# Forecasting Bicycle Demand

To forecast bicycle, demand every station must be taken separately. Location of a station is one of the prime factors when it comes to bicycle demand. So, take every station and forecast for it. Assuming one station is independent from the other.

To evaluate how different variables are related to number of transactions at a station. Start with a simple linear regression model and then move to non-linear machine learning techniques depending on the results. In this case linear regression then, random forest and then time-series analysis is used. All models are compared by their performance on test-data.

After doing the data pre-processing dummy variables for pass-type (walkup, one day, monthly and annually) and trip-type (round trip & one-way trip) are added. Then total number of combinations are 8 for every pass-type and trip-type.

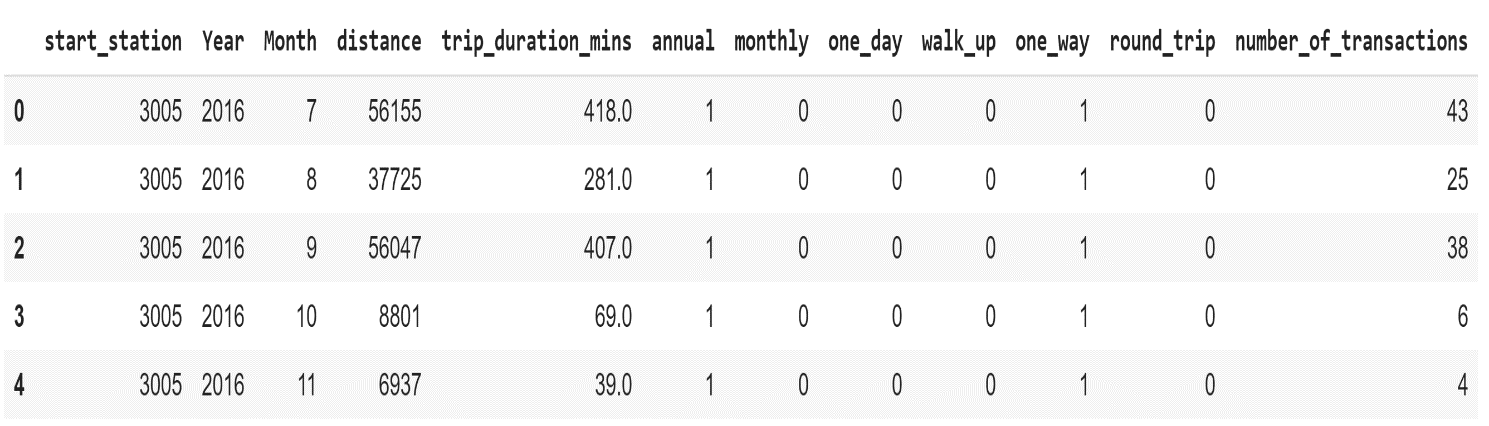
Further data for every station is separated. Then monthly group the transactions to do the final analysis. For e.g. if data is of 3 years i.e. from 2016 to 2018 (36 months) then for every station there are 36\*8 rows (eight rows corresponding to every month). Below is the figure which shows the final dataset which will be used for forecasting. ‘Number of transactions’ will be response variable and all other variable will be predictors.

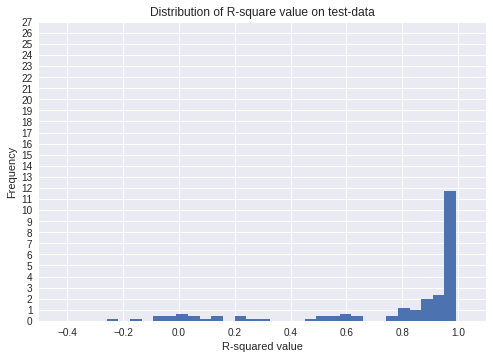
Figure 1:- Format of final data set

## Linear Regression

Starting with the simple linear regression model with only one predictor i.e. ‘distance’. Performance is evaluated based on r-square value on test-data.

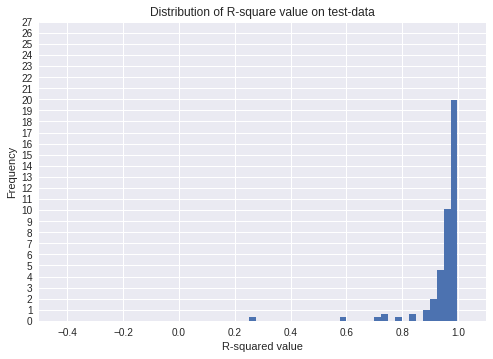
R-square value is not a good measure especially when comparing two models with different number of predictors. But here performance is evaluated looking at the r-square value of test data not on trained data.

Data corresponding to every station is divided into 80%(training) and 20%(testing). To have enough rows in test-data only those stations which have more than 6 months of data are considered (125 stations).

* **Model 1**: - y= Trip id, X=Distance (Distribution of R-square values corresponding to all stations is shown below in the figure)

As observed from the histogram that r-square values are spread all over the number line.

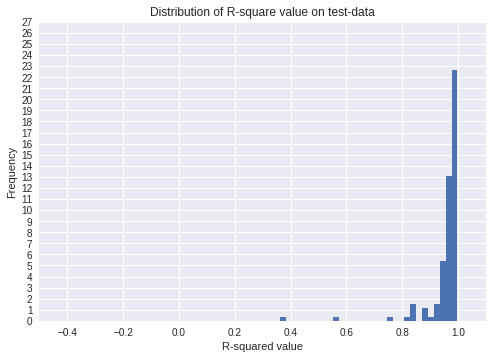
Further add another variable and then compare the performance.

* **Model 2: -** y= Trip id, X=Distance, Duration.

As observed from the histogram that only one station (station 3013) has r-square value less than 0.5.

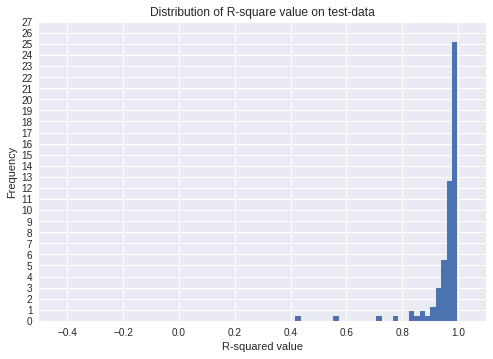
After observing station 3013 it is clear that there are outliers in the distance column due to which it has low r-square value.

Hence, Model 2 is much better than Model 1.



* **Model 3: -** y =Trip id, X = Distance, duration, pass-type

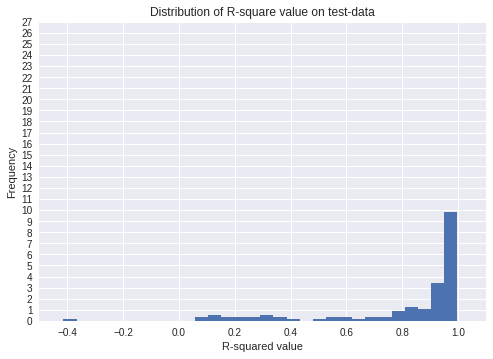
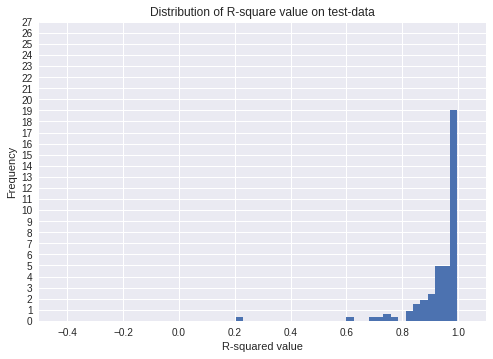
R-square value on test-data after adding pass-type improved a bit. But still station 3013 is have a low r-square value. But after comparing the histograms it is evident that model 3 is better than model 2.

* **Model 4: -** y= Trip id, X = = Distance, duration, pass-type, trip-type

After adding trip-type to the regression r-square value of station 3013 improved further and same can be seen for other stations also. As all the r-square values are on the test data so it proves that model 4 is performing better than model 3.

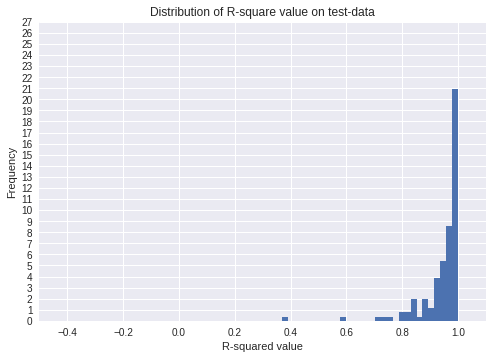
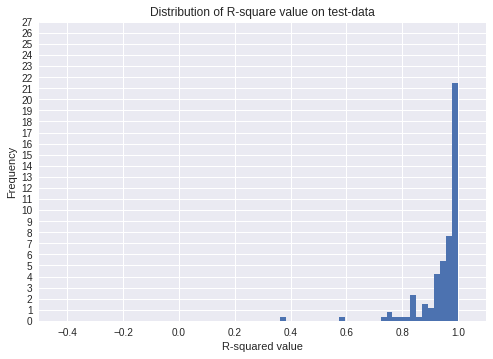
## Random Forest

Random-forest will take in account non-linearities which are there in the data much better than linear regression. Building the same 4 models for random forest which were built for linear regression and comparing the performance on test data using r-square value.



R-square values for Model 1 with random forest

R-square values for Model 2 with random forest



R-square values for Model 4 with random forest

R-square values for Model 3 with random forest

It can be observed from the histograms above that none of the model using random forest is performing better than Model 4 with Linear regression.

After doing the time-series analysis in the next part. Comparison will be done between Model 4 of linear regression and time-series by using the RMSE values of the test data. Depending on the that final model will be selected for forecasting.