# NEURAL NETWORK

## QUESTION NO 1:

Build a Neural Network model for 50_startups data to predict profit

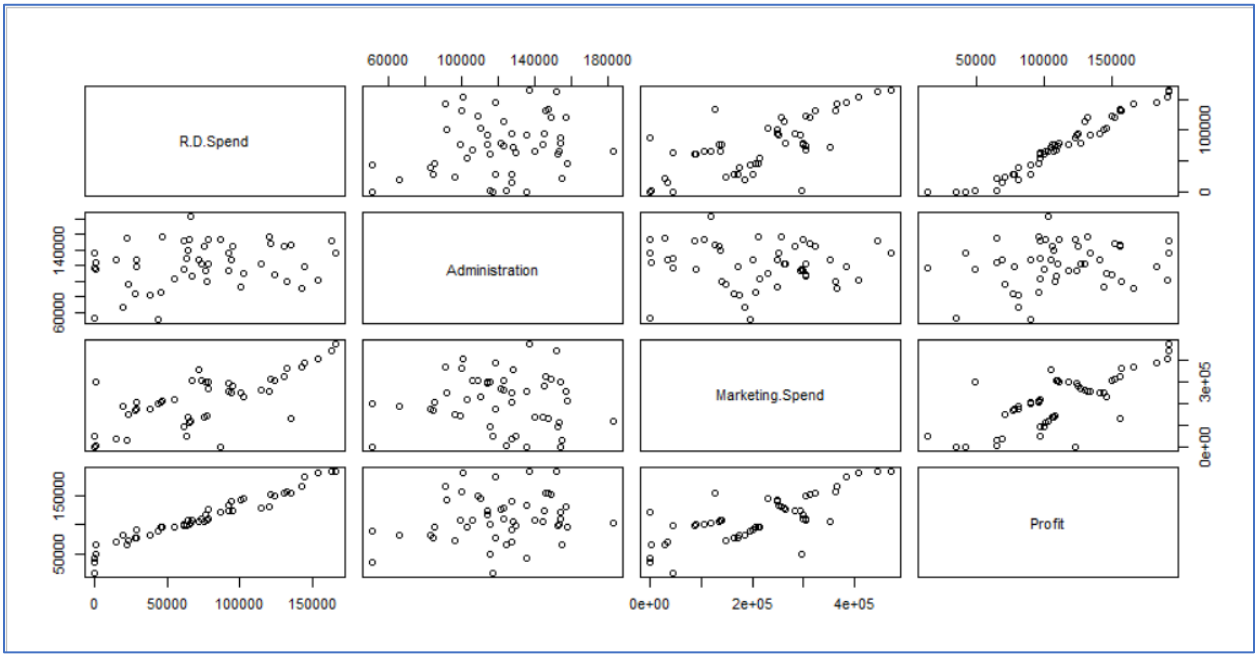
Let's Have a look on structure of the data

```
'data.frame':    50 obs. of  5 variables:
 $ R.D.Spend      : num  165349 162598 153442 144372 142107 ...
 $ Administration : num  136898 151378 101146 118672 91392 ...
 $ Marketing.Spend: num  471784 443899 407935 383200 366168 ...
 $ State          : Factor w/ 3 levels "California","Florida",..: 3 1 2 3 2 3 1 2 3 1
 $ Profit         : num  192262 191792 191050 182902 166188 ...
```

Our data contains one variable factor i.e. State and other are numeric.

Target variable is profit in this scenario. Which is numeric and continuous in nature

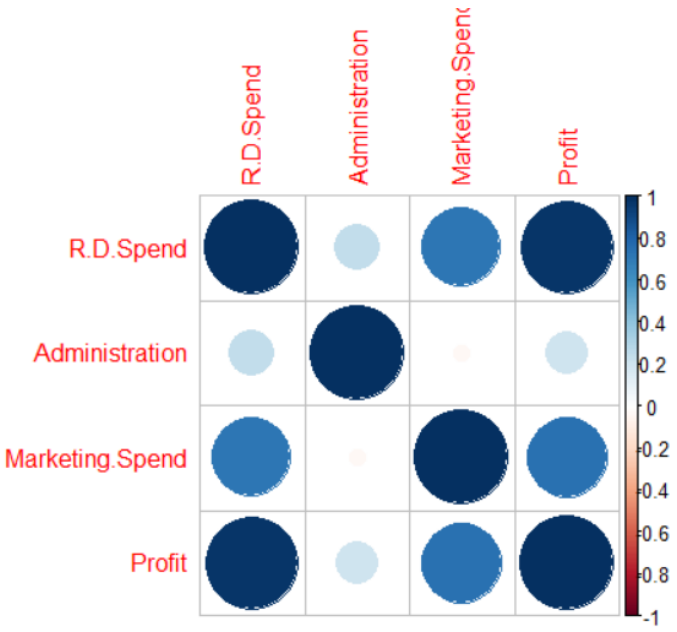
## PAIRS PLOT :



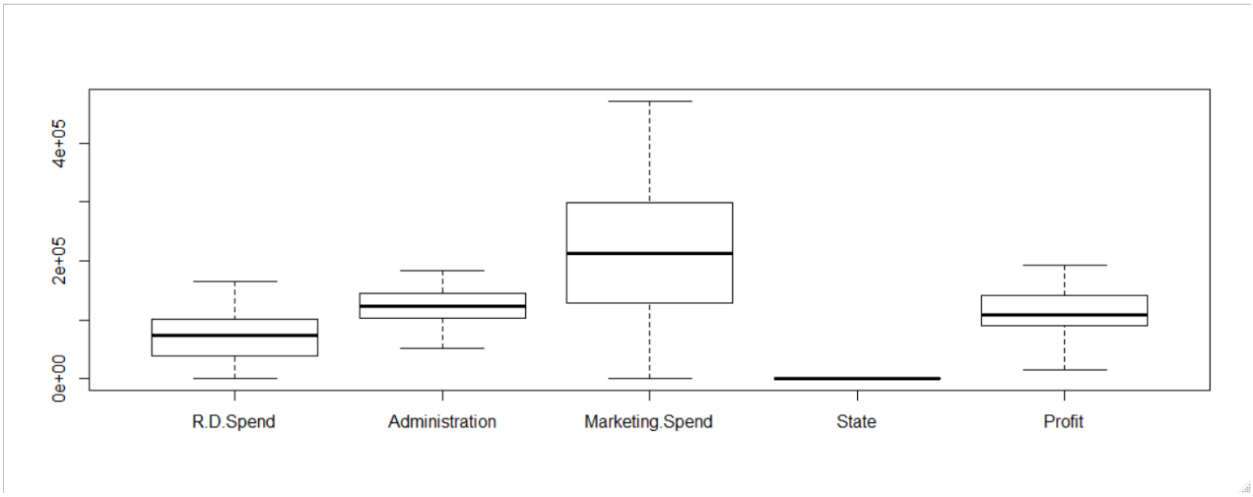

from the plot we can see that there R.D.Spend and Marketing.Spend are positively correlated with profit, and rest are showing week correlation among themselves with each other.

## CORRELATION PLOT:

Ignoring the diagonal elements.



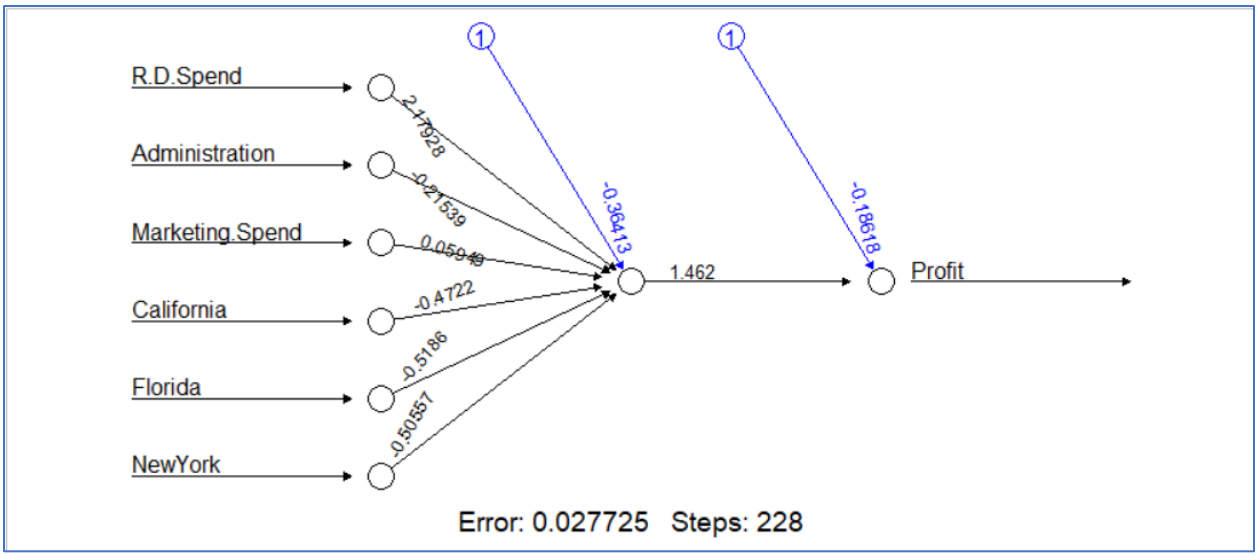

## BOXPLOT




From the boxplot we can see that there is no outlier inside the data set.

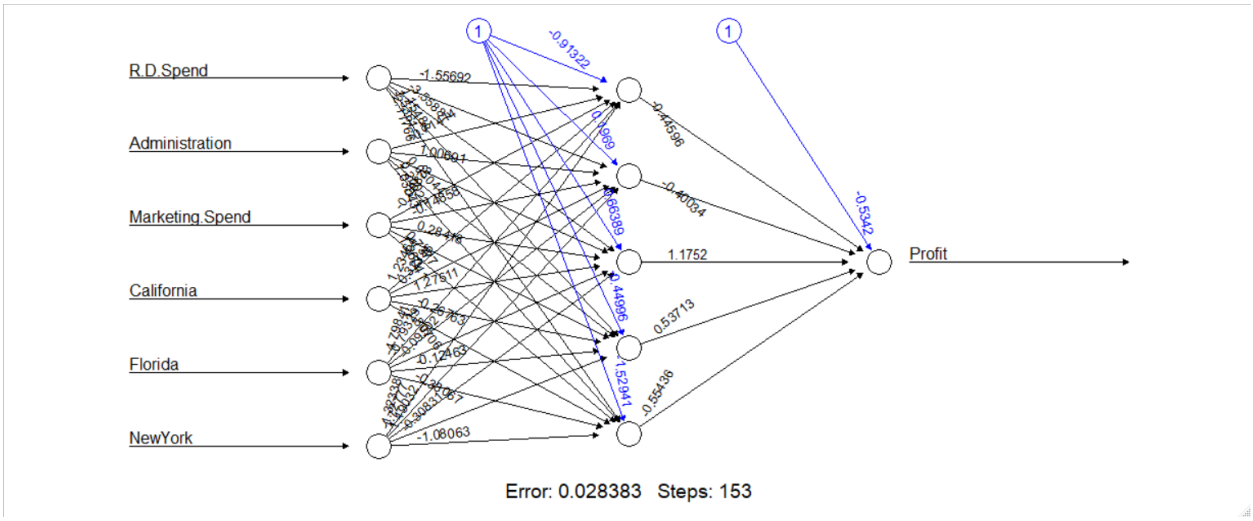
## MODEL NO 1:

In my model 1 I have no hidden layer in my model. And My accuracy is about 0.9563 as all accuracy is due to my Variable R.D.Client, which is highly correlated with the profit.




## MODEL NO 2:

In my second model I have considered 5 hidden layers.




Here I am getting my efficiency 0.96 i.e. slightly increased.

## CONCLUSION:

I can go for any of the model mention above during my prediction, for Time consumption purpose I may consider my 2nd model with hidden layer. And if I need to decrease the complexity in my model, I may consider my 1st model without hidden layers. Model 1 is with RMSE 14478 and model 2 is with RMSE 13479.

In both the model error value is too low i.e. 0.02 and steps is between them is slightly different, in case of hidden layer Our job is done within 153 steps, but incase of no hidden layer we may face some extra steps, which my decrease our computation speed. Also Looking at RMSE we may consider our model 2 with hidden layers.

## QUESTION NO 2:

PREDICT THE BURNED AREA OF FOREST FIRES WITH NEURAL NETWORKS

### STRUCTURE OF THE FOREST DATA

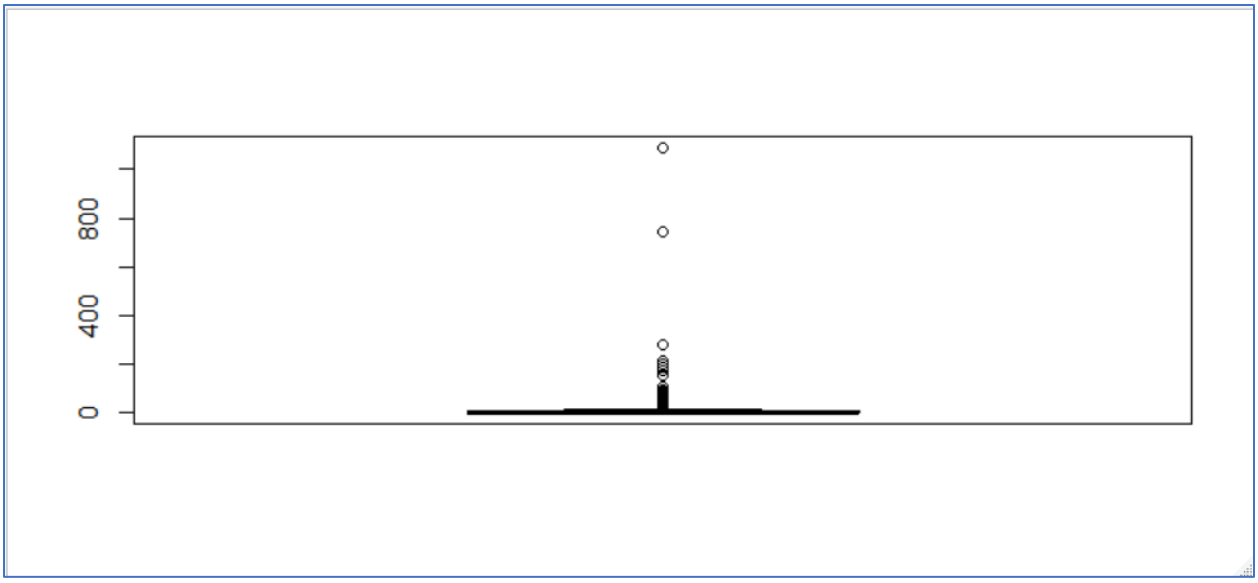
```
$data.frame':    517 obs. of  31 variables:
 $ month        : Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12 12
 $ day          : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...
 $ FFMC         : num  86.2 90.6 90.6 91.7 89.3 92.3 92.3 91.5 91 92.5 ...
 $ DMC          : num  26.2 35.4 43.7 33.3 51.3 ...
 $ DC           : num  94.3 669.1 686.9 77.5 102.2 ...
 $ ISI          : num  5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...
 $ temp         : num  8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...
 $ RH           : int  51 33 33 97 99 29 27 86 63 40 ...
 $ wind         : num  6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...
 $ rain         : num  0 0 0 0.2 0 0 0 0 0 0 ...
 $ area         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ dayfri       : int  1 0 0 1 0 0 0 0 0 0 ...
 $ daymon       : int  0 0 0 0 0 0 1 1 0 0 ...
 $ daysat       : int  0 0 1 0 0 0 0 0 0 1 ...
 $ daysun       : int  0 0 0 0 1 1 0 0 0 0 ...
 $ daythu       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ daytue       : int  0 1 0 0 0 0 0 0 1 0 ...
 $ daywed       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthapr     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthaug     : int  0 0 0 0 0 1 1 1 0 0 ...
 $ monthdec     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthfeb     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthjan     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthjul     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthjun     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthmar     : int  1 0 0 1 1 0 0 0 0 0 ...
 $ monthmay     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthnov     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ monthoct     : int  0 1 1 0 0 0 0 0 0 0 ...
 $ monthsep     : int  0 0 0 0 0 0 0 0 1 1 ...
 $ size_category: Factor w/ 2 levels "large","small": 2 2 2 2 2 2 2 2 2 2
```

```
My data has 3 factor variables month, day, size_category. All other variables are num
eric
```

Our target variable is area i.e. numeric in nature.

Here in our data set there is no missing values.

BOXPLOT:

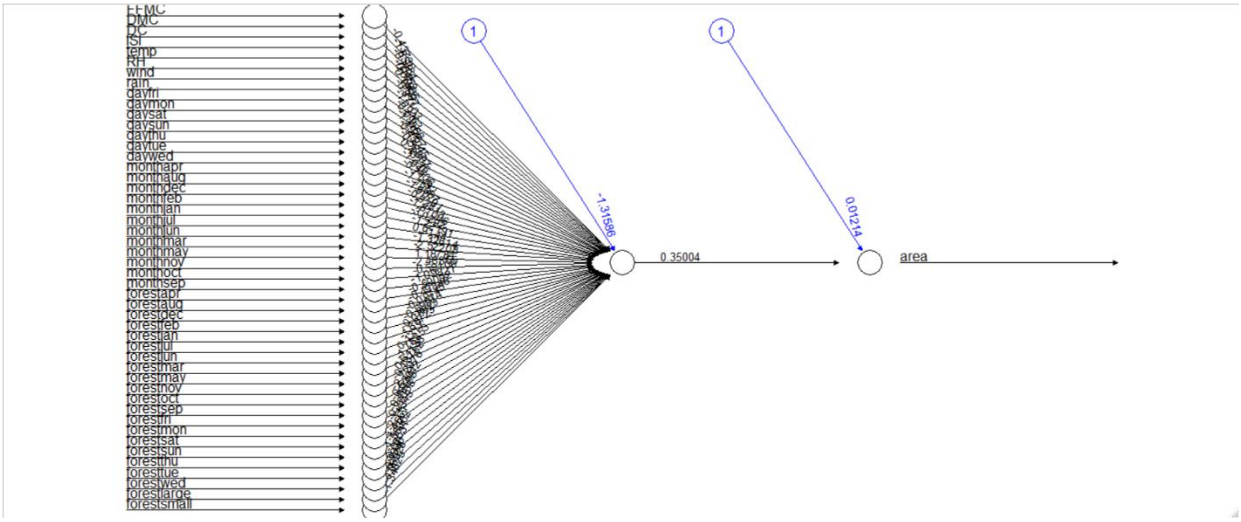



The Above boxplot is for the variable area in our forest data, which is the target variable in our business scenario.

So I am just ignoring my outliers.

Model no 1:

In my first model I have not considered any hidden layer.

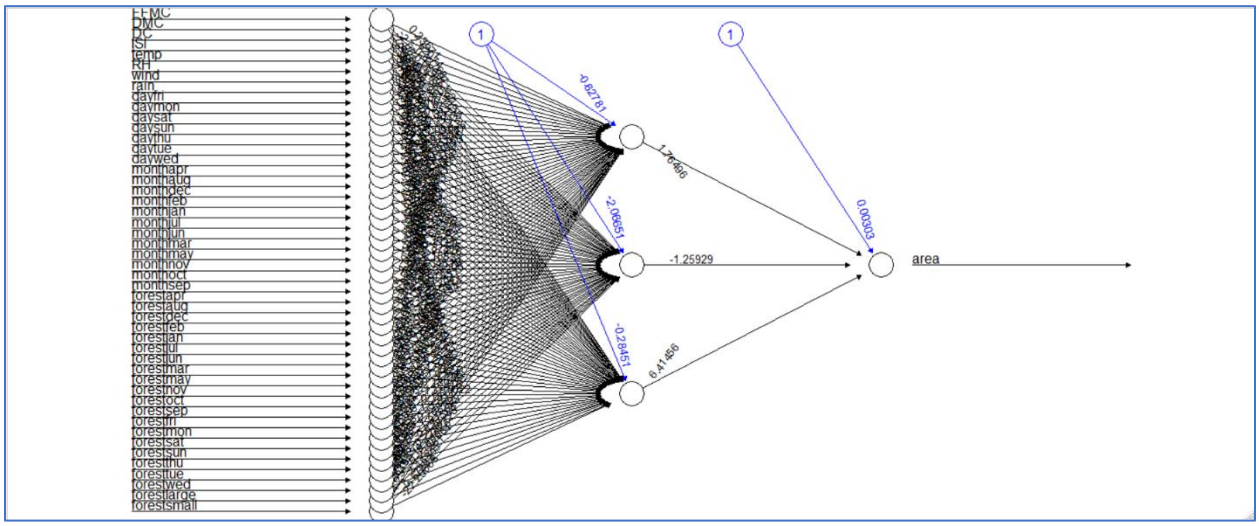



Here in my 1st model I am getting a very poor accuracy in my model as 0.24 i.e. only 24%

So, in my model 2 I may consider my hidden layer.

## MODEL NO 2:

I have increased my no of hidden layers to 3.




Here I am getting lesser accuracy compared to the model 1 i.e. 0.12 i.e. only 12%

I can say that the variable area is not correlated with any of the columns in my data set if than it must be very week correlation among them.

## MODEL 3:

In my model 3 I removed all my records where area is 0, and build my model.

Here I come up with 0.06 Accuracy,

In some cases, after denormalization I am getting some negative values in my Predicted variables. So, I may not relay on the data set columns (independent variables, predictors) i.e. the columns may not able to express the variation in target variable.

## CONCLUSION

I am not capable of getting relevant result from this particular data set, as the target variable has very week or no relation with other predictor variable in my data set, so It will be a bad model if we rely on the available columns of our data set. Its better to go for joining more informative columns in my data if possible, so that we will come up with a good predictive model.

If any technique I am lagging with than kindly inform me.

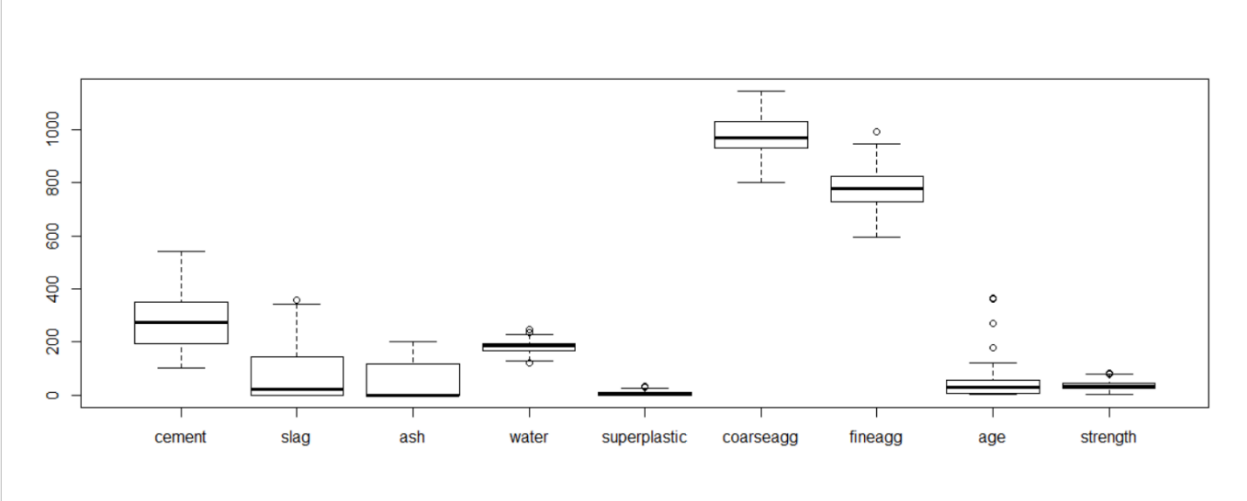
## QUESTION NO 3:

Prepare a model for strength of concrete data using Neural Networks

```
'data.frame':    1030 obs. of  9 variables:
 $ cement      : num  141 169 250 266 155 ...
 $ slag        : num  212 42.2 0 114 183.4 ...
 $ ash         : num  0 124.3 95.7 0 0 ...
 $ water       : num  204 158 187 228 193 ...
 $ superplastic: num  0 10.8 5.5 0 9.1 0 0 6.4 0 9 ...
 $ coarseagg   : num  972 1081 957 932 1047 ...
 $ fineagg     : num  748 796 861 670 697 ...
 $ age         : int  28 14 28 28 28 90 7 56 28 28 ...
 $ strength    : num  29.9 23.5 29.2 45.9 18.3 ...
```

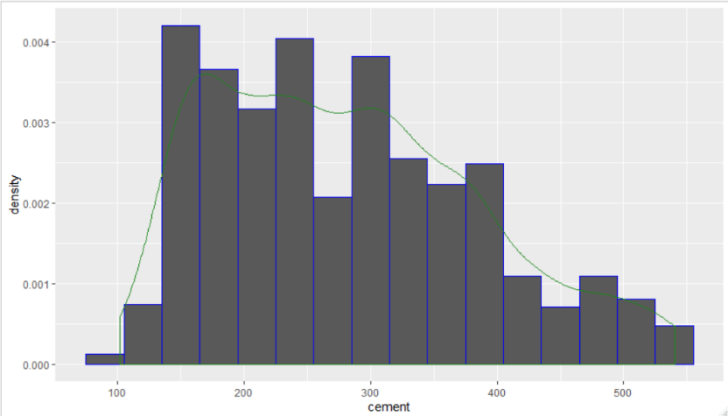
Our data contains all numeric variables. Our Target variable is strength

## BOXPLOT:




Here from the boxplot we can see that age, water, fineadd, and strength contains outliers. Here we can see very less numbers of outliers, so I may remove the outliers.

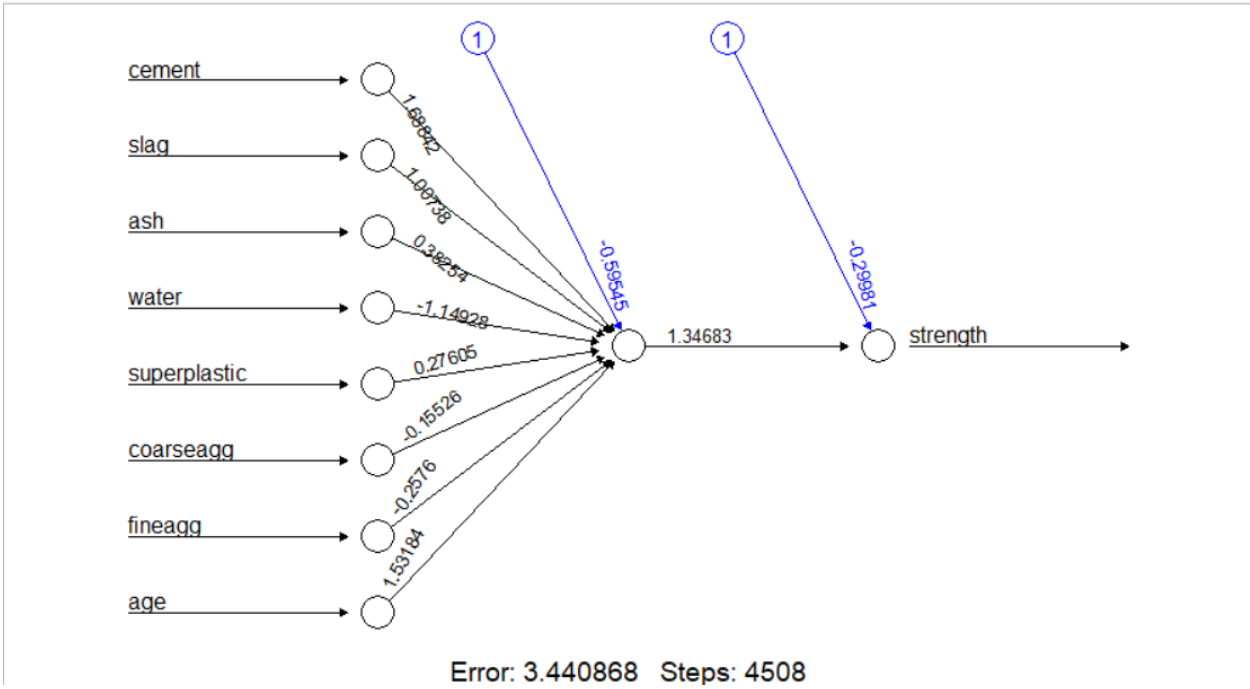
As after removing my outliers again I am getting another set of outliers. So I am stick to all my records.




Looking at the histogram of my cement variable we can say that its right skewed.

## MODEL NO 1

Here I have considered all my normalized data for model fitting i.e. Strength as Dependent and all rest data as Independent. Without considering any hidden layer, I have fitted my model.

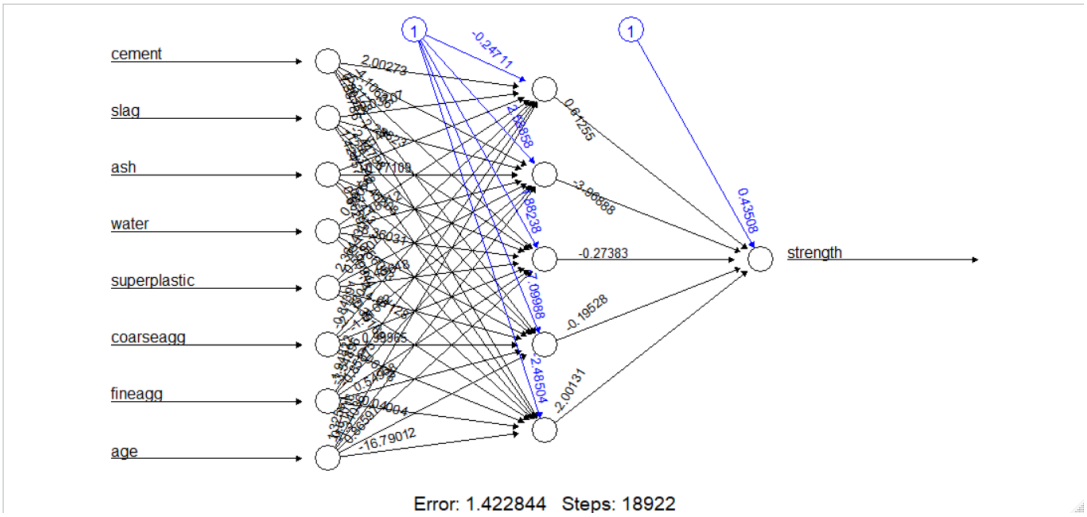



In this model I come up with correlation between the Actual and predicted model is 0.86051 and RMSE value is 8.731773

To Improve my accuracy in the model, I may introduce hidden layer to my model.

## MODEL NO 2

In my second model I have considered hidden layer as 5



Here In my second model my accuracy is 0.9361695 and RMSE is 5.649183 which is decreased from the previous model.

## CONCLUSION

Now I can safely choose my second model as my final model for prediction purpose as It has the least accuracy among the two model.