Datasets:

The seven input features are:

* id: a unique identifier for each trip.
* vendor\_id: a code indicating the provider associated with the trip record.
* pickup\_datetime: date & time when the meter was engaged.
* dropoff\_datetime: date & time when the meter was disengaged.
* passenger\_count: the number of passengers in the vehicle.
* pickup\_longitude: the longitude when the meter was engaged.
* pickup\_latitude: the latitude when the meter was engaged.
* dropoff\_longitude: the longitude when the meter was disengaged
* dropoff\_latitude: the latitude when the meter was disengaged.
* store\_and\_fwd\_flag: the flag telling whether the trip was recorded in vehicle memory.

The output label is:

* y: the total duration of the trip.

From our point of view after seeing the data briefly, we can see that it has more than million records, so we will not have any issues with dropping of some records. There are two columns which is in datetime format (YYYY-MM-DD HH:MM:SS). Four columns containing the geo location of the trip. The `store\_and\_fwd\_flag` contains binary values, which can be label-encoded. The `passenger\_count` columns is important for now, and let us not decide on `vendor\_id` that quick as well.

Datasets Preprocessing:

1. Check the shape of our datasets. They are in valid shape.
2. Check the null values in dataset. It appears that there is no null values in our dataset.
3. Label-encode the `store\_and\_fwd\_flag`.
4. Convert the `pickup\_datetime` and `dropoff\_datetime` into actual datetime objects.
5. Drop the `id` and `vendor\_id`, because it feels like they are just a placeholder at this point.
6. Feature engineering: From datetime columns we had earlier, we will now take break down datetime into a couple of important features such as `day` & `hour`. I thought of taking months as well, but it seems there is only data from first six months. Now, we should have four additional features: `pickup\_day`, `pickup\_hour`, `dropoff\_day`, `dropoff\_hour`. We can drop `dropoff\_day`, because I feel like there is no ride which takes more than 1 days to complete. Further, `dropoff\_hour` can be dropped as well as it has no significance. Now, I have created a new features based upon the time, categorizing time into four different settings: morning, noon, evening & night
7. Standardization:
   1. Here, we will standardize all the numerical columns we have. And, then for categorical columns, we will one hot encode them.