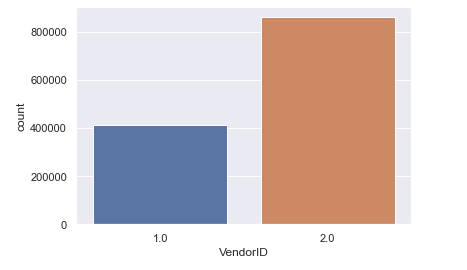
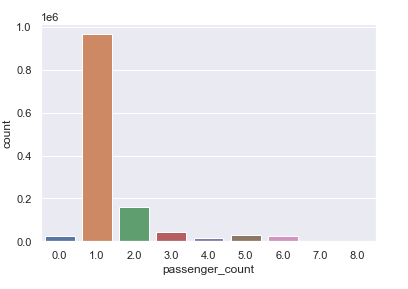
**9. Result**

All graphs are results and all the comments are optimization and evaluation.

**VENDOR ID** - From univariate analysis we can see there is a difference of more than 200000 trips taken by both vendor



**Passenger Count** - From univariate analysis there are some trip with 0 passenger which is an incorrect data or driver has purposely added it to complete a trip We see the highest amount of trips are with 1 passenger. Let us remove the rows which have 0 or 7 or 8passenger count.



**Store and Forward Flag** - From univariate analysis We can observe that less than 1% of journeys were saved before being sent. The number of N flags is significantly higher. We'll see if they have anything to do with the trip's duration afterwards.

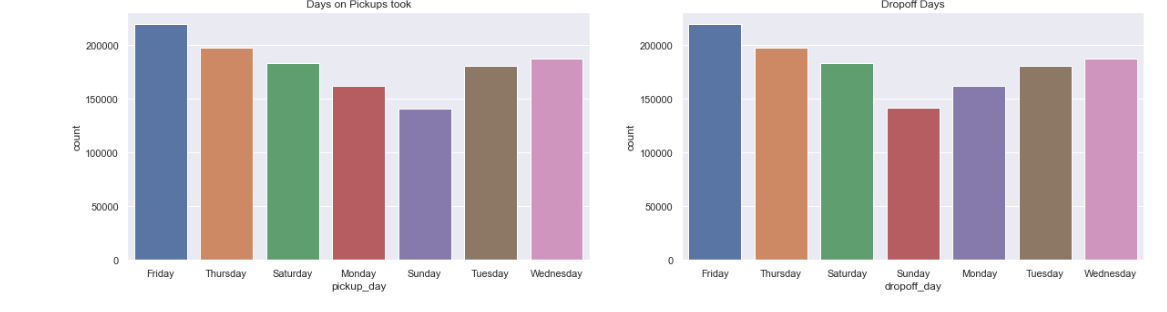
**Distance** - From univariate analysis, In order to find whether there are any trip with 0 kms we need to analyze the data using bivariate analysis. There can be many reason for trips with 0 kms some are mentioned below:

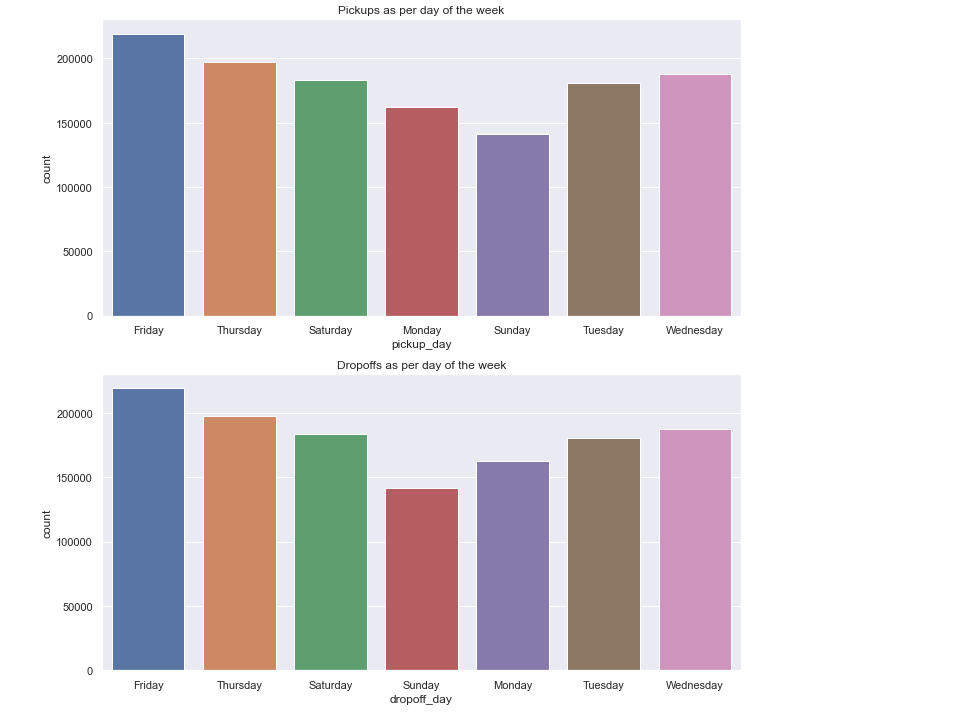
1. Since we don't any latitude and longitude it is possible that drop location was not identified

2. The driver in order to complete required number of ride must have purposely taken trip with 0kms

3. The passenger canceled the trip

**Trips by Day** - From univariate analysis we can see that the day Friday is the day where maximum pickup and dropoff took place. It is mostly likely because of the start of the weekend.

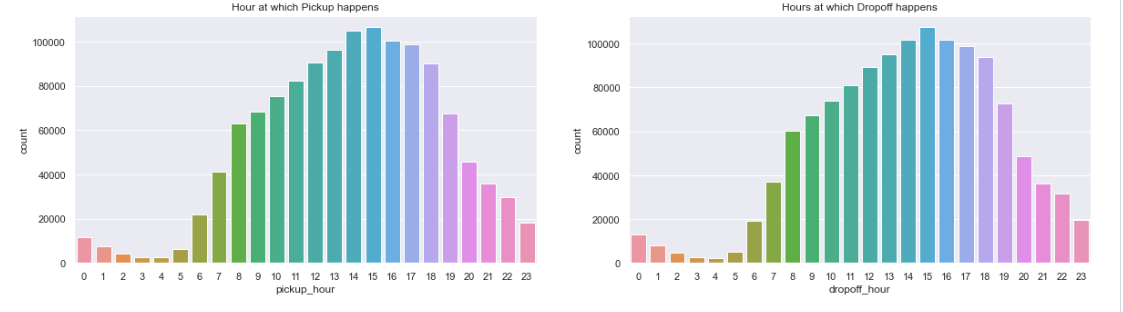


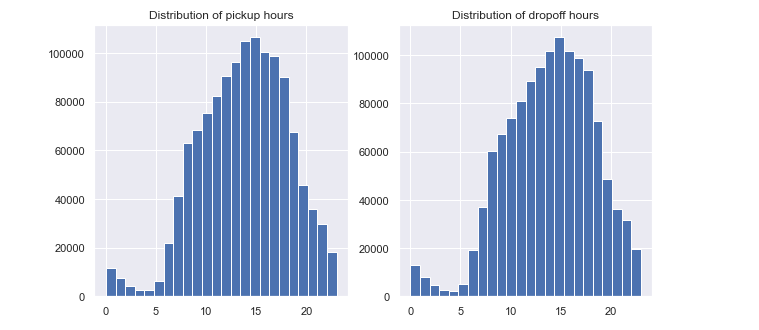


**Trips per Hour** - From univariate analysis we can see the busiest hours is between 9 am till 6 pm which is our office hours

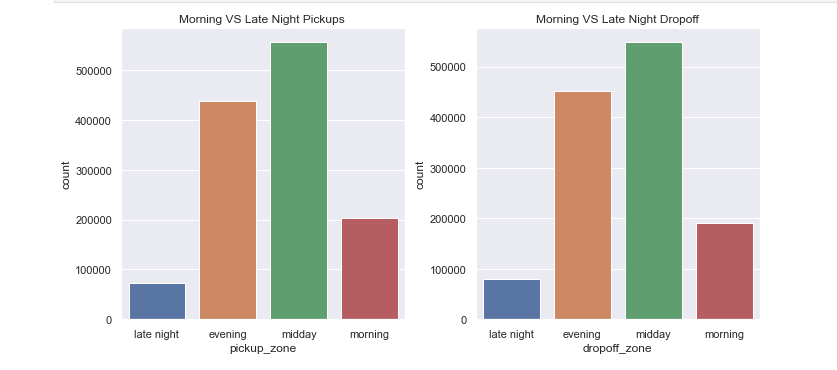
Since it is difficult to analysis time, hence we divide the time in 4 zones:

1. morning (4 hrs to 10 hrs)
2. midday (10 hrs to 16 hrs)
3. evening (16 hrs to 22 hrs)
4. late night (22 hrs to 4 hrs)

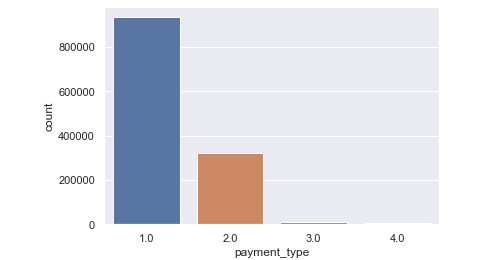




**Trips per Time of Day** - From univariate analysis we can see from the most common form of payment is credit card and there are no trip which is whose payment is unknown or voided

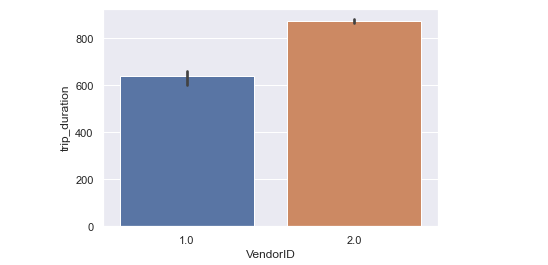


**Payment\_type -** From univariate analysis we can see from the most common form of payment is credit card and there are no trip which is whose payment is unknown or voided

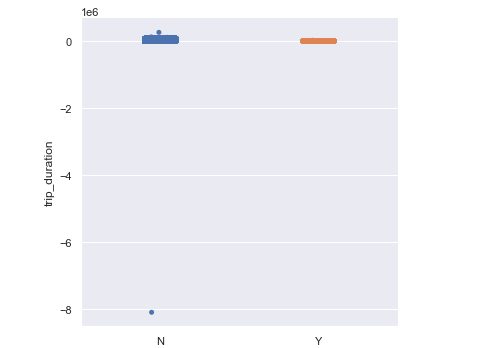


**Trip Duration per Vendor** - From Bivariate Analysis

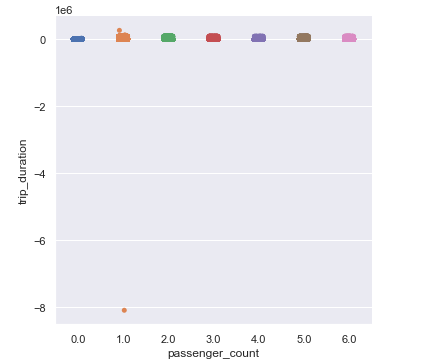
VendorId with value 2 takes longer trips compared to vendor id with value 1



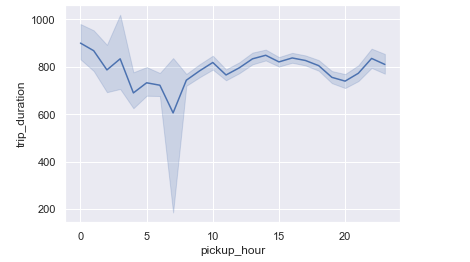
**Trip Duration per Store and Forward Flag** - From Bivariate Analysis we can see trip duration is longer for those whose has unstored flag



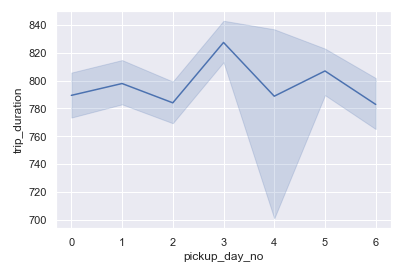
**Trip Duration per passenger count -** From Bivariate Analysis maximum rides are done by single passenger no relationship can be identified

****

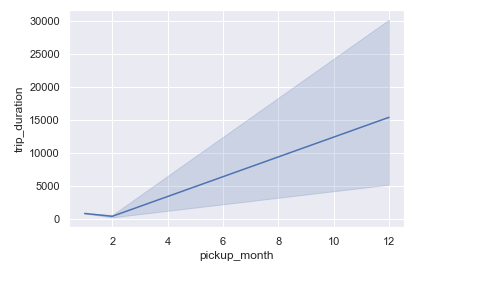
**Trip Duration per hour -** From Bivariate Analysis maximum duration is around 9 am in the morning which is because of office start time causing a lot of traffic



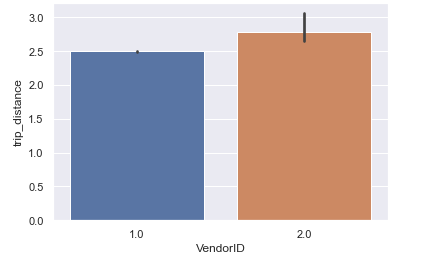
**Trip Duration per time of day -** From Bivariate Analysis we can see trip duration is maximum in late night, evening and afternoon and least in morning



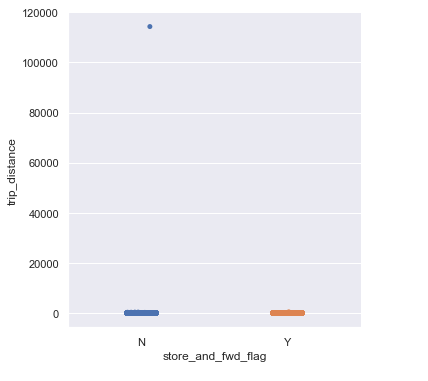
**Trip Duration per month -** From Bivariate Analysis We can see trip duration rising every month



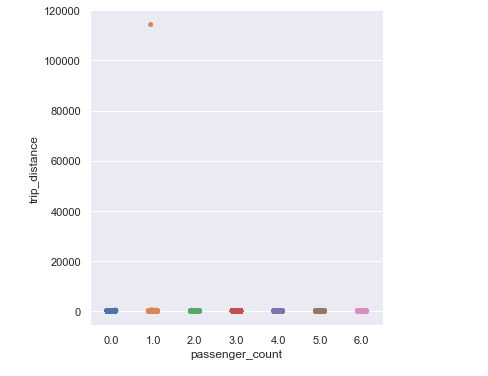
**Distance and Vendor -** From Bivariate Analysis we can see vendor with id as 2 has traveled little more distance than vendor with id as 1



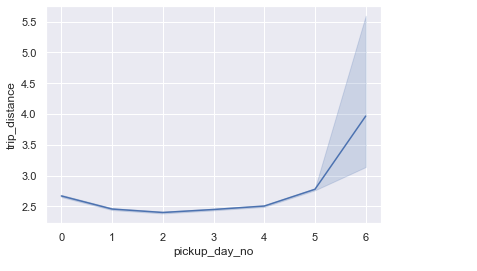
**Distance and Store and Forward Flag -** From Bivariate Analysis we can see longer distance trip are not sorted



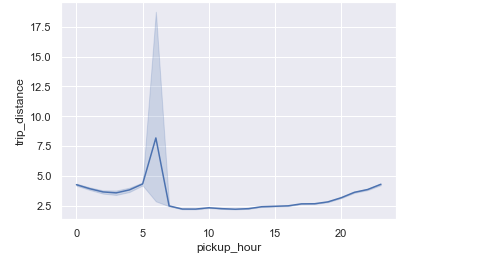
**Distance per passenger count -** From Bivariate Analysis we can see longer distance is traveled when passenger count is 1



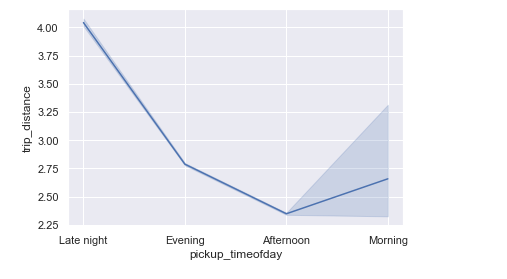
**Distance per day of week -** From Bivariate Analysis we can see Saturday have longer distances, owing to the fact that it is the weekend.

1. The distances traveled on Mondays are also rather shorts.
2. This indicates that outstation travels are possible on these days, and/or that the streets are busier.
3. 

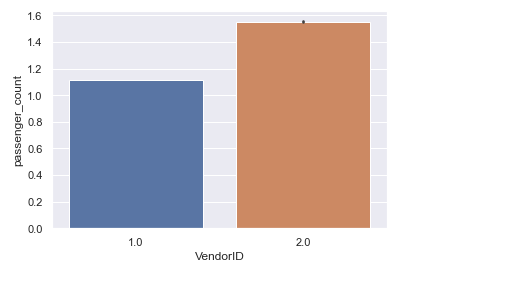
**Distance per hour of day -** From Bivariate Analysis distance is longer around 9am when office time start



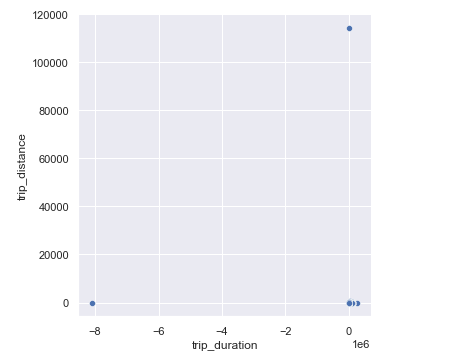
**Distance per time of day -** From Bivariate Analysis distance traveled in late night and evening are more due to longer trips



**Passenger Count and Vendor id -** From Bivariate Analysis this demonstrates that vendor 2 typically transports two people, but vendor 1 only transports one.



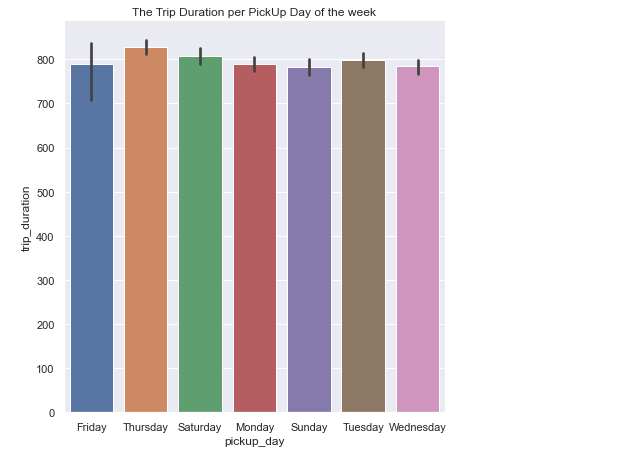
**Trip Duration and Distance -** From Bivariate Analysis we can observe that there are flights that last as little as 0 seconds yet cover a big distance. Trips with zero kilometers traveled and negative trip durations are also possible.

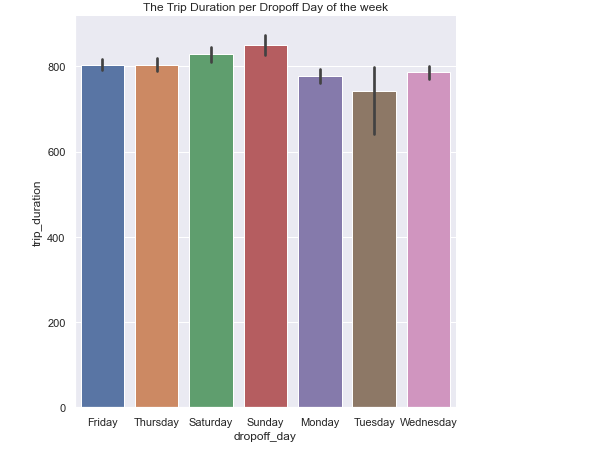


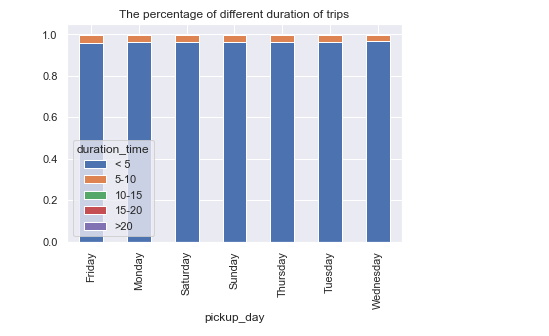
**Trip Duration and The day of the week -** From Bivariate Analysis the graphs show the average travel estimate for each day of the week. The error bars show the degree of uncertainty surrounding that estimate.

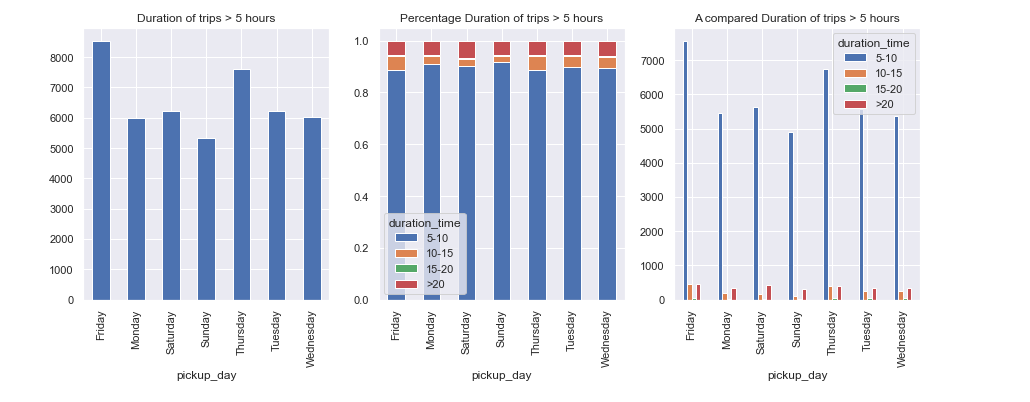
The highest completion time is on Monday while rest other takes lesser time

But this is not enough. We must also take into consideration the percentage of short, medium and long trips taken on each day.

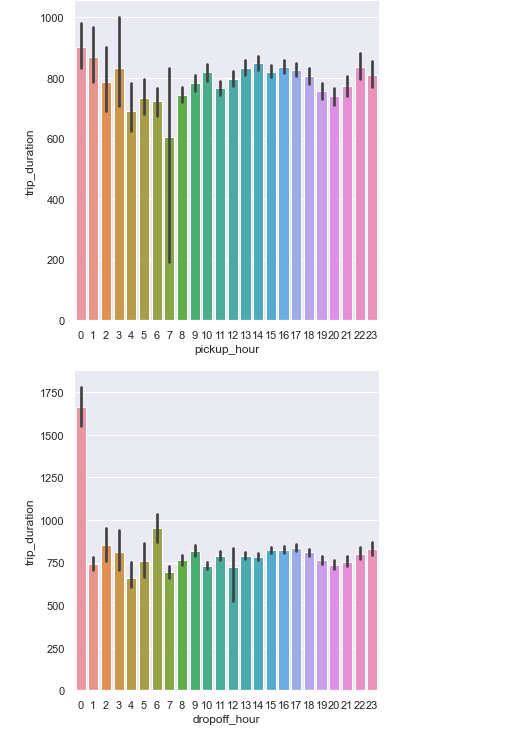




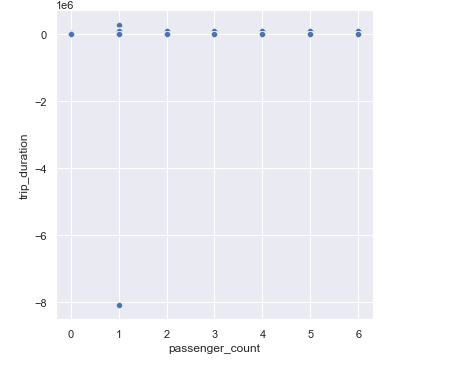




**Trip Duration and The time of the day -**



**Passenger count and duration -**



Future Work, Comments - students may want to consider the following questions

1. What was unique about the data? Did you have to deal with imbalance? What data cleaning did you do? Outlier treatment? Imputation?

The data was unique with respect to every cab taken in NewYork, there are total of 4 cab services and each of them have their own uniqueness. The data consist of various null or empty values where we implement feature scaling and data cleaning process to get desire value. The below are the null values we treated

a. VendorID 98352

b. passenger\_count 98352

c. RatecodeID 98352

d. store\_and\_fwd\_flag 98352

e. payment\_type 98352

Our target variable trip\_duration has extreme right skewness hence it has outlier. We decided to check with largest 10 in them and figure out that the 1st entry is way far than 2nd and 3rd. The highest figure is much higher than the second and third highest trip time values. This might be due to data gathering mistakes, or it could be legitimate information. Because such a large value is unlikely to occur, this row should be dropped before further research. The value can be replaced by the mode or median of trip duration as well

data\_yellow\_cab\_td=data\_yellow\_cab[data\_yellow\_cab.trip\_duration!=data\_yellow\_cab.trip\_duration.max()]

Since there are still extreme right skewness we have divided the trip duration into 5 intervals

The intervals are decided as follows:

1.less than 5 hours

2.5–10 hours

3.10–15 hours

4.15–20 hours

5.more than 20 hours

bins=np.array([0,1800,3600,5400,7200,90000])

data\_yellow\_cab\_td['duration\_time']=pd.cut(data\_yellow\_cab\_td.trip\_duration.astype('timedelta64[s]'),bins,labels=["< 5", "5-10", "10-15","15-20",">20"]

2. Did you create any new additional features / variables?

Since our analysis and machine learning model was completely revolving around the trip\_duration we have added that as field. In addition to this other new variable created are as follow:-

Variable Data Type Description

pickuptime datetime64[ns] Same as tpep\_pickup\_datetime

droptime datetime64[ns] Same as tpep\_dropoff\_datetime

trip\_duration timedelta64[ns]. Difference between pickuptime and droptime

pickuptime\_datetime datetime64[ns]. Same as tpep\_pickup\_datetime

droptime\_datetime datetime64[ns] Same as tpep\_dropoff\_datetime

pickup\_day object Days of week when cab was pickup

dropoff\_day object Days of week when cab was drop-off

pickup\_day\_no int64 Number assigned to each day for pickup

dropoff\_day\_no int64 Number assigned to each day for drop-off

pickup\_hour int64 Hour at which cab was pickup

dropoff\_hour int64 Hour at which cab was dropped

pickup\_month int64 Month at which cab was pickup

dropoff\_month int64 Month at which cab was dropped

pickup\_timeofday object Time at which cab was pickup

dropoff\_timeofday object Time at which cab was Dropped

duration\_time category Time in bins

3. What was the process you used for evaluation? What was the best result?

We used univariate and bivariate analysis for evaluation. Both gave us best result, but bivariate analysis helps us to identify the finding links, patterns, and correlations between two variables. With bivariate analysis we were able to identify wrongly inputted data, relationship between different variable,etc

4. What were the problems you faced? How did you solve them?

The location where provided in form of ID not in form of latitude and longitude. Hence creating map of newyork to show the distribution was not possible.

The time provide was also in datetime , hence the target variable time\_duration was creating by timedelta which needs to be converted to datetime every single time. There were extreme right skewness causing outlines which was resolved by dividng the time interval further but it couldn’t helped.

5. What future work would you like to do?

Since the data present is in huge amount from Jan 2022 till Dec 2000, we will in future stored the data in S3 , create our jupyter notebook on AWS and by using Sagemake studio and notebook instance we will create automated ML models.