EDA PROJECT REPORT

on

NYC Taxi Trip Duration

Submitted by

Yamana Pavan Kumar

Registration Number: 12016022

Program Name: B.Tech Data Science(ML and AI)

Under the Guidance of

Mr. Abhijeet Dutta

School of Computer Science & Engineering

Lovely Professional University, Phagwara

Introduction:

Taxi is one of the urban public transports in many busy countries. Unlike other public transports, taxi rides provide accessibility, convenience, yet privacy to passengers. A competitive and reasonable taxi pricing is worth the ride for private car users to switch to a taxi service. Millions of taxi trips data are generated on monthly basis, which this data can be useful to gain the insight of the traffic patterns and obtain a clear view of urban city life. Not only that, by leveraging the given dataset, taxi demand on major events like Christmas and New Year's Eve, can be studied in order to make a better decision making.

In this project, we examine the urban dataset of New York City taxi trips. According to taxi factbook report release by NYC Taxi & Limousine Commission (2018), TLC taxi trips in New York City has a total of over 41 million trips a year between June 2017 and June 2018. In general, taxi trips comprise spatial elements like longitude and latitude points. Thus, this data enables to encode geolocation information into an insight of urban traffic movement and activities.

Not only that, other attributes like taxi Id, number of passengers are also recorded which allows to study the traffic congestion, optimal fleet size.

Data is obtained through Kaggle repository titled "NYC Taxi Trip Duration" which the data composes of taxi ride in New York City from 2009 to 2015. The dataset attributes including key, pickup and drop-off geolocation, date, time, passenger count and finally fare amount will be analysed.

Problem Context:

A typical taxi company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. One of main issue is determining the duration of the current trip so it can predict when the cab will be free for the next trip. The data set contains the data regarding several taxi trips and its duration in New York City. I will now try and apply different techniques of Data Analysis to get insights about the data and determine how different variables are dependent on the target variable Trip Duration.

The insights I want find from this dataset are:

- How much can the industry expect to earn from a day
- Does the number of customers change from start to the end of the day
- How to plan the work schedule so that the driver works in best possible time and day and increase his profit considerably
- Find the best and worst location to get customers at given time and day of the month
- Analyse the location where people most likely go when boarding the taxi depending on their pickup location.
- A model can be designed for increasing the growth of the companies in this field by making them spend their resources and money wisely

Let's have look on the columns of dataset:

Demographic information of Customer & Vendor:

- id : a unique identifier for each trip
- vendor_id : a code indicating the provider associated with the trip record
- passenger_count : the number of passengers in the vehicle (driver entered value)

Information about the Trip:

- pickup_longitude : date and time when the meter was engaged
- pickup_latitude : date and time when the meter was disengaged
- dropoff_longitude : the longitude where the meter was disengaged
- dropoff_latitude : the latitude where the meter was disengaged
- store_and_fwd_flag : This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server (Y=store and forward; N=not a store and forward trip
- trip_duration : (target) duration of the trip in seconds

Thus, we have a data set with 1458644 rows and 11 columns. There are 10 features and 1 target variable which is trip_duration.

Let's get a glimpse of the data set by looking at the first 5 rows.

	id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_f
0	id2875421	2	2016-03-14 17:24:55	2016-03-14 17:32:30	1	-73.982155	40.767937	-73.964630	40.765602	
1	id2377394	1	2016-06-12 00:43:35	2016-06-12 00:54:38	1	-73.980415	40.738564	-73.999481	40.731152	
2	id3858529	2	2016-01-19 11:35:24	2016-01-19 12:10:48	1	-73.979027	40.763939	-74.005333	40.710087	
3	id3504673	2	2016-04-06 19:32:31	2016-04-06 19:39:40	1	-74.010040	40.719971	-74.012268	40.706718	
4	id2181028	2	2016-03-26 13:30:55	2016-03-26 13:38:10	1	-73.973053	40.793209	-73.972923	40.782520	

Let's look into type of columns in dataset :

```
id                    object
vendor_id              int64
pickup_datetime       object
dropoff_datetime      object
passenger_count        int64
pickup_longitude     float64
pickup_latitude      float64
dropoff_longitude    float64
dropoff_latitude     float64
store_and_fwd_flag    object
trip_duration          int64
dtype: object
```

- The columns id and vendor_id are nominal.
- The columns pickup_datetime and dropoff_datetime are stored as object which must be converted to datetime for better analysis.
- The column store_and_fwd_flag is categorical

Check for a statistical summary of our dataset:

	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	trip_duration
count	1458644.00	1458644.00	1458644.00	1458644.00	1458644.00	1458644.00	1458644.00
mean	1.53	1.66	-73.97	40.75	-73.97	40.75	959.49
std	0.50	1.31	0.07	0.03	0.07	0.04	5237.43
min	1.00	0.00	-121.93	34.36	-121.93	32.18	1.00
25%	1.00	1.00	-73.99	40.74	-73.99	40.74	397.00
50%	2.00	1.00	-73.98	40.75	-73.98	40.75	662.00
75%	2.00	2.00	-73.97	40.77	-73.96	40.77	1075.00
max	2.00	9.00	-61.34	51.88	-61.34	43.92	3526282.00

The returned table gives certain insights:

- Passenger count has a minimum of 0 which means either it is an error entered or the drivers deliberately entered 0 to complete a target number of rides.
- The passenger count varies between 1 and 9 with most people number of people being 1 or 2
- The trip duration varying from 1s to 3526282s~ 979 hrs. There are definitely some outliers present which must be treated.

Let us see if there are any null values in our dataset.

```
id                    0
vendor_id             0
pickup_datetime       0
dropoff_datetime      0
passenger_count       0
pickup_longitude      0
pickup_latitude       0
dropoff_longitude     0
dropoff_latitude      0
store_and_fwd_flag    0
trip_duration         0
dtype: int64
```

➢ There are no null values in this dataset which saves us a step of imputing.

Let us check for unique values of all columns.

```
id                  1458644
vendor_id                 2
pickup_datetime     1380222
dropoff_datetime    1380377
passenger_count          10
pickup_longitude      23047
pickup_latitude       45245
dropoff_longitude     33821
dropoff_latitude      62519
store_and_fwd_flag        2
trip_duration          7417
dtype: int64
```

- We see that id has 729322 unique values which are equal to the number of rows in our dataset.
- There are 2 unique vendor ids.
- There are 9 unique passenger counts.
- There are 2 unique values for store_and_fwd_flag, that we also saw in the description of the variables, which are Y and N.

Feature Creation:

Let us create some new features from the existing variables so that we can gain more insights from the data.

Remember pickup_datetime and dropoff_datetime were both of type object. If we want to make use of this data, we can convert it to datetime object which contains numerous functions with which we can create new features that we will see soon.

Let us extract and create new features from this datetime features

We have created the following features:

- pickup_day and dropoff_day which will contain the name of the day on which the ride was taken.
- pickup_day_no and dropoff_day_no which will contain the day number instead of characters with Monday=0 and Sunday=6.
- pickup_hour and dropoff_hour with an hour of the day in the 24-hour format.
- pickup_month and dropoff_month with month number with January=1 and December=12.

Lets us determine what time of the day the ride was taken. I have created 4 time zones 'Morning' (from 6:00 am to 11:59 pm), 'Afternoon' (from 12 noon to 3:59 pm), 'Evening' (from 4:00 pm to 9:59 pm), and 'Late Night' (from 10:00 pm to 5:59 am).

We have created the following features:

- pickup_timeofday which will contain the time zone at the time of pickup
- dropoff_timeofday which will contain the time zone at the time of dropping

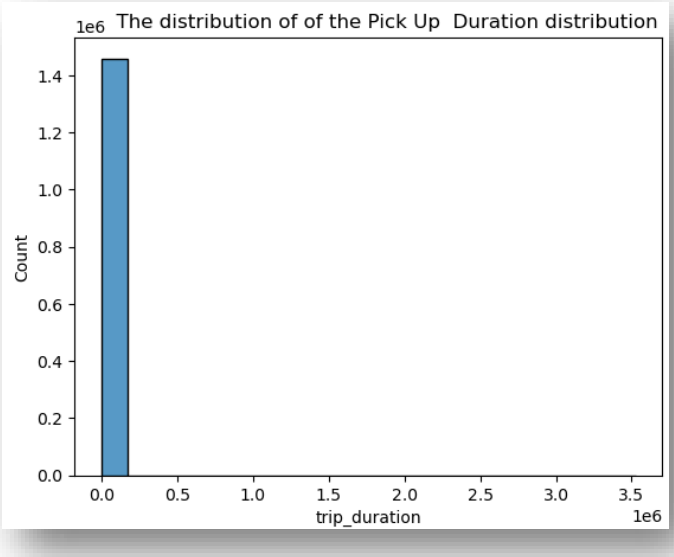
We also saw during dataset exploration that we have coordinates in the form of longitude and latitude for pickup and dropoff. But we can't really gather any insights or draw conclusions from that. So, the most obvious feature that we can extract from this is distance.

We have created the following features:

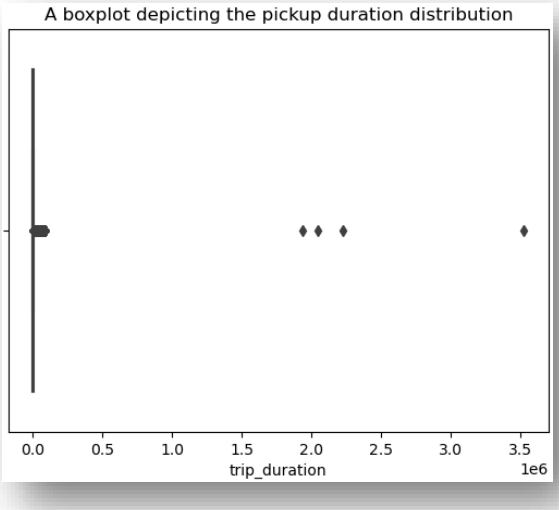
- Distance which contains how much distance covered by each trip

Univariate Analysis:

Histplot of Trip Duration (Target Variable):



➢ This histogram shows extreme right skewness, hence there are outliers. Let's see the boxplot of this variable.



➢ Thus we see there is only value near 2000000 while all the others are somewhere between 0 and 100000. The one near 2000000 is definitely an outlier which must be treated. We can clearly see an outlier.¶

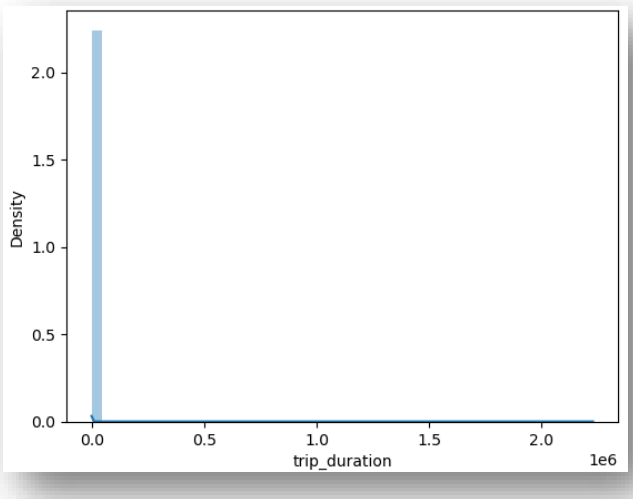
Let's have a look at the 10 largest value of trip_duration :

```
978383     3526282
924150     2227612
680594     2049578
355003     1939736
1234291      86392
295382       86391
73816        86390
59891        86387
1360439      86385
753765       86379
Name: trip_duration,
```

➢ The largest value is much greater than the 2nd and 3rd largest trip duration value. This might be because of some errors which typically occurs during data collection, or this might be a legit data. Since the occurrence of such a huge value is unlikely so it's better to drop this row before further analysis.

- The value can be replaced by the mode or median of trip duration as well.

Distribution of the trip_duration after we have dropped the outlier:



➢ Still there is an extreme right skewness. Thus, we will divide the trip_duration column into some interval.

The intervals are decided as follows:

- less than 5 hours
- 5–10 hours
- 10–15 hours
- 15–20 hours
- more than 20 hours

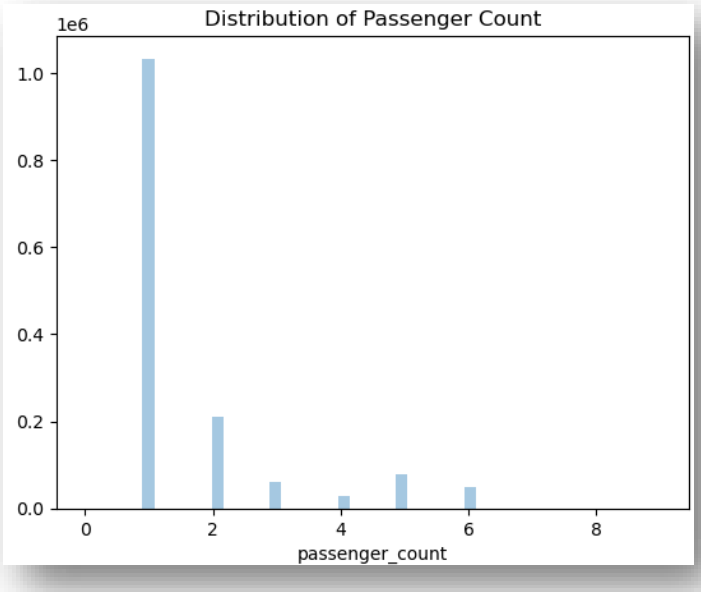
Passenger Count:

Count of unique values of Passenger Count Column:

```
1    1033540
2     210318
5      78088
3      59896
6      48333
4      28404
0         60
7          3
9          1
8          1
Name: passenger_count
```

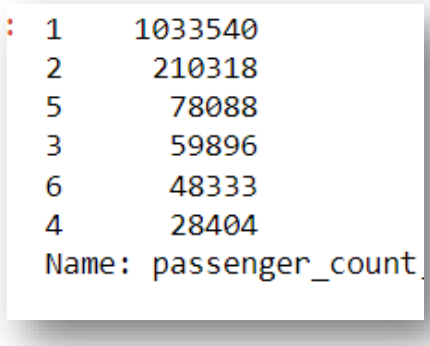
- There are some trips with even 0 passenger count that means the booking might be cancelled due to some reason.
- There is only 1 trip each for 7 and 9 passengers.

Histogram of the passengers in each trip :



- Here we see that the mostly 1 or 2 passengers avail the cab. The instance of large group of people travelling together is rare.

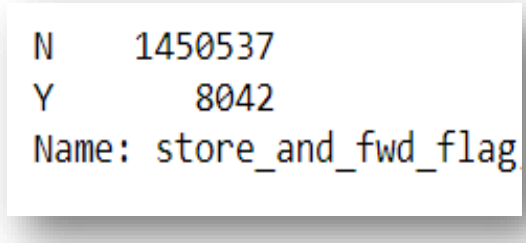
Let us remove the rows which have 0 or 7 or 8 or 9 passenger count.



- Now, that seems like a fair distribution.

Store and Forward Flag:

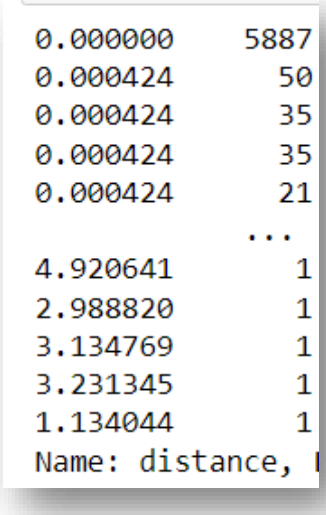
Frequency distribution of the Yes/No Flag



➢ The number of N flag is much larger. We can later see whether they have any relation with the duration of the trip.

Distance:

Count of unique values of Distance Column :



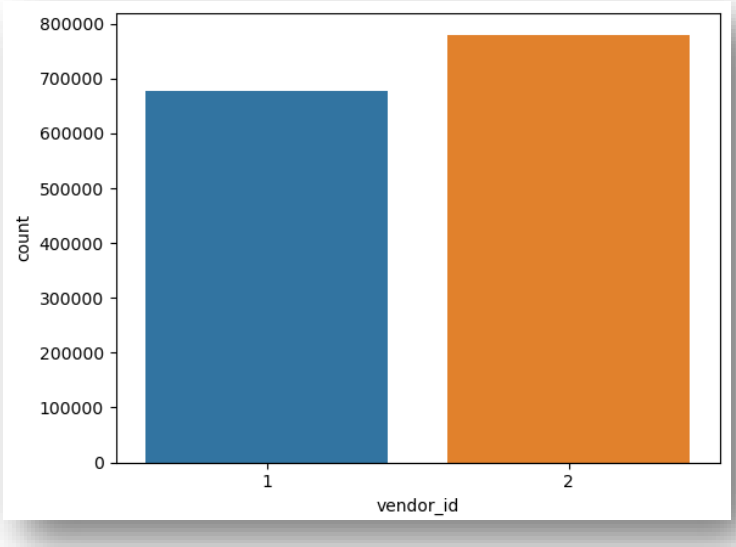
➢ We see there are 5887 trips with 0 km distance.

➢ The reasons for 0 km distance can be:

✓ The drop-off location couldn't be tracked.
✓ The driver deliberately took this ride to complete a target ride number.
✓ The passengers cancelled the trip.

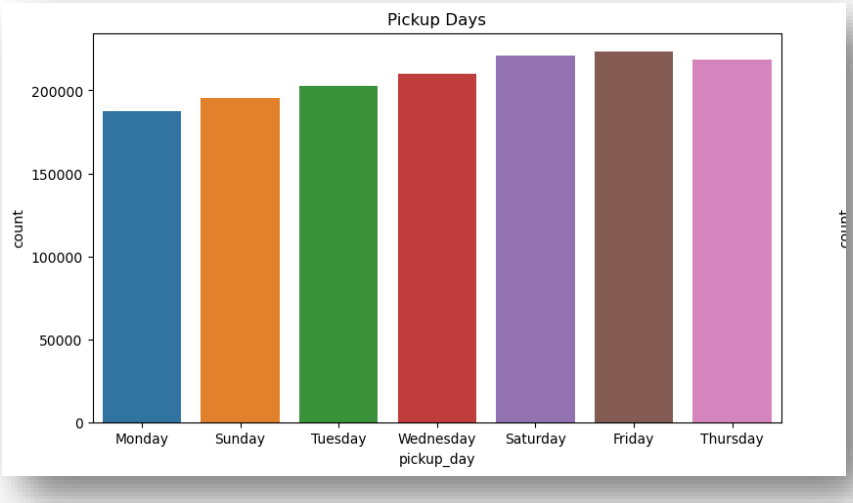
Vendor id:

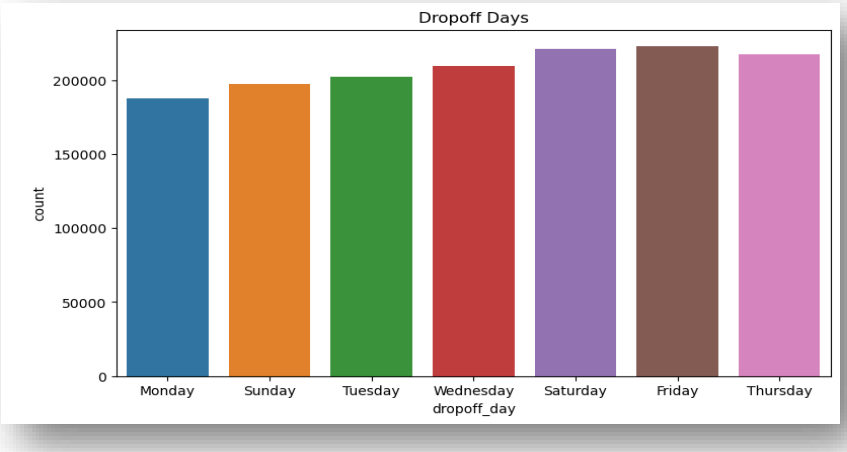
Countplot of Vendor id column



- We see that there is not much difference between the trips taken by both vendors.

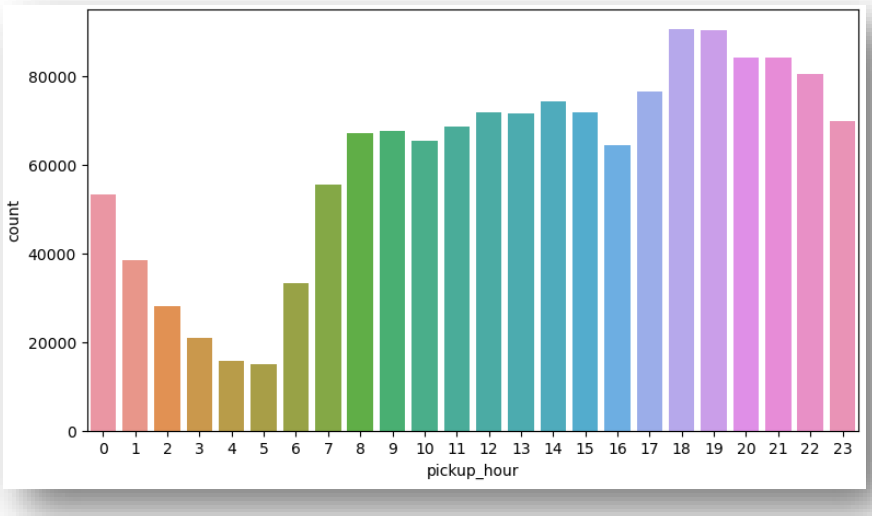
Trips per Day:

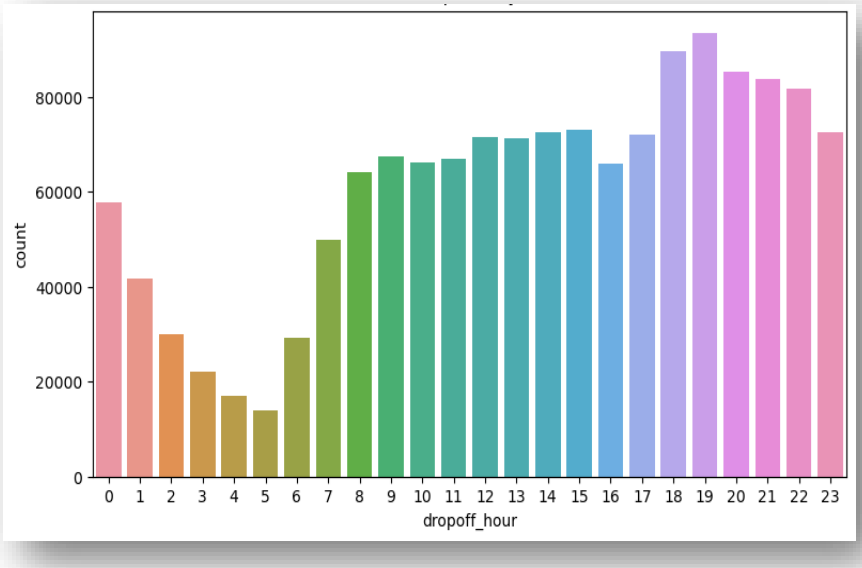




- Thus we see most trips were taken on Friday and Monday being the least.
- We see Fridays are the busiest days followed by Saturdays. That is probably because it's weekend.

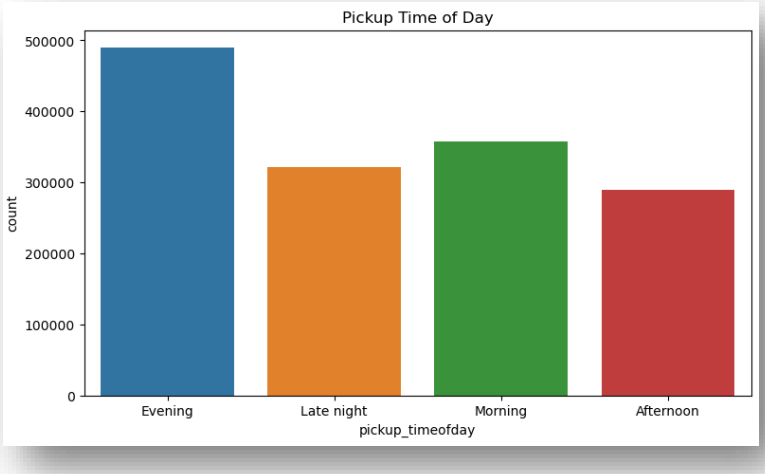
Trips per Hour :

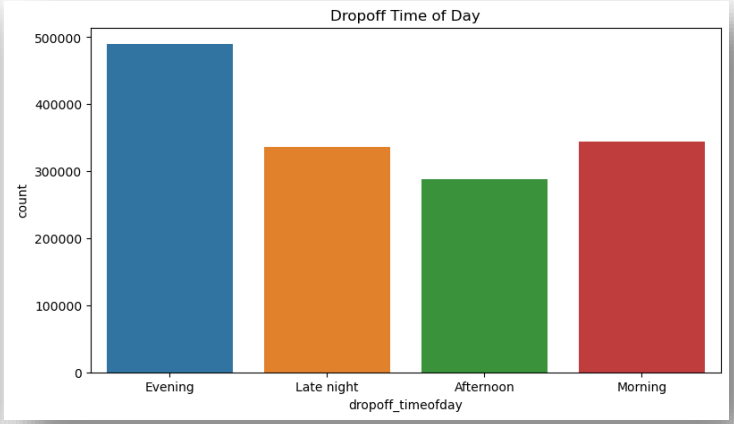




➢ We see the busiest hours are 6:00 pm to 7:00 pm and that makes sense as this is the time when people return from their offices.

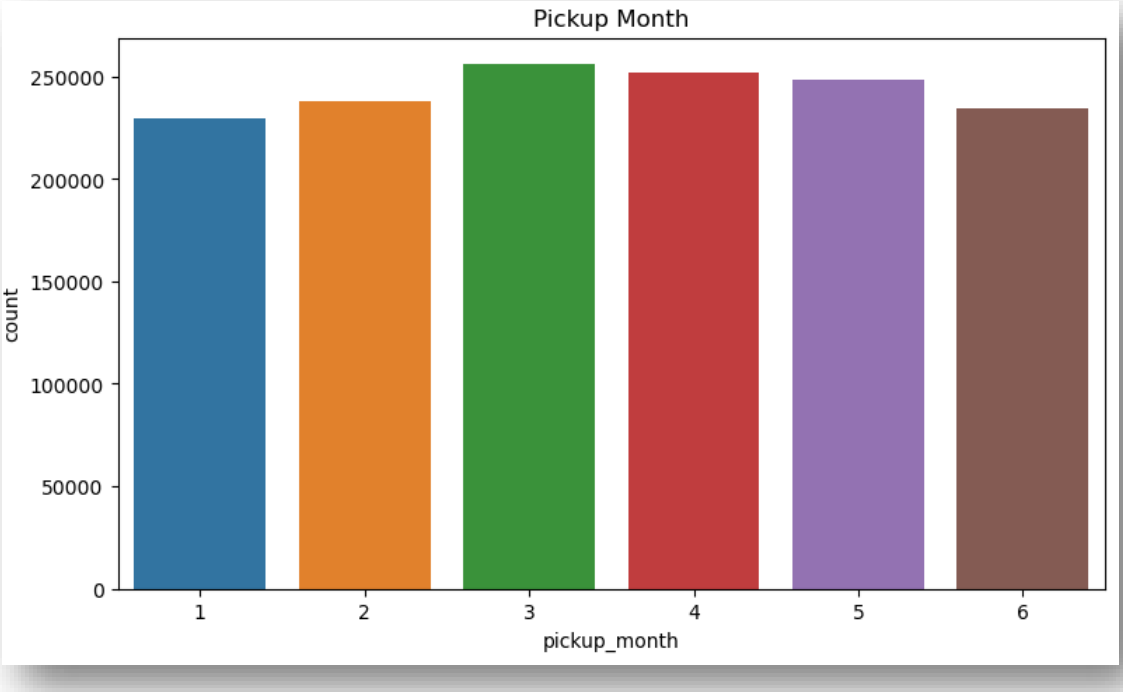
Trips per Time of Day:

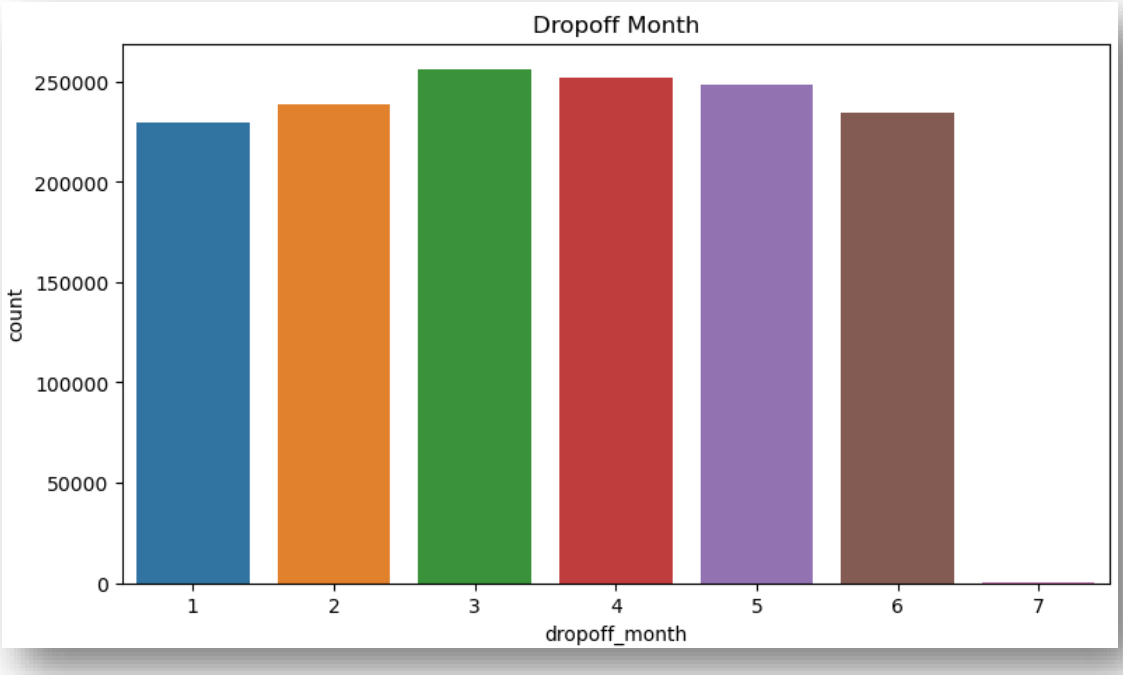




- Thus, we observe that most pickups and drops occur in the evening. While the least drops and pickups occur during morning.

Trips per month:

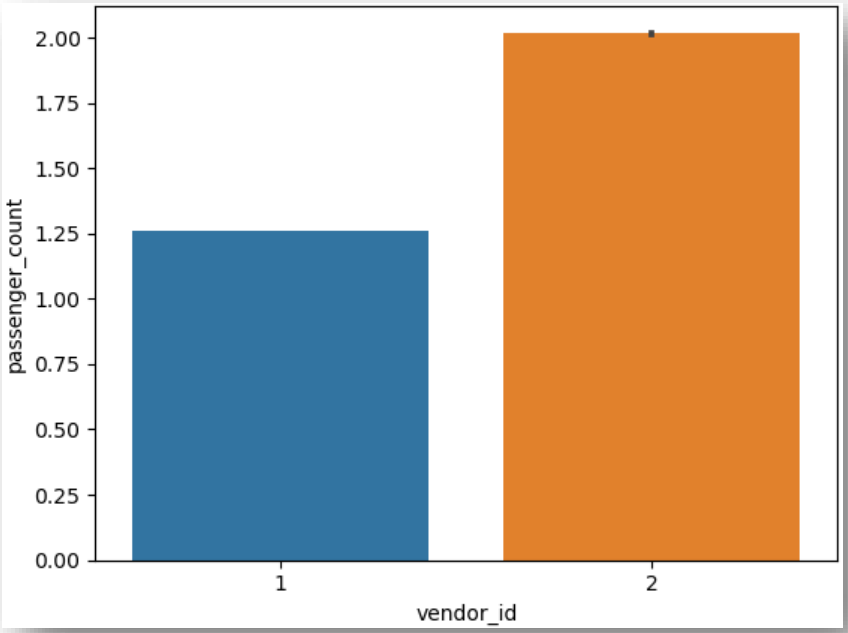




- There is not much difference in the number of trips across months

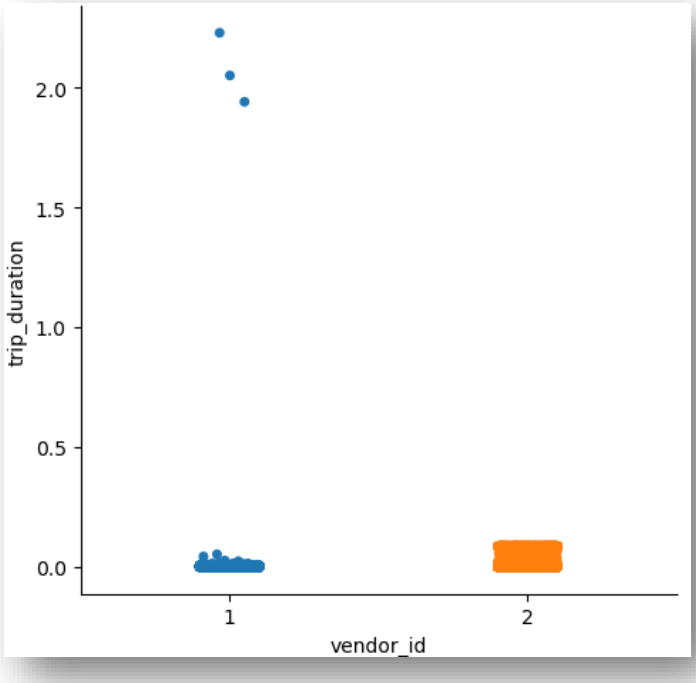
Bivariate Analysis:

Distribution of Passenger Count and Vendor id



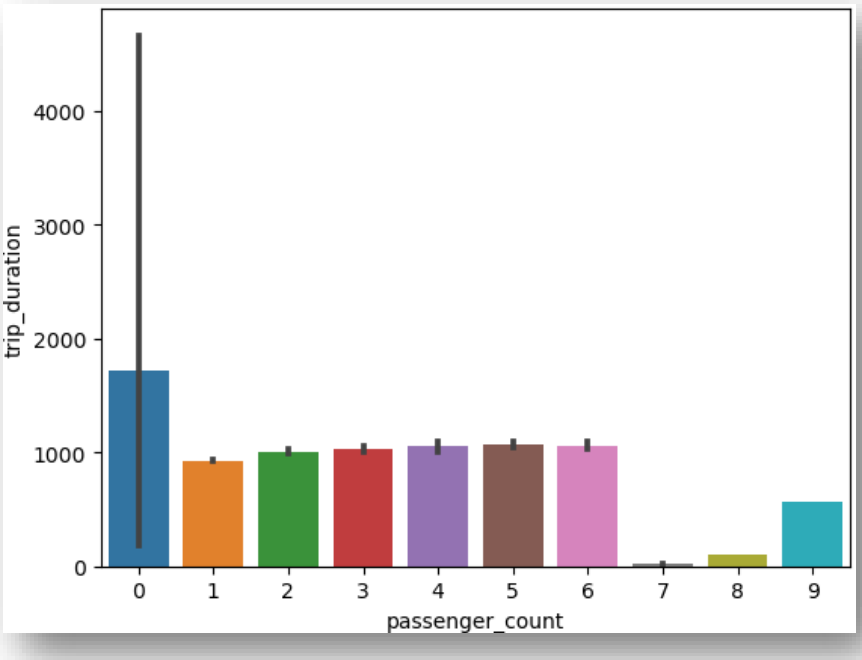
- Most People Preferred Vendor 2 For Booking Cab Services that can lead to the Thinking part:
- There might be shortage of Taxi provided by vendor 1.
- Or There might be good service Provided by Vendor B as compared to A.

Lets investigate by drawing catplot.



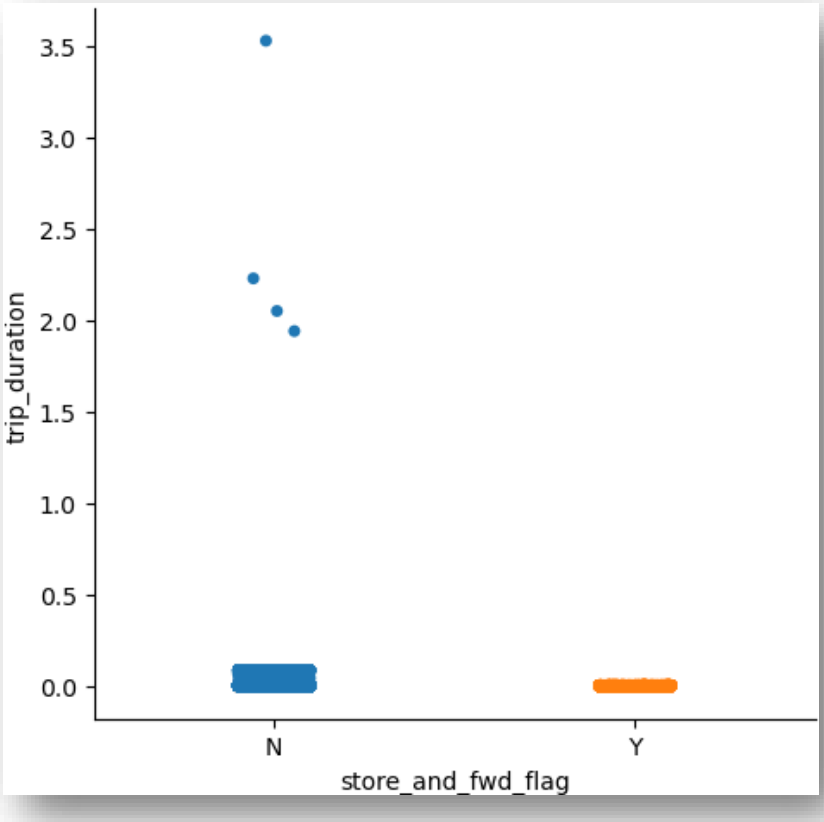
- Vendor 1 Offer Short Trips. Whereas Vendor 2 offer short as well as long trips that's why people preferred more Vendor 2.
- Vendor id 2 takes longer trips as compared to vendor 1.

Distribution of Passenger Count and Trip Duration:



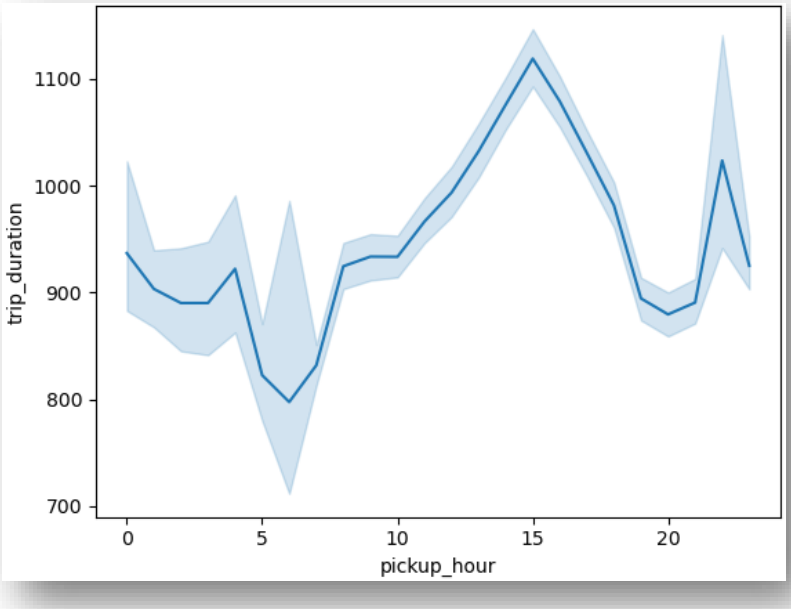
- Most of the Passenger Trip Duration is b/w 800 to 1000 sec.
- The lowest trip duration is around 0 that might be an outlier or people might have cancelled after booking.

Distribution of Trip Duration per Store and Forward Flag:



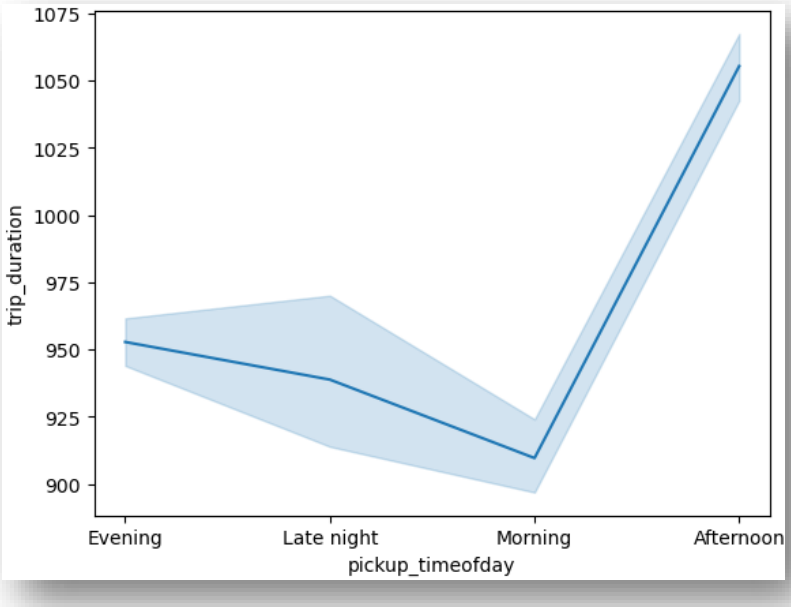
- Trip duration is generally longer for trips whose flag was not stored.

Distribution of Trip Duration per hour:



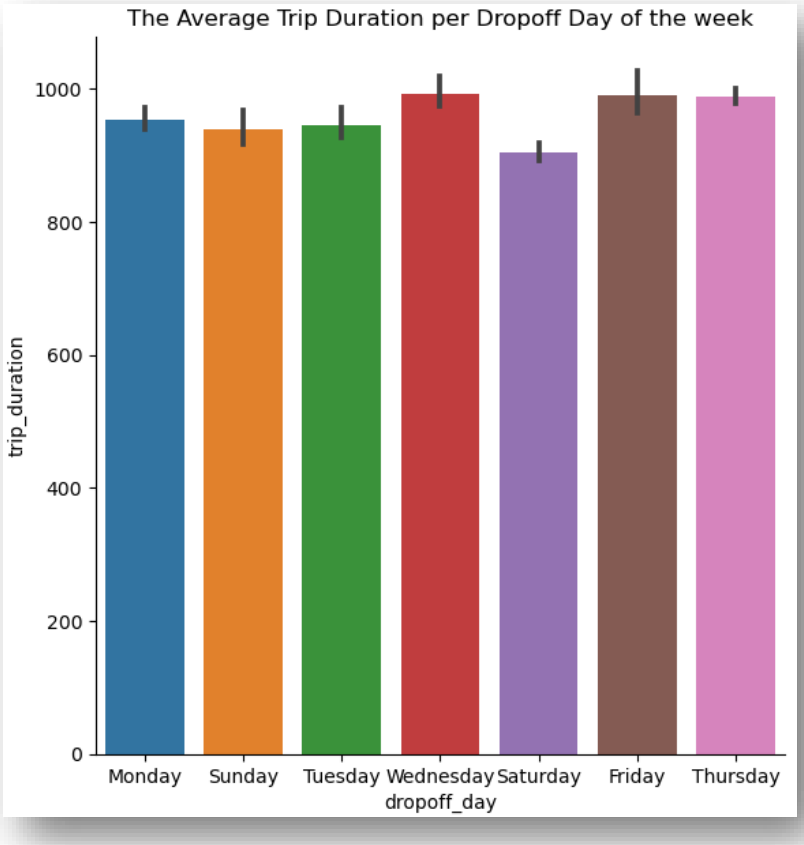
- We see the trip duration is the maximum around 3 pm which may be because of traffic on the roads. Trip duration is the lowest around 6 am as streets may not be busy.

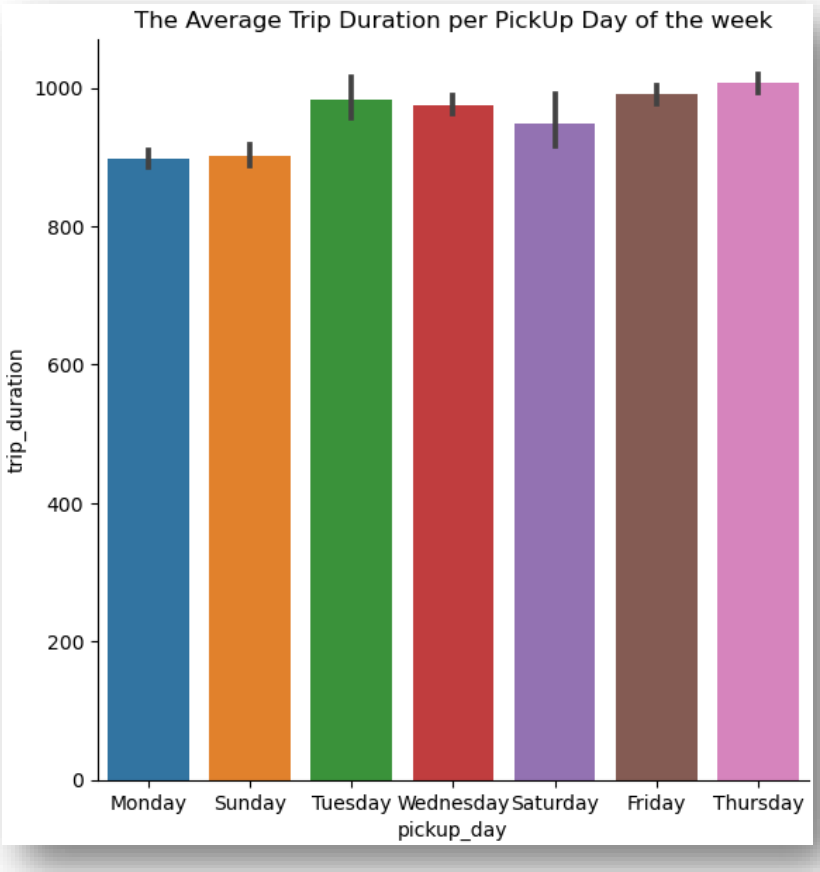
Distribution of Trip Duration per time of day



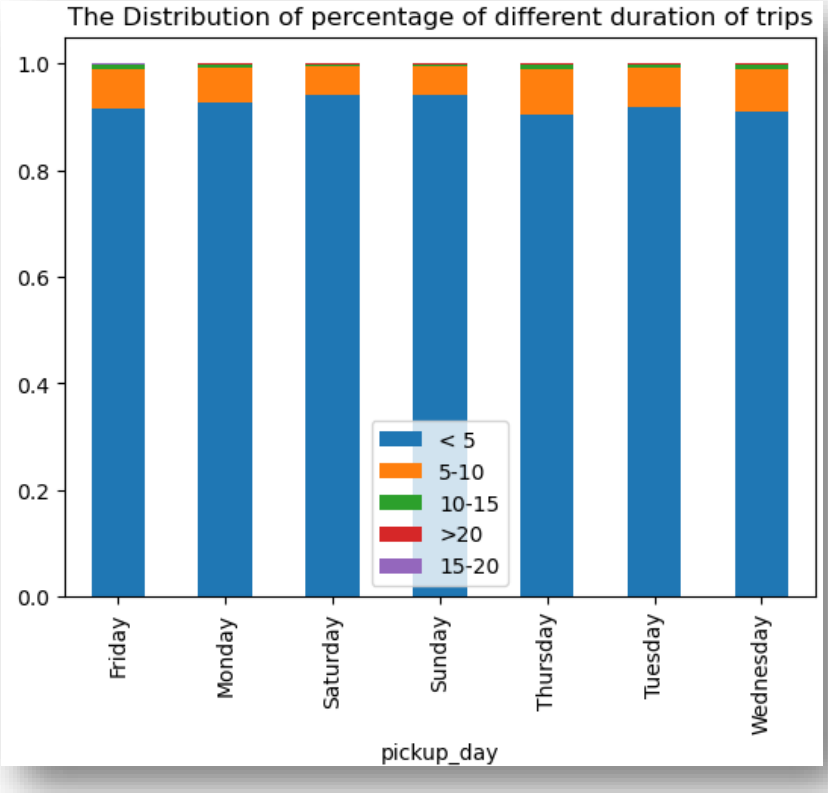
- As we saw above, trip duration is the maximum in the afternoon and lowest between late night and morning.

Distribution of Trip Duration per Day of Week:

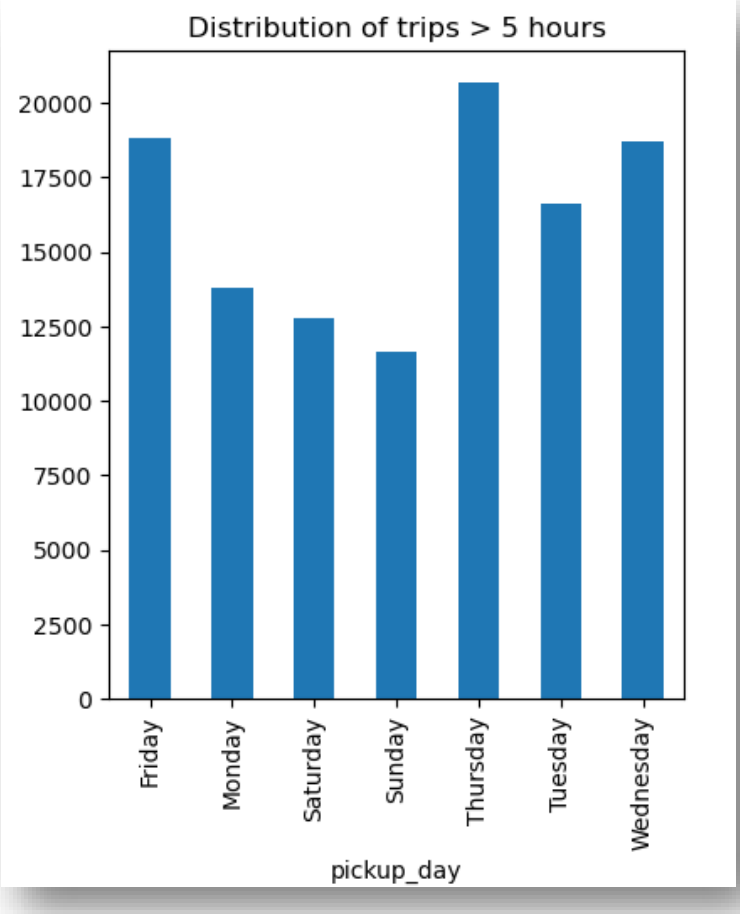




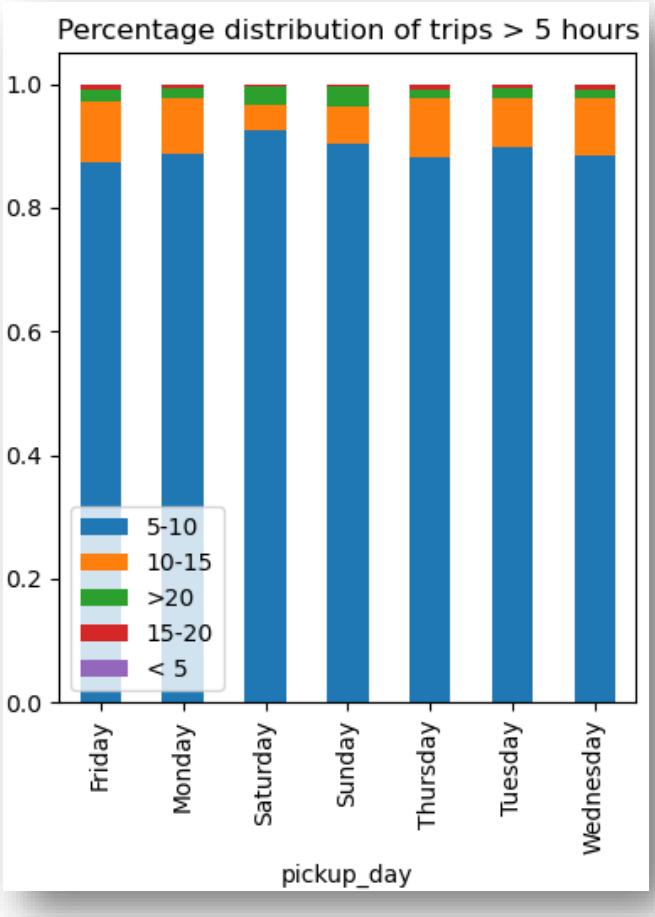
- The graphs denote the average estimate of a trip for each day of the week. The error bars provide some indication of the uncertainty around that estimate
- Thus, the highest average time taken to complete a trip is on Thursday while Monday, Saturday and Sunday take the least time.
- But this is not enough. We must also take into consideration the percentage of short, medium and long trips taken on each day.



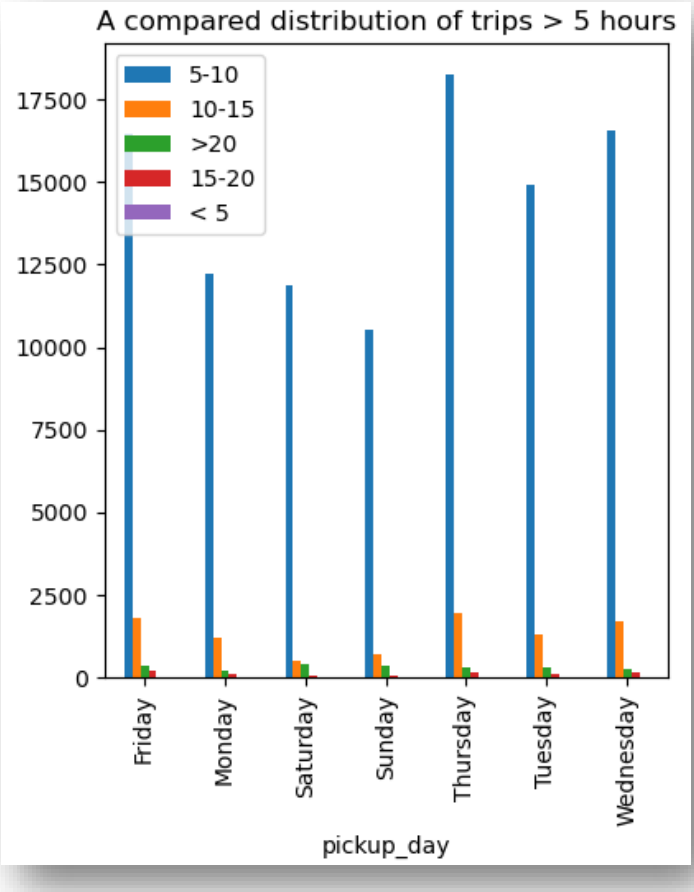
- The graph shows a percentage distribution of the trips of different duration within each day of the week.
- This does not give many insights as the number of trips within 0–5 hours range is much larger for all the days,
- Let's look at the percentage of only longer trips (with duration time > 5 hours)



- Graph shows a frequency distribution of the number of trips (> 5 hours) taken on each day of the week



➢ Graph shows a percentage distribution of the trips of different duration (> 5 hours) within each day of the week.

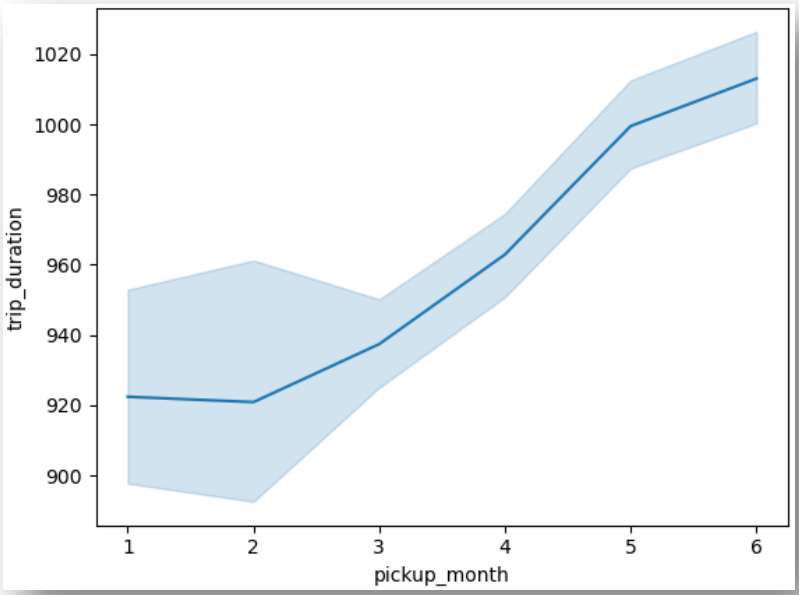


➢ Graph shows the frequency distribution of the trips of different duration (> 5 hours) within each day of the week.

Some key points:

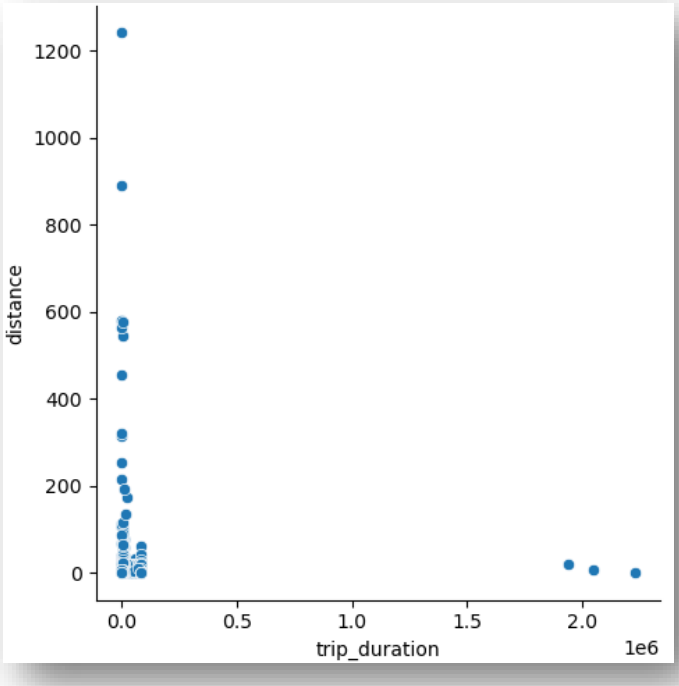
- The most number trips which lasts > 5 hours were taken on Thursday followed by Friday and Wednesday. (Left graph)
- The most number of trips of duration 5–10, 10–15 was taken on Thursday. (right graph)
- But the highest percentage of trips longer than 20 hours was taken on Sunday and Saturday. (middle graph)

Distribution of Trip Duration per month:



➢ From February, we can see trip duration rising every month.

Distribution of Trip Duration and Distance:



➢ We can see there are trips which trip duration as short as 0 seconds and yet covering a large distance. And trips with 0 km distance and long trip durations.

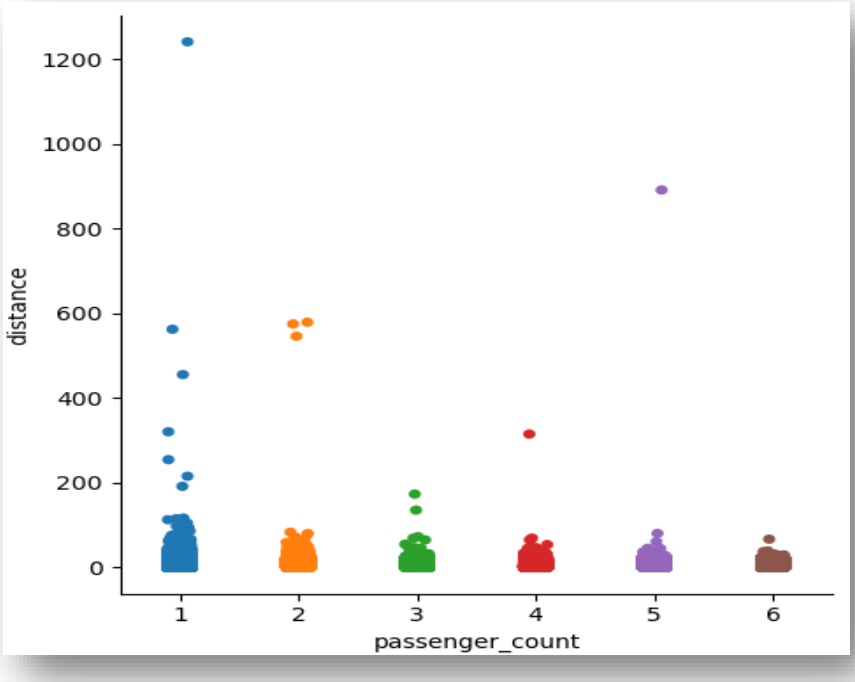
Let us see few rows whose distances are 0.

flag	...	pickup_day_no	dropoff_day_no	pickup_hour	dropoff_hour	pickup_month	dropoff_month	pickup_timeofday	dropoff_timeofday	distance	duration_time
N	...	0	0	18	18	2	2	Evening	Evening	0.0	< 5
N	...	1	1	18	18	5	5	Evening	Evening	0.0	< 5
N	...	0	0	23	23	5	5	Late night	Late night	0.0	< 5
N	...	0	0	19	19	1	1	Evening	Evening	0.0	< 5
N	...	2	2	22	22	1	1	Late night	Late night	0.0	< 5

We can see even though distance is recorded as 0 but trip duration is definitely more.

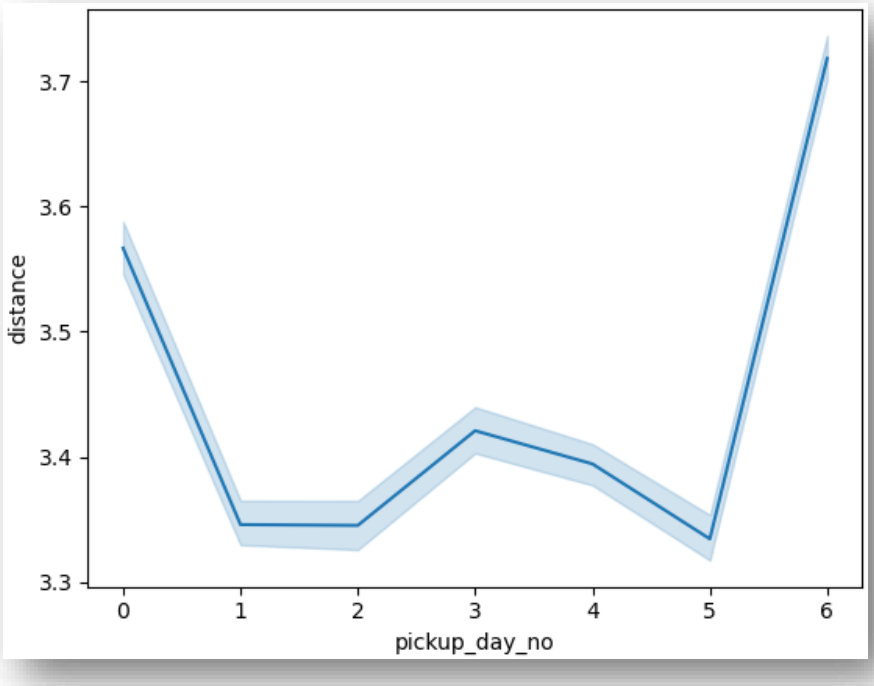
- One reason can be that the drop-off coordinates weren't recorded.
- Another reason one can think is that for short trip durations, maybe the passenger changed their mind and cancelled the ride after some time.

Distribution of Distance per passenger count:



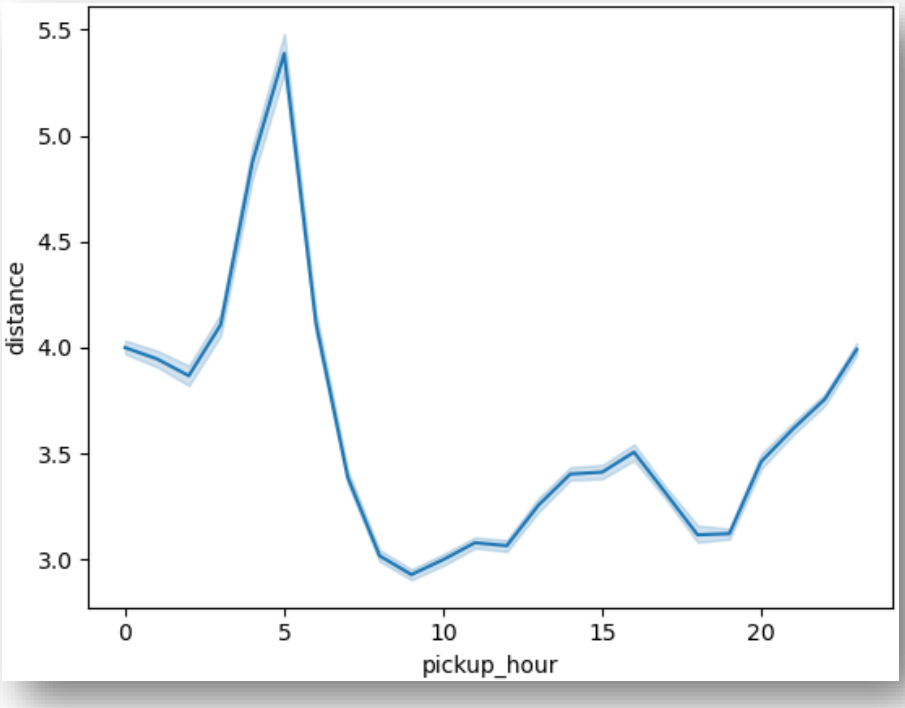
➢ We see some of the longer distances are covered by either 1 or 2 or 4 or 5 passenger rides.

Distribution of Distance per day of week:



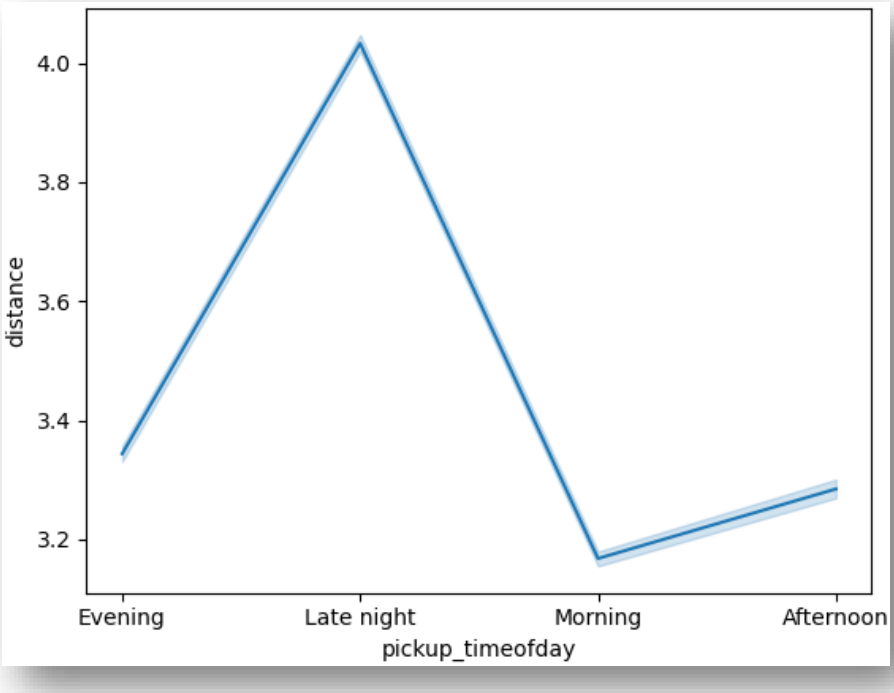
- Distances are longer on Sundays probably because it's weekend.
- Monday trip distances are also quite high.
- This probably means that there can be outstation trips on these days and/or the streets are busier.

Distance per hour of day:



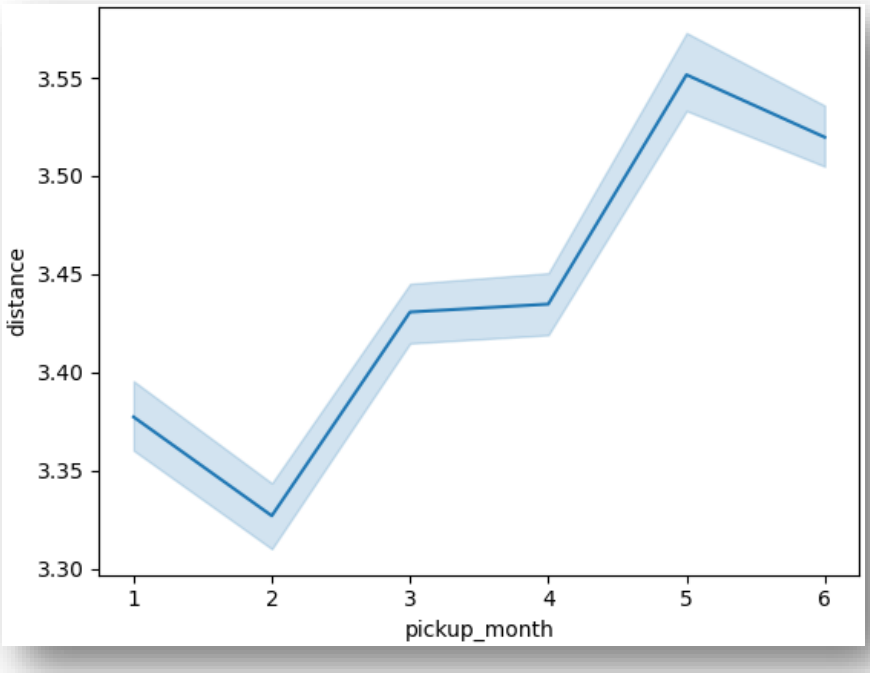
- Distances are the longest around 5 am.

Distance per time of day:



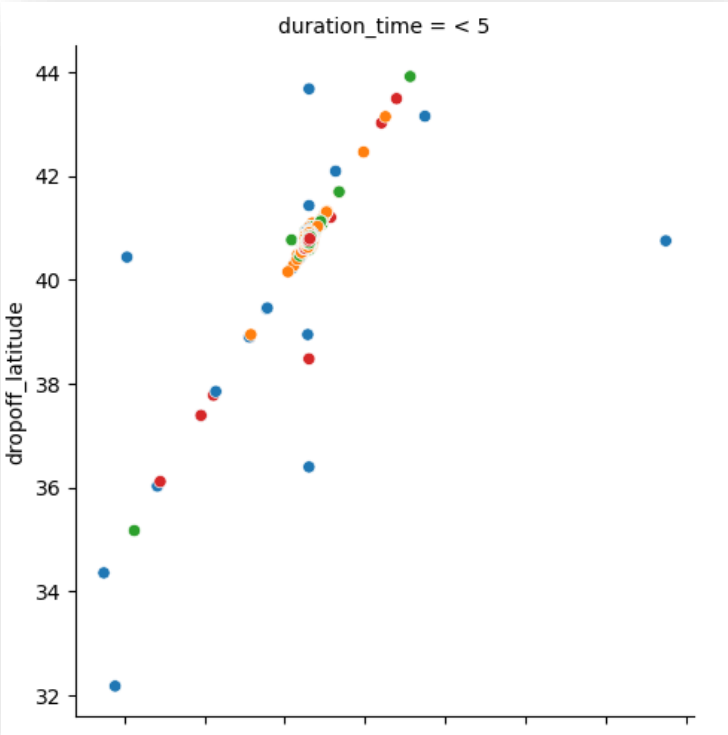
➢ As seen above also, distances being the longest during late night or it may be called as early morning too. This can probably point to outstation trips where people start early for the day.

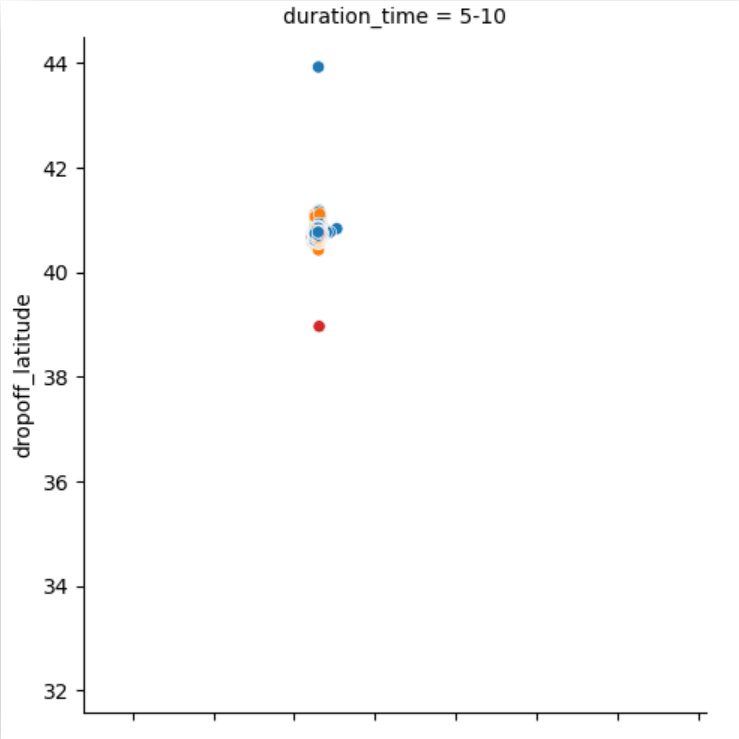
Distance per month:

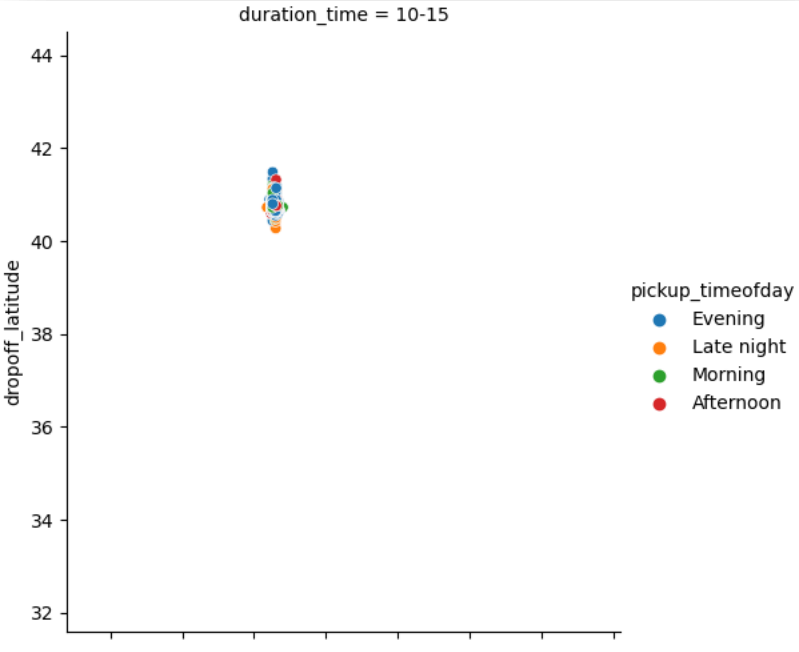


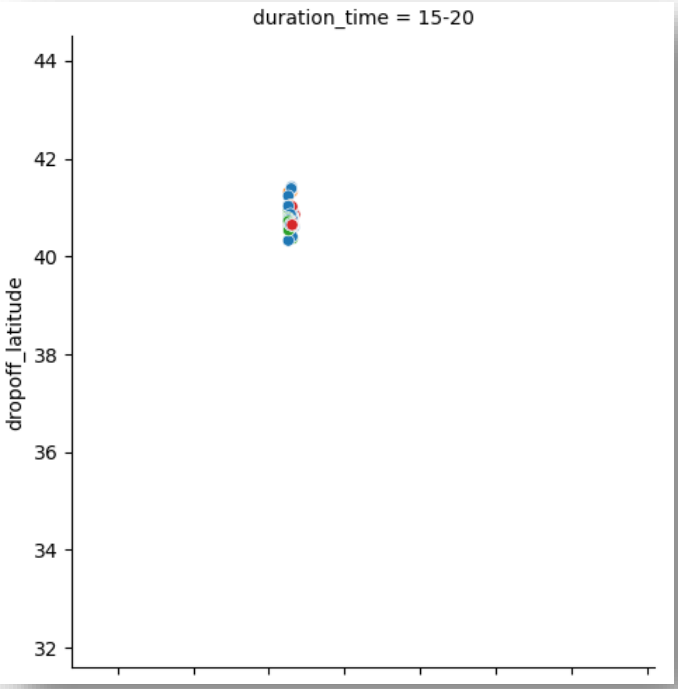
➢ As we also saw during trip duration per month, similarly, trip distance is the lowest in February and the maximum in June.

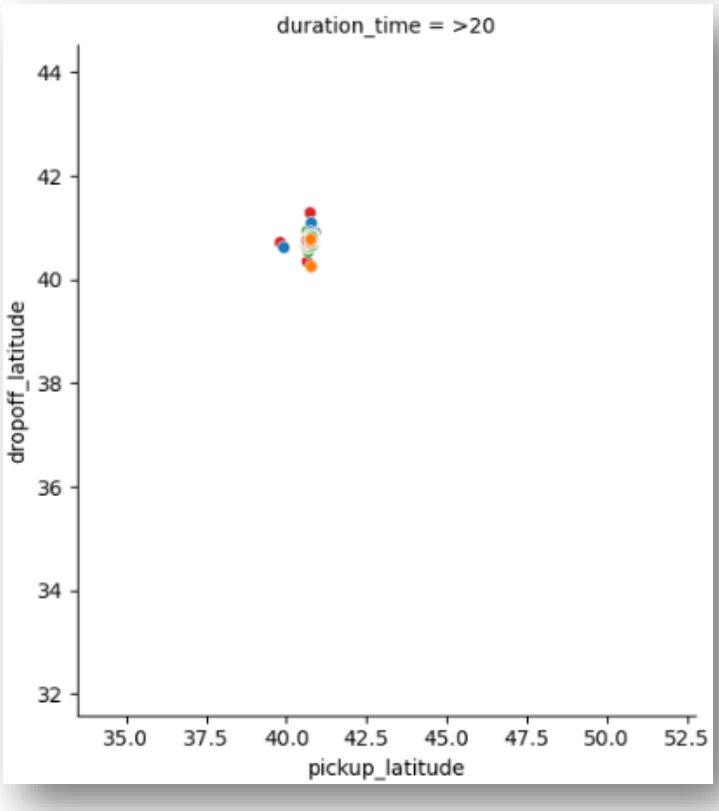
Duration and Geographical Location:



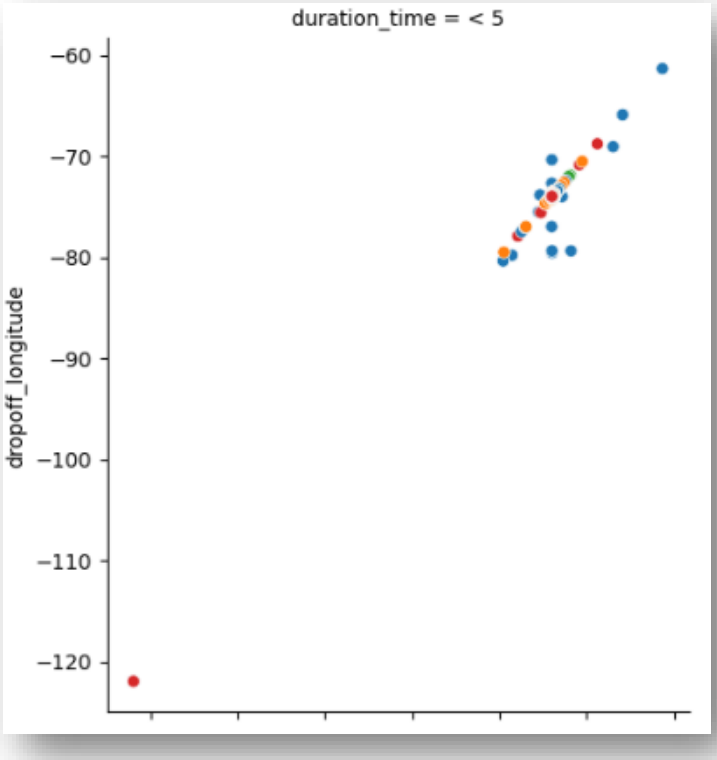


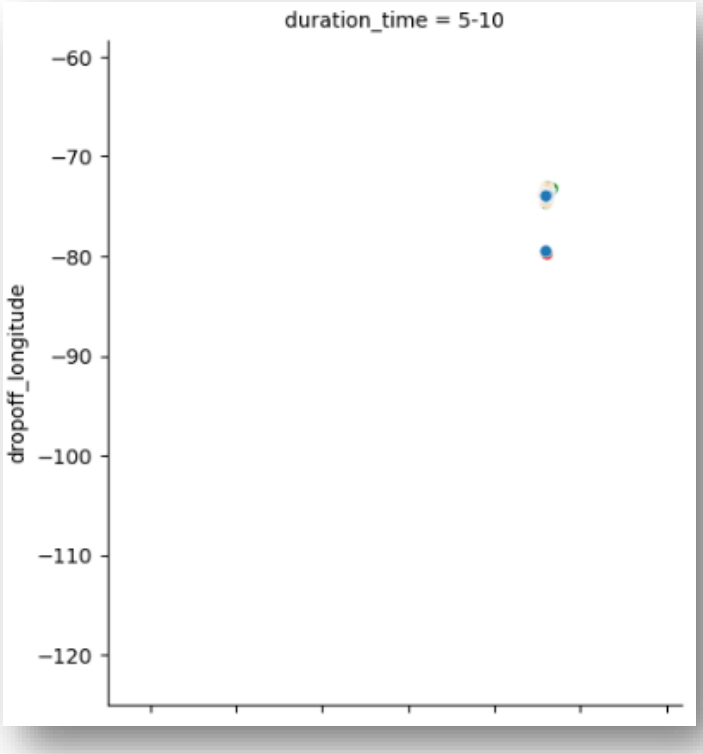


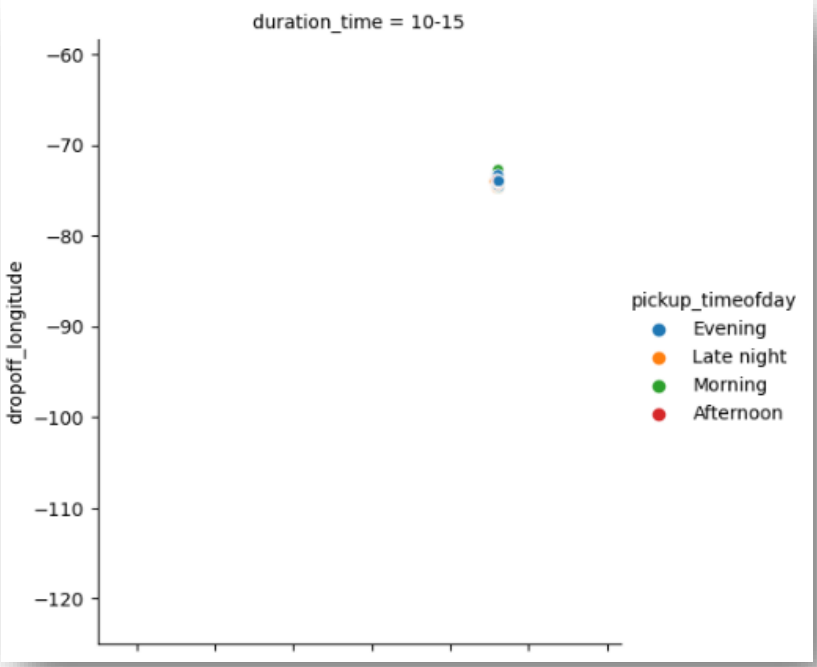


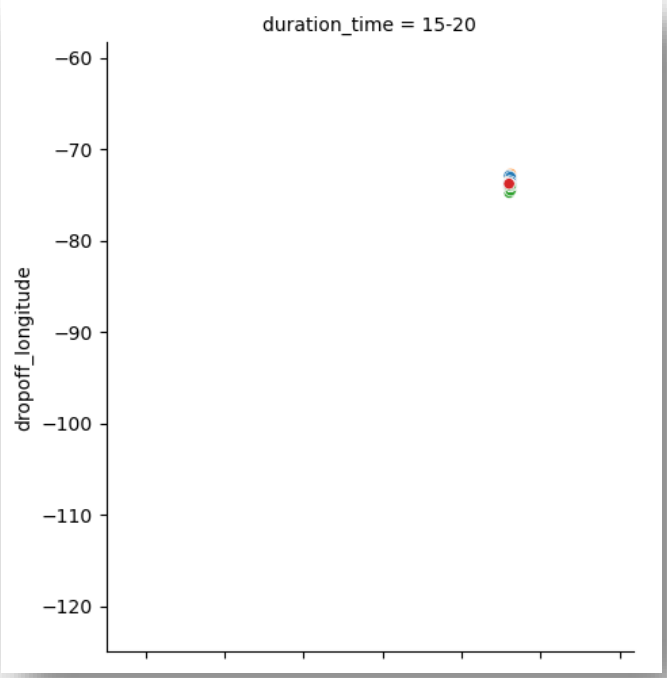


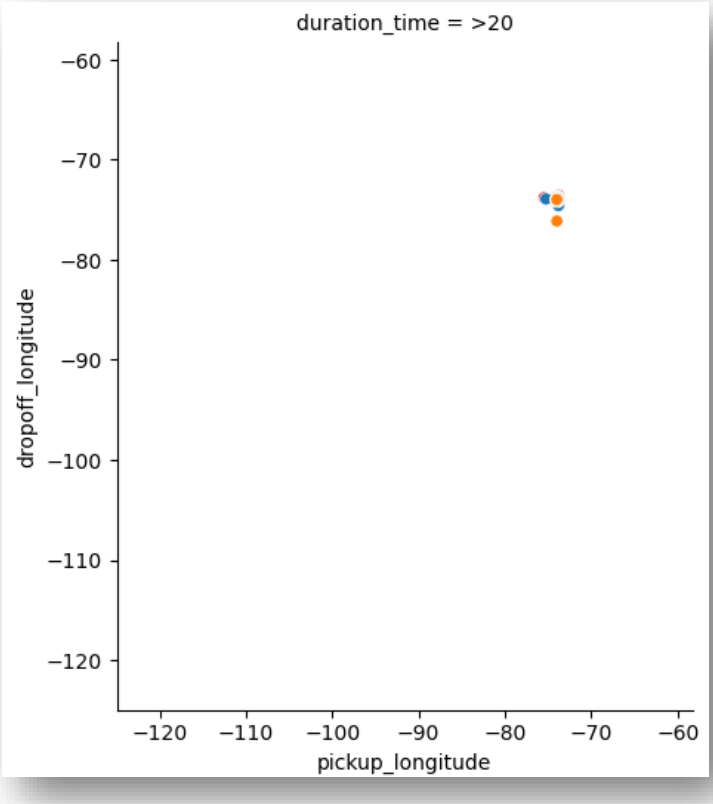
- For shorter trips (<5 hours), the pickup and drop-off latitude is more or less evenly distributed between 30 ° and 40 °
- For longer trips (>5 hours) the pickup and drop-off latitude is all concentrated between 40 ° and 42 ° degrees.







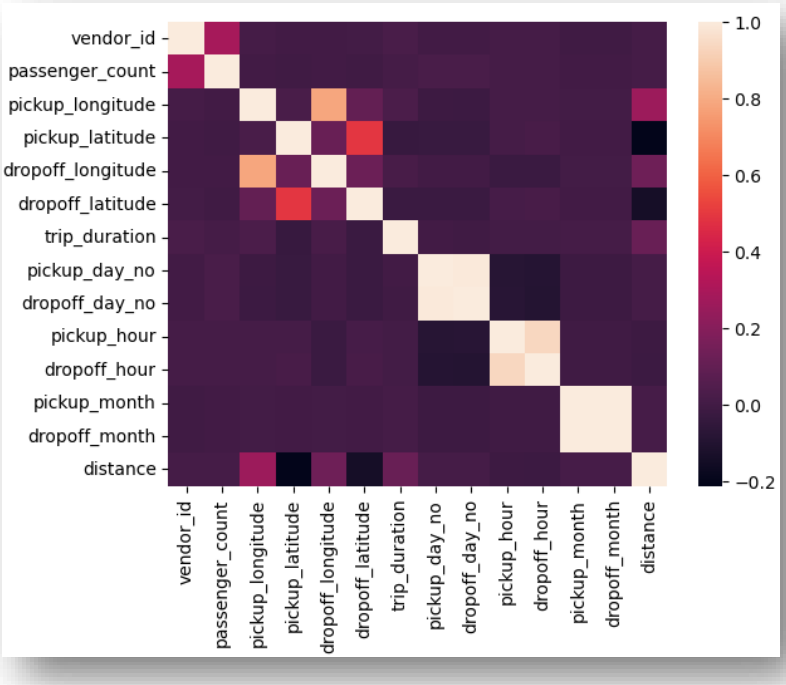




- For shorter trips (<5), the pickup and drop-off longitude are more or less evenly distributed between -80 ° and -65 ° with one outlier near -120 °.
- For longer trips (>5) the pickup and drop-off longitude are all concentrated near -75 °

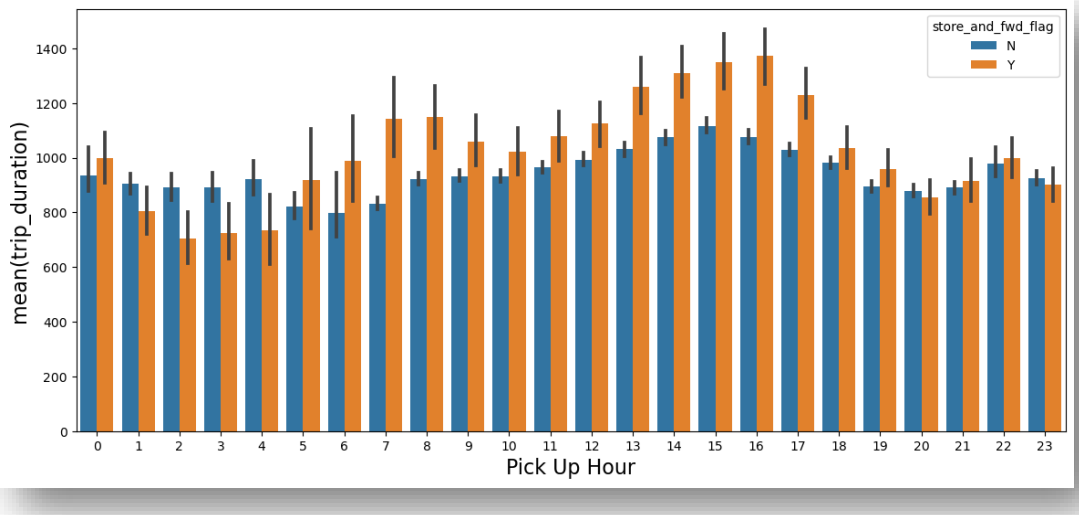
Multivariate Analysis:

Correlation of columns



- Strong Relationship b/w Pickup Day and Drop off Day.
- Moderate Relationship b/w average speed and Distance.

Distribution of Pickup hours, Trip Duration and Store and FWD Flag



- For Longer Trips the Flags were Recorded as compared to shorter trips we can Infer from here.

Conclusion

- Mostly 1 or 2 passengers avail the cab. The instance of large group of people travelling together is rare.
- Vendor 2 mostly provides the longer trips
- Most trips are taken on Friday, Saturday and Thursday
- The most number trips which lasts > 5 hours were taken on Thursday followed by Friday and Wednesday.
- The most number of trips of duration 5–10, 10–15 was taken on Thursday.
- But the highest percentage of trips longer than 20 hours was taken on Sunday and Saturday.
- Distances are longer on Sundays probably because it's weekend.
- Thus, we observe that most pickups and drops occur in the evening. While the least drops and pickups occur during morning.
- The average duration of trips started in between 14 hours and 17 hours is the largest.
- Most of the Passenger Trip Duration is b/w 800 to 1000 sec.
- The long duration trips (> 5 hours) are mostly concentrated with their pickup region near (40 °,75 °) to (42°,75°)

Suggestions

- we see that 1 or 2 passengers' trip are much more than other trip so 4-seater cabs are more prefer than 5- or 6-seater cabs for 1 or 2 passengers
- Thursday, Friday and Saturday are peak days for more trips, so if we provide basic discount then we can garb more trips
- In morning section trips are very less may be having multiple options to travel like Government transport, if we give discount, we can also increase trips in morning section

Dataset: https://www.kaggle.com/competitions/nyc-taxi-trip-duration/data

GitHub: Yamana06/NYC-Taxi-Trip-Duration: Exploratory Data Analysis on NYC Taxi Trip Duration Dataset (github.com)