Probability Project

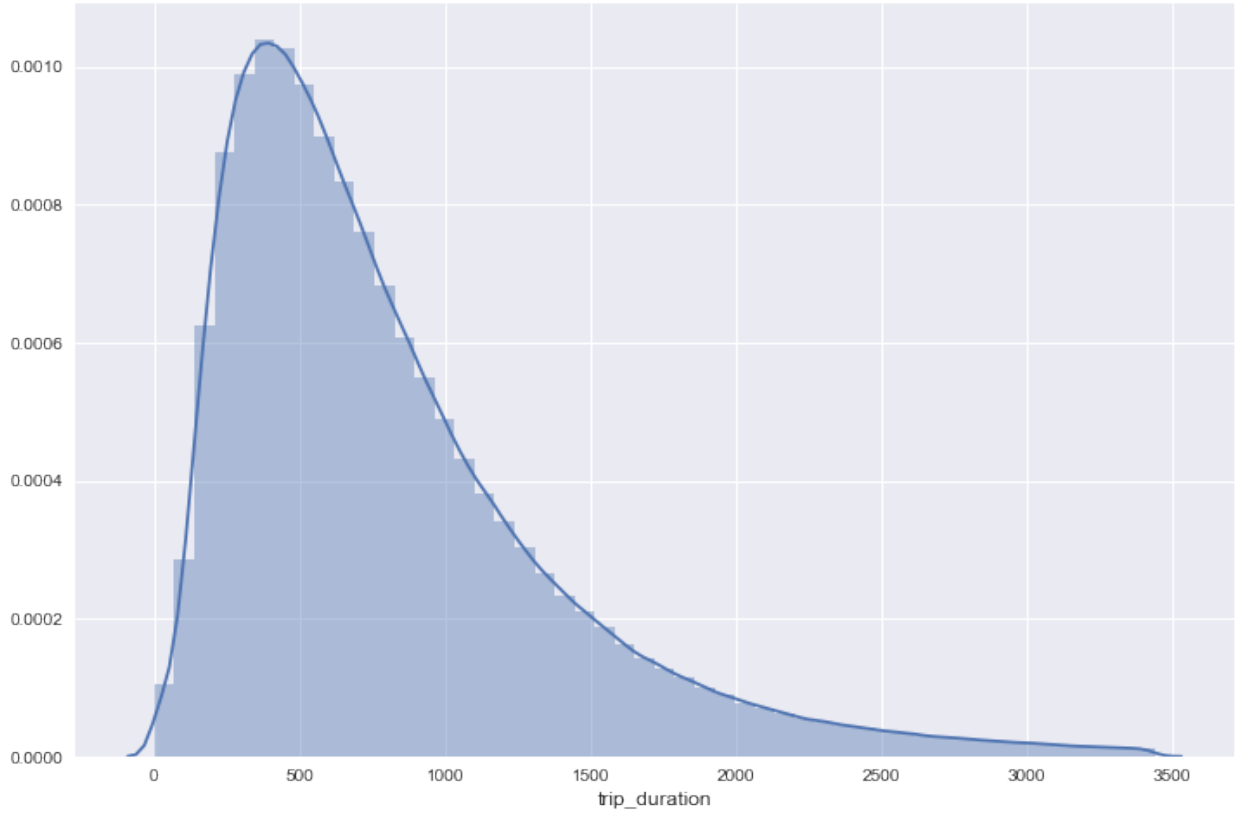
Team Members:

Likhith Kumar
Hari Krishna
Sai Kiran

Q1) If you had to fit a probability distribution to the trip duration, how would you do it and why ?

The original data had extreme outliers – In the order of 2000000 in the feature trip duration. Hence we have removed them before further analysis.

We plotted a density curve for the trip duration. It has Gaussian Distribution skewed to the right with mean 807.46 and standard deviation 575.42



The distribution is a right tailed. So, using mean would be inappropriate. Hence using median for the analysis.

Q2) Divide the data you are using into two parts (Training and Testing), and analyze it using the distribution you mentioned in Q1. Basically you have to learn the parameters from the training set and analyze the applicability of the fitted model on the testing set.

Using the GaussianProcessRegressor in the sklearn library to model this data, we got the following results.
The root mean square error for the training set is 333.008
The root mean square error for the test set is 21337.652

Since the RMSE difference of train and test sets is high using GaussianProcessRegressor for this data with this specific train, test split it is not appropriate to use GaussianProcessRegressor to model this data.

Q3) Given you have to reach a certain location (any area code of your choice) at 2 pm what should be the estimated start time?

We do not have area codes for our data. Hence clustering on latitudes and longitudes to come up with clusters (we chose 15 clusters).

The Median time taken (choosing cluster 3 as our destination) from any cluster to cluster – 3 is approximately 1000 seconds i.e. 16.6 minutes. Therefore, the estimated start time should be 1:43:24 PM.

Alternatively, if we want more accuracy with estimated start time, we can have estimated start times by pickup_cluster number.
For example,
Eg – 1) Median trip_duration from cluster – 0 to cluster – 3 is 1107 seconds ~ 18.45 minutes.
Hence the estimated start time should be 1:41:33 PM.

Eg – 2) Median trip_duration from cluster – 1 to cluster – 3 is 723 seconds ~ 12 minutes.
Hence the estimated start time should be 1:48:00 PM.
Sample of the matrix is below:

	pickup_cluster	dropoff_cluster	trip_duration
3	0	3	1107.5
16	1	3	723.0
29	2	3	715.0
56	4	3	2521.0
70	5	3	1246.0

Work Distribution:

Likith Kumar – Question1

Hari Krishna – Question2

Sai Kiran – Question3