Exploratory Data Analysis
Project Report
NYC Taxi Dataset

Authors : dvasani,pphadke

Statement of goals

The goal of this project is to determine the trip duration of taxi rides in New York City. New York City is one of the busiest cities not only in the United States but globally. Hence, it is helpful to analyze the duration of a ride and then decide if the user wants to take a cab or choose an alternate mode of transportation like a subway or walk. The questions we are trying to address is to figure out which factors affect the taxi ride trip duration.

Data description

We have trip data from January to June of 2016. Let's take a look at the data fields and the first few rows:

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - start date and time of the trip
- dropoff_datetime - end date and time of the trip
- passenger_count - the number of passengers
- pickup_longitude - start longitude of the trip
- pickup_latitude - start latitude of the trip
- dropoff_longitude - end longitude trip
- dropoff_latitude - end latitude trip
- 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 - duration of the trip in seconds

```{r}
glimpse(train)
```

```
Rows: 1,428,714
Columns: 11
$ id                <chr> "id2875421", "id2377394", "id3858529", "id3504673", "id2181028", "id0801584", "id18132...
$ vendor_id         <int> 2, 1, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 1, 2, 1, 2, 1, 1, 2, 2, 1, 1, 2,...
$ pickup_datetime   <chr> "2016-03-14 17:24:55", "2016-06-12 00:43:35", "2016-01-19 11:35:24", "2016-04-06 19:32...
$ dropoff_datetime  <chr> "2016-03-14 17:32:30", "2016-06-12 00:54:38", "2016-01-19 12:10:48", "2016-04-06 19:39...
$ passenger_count   <int> 1, 1, 1, 1, 1, 6, 4, 1, 1, 1, 1, 4, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
$ pickup_longitude  <dbl> -73.98215, -73.98042, -73.97903, -74.01004, -73.97305, -73.98286, -73.96902, -73.96928...
$ pickup_latitude   <dbl> 40.76794, 40.73856, 40.76394, 40.71997, 40.79321, 40.74220, 40.75784, 40.79778, 40.738...
$ dropoff_longitude <dbl> -73.96463, -73.99948, -74.00533, -74.01227, -73.97292, -73.99208, -73.95741, -73.92247...
$ dropoff_latitude  <dbl> 40.76560, 40.73115, 40.71009, 40.70672, 40.78252, 40.74918, 40.76590, 40.76056, 40.732...
$ store_and_fwd_flag <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "...
$ trip_duration     <int> 455, 663, 2124, 429, 435, 443, 341, 1551, 255, 1225, 1274, 1128, 1114, 260, 1414, 211,...
```

As we can see from the picture above there are 1428714 rows and 11 columns in the dataset. We have abundant data and the next step was to check which columns would help train the model well or any feature engineering is necessary.

Feature Engineering:

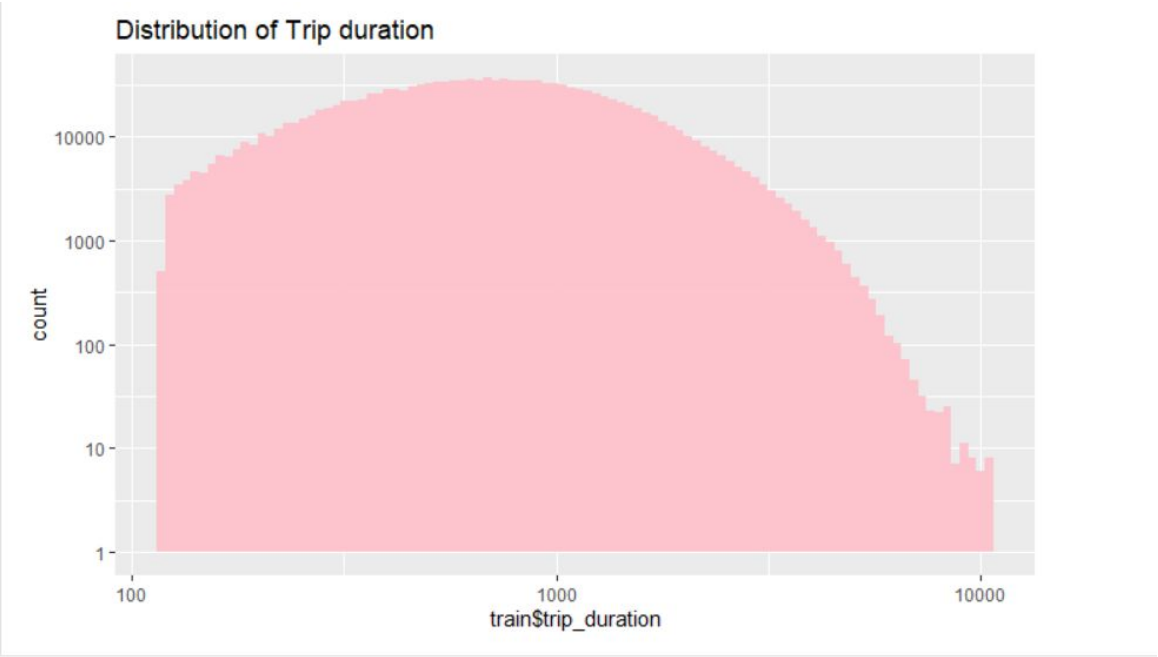
The first step in feature engineering is to extract various attributes from the date-time columns i.e. pickup datetime and dropoff datetime. We extract attributes like the hour of the day, day of the week, week number, and was it a weekday or a weekend. We expect these attributes to capture various seasonal trends in the duration of the taxi ride. We believe that by plotting some of these trends we will get a better idea about the factors that affect ride duration.
We also use the pickup and dropoff longitudes and latitudes to extract the distance in kilometers of these rides. We expect rides with greater distances to have a higher duration however it is not an ideal system and there are often other known or unknown factors that affect the typical behavior of our dependent variable.
The last piece of feature engineering is to extract the speed based on distance and time. The unit of speed is kilometer/hour.

Data cleaning and exploration:

Even though our dataset did not contain any null values, a lot of samples did not fall in a reasonable range. An example of the same would be a taxi ride that lasted 40 days. In order to handle these outliers, we studied the various data fields.

New York is a long city. The farthest points in New York are about 61 kilometers apart. Hence we only considered trips under 100 kilometers to keep a safe margin. The next field we rectified was passenger count. Some of the rides had up to 9 passengers. We set a limit on the number of passengers to 6 per ride. Also, there are entries in the dataset with 0 passenger count. We were not sure if these were in fact outliers or the taxi driver going to fill gas or making a trip home. Hence, we keep these entries.

The final field we corrected was our target variable trip duration. We limit it to a minimum of 2 minutes and a maximum of 3 hours.New York is also known for the traffic and to account for that even though 3 hours might seem a lot we have set that limit. It could happen that the taxi was stuck even for a short ride.
Let’s check the distribution of trip duration to understand if our data is balanced or not.

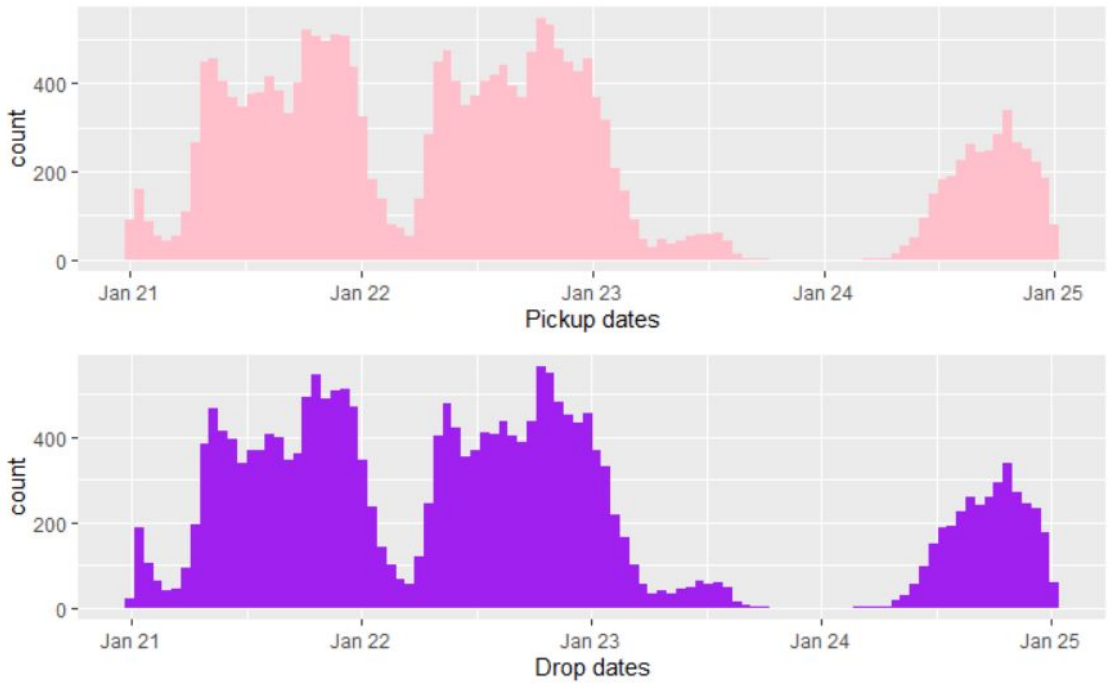


As we can see, there are more trips on the shorter duration side till the mid after which the trend drops. This means our data is a little skewed.

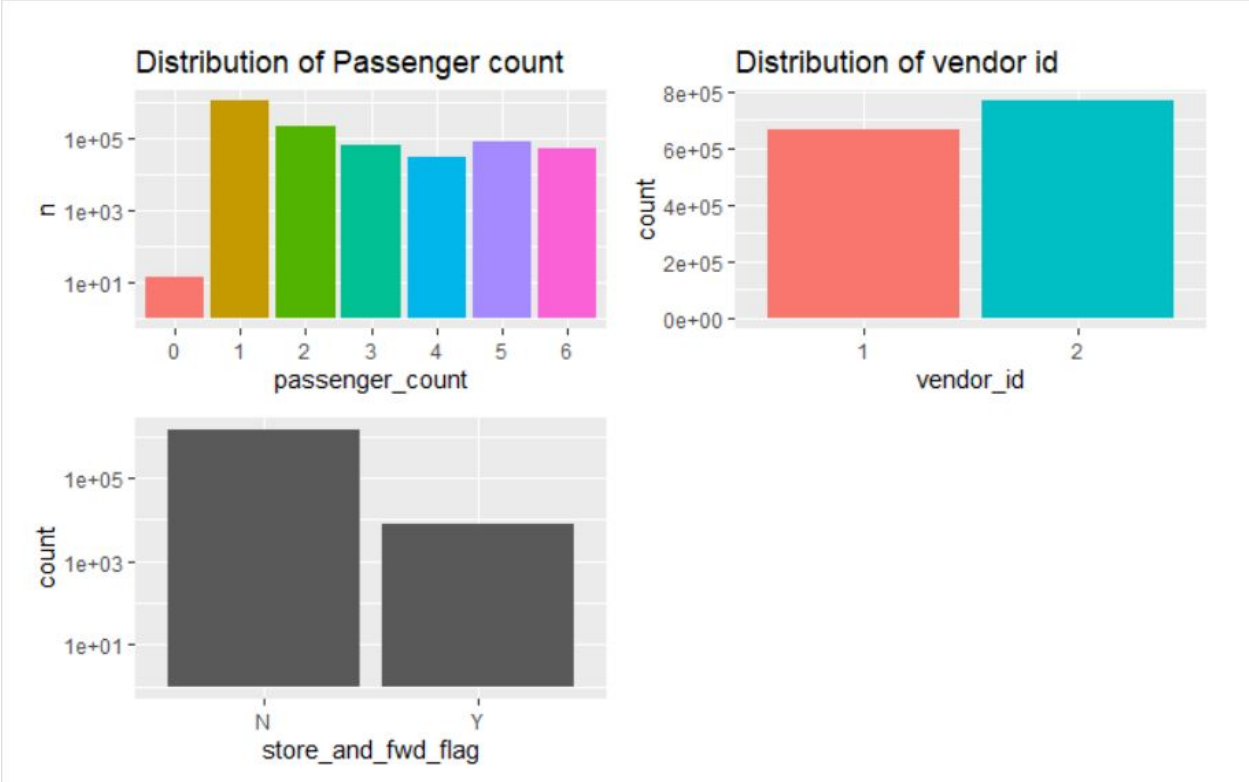
Date distribution

We now check the number of trips with respect to pick up and drop off dates.

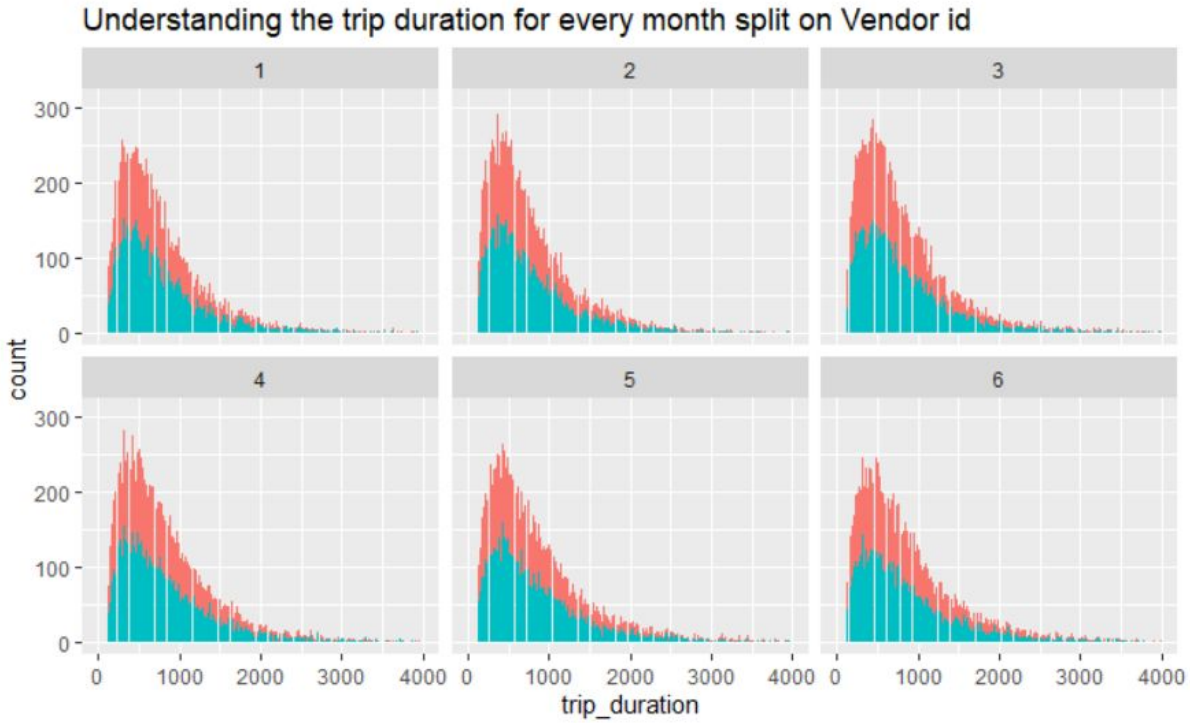
We see that the trend is quite flat except for somewhere near the end of January.
Hence we take a closer look at the sudden drop in rides for the specific period in January.



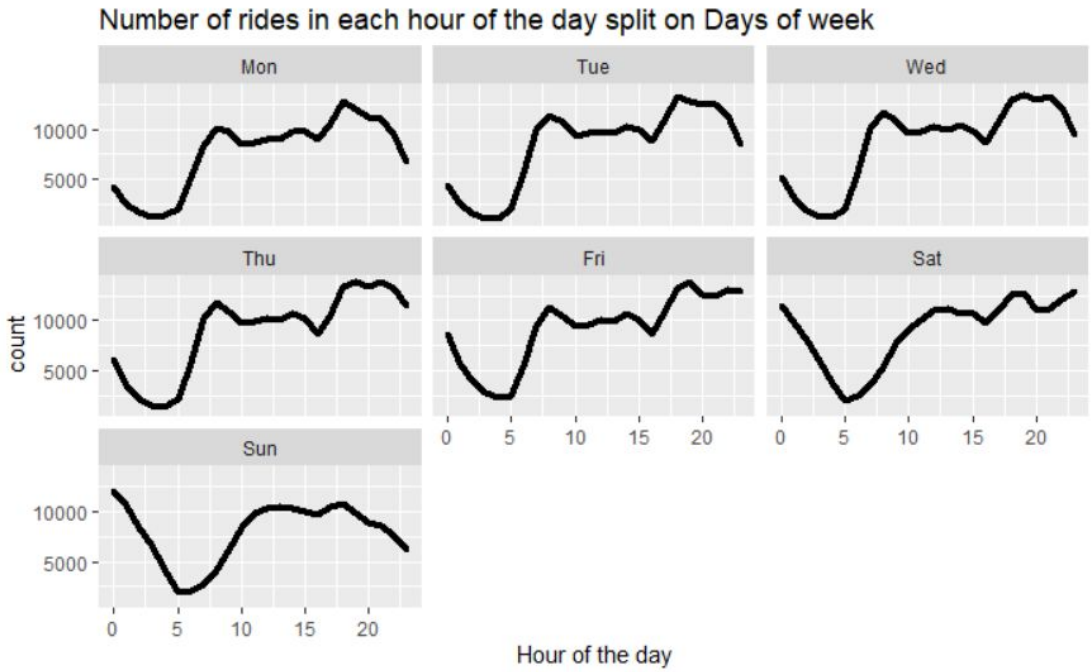
A closer look tells us this dip is between 22nd and 24th January. A quick Google search confirmed that between 22nd and 24th January 2016, there was a blizzard in New York. That explains the dip. Next, we take a look at histograms for passenger count and vendor id. This is an assurance that the data is valid and mirrors the real world scenarios.



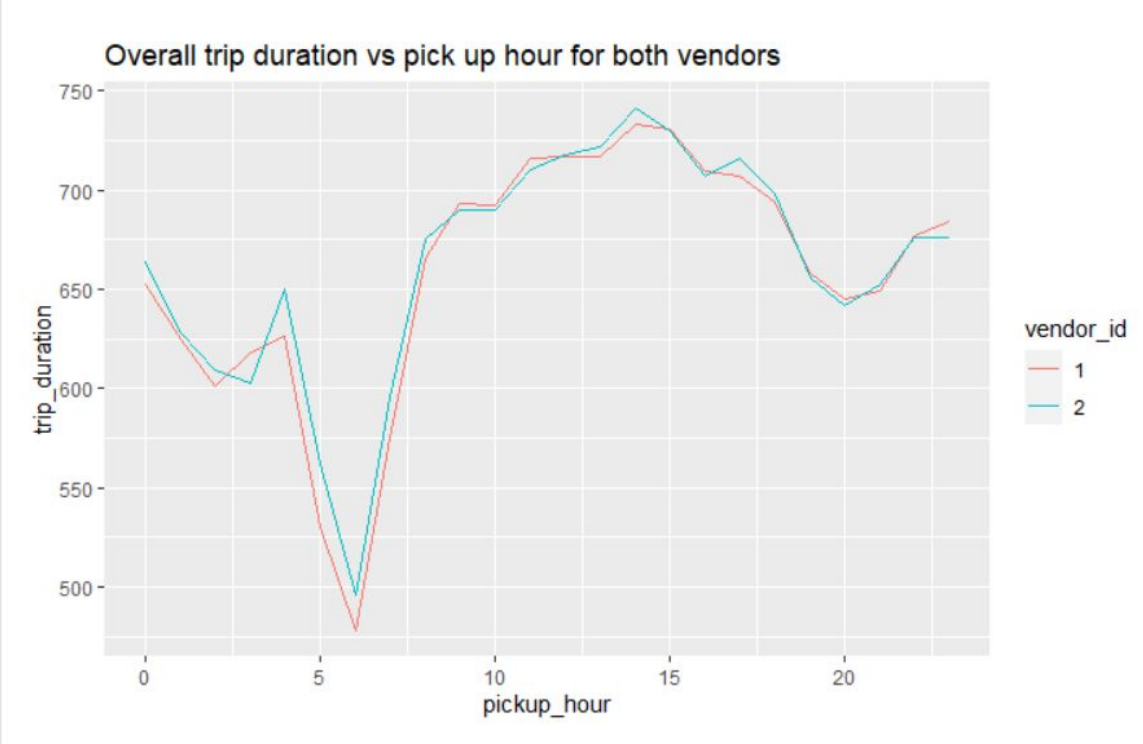
We don’t see any trend in any of these columns that might help our model. To think of reasonably, the trip duration does not change no matter how many passengers are present in the car. Thus, it makes sense to not include these variables in the model. Hence we exclude them from the modeling process. We now check the trip duration according to the month to see if it varies based on the weather.



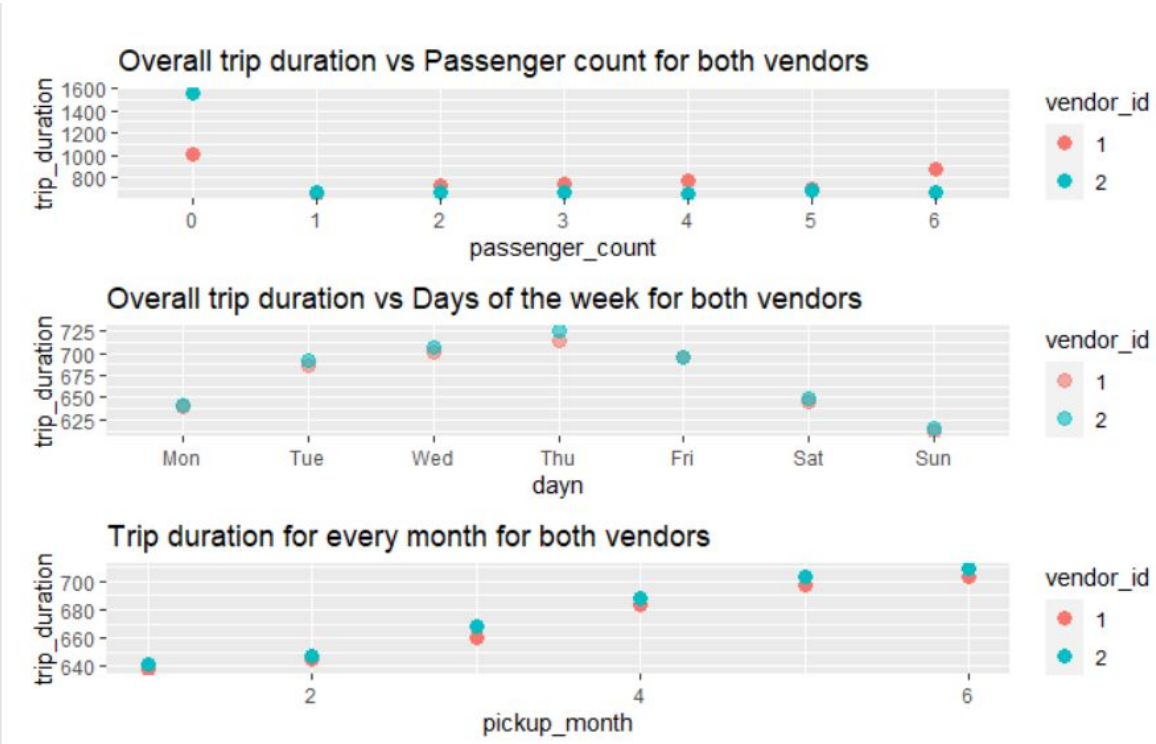
The trip duration follows a pretty common trend across all 6 months. Hence, including the month in the model might not be useful as well. Now let’s check the number of rides per day. We expect more rides on weekdays than weekends. And we see that it does follow that trend. The number of rides on weekends is slightly lesser than weekdays.



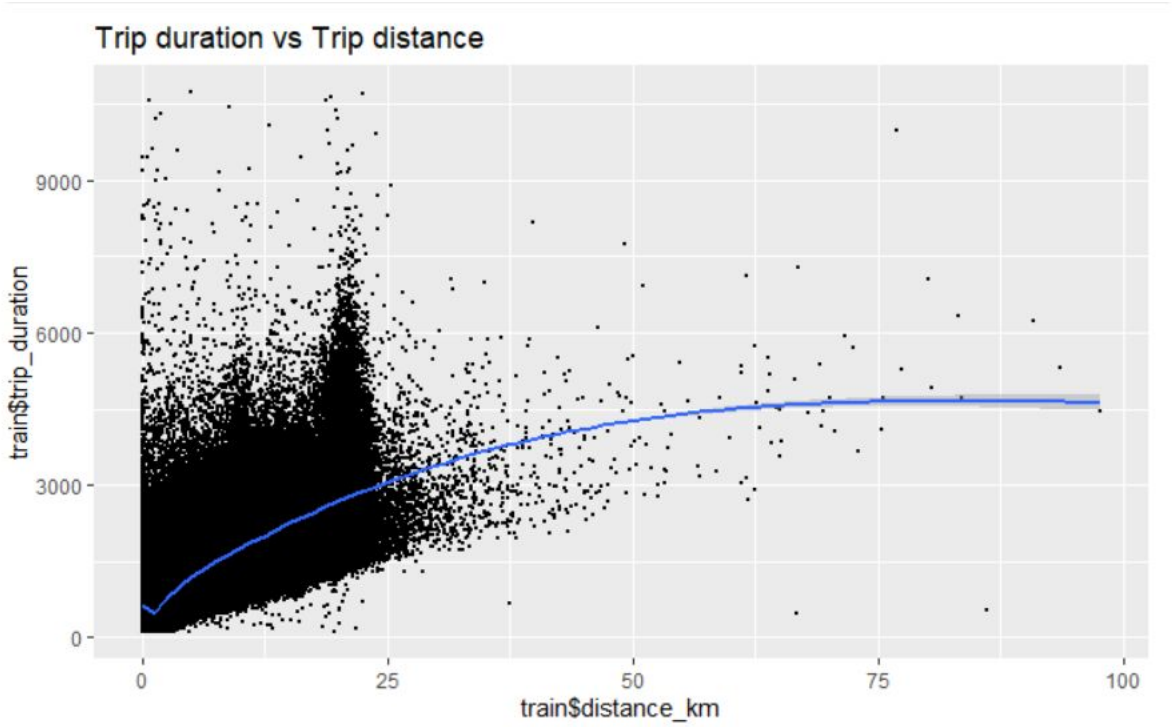
It also shows peaks or troughs where expected. Early morning the number of rides decreases drastically while during early office hours and late evening they peak. We now plot the pickup hour versus trip duration and facet it on vendor id. We don't see any difference between the two vendors.



We check for more trends of trip duration based on other factors.



The plots show a pattern however trip duration is in seconds and a variation of 5-10 seconds doesn't account for much hence these plots are essentially flat.
Finally, we plot the distance vs trip duration.



Distance is the most important factor in predicting trip duration. The plot above confirms that. As in the real world, this scenario is reasonable which assures us that the data is valid and represents real world taxi trends.
Now that we have analyzed our data, we can move on to training a model. In order to do so, we convert our dependent variable into 5 categories since accurately predicting a trip duration in seconds is an extremely difficult task.

The 5 categories are as follows:

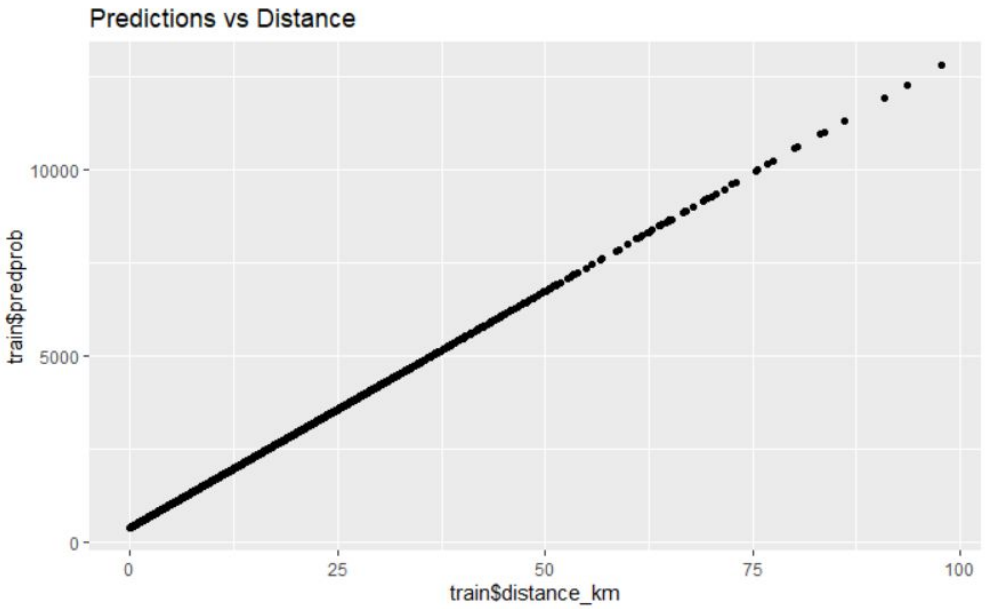
- Very short : 2 to 10 minutes
- Short : 10 to 30 minutes
- Medium : 30 to 60 minutes
- Long : 1 hour to 2 hours
- Very long : 2 hours to 3 hours

Modeling

From the plots, it seems clear that distance is a driving factor to predict the trip duration. They follow a linear relationship with larger distances having longer trip durations. Hence we train two models with our data.

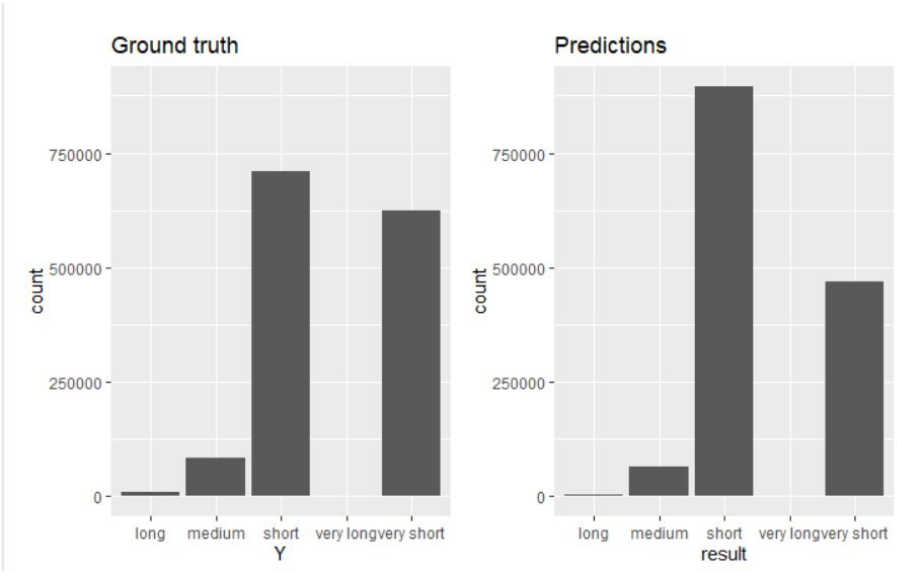
Model 1:

In the first model, we try to predict the trip duration solely based on distance. We train a basic linear model without any transformations on the data and we are able to achieve an accuracy of about 72%. The plot of the predicted probabilities vs distance looks as follows:



The relationship again makes a lot of sense. The slope is evenly increasing and seems to have a good linear relationship. As mentioned before, we expected distance to be the main predictor but, there are interferences of other factors usually and to make sure that we consider these, we will train another model inclusive of a few more predictors.

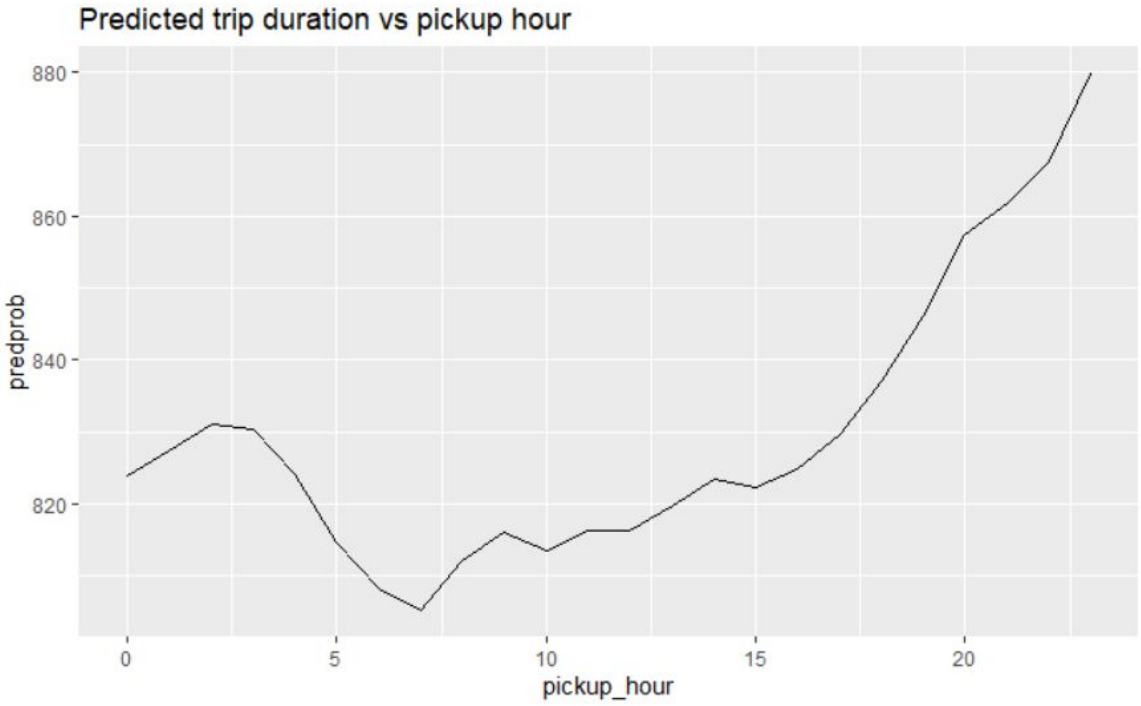
We also plot histograms of our ground truths and predictions and see that they look quite similar. Ground Truth is the actual trip duration in the data (left) and predictions (right).



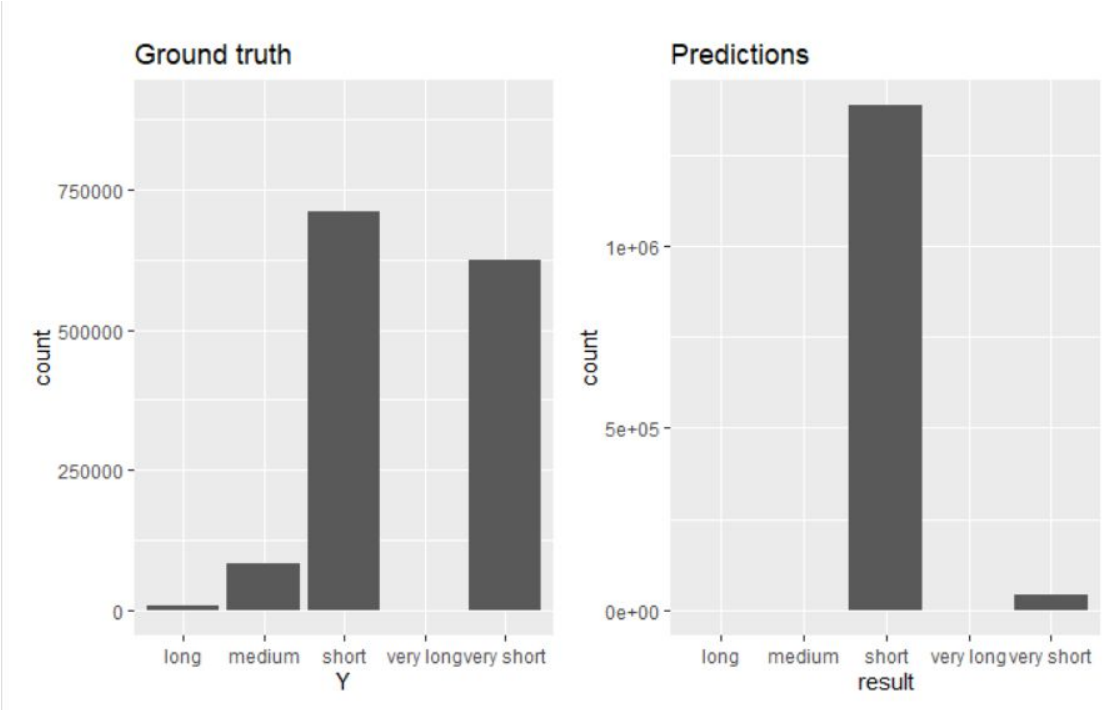
Model 2:

The second model we train is a glm based on most of the other factors except distance. We want to check the relationship of trip duration with the other factors. We believe that distance would dominate the model's predictions hence we keep it separate. We also exclude the dropoff datetime because that is usually not known when we start a ride, and if we include it, we basically give the model the trip duration and our model will horribly overfit.

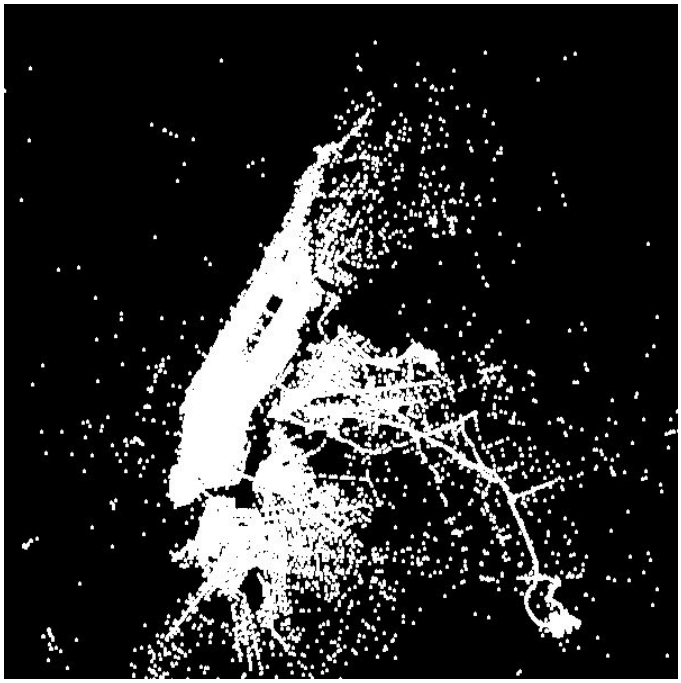
This model doesn't perform very well and gives an accuracy of about 50%. To check this, we plot the predicted probabilities vs hour of the day.



And we see that our model clearly misses the trend. Early morning at 5 am it shows longer trip durations when we've already seen before that there were hardly any trips early morning.



Output



The plot above represents in white, the trip start points in the expanse of New York City. The completely white dense part shows the concentration of trips originating in the West side.

Conclusion & Limitation

Distance is a very important factor in determining trip duration. Although, there are factors not present in the dataframe that are driving the results to be away from the fitted model. It might be Traffic, some holiday or summer break season. The point to be understood here is that the data collected needs to capture more aspects of the world.
The model also is not tuned perfectly. That can be improved as well.

Future work

Improving the metric for evaluation of the model.
Using data for factors like traffic which affect the trends in the data.
Building a better way to visualize the results.