ratio between the day and night trips are almost similar with 60-40 split between day and night trips.

Drilling down further,

How many trips had tips and how many didn't ?

This is important beacsue, unlike Asian countries , Tip is an important parameter in America.

Trips with Tip: 73 M
Percentage: 64.32 %

Trips without Tip: 40M
Percentage: 35.68 %

Trips with Tip: 4.8 M
Percentage: 41.49 %

Trips without Tip:6.8M
Percentage: 58.51 %

As expected, we can see that about 65 % of the yellow cab trips had a tip! Interestingly, for green cabs, trips without tip is high with about 58%

How many passengers did the cabs serve?

Probably the most important factor. How many customers did they serve. At the end of the day , this determines the success of the business.

Total number of Passengers: 184 M

Total number of Passengers: 16 M

The yellow cab service has attended a whopping 184 million customers for 2017 whereas green has aided for 16 million customers.

Finally, the most important metric, earnings. How much did they clock for the year 2017 ?

Total Earnings: 1.8 B

Total Earnings: 167 M

Yellow cab has clocked a humongous 1.8 B for the year 2017 while green has managed to make 167M.

What about the tip amount for each can types ?

Total Tips : 208.5 M

Total Tips: 13.5 M

Yellow cabs has received about 209 M ie., 11.5% of their entire earnings via tip while the green cab has received about 8% of the earnings via tip. This might be because of the brand value associated with people or may be because the yellow cab drivers are more friendly so that the customers tip them more.

The above numbers give the overall statistics for the yellow and green cabs. We can now drill down further to see the per trip average for various metrics. This is important because the overall numbers will give only the top level details whereas the per trip analysis gives you more granular information about both these services.

Per trip average analysis:

We take into account , the following metrics for per trip analysis:

- Avg trip distance
- Avg Passenger count
- Avg fare per ride
- Avg tip per ride
- Avg time per trip
- Avg speed of the trip

Avg trip distance: 2.93m

Avg Passenger cnt: 1.63

Avg fare per ride: $16.34

Avg tip per ride: $1.83

Avg time per trip: 16.32 min

Avg speed of trip : 13.3 m/h

Avg trip distance : 2.68m

Avg Passenger cnt: 1.36

Avg fare per ride : $14.24

Avg tip per ride : $1.15

Avg time per trip : 20.38 min

Avg speed of trip : 14.3 m/h

This per trip analysis gives us more granularity into the data understanding how these services stand in the market.

Even though the numbers are close still yellow cab marginally outclasses the green cab in all categories except the average speed per trip. Again, higher average speed can either be good or bad depending on the scenario. But again, since they serve different population it will not be correct to compare the services directly. But on the whole, yellow cab services are in a way better than the green cab if we look at the numbers from per trip analysis.

Another metrics that might catch eyes is the payment type. Which is the most preferred payment type with respect to user.

Credit card : 76244327

Cash : 36479462

No charge : 599001

Dispute : 174081

Unknown : 3

Credit card : 5914813

Cash : 5740714

No charge : 54447

Dispute : 30151

Unknown : 542

If we see the above analysis, as expected credit cards and cash are the most preferred payment modes. In fact, more than cash people nowadays prefer paying via card. So it would be advisable for the both the vendors have credit card payment option available for all their services.

Especially in yellow cab service, the ratio to card to cash payment is almost double while in green cab its almost the same. This might infer that may be most of the people using yellow cab are working professionals while the green cab caters equally to both professionals and others as they operate in somewhat remote location when compared to the yellow cabs.

Another important factors is the percentage of disputes. Yellow cabs have 0.15% of total trips in dispute where as green has 0.3% even though they cater to less population. Green cab services providers have to have a look at why the dispute ratio is high so that they don't lose their customers.

Which pickup and dropoff location is in most demand?

This is another important question because this will help them to understand which location to have more services to.

Pickup location:

Zone:Upper East side south

Borough: Manhattan

Dropoff location:

Zone:Upper Midtown Center

Borough: Manhattan

Pickup location:

Zone : East Harlem North

Borough: Manhattan

Dropoff location:

Zone: East Harlem North

Borough: Manhattan

As far as the location details are concerned, the most demanding borough is Manhattan. Meaning most of the trips catering to Yellow and Green cabs are in Manhattan.

That too, for green cab the most important pickup and dropoff location is the same, East Harlem North Zone. Meaning most customers demand their services in that region

Now we try break down data into different time related aspects. Here we slice the data into the following aspects to understand the trends.

- Quarter
- Day of week
- Hour of Day

This breakdown will gives us more details about the data making it even more granular than the previous analysis.

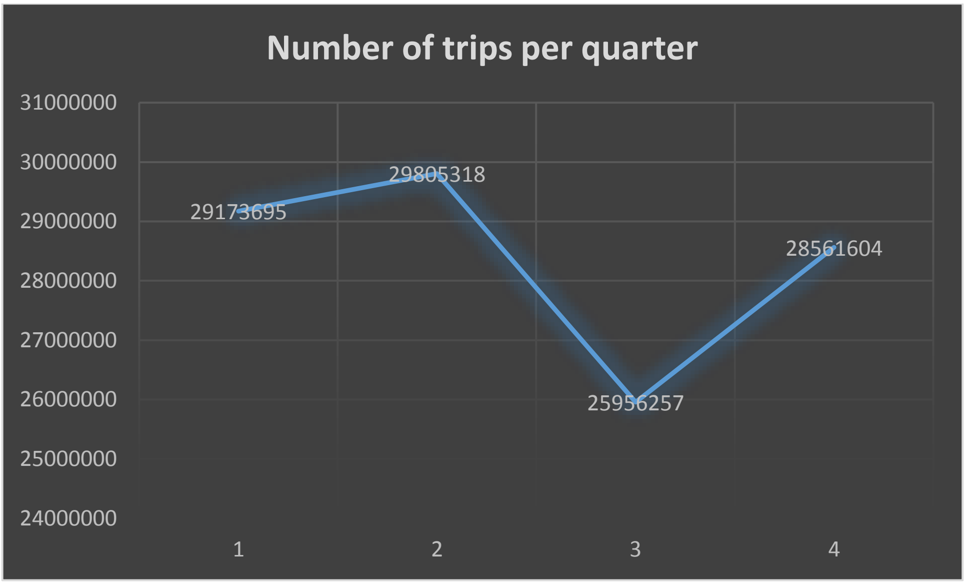
Metrics considered:

- Number of trips
- Avg. fare per ride
- Avg. trip per ride
- Avg. Passenger per ride

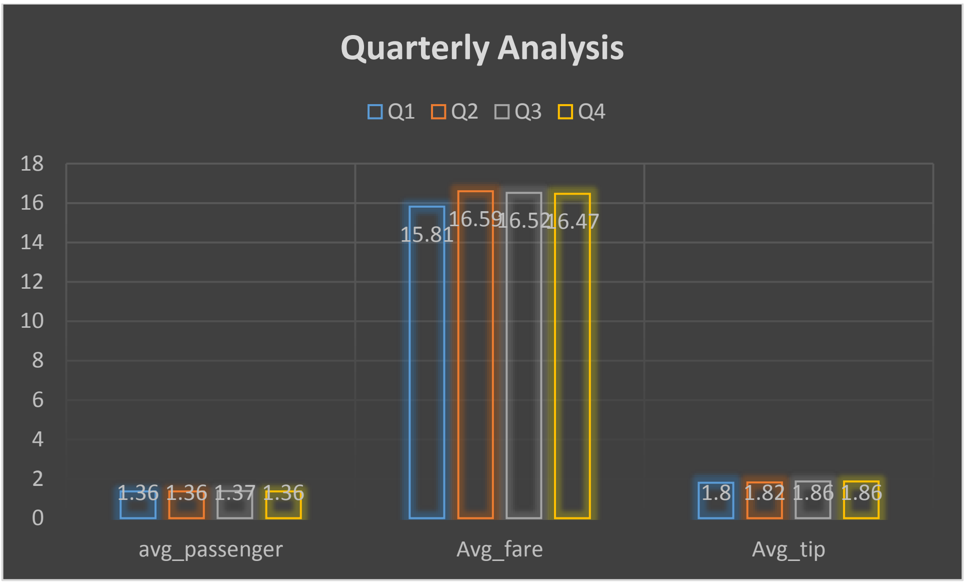
Quarterly Analysis:

In this section, we try to analyse the trends in different quarters of the year to identify the best performing quarter and how the others quarters perform.

For Yellow Taxi :

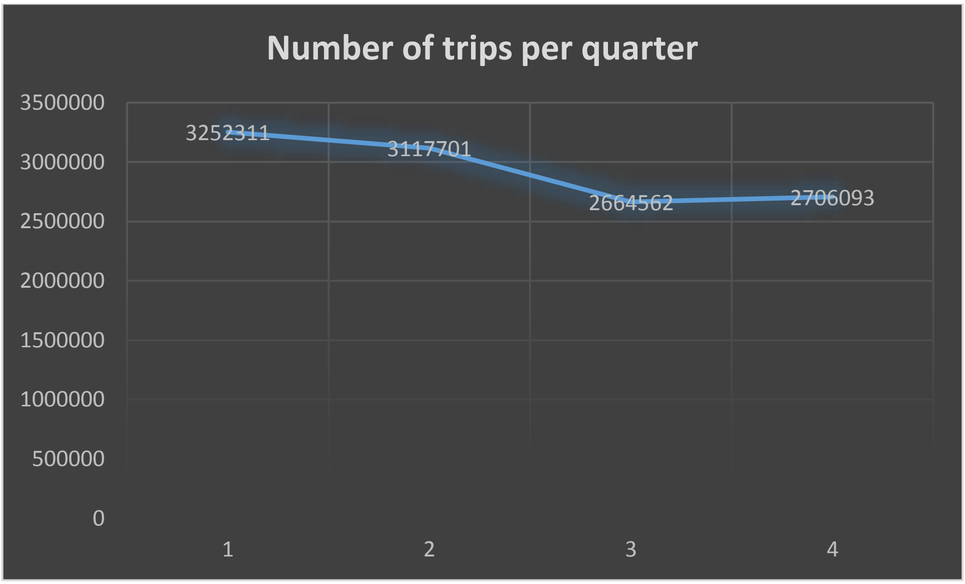


Number of trips per quarter



Quarterly analysis of other metrics for yellow taxi

For Green Taxi:



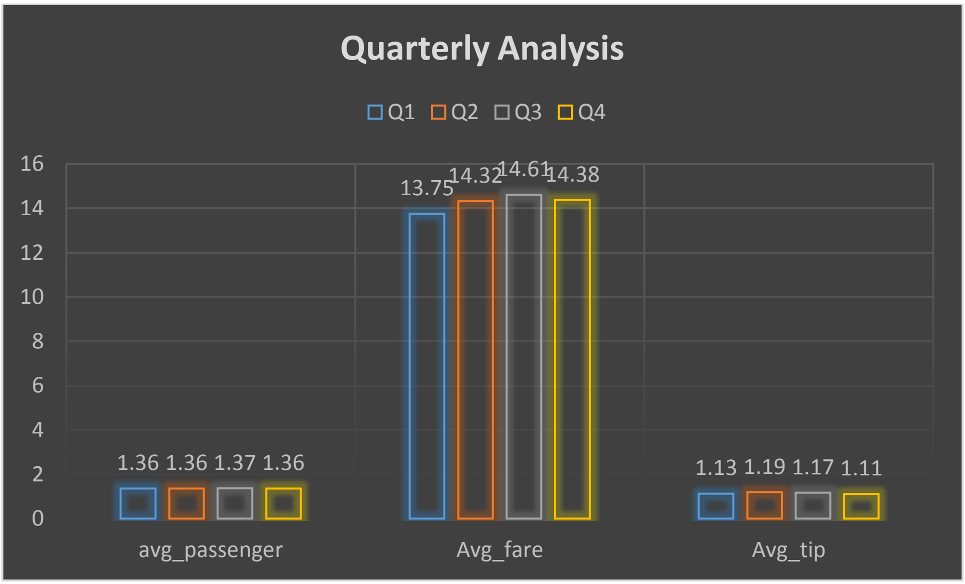
Number of trips in each quarter

If we analyse the trips graphs above, it clear that quarter 3 is the least performing quarter for both yellow and green cabs. While, the initial quarters produce better results for both the services. The ending quarter is average performing quarters for both services with respect to the number of trips are concerned.

The other metrics in consideration has not changed much over the quarters and has been consistent with little deviations which is acceptable.

If we also look at the data along with the monthly statistics available in the technical document, we can see that the most profitable or the month in which the demand was high were Feb, March, April and May which come under first and second quarter.

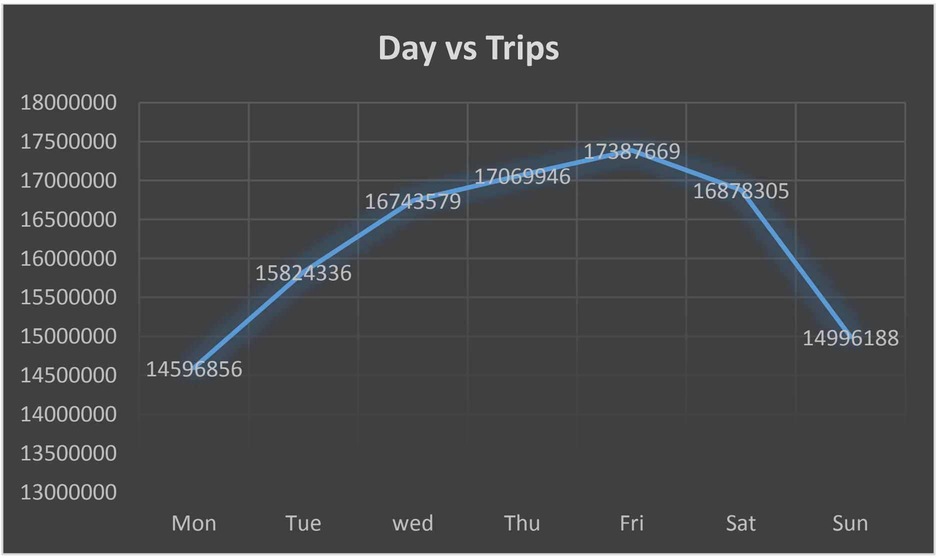
In 2017, these 4 months have been the most demanding for the service providers. It is also understandable because, it's the closing of the financial year and hence lot of activities go around during that time. The service providers can plan accordingly for the next year to meet the demands around the same months.



Quarterly analysis of other metrics for green taxi

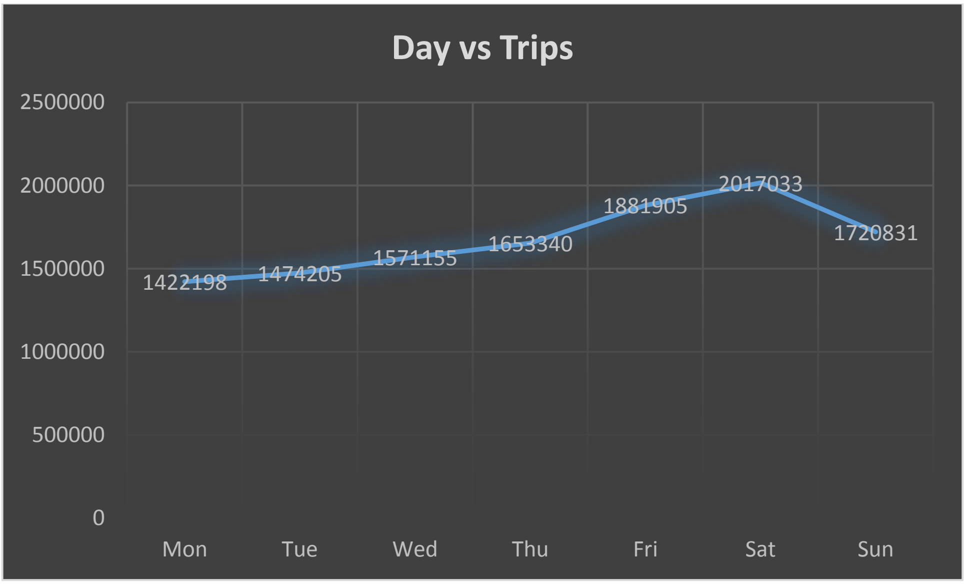
Daywise Statistics:

Yellow Taxi:



Day vs Trips

Green Taxi:



Day vs Trips

As expected , the trends in the daywise statistics is as follows.

Number of trips are relatively less during the start of the week (Mon and Tue), picks up during the nidweek (Wed,Thu), reaches the high during the weekend (Friday ,Sat) and again faces the drop during Sunday.

This is completely understandable as most people uses cab during weekend to go home as well as going out with family and other things. And in Sunday generally people prefer rest and the week begins slowly by Monday.

This trend is similar to both services.

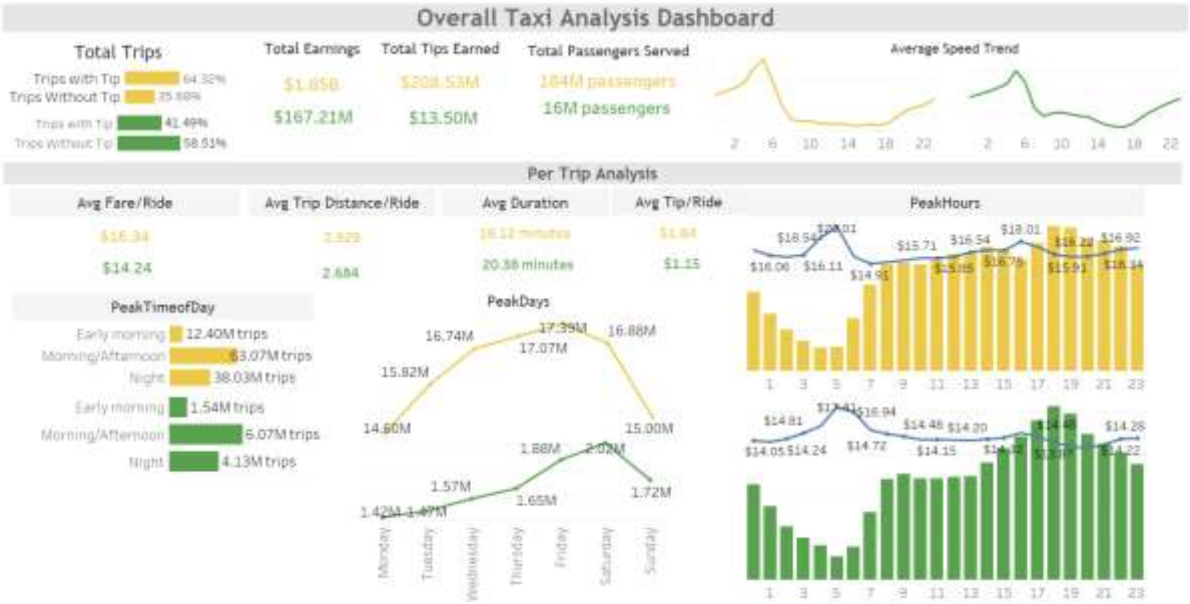
So the service providers should be well prepared to meet the demand during weekend and should manage their services accordingly.

Likewise the Quarterly statistics, the other metrics remain consistent through the day of the week with little fluctuations. Further details on these numbers can be found in the technical document , under daywise statistics for both yellow and green taxis.

Hourwise statistics:

Now we drill further deep hourwise to see the trends.

Here we consider no.of trips, average_fare_trip and avg_speed of the trip for analysis.



Overall Taxi Analysis Dashboard

If we look at the PeakHours in above dashboard for both yellow and green cabs, it is clearly evident that the peak hours fall between 4 PM to 8 PM in the eve with the highest peak recorded at 6 PM which is generally the time most people leave work or go for dinner. So this is the time where demand for the cabs are really high.

Similarly, if we look at the entire peak hour graph, another common trend is day starts slowly at 8 AM and then saturates in the noon (12 PM – 3 PM), then peaks in the eve (4 PM – 8 PM) and slowly goes down in the night and midnight.

If we plot the avg_fare amount along with the hourly trip breakdown, we get another insight. Eventhough the peak hour is in the eve, the highest average_fare_per_ride comes for the early morning trips around (4 AM - 5 AM). Driving at these time supposedly gives more average fare since not many cabs will be available during that time.

Highest average_fare for yellow cab is $20.41at 5 AM whereas its $17.41 for green cab around the same time.

What about the speed at which drivers drive ?

Another important metric to consider is speed. How soeedy are your drivers? This is important for two aspects.

1) **Time management of the trips**
2) **Safety of the trip**

Again like the above analysis, if we look at the Average speed trend graph in the dashboard, we can see that the maximum speed is during midnight to early morning (2 AM – 6 AM) when the traffic on road is less.

As the day progresses, the speed also decreases reaching the lowest at the peak hour of 6 PM which is because of the number of vehicles on road and slowly picks up at night again.

This trend can help drivers manage their time accordingly understanding the trend in which speed behaves throughout the day.

Vendor Analysis:

Another analysis standpoint that we have considered is the by the vendor type.

There are two different vendors providing the yellow and the green cab services.

1) **Creative Mobile Technologies, LLC;**
2) **VeriFone Inc.**

Here we will just do a basic exploratory analysis of the vendor across the key metrics to understand the vendor performance.

Number of trips served by each vendor:

Total number of trips:

1: 51 M

2: 62 M

Total number of trips:

1: 2.2 M

2: 9.4 M

In both the services, its **VeriFone Inc.** this is leading the race. Especially in the **green trips market**, they hold the market share of **82 %**

Total Earnings:

Total Earnings:

1: 800 M

2: 1 B

Total Earnings:

1: 31.5 M

2: 135 M

As expected , **VeriFone Inc** contributes to the majority of the green taxi revenue where as in yellow taxi is bit more evenly distributed.

Per trip details for both the vendors:

Vendor 1 :

Avg trip distance: 2.84m

Avg Passenger cnt: 1.25

Avg fare per ride: $16.16

Avg tip per ride: $1.83

Avg time per trip: 13.57 min

Avg speed of trip : 15.2 m/h

Vendor 2 :

Avg trip distance: 3.01m

Avg Passenger cnt: 1.93

Avg fare per ride: $16.5

Avg tip per ride: $1.87

Avg time per trip: 18.58 min

Avg speed of trip : 11.7 m/h

Although most of the metrics are same, vendor 2 serves more customers on average and spend more time on road.

Vendor 1:

Avg trip distance : 2.61

Avg Passenger cnt: 1.17

Avg fare per ride : $13.77

Avg tip per ride : $1.14

Avg time per trip : 12.48 min

Avg speed of trip : 19.3 m/h

Vendor 2:

Avg trip distance : 2.7m

Avg Passenger cnt: 1.41

Avg fare per ride : $14.36

Avg tip per ride : $1.15

Avg time per trip : 22.39 min

Avg speed of trip : 13.0 m/h

Not surprisingly, vendor 2 leads vendor 1 in important aspects like the passengers served, average fare per trip. But vendor 1 leads vendor 2 highly on the average speed. This might be because of several factors like vendor 1 might not get enough trips during the peak day, most of their trips might be during late night and also vendor 2 holds 80% share in green taxi market so they attend lot of trips and the speed is supposedly should go down. All in all, vendor 2 outclasses vendor 1 in green market.

What about the disputes between passenger and driver?

This is one area where vendor 2 has a clean slate in the green taxi data provided where as vendor 1 as 542 cases of disputes even though they serve a small population.

This is one area **Creative Mobile Technologies, LLC** should take a look at , as they are still finding their feet in the green market. These many high dispute cases with passengers can have an impact on their brand and growth as well.

Where as in the yellow market both vendor does not have any disputes as their market is quite stable and almost equally split.

Conclusion:

This concludes the analysis done on the yellow and green cabs for the year 2017.The aim of this analysis is to explore the data set on different metrics and to summarize the data as closely as possible. Further analysis on other metrics like tip amount, tolls amount has been done using SQL and it is available in the technical document attached. The tableau dashboard shows the overall taxi analysis for both green and yellow highlighting the key metrics.

Additional Analysis:

Irregular Data:

The last bit of analysis is the irregular data. When doing analysis, there we lot of data points which were having irregularities. Few of which are highlighted here.

Faulty records:

There are about 1447 records in green taxi and 51 records in Yellow taxi data where the dropoff date time is less than the pickup datetime.

Longest trip in green data:

Row	pickup_datetime	dropoff_datetime	longest_trip_in_days	trip_distance	total_amount
1	2017-04-08T23:35:14	2017-04-10T14:54:34	2	6.9	28.8

The longest trip shows the trip occurred for 2 days but the distance covered is only 6.9 miles and charged $28.8 . Definitely there is something wrong with this data point as this logically doesn’t make any sense.

Highest paid trip:

Row	vendor_id	pickup_datetime	dropoff_datetime	fare_amount	tolls_amount	total_amount
1	1	2017-03-28T15:32:05	2017-03-28T15:34:40	999.99	7999.92	8999.91

If we look at the highest paid trip above , the trip is billed at $8999.91 for a trip that lasted for 12 mins. The major reason for this high total is because of the tolls amount which is charged at 7999.92.This is again a data point that needs further investigation as something has gone wrong with the data.