1.) Build a Neural Network model for 50\_startups data to predict profit

Solution:

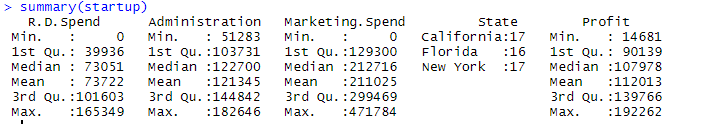
**Business Objective:** To prepare a model to predict profit of startups

**Datasets:**

Dependent variable: Profit (Continous)

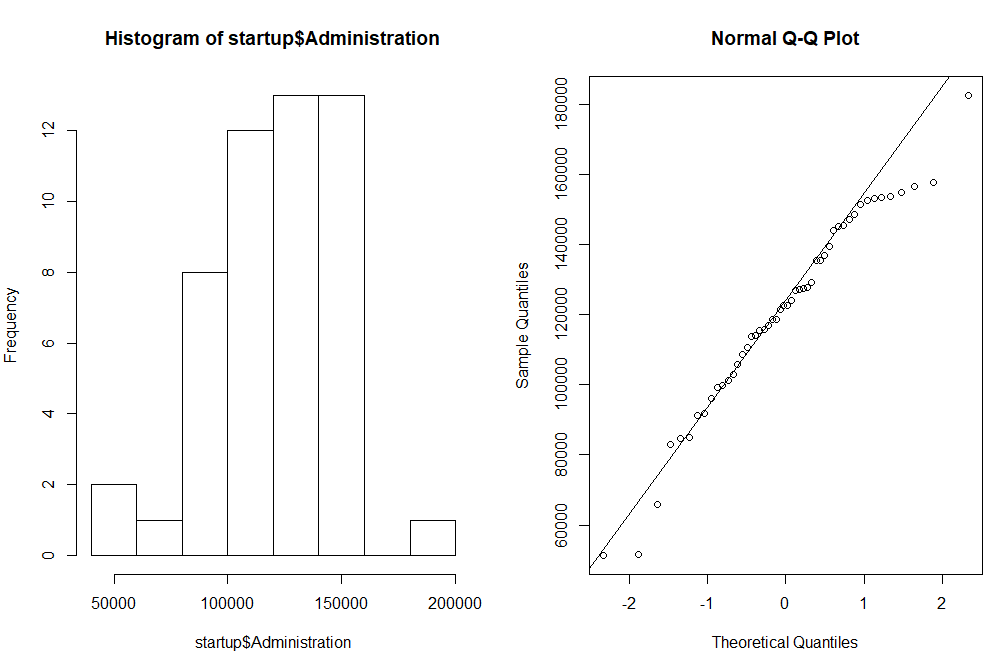
Independent variable: R.D.Spend (Continous), Administration (Continous), Marketing.Spend (Continous), State (Categorical)

**EDA:**



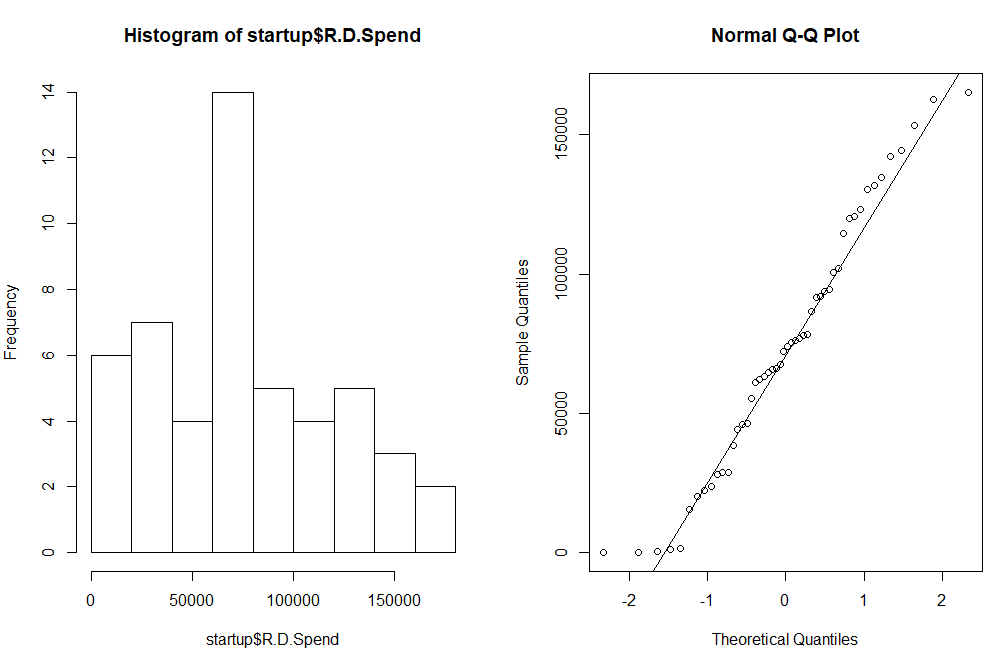
**Graphical representation**

**For Administartion**



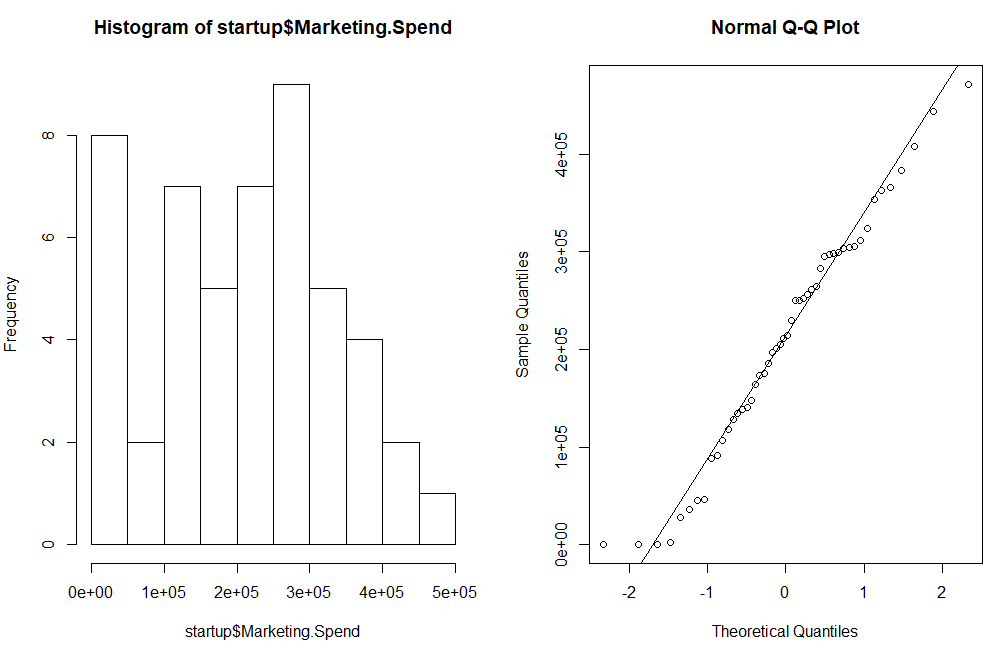
This is negatively skewed data and is not normally distributed. There is no outliers.

**For R.DSpent**



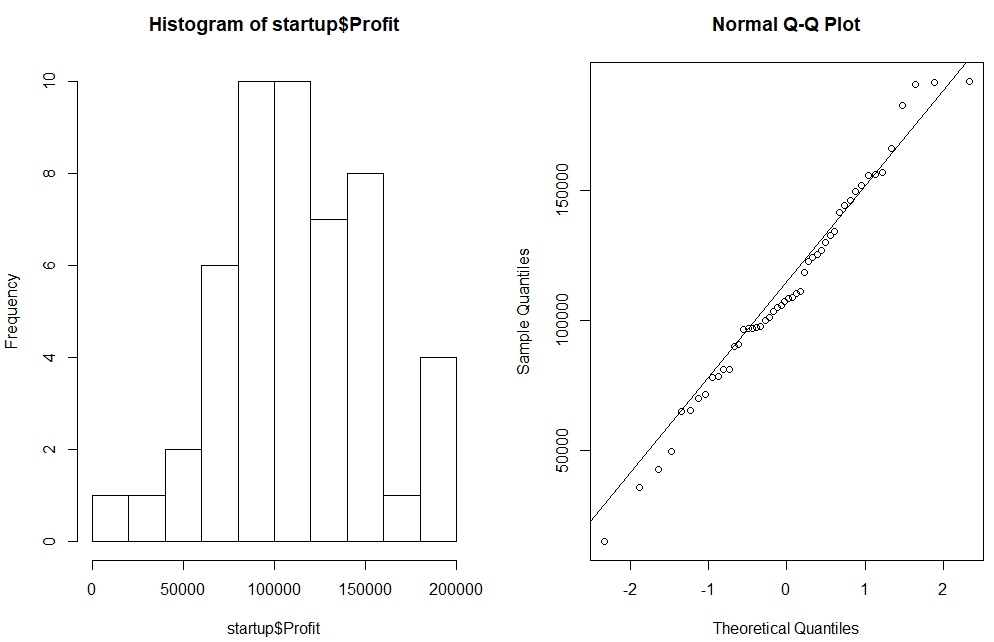
This is positively skewed data and is no normally distributed. There is not outliers.

**For Marketing.Spend**



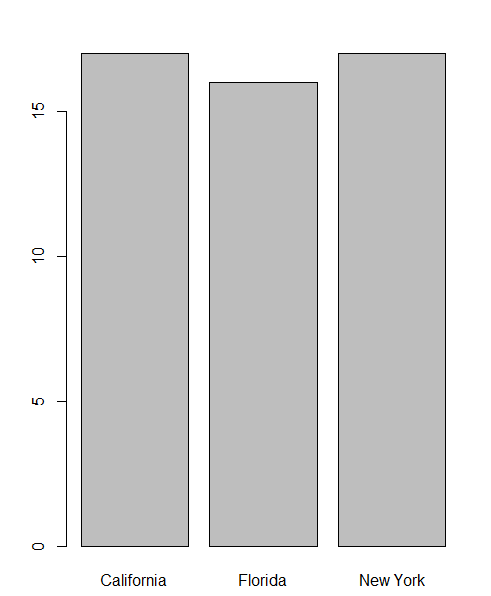
This is negatively skewed data and is not normally distributed. There is no outliers.

**For Profit**



This is positively skewed data and is not normally distributed. There is no outlier.

**For States**



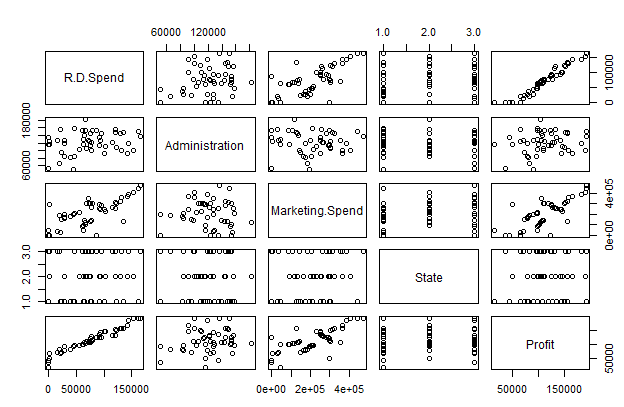
California= 17

Florida = 16

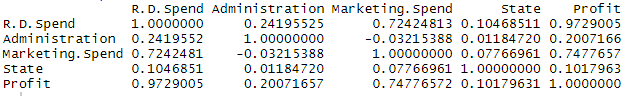
New York =17

Correlation between profit and other variables(Administration, R.D Spend,

Marketing Spend, States)



Correlation coefficient – strength and direction of correlation

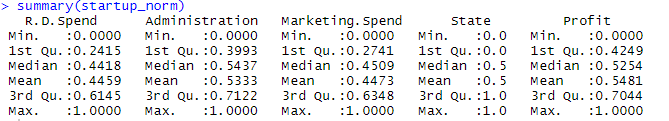


From the summary above, it can be said that the data of different variables is on different

scale so need to normalize the data

Normalization

**Summary of Normalized data**



From the above summary, it can be said that the data of all the variables have same

common scale from 0 to 1.

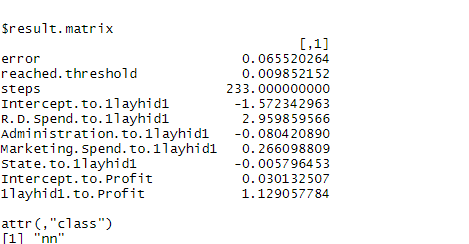
Data Partitioning

The normalized startup data is divided into two sets of train and test with 70: 30 ratio.

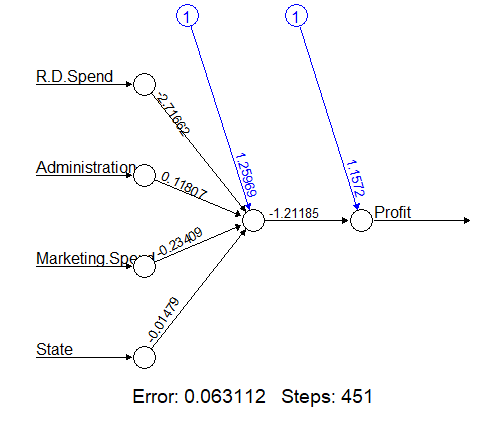
**Model Building:**

Building a neural network model on training data using single hidden neuron

Summary



**Visualization of matrix**



The above network topology shows the different weights involved for each variable to

give the predicted output. The error rate is 0.063112 which we need to further reduce.

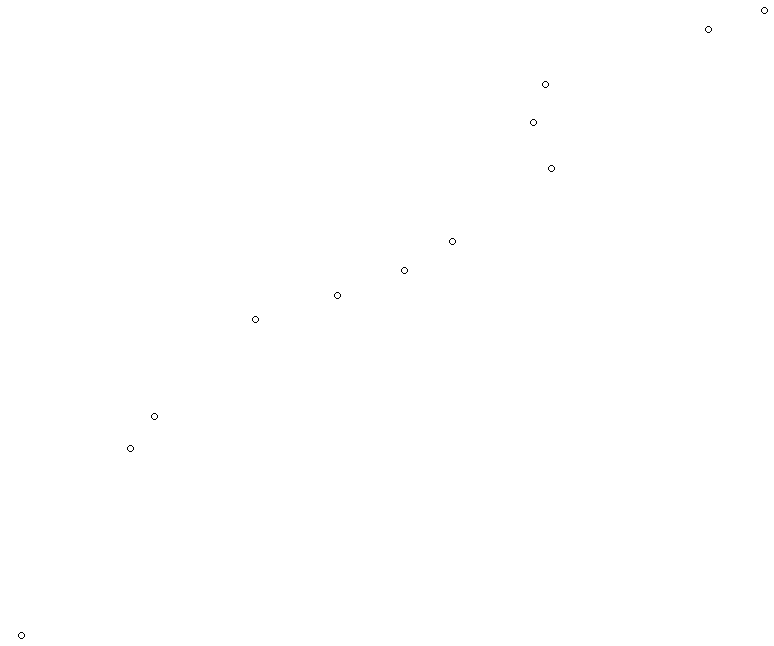
**Evaluating model performance:**

The model performance is evaluated using test datasets and predicted value for profit is

computed using the model built above.

Correlation coefficient between predicted and actual values

r= 0.9748265. Since the value is very high, it seems the model is good

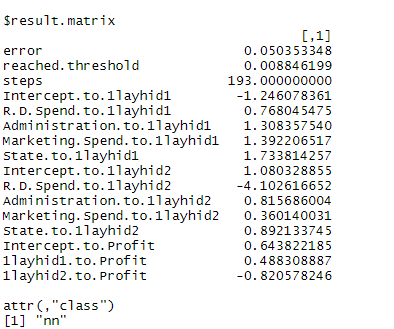


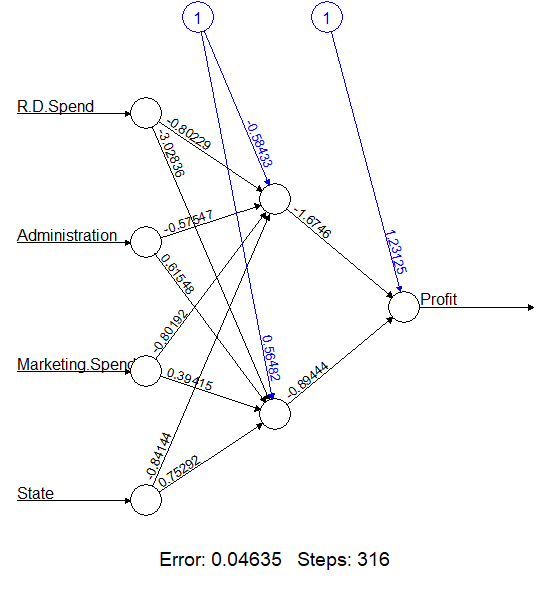
The above plot between predicted and actual values shows linear trend between them.

Improving model performance

**1.) A model is prepared using 2 hidden layers**

**Model summary**

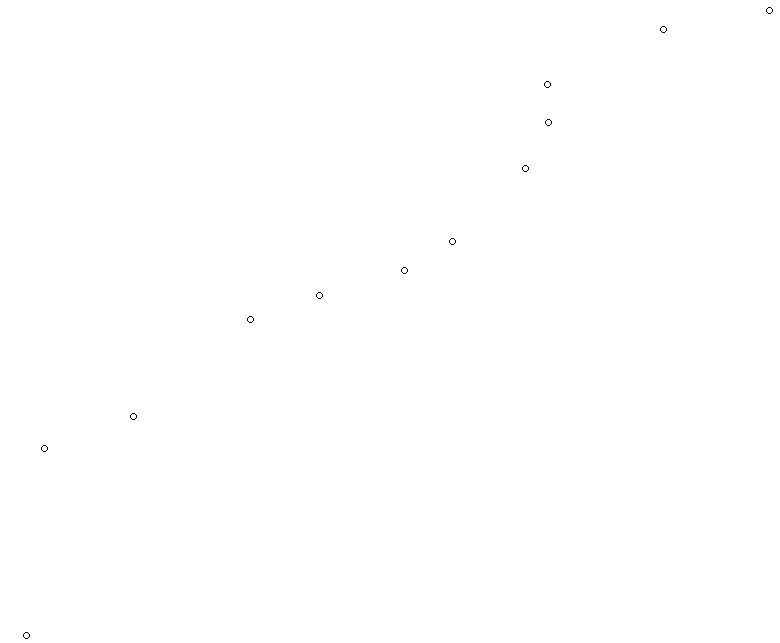




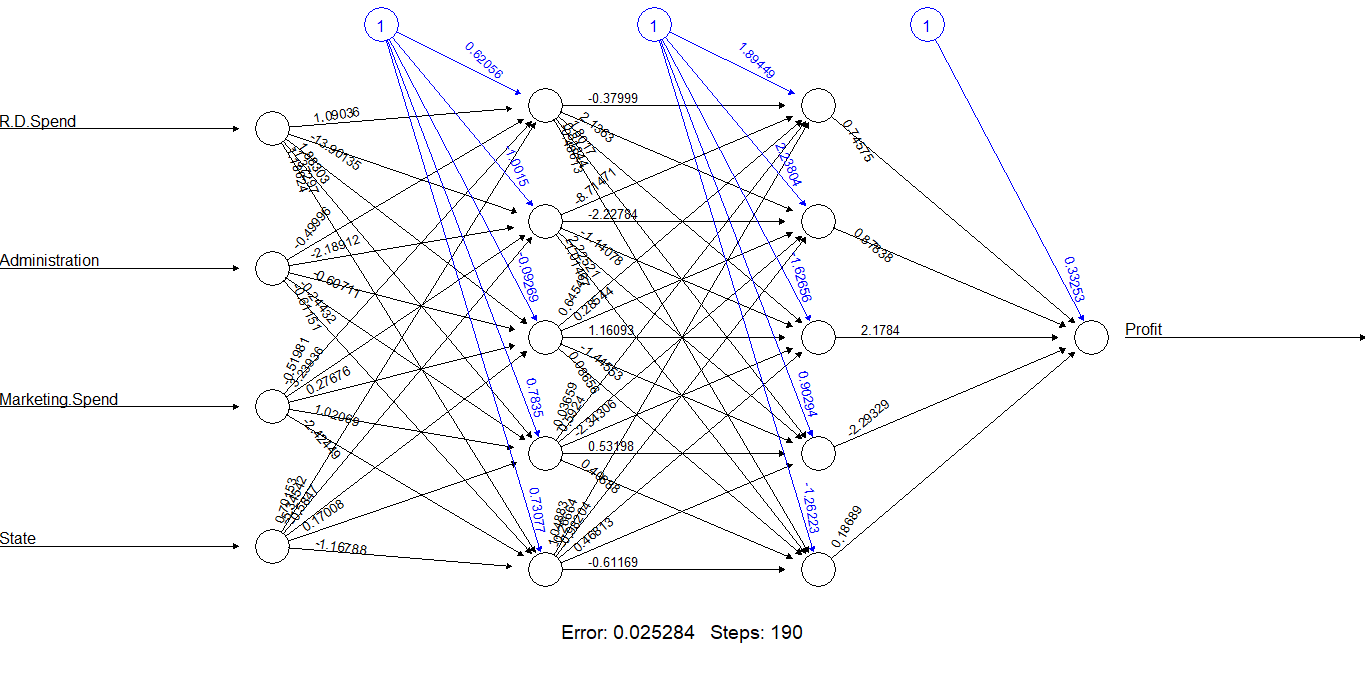
The error rate is = 0.04635 which is lower then the previous model.

r= 0.9654

Plot between actual and predicted values.



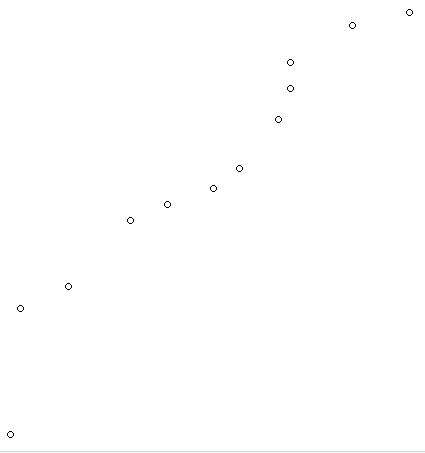
**2.) A model is prepared using 5 and 5 hidden layers**



**Error rate is 0.02528,**

**r= 0.966**

Plot between actual and predicted output



SSE (error) is reducing, as no. of hidden layers are increased.

**2.) PREDICT THE BURNED AREA OF FOREST FIRES WITH NEURAL NETWORKS**

**Solution:**

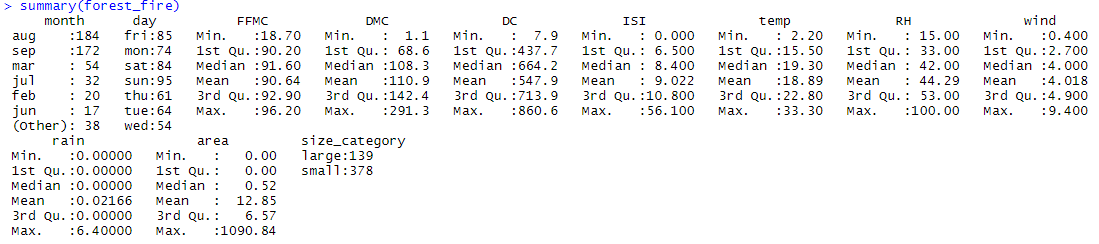
**Business Problem:** To prepare a model to predict the burned area

**Datasets:**

Independent Variable: month, day, FFMC, DMC, DC, ISI, temp, RH, wind, rain, area,

Dependent Variable : size\_category

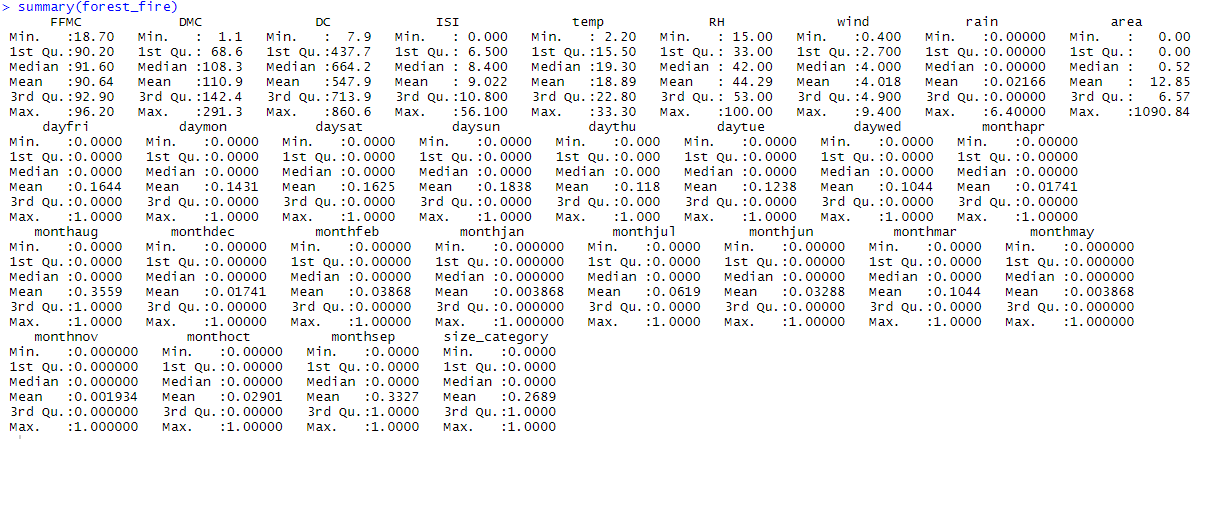
**EDA:**



The output variable (size\_category) is converted to numeric from factor

Large is encoded as 1 and small as 0.

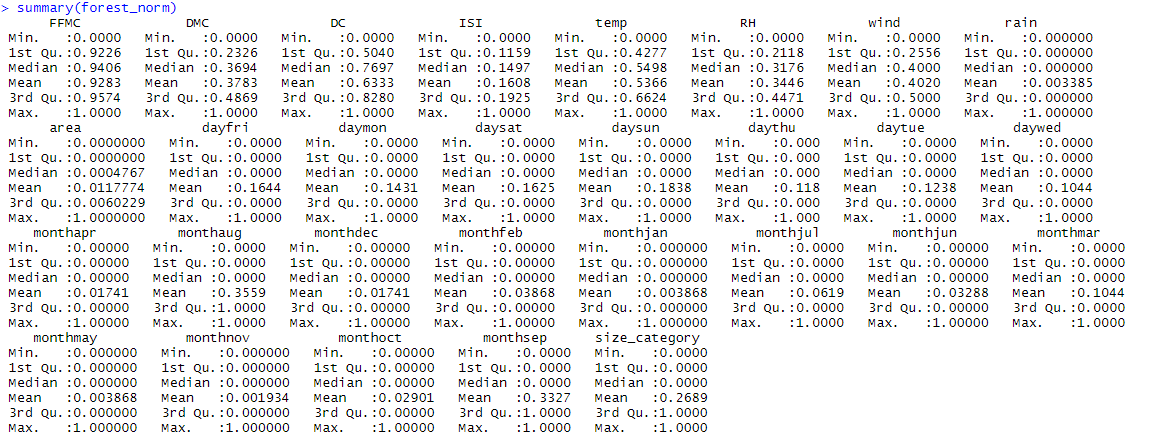
**Summary of datasets including dummy variable**



**Normalization**

Since data is not having similar scale and unit we have to normalize the data.

The following is the summary of all normalized numerical variables



From the above summary, it can be said that the data of all the variables have same

common scale and units and data value ranges from 0 to 1.

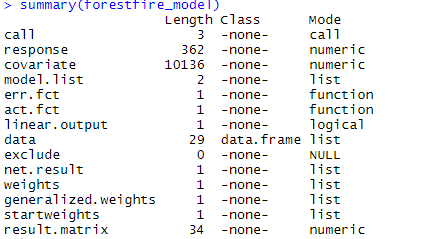
**Data Partitioning:**

The normalized forest\_fire data is divided into two sets of train and test by 7: 3 ratio.

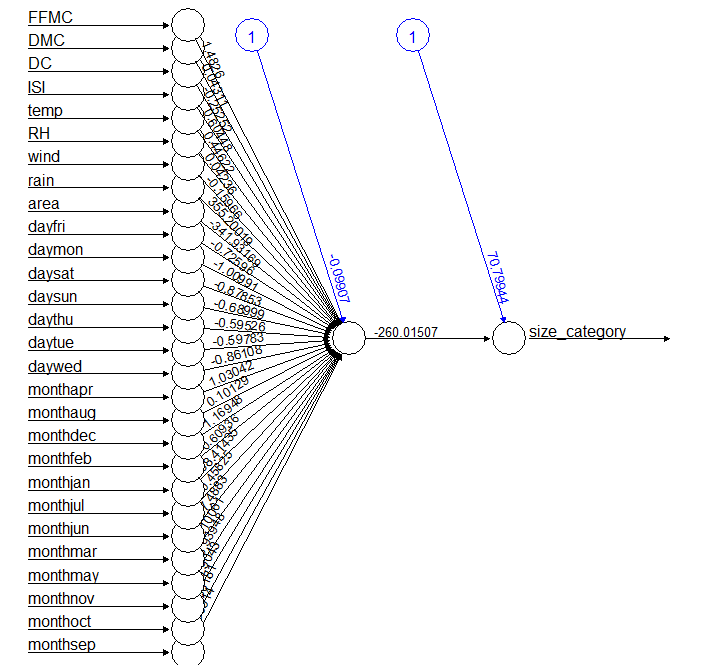
**Model Building:**

Building a neural network model on training data using single hidden neuron

Summary



**Visualization of matrix**



The above network topology shows the different weights involved for each variable to

give the predicted output. The error rate is 9.402338e-03 which we need to further reduce

**Evaluating model performance:**

The model performance is evaluated using test datasets and predicted value for

size\_category is computed using the model built above.

Correlation coefficient between predicted and actual values.

r= 0.867. Since the value is very high, it seems the model is good.

**Confusion Matrix and Misclassification**

We found the value which is having probability >0.5 as “1” and the value having probability <0.5 as 0



Here 111 and 39 are truly classified as Large and Small size respectively

4 data points are misclassified.

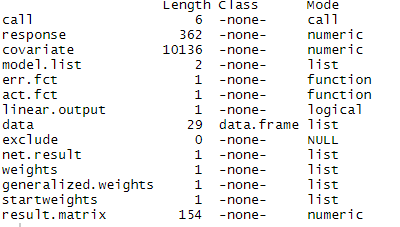
Accuracy = 96.77%

Error = 3.22%

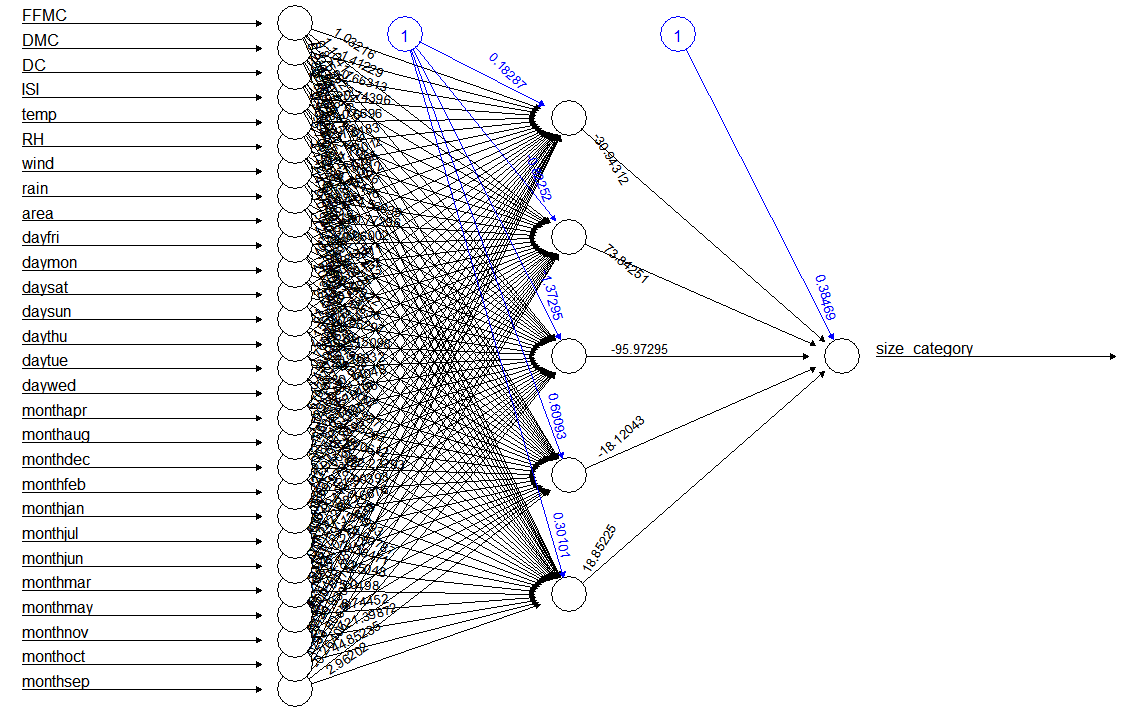
**Improving model performance**

**1.) A model is prepared using 5 hidden layers**

**Model summary**



**Visualiztion**



**Confusion Matrix and Misclassification**

We found the value which is having probability >0.5 as “1” and the value having probability <0.5 as 0



Here 113 and 38 are truly classified as Large and Small size respectively

4 data points are misclassified.

Accuracy = 97.41

Error = 2.5 %

The error reduces as we increases the no. of hidden layers.

We will chose the 2nd model as it is having less error.

**3.) Prepare a model for strength of concrete data using Neural Networks**

**Solution:**

**Business Problem:** To prepare a model to predict strength.

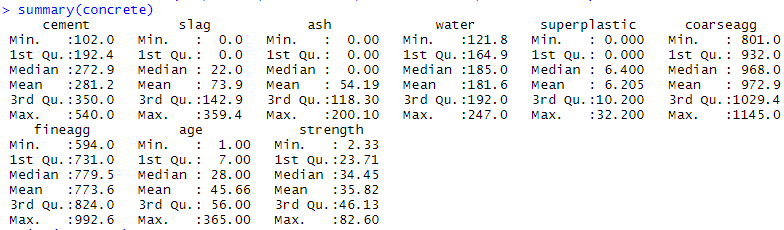
**Datasets:**

Dependent Variable: strength

Independent Variable: cement, slag, ash, water, superplastic,

coarseagg, fineagg, age

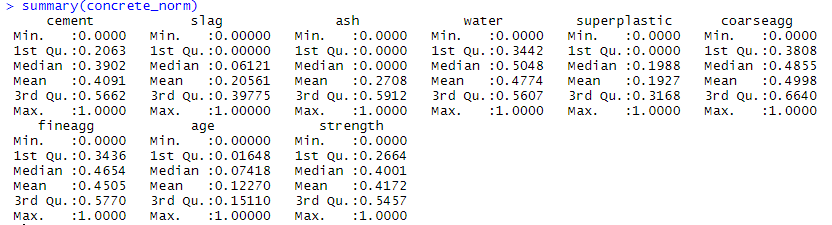
**EDA:**



**Normalization:**

To make the data scale free and unit less, we have to bring the data in common range i.e is 0 to 1.

Summary of Normalized data



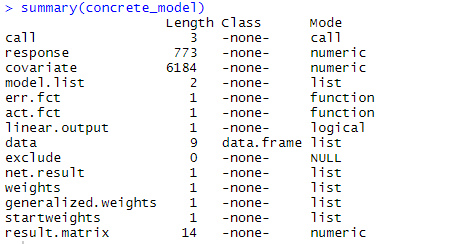
**Data Partitioning**

The normalized concrete data is divided into two sets of train and test with 7: 3 ratio.

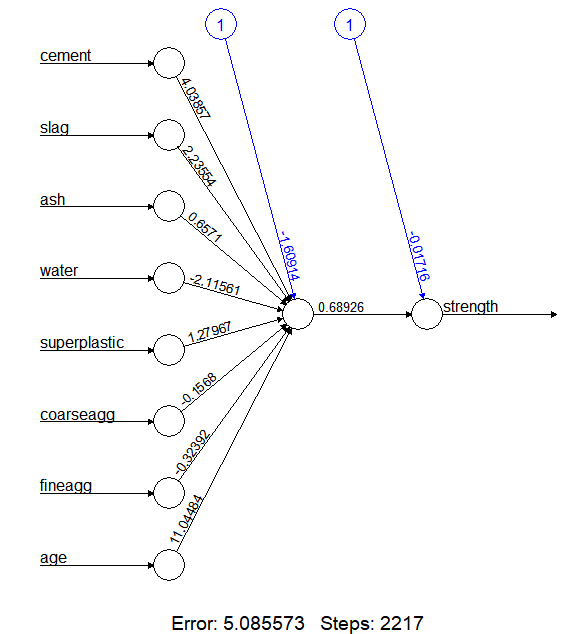
**Model Building:**

Building a neural network model on training data using single hidden neuron

Summary



**Visualization of matrix**



The above network topology shows the different weights involved for each variable to

give the predicted output. The error rate is 5.085573 which we need to further reduce.

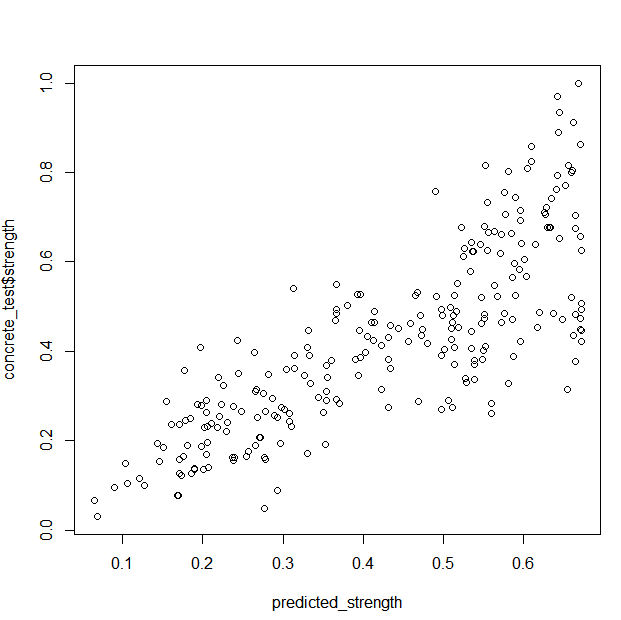
**Evaluating model performance:**

The model performance is evaluated using test datasets and predicted value for strength is

computed using the model built above.

Correlation coefficient between predicted and actual values

r= 0.8056. Since the value is very high, it seems the model is average

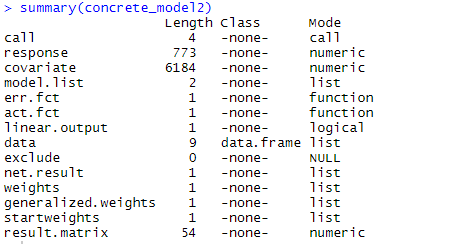


The above plot between predicted and actual values shows linear trend between them.

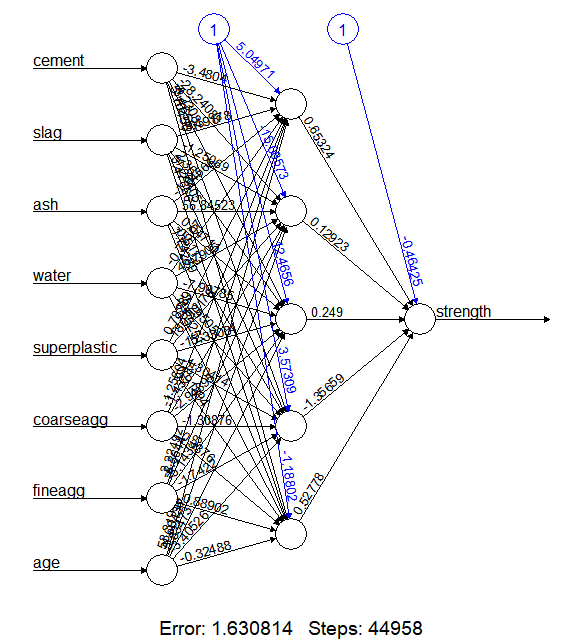
**Improving model performance**

**1.) A model is prepared using 5 hidden layers**

**Model summary**



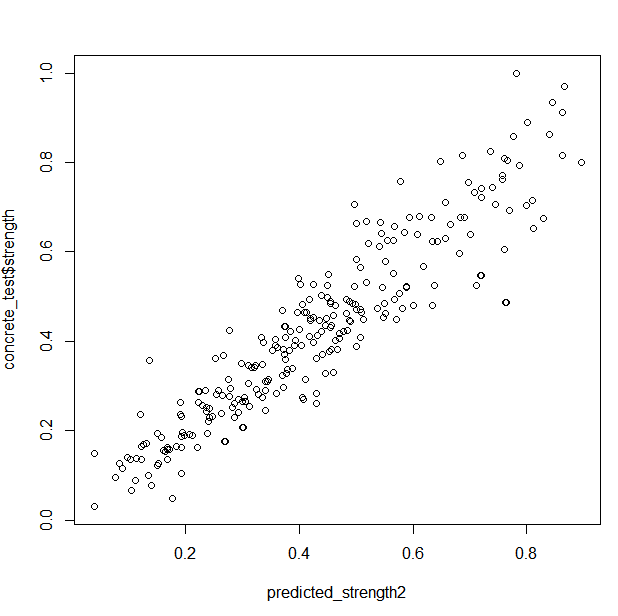
**Visualization**



The error rate is = 1.6308 which is lower then the previous model.

r= 0.9267

Plot between actual and predicted values



SSE (error) is reducing, as no. of hidden layers are increased. Since error is less in 2nd model we will choose the second model for prediction.