Mini Project - Indian Credit Risk

Name : Numer P

Table of Contents

1. Project Objective

The main objective of the report is to explore the Indian Credit Risk ("raw-data.xlsx") with the validation dataset ("validation_data.xlsx") in R and generate insights about the data set. This exploration report will consist of the following,

- Importing dataset in R
- Understanding the structure of Dataset
- Graphical exploration
- Descriptive Statistics
- Default Risk Modelling using Logistics Regression

2. Assumptions

The Indian Credit Risk is performed under the Logistics Regression model to interpret the default credit risk companies. The Credit Risk System is validated with Defaulters and Non-Defaulters list among the companies and ranked the companies with highest returns and risk measured with the default status. Indian Credit Risk model are performed with two different datasets.

The Dataset is explored with numeric variables associated with Net Worth Next Year for the various variables is bucketed like profitability ratio, leverage ratio, liquidity ratio and company size of the Indian credit companies. The credit risk is carried by the negative growth and positive growth for the next year is diversified by non-defaulters for the positive growth in next year and defaulters with negative growth in the next year. Indian Credit Risk model is performed other ratios include Debt to Equity Ratio, Return on Assets, Return on Equity Ratio, Current Ratio, Asset Turnover Ratio from the variables in the datasets.

3. Exploratory Data Analysis – Step by Step Approach

1. Environment Set up and Data Import
2. Variable Identification
3. Univariate Analysis
4. Bivariate Analysis
5. Variable Transformation
6. Feature Exploration

3.1 Environment Setup and Data Import

3.1.1 Install necessary packages and Import Libraries

This section is used to install packages and invoke the associated libraries. Having all packages at the same places increase code readability.

3.1.2 Setup Working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

Please refer Appendix A for source code.

3.1.3 Import and read the dataset

The given dataset is in .xlsx format. Hence the command 'read_xlsx' is used for import the dataset.

Please refer Appendix A for source code.

3.2 Variable Identification

- setwd() used for setup working directory to export data and files from the folder or location in PC.
- getwd() used to identify the location was correctly entered or not.
- Library function is used to load the installed packages like ggplot2, readxl, car, DataExplorer, ggcorrplot, devtools, lares, Data.table, ROCR, pROC, caret, ineq, pscl.
- read_xlsx function is used to load the excel files in the path.
- attach function is used to reduce the reusability of variable name to enter each time.
- str function is used to check the category variables formats.
- summary function is used to get the summarised value like length, class and basic statistics values with quartile ranges.
- dim function is used to find the total observations and variables.
- Predict function is used to predict the train dataset and the validation dataset.
- glm is used for the predictive modelling for the logistic regression.
- Loglik is the function used to predict the likelihood of the predicted dataset.
- exp is used to examine the exponential values of the logistics regression models.
- McFadden is used to define how the good model is form the logistics regression.
- Decile are used to create the rank order of credit risk of companies in descending order.
- ROC is used to build the ROC curve of the model developed predicted values.
- Corr_cross is used to find out the top correlation variables in the model.

3.2.1 Variable Identification – Inferences

#getwd()

It shows the working directory dataset

#library(readxl)

(readxl) helps in reading the files in excel formats.

#library(ggplot2)

(ggplot2) helps is visualise the datasets in boxplot, histogram and graphical representation.

#str

str shows the variables along with class of the data. It shows some samples to understand the data.

#class

class function describes the full file in data.frame format. As the files includes category in season variables, it shows the values as character format.

#attach

Variable is attached to reduce the reusability in following syntax.

#dim

It shows the number observations and the variables associated in the file.

#summary

It produces the results as summarised format for each variable.

#names

It gives the column names from the dataset.

#plot_boxplot

Boxplot graph is used for the dataset to find the outliers and the values are grouped.

#plot_histogram

This function is used to plot the histogram for the variables.

#plot_bar

This function is used to build bar charts for the category variables.

#plot_scatterplot

Scatterplot of all grouped variables are built in scattered point of the charts.

#plot_density

Density plot is defined with the density of the values for the variables.

#plot_qq

Quantile charts are used to build the percentile of the values in the variables.

#glm

It is used for the generalized linear models and specifies the linear prediction for the variables and its error.

#vif

The variance inflation factor function is used for the calculation of variance in the linear models.

#loglik

The function is used for the likelihood prediction of the variables.

#predict

The predict function is used for the fitting the model function used for the datasets.

#table

The table is used for the cross classifying the variables in the contingency tables.

#cor

The correlation function is used to understand the correlation between the variables.

#confusionmatrix

Calculated for the cross validation in the observed and predicted values from the models.

#corr_cross

Correlation of the top variables are defined by the highest correlation values.

#deciles

Deciles are used to create the ranked variables of the predicted model values.

#rank

Rank is used in order to measure the variables associated for the KS values.

#quantile

Quantile function is used to treat the variables from 1% and 99% in the variables.

#predict.glm

The prediction of the fitted logistics regression model is built among the significant variables.

3.3 Univariate Analysis

Univariate analysis is the analysis of data of one variable at time and it involves whether the datasets are descriptive or inferential statistics.

1. Exploratory Analysis

```
> library(readxl)
> my_train=read_excel("raw-data.xlsx",sheet = 1)
> my_test=read_excel("validation_data.xlsx",sheet = 1)
```

The datasets are imported to the R and read the raw data file as my_train and validation dataset as my_test.

```
> dim(my_train)
```

[1] 3541 52

The dimensionality of the train dataset has 3541 observations and 52 Variables.

```
> colnames(my_train)=make.names(colnames(my_train))
```

The column names of the train dataset are converted into column names with space replaced by "." As it is easy to read the column names and take it further for the analysis of the raw data.

```
> names(my_train)
```

[1]	"Num"	"Networth.Next.Year"	"Total.assets"
[4]	"Net.worth"	"Total.income"	"Change.in.stock"
[7]	"Total.expenses"	"Profit.after.tax"	"PBDITA"
[10]	"PBT"	"Cash.profit"	"PBDITA.as...of.total.income"
[13]	"PBT.as...of.total.income"	"PAT.as...of.total.income"	"Cash.profit.as...of.total.income"
[16]	"PAT.as...of.net.worth"	"Sales"	"Income.from.financial.services"
[19]	"Other.income"	"Total.capital"	"Reserves.and.funds"
[22]	"Deposits..accepted.by.commercial.banks."	"Borrowings"	"Current.liabilities...provisions"
[25]	"Deferred.tax.liability"	"Shareholders.funds"	"Cumulative.retained.profits"
[28]	"Capital.employed"	"TOL.TNW"	"Total.term.liabilities...tangible.net.worth"
[31]	"Contingent.liabilities...Net.worth...."	"Contingent.liabilities"	"Net.fixed.assets"
[34]	"Investments"	"Current.assets"	"Net.working.capital"
[37]	"Quick.ratio..times."	"Current.ratio..times."	"Debt.to.equity.ratio..times."
[40]	"Cash.to.current.liabilities..times."	"Cash.to.average.cost.of.sales.per.day"	"Creditors.turnover"
[43]	"Debtors.turnover"	"Finished.goods.turnover"	"WIP.turnover"
[46]	"Raw.material.turnover"	"Shares.outstanding"	"Equity.face.value"
[49]	"EPS"	"Adjusted.EPS"	"Total.liabilities"
[52]	"PE.on.BSE"		

The column names shows the variables are identified in the train datasets, in which Net Worth Next Year is the calculation of overall ratios, liabilities, profitability, liquidity, assets and working capital of the companies. The variables are identified as variables spaces are replaced by "." in train dataset.

> str(my_train)

```
tibble [3,541 × 52] (S3: tbl_df/tbl/data.frame)
$ Num                                    : num [1:3541] 1 2 3 4 5 6 7 8 9 10 ...
$ Networth.Next.Year                     : num [1:3541] 8890.6 394.3 92.2 2.7 109 ...
$ Total.assets                           : num [1:3541] 17512.3 941 232.8 2.7 478.5 ...
$ Net.worth                              : num [1:3541] 7093.2 351.5 100.6 2.7 107.6 ...
$ Total.income                           : num [1:3541] 24965 1527 477 NA 1580 ...
$ Change.in.stock                        : num [1:3541] 235.8 42.7 -5.2 NA -17 ...
$ Total.expenses                         : num [1:3541] 23658 1455 479 NA 1558 ...
$ Profit.after.tax                       : num [1:3541] 1543.2 115.2 -6.6 NA 5.5 ...
$ PBDITA                                 : num [1:3541] 2860.2 283 5.8 NA 31 ...
$ PBT                                    : num [1:3541] 2417.2 188.4 -6.6 NA 6.3 ...
$ Cash.profit                            : num [1:3541] 1872.8 158.6 0.3 NA 11.9 ...
$ PBDITA.as...of.total.income            : num [1:3541] 11.46 18.53 1.22 0 1.96 ...
$ PBT.as...of.total.income               : num [1:3541] 9.68 12.33 -1.38 0 0.4 ...
$ PAT.as...of.total.income               : num [1:3541] 6.18 7.54 -1.38 0 0.35 2.81 0 0.72 8.29 -2.88 ...
$ Cash.profit.as...of.total.income       : num [1:3541] 7.5 10.38 0.06 0 0.75 ...
$ PAT.as...of.net.worth                  : num [1:3541] 23.78 38.08 -6.35 0 5.25 ...
$ Sales                                  : num [1:3541] 24458 1504 476 NA 1575 ...
$ Income.from.financial.services         : num [1:3541] 158 4 1.5 NA 3.9 6.4 NA NA 7.3 NA ...
$ Other.income                           : num [1:3541] 297.2 15.9 0.2 NA 0.9 ...
$ Total.capital                          : num [1:3541] 423.8 115.5 81.4 0.5 6.2 ...
$ Reserves.and.funds                     : num [1:3541] 6822.8 257.8 19.2 2.2 161.8 ...
$ Deposits..accepted.by.commercial.banks. : logi [1:3541] NA NA NA NA NA NA ...
$ Borrowings                             : num [1:3541] 14.9 272.5 35.4 NA 193.1 ...
$ Current.liabilities...provisions       : num [1:3541] 9965.9 210 96.8 NA 112.8 ...
$ Deferred.tax.liability                 : num [1:3541] 284.9 85.2 NA NA 4.6 ...
$ Shareholders.funds                     : num [1:3541] 7093.2 351.5 100.6 2.7 107.6 ...
$ Cumulative.retained.profits            : num [1:3541] 6263.3 247.4 32.4 2.2 82.7 ...
$ Capital.employed                       : num [1:3541] 7108.1 624 136 2.7 300.7 ...
$ TOL.TNW                                : num [1:3541] 1.33 1.23 1.44 0 2.83 1.8 0.03 5.17 1.05 3.25 ...
$ Total.term.liabilities...tangible.net.worth : num [1:3541] 0 0.34 0.29 0 1.59 0.37 0.03 0.94 0.3 0.54 ...
$ Contingent.liabilities...Net.worth....  : num [1:3541] 14.8 19.2 45.8 0 34.9 ...
$ Contingent.liabilities                 : num [1:3541] 1049.7 67.6 46.1 NA 37.6 ...
$ Net.fixed.assets                       : num [1:3541] 1900.2 286.4 38.7 2.5 94.8 ...
$ Investments                            : num [1:3541] 1069.6 2.2 4.3 NA 7.4 ...
$ Current.assets                         : num [1:3541] 13277.5 563.9 167.5 0.2 349.7 ...
$ Net.working.capital                    : num [1:3541] 3588.5 203.5 59.6 0.2 215.8 ...
$ Quick.ratio..times.                    : num [1:3541] 1.18 0.95 1.11 NA 1.41 0.48 NA 0.54 0.59 0.39 ...
$ Current.ratio..times.                  : num [1:3541] 1.37 1.56 1.55 NA 2.54 1.27 NA 1.15 1.58 0.5 ...
$ Debt.to.equity.ratio..times.           : num [1:3541] 0 0.78 0.35 0 1.79 1.09 0.32 2.31 0.94 3.13 ...
$ Cash.to.current.liabilities..times.    : num [1:3541] 0.43 0.06 0.21 NA 0 0.11 NA 0.04 0.19 0 ...
$ Cash.to.average.cost.of.sales.per.day  : num [1:3541] 68.21 5.96 17.07 NA 0 ...
$ Creditors.turnover                     : chr [1:3541] "3.62" "9.8000000000000007" "5.28" "0" ...
$ Debtors.turnover                       : chr [1:3541] "3.85" "5.7" "5.07" "0" ...
$ Finished.goods.turnover                : chr [1:3541] "200.55" "14.21" "9.24" NA ...
$ WIP.turnover                           : chr [1:3541] "21.78" "7.49" "0.23" NA ...
$ Raw.material.turnover                  : chr [1:3541] "7.71" "11.46" NA "0" ...
$ Shares.outstanding                     : chr [1:3541] "42381675" "11550000" "8149090" "52404" ...
$ Equity.face.value                      : chr [1:3541] "10" "10" "10" "10" ...
$ EPS                                    : num [1:3541] 35.52 9.97 -0.5 0 7.91 ...
$ Adjusted.EPS                           : num [1:3541] 7.1 9.97 -0.5 0 7.91 ...
$ Total.liabilities                      : num [1:3541] 17512.3 941 232.8 2.7 478.5 ...
$ PE.on.BSE                              : chr [1:3541] "27.31" "8.17" "-5.76" "NA" ...
```

The column names are identified with the nature of its type as character and numeric values in the table and data frame structure of dataset. Character variables are measured in Creditors Turnover, Debtor Turnover, Finished Goods Turnover, WIP Turnover, Raw material turnover, share outstanding, equity face value and PE on BSE. Missing values are also identified in the variables.

Logical values are presented in the Deposits accepted by commercial banks are measured and the variables are mentioned as Missing values in the dataset.

```
> my_train$Creditors.turnover=as.numeric(my_train$Creditors.turnover)
> my_train$Debtors.turnover=as.numeric(my_train$Debtors.turnover)
> my_train$Finished.goods.turnover=as.numeric(my_train$Finished.goods.turnover)
> my_train$WIP.turnover=as.numeric(my_train$WIP.turnover)
> my_train$Raw.material.turnover=as.numeric(my_train$Raw.material.turnover)
> my_train$Shares.outstanding=as.numeric(my_train$Shares.outstanding)
> my_train$Equity.face.value=as.numeric(my_train$Equity.face.value)
> my_train$PE.on.BSE=as.numeric(my_train$PE.on.BSE)
Warning message:
NAs introduced by coercion
```

The character variables are converted to numeric variables as they are continuous variables and the values are further interpretable and correlated for the analysis of the values in logistics regression model.

The variables are introduced by NAs in coercion and the values are taken as the original values find in the dataset.

```
> str(my_train)
```

The train dataset variables are converted into numeric variables and the continuous variables are further analysed with the logistics regression and prediction of the variables in the fitted model of the values. The logistics values are measured with the deposits of the customer in various companies. The values are observed with Missing values and outlier identification of the variables. Sales, Share outstanding, liabilities, Net worth next year are the basics of the numeric variables. The variables are correlated to the various variables in the correlation of the values.

```
tibble [3,541 × 52] (S3: tbl_df/tbl/data.frame)
$ Num                                  : num [1:3541] 1 2 3 4 5 6 7 8 9 10 ...
$ Networth.Next.Year                   : num [1:3541] 8890.6 394.3 92.2 2.7 109 ...
$ Total.assets                         : num [1:3541] 17512.3 941 232.8 2.7 478.5 ...
$ Net.worth                            : num [1:3541] 7093.2 351.5 100.6 2.7 107.6 ...
$ Total.income                         : num [1:3541] 24965 1527 477 NA 1580 ...
$ Change.in.stock                      : num [1:3541] 235.8 42.7 -5.2 NA -17 ...
$ Total.expenses                       : num [1:3541] 23658 1455 479 NA 1558 ...
$ Profit.after.tax                     : num [1:3541] 1543.2 115.2 -6.6 NA 5.5 ...
$ PBDITA                               : num [1:3541] 2860.2 283 5.8 NA 31 ...
$ PBT                                  : num [1:3541] 2417.2 188.4 -6.6 NA 6.3 ...
$ Cash.profit                          : num [1:3541] 1872.8 158.6 0.3 NA 11.9 ...
$ PBDITA.as...of.total.income          : num [1:3541] 11.46 18.53 1.22 0 1.96 ...
$ PBT.as...of.total.income             : num [1:3541] 9.68 12.33 -1.38 0 0.4 ...
$ PAT.as...of.total.income             : num [1:3541] 6.18 7.54 -1.38 0 0.35 2.81 0 0.72 8.29 -2.88 ...
$ Cash.profit.as...of.total.income     : num [1:3541] 7.5 10.38 0.06 0 0.75 ...
$ PAT.as...of.net.worth                : num [1:3541] 23.78 38.08 -6.35 0 5.25 ...
$ Sales                                : num [1:3541] 24458 1504 476 NA 1575 ...
$ Income.from.financial.services       : num [1:3541] 158 4 1.5 NA 3.9 6.4 NA NA 7.3 NA ...
$ Other.income                         : num [1:3541] 297.2 15.9 0.2 NA 0.9 ...
$ Total.capital                        : num [1:3541] 423.8 115.5 81.4 0.5 6.2 ...
$ Reserves.and.funds                   : num [1:3541] 6822.8 257.8 19.2 2.2 161.8 ...
$ Deposits..accepted.by.commercial.banks. : logi [1:3541] NA NA NA NA NA NA ...
$ Borrowings                           : num [1:3541] 14.9 272.5 35.4 NA 193.1 ...
$ Current.liabilities...provisions     : num [1:3541] 9965.9 210 96.8 NA 112.8 ...
$ Deferred.tax.liability               : num [1:3541] 284.9 85.2 NA NA 4.6 ...
$ Shareholders.funds                   : num [1:3541] 7093.2 351.5 100.6 2.7 107.6 ...
$ Cumulative.retained.profits          : num [1:3541] 6263.3 247.4 32.4 2.2 82.7 ...
$ Capital.employed                     : num [1:3541] 7108.1 624 136 2.7 300.7 ...
$ TOL.TNW                              : num [1:3541] 1.33 1.23 1.44 0 2.83 1.8 0.03 5.17 1.05 3.25 ...
$ Total.term.liabilities...tangible.net.worth : num [1:3541] 0 0.34 0.29 0 1.59 0.37 0.03 0.94 0.3 0.54 ...
$ Contingent.liabilities...Net.worth....  : num [1:3541] 14.8 19.2 45.8 0 34.9 ...
$ Contingent.liabilities               : num [1:3541] 1049.7 67.6 46.1 NA 37.6 ...
$ Net.fixed.assets                     : num [1:3541] 1900.2 286.4 38.7 2.5 94.8 ...
$ Investments                          : num [1:3541] 1069.6 2.2 4.3 NA 7.4 ...
$ Current.assets                       : num [1:3541] 13277.5 563.9 167.5 0.2 349.7 ...
$ Net.working.capital                  : num [1:3541] 3588.5 203.5 59.6 0.2 215.8 ...
$ Quick.ratio..times.                  : num [1:3541] 1.18 0.95 1.11 NA 1.41 0.48 NA 0.54 0.59 0.39 ...
$ Current.ratio..times.                : num [1:3541] 1.37 1.56 1.55 NA 2.54 1.27 NA 1.15 1.58 0.5 ...
$ Debt.to.equity.ratio..times.         : num [1:3541] 0 0.78 0.35 0 1.79 1.09 0.32 2.31 0.94 3.13 ...
$ Cash.to.current.liabilities..times.  : num [1:3541] 0.43 0.06 0.21 NA 0 0.11 NA 0.04 0.19 0 ...
$ Cash.to.average.cost.of.sales.per.day : num [1:3541] 68.21 5.96 17.07 NA 0 ...
$ Creditors.turnover                   : num [1:3541] 3.62 9.8 5.28 0 13 ...
$ Debtors.turnover                     : num [1:3541] 3.85 5.7 5.07 0 9.46 ...
$ Finished.goods.turnover              : num [1:3541] 200.55 14.21 9.24 NA 12.68 ...
$ WIP.turnover                         : num [1:3541] 21.78 7.49 0.23 NA 7.9 ...
$ Raw.material.turnover                : num [1:3541] 7.71 11.46 NA 0 17.03 ...
$ Shares.outstanding                   : num [1:3541] 42381675 11550000 8149090 52404 619635 ...
$ Equity.face.value                    : num [1:3541] 10 10 10 10 10 10 10 NA 10 10 ...
$ EPS                                  : num [1:3541] 35.52 9.97 -0.5 0 7.91 ...
$ Adjusted.EPS                         : num [1:3541] 7.1 9.97 -0.5 0 7.91 ...
$ Total.liabilities                    : num [1:3541] 17512.3 941 232.8 2.7 478.5 ...
$ PE.on.BSE                            : num [1:3541] 27.31 8.17 -5.76 NA NA ...
```

> summary(my_train)

Num Min. : 1 1st Qu.: 886 Median :1773 Mean :1772 3rd Qu.:2658 Max. :3545	Networth.Next.Year Min. :-74265.6 1st Qu.: 31.7 Median : 116.3 Mean : 1616.3 3rd Qu.: 456.1 Max. :805773.4	Total.assets Min. : 0.1 1st Qu.: 91.3 Median : 309.7 Mean : 3443.4 3rd Qu.: 1098.7 Max. :1176509.2	Net.worth Min. : 0.0 1st Qu.: 31.3 Median : 102.3 Mean : 1295.9 3rd Qu.: 377.3 Max. :613151.6	Total.income Min. : 0.0 1st Qu.: 106.4 Median : 444.9 Mean : 4582.8 3rd Qu.: 1440.9 Max. :2442828.2 NA's :198	Change.in.stock Min. :-3029.40 1st Qu.: -1.80 Median : 1.60 Mean : 41.49 3rd Qu.: 18.05 Max. :14185.50 NA's :458	Total.expenses Min. : -0.1 1st Qu.: 95.8 Median : 407.7 Mean : 4262.9 3rd Qu.: 1359.8 Max. :2366035.3 NA's :139
Profit.after.tax Min. : -3908.30 1st Qu.: 0.50 Median : 8.80 Mean : 277.36 3rd Qu.: 52.27 Max. :119439.10 NA's :131	PBDITA Min. : -440.7 1st Qu.: 6.9 Median : 35.4 Mean : 578.1 3rd Qu.: 150.2 Max. :208576.5 NA's :131	PBT Min. : -3894.80 1st Qu.: 0.70 Median : 12.40 Mean : 383.81 3rd Qu.: 71.97 Max. :145292.60 NA's :131	Cash.profit Min. : -2245.70 1st Qu.: 2.90 Median : 18.85 Mean : 392.07 3rd Qu.: 93.20 Max. :176911.80 NA's :131	PBDITA.as...of.total.income Min. :-6400.000 1st Qu.: 5.000 Median : 9.660 Mean : 4.571 3rd Qu.: 16.390 Max. : 100.000 NA's :68	PBT.as...of.total.income Min. :-21340.00 1st Qu.: 0.55 Median : 3.31 Mean : -17.28 3rd Qu.: 8.80 Max. : 100.00 NA's :68	
PAT.as...of.total.income Min. :-21340.00 1st Qu.: 0.35 Median : 2.34 Mean : -19.20 3rd Qu.: 6.34 Max. : 150.00 NA's :68	Cash.profit.as...of.total.income Min. :-15020.000 1st Qu.: 2.020 Median : 5.640 Mean : -8.229 3rd Qu.: 10.700 Max. : 100.000 NA's :68	PAT.as...of.net.worth Min. :-748.72 1st Qu.: 0.00 Median : 7.92 Mean : 10.27 3rd Qu.: 20.19 Max. :2466.67	Sales Min. : 0.1 1st Qu.: 112.7 Median : 453.1 Mean : 4549.5 3rd Qu.: 1433.6 Max. :2384984.4 NA's :259	Income.from.financial.services Min. : 0.00 1st Qu.: 0.40 Median : 1.80 Mean : 80.84 3rd Qu.: 9.68 Max. :51938.20 NA's :935		
Other.income Min. : 0.00 1st Qu.: 0.40 Median : 1.40 Mean : 41.36 3rd Qu.: 5.97 Max. :42856.70 NA's :1295	Total.capital Min. : 0.1 1st Qu.: 13.1 Median : 42.1 Mean : 216.6 3rd Qu.: 100.3 Max. :78273.2 NA's :4	Reserves.and.funds Min. : -6525.9 1st Qu.: 5.0 Median : 54.8 Mean : 1163.8 3rd Qu.: 277.3 Max. :625137.8 NA's :85	Deposits..accepted.by.commercial.banks. Mode:logical NA's:3541	Borrowings Min. : 0.10 1st Qu.: 23.95 Median : 99.20 Mean : 1122.28 3rd Qu.: 352.60 Max. :278257.30 NA's :366		
Current.liabilities...provisions Min. : 0.1 1st Qu.: 17.8 Median : 69.4 Mean : 940.6 3rd Qu.: 261.7 Max. :352240.3 NA's :96	Deferred.tax.liability Min. : 0.1 1st Qu.: 3.2 Median : 13.4 Mean : 227.2 3rd Qu.: 50.0 Max. :72796.6 NA's :1140	Shareholders.funds Min. : 0.0 1st Qu.: 32.0 Median : 105.6 Mean : 1322.1 3rd Qu.: 393.2 Max. :613151.6	Cumulative.retained.profits Min. : -6534.3 1st Qu.: 1.1 Median : 37.1 Mean : 890.5 3rd Qu.: 202.3 Max. :390133.8 NA's :38	Capital.employed Min. : 0.0 1st Qu.: 60.8 Median : 214.7 Mean : 2328.3 3rd Qu.: 767.3 Max. :891408.9	TOL.TNW Min. :-350.480 1st Qu.: 0.600 Median : 1.430 Mean : 3.994 3rd Qu.: 2.830 Max. : 473.000	
Total.term.liabilities...tangible.net.worth Min. :-325.600 1st Qu.: 0.050 Median : 0.340 Mean : 1.844 3rd Qu.: 1.000 Max. : 456.000	Contingent.liabilities....Net.worth.... Min. : 0.00 1st Qu.: 0.00 Median : 5.33 Mean : 53.94 3rd Qu.: 30.76 Max. :14704.27	Contingent.liabilities Min. : 0.1 1st Qu.: 6.3 Median : 38.0 Mean : 932.9 3rd Qu.: 192.7 Max. :559506.8 NA's :1188	Net.fixed.assets Min. : 0.0 1st Qu.: 26.0 Median : 93.5 Mean : 1189.7 3rd Qu.: 344.9 Max. :636604.6 NA's :118	Investments Min. : 0.00 1st Qu.: 1.00 Median : 8.35 Mean : 694.73 3rd Qu.: 64.30 Max. :199978.60 NA's :1435		
Current.assets Min. : 0.1 1st Qu.: 36.2 Median : 145.1 Mean : 1293.4 3rd Qu.: 502.2 Max. :354815.2 NA's :66	Net.working.capital Min. :-63839.0 1st Qu.: -1.1 Median : 16.2 Mean : 138.6 3rd Qu.: 84.2 Max. : 85782.8 NA's :32	Quick.ratio..times. Min. : 0.000 1st Qu.: 0.410 Median : 0.670 Mean : 1.401 3rd Qu.: 1.030 Max. :341.000 NA's :93	Current.ratio..times. Min. : 0.00 1st Qu.: 0.93 Median : 1.23 Mean : 2.13 3rd Qu.: 1.71 Max. :505.00 NA's :93	Debt.to.equity.ratio..times. Min. : 0.00 1st Qu.: 0.22 Median : 0.79 Mean : 2.78 3rd Qu.: 1.75 Max. :456.00	Cash.to.current.liabilities..times. Min. : 0.0000 1st Qu.: 0.0200 Median : 0.0700 Mean : 0.4904 3rd Qu.: 0.1900 Max. :165.0000 NA's :93	
Cash.to.average.cost.of.sales.per.day Min. : 0.00 1st Qu.: 2.79 Median : 8.03 Mean : 158.44 3rd Qu.: 21.79 Max. :128040.76 NA's :85	Creditors.turnover Min. : 0.000 1st Qu.: 3.700 Median : 6.095 Mean : 15.446 3rd Qu.: 11.490 Max. :2401.000 NA's :333	Debtors.turnover Min. : 0.00 1st Qu.: 3.76 Median : 6.32 Mean : 17.04 3rd Qu.: 11.68 Max. :3135.20 NA's :328	Finished.goods.turnover Min. : -0.09 1st Qu.: 8.20 Median : 17.27 Mean : 87.08 3rd Qu.: 40.35 Max. :17947.60 NA's :740	WIP.turnover Min. : -0.18 1st Qu.: 5.10 Median : 9.76 Mean : 27.93 3rd Qu.: 20.24 Max. :5651.40 NA's :640	Raw.material.turnover Min. : -2.00 1st Qu.: 2.99 Median : 6.40 Mean : 19.09 3rd Qu.: 11.85 Max. :21092.00 NA's :361	
Shares.outstanding Min. :-2.147e+09 1st Qu.: 1.316e+06 Median : 4.672e+06 Mean : 2.207e+07 3rd Qu.: 1.065e+07 Max. : 4.130e+09 NA's :692	Equity.face.value Min. :-999999 1st Qu.: 10 Median : 10 Mean : -1334 3rd Qu.: 10 Max. : 100000 NA's :692	EPS Min. :-843181.8 1st Qu.: 0.0 Median : 1.4 Mean : -220.3 3rd Qu.: 9.6 Max. : 34522.5	Adjusted.EPS Min. :-843181.8 1st Qu.: 0.0 Median : 1.2 Mean : -221.5 3rd Qu.: 7.5 Max. : 34522.5	Total.liabilities Min. : 0.1 1st Qu.: 91.3 Median : 309.7 Mean : 3443.4 3rd Qu.: 1098.7 Max. :1176509.2	PE.on.BSE Min. :-1116.64 1st Qu.: 3.28 Median : 9.10 Mean : 63.91 3rd Qu.: 17.79 Max. :51002.74 NA's :2194	

Summarization of the variables are measured for the 1st quartile, 3rd Quartile, Minimum value, maximum value, median, mean for each variables and the missing values are identified with the total counts of each variables. The minimum of some variables are measured with negative values and the maximum of some variables are measured with higher values are taken as outlier identification. The net worth next year is identified with negative values as defaulters which reflects the loss of the profit for the next year and net worth next year is identified by the positive values is take as non-defaulters for the company with positive growth of the variables.

> dim(my_test)

[1] 715 52

The dimension of the test variables is with 715 observations and 52 variables.

> colnames(my_test)=make.names(colnames(my_test))

The column names are converted with the spaces to “.”

> names(my_test)

```
[1]  "Num"                                  "Default...1"                          "Total.assets"
[4]  "Net.worth"                            "Total.income"                         "Change.in.stock"
[7]  "Total.expenses"                       "Profit.after.tax"                     "PBDITA"
[10] "PBT"                                  "Cash.profit"                          "PBDITA.as...of.total.income"
[13] "PBT.as...of.total.income"             "PAT.as...of.total.income"             "Cash.profit.as...of.total.income"
[16] "PAT.as...of.net.worth"                "Sales"                                "Income.from.financial.services"
[19] "Other.income"                         "Total.capital"                        "Reserves.and.funds"
[22] "Deposits..accepted.by.commercial.b"   "Borrowings"                           "Current.liabilities...provisions"
[25] "Deferred.tax.liability"               "Shareholders.funds"                   "Cumulative.retained.profits"
[28] "Capital.employed"                     "TOL.TNW"                              "Total.term.liabilities...tangible.net.worth"
[31] "Contingent.liabilities...Net.worth...." "Contingent.liabilities"             "Net.fixed.assets"
[34] "Investments"                          "Current.assets"                       "Net.working.capital"
[37] "Quick.ratio..times."                  "Current.ratio..times."                "Debt.to.equity.ratio..times."
[40] "Cash.to.current.liabilities..times."  "Cash.to.average.cost.of.sales.per.day" "Creditors.turnover"
[43] "Debtors.turnover"                     "Finished.goods.turnover"              "WIP.turnover"
[46] "Raw.material.turnover"                "Shares.outstanding"                   "Equity.face.value"
[49] "EPS"                                  "Adjusted.EPS"                         "Total.liabilities"
[52] "PE.on.BSE"
```

The variables are measured with Default variables which is used to predict the model and the values of the variables in the train as well as in validation dataset.

> str(my_test)

The column names are identified with the nature of its type as character and numeric values in the table and data frame structure of dataset. Character variables are measured in Creditors Turnover, Debtor Turnover, Finished Goods Turnover, WIP Turnover, Raw material turnover, share outstanding, equity face value and PE on BSE. Missing values are also identified in the variables. The logical variable is identified with deposits of the costumer in the companies and the values are mentioned with missing values.

```
tibble [715 × 52] (S3: tbl_df/tbl/data.frame)
$ Num                                : num [1:715] 1 2 3 4 5 6 7 8 9 10 ...
$ Default...1                        : num [1:715] 0 0 1 0 0 0 0 0 0 0 ...
$ Total.assets                       : num [1:715] 971 675 532 858 823 ...
$ Net.worth                          : num [1:715] 276 212 120 201 349 ...
$ Total.income                       : num [1:715] 2185 819 564 3576 1034 ...
$ Change.in.stock                    : num [1:715] 14.2 10.4 -28.1 -0.6 28.9 -0.5 NA -7.7 27.2 -0.2 ...
$ Total.expenses                     : num [1:715] 2099 810 578 3613 1042 ...
$ Profit.after.tax                   : num [1:715] 100.2 19.7 -42.4 -37.5 21.4 ...
$ PBDITA                             : num [1:715] 285.6 116 -31 68.2 90.1 ...
$ PBT                                : num [1:715] 152.1 33.7 -56 25.7 29.7 ...
$ Cash.profit                        : num [1:715] 182.3 50.5 -35.3 37.3 62.7 ...
$ PBDITA.as...of.total.income        : num [1:715] 13.07 14.16 -5.5 1.91 8.71 ...
$ PBT.as...of.total.income           : num [1:715] 6.96 4.11 -9.94 0.72 2.87 ...
$ PAT.as...of.total.income           : num [1:715] 4.59 2.4 -7.52 -1.05 2.07 ...
$ Cash.profit.as...of.total.income   : num [1:715] 8.34 6.16 -6.26 1.04 6.06 ...
$ PAT.as...of.net.worth              : num [1:715] 42.11 10.66 -31.2 0 6.31 ...
$ Sales                              : num [1:715] 2171 817 552 3573 1027 ...
```

```
$ Income.from.financial.services                : num [1:715] 2.3 0.8 9.1 1 0.7 ...
$ Other.income                                  : num [1:715] NA 0.2 2.1 1.5 2.3 0.1 NA NA 0.1 0.1 ...
$ Total.capital                                 : num [1:715] 48 114 47.1 50.5 33 ...
$ Reserves.and.funds                            : num [1:715] 413.1 97.6 227.4 150.9 316.2 ...
$ Deposits..accepted.by.commercial.banks.       : logi [1:715] NA NA NA NA NA NA ...
$ Borrowings                                    : num [1:715] 177.3 339.8 17.5 524.2 162.3 ...
$ Current.liabilities...provisions              : num [1:715] 328.5 100.5 240.1 75.2 299.6 ...
$ Deferred.tax.liability                        : num [1:715] 3.7 23.1 NA 56.7 12.2 2.1 1.9 4.4 2.9 NA ...
$ Shareholders.funds                            : num [1:715] 276 212 120 201 349 ...
$ Cumulative.retained.profits                   : num [1:715] 227.8 97.6 69.9 150.9 316.2 ...
$ Capital.employed                              : num [1:715] 453 551 138 726 512 ...
$ TOL.TNW                                       : num [1:715] 1.8 2.01 1.73 2.94 1.02 0.86 0.06 1.92 0.37 1.96 ...
$ Total.term.liabilities...tangible.net.worth   : num [1:715] 0.27 0.72 0.09 0.81 0.1 0.11 0.05 0.78 0 1.81 ...
$ Contingent.liabilities...Net.worth....        : num [1:715] 112.94 5.77 102.83 0.65 28.78 ...
$ Contingent.liabilities                        : num [1:715] 311.5 12.2 123.6 1.3 100.5 ...
$ Net.fixed.assets                              : num [1:715] 332 199 270 263 191 ...
$ Investments                                   : num [1:715] NA NA 0.7 NA NA NA 17.3 2.6 NA NA ...
$ Current.assets                                : num [1:715] 560 407 148 536 472 ...
$ Net.working.capital                           : num [1:715] 134.2 123.6 -97.1 99.6 75.3 ...
$ Quick.ratio..times.                           : num [1:715] 0.92 0.48 0.32 0.51 0.58 0.97 166 0.52 0.88 0.6 ...
$ Current.ratio..times.                         : num [1:715] 1.31 1.39 0.6 1.23 1.19 1.86 166 1.56 1.19 0.55 ...
$ Debt.to.equity.ratio..times.                  : num [1:715] 0.64 1.61 0.15 2.6 0.46 0.32 0.05 1.24 0 1.81 ...
$ Cash.to.current.liabilities..times.           : num [1:715] 0.09 0.03 0.04 0.08 0.08 0 165 0.03 0.35 0.23 ...
$ Cash.to.average.cost.of.sales.per.day         : num [1:715] 7.56 3.88 4.63 3.71 11.15 ...
$ Creditors.turnover                            : chr [1:715] "5.94" "10.59" "2.35" "NA" ...
$ Debtors.turnover                              : chr [1:715] "5.74" "6.03" "9.6" "NA" ...
$ Finished.goods.turnover                       : chr [1:715] "25.11" "28.96" "8.23" "NA" ...
$ WIP.turnover                                  : chr [1:715] "20.010000000000002" "18.649999999999999" "6.6" "NA" ...
$ Raw.material.turnover                         : chr [1:715] "17.579999999999998" "2.67" "3.77" "NA" ...
$ Shares.outstanding                            : chr [1:715] "4800000" "11400000" "471285" "5050000" ...
$ Equity.face.value                             : chr [1:715] "10" "10" "100" "10" ...
$ EPS                                           : num [1:715] 18.6 1.65 -90.39 -7.09 5.9 ...
$ Adjusted.EPS                                  : num [1:715] 18.6 1.65 -90.39 -7.09 5.9 ...
$ Total.liabilities                             : num [1:715] 971 675 532 858 823 ...
$ PE.on.BSE                                     : chr [1:715] "NA" "NA" "-15.5" "-0.16" ...
```

```
> my_test$Creditors.turnover=as.numeric(my_test$Creditors.turnover)
> my_test$Debtors.turnover=as.numeric(my_test$Debtors.turnover)
> my_test$Finished.goods.turnover=as.numeric(my_test$Finished.goods.turnover)
> my_test$WIP.turnover=as.numeric(my_test$WIP.turnover)
> my_test$Raw.material.turnover=as.numeric(my_test$Raw.material.turnover)
> my_test$Shares.outstanding=as.numeric(my_test$Shares.outstanding)
> my_test$Equity.face.value=as.numeric(my_test$Equity.face.value)
> my_test$PE.on.BSE=as.numeric(my_test$PE.on.BSE)
Warning message:
NAs introduced by coercion
```

The character type variables are converted to the numeric variables with the missing values and outlier of the variables.

> str(my_test)

```
tibble [715 × 52] (S3: tbl_df/tbl/data.frame)
$ Num                                    : num [1:715] 1 2 3 4 5 6 7 8 9 10 ...
$ Default...1                            : num [1:715] 0 0 1 0 0 0 0 0 0 0 ...
$ Total.assets                           : num [1:715] 971 675 532 858 823 ...
$ Net.worth                              : num [1:715] 276 212 120 201 349 ...
$ Total.income                           : num [1:715] 2185 819 564 3576 1034 ...
$ Change.in.stock                        : num [1:715] 14.2 10.4 -28.1 -0.6 28.9 -0.5 NA -7.7 27.2 -0.2 ...
$ Total.expenses                         : num [1:715] 2099 810 578 3613 1042 ...
$ Profit.after.tax                       : num [1:715] 100.2 19.7 -42.4 -37.5 21.4 ...
$ PBDITA                                 : num [1:715] 285.6 116 -31 68.2 90.1 ...
$ PBT                                    : num [1:715] 152.1 33.7 -56 25.7 29.7 ...
$ Cash.profit                            : num [1:715] 182.3 50.5 -35.3 37.3 62.7 ...
$ PBDITA.as...of.total.income            : num [1:715] 13.07 14.16 -5.5 1.91 8.71 ...
$ PBT.as...of.total.income               : num [1:715] 6.96 4.11 -9.94 0.72 2.87 ...
$ PAT.as...of.total.income               : num [1:715] 4.59 2.4 -7.52 -1.05 2.07 ...
$ Cash.profit.as...of.total.income       : num [1:715] 8.34 6.16 -6.26 1.04 6.06 ...
$ PAT.as...of.net.worth                  : num [1:715] 42.11 10.66 -31.2 0 6.31 ...
$ Sales                                  : num [1:715] 2171 817 552 3573 1027 ...
$ Income.from.financial.services         : num [1:715] 2.3 0.8 9.1 1 0.7 ...
$ Other.income                           : num [1:715] NA 0.2 2.1 1.5 2.3 0.1 NA NA 0.1 0.1 ...
$ Total.capital                          : num [1:715] 48 114 47.1 50.5 33 ...
$ Reserves.and.funds                     : num [1:715] 413.1 97.6 227.4 150.9 316.2 ...
$ Deposits..accepted.by.commercial.banks. : logi [1:715] NA NA NA NA NA NA ...
$ Borrowings                             : num [1:715] 177.3 339.8 17.5 524.2 162.3 ...
$ Current.liabilities...provisions       : num [1:715] 328.5 100.5 240.1 75.2 299.6 ...
$ Deferred.tax.liability                 : num [1:715] 3.7 23.1 NA 56.7 12.2 2.1 1.9 4.4 2.9 NA ...
$ Shareholders.funds                     : num [1:715] 276 212 120 201 349 ...
$ Cumulative.retained.profits            : num [1:715] 227.8 97.6 69.9 150.9 316.2 ...
$ Capital.employed                       : num [1:715] 453 551 138 726 512 ...
$ TOL.TNW                                : num [1:715] 1.8 2.01 1.73 2.94 1.02 0.86 0.06 1.92 0.37 1.96 ...
$ Total.term.liabilities...tangible.net.worth : num [1:715] 0.27 0.72 0.09 0.81 0.1 0.11 0.05 0.78 0 1.81 ...
$ Contingent.liabilities...Net.worth....  : num [1:715] 112.94 5.77 102.83 0.65 28.78 ...
$ Contingent.liabilities                 : num [1:715] 311.5 12.2 123.6 1.3 100.5 ...
$ Net.fixed.assets                       : num [1:715] 332 199 270 263 191 ...
$ Investments                            : num [1:715] NA NA 0.7 NA NA NA 17.3 2.6 NA NA ...
$ Current.assets                         : num [1:715] 560 407 148 536 472 ...
$ Net.working.capital                    : num [1:715] 134.2 123.6 -97.1 99.6 75.3 ...
$ Quick.ratio..times.                    : num [1:715] 0.92 0.48 0.32 0.51 0.58 0.97 166 0.52 0.88 0.6 ...
$ Current.ratio..times.                  : num [1:715] 1.31 1.39 0.6 1.23 1.19 1.86 166 1.56 1.19 0.55 ...
$ Debt.to.equity.ratio..times.           : num [1:715] 0.64 1.61 0.15 2.6 0.46 0.32 0.05 1.24 0 1.81 ...
$ Cash.to.current.liabilities..times.    : num [1:715] 0.09 0.03 0.04 0.08 0.08 0 165 0.03 0.35 0.23 ...
$ Cash.to.average.cost.of.sales.per.day  : num [1:715] 7.56 3.88 4.63 3.71 11.15 ...
$ Creditors.turnover                     : num [1:715] 5.94 10.59 2.35 NA 5.48 ...
$ Debtors.turnover                       : num [1:715] 5.74 6.03 9.6 NA 4.78 ...
$ Finished.goods.turnover                : num [1:715] 25.11 28.96 8.23 NA 6.28 ...
$ WIP.turnover                           : num [1:715] 20 18.6 6.6 NA 6.7 ...
$ Raw.material.turnover                  : num [1:715] 17.58 2.67 3.77 NA 3.7 ...
$ Shares.outstanding                     : num [1:715] 4800000 11400000 471285 5050000 3205946 ...
$ Equity.face.value                      : num [1:715] 10 10 100 10 10 100 10 NA 10 10 ...
$ EPS                                    : num [1:715] 18.6 1.65 -90.39 -7.09 5.9 ...
$ Adjusted.EPS                           : num [1:715] 18.6 1.65 -90.39 -7.09 5.9 ...
$ Total.liabilities                      : num [1:715] 971 675 532 858 823 ...
$ PE.on.BSE                              : num [1:715] NA NA -15.5 -0.16 NA NA NA NA NA NA ...
```

The variables are converted in to numeric variables and the structure of each variables are identified in the datasets.

> summary(my_test)

Num Min. : 1.0 1st Qu.:179.5 Median :358.0 Mean :358.0 3rd Qu.:536.5 Max. :715.0	Default...1 Min. :0.00000 1st Qu.:0.00000 Median :0.00000 Mean :0.07552 3rd Qu.:0.00000 Max. :1.00000	Total.assets Min. : 0.1 1st Qu.: 93.2 Median : 347.7 Mean : 4218.6 3rd Qu.: 1315.3 Max. :354727.3	Net.worth Min. : 0.1 1st Qu.: 34.4 Median : 120.9 Mean : 1629.7 3rd Qu.: 451.5 Max. :171840.0	Total.income Min. : 0.0 1st Qu.: 110.8 Median : 536.0 Mean : 5204.7 3rd Qu.: 1727.1 Max. :1028087.4 NA's :33	Change.in.stock Min. :-488.10 1st Qu.: -1.90 Median : 1.80 Mean : 54.66 3rd Qu.: 19.35 Max. :7540.00 NA's :92	Total.expenses Min. : 0.0 1st Qu.: 104.1 Median : 511.1 Mean : 4817.3 3rd Qu.: 1642.3 Max. :1014813.1 NA's :26
Profit.after.tax Min. : -998.00 1st Qu.: 0.68 Median : 10.20 Mean : 382.22 3rd Qu.: 68.95 Max. :62022.90 NA's :23	PBDITA Min. : -393.90 1st Qu.: 7.15 Median : 42.20 Mean : 743.35 3rd Qu.: 192.82 Max. :110557.10 NA's :23	PBT Min. : -993.90 1st Qu.: 1.00 Median : 14.25 Mean : 540.59 3rd Qu.: 90.50 Max. :94565.20 NA's :23	Cash.profit Min. : -894.60 1st Qu.: 3.28 Median : 22.05 Mean : 488.11 3rd Qu.: 120.30 Max. :71581.60 NA's :23	PBDITA.as...of.total.income Min. :-6400.000 1st Qu.: 4.702 Median : 9.780 Mean : -3.681 3rd Qu.: 16.753 Max. : 100.000 NA's :11	PBT.as...of.total.income Min. :-9700.000 1st Qu.: 0.623 Median : 3.450 Mean : -22.725 3rd Qu.: 9.725 Max. : 100.000 NA's :11	
PAT.as...of.total.income Min. :-9700.000 1st Qu.: 0.390 Median : 2.405 Mean : -24.147 3rd Qu.: 6.790 Max. : 100.000 NA's :11	Cash.profit.as...of.total.income Min. :-6400.000 1st Qu.: 1.930 Median : 5.835 Mean : -12.929 3rd Qu.: 10.982 Max. : 100.000 NA's :11	PAT.as...of.net.worth Min. :-194.520 1st Qu.: 0.000 Median : 8.710 Mean : 9.666 3rd Qu.: 20.215 Max. : 441.670	Sales Min. : 0.1 1st Qu.: 120.8 Median : 552.5 Mean : 5117.5 3rd Qu.: 1721.3 Max. :976884.0 NA's :46	Income.from.financial.services Min. : 0.10 1st Qu.: 0.50 Median : 2.00 Mean : 83.86 3rd Qu.: 10.10 Max. :8097.20 NA's :176		
Other.income Min. : 0.00 1st Qu.: 0.32 Median : 1.65 Mean : 128.16 3rd Qu.: 7.25 Max. :42856.70 NA's :261	Total.capital Min. : 0.1 1st Qu.: 14.1 Median : 45.3 Mean : 263.9 3rd Qu.: 121.1 Max. :41304.0 NA's :1	Reserves.and.funds Min. : -1125.00 1st Qu.: 7.32 Median : 57.45 Mean : 1440.70 3rd Qu.: 334.80 Max. :133684.20 NA's :13	Deposits..accepted.by.commercial.banks. Mode:logical NA's:715	Borrowings Min. : 0.20 1st Qu.: 25.92 Median : 105.50 Mean : 1439.86 3rd Qu.: 391.82 Max. :105175.30 NA's :65		
Current.liabilities...provisions Min. : 0.1 1st Qu.: 16.8 Median : 75.2 Mean : 1058.9 3rd Qu.: 300.4 Max. :112712.7 NA's :14	Deferred.tax.liability Min. : 0.10 1st Qu.: 3.10 Median : 14.70 Mean : 270.45 3rd Qu.: 62.42 Max. :27077.90 NA's :229	Shareholders.funds Min. : 0.1 1st Qu.: 35.5 Median : 124.0 Mean : 1646.0 3rd Qu.: 478.4 Max. :171840.0	Cumulative.retained.profits Min. : -2582.4 1st Qu.: 0.8 Median : 40.6 Mean : 1168.1 3rd Qu.: 244.5 Max. :128183.1 NA's :7	Capital.employed Min. : 0.10 1st Qu.: 64.35 Median : 246.10 Mean : 2954.96 3rd Qu.: 913.65 Max. :235389.50	TOL.TNW Min. :-350.480 1st Qu.: 0.595 Median : 1.400 Mean : 4.181 3rd Qu.: 2.800 Max. : 411.270	
Total.term.liabilities...tangible.net.worth Min. :-325.600 1st Qu.: 0.060 Median : 0.350 Mean : 1.906 3rd Qu.: 1.005 Max. : 292.020	Contingent.liabilities....Net.worth.... Min. : 0.00 1st Qu.: 0.00 Median : 5.52 Mean : 64.47 3rd Qu.: 31.49 Max. :6295.24	Contingent.liabilities Min. : 0.1 1st Qu.: 5.1 Median : 37.5 Mean : 1022.0 3rd Qu.: 217.1 Max. :72620.8 NA's :214	Net.fixed.assets Min. : 0.1 1st Qu.: 27.2 Median : 95.0 Mean : 1306.2 3rd Qu.: 409.2 Max. :115737.5 NA's :14	Investments Min. : 0.0 1st Qu.: 0.9 Median : 7.8 Mean : 853.2 3rd Qu.: 61.6 Max. :88047.8 NA's :280		
Current.assets Min. : 0.1 1st Qu.: 38.9 Median : 165.6 Mean : 1632.9 3rd Qu.: 578.0 Max. :196614.6 NA's :14	Net.working.capital Min. :-41908.3 1st Qu.: -1.3 Median : 20.1 Mean : 283.0 3rd Qu.: 99.2 Max. : 85782.8 NA's :5	Quick.ratio..times. Min. : 0.000 1st Qu.: 0.410 Median : 0.660 Mean : 1.968 3rd Qu.: 1.020 Max. :341.000 NA's :12	Current.ratio..times. Min. : 0.000 1st Qu.: 0.920 Median : 1.230 Mean : 2.880 3rd Qu.: 1.725 Max. :505.000 NA's :12	Debt.to.equity.ratio..times. Min. : 0.000 1st Qu.: 0.220 Median : 0.800 Mean : 3.327 3rd Qu.: 1.700 Max. :341.180	Cash.to.current.liabilities..times. Min. : 0.0000 1st Qu.: 0.0300 Median : 0.0800 Mean : 0.7149 3rd Qu.: 0.1900 Max. :165.0000 NA's :12	
Cash.to.average.cost.of.sales.per.day Min. : 0.000 1st Qu.: 3.248 Median : 8.130 Mean : 79.565 3rd Qu.: 22.645 Max. :15999.170 NA's :15	Creditors.turnover Min. : 0.00 1st Qu.: 3.84 Median : 6.49 Mean : 23.48 3rd Qu.: 12.90 Max. :1934.00 NA's :58	Debtors.turnover Min. : 0.000 1st Qu.: 4.133 Median : 7.050 Mean : 22.264 3rd Qu.: 12.920 Max. :2473.040 NA's :57	Finished.goods.turnover Min. : -0.09 1st Qu.: 8.06 Median : 17.49 Mean : 71.31 3rd Qu.: 38.67 Max. :5614.80 NA's :134	WIP.turnover Min. : 0.000 1st Qu.: 5.135 Median : 10.710 Mean : 32.386 3rd Qu.: 20.130 Max. :5651.400 NA's :124	Raw.material.turnover Min. : 0.000 1st Qu.: 3.190 Median : 6.445 Mean : 11.087 3rd Qu.: 11.650 Max. :279.960 NA's :67	
Shares.outstanding Min. :1.280e+02 1st Qu.:1.262e+06 Median :4.940e+06 Mean :3.187e+07 3rd Qu.:1.282e+07 Max. :4.130e+09 NA's :118	Equity.face.value Min. : 1.0 1st Qu.: 10.0 Median : 10.0 Mean : 45.1 3rd Qu.: 10.0 Max. :10000.0 NA's :118	EPS Min. :-72750.00 1st Qu.: 0.00 Median : 1.83 Mean : -76.87 3rd Qu.: 11.46 Max. : 8784.00	Adjusted.EPS Min. :-72750.00 1st Qu.: 0.00 Median : 1.50 Mean : -78.74 3rd Qu.: 8.35 Max. : 8784.00	Total.liabilities Min. : 0.1 1st Qu.: 93.2 Median : 347.7 Mean : 4218.6 3rd Qu.: 1315.3 Max. :354727.3	PE.on.BSE Min. :-263.920 1st Qu.: 1.863 Median : 7.670 Mean : 15.132 3rd Qu.: 14.665 Max. :1478.420 NA's :433	

Summarization of the variables are measured for the 1st quartile, 3rd Quartile, Minimum value, maximum value, median, mean for each variables and the missing values are identified with the total counts of each variables. The minimum of some variables are measured with negative values and the maximum of some variables are measured with higher values are taken as outlier identification.

The values are measured with the observations and 52 variables. The Default variable is used as predictor for the validation dataset as well as in the train datasets also.

1.1 Outlier Treatment

```
> library(scales)
```

Library Scales is loaded to treat the each variables of the datasets.

```
> for(i in 3:ncol(my_train)){
+ q=quantile(my_train[,i],c(0.1,0.99))
+ my_train[,i]=squish(my_train[,i],q)
+ }
```

The quantile function is used to treat the outlier of variables form 1% to 99% of the values in the range of the variables found.

```
> summary(my_train)
```

Num	Networth.Next.Year	Total.assets	Net.worth	Total.income	Change.in.stock	Total.expenses	Profit.after.tax
Min. : 1 1st Qu.: 886 Median :1773 Mean :1772 3rd Qu.:2658 Max. :3545	Min. :-74265.6 1st Qu.: 31.7 Median : 116.3 Mean : 1616.3 3rd Qu.: 456.1 Max. :805773.4	Min. : 25.5 1st Qu.: 91.3 Median : 309.7 Mean : 2052.4 3rd Qu.: 1098.7 Max. :51658.8	Min. : 8.4 1st Qu.: 31.3 Median : 102.3 Mean : 754.5 3rd Qu.: 377.3 Max. :20920.8	Min. : 21.4 1st Qu.: 121.2 Median : 444.9 Mean : 2042.0 3rd Qu.: 1340.3 Max. :42282.8	Min. :-12.9 1st Qu.: -0.7 Median : 1.6 Mean : 25.3 3rd Qu.: 13.4 Max. :600.2	Min. : 16.6 1st Qu.: 104.1 Median : 407.7 Mean : 1918.8 3rd Qu.: 1284.6 Max. :38283.8	Min. : -2.7 1st Qu.: 0.6 Median : 8.8 Mean : 130.3 3rd Qu.: 48.1 Max. :4069.9
PBDITA	PBT	Cash.profit	PBDITA.as...of.total.income	PBT.as...of.total.income	PAT.as...of.total.income		
Min. : 0.6 1st Qu.: 7.3 Median : 35.4 Mean : 285.9 3rd Qu.: 139.1 Max. :7370.1	Min. : -2.7 1st Qu.: 0.9 Median : 12.4 Mean : 176.4 3rd Qu.: 67.5 Max. :5421.4	Min. : -0.10 1st Qu.: 3.10 Median : 18.85 Mean : 189.22 3rd Qu.: 86.80 Max. :5453.78	Min. : 0.38 1st Qu.: 5.07 Median : 9.66 Mean :12.63 3rd Qu.:16.15 Max. :80.81	Min. :-3.44 1st Qu.: 0.60 Median : 3.31 Mean : 5.97 3rd Qu.: 8.63 Max. :51.66	Min. :-3.180 1st Qu.: 0.390 Median : 2.340 Mean : 4.441 3rd Qu.: 6.250 Max. :43.550		
Cash.profit.as...of.total.income	PAT.as...of.net.worth	Sales	Income.from.financial.services	Other.income	Total.capital		
Min. : 0.000 1st Qu.: 2.090 Median : 5.640 Mean : 7.972 3rd Qu.:10.560 Max. :56.244	Min. :-4.94 1st Qu.: 0.00 Median : 7.92 Mean :13.32 3rd Qu.:20.19 Max. :97.36	Min. : 25.6 1st Qu.: 133.3 Median : 453.1 Mean : 1989.1 3rd Qu.: 1314.7 Max. :40605.1	Min. : 0.2 1st Qu.: 0.7 Median : 1.8 Mean : 23.4 3rd Qu.: 5.4 Max. :769.9	Min. : 0.200 1st Qu.: 0.800 Median : 1.400 Mean : 9.497 3rd Qu.: 2.500 Max. :286.160	Min. : 3.6 1st Qu.: 13.1 Median : 42.1 Mean : 150.0 3rd Qu.: 100.3 Max. :2936.7		
Reserves.and.funds	Borrowings	Current.liabilities...provisions	Deferred.tax.liability	Shareholders.funds	Cumulative.retained.profits		
Min. : -10.7 1st Qu.: 5.8 Median : 54.8 Mean : 614.2 3rd Qu.: 263.2 Max. :17416.7	Min. : 5.2 1st Qu.: 29.7 Median : 99.2 Mean : 601.0 3rd Qu.: 296.0 Max. :14803.4	Min. : 3.9 1st Qu.: 18.7 Median : 69.4 Mean : 473.4 3rd Qu.: 249.1 Max. :11092.0	Min. : 1.40 1st Qu.: 6.70 Median : 13.40 Mean : 76.59 3rd Qu.: 26.90 Max. :2226.60	Min. : 8.6 1st Qu.: 32.0 Median : 105.6 Mean : 770.0 3rd Qu.: 393.2 Max. :20920.8	Min. : -20.5 1st Qu.: 1.3 Median : 37.1 Mean : 456.4 3rd Qu.: 199.4 Max. :13027.3		
Capital.employed	TOL.TNW	Total.term.liabilities...tangible.net.worth	Contingent.liabilities...Net.worth....	Contingent.liabilities			
Min. : 17.2 1st Qu.: 60.8 Median : 214.7 Mean : 1402.5 3rd Qu.: 767.3 Max. :34914.6	Min. : 0.150 1st Qu.: 0.600 Median : 1.430 Mean : 3.223 3rd Qu.: 2.830 Max. :55.958	Min. : 0.000 1st Qu.: 0.050 Median : 0.340 Mean : 1.243 3rd Qu.: 1.000 Max. :29.464	Min. : 0.00 1st Qu.: 0.00 Median : 5.33 Mean : 36.06 3rd Qu.: 30.76 Max. :773.79	Min. : 2.4 1st Qu.: 16.0 Median : 38.0 Mean : 247.5 3rd Qu.: 84.2 Max. :6177.7			
Net.fixed.assets	Investments	Current.assets	Net.working.capital	Quick.ratio..times.	Current.ratio..times.	Debt.to.equity.ratio..times.	
Min. : 7.6 1st Qu.: 27.3 Median : 93.5 Mean : 648.9 3rd Qu.: 328.8 Max. :16862.2	Min. : 0.40 1st Qu.: 4.60 Median : 8.35 Mean : 147.13 3rd Qu.: 16.10 Max. :5024.04	Min. : 6.6 1st Qu.: 37.2 Median : 145.1 Mean : 779.1 3rd Qu.: 485.9 Max. :17377.1	Min. : -53.2 1st Qu.: -1.0 Median : 16.2 Mean : 144.7 3rd Qu.: 81.6 Max. :3688.2	Min. : 0.190 1st Qu.: 0.420 Median : 0.670 Mean : 1.064 3rd Qu.: 1.020 Max. :14.000	Min. : 0.620 1st Qu.: 0.940 Median : 1.230 Mean : 1.774 3rd Qu.: 1.690 Max. :18.920	Min. : 0.000 1st Qu.: 0.220 Median : 0.790 Mean : 1.954 3rd Qu.: 1.750 Max. :37.104	
Cash.to.current.liabilities..times.	Cash.to.average.cost.of.sales.per.day	Creditors.turnover	Debtors.turnover	Finished.goods.turnover			
Min. :0.0100 1st Qu.:0.0200 Median :0.0700 Mean :0.2891 3rd Qu.:0.1900 Max. :5.7780	Min. : 0.850 1st Qu.: 2.890 Median : 8.025 Mean : 47.483 3rd Qu.: 21.150 Max. :1277.500	Min. : 1.740 1st Qu.: 3.940 Median : 6.095 Mean : 11.736 3rd Qu.: 10.550 Max. :131.374	Min. : 1.69 1st Qu.: 4.01 Median : 6.32 Mean : 12.83 3rd Qu.: 10.82 Max. :182.90	Min. : 5.04 1st Qu.: 10.21 Median : 17.27 Mean : 46.58 3rd Qu.: 30.72 Max. :857.22			
WIP.turnover	Raw.material.turnover	Shares.outstanding	Equity.face.value	EPS	Adjusted.EPS	Total.liabilities	PE.on.BSE
Min. : 3.00 1st Qu.: 5.93 Median : 9.76 Mean : 18.76 3rd Qu.: 16.94 Max. :218.88	Min. : 0.00 1st Qu.: 3.41 Median : 6.40 Mean :10.06 3rd Qu.:10.92 Max. :97.82	Min. : 350000 1st Qu.: 2209860 Median : 4672063 Mean : 14455333 3rd Qu.: 8320000 Max. :268509911	Min. : 10.00 1st Qu.: 10.00 Median : 10.00 Mean : 18.24 3rd Qu.: 10.00 Max. :190.00	Min. : -0.69 1st Qu.: 0.00 Median : 1.43 Mean : 27.08 3rd Qu.: 9.62 Max. :896.14	Min. : -0.66 1st Qu.: 0.00 Median : 1.18 Mean : 25.92 3rd Qu.: 7.48 Max. :896.14	Min. : 25.5 1st Qu.: 91.3 Median : 309.7 Mean : 2052.4 3rd Qu.: 1098.7 Max. :51658.8	Min. : 3.70 1st Qu.: 9.10 Median : 9.10 Mean : 12.65 3rd Qu.: 9.10 Max. :156.05

The summarization of each variables are identified in the datasets and the outliers are treated within the range of 1% and 99% as treated as minimum outlier and maximum outlier of the variables.

```
> for(i in 3:ncol(my_test)){
+   q=quantile(my_test[,i],c(0.1,0.99))
+   my_test[,i]=squish(my_test[,i],q)
+ }
```

The variables of the test dataset also treated with outliers for values above 99% and 1% in test.

> summary(my_test)

Num Min. : 1.0 1st Qu.:179.5 Median :358.0 Mean :358.0 3rd Qu.:536.5 Max. :715.0	Default...1 Min. :0.00000 1st Qu.:0.00000 Median :0.00000 Mean :0.07552 3rd Qu.:0.00000 Max. :1.00000	Total.assets Min. : 32.84 1st Qu.: 93.20 Median : 347.70 Mean : 2550.96 3rd Qu.: 1315.35 Max. :57604.77	Net.worth Min. : 8.5 1st Qu.: 34.4 Median : 120.9 Mean : 1007.3 3rd Qu.: 451.5 Max. :27995.3	Total.income Min. : 22.6 1st Qu.: 122.0 Median : 536.0 Mean : 2544.5 3rd Qu.: 1633.4 Max. :53816.7	Change.in.stock Min. :-16.68 1st Qu.: -0.75 Median : 1.80 Mean : 31.53 3rd Qu.: 15.20 Max. :845.68	Total.expenses Min. : 16.88 1st Qu.: 113.40 Median : 511.10 Mean : 2330.31 3rd Qu.: 1578.60 Max. :46388.99	Profit.after.tax Min. : -3.26 1st Qu.: 0.80 Median : 10.20 Mean : 197.15 3rd Qu.: 64.00 Max. :6111.71
PBDITA Min. : 0.34 1st Qu.: 7.30 Median : 42.20 Mean : 387.07 3rd Qu.: 181.00 Max. :9815.61	PBT Min. : -3.18 1st Qu.: 1.15 Median : 14.25 Mean : 248.89 3rd Qu.: 83.10 Max. :6679.83	Cash.profit Min. : -0.46 1st Qu.: 3.45 Median : 22.05 Mean : 255.86 3rd Qu.: 114.10 Max. :6728.99	PBDITA.as...of.total.income Min. : 0.132 1st Qu.: 4.805 Median : 9.780 Mean :12.689 3rd Qu.:16.715 Max. :73.445	PBT.as...of.total.income Min. :-3.942 1st Qu.: 0.660 Median : 3.450 Mean : 6.254 3rd Qu.: 9.645 Max. :48.083	PAT.as...of.total.income Min. :-3.642 1st Qu.: 0.410 Median : 2.405 Mean : 4.710 3rd Qu.: 6.660 Max. :43.582		
Cash.profit.as...of.total.income Min. :-0.216 1st Qu.: 1.950 Median : 5.835 Mean : 8.001 3rd Qu.:10.800 Max. :54.217	PAT.as...of.net.worth Min. : -8.306 1st Qu.: 0.000 Median : 8.710 Mean : 14.031 3rd Qu.: 20.215 Max. :124.276	Sales Min. : 30.16 1st Qu.: 138.90 Median : 552.50 Mean : 2485.38 3rd Qu.: 1613.70 Max. :53045.61	Income.from.financial.services Min. : 0.20 1st Qu.: 0.85 Median : 2.00 Mean : 28.83 3rd Qu.: 6.05 Max. :822.10	Other.income Min. : 0.20 1st Qu.: 0.90 Median : 1.65 Mean : 13.63 3rd Qu.: 2.90 Max. :368.64	Total.capital Min. : 2.74 1st Qu.: 14.10 Median : 45.30 Mean : 184.33 3rd Qu.: 120.85 Max. :3386.43		
Reserves.and.funds Min. : -16.34 1st Qu.: 7.85 Median : 57.45 Mean : 849.52 3rd Qu.: 315.65 Max. :24622.83	Borrowings Min. : 5.2 1st Qu.: 32.1 Median : 105.5 Mean : 914.7 3rd Qu.: 344.4 Max. :30763.3	Current.liabilities...provisions Min. : 5.14 1st Qu.: 17.40 Median : 75.20 Mean : 638.32 3rd Qu.: 291.15 Max. :15572.95	Deferred.tax.liability Min. : 1.50 1st Qu.: 6.80 Median : 14.70 Mean : 98.73 3rd Qu.: 29.90 Max. :3345.24	Shareholders.funds Min. : 8.68 1st Qu.: 35.50 Median : 124.00 Mean : 1023.57 3rd Qu.: 478.35 Max. :27996.67	Cumulative.retained.profits Min. : -27.72 1st Qu.: 1.00 Median : 40.60 Mean : 684.01 3rd Qu.: 241.20 Max. :22751.99		
Capital.employed Min. : 18.72 1st Qu.: 64.35 Median : 246.10 Mean : 1839.62 3rd Qu.: 913.65 Max. :46687.12	TOL.TNW Min. : 0.194 1st Qu.: 0.595 Median : 1.400 Mean : 3.420 3rd Qu.: 2.800 Max. :54.946	Total.term.liabilities...tangible.net.worth Min. : 0.000 1st Qu.: 0.060 Median : 0.350 Mean : 1.414 3rd Qu.: 1.005 Max. :30.720	Contingent.liabilities...Net.worth.... Min. : 0.00 1st Qu.: 0.00 Median : 5.52 Mean : 54.25 3rd Qu.: 31.49 Max. :1550.81	Contingent.liabilities Min. : 1.88 1st Qu.: 12.25 Median : 37.50 Mean : 441.20 3rd Qu.: 112.25 Max. :17684.28			
Net.fixed.assets Min. : 7.60 1st Qu.: 27.95 Median : 95.00 Mean : 849.98 3rd Qu.: 403.10 Max. :24061.44	Investments Min. : 0.40 1st Qu.: 4.25 Median : 7.80 Mean : 225.86 3rd Qu.: 15.20 Max. :7463.18	Current.assets Min. : 6.94 1st Qu.: 40.50 Median : 165.60 Mean : 965.46 3rd Qu.: 561.15 Max. :21451.81	Net.working.capital Min. : -53.12 1st Qu.: -1.20 Median : 20.10 Mean : 203.32 3rd Qu.: 97.75 Max. :5588.83	Quick.ratio..times. Min. : 0.170 1st Qu.: 0.410 Median : 0.660 Mean : 1.087 3rd Qu.: 1.020 Max. :16.095	Current.ratio..times. Min. : 0.560 1st Qu.: 0.920 Median : 1.230 Mean : 1.848 3rd Qu.: 1.720 Max. :24.482	Debt.to.equity.ratio..times. Min. : 0.010 1st Qu.: 0.220 Median : 0.800 Mean : 2.179 3rd Qu.: 1.700 Max. :35.002	
Cash.to.current.liabilities..times. Min. :0.010 1st Qu.:0.030 Median :0.080 Mean :0.310 3rd Qu.:0.185 Max. :5.894	Cash.to.average.cost.of.sales.per.day Min. : 0.904 1st Qu.: 3.355 Median : 8.130 Mean : 39.924 3rd Qu.: 21.715 Max. :817.195	Creditors.turnover Min. : 2.04 1st Qu.: 4.03 Median : 6.49 Mean : 17.96 3rd Qu.: 12.15 Max. :402.21	Debtors.turnover Min. : 2.004 1st Qu.: 4.460 Median : 7.050 Mean : 14.548 3rd Qu.: 11.820 Max. :213.535	Finished.goods.turnover Min. : 4.974 1st Qu.: 9.970 Median : 17.490 Mean : 43.570 3rd Qu.: 31.385 Max. :757.770			
WIP.turnover Min. : 2.882 1st Qu.: 6.100 Median : 10.710 Mean : 18.970 3rd Qu.: 17.335 Max. :256.353	Raw.material.turnover Min. : 0.004 1st Qu.: 3.625 Median : 6.445 Mean : 9.784 3rd Qu.:10.810 Max. :83.862	Shares.outstanding Min. : 253668 1st Qu.: 2243890 Median : 4939709 Mean : 17992122 3rd Qu.: 10225000 Max. :356369400	Equity.face.value Min. : 10.0 1st Qu.: 10.0 Median : 10.0 Mean : 18.2 3rd Qu.: 10.0 Max. :100.0	EPS Min. : -0.988 1st Qu.: 0.000 Median : 1.830 Mean : 34.585 3rd Qu.: 11.460 Max. :1343.554	Adjusted.EPS Min. : -0.918 1st Qu.: 0.000 Median : 1.500 Mean : 32.942 3rd Qu.: 8.350 Max. :1343.554	Total.liabilities Min. : 32.84 1st Qu.: 93.20 Median : 347.70 Mean : 2550.96 3rd Qu.: 1315.35 Max. :57604.77	
PE.on.BSE Min. : 2.228 1st Qu.: 7.670 Median : 7.670 Mean : 9.886 3rd Qu.: 7.670 Max. :87.112							

The summarization of each variables are identified in the datasets and the outliers are treated within the range of 1% and 99% as treated as minimum outlier and maximum outlier of the variables.

1.2 Missing Value Treatment

```
> colSums(is.na(my_train))
```

Num 0	Networth.Next.Year 0	Total.assets 0
Net.worth 0	Total.income 198	Change.in.stock 458
Total.expenses 139	Profit.after.tax 131	PBDITA 131
PBT 131	Cash.profit 131	PBDITA.as...of.total.income 68
PBT.as...of.total.income 68	PAT.as...of.total.income 68	Cash.profit.as...of.total.income 68
PAT.as...of.net.worth 0	Sales 259	Income.from.financial.services 935
Other.income 1295	Total.capital 4	Reserves.and.funds 85
Deposits..accepted.by.commercial.banks. 3541	Borrowings 366	Current.liabilities...provisions 96
Deferred.tax.liability 1140	Shareholders.funds 0	Cumulative.retained.profits 38
Capital.employed 0	TOL.TNW 0	Total.term.liabilities...tangible.net.worth 0
Contingent.liabilities...Net.worth.... 0	Contingent.liabilities 1188	Net.fixed.assets 118
Investments 1435	Current.assets 66	Net.working.capital 32
Quick.ratio..times. 93	Current.ratio..times. 93	Debt.to.equity.ratio..times. 0
Cash.to.current.liabilities..times. 93	Cash.to.average.cost.of.sales.per.day 85	Creditors.turnover 333
Debtors.turnover 328	Finished.goods.turnover 740	WIP.turnover 640
Raw.material.turnover 361	Shares.outstanding 692	Equity.face.value 692
EPS 0	Adjusted.EPS 0	Total.liabilities 0
PE.on.BSE 2194		

The missing values are measured for each variables and the count of missing values is displayed for the treatment of the variables in the train datasets.

```
> my_train=subset(my_train,select = -c(22))
```

The variable Deposit accepted by commercial banks is full of missing variables, so the variable is dropped from the datasets.

```
> class(my_train)
```

[1] "tbl_df" "tbl" "data.frame"

The class of the variable is identify by the table format and data frame as the variables are assigned.

```
> my_train=as.data.frame(my_train)
```

The variables are converted to data frame of the dataset.

```
> for(i in 1:ncol(my_train)){
+   my_train[,i]=as.numeric(unlist(my_train[,i]))
+   my_train[is.na(my_train[,i]),i]=median(my_train[,i],na.rm = TRUE)
+ }
```

The variables are measured for the missing values in the data frame of the dataset and the variables are treated the missing values with the “median” of the each variables.

```
> any(is.na(my_train))
```

[1] FALSE

There is no missing values in the variable after the treatment of the missing values with median.

```
> colSums(is.na(my_test))
```

Num 0	Default...1 0	Total.assets 0
Net.worth 0	Total.income 33	Change.in.stock 92
Total.expenses 26	Profit.after.tax 23	PBDITA 23
PBT 23	Cash.profit 23	PBDITA.as...of.total.income 11
PBT.as...of.total.income 11	PAT.as...of.total.income 11	Cash.profit.as...of.total.income 11
PAT.as...of.net.worth 0	Sales 46	Income.from.financial.services 176
Other.income 261	Total.capital 1	Reserves.and.funds 13
Deposits..accepted.by.commercial.banks. 715	Borrowings 65	Current.liabilities...provisions 14
Deferred.tax.liability 229	Shareholders.funds 0	Cumulative.retained.profits 7
Capital.employed 0	TOL.TNW 0	Total.term.liabilities...tangible.net.worth 0
Contingent.liabilities...Net.worth.... 0	Contingent.liabilities 214	Net.fixed.assets 14
Investments 280	Current.assets 14	Net.working.capital 5
Quick.ratio..times. 12	Current.ratio..times. 12	Debt.to.equity.ratio..times. 0
Cash.to.current.liabilities..times. 12	Cash.to.average.cost.of.sales.per.day 15	Creditors.turnover 58
Debtors.turnover 57	Finished.goods.turnover 134	WIP.turnover 124
Raw.material.turnover 67	Shares.outstanding 118	Equity.face.value 118
EPS 0	Adjusted.EPS 0	Total.liabilities 0
PE.on.BSE 433		

Each variables has been identified by the number of the missing values in the datasets.

```
> my_test=subset(my_test,select = -c(22))
```

The test dataset is identified by the variable deposits by commercial bank is full of missing values and the variables is dropped down from the variables.

```
> class(my_test)
```

[1] "tbl_df" "tbl" "data.frame"

The class of the variable is identify by the table format and data frame as the variables are assigned.

```
> my_test=as.data.frame(my_test)
```

The test dataset is converted into data frame structure of overall datasets.

```
> for(i in 1:ncol(my_test)){
+   my_test[,i]=as.numeric(unlist(my_test[,i]))
+   my_test[is.na(my_test[,i]),i]=median(my_test[,i],na.rm = TRUE)
+ }
```

The missing values are treated with the median of the variables in the datasets.

```
> any(is.na(my_test))
```

[1] FALSE

There are no missing values in the datasets and this data frame can be used in the validation of the datasets.

1.3 New Variables Creation (One ration for profitability, leverage, liquidity and company's size each)

```
> my_train$Default...1=ifelse(my_train$Networth.Next.Year>0,0,1)
```

The default variable is created on the basis of the net worth next ear. If the next year net worth of the company is negative then the company will have defaulters and then the next year net worth of the company is positive then the company will have defaulters in the datasets.

```
> summary(my_train$Default...1)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00000	0.00000	0.00000	0.06862	0.00000	1.00000

The default variable is measured with maximum of 1 and minimum value is 0. The mean of the variable is measure with 0.0682 for the default variable.

```
> train_data=subset(my_train,select = -c(1,2))
```

The train dataset is dropped with the num column and the net worth next year, as the variables are not use in the further analysis.

```
> test_data=subset(my_test,select = -c(1))
```

The test dataset is dropped with the num column as the variable is not use in the further analysis.

```
> train_data=train_data[,c(50,1:49)]
```

Train Dataset is assigned by the last column defaulter to first column as it take and matches the column names in both test and train dataset.

```
> train_data$Default...1=as.factor(train_data$Default...1)
```

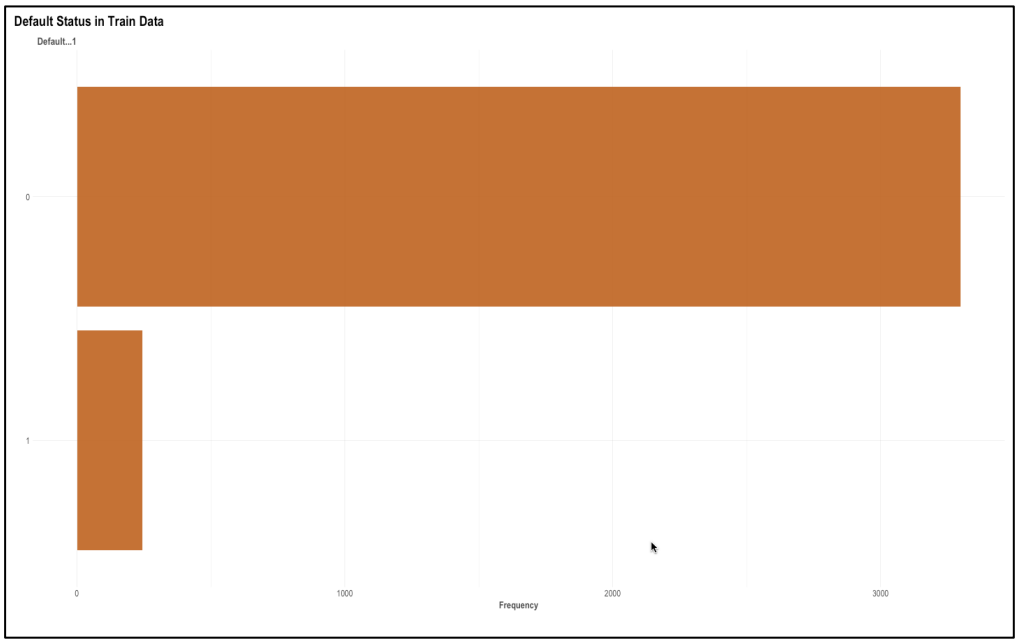
The default variables are treated and converted in to factor variables with two levels in the train dataset.

```
> test_data$Default...1=as.factor(test_data$Default...1)
```

The default variables are treated and converted in to factor variables with two levels in the test dataset.

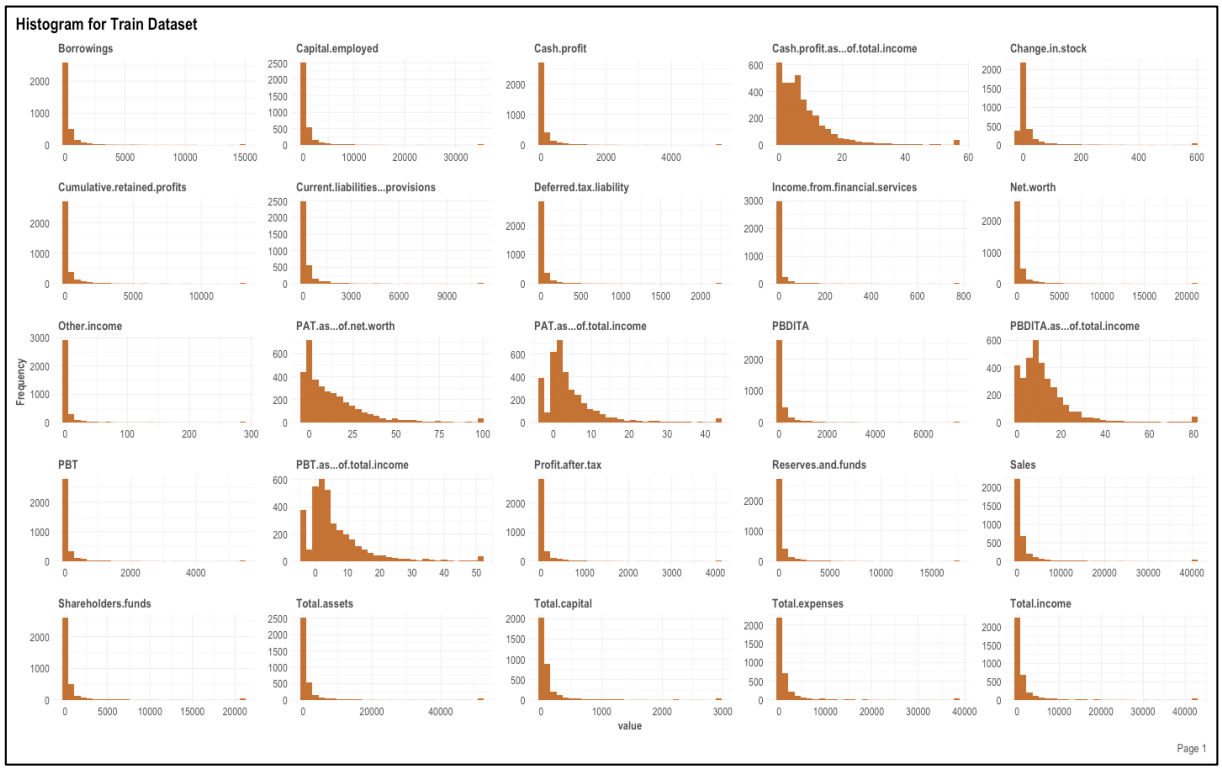
1.4 EDA Visualisation - Univariate & bivariate analysis

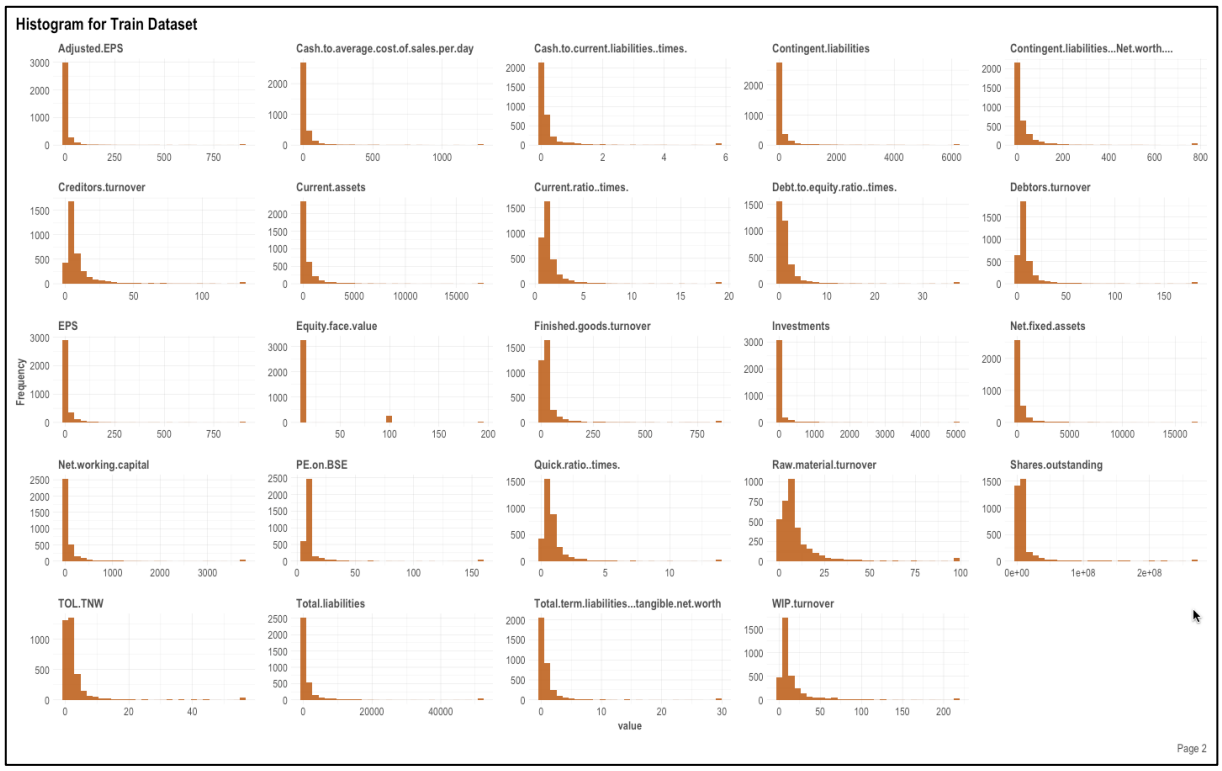
```
> plot_bar(data = train_data,ggtheme = theme_lares(),
+          title = "Default Status in Train Data")
```



Bar plot for the defaulter status is measured for the Train Dataset Default variable. The Defaulter are in range of the 243 and the non-defaulters are 3298.

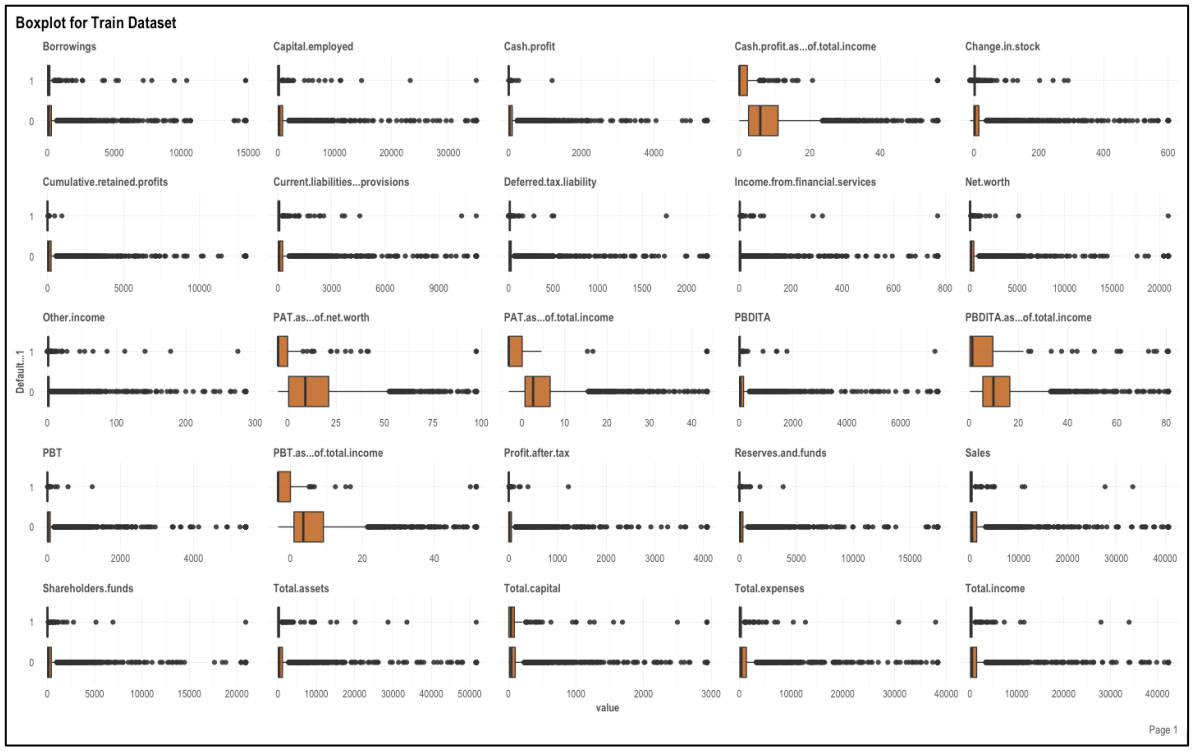
```
> plot_histogram(data = train_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
+                title = "Histogram for Train Dataset")
```

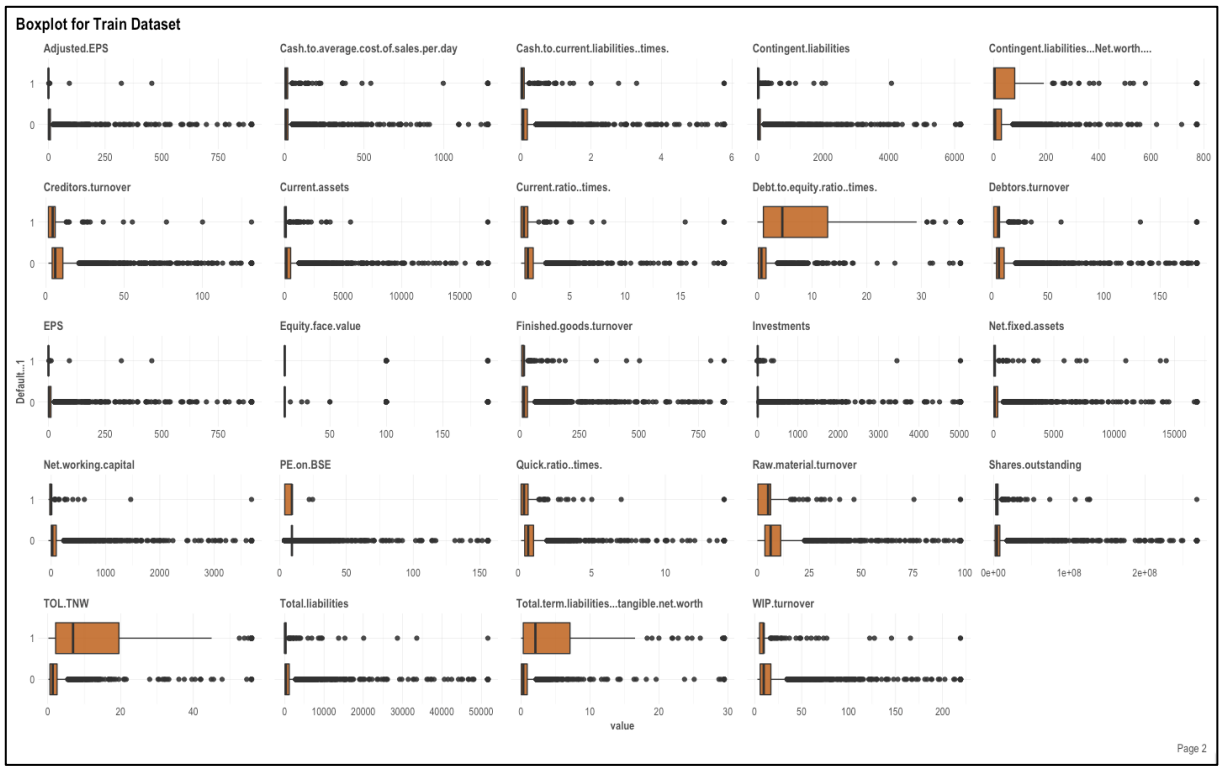




The histogram shows the variable values with the measured range. The variables are mostly common in increased width of the initial stage of the credit as it indicates that most of the variable values are within the range of the 25% to 50% and the above values are measured with the outliers of the variables.

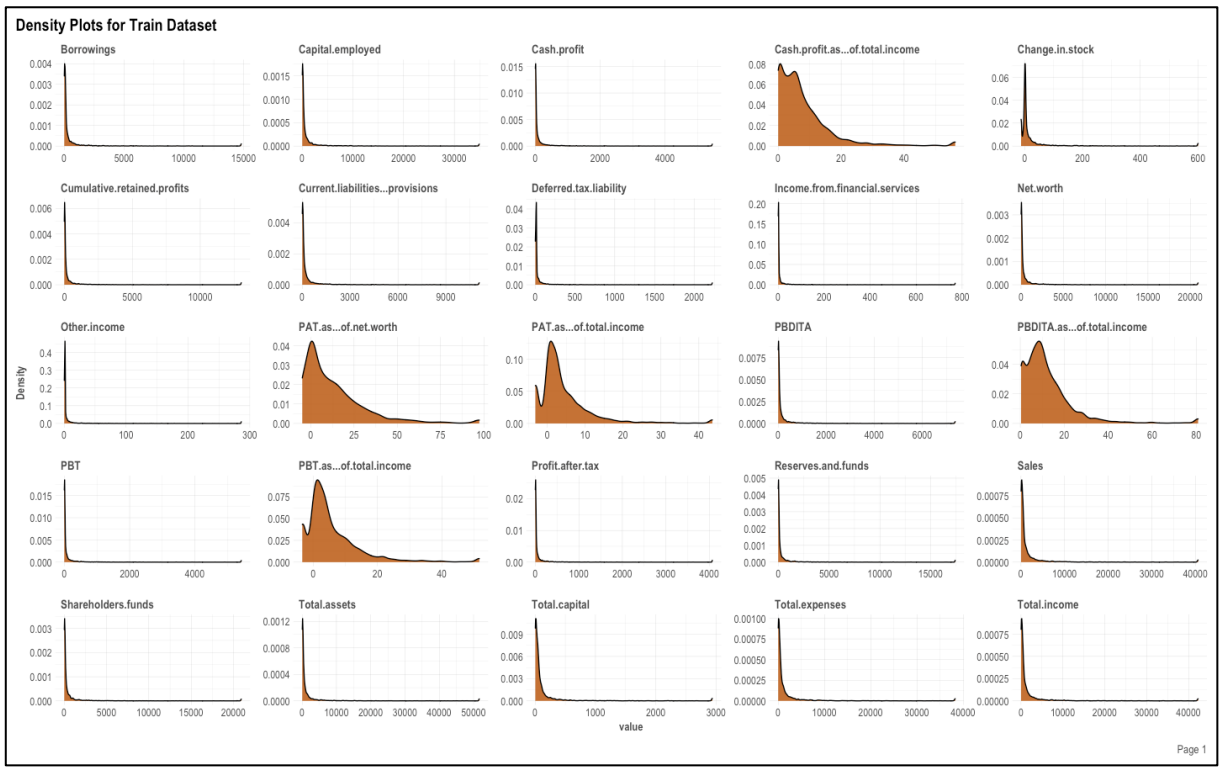
```
> plot_boxplot(data = train_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol = 5, title = "Boxplot for Train Dataset")
```

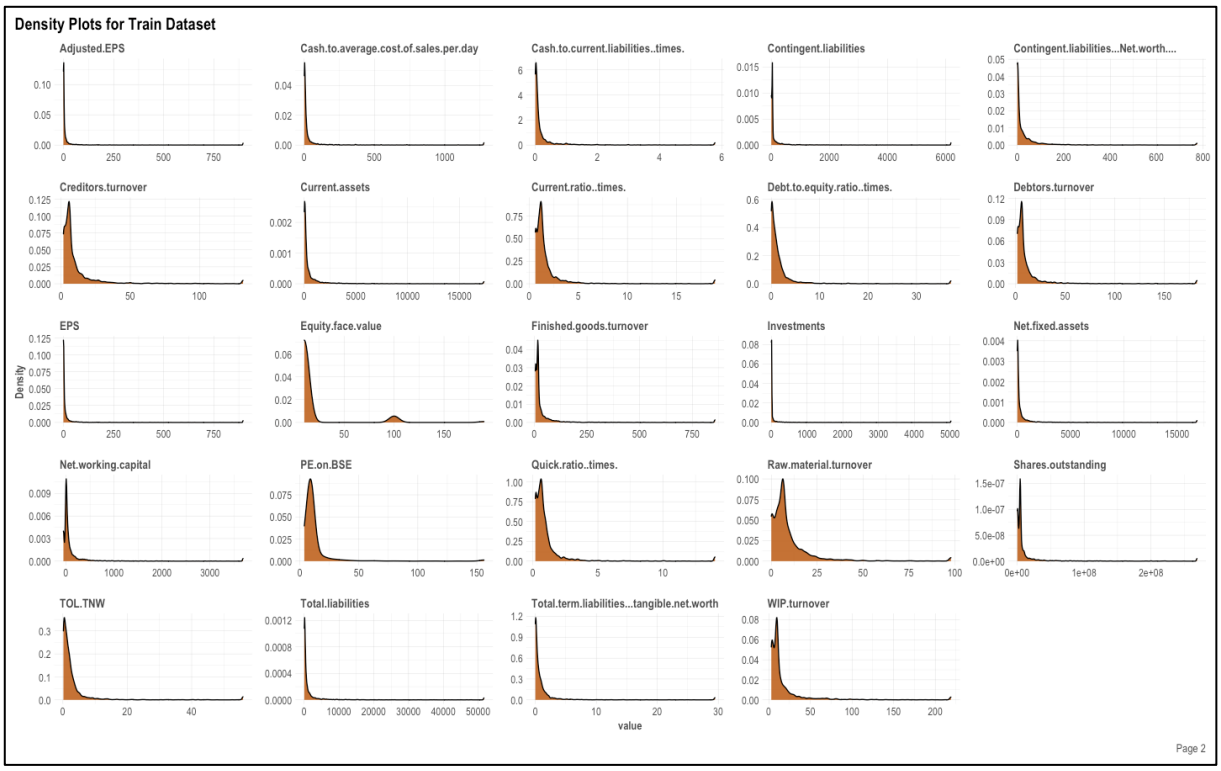




The boxplot shows the outlier in each variables and the outliers are identified in the ratio for each variables.

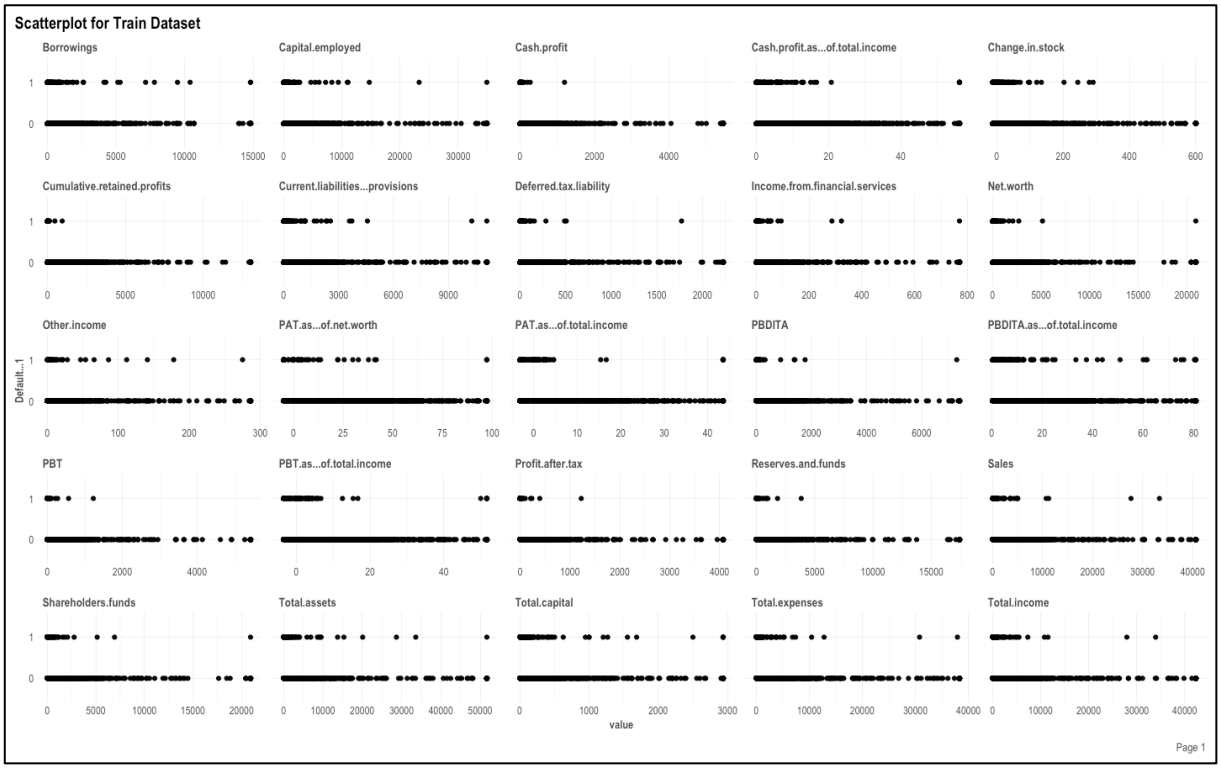
```
> plot_density(data = train_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
+              title = "Density Plots for Train Dataset")
```

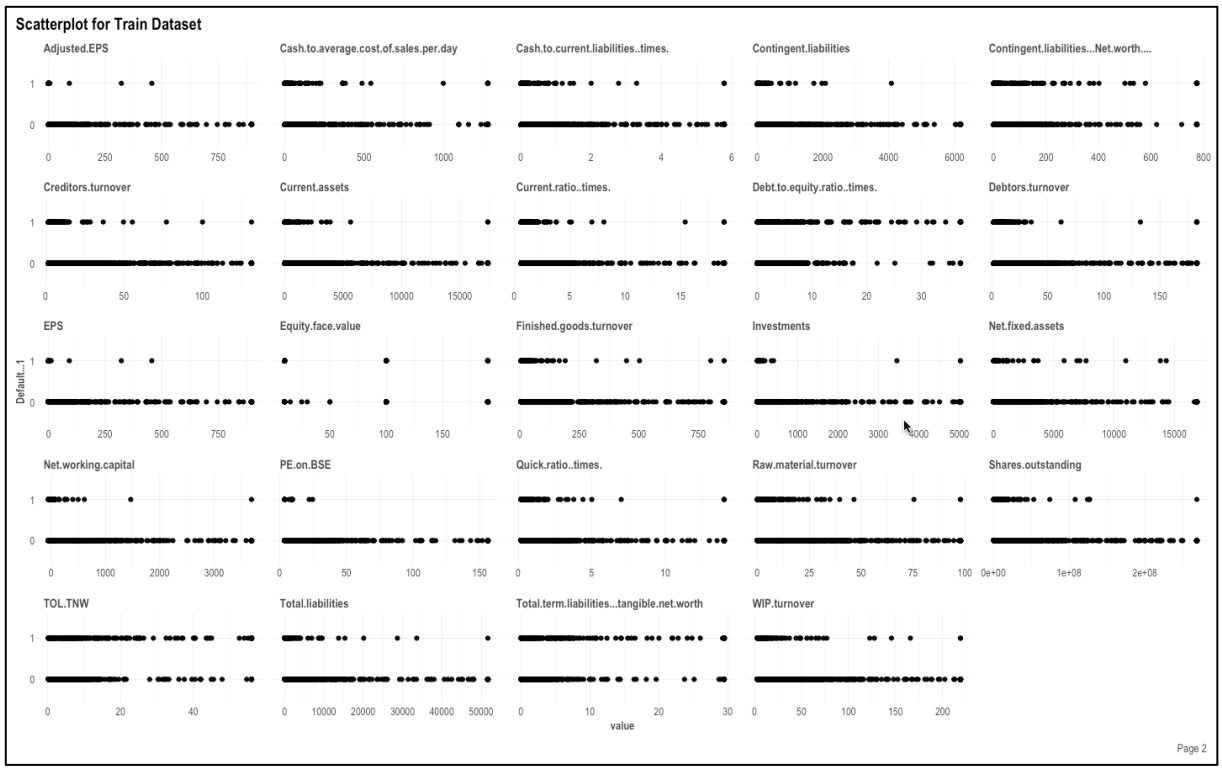




Density plot shows the area of the each variables in the where the areas are filled with initial part of the variables and it reflects the same of the histogram plots in the datasets.

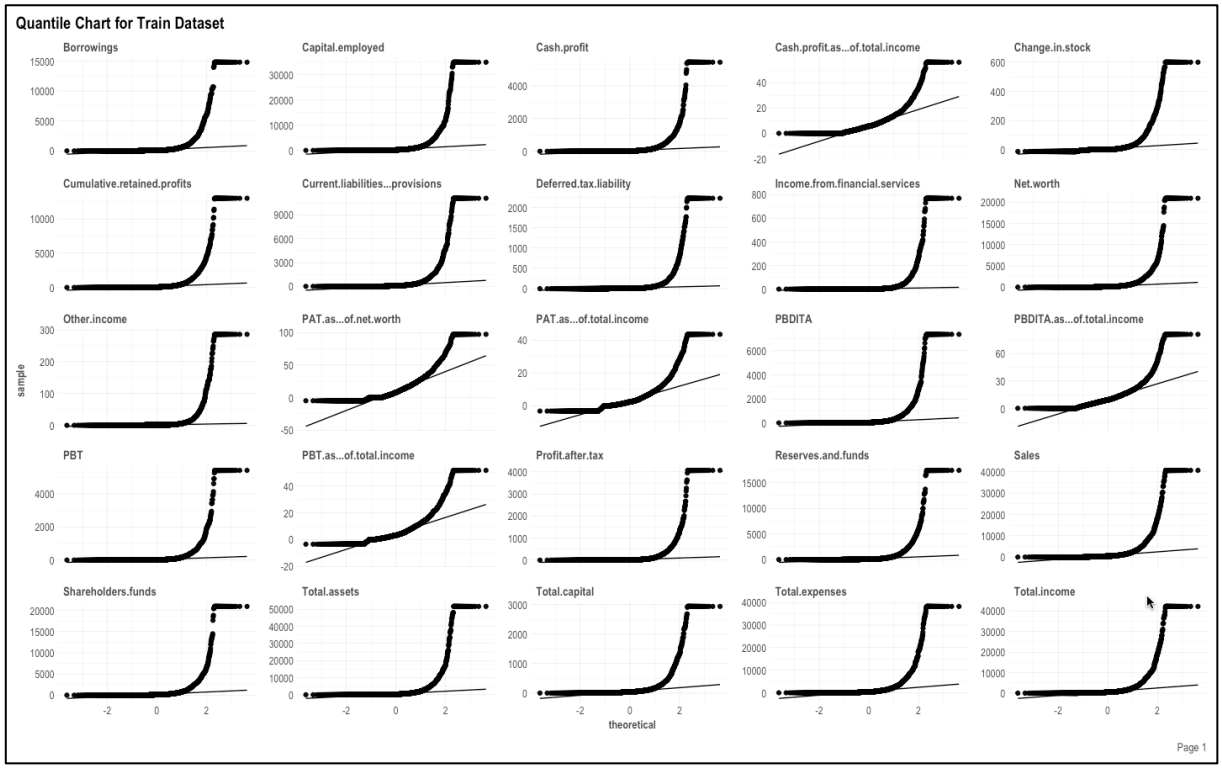
```
> plot_scatterplot(data = train_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol = 5,+                title = "Scatterplot for Train Dataset")
```

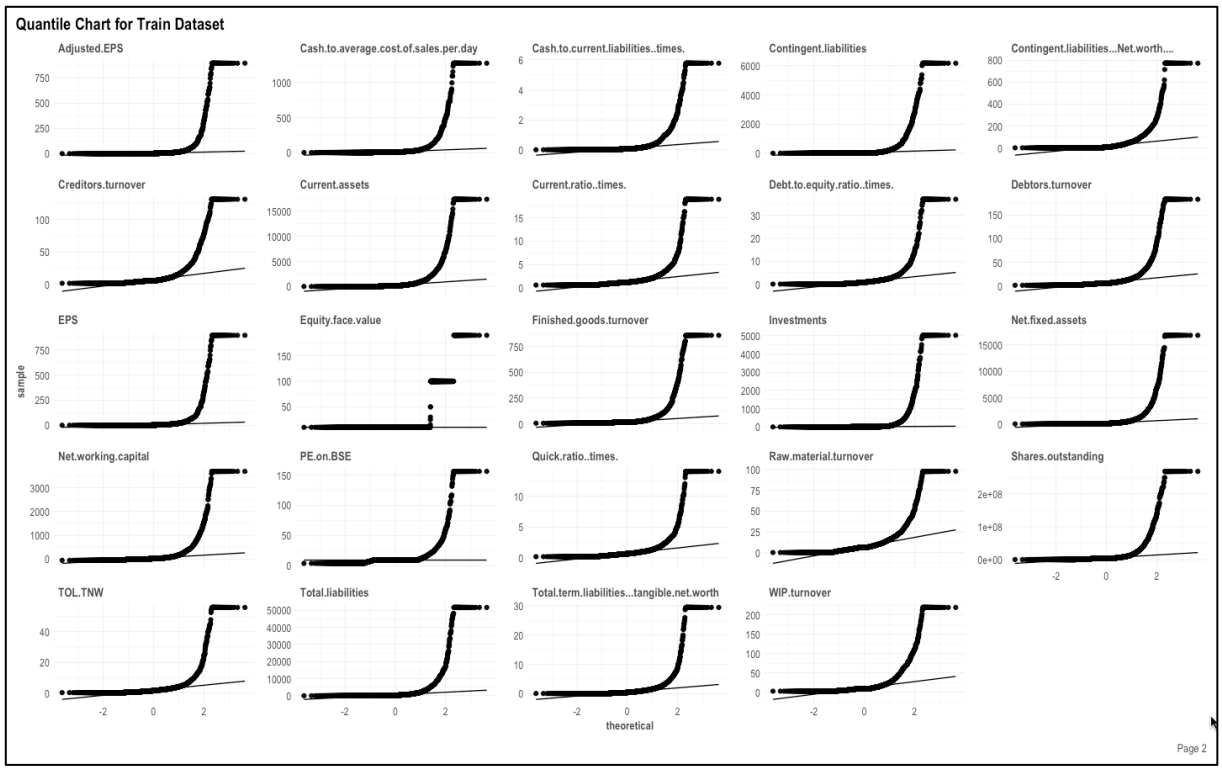




The scatter plot shows the scatter points in the defaulter variables of the train dataset. The scatter points reflects how the values are dispersed in the variables.

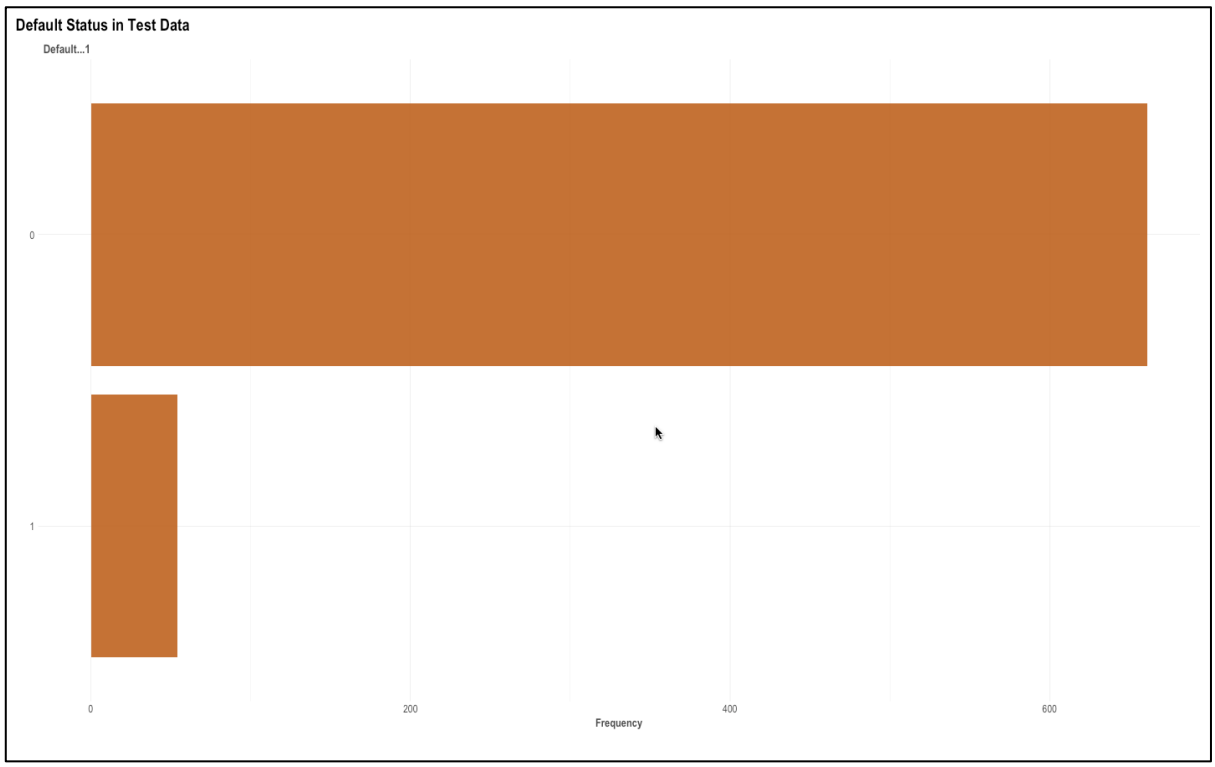
```
> plot_qq(data = train_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
+         title = "Quantile Chart for Train Dataset")
```





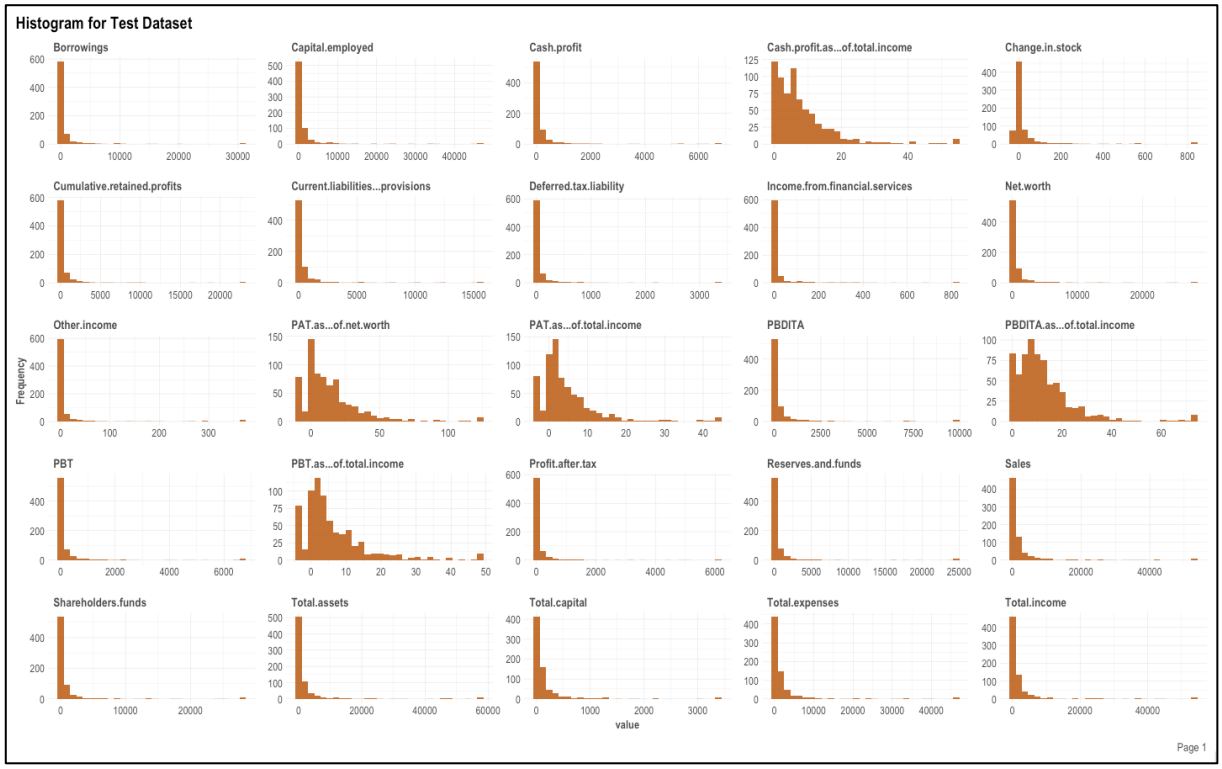
The plot shows the quantile and quantile area of the dataset is measured with the linear line in the dataset. The line above the quantile points are measured with increased curve of the points.

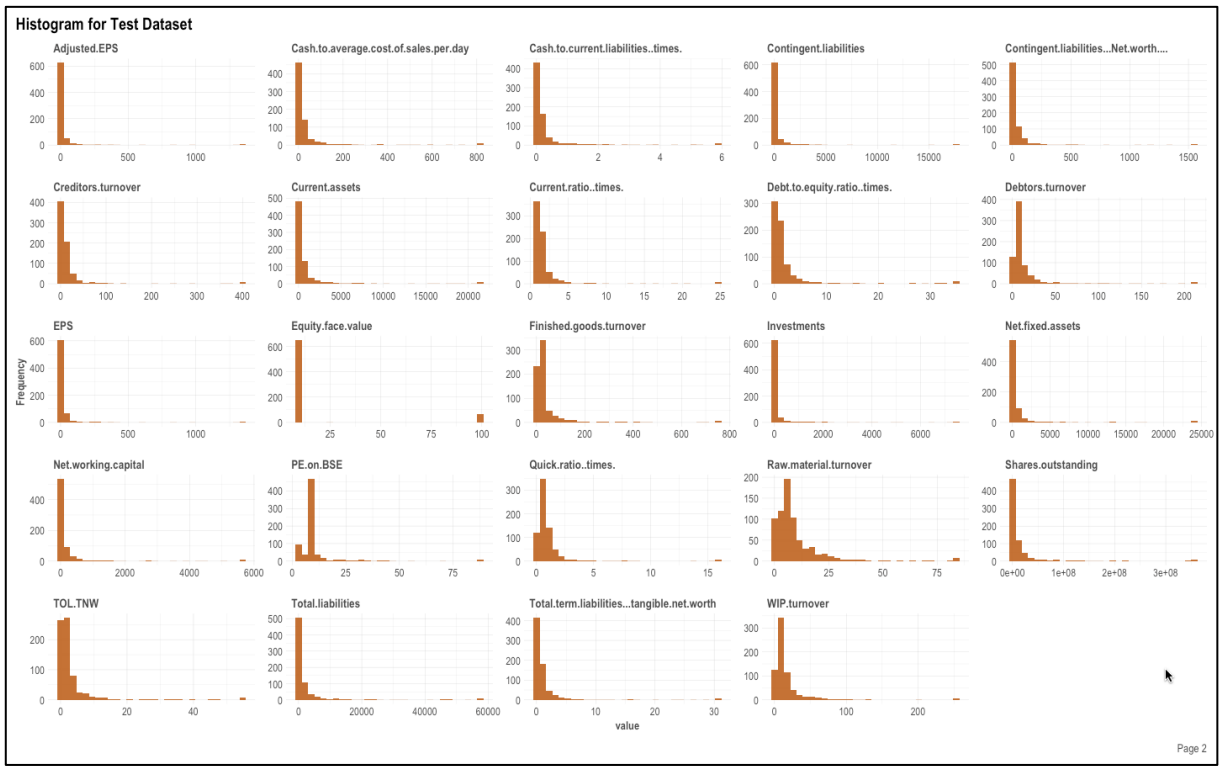
```
> plot_bar(data = test_data,ggtheme = theme_lares(),
+          title = "Default Status in Test Data")
```



The bar plot shows the defaulters of the variables in the test datasets. The defaulters are more than the non-defaulters in the dataset.

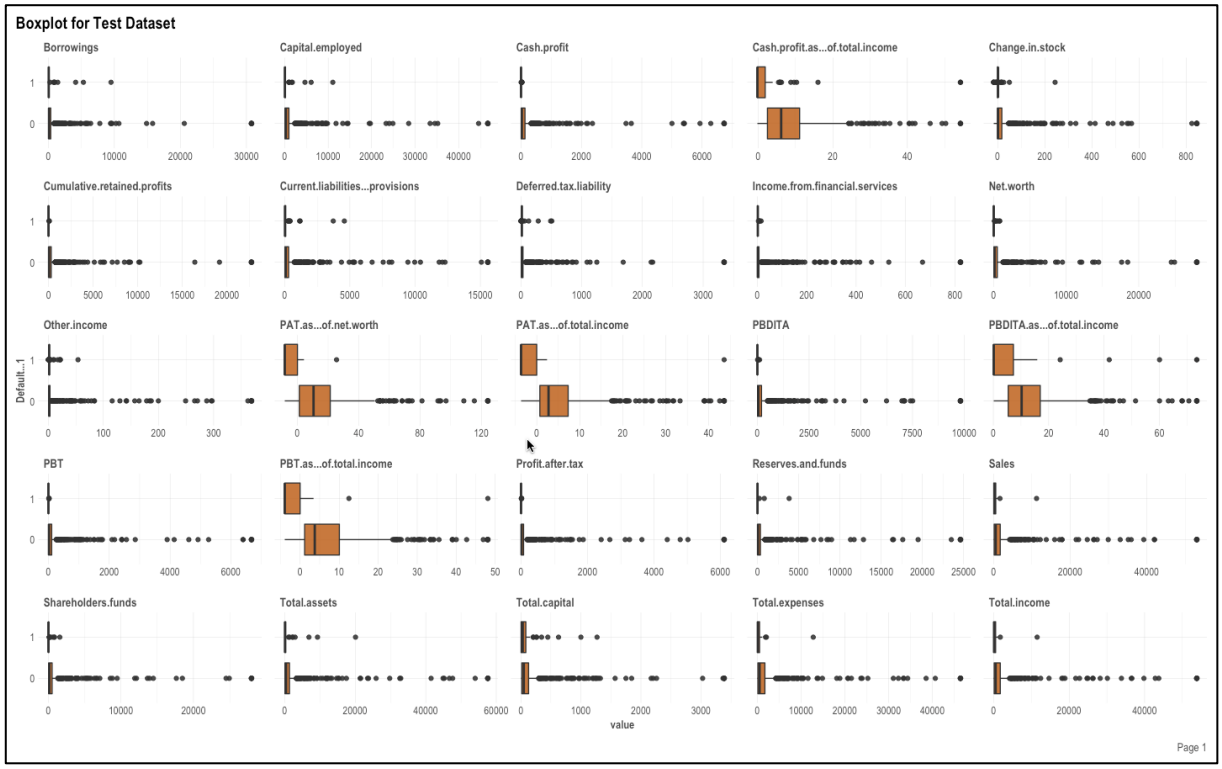
```
> plot_histogram(data = test_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
+                title = "Histogram for Test Dataset")
```

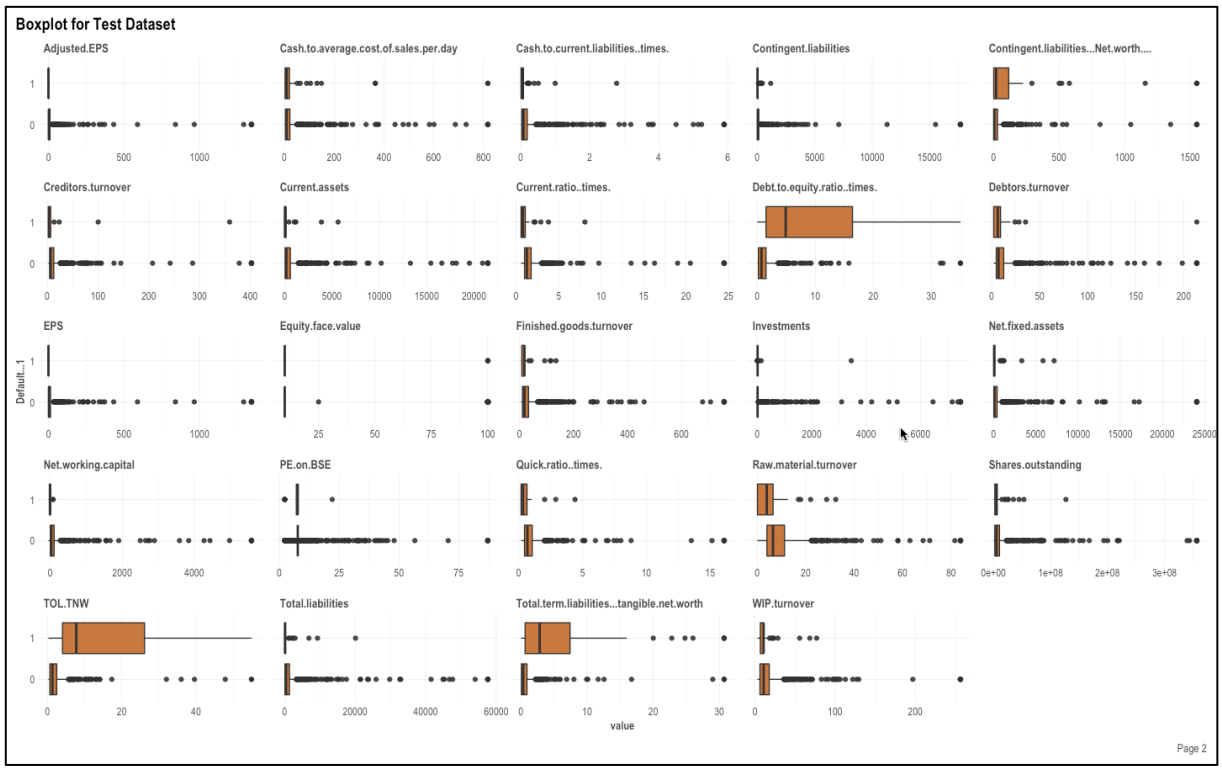




The histogram shows the variables for the identification of values in the datasets and the bar reflects the size of the variables. The plots explains it values are increased in the initial stage of the values and decreased at the values, it shows increased frequency decreased value.

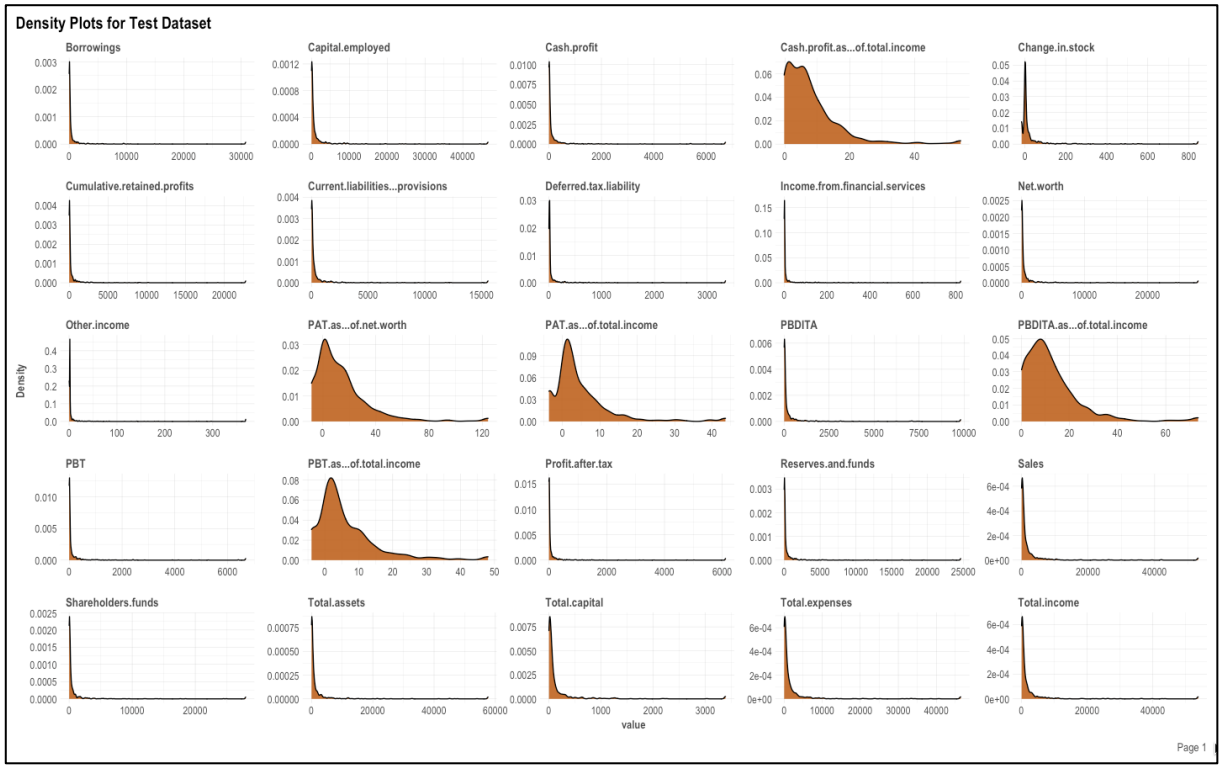
```
> plot_boxplot(data = test_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol = 5,
+              title = "Boxplot for Test Dataset")
```

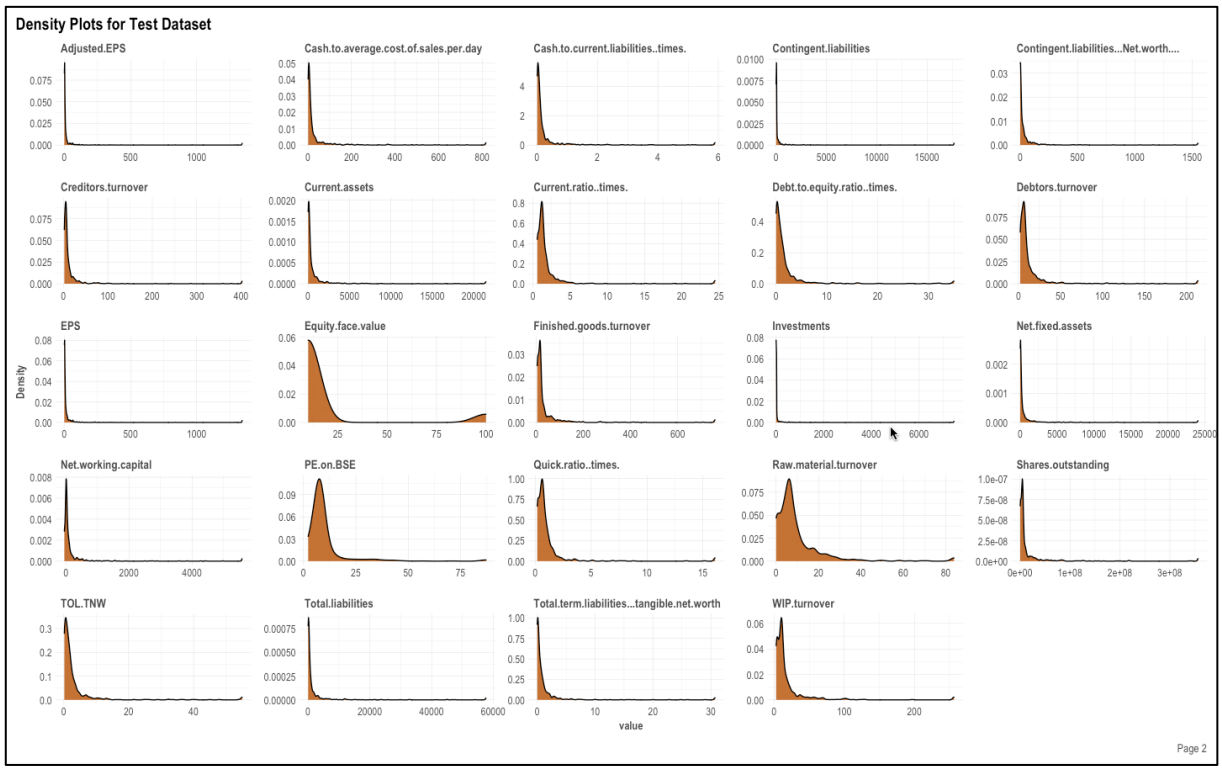




The boxplot provide the details of the outliers present in the variables. The values are maximum in some of the variables which reflects the company size, profitability, leverage, liquidity of the variables in the dataset as the values are representing the larger in size are maximum outliers.

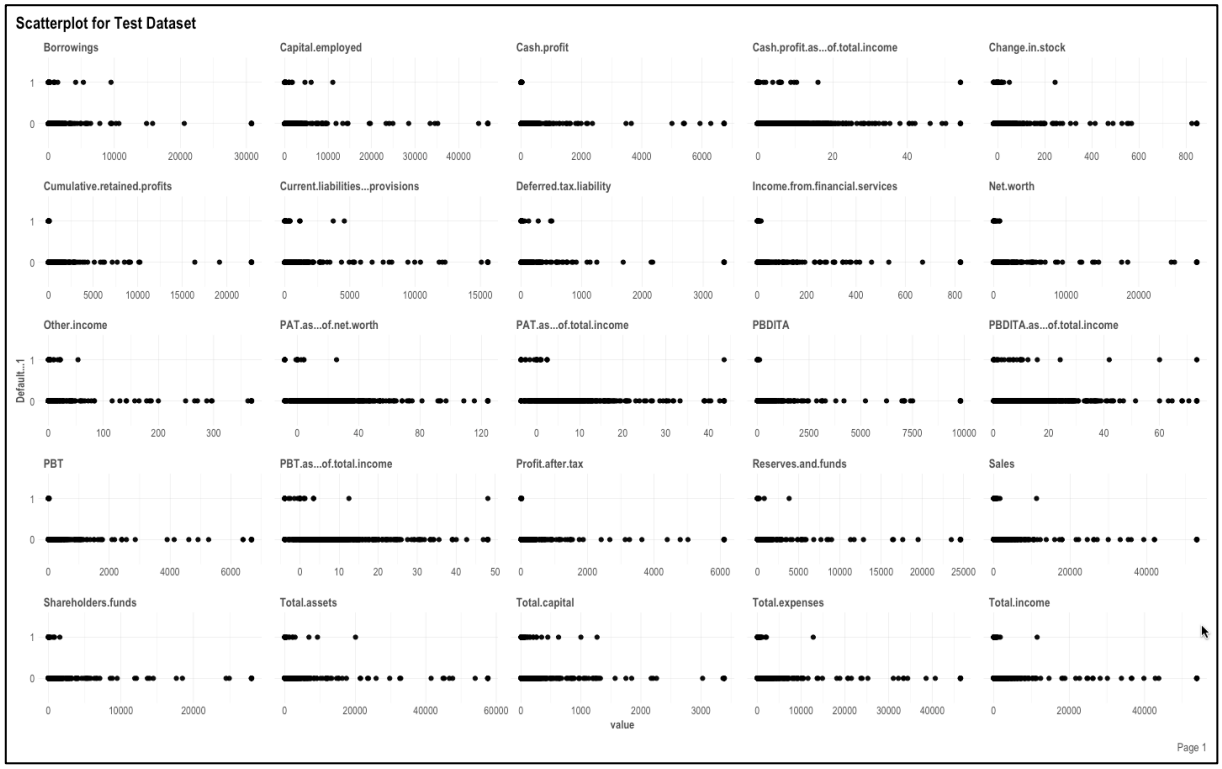
```
> plot_density(data = test_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
+              title = "Density Plots for Test Dataset")
```

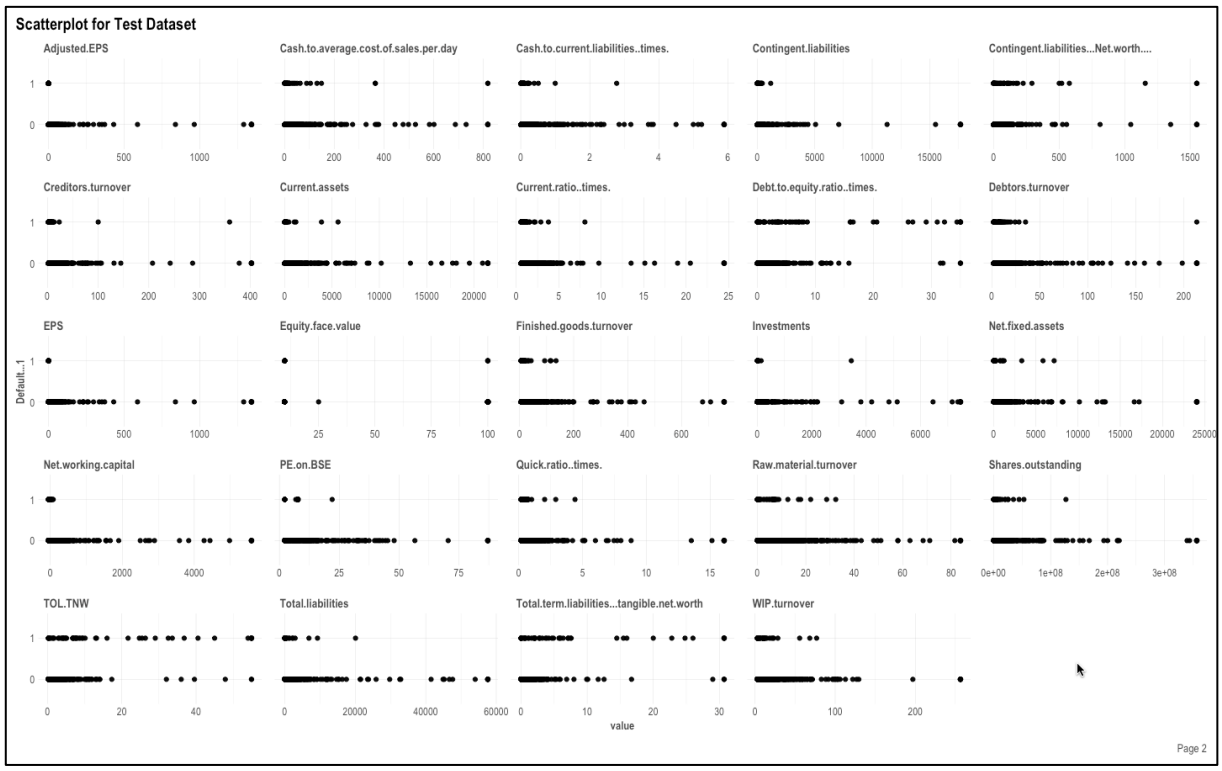




The density plot shows the area of the plots and the maximum are covered in the initial stage of the values. The frequency of are is increasing in the lower values and decreased frequency is changing in the higher values of each variables.

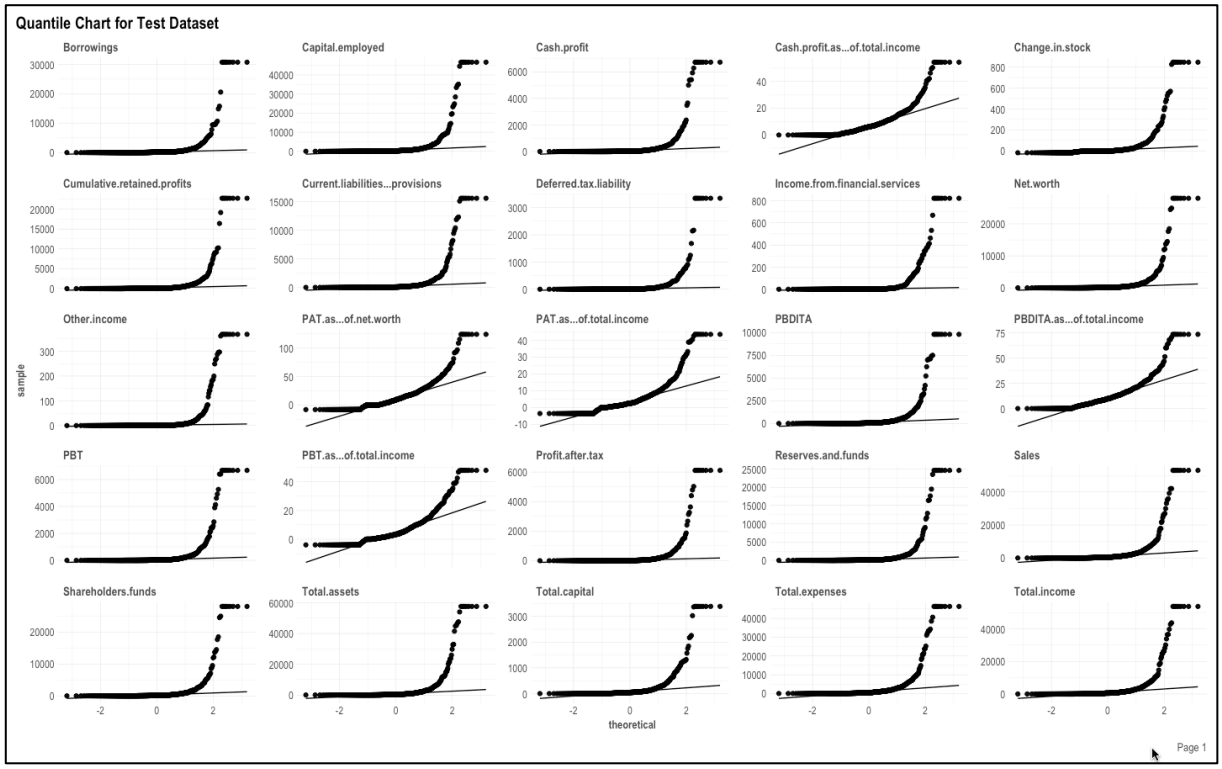
```
> plot_scatterplot(data = test_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol =
5,+                  title = "Scatterplot for Test Dataset")
```

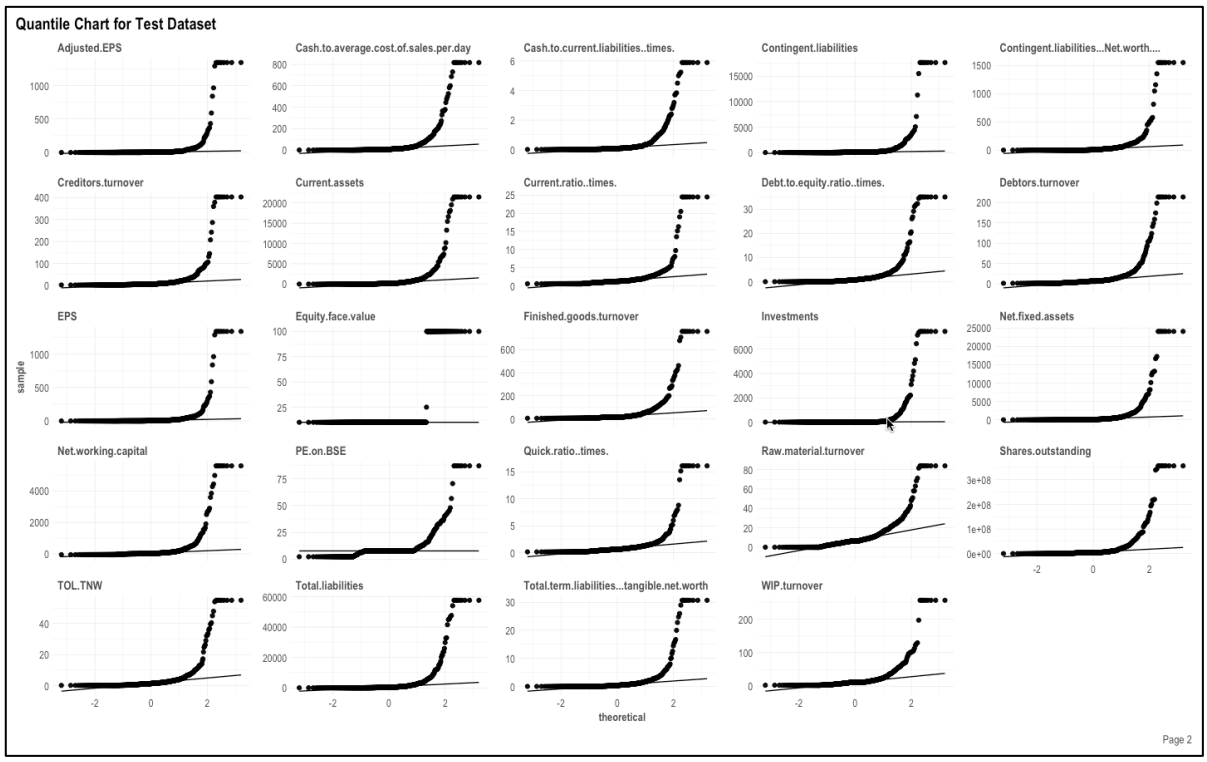




The scatterplot defines the scatter points is distributing by the defaulters list on the each variables and the variables are constructed in the increased frequency for the non-defaulters and the decreased scatter points in the defaulters. The scatter points are increased with both defaults.

```
> plot_qq(data = test_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
+         title = "Quantile Chart for Test Dataset")
```





The quantile charts explains the frequency are increased in the highest correlation with the values. The variables are measured to the decreased frequency in the negative values for the scatter points. QQ Plot explains the frequency distribution in the charts and explain the various values in the area of the plots.

3.4 Bivariate Analysis

Correlation Check

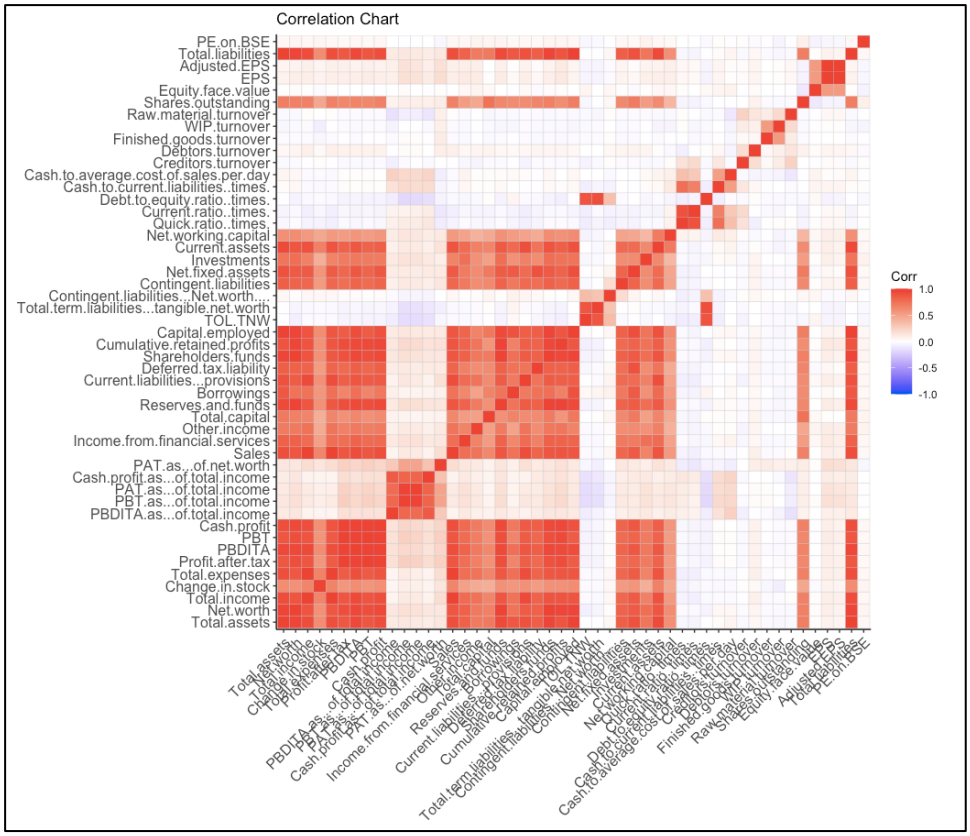
```
> correlation=cor(train_data[,-1])
```

The correlation of variables are identified and find the relationship between all variables with default variables.

The correlation of all variables are measured within the value take as negative correlation for -1 and positive correlation on +1.

```
> ggcorrplot::ggcorrplot(correlation,method = "square",ggtheme = theme_classic(),
+                        title = "Correlation Chart")
```

The ggcorrplot is used to plot the larger variables and the variables are plotted by,



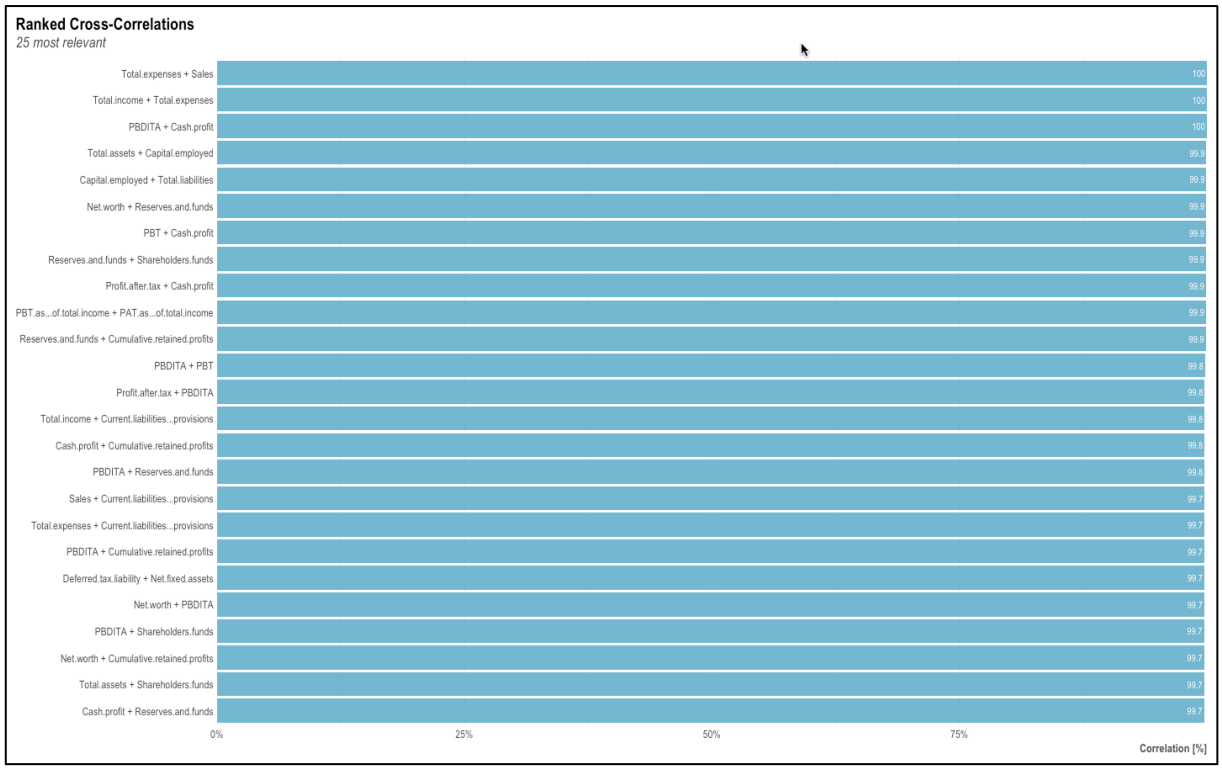
The highest correlation is found among various variables and the variables are used to find the largest relationship between two variables and the colour indicates the increased or decreased correlation values in the variables.

Most of the correlated variables values are between the range of -0.5 to +0.5 which shows the correlation have good prediction in the highest correlation values and lowest correlation values. Hence the rank of cross correlation is studied further.

> library(lares)

> corr_cross(correlation)

The correlation values and the top highest correlated variables are plotted using this function from the package lares. The plot shows the top 25 highest correlation between variables and the variables are used in predicting the most possible relationship of the variables. Cross validation can be done on the basis of the highest correlated variables.



The correlation is reflected with percentage and the highest correlation exists with the Total Expenses and Sales with showing the correlation of 0.9975872 is the highest correlation of variables in the datasets.

1. 4 Check for multicollinearity

Multicollinearity is checked for the prediction of the variables how related to each other in the linear model build. The multicollinearity values are related in predicting the values between -1 to 1 in P Values generated in the logistics regression model.

> model1=glm(Default...1~. -Default...1,data = train_data,family = binomial(link = logit))

Logistics regression model is created for the original dataset without creation of new ratios. The model is checked for the multicollinearity presented in the data.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.15E+00	3.14E-01	-3.68	0.000233	***
Total.assets	4.00E-03	1.16E-03	3.449	0.000563	***
Net.worth	1.17E-03	9.18E-04	1.276	0.201949	
Total.income	-4.40E-04	1.32E-03	-0.334	0.73827	
Change.in.stock	4.26E-04	3.68E-03	0.116	0.907852	
Total.expenses	3.89E-04	9.01E-04	0.432	0.66561	
Profit.after.tax	-4.62E-03	1.17E-02	-0.393	0.694152	
PBDITA	-6.40E-03	2.22E-03	-2.88	0.003971	**
PBT	1.23E-02	1.15E-02	1.066	0.286542	
Cash.profit	-4.93E-04	3.58E-03	-0.138	0.890536	
PBDITA.as...of.total.income	-2.96E-03	1.03E-02	-0.287	0.774155	
PBT.as...of.total.income	-2.17E-03	4.96E-02	-0.044	0.965162	
PAT.as...of.total.income	-9.44E-03	5.62E-02	-0.168	0.866635	
Cash.profit.as...of.total.income	-2.06E-02	1.82E-02	-1.128	0.259419	
PAT.as...of.net.worth	-6.03E-02	1.02E-02	-5.909	3.44E-09	***
Sales	-1.18E-04	1.10E-03	-0.108	0.914042	
Income.from.financial.services	4.83E-03	9.04E-03	0.534	0.593234	
Other.income	1.64E-02	9.51E-03	1.721	0.085261	.
Total.capital	2.68E-05	8.00E-04	0.033	0.97333	
Reserves.and.funds	-5.44E-03	1.55E-03	-3.509	0.000449	***
Borrowings	9.32E-04	1.48E-03	0.631	0.527907	
Current.liabilities...provisions	-3.11E-03	1.21E-03	-2.564	0.010338	*
Deferred.tax.liability	-4.23E-04	1.68E-03	-0.251	0.801428	
Shareholders.funds	-3.81E-04	1.44E-03	-0.264	0.791482	
Cumulative.retained.profits	-6.76E-03	2.82E-03	-2.395	0.016642	*
Capital.employed	-4.97E-03	1.99E-03	-2.492	0.012702	*
TOL.TNW	7.20E-02	1.82E-02	3.95	7.81E-05	***
Total.term.liabilities...tangible.net.worth	-1.44E-01	4.24E-02	-3.388	0.000704	***
Contingent.liabilities...Net.worth....	1.18E-03	7.91E-04	1.492	0.135679	
Contingent.liabilities	-1.44E-03	7.85E-04	-1.836	0.066333	.
Net.fixed.assets	9.64E-04	3.94E-04	2.446	0.014449	*
Investments	-1.97E-04	1.20E-03	-0.164	0.869972	
Current.assets	-6.59E-04	7.91E-04	-0.833	0.404568	
Net.working.capital	-2.40E-05	1.29E-03	-0.019	0.985182	
Quick.ratio..times.	-2.22E-01	1.45E-01	-1.531	0.125804	
Current.ratio..times.	-2.45E-02	7.38E-02	-0.332	0.739763	
Debt.to.equity.ratio..times.	1.48E-01	3.28E-02	4.505	6.63E-06	***
Cash.to.current.liabilities..times.	3.48E-01	2.30E-01	1.513	0.130298	
Cash.to.average.cost.of.sales.per.day	7.71E-04	4.50E-04	1.714	0.086579	.
Creditors.turnover	-1.39E-02	7.96E-03	-1.739	0.081963	.
Debtors.turnover	-5.28E-04	3.74E-03	-0.141	0.887827	
Finished.goods.turnover	6.66E-04	9.71E-04	0.686	0.492986	
WIP.turnover	-4.84E-03	3.93E-03	-1.232	0.217902	
Raw.material.turnover	-9.21E-03	8.52E-03	-1.081	0.279612	
Shares.outstanding	-2.11E-09	7.11E-09	-0.297	0.766317	
Equity.face.value	-2.75E-03	3.51E-03	-0.783	0.433636	
EPS	-2.93E-02	1.28E-01	-0.228	0.819359	
Adjusted.EPS	3.05E-02	1.28E-01	0.237	0.812329	
Total.liabilities	NA	NA	NA	NA	
PE.on.BSE	-7.45E-02	2.97E-02	-2.508	0.01214	*

The significance variables are identified and the insignificance variables are removed as per multicollinearity of the variables. The AIC shows the values of 1134.9 which higher on the significance variables.

```
>model2=glm(Default...1~Total.assets+PBDITA+PAT.as...of.net.worth+Other.income+Reser
ves.and.funds
+
+Current.liabilities...provisions+Cumulative.retained.profits+Capital.employed+TOL.TNW
+          +Total.term.liabilities...tangible.net.worth+Contingent.liabilities+Net.fixed.assets
+          +Debt.to.equity.ratio..times.+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
+          +PE.on.BSE,data = train_data,family = binomial)
```

The logistics model is prepared for the significant variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.59514	0.255052	-6.254	4.00E-10	***
Total.assets	0.001982	0.000751	2.638	0.00834	**
PBDITA	-0.00204	0.001227	-1.663	0.09627	.
PAT.as...of.net.worth	-0.06929	0.010148	-6.828	8.61E-12	***
Other.income	0.01067	0.007483	1.426	0.15389	
Reserves.and.funds	-0.00261	0.001063	-2.455	0.0141	*
Current.liabilities...provisions	-0.00177	0.000818	-2.163	0.03054	*
Cumulative.retained.profits	-0.00953	0.002933	-3.248	0.00116	**
Capital.employed	-0.00233	0.000842	-2.768	0.00564	**
TOL.TNW	0.075849	0.017252	4.397	1.10E-05	***
Total.term.liabilities...tangible.net.worth	-0.16166	0.040216	-4.02	5.83E-05	***
Contingent.liabilities	-0.00064	0.000463	-1.394	0.16346	
Net.fixed.assets	0.000709	0.000266	2.67	0.00759	**
Debt.to.equity.ratio..times.	0.161914	0.030957	5.23	1.69E-07	***
Cash.to.average.cost.of.sales.per.day	0.000612	0.000321	1.908	0.05633	.
Creditors.turnover	-0.0185	0.007655	-2.417	0.01565	*
PE.on.BSE	-0.07818	0.028077	-2.784	0.00536	**

The Significant variables are identified after removing the variables in the logistics model and the AIC value is measured with 1108.9 which is lesser then the previous model and shows the fitted values in the model for the better understanding of the variables.

```
> model3=glm(Default...1~Total.assets+PBDITA+PAT.as...of.net.worth+Reserves.and.funds
+
+Current.liabilities...provisions+Cumulative.retained.profits+Capital.employed+TOL.TNW
+          +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
+          +Debt.to.equity.ratio..times.+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
+          +PE.on.BSE,data = train_data,family = binomial)
```

The significant variables are created for the another model in the variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.63054	0.253201	-6.44	1.20E-10	***
Total.assets	0.001601	0.000696	2.299	0.021487	*
PBDITA	-0.0015	0.001114	-1.344	0.178974	
PAT.as...of.net.worth	-0.06984	0.010178	-6.862	6.80E-12	***
Reserves.and.funds	-0.00209	0.001012	-2.068	0.038622	*
Current.liabilities...provisions	-0.00137	0.000733	-1.875	0.060815	.
Cumulative.retained.profits	-0.0098	0.002918	-3.357	0.000788	***
Capital.employed	-0.00185	0.000772	-2.397	0.016513	*
TOL.TNW	0.074404	0.017082	4.356	1.33E-05	***
Total.term.liabilities...tangible.net.worth	-0.16345	0.040346	-4.051	5.09E-05	***
Net.fixed.assets	0.000524	0.000247	2.124	0.033688	*
Debt.to.equity.ratio..times.	0.164791	0.031136	5.293	1.21E-07	***
Cash.to.average.cost.of.sales.per.day	0.000626	0.00032	1.958	0.050245	.
Creditors.turnover	-0.01863	0.007641	-2.438	0.014757	*
PE.on.BSE	-0.07684	0.027898	-2.754	0.005882	**

The model shows the variables are significant with values and PBDITA is observed with insignificance when compared with previous model with AIC value of 1109.3.

```
> model4=glm(Default...1~Total.assets+PAT.as...of.net.worth+Reserves.and.funds
+
+Current.liabilities...provisions+Cumulative.retained.profits+Capital.employed+TOL.TNW
+          +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
+          +Debt.to.equity.ratio..times.+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
+          +PE.on.BSE,data = train_data,family = binomial)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.33E+14	1.94E+06	-1E+08	<2e-16	***
Total.assets	3.13E+11	1.96E+03	2E+08	<2e-16	***
PAT.as...of.net.worth	-3.03E+13	6.22E+04	-5E+08	<2e-16	***
Reserves.and.funds	-3.71E+11	2.45E+03	-2E+08	<2e-16	***
Current.liabilities...provisions	-3.90E+11	2.69E+03	-1E+08	<2e-16	***
Cumulative.retained.profits	-1.18E+11	2.67E+03	-4E+07	<2e-16	***
Capital.employed	-3.48E+11	2.35E+03	-1E+08	<2e-16	***
TOL.TNW	4.80E+13	3.78E+05	1E+08	<2e-16	***
Total.term.liabilities...tangible.net.worth	-5.87E+13	8.30E+05	-7E+07	<2e-16	***
Net.fixed.assets	7.48E+10	1.30E+03	6E+07	<2e-16	***
Debt.to.equity.ratio..times.	5.03E+12	6.53E+05	8E+06	<2e-16	***
Cash.to.average.cost.of.sales.per.day	3.13E+11	7.15E+03	4E+07	<2e-16	***
Creditors.turnover	-9.11E+12	6.01E+04	-2E+08	<2e-16	***
PE.on.BSE	-1.08E+13	6.07E+04	-2E+08	<2e-16	***

All the variables are significant but the P values are very less that often gets the variables are not fitted to the variables in the dataset. The variance of the fitted model is defined by the Variance Inflation Factor.

```
> car::vif(model4)
```

Total.assets	PAT.as...of.net.worth
130.555241	1.0533
Reserves.and.funds	Current.liabilities...provisions
21.876383	12.20843
Cumulative.retained.profits	Capital.employed
14.215531	86.158994
TOL.TNW	Total.term.liabilities...tangible.net.worth
5.865426	7.063598
Net.fixed.assets	Debt.to.equity.ratio..times.
6.04492	7.372041
Cash.to.average.cost.of.sales.per.day	Creditors.turnover
1.043003	1.017126
PE.on.BSE	
1.009053	

The Variance Inflation factor should be less than 5 to consider it as good model. Hence the variables with higher inflation rate are removed from the following model and check it for the significance of the variables developed in the model.

```
> model5=glm(Default...1~PAT.as...of.net.worth+Reserves.and.funds
+          +Cumulative.retained.profits+TOL.TNW
+          +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
+          +Debt.to.equity.ratio..times.
+          +Cash.to.average.cost.of.sales.per.day+Creditors.turnover
+          +PE.on.BSE,data = train_data,family = binomial)
> summary(model5)
```

The model is developed without the higher Variance Factors.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.711057	0.245335	-6.974	3.07E-12	***
PAT.as...of.net.worth	-0.071543	0.010179	-7.029	2.09E-12	***
Reserves.and.funds	-0.000546	0.00052	-1.051	0.29347	
Cumulative.retained.profits	-0.011531	0.002645	-4.359	1.31E-05	***
TOL.TNW	0.0781147	0.017078	4.574	4.79E-06	***
Total.term.liabilities...tangible.net.worth	-0.174517	0.040236	-4.337	1.44E-05	***
Net.fixed.assets	0.0002372	0.000115	2.062	0.03924	*
Debt.to.equity.ratio..times.	0.1662786	0.031148	5.338	9.38E-08	***
Cash.to.average.cost.of.sales.per.day	0.0007524	0.00031	2.431	0.01507	*
Creditors.turnover	-0.018922	0.007667	-2.468	0.01359	*
PE.on.BSE	-0.073005	0.026863	-2.718	0.00657	**

The model is defined with insignificant for the Reserves and funds which is no longer in relationship of the variables in the dataset. Hence the reserves and funds is not selected in the further model definition of the dataset.

```
> model6=glm(Default...1~PAT.as...of.net.worth
+          +Cumulative.retained.profits+TOL.TNW
+          +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
+          +Debt.to.equity.ratio..times.
+          +Cash.to.average.cost.of.sales.per.day+Creditors.turnover
+          +PE.on.BSE,data = train_data,family = binomial)
```

The model is developed for the significant variables.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.71E+00	2.45E-01	-6.977	3.01E-12	***
PAT.as...of.net.worth	-7.15E-02	1.02E-02	-7.022	2.18E-12	***
Cumulative.retained.profits	-1.22E-02	2.58E-03	-4.725	2.30E-06	***
TOL.TNW	7.86E-02	1.71E-02	4.599	4.24E-06	***
Total.term.liabilities...tangible.net.worth	-1.74E-01	4.02E-02	-4.334	1.47E-05	***
Net.fixed.assets	1.67E-04	8.98E-05	1.858	0.0631	.
Debt.to.equity.ratio..times.	1.66E-01	3.11E-02	5.316	1.06E-07	***
Cash.to.average.cost.of.sales.per.day	7.09E-04	3.08E-04	2.301	0.0214	*
Creditors.turnover	-1.87E-02	7.65E-03	-2.448	0.0143	*
PE.on.BSE	-7.34E-02	2.68E-02	-2.737	0.0062	**

The model is defined with the significant variables and the variables are understandable with the defaulters for the higher significance. This variables are contributing the higher values in the default status of the creditors.

```
> car::vif(model6)
```

PAT.as...of.net.worth	Cumulative.retained.profits
1.177932	1.112141
TOL.TNW	Total.term.liabilities...tangible.net.worth
3.771168	5.879349
Net.fixed.assets	Debt.to.equity.ratio..times.
1.019773	5.622013
Cash.to.average.cost.of.sales.per.day	Creditors.turnover
1.022241	1.038673
PE.on.BSE	
1.043259	

The variance inflation of the variables are checked for the significant model in which total term liabilities and tangible net worth is having higher inflation rate and the variables are making the less significant in the Net fixed assets.

```
> model7=glm(Default...1~PAT.as...of.net.worth+Cumulative.retained.profits+TOL.TNW
+          +Net.fixed.assets+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
+          +PE.on.BSE,data = train_data,family = binomial)
```

The higher inflation rate variables, less significant variables are removed with the multicollinearity and the significant variables of the final model is created.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.55E+00	2.43E-01	-6.38	1.77E-10	***
PAT.as...of.net.worth	-6.52E-02	1.04E-02	-6.271	3.60E-10	***
Cumulative.retained.profits	-1.32E-02	2.63E-03	-5.026	5.02E-07	***
TOL.TNW	8.98E-02	8.50E-03	10.562	< 2e-16	***
Net.fixed.assets	1.78E-04	8.36E-05	2.133	0.03293	*
Cash.to.average.cost.of.sales.per.day	5.74E-04	3.09E-04	1.86	0.06286	.
Creditors.turnover	-1.65E-02	7.15E-03	-2.302	0.02135	*
PE.on.BSE	-8.10E-02	2.72E-02	-2.973	0.00295	**

The significant variables are identified and the p values are measured with lesser values on Cash to average cost of sales per day on the model.

```
> car::vif(model7)
```

PAT.as...of.net.worth	Cumulative.retained.profits
1.12227	1.127346
TOL.TNW	Net.fixed.assets
1.02108	1.013057
Cash.to.average.cost.of.sales.per.day	Creditors.turnover
1.014794	1.016374
PE.on.BSE	
1.040168	

The variance of the each significant variables are measured with good fitted values and the model is best fitted for the dataset without any creations of ratios in the datasets.

- The Profitability is measured with Profit After Tax as of Net Worth and Cumulative Retained Profits.
- The leverage is measured with the variables Total Liabilities of the customers by the Total Net worth of the company.
- The liquidity is measured with the variables Cash to average cost of sales per day and creditors turnover in the logistics model.
- The company size variables are identified by the Net Fixed Assets. The market sentiment is mentioned by the company current stock.

After removing variables for multicollinearity, try to take at least one variable for creating the model from each of the 4 factors namely -

1) Profitability
2) Leverage
3) Liquidity
4) Company's size

Creation of new variables

```
> #Profitability Ratio - Gross Margin Ratio
> train_data$Gross.Margin.Ratio=train_data$Total.income/train_data$Sales
```

Gross Margin Ratio is measured with ratio of Total Income and Sales from the companies.

```
> #Profitability Ratio - Return on Assets Ratio
> train_data$Return.on.assets.ratio=train_data$Total.income/train_data$Total.assets
```

Return on Assets is calculated from the ratio of Total Income and Total Assets in the companies which is measured for the retuned assets in the dataset.

```
> #Profitability - Return on Equity Ratio
> train_data$Return.on.equity.ratio=train_data$Total.income/train_data$Shareholders.funds
```

Return on equity ratio are created and the variables are identified from the ratio of Total Income and Shareholders funds of the variables.

```
> #Leverage Financial Ratio - Debt to Equity Ratio
> train_data$Debt.to.equity.ratio=train_data$Total.liabilities/train_data$Shareholders.funds
```

The leverage ratio is identified by the ratio of Total Liabilities and Shareholder's Fund in the companies.

```
> #Liquidity Ratio - Current Ratio
>train_data$Current.ratio=train_data$Current.assets/train_data$Current.liabilities...provisions
```

The current ratio is defined by the Current Assets and Current Liabilities Provisions in the data.

```
> #Liquidity Ratio - Cash Ratio
> train_data$Cash.ratio=train_data$Cash.profit/train_data$Current.liabilities...provisions
```

The cash ratio is the working ratio of Cash Profit and Current Liabilities Provisions in the data for identify the current flow of cash in the company.

```
> #Efficiency Ratio - Asset Turnover Ratio
> train_data$Asset.turnover.ratio=train_data$Sales/train_data$Total.assets
```

The Efficiency ratio defines the working finance capability of the variables in the various variables by Sales and Total Assets.

```
> #Efficiency Ratio - Day Sales in Inventory Ratio
> train_data$Day.sales.in.inventory.ratio=365/train_data$Finished.goods.turnover
```

Day Sales in Inventory ratio is calculated for yearly basis with the finished goods turnover in the company of the dataset.

Similarly the ratio are created for the Test Dataset to make the final credit risk model.

```
> #Profitability Ratio - Gross Margin Ratio
> test_data$Gross.Margin.Ratio=test_data$Total.income/test_data$Sales
```

The Gross Margin Ratio is further created with Total Income and Sales of the test datasets.

```
> #Profitability Ratio - Return on Assets Ratio
> test_data$Return.on.assets.ratio=test_data$Total.income/test_data$Total.assets
```

Return On Assets is calculated for the ratio of the Total Income and Total sales of the companies in the test data.

```
> #Profitability - Retun on Equity Ratio
> test_data$Return.on.equity.ratio=test_data$Total.income/test_data$Shareholders.funds
```

Return on Equity Ratio is defined by the ratio of Total Income and Shareholder's funds in the Test Dataset.

```
> #Leverage Financial Ratio - Debt to Equity Ratio
> test_data$Debt.to.equity.ratio=test_data$Total.liabilities/test_data$Shareholders.funds
```

Debt to equity ratio is measured with the ratio of total liabilities and shareholder's funds in the test dataset.

```
> #Liquidity Ratio - Current Ratio
> test_data$Current.ratio=test_data$Current.assets/test_data$Current.liabilities...provisions
```

Current ratio is identified by the ratio measured with the current assets and current liabilities with provisions of the test data.

```
> #Liquidity Ratio - Cash Ratio
> test_data$Cash.ratio=test_data$Cash.profit/test_data$Current.liabilities...provisions
```

The test data variables are calculated for the ratio of cash profit and current liabilities and provisions.

```
> #Efficiency Ratio - Asset Turnover Ratio
> test_data$Asset.turnover.ratio=test_data$Sales/test_data$Total.assets
```

Asset Turnover Ratio is measured for the ratio of the Sales and Total assets in the test data.

```
> #Efficiency Ratio - Day Sales in Inventory Ratio
> test_data$Day.sales.in.inventory.ratio=365/test_data$Finished.goods.turnover
```

Day Sales in Inventory ratio is calculated by year and finished goods turnover of the variables.

2. Modelling

2.1 Build Logistic Regression Model on most important variables

```
> log.reg1=glm(Default...1~.,data = train_data,family = binomial)
```

The Logistics model is created with the new variables ratio of Gross Margin Ratio, Return on Assets Ratio, Return on Equity Ratio, Debt to Equity Ratio, Current Ratio, Cash Ratio, Asset Turnover Ratio, Day Sales in Inventory Rate which is diversely classified in the Profitability Ratio, Leverage Ratio, Liquidity Ratio and Efficiency Ratio of basic financial ratios for both train data and validation dataset. Hence all the LR Model has been built on each variables.

> summary(log.reg1)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.98E+13	3.82E+06	1E+07	<2e-16	***
Total.assets	2.18E+11	2.46E+03	9E+07	<2e-16	***
Net.worth	1.41E+11	9.50E+03	1E+07	<2e-16	***
Total.income	-2.30E+11	1.03E+04	-2E+07	<2e-16	***
Change.in.stock	-4.25E+11	2.09E+04	-2E+07	<2e-16	***
Total.expenses	2.79E+11	6.16E+03	5E+07	<2e-16	***
Profit.after.tax	-8.55E+11	3.21E+04	-3E+07	<2e-16	***
PBDITA	-1.54E+12	1.19E+04	-1E+08	<2e-16	***
PBT	2.54E+12	2.54E+04	1E+08	<2e-16	***
Cash.profit	1.44E+12	1.58E+04	9E+07	<2e-16	***
PBDITA.as...of.total.income	-1.56E+13	2.08E+05	-7E+07	<2e-16	***
PBT.as...of.total.income	-4.56E+12	4.80E+05	-9E+06	<2e-16	***
PAT.as...of.total.income	-6.57E+12	5.71E+05	-1E+07	<2e-16	***
Cash.profit.as...of.total.income	1.17E+13	3.02E+05	4E+07	<2e-16	***
PAT.as...of.net.worth	-2.94E+13	7.86E+04	-4E+08	<2e-16	***
Sales	-1.51E+11	8.83E+03	-2E+07	<2e-16	***
Income.from.financial.services	9.58E+11	2.85E+04	3E+07	<2e-16	***
Other.income	2.78E+12	5.38E+04	5E+07	<2e-16	***
Total.capital	1.40E+11	6.60E+03	2E+07	<2e-16	***
Reserves.and.funds	-2.70E+11	3.81E+03	-7E+07	<2e-16	***
Borrowings	-1.01E+11	2.56E+03	-4E+07	<2e-16	***
Current.liabilities...provisions	-7.65E+10	3.88E+03	-2E+07	<2e-16	***
Deferred.tax.liability	2.25E+11	1.11E+04	2E+07	<2e-16	***
Shareholders.funds	-2.28E+11	9.72E+03	-2E+07	<2e-16	***
Cumulative.retained.profits	-3.09E+11	3.63E+03	-9E+07	<2e-16	***
Capital.employed	-1.57E+11	3.76E+03	-4E+07	<2e-16	***
TOL.TNW	2.31E+13	4.08E+05	6E+07	<2e-16	***
Total.term.liabilities...tangible.net.worth	-2.97E+13	8.58E+05	-3E+07	<2e-16	***
Contingent.liabilities...Net.worth....	1.05E+11	1.37E+04	8E+06	<2e-16	***
Contingent.liabilities	-2.83E+11	2.97E+03	-1E+08	<2e-16	***
Net.fixed.assets	1.64E+11	2.34E+03	7E+07	<2e-16	***
Investments	1.12E+11	3.75E+03	3E+07	<2e-16	***
Current.assets	-5.73E+10	2.46E+03	-2E+07	<2e-16	***
Net.working.capital	-3.90E+11	4.19E+03	-9E+07	<2e-16	***
Quick.ratio..times.	-6.58E+13	1.81E+06	-4E+07	<2e-16	***
Current.ratio..times.	1.92E+12	1.18E+06	2E+06	<2e-16	***
Debt.to.equity.ratio..times.	2.83E+13	6.71E+05	4E+07	<2e-16	***
Cash.to.current.liabilities..times.	1.08E+14	2.60E+06	4E+07	<2e-16	***
Cash.to.average.cost.of.sales.per.day	-6.33E+10	9.20E+03	-7E+06	<2e-16	***
Creditors.turnover	-5.05E+12	6.59E+04	-8E+07	<2e-16	***
Debtors.turnover	-3.59E+11	4.90E+04	-7E+06	<2e-16	***
Finished.goods.turnover	3.08E+11	1.21E+04	3E+07	<2e-16	***
WIP.turnover	-9.93E+11	4.71E+04	-2E+07	<2e-16	***
Raw.material.turnover	-6.16E+10	8.71E+04	-7E+05	<2e-16	***
Shares.outstanding	-3.52E+06	5.42E-02	-6E+07	<2e-16	***
Equity.face.value	-1.50E+11	4.90E+04	-3E+06	<2e-16	***
EPS	-1.58E+13	1.20E+05	-1E+08	<2e-16	***
Adjusted.EPS	1.48E+13	1.20E+05	1E+08	<2e-16	***
Total.liabilities	NA	NA	NA	NA	
PE.on.BSE	-2.91E+13	6.16E+04	-5E+08	<2e-16	***
Gross.Margin.Ratio	4.17E+13	4.42E+05	9E+07	<2e-16	***
Return.on.assets.ratio	-5.43E+13	1.09E+06	-5E+07	<2e-16	***
Return.on.equity.ratio	8.98E+12	2.57E+05	3E+07	<2e-16	***
Debt.to.equity.ratio	-8.60E+12	3.16E+05	-3E+07	<2e-16	***
Current.ratio	-2.87E+12	2.63E+05	-1E+07	<2e-16	***
Cash.ratio	9.93E+13	1.62E+06	6E+07	<2e-16	***
Asset.turnover.ratio	-7.05E+12	7.20E+05	-1E+07	<2e-16	***
Day.sales.in.inventory.ratio	2.00E+12	6.22E+04	3E+07	<2e-16	***

The logistics regression model is developed with all variables and the significance of the variables are measured for the P value where the hypothesis of the P Value is measured for -1 to +1 the values measured with the variables for the very less values for the variables. The logistics regression model is also build to predict the significant and insignificant variables of the data.

```
>log.reg2=glm(Default...1~Total.assets+Net.worth+Total.income+Change.in.stock+Total.exp
enses+
+              Profit.after.tax+PBDITA+PBT+Cash.profit+PBDITA.as...of.total.income+
+             PBT.as...of.total.income+PAT.as...of.total.income+Cash.profit.as...of.total.income+
+              PAT.as...of.net.worth+Sales+Income.from.financial.services+Other.income+
+              Total.capital+Reserves.and.funds+Borrowings+Current.liabilities...provisions+
+              Deferred.tax.liability+Shareholders.funds+Cumulative.retained.profits+
+              Capital.employed+TOL.TNW+Total.term.liabilities...tangible.net.worth+
+              Contingent.liabilities...Net.worth....+Contingent.liabilities+Net.fixed.assets+
+              Investments+Current.assets+Net.working.capital+Quick.ratio..times.+
+              Current.ratio..times.+Debt.to.equity.ratio..times.+
+              Cash.to.current.liabilities..times.+Cash.to.average.cost.of.sales.per.day+
+              Creditors.turnover+Debtors.turnover+Finished.goods.turnover+WIP.turnover+
+            Raw.material.turnover+Shares.outstanding+Equity.face.value+EPS+Adjusted.EPS+
+              PE.on.BSE+Gross.Margin.Ratio+Return.on.assets.ratio+Return.on.equity.ratio+
+              Debt.to.equity.ratio+Current.ratio+Cash.ratio+Asset.turnover.ratio+
+              Day.sales.in.inventory.ratio,data = train_data,
+            family = binomial)
```

The logistics regression model is built with the significant variables from the previous model. The variables are measured with the default status from the data. Since the previous model significant variables are measured with very less P value hence model can be less fitted values and the model is not measured for the all values in the variables.

```
> summary(log.reg2)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.98E+13	3.82E+06	1E+07	<2e-16	***
Total.assets	2.18E+11	2.46E+03	9E+07	<2e-16	***
Net.worth	1.41E+11	9.50E+03	1E+07	<2e-16	***
Total.income	-2.30E+11	1.03E+04	-2E+07	<2e-16	***
Change.in.stock	-4.25E+11	2.09E+04	-2E+07	<2e-16	***
Total.expenses	2.79E+11	6.16E+03	5E+07	<2e-16	***
Profit.after.tax	-8.55E+11	3.21E+04	-3E+07	<2e-16	***
PBDITA	-1.54E+12	1.19E+04	-1E+08	<2e-16	***

PBT	2.54E+12	2.54E+04	1E+08	<2e-16	***
Cash.profit	1.44E+12	1.58E+04	9E+07	<2e-16	***
PBDITA.as...of.total.income	-1.56E+13	2.08E+05	-7E+07	<2e-16	***
PBT.as...of.total.income	-4.56E+12	4.80E+05	-9E+06	<2e-16	***
PAT.as...of.total.income	-6.57E+12	5.71E+05	-1E+07	<2e-16	***
Cash.profit.as...of.total.income	1.17E+13	3.02E+05	4E+07	<2e-16	***
PAT.as...of.net.worth	-2.94E+13	7.86E+04	-4E+08	<2e-16	***
Sales	-1.51E+11	8.83E+03	-2E+07	<2e-16	***
Income.from.financial.services	9.58E+11	2.85E+04	3E+07	<2e-16	***
Other.income	2.78E+12	5.38E+04	5E+07	<2e-16	***
Total.capital	1.40E+11	6.60E+03	2E+07	<2e-16	***
Reserves.and.funds	-2.70E+11	3.81E+03	-7E+07	<2e-16	***
Borrowings	-1.01E+11	2.56E+03	-4E+07	<2e-16	***
Current.liabilities...provisions	-7.65E+10	3.88E+03	-2E+07	<2e-16	***
Deferred.tax.liability	2.25E+11	1.11E+04	2E+07	<2e-16	***
Shareholders.funds	-2.28E+11	9.72E+03	-2E+07	<2e-16	***
Cumulative.retained.profits	-3.09E+11	3.63E+03	-9E+07	<2e-16	***
Capital.employed	-1.57E+11	3.76E+03	-4E+07	<2e-16	***
TOL.TNW	2.31E+13	4.08E+05	6E+07	<2e-16	***
Total.term.liabilities...tangible.net.worth	-2.97E+13	8.58E+05	-3E+07	<2e-16	***
Contingent.liabilities...Net.worth....	1.05E+11	1.37E+04	8E+06	<2e-16	***
Contingent.liabilities	-2.83E+11	2.97E+03	-1E+08	<2e-16	***
Net.fixed.assets	1.64E+11	2.34E+03	7E+07	<2e-16	***
Investments	1.12E+11	3.75E+03	3E+07	<2e-16	***
Current.assets	-5.73E+10	2.46E+03	-2E+07	<2e-16	***
Net.working.capital	-3.90E+11	4.19E+03	-9E+07	<2e-16	***
Quick.ratio..times.	-6.58E+13	1.81E+06	-4E+07	<2e-16	***
Current.ratio..times.	1.92E+12	1.18E+06	2E+06	<2e-16	***
Debt.to.equity.ratio..times.	2.83E+13	6.71E+05	4E+07	<2e-16	***
Cash.to.current.liabilities..times.	1.08E+14	2.60E+06	4E+07	<2e-16	***
Cash.to.average.cost.of.sales.per.day	-6.33E+10	9.20E+03	-7E+06	<2e-16	***
Creditors.turnover	-5.05E+12	6.59E+04	-8E+07	<2e-16	***
Debtors.turnover	-3.59E+11	4.90E+04	-7E+06	<2e-16	***
Finished.goods.turnover	3.08E+11	1.21E+04	3E+07	<2e-16	***
WIP.turnover	-9.93E+11	4.71E+04	-2E+07	<2e-16	***
Raw.material.turnover	-6.16E+10	8.71E+04	-7E+05	<2e-16	***
Shares.outstanding	-3.52E+06	5.42E-02	-6E+07	<2e-16	***
Equity.face.value	-1.50E+11	4.90E+04	-3E+06	<2e-16	***
EPS	-1.58E+13	1.20E+05	-1E+08	<2e-16	***
Adjusted.EPS	1.48E+13	1.20E+05	1E+08	<2e-16	***
PE.on.BSE	-2.91E+13	6.16E+04	-5E+08	<2e-16	***
Gross.Margin.Ratio	4.17E+13	4.42E+05	9E+07	<2e-16	***
Return.on.assets.ratio	-5.43E+13	1.09E+06	-5E+07	<2e-16	***
Return.on.equity.ratio	8.98E+12	2.57E+05	3E+07	<2e-16	***
Debt.to.equity.ratio	-8.60E+12	3.16E+05	-3E+07	<2e-16	***
Current.ratio	-2.87E+12	2.63E+05	-1E+07	<2e-16	***
Cash.ratio	9.93E+13	1.62E+06	6E+07	<2e-16	***
Asset.turnover.ratio	-7.05E+12	7.20E+05	-1E+07	<2e-16	***
Day.sales.in.inventory.ratio	2.00E+12	6.22E+04	3E+07	<2e-16	***

The model is measured for the variables are being dependent with the default variables. The variables with less very less p values are measured further for the variance inflation factor is higher when compared to the basic variables of the significance variables.

The higher variation factor variables may results in the less p values and make the models to not good fit.

> car::vif(log.reg2)

Total.assets	Net.worth
205.046738	466.585786
Total.income	Change.in.stock
2605.525814	2.440045
Total.expenses	Profit.after.tax
797.713305	194.254596
PBDITA	PBT
98.182122	217.794673
Cash.profit	PBDITA.as...of.total.income
87.828867	5.619524
PBT.as...of.total.income	PAT.as...of.total.income
15.453068	14.472693
Cash.profit.as...of.total.income	PAT.as...of.net.worth
6.0171	1.679502
Sales	Income.from.financial.services
1788.009499	5.821039
Other.income	Total.capital
2.974837	5.283035
Reserves.and.funds	Borrowings
52.940367	18.508734
Current.liabilities...provisions	Deferred.tax.liability
25.486552	7.523562
Shareholders.funds	Cumulative.retained.profits
492.464455	26.316973
Capital.employed	TOL.TNW
220.850524	6.854175
Total.term.liabilities...tangible.net.worth	Contingent.liabilities...Net.worth....
7.548341	1.369739
Contingent.liabilities	Net.fixed.assets
4.622284	19.662065
Investments	Current.assets
4.496111	24.107962
Net.working.capital	Quick.ratio..times.
3.238569	7.649234
Current.ratio..times.	Debt.to.equity.ratio..times.
5.864808	7.776218
Cash.to.current.liabilities..times.	Cash.to.average.cost.of.sales.per.day
3.150497	1.725617
Creditors.turnover	Debtors.turnover
1.222476	1.152225
Finished.goods.turnover	WIP.turnover
1.503317	1.610072
Raw.material.turnover	Shares.outstanding
1.165363	3.183915
Equity.face.value	EPS
1.578459	141.691541
Adjusted.EPS	PE.on.BSE
140.124623	1.039231
Gross.Margin.Ratio	Return.on.assets.ratio
3.595957	8.825794
Return.on.equity.ratio	Debt.to.equity.ratio
5.842552	2.727076
Current.ratio	Cash.ratio
1.324189	1.643634
Asset.turnover.ratio	Day.sales.in.inventory.ratio
5.147193	1.35755

The variables with higher inflation factor are removed from further model building.

```
> log.reg3=glm(Default...1~Change.in.stock+
+              PBDITA+Cash.profit+PBDITA.as...of.total.income+
+             PBT.as...of.total.income+PAT.as...of.total.income+Cash.profit.as...of.total.income+
+              PAT.as...of.net.worth+Income.from.financial.services+Other.income+
+              Total.capital+Reserves.and.funds+Borrowings+Current.liabilities...provisions+
+              Deferred.tax.liability+Cumulative.retained.profits+
+              TOL.TNW+Total.term.liabilities...tangible.net.worth+
+              Contingent.liabilities...Net.worth....+Contingent.liabilities+Net.fixed.assets+
+              Investments+Current.assets+Net.working.capital+Quick.ratio..times.+
+              Current.ratio..times.+Debt.to.equity.ratio..times.+
+              Cash.to.current.liabilities..times.+Cash.to.average.cost.of.sales.per.day+
+              Creditors.turnover+Debtors.turnover+Finished.goods.turnover+WIP.turnover+
+              Raw.material.turnover+Shares.outstanding+Equity.face.value+
+              PE.on.BSE+Gross.Margin.Ratio+Return.on.assets.ratio+Return.on.equity.ratio+
+              Debt.to.equity.ratio+Current.ratio+Cash.ratio+Asset.turnover.ratio+
+              Day.sales.in.inventory.ratio,data = train_data,
+            family = binomial)
```

The higher variance inflation factor variables with more than 1000 are removed from further analysis which results in the variables are significant in one p values and the most important variables are derived from the linear relationship of the default companies. The defaulted status is predicted with lesser inflation factor and the variables are associated for the further relations with default companies.

```
> summary(log.reg3)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.01E+00	3.84E-01	-5.234	1.66E-07	***
Change.in.stock	-9.26E-04	3.15E-03	-0.294	0.769015	
PBDITA	-1.16E-03	1.71E-03	-0.677	0.498376	
Cash.profit	-4.24E-03	3.76E-03	-1.126	0.260007	
PBDITA.as...of.total.income	-4.06E-03	1.06E-02	-0.383	0.701671	
PBT.as...of.total.income	-9.21E-04	4.65E-02	-0.02	0.984191	
PAT.as...of.total.income	-2.52E-02	5.32E-02	-0.474	0.635229	
Cash.profit.as...of.total.income	-7.62E-03	1.82E-02	-0.419	0.67525	
PAT.as...of.net.worth	-5.40E-02	1.02E-02	-5.317	1.05E-07	***
Income.from.financial.services	1.95E-03	7.97E-03	0.245	0.806281	
Other.income	1.06E-02	8.62E-03	1.232	0.218048	
Total.capital	4.40E-04	6.13E-04	0.717	0.473327	
Reserves.and.funds	-5.32E-04	8.17E-04	-0.65	0.515589	

Borrowings	-1.89E-04	2.71E-04	-0.7	0.483924	
Current.liabilities...provisions	1.51E-03	7.63E-04	1.981	0.047592	*
Deferred.tax.liability	1.72E-03	1.59E-03	1.077	0.28138	
Cumulative.retained.profits	-1.08E-02	3.00E-03	-3.594	0.000326	***
TOL.TNW	7.58E-02	1.92E-02	3.943	8.04E-05	***
Total.term.liabilities...tangible.net.worth	-1.62E-01	4.33E-02	-3.748	0.000178	***
Contingent.liabilities...Net.worth....	1.37E-03	7.86E-04	1.741	0.081756	.
Contingent.liabilities	-1.37E-03	6.85E-04	-1.994	0.046182	*
Net.fixed.assets	5.73E-04	3.44E-04	1.666	0.09571	.
Investments	-7.51E-04	8.35E-04	-0.9	0.368359	
Current.assets	-1.08E-03	5.99E-04	-1.8	0.071846	.
Net.working.capital	1.44E-04	1.11E-03	0.13	0.896276	
Quick.ratio..times.	-1.88E-01	1.47E-01	-1.276	0.201961	
Current.ratio..times.	-5.50E-02	8.00E-02	-0.688	0.491561	
Debt.to.equity.ratio..times.	1.70E-01	3.34E-02	5.101	3.37E-07	***
Cash.to.current.liabilities..times.	3.65E-01	2.23E-01	1.636	0.101742	
Cash.to.average.cost.of.sales.per.day	5.50E-04	4.18E-04	1.315	0.188491	
Creditors.turnover	-1.16E-02	7.86E-03	-1.476	0.13984	
Debtors.turnover	-7.56E-04	3.71E-03	-0.204	0.838334	
Finished.goods.turnover	9.81E-04	9.73E-04	1.008	0.313361	
WIP.turnover	-1.95E-03	4.00E-03	-0.488	0.62527	
Raw.material.turnover	-6.05E-03	8.26E-03	-0.732	0.464022	
Shares.outstanding	-5.97E-09	7.55E-09	-0.791	0.429	
Equity.face.value	-5.80E-04	3.22E-03	-0.18	0.856866	
PE.on.BSE	-7.79E-02	3.05E-02	-2.555	0.010627	*
Gross.Margin.Ratio	1.58E-01	8.50E-02	1.857	0.063261	.
Return.on.assets.ratio	-3.00E-02	4.80E-02	-0.626	0.53154	
Return.on.equity.ratio	9.80E-03	1.14E-02	0.858	0.391137	
Debt.to.equity.ratio	-2.71E-02	1.28E-02	-2.129	0.033262	*
Current.ratio	-6.57E-03	1.51E-02	-0.435	0.663611	
Cash.ratio	1.41E-01	9.95E-02	1.419	0.155889	
Asset.turnover.ratio	4.84E-02	3.28E-02	1.476	0.139867	
Day.sales.in.inventory.ratio	1.18E-02	4.32E-03	2.725	0.006439	**

The variables are measured for the less significant p values and the lesser significant variables are considered as the non-important variables in creating the logistics regression fitted model for the prediction of the credit risk. Hence the not fitted variables are removed in the further analysis.

```
> log.reg4=glm(Default...1~PAT.as...of.net.worth+Current.liabilities...provisions+
+              Cumulative.retained.profits+TOL.TNW+Total.term.liabilities...tangible.net.worth+
+              Contingent.liabilities...Net.worth....+Contingent.liabilities+Net.fixed.assets+
+              Current.assets+Debt.to.equity.ratio..times.+PE.on.BSE+Gross.Margin.Ratio+
+              Debt.to.equity.ratio+Day.sales.in.inventory.ratio,data = train_data,
+            family = binomial)
```

> summary(log.reg4)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.0921483	0.271244	-7.713	1.23E-14	***
PAT.as...of.net.worth	-0.0683922	0.010081	-6.784	1.17E-11	***
Current.liabilities...provisions	0.0015038	0.000458	3.281	0.001035	**
Cumulative.retained.profits	-0.0126703	0.002664	-4.757	1.97E-06	***
TOL.TNW	0.085587	0.018647	4.59	4.44E-06	***
Total.term.liabilities...tangible.net.worth	-0.1776988	0.041002	-4.334	1.46E-05	***
Contingent.liabilities...Net.worth....	0.000757	0.000713	1.062	0.288141	
Contingent.liabilities	-0.0009327	0.000464	-2.009	0.04458	*
Net.fixed.assets	0.0002147	0.000117	1.828	0.067519	.
Current.assets	-0.0010879	0.000325	-3.344	0.000827	***
Debt.to.equity.ratio..times.	0.1707166	0.031779	5.372	7.79E-08	***
PE.on.BSE	-0.0758828	0.027051	-2.805	0.005029	**
Gross.Margin.Ratio	0.0772978	0.028007	2.76	0.005781	**
Debt.to.equity.ratio	-0.020263	0.011167	-1.815	0.069587	.
Day.sales.in.inventory.ratio	0.0097099	0.003608	2.691	0.007115	**

The variables with significance are measured as the important reflector of the credit default risk companies affected in the variables. The measure of insignificant variables is removed from the analysis and the logistics model is built once again to fit the variables.

```
> log.reg5=glm(Default...1~PAT.as...of.net.worth+Current.liabilities...provisions+
+              Cumulative.retained.profits+TOL.TNW+Total.term.liabilities...tangible.net.worth+
+              Contingent.liabilities+Net.fixed.assets+
+              Current.assets+Debt.to.equity.ratio..times.+PE.on.BSE+Gross.Margin.Ratio+
+              Debt.to.equity.ratio+Day.sales.in.inventory.ratio,data = train_data,
+            family = binomial)
> summary(log.reg5)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.0717688	0.270465	-7.66	1.86E-14	***
PAT.as...of.net.worth	-0.0689116	0.010075	-6.84	7.94E-12	***
Current.liabilities...provisions	0.0014588	0.000456	3.197	0.00139	**
Cumulative.retained.profits	-0.0127499	0.002676	-4.765	1.88E-06	***
TOL.TNW	0.0872511	0.018467	4.725	2.30E-06	***
Total.term.liabilities...tangible.net.worth	-0.1780281	0.04072	-4.372	1.23E-05	***
Contingent.liabilities	-0.0007614	0.000413	-1.842	0.06548	.
Net.fixed.assets	0.0001998	0.000116	1.728	0.08395	.
Current.assets	-0.0010634	0.000326	-3.266	0.00109	**
Debt.to.equity.ratio..times.	0.1724242	0.031662	5.446	5.16E-08	***
PE.on.BSE	-0.0766271	0.027081	-2.83	0.00466	**
Gross.Margin.Ratio	0.0760328	0.028334	2.683	0.00729	**
Debt.to.equity.ratio	-0.0178668	0.010745	-1.663	0.09634	.
Day.sales.in.inventory.ratio	0.009386	0.003596	2.61	0.00905	**

The variables are significant and the model fit is observed with the variance inflation factor and the higher inflation rate is removed for the LR model in further to make the important variables.

> car::vif(log.reg5)

PAT.as...of.net.worth	Current.liabilities...provisions
1.185514	8.256895
Cumulative.retained.profits	TOL.TNW
1.117648	4.341815
Total.term.liabilities...tangible.net.worth	Contingent.liabilities
5.826933	1.661877
Net.fixed.assets	Current.assets
1.946158	7.52947
Debt.to.equity.ratio..times.	PE.on.BSE
5.505472	1.046153
Gross.Margin.Ratio	Debt.to.equity.ratio
1.299824	1.66998
Day.sales.in.inventory.ratio	
1.022709	

The variance inflation factor of the variables are measured for the significant variables in the model developed. The variables with variance above 5 are removed from the further analysis to get the better model in the default credit status.

```
> log.reg6=glm(Default...1~PAT.as...of.net.worth+
+              Cumulative.retained.profits+TOL.TNW+
+              Contingent.liabilities+Net.fixed.assets+
+              PE.on.BSE+Gross.Margin.Ratio+
+              Debt.to.equity.ratio+Day.sales.in.inventory.ratio,data = train_data,
+            family = binomial)
```

The model is developed for the less variance inflation factor variables. The variables are checked for the significance of p values and the intercept of the default credit status.

> summary(log.reg6)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.99E+00	2.75E-01	-7.261	3.83E-13	***
PAT.as...of.net.worth	-6.43E-02	1.04E-02	-6.214	5.16E-10	***
Cumulative.retained.profits	-1.39E-02	2.68E-03	-5.208	1.91E-07	***
TOL.TNW	9.31E-02	1.01E-02	9.219	< 2e-16	***
Contingent.liabilities	-3.72E-04	3.77E-04	-0.987	0.3238	
Net.fixed.assets	2.42E-04	9.56E-05	2.529	0.01142	*
PE.on.BSE	-7.99E-02	2.68E-02	-2.978	0.0029	**
Gross.Margin.Ratio	5.86E-02	5.47E-02	1.07	0.28452	
Debt.to.equity.ratio	-7.17E-03	1.03E-02	-0.696	0.48614	
Day.sales.in.inventory.ratio	1.01E-02	3.50E-03	2.879	0.00399	**

The higher inflation rate variables are removed and the significant variable contingent liabilities, gross margin ratio and debt equity ratio is insignificant from the model. The model is further developed with the logistic model for the default status of the companies.

```
> log.reg7=glm(Default...1~PAT.as...of.net.worth+
+              Cumulative.retained.profits+TOL.TNW+
+              Net.fixed.assets+
+              PE.on.BSE+
+              Day.sales.in.inventory.ratio,data = train_data,
+            family = binomial)
```

The model is developed from the significant variables from the previous model of the logistics model build.

```
> summary(log.reg7)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.96E+00	2.62E-01	-7.463	8.44E-14	***
PAT.as...of.net.worth	-6.50E-02	1.03E-02	-6.304	2.89E-10	***
Cumulative.retained.profits	-1.39E-02	2.65E-03	-5.225	1.74E-07	***
TOL.TNW	8.91E-02	8.33E-03	10.7	< 2e-16	***
Net.fixed.assets	1.87E-04	8.27E-05	2.263	0.02361	*
PE.on.BSE	-7.92E-02	2.66E-02	-2.975	0.00293	**
Day.sales.in.inventory.ratio	9.91E-03	3.50E-03	2.836	0.00456	**

The significant variables are measured for the variables PAT as of net worth, cumulative retained profits, TOL, net fixed assets, PE on BSE, Day sales in inventory ratio.

The significant variables are measured for the variance inflation factor to check the fit of the model in the default credit status.

```
> car::vif(log.reg7)
```

PAT.as...of.net.worth	Cumulative.retained.profits	TOL.TNW
1.124491	1.120875	1.016318
Net.fixed.assets	PE.on.BSE	Day.sales.in.inventory.ratio
1.012056	1.041306	1.009711

The variance inflation factor is measured with good fit model and the model is further used in prediction of the train dataset and validation dataset. The model is developed with good fit of inflation factor.

2.2 Analyze coefficient & their signs

```
> exp(coef(log.reg7))
```

(Intercept)	PAT.as...of.net.worth	Cumulative.retained.profits
0.1413645	0.9370483	0.9862309
TOL.TNW	Net.fixed.assets	PE.on.BSE
1.0931769	1.0001871	0.9238188
Day.sales.in.inventory.ratio		
1.009963		

The exponential values of the model is identified by the values are intercepted by 0.14 as the variables are mostly related in the linear function of the model. The model is developed in the function of the exponential values which derives the basic relationship between the variables.

> exp(coef(log.reg7))/(1+exp(coef(log.reg7)))

(Intercept)	PAT.as...of.net.worth	Cumulative.retained.profits
0.1238557	0.4837506	0.4965339
TOL.TNW	Net.fixed.assets	PE.on.BSE
0.5222573	0.5000468	0.4802005
Day.sales.in.inventory.ratio		
0.5024784		

The exponential of the coefficients are measured for the model and the model validates is identified by the very less values of the intercept is 0.12 in the variables for the measure of significant variables and the model is validated for the original model.

> pscl::pR2(log.reg7)["McFadden"]

fitting null model for pseudo-r2
McFadden
0.3622924

The McFadden value of the good fit model is examined for 0.3622924 which is measured as good fit of the logistics model in the pseudo-r squared values. The McFadden values are basic definition of the good model and the values are measured with significant variables.

> logLik(log.reg7)

'log Lik.' -564.6815 (df=7)

The likelihood of the measured model is marked degrees of freedom 7 which is measured with the good fit of the model. The model is classified by the credit default model where the measurement of the likelihood model is measured for the gaussian methodology is validated with the good fit of the model.

- The coefficients of the model is described by the p value, z value, standard error of the variables and estimated values for the variables. The variables are varied I the significance level of the measured variables in the model.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.96E+00	2.62E-01	-7.463	8.44E-14	***

- The Intercept of the model is defined for the variables measured in the values interpreted with estimate of very less values predicted for the variables and the z value is measured for -7.4 where the p value is very less and the compared significant variables are measure in the good fit of the interception in standard error of 2.62 was increased for the model developing measures in the fitted values.

	Estimate	Std. Error	z value	Pr(>|z|)	
PAT.as...of.net.worth	-6.50E-02	1.03E-02	-6.304	2.89E-10	***

- The variable Profit After Tax as of the net worth in the estimation of the values is very less very on the variables for the standard error which is confined by the z value in the -6.3 in the values are controlled by the very less p factor for the good model built on the logit regressions.

	Estimate	Std. Error	z value	Pr(>|z|)	
Cumulative.retained.profits	-1.39E-02	2.65E-03	-5.225	1.74E-07	***

- Variable cumulative retained profits is measured for the positive estimation in the models and the values are measured around for the z values and the P value is significant of the very less values and the measured values are increased in the fitness of the model.

	Estimate	Std. Error	z value	Pr(>|z|)	
TOL.TNW	8.91E-02	8.33E-03	10.7	< 2e-16	***

- The Total Liabilities and the ratio of the total net worth of the assets are ratio of the measured values are estimated factors are measured in the increased factor values is 8.9 and the values are increased for the standard error of the increased values and the values are measured with the various values are increased for the z value is 10.7 and the very less values are measured for the good fit in the values are increased in the less values in P value is very less and contributes more in good fit.

	Estimate	Std. Error	z value	Pr(>|z|)	
Net.fixed.assets	1.87E-04	8.27E-05	2.263	0.02361	*

- The net fixed assets are measured for the increased p value which is positive values for the intercepts and the coefficient of the variables are measured with 0.02 p value contributes for the good fit model and the significance level is increased by the ratio of the company size default companies.

	Estimate	Std. Error	z value	Pr(>|z|)	
PE.on.BSE	-7.92E-02	2.66E-02	-2.975	0.00293	**

- Market sentiment of the companies are measured for the p values is observed for the increased significance of the variables in the p value is higher with the increased estimation of the negative less values.

	Estimate	Std. Error	z value	Pr(>|z|)	
Day.sales.in.inventory.ratio	9.91E-03	3.50E-03	2.836	0.00456	**

- The day sales in inventory ratio is measured for the efficiency of the variables is increased for the values are observed with p values is very higher positive values for the significant values in the measured good fit.

3. Model Performance Measures

3.1 Predict accuracy of model on dev and validation datasets

```
> log.reg=predict.glm(log.reg7,train_data,type = "response")
```

The model is predicted for the dataset in the values is measured for the type response in the original datasets and the values are increased for the fitted generalized linear model.

```
> tab.train=table(train_data$Default...1,log.reg>0.3)
```

The table is created for the default credit model in the prediction of the model is above 3% of the predicted values in the model.

```
> tab.train
```

	FALSE	TRUE
0	3245	53
1	136	107

The table shows the variables are measured for the False and True values are considered for the true false positive rate and the true negative values are measured for the positive values and negative values in the default companies.

```
> sum(diag(tab.train))/sum(tab.train)
```

[1] 0.9466252

The accuracy of the defaulted companies in the variables are measured for the 95% in the predicted logistics model and the values are increased for the train dataset in the values. The values are increased for the good fit model and the precisions are measure for the 64% of the model. The original datasets are controlled in the variables of the increased fit of the model.

- The Defaulted companies will have negative impact of the credit structures from the borrowers of the insurance of the values.

```
> pred1=ifelse(log.reg>0.68,0,1)
```

The predicted values are measured for the values are observed for the threshold companies is 0.68 with the default status and non-default status in the companies.

```
> actual1=train_data$Default...1
```

The actual values is taken with the default status of the companies in the train datasets in the variables are increased for the values in the variables.

```
> cm_log_train=confusionMatrix(as.factor(pred1),actual1,positive = "1")
```

The confusion matrix is performed for the actual and predicted values for the positive value 1 and the confusion matrix is developed for the accuracy measures of the observed threshold above 0.68 and the values are increased for the trend of the values in specificity, sensitivity.

> cm_log_train

Prediction	0	1
0	15	47
1	3283	196

Accuracy	: 0.0596
95% CI	: (0.052, 0.0679)
No Information Rate	: 0.9314
P-Value [Acc > NIR]	: 1
Kappa	: -0.0263
Mcnemar's Test P-Value	: <2e-16
Sensitivity	: 0.806584
Specificity	: 0.004548
Pos Pred Value	: 0.056338
Neg Pred Value	: 0.241935
Prevalence	: 0.068625
Detection Rate	: 0.055352
Detection Prevalence	: 0.982491
Balanced Accuracy	: 0.405566
'Positive' Class	: 1

The sensitivity and specificity of the model is measured for the 80% and 0.004 respectively and values is measure for the Kappa values is very less in the predicted for the accuracy of 59% in the true positive values and true false positive values in the confusion matrix.

> log.reg.pred=predict.glm(log.reg7,test_data,type = "response")

The prediction is measured for the original dataset and the model is validated with the validation datasets in the variables for the type of the predict is response.

> table.log=table(test_data$Default...1,log.reg.pred>0.3)

The table is created for the predicted model in the threshold above 0.3 and the prediction of the model in validation data is measured for the True Positive Factor and Ture Negative Factor in the datasets. The datasets are measured for the defaulted companies in the prediction of the model.

> table.log

	FALSE	TRUE
0	641	20
1	23	31

The true positive value of the default case is measured for the 31 values in the defaulted companies. The credit is measured for the 31 and returns the company's credit money of the variables. The values are measured for accuracy of the credit lenders.

> sum(diag(table.log))/sum(table.log)

[1] 0.9398601

94% of the validation dataset is measured for the variables in the significant model and the values are measured in the validation data and the default status of the companies is marked for the increased values in the values are counted for the accuracy of the model. In which the model is best predicted for the train dataset and the accuracy of the validation data reflects the values are increased for the fit of the model.

> pred=ifelse(log.reg.pred>0.68,0,1)

The predicted logistics model is measured for the threshold of 0.68 in the values.

> actual=test_data$Default...1

The actual values of the defaulted companies is measured for the datasets in the values.

> cm_log=confusionMatrix(as.factor(pred),actual,positive = "1")

The confusion matrix of the predicted model is validated with the validation of the datasets in the values are measured for the accuracy, sensitivity, specificity of the variables in the validation of the defaulted cases. The confusion matrix is cross validated in the related variables of the measures.

> cm_log

	Reference	
Prediction	0	1
0	4	13
1	657	41

Accuracy	: 0.0629
95% CI	: (0.0463, 0.0833)
No Information Rate	: 0.9245
P-Value [Acc > NIR]	: 1
Kappa	: -0.0362
Mcnemar's Test P-Value	: <2e-16
Sensitivity	: 0.759259
Specificity	: 0.006051
Pos Pred Value	: 0.058739
Neg Pred Value	: 0.235294
Prevalence	: 0.075524
Detection Rate	: 0.057343
Detection Prevalence	: 0.976224
Balanced Accuracy	: 0.382655
'Positive' Class	: 1

The accuracy of the model is increased by 62% and makes the validation datasets is predicted for good fit model. The Kappa value is measured for -0.03 for the validated dataset.

3.2 Sort the data in descending order based on probability of default and then divide into 10 docile based on probability & check how well the model has performed

```
> train_data$pred=predict(log.reg7,train_data,type = "response")
```

The predicted values are created with object pred for the further interpretation of models in the decile creations for the train and test datasets.

```
> decile = function(x){
+   deciles = vector(length=10)
+   for (i in seq(0.1,1,.1)){
+     deciles[i*10] = quantile(x, i, na.rm=T)
+   }
+   return (
+     ifelse(x<deciles[1], 1,
+            ifelse(x<deciles[2], 2,
+                   ifelse(x<deciles[3], 3,
+                          ifelse(x<deciles[4], 4,
+                                 ifelse(x<deciles[5], 5,
+                                        ifelse(x<deciles[6], 6,
+                                               ifelse(x<deciles[7], 7,
+                                                      ifelse(x<deciles[8], 8,
+                                                             ifelse(x<deciles[9], 9,
10
+                                                             ))))))))))
```

The decile function is created for the original model predicting with the train datasets. The model is predicted for the descending order in the association of rank ordered created values for the variables.

```
> train_data$deciles=decile(train_data$pred)
```

The decile function is further predicted with the train dataset for the rank ordered list of creating variables in the model.

```
> m=data.table::data.table(train_data)
```

The train data is created into further of data table with the creation of the rank ascending variables.

```
> rank.default = m[, list(cnt=length(Default...1),
+                         cnt_resp=sum(Default...1==1),
+                         cnt_non_resp=sum(Default...1==0)
+                         ), by=deciles][order(-deciles)]
```

The rank ordered variables are measured for the model in the prediction of the dataset which is measured for the data table format of the variables in the train data.

```
> rank.default$rrate=round(rank.default$cnt_resp/rank.default$cnt,4)
> rank.default$cum_resp=cumsum(rank.default$cnt_resp)
> rank.default$cum_non_resp=cumsum(rank.default$cnt_non_resp)
>
rank.default$cum_rel_resp=round(rank.default$cum_resp/sum(rank.default$cnt_non_resp),4)
>rank.default$cum_rel_non_resp=round(rank.default$cum_non_resp/sum(rank.default$cnt_non_resp),4)
> rank.default$ks=abs(rank.default$cum_rel_resp - rank.default$cum_rel_non_resp)*100
> rank.default$rrate=scales::percent(rank.default$rrate)
> rank.default$cum_rel_resp=scales::percent(rank.default$cum_rel_resp)
> rank.default$cum_rel_non_resp=scales::percent(rank.default$cum_rel_non_resp)
```

The rank of the deciles are created for the default cases in the variables and the gain of the rank is measured for the deciles in the train model.

```
> train_data_rank=rank.default
```

The train data rank is created for the rank in the default cases in the variables are measured for the defaulted cases in the model.

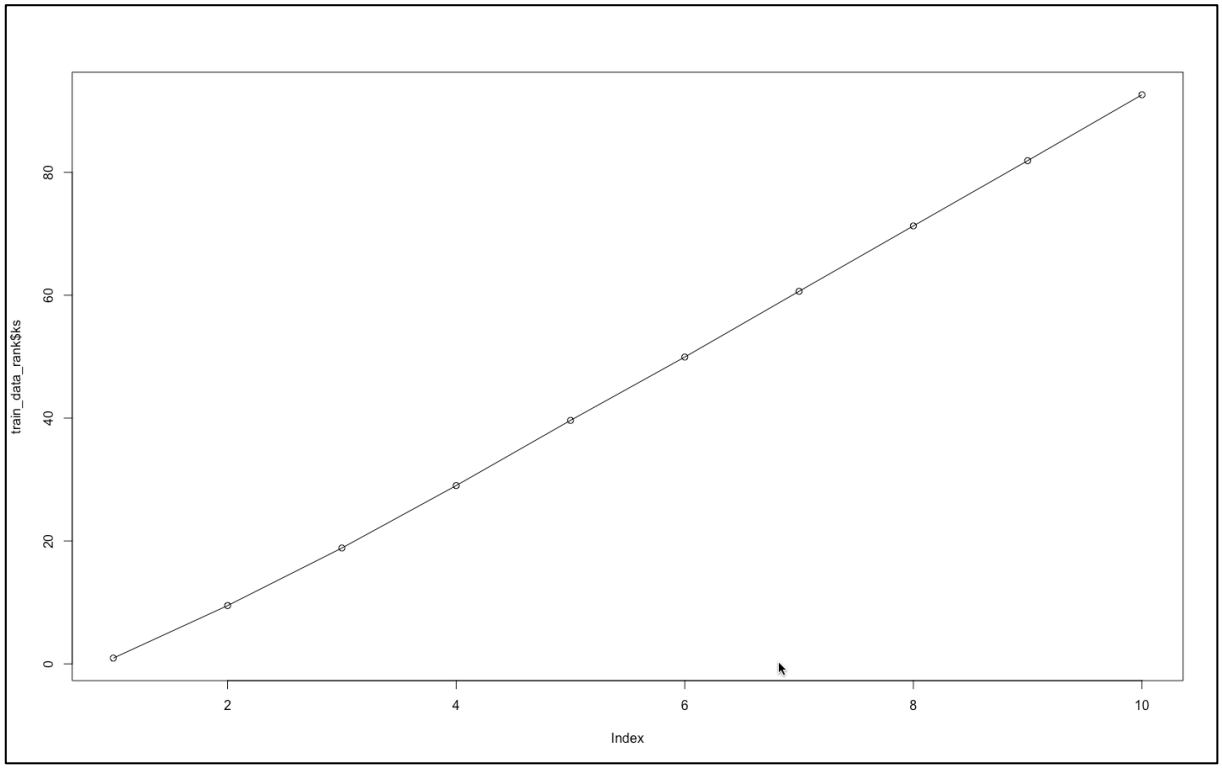
```
> print(train_data_rank)
```

	deciles	cnt	cnt_resp	cnt_non_resp	rrate	cum_resp	cum_non_resp	cum_rel_resp	cum_rel_non_resp	ks
1	10	355	162	193	45.63%	162	193	4.91%	5.80%	0.94
2	9	354	36	318	10.17%	198	511	6.00%	15.50%	9.49
3	8	353	22	331	6.23%	220	842	6.67%	25.50%	18.86
4	7	355	10	345	2.82%	230	1187	6.97%	36.00%	29.02
5	6	354	2	352	0.56%	232	1539	7.03%	46.70%	39.63
6	5	354	7	347	1.98%	239	1886	7.25%	57.20%	49.94
7	4	353	0	353	0.00%	239	2239	7.25%	67.90%	60.64
8	3	355	2	353	0.56%	241	2592	7.31%	78.60%	71.28
9	2	354	2	352	0.56%	243	2944	7.37%	89.30%	81.90
10	1	354	0	354	0.00%	243	3298	7.37%	100.00%	92.63

The table shows the descending order of the rank deciles in the variables created for the cumulative response, non-cumulative response, return rate, counted response, counted non response, cumulative relative response, cumulative non relative response to the KS chart values. The cumulative relative response is achieved the value of 4.91% in the model fitted values. Values which are controlled in the rank 1 has highest cumulative non-responds in the value of 3298 with the values are mentioned as 100% of the credit risk defaulters for the model predicted. The KS Values will plotted for the accuracy of the charts in the predicted model.

> plot(train_data_rank$ks)

> lines(train_data_rank$ks)



The plot displayed the values are increased for the decile ranks in the predicted model of the variables. The KS plot are measured for the increased rank for the increased values in the default predictions.

> test_data$pred=predict(log.reg7,test_data,type = "response")

The predicted model is built for the test dataset in the validation of the model.

```
> decile = function(x){
+   deciles = vector(length=10)
+   for (i in seq(0.1,1,.1)){
+     deciles[i*10] = quantile(x, i, na.rm=T)
+   }
+   return (
+     ifelse(x<deciles[1], 1,
+            ifelse(x<deciles[2], 2,
+                   ifelse(x<deciles[3], 3,
+                          ifelse(x<deciles[4], 4,
+                                 ifelse(x<deciles[5], 5,
+                                        ifelse(x<deciles[6], 6,
+                                               ifelse(x<deciles[7], 7,
+                                                      ifelse(x<deciles[8], 8,
+                                                             ifelse(x<deciles[9],
9, 10
+                                                             ))))))))))
+ }
```

The deciles are created for the descending rank orders in the validation datasets are created.

```
> test_data$deciles=decile(test_data$pred)
```

The deciles are created for the validation of the predicted validation datasets are generated for the deciles in predicting the model.

```
> n=data.table::data.table(test_data)
```

The object is created for the data table formatted validation datasets in the deciles for the rank order creating.

```
> rank.test.default = n[, list(cnt=length(Default...1),
+                              cnt_resp=sum(Default...1==1),
+                              cnt_non_resp=sum(Default...1==0)
+ ), by=deciles][order(-deciles)]
```

The ranks are created for the decile prediction in the variables and the ranked variables are measured for the deciles in descending the ordered values of the variables.

```
> rank.test.default$rrate=round(rank.test.default$cnt_resp/rank.test.default$cnt,4)
> rank.test.default$cum_resp=cumsum(rank.test.default$cnt_resp)
> rank.test.default$cum_non_resp=cumsum(rank.test.default$cnt_non_resp)
> rank.test.default$cum_rel_resp=round(rank.test.default$cum_resp/
+                                      sum(rank.test.default$cnt_non_resp),4)
> rank.test.default$cum_rel_non_resp=round(rank.test.default$cum_non_resp/
+                                          sum(rank.test.default$cnt_non_resp),4)
>                          rank.test.default$ks=abs(rank.test.default$cum_rel_resp                    -
rank.test.default$cum_rel_non_resp)*100
> rank.test.default$rrate=percent(rank.test.default$rrate)
> rank.test.default$cum_rel_resp=percent(rank.test.default$cum_rel_resp)
> rank.test.default$cum_rel_non_resp=percent(rank.test.default$cum_rel_non_resp)
```

The ranked variables are created for the variables that impact the variables in the creation of the datasets.

```
> test_data_rank=rank.test.default
```

The variables is created for the variables created in the default cases of the company on the variables. The variables are measured for the respondents in the KS chart values.

> print(test_data_rank)

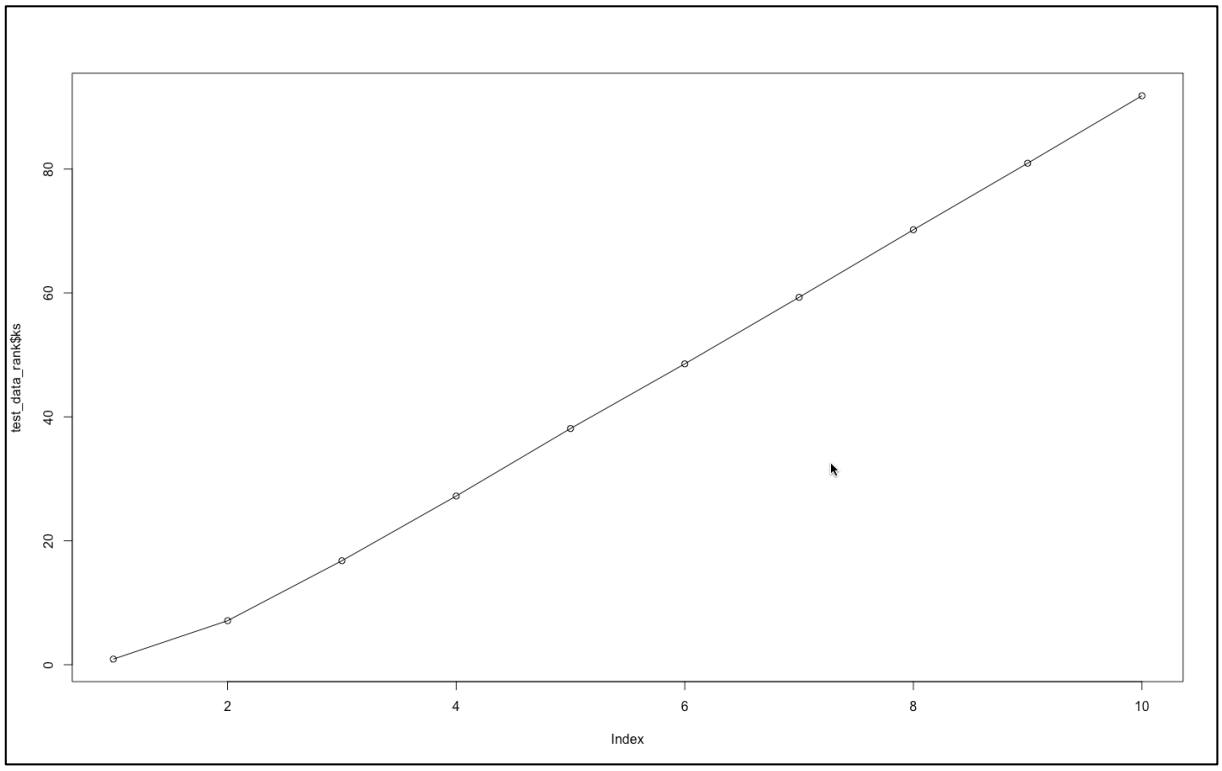
	deciles	cnt	cnt_resp	cnt_non_resp	rrate	cum_resp	cum_non_resp	cum_rel_resp	cum_rel_non_resp	ks
1	10	72	39	33	54.20%	39	33	5.90%	5.00%	0.91
2	9	71	9	62	12.70%	48	95	7.26%	14.40%	7.11
3	8	72	4	68	5.60%	52	163	7.87%	24.70%	16.79
4	7	71	1	70	1.40%	53	233	8.02%	35.20%	27.23
5	6	72	0	72	0.00%	53	305	8.02%	46.10%	38.12
6	5	71	1	70	1.40%	54	375	8.17%	56.70%	48.56
7	4	71	0	71	0.00%	54	446	8.17%	67.50%	59.3
8	3	72	0	72	0.00%	54	518	8.17%	78.40%	70.2
9	2	71	0	71	0.00%	54	589	8.17%	89.10%	80.94
10	1	72	0	72	0.00%	54	661	8.17%	100.00%	91.83

The deciles are created for the counts in the rank, counted respondents, count non-responds, r rate, cumulative responds, cumulative non responds, cumulative relative non-responds and the KS charts for the validation of datasets in the variables. The deciles are measure for the values and the KS chart are compared with the train dataset for the difference values in the relations of the values are controlled in the measures.

The r rate is increasing from the ranks and the increased rank is reflected with the consistent cumulative relative responds are measured in the first five ranks of the deciles. The decile ranks are measured for the KS chart.

> plot(test_data_rank$ks)

> lines(test_data_rank$ks)



The KS chart is explained for the increased trends in the values are measured form the values and increased with frequency of the values.

> rocpred=prediction(log.reg.pred,test_data$Default...1)

The ROC Prediction is predicted for the validation datasets and the measures are understand by the values in the fit model.

> ROC=pROC::roc(Default...1,log.reg7$fitted.values)

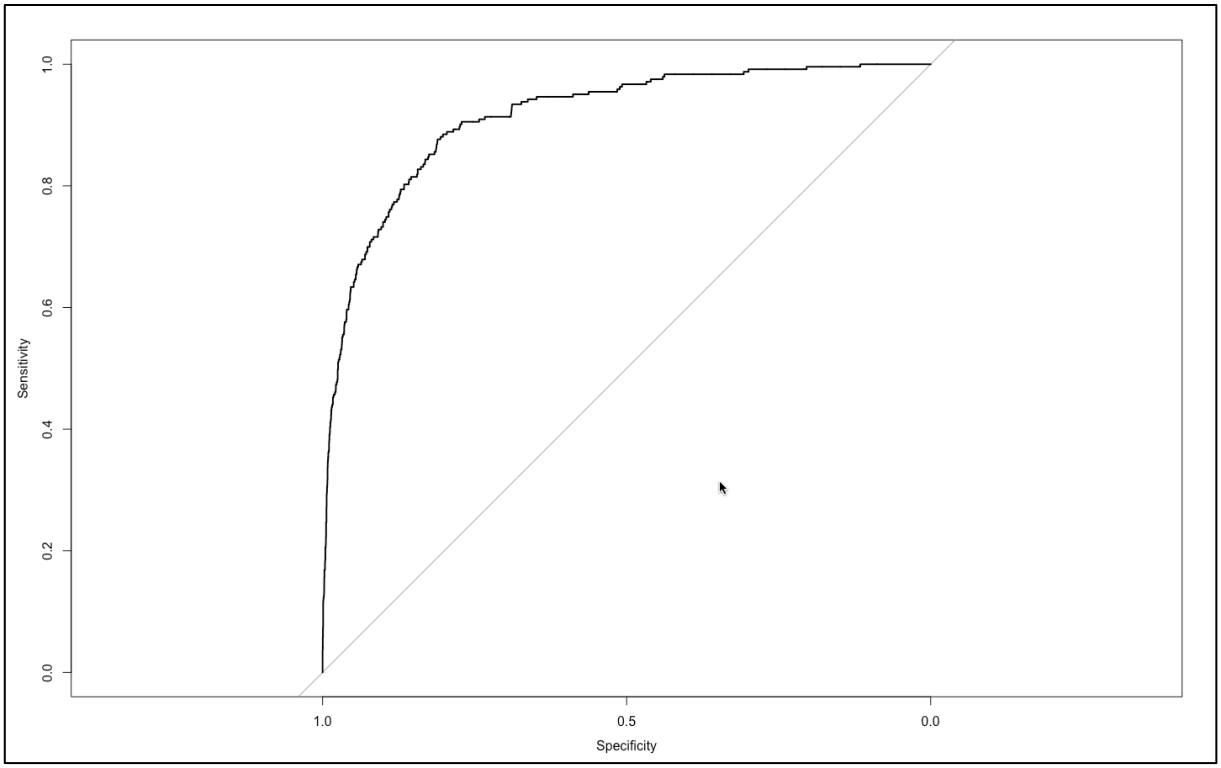
The ROC is predicted for the charts with variables plots in the values are increased for the fitted values in the logistics regression of the increased values with the defaulted status of the companies.

> ROC

Area under the curve: 0.9139

The ROC predicted for the area under the curve is about 91% in the validation dataset. The AUC values are predicted for the variables with significant predicted values.

> plot.roc(Default...1,log.reg7$fitted.values)



The ROC plots the values are measured for the variables in the validation datasets and the values are measured for the Area Under the Curve for the predicted values from 0 to 10 in increased plots for the curve. The curve is diverged for the single line which shows the area is under the curve with 91% of the defaulted cases are presented for the companies.

> AUC=as.numeric(performance(rocpred, "auc")@y.values)

> AUC

[1] 0.9468258

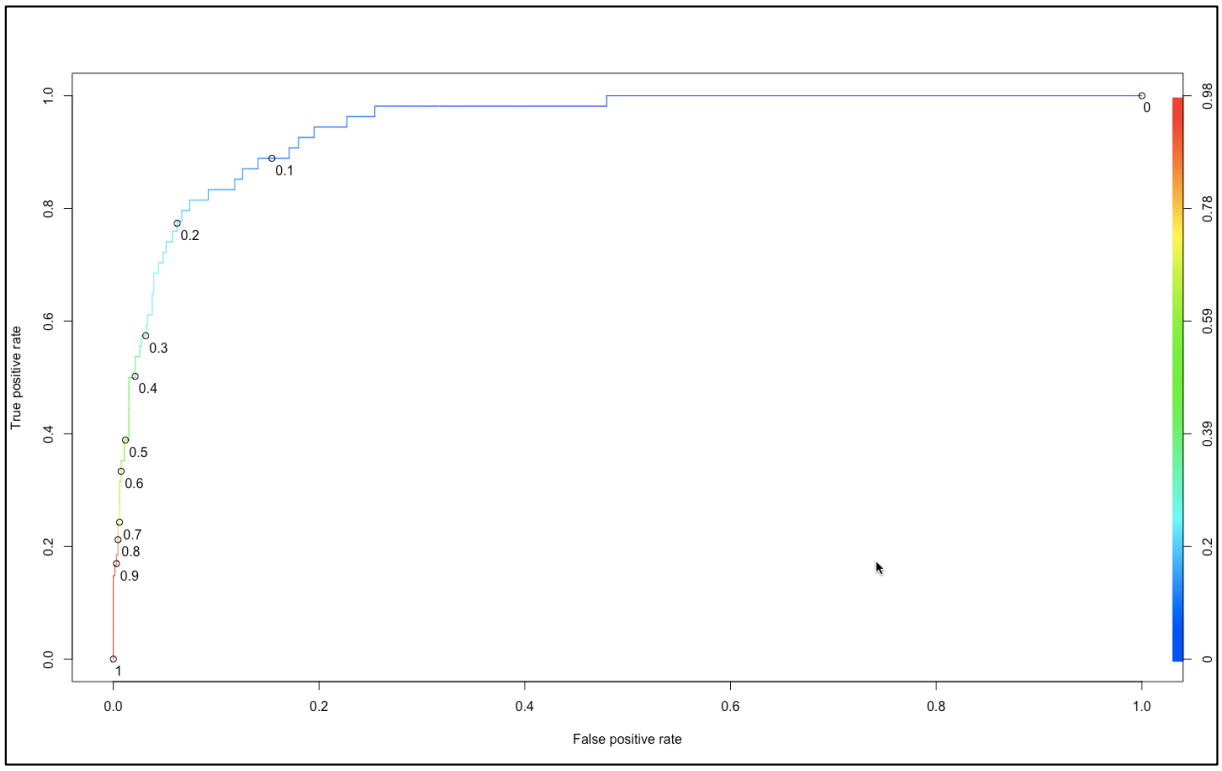
The area of the curve is measured for the validation datasets in the variables are measured of 94% in the validated datasets.

```
> KS=max(attr(perf, 'y.values')[[1]]-attr(perf, 'x.values')[[1]])
> KS
```

[1] 0.7492856

The KS values is determined by the value of 74% in the validation datasets in the variables are measured for the increased curve in the plotted fit values.

```
> plot(perf, colorize =T, print.cutoffs.at= seq(0, 1, .1), text.adj = c(-.2, 1.7))
```



The plot demonstrate the values are measured for the relation of the variables in the controlled values for the demonstration of the increased values to the particular intense variables of the associated variables for creating the increased frequency of the default cases in the company.

```
> gini=ineq(log.reg.pred, type="Gini")
> gini
```

[1] 0.7506139

The Gini values of the validation datasets are predicted for the 75% in the variables and the values are demonstrated for the Gini values.

The models are created and the model performance measures are increased for the logistics model of the variables and the values are best fitted in the AUC and the values ranked in deciles are interpreting the values of the measures in the validation datasets.

3.5 Outlier Identification

The outlier identification for the temperature is found as the maximum values in the variables.

```
outlier=boxplot(my_train[,-c(1,2)],plot = FALSE)$out
print(outlier)
```

```
> print(outlier)
  [1] 17512.3  5669.1  3142.5  7394.3  9076.1  6474.7  5763.4  2956.1  2843.4  3245.7  3582.0  5350.3 10837.0  5131.4
 [15] 11876.4  4379.5  2965.4  3193.7  4291.3  4326.1  5731.8  5565.8  8022.9  4362.0  8183.1  4636.2  6429.7  7721.9
 [29] 17330.1  3240.6  3356.1  5385.9 10652.6  5577.4 10744.0  2760.3  5420.2  4553.0 20099.1  9182.6 11807.2  3771.1
 [43]  2885.0  2699.2  3155.2 14101.6  5977.0  4825.0  7011.9  3316.3 10335.2 10751.0  5164.9  7493.0  8461.7 17014.5
 [57]  8407.1  3080.6  7729.5  2822.6  3819.9  3222.6 12102.6  6883.3 15576.6  2724.7  3214.9  4012.8  3916.3  6989.1
 [71]  4391.6 12072.9  5352.3  2717.0  3099.9  9353.6  3073.3 11300.7 16431.9  5263.3 12259.9  3117.7  5226.6  3724.0
 [85] 13354.5  6112.9  6357.3  6634.6  5277.0  2776.5  4643.1  2981.9  6412.8  4231.9  3574.9  3075.9 13724.8  5481.7
 [99]  5382.8 16861.2  2851.4  3401.4  6360.8  2843.7  4018.9  8378.8  3239.7  3276.8  5775.0  9310.8 15375.2  4313.3
[113]  2662.5  4023.1  3537.3  9411.7  7488.9  8556.0  7014.2  3467.3 21411.6 10099.2  8993.2  2831.0  4189.8  3334.5
[127]  7363.7  2624.5 13013.5  4352.9  5481.1 12018.0  8597.9 10250.2  2643.8  6669.4  3196.1  2703.5  7299.4  3410.5
[141]  3886.1  8447.2  3785.9  2843.7  7756.7  4463.6 12933.4 10131.2  5968.0  3354.4  7025.3  7748.7  4295.2  3219.7
[155] 10768.6  2729.1  5320.3  3037.8 12569.9  6382.2  2629.0  4995.9  7997.0  5092.0  3544.6  5690.2  5433.6 20345.0
[169]  3001.3 15376.0 11284.9  9548.8  3021.0  3231.8  4136.3  3959.3  9508.0 11618.1  5041.4 12729.9 11705.4 19046.1
[183] 12912.5  6289.1  3669.6 19450.6  2771.4  2610.9 15120.4 11900.9  5397.1  9322.7  9275.6  3682.3  3378.5  3479.7
[197]  3292.8  5592.3  5782.4  9174.2 31417.0  9795.3 13241.2  9508.6 21552.1  6215.5  3411.8 19178.6  9017.5 16986.8
[211]  2713.6 11741.7  6644.6  7415.8  3133.4 25236.8  6368.9  4525.0 16535.0  4180.5  3948.7  7936.6  6662.1  5374.9
[225]  3032.9 12989.0  3733.7 21565.9  4038.0  8211.2  2680.9  6378.7  3647.0  4107.6  3098.1  3192.3 15225.7 25936.7
[239]  2801.4  5930.5  6326.3  5406.9  2615.9 16616.9  7731.7  6547.7  6168.0 12895.0  8397.6  4770.6  8369.2  5212.4
[253] 19735.0  3037.8  5166.3  5934.8  5073.5  9588.9  5682.3 11910.0  9090.0  3093.6  6799.4  4664.6  2817.4  2913.2
[267]  2963.1  4217.6 25221.9  3513.0 14311.6  4126.3  7875.7  2853.7  3309.1 12511.9  2652.3  3864.3  6454.1  4800.4
[281]  4606.5  4433.8 29229.0  4499.2 22225.0  5964.9  2992.5  4433.3 10170.3  6294.9  4130.0  2814.9  7601.3  7103.3
[295] 15688.1  3534.1  6340.7 11638.9  5112.2  2676.5  3300.8  3614.5  5909.5 23779.2  5573.0 15161.3  4719.4  9353.6
[309] 13952.5  5596.8 15223.4  2933.4  2743.6  3235.7  4465.6  3750.0  4491.5  4701.0  2622.9  8384.2  3222.7  2955.0
[323]  5135.5  3691.5  3441.5  3386.7 17085.1 13358.8  3923.4 13616.2  3869.4  3009.4  5620.4  8961.8  8452.9 11824.7
[337]  3121.8  2914.6  9319.9  3912.6 28671.9  3485.5 10378.5  4096.5  2835.6  3936.8  3224.2  2621.9  4894.2 16359.9
[351] 60132.0  5653.3  3853.5 66167.8  5747.5  4065.0  4747.0 47998.1 54654.0 25735.6 37982.5 45250.2  4253.4  3695.4
[365]  4689.1 44911.9 46799.4 69095.4 37345.8 342407.2 60312.3 54022.1 14718.2  3665.6 33520.7  4041.1 11153.9 36558.3
[379] 84791.3  7447.5  7369.2 12087.9  4415.3  6888.9 58188.0  6798.9 156112.7 161006.8 12757.3 14632.7  9542.2  3592.1
[393] 26127.1  6883.5  4575.3  5806.7 32654.9  6077.3  5701.4 70325.5  2642.8 42463.4 250277.1 351433.6 318611.7 75319.4
[407] 87602.0  7785.8  4444.5  3040.6 152522.7 47628.8 12006.6 118940.0 1028927.7  9739.8 23511.6 132044.6 111972.1 15318.7
[421]  4130.9 64746.9 76375.3  3438.1  7233.3 15149.3 20090.8  6894.0  5770.7 15687.0 29792.3  5950.0  9981.1  2983.0
[435] 93634.1  3078.3 44197.0 24626.0  2904.6  5924.6 36186.3  3826.0 112637.2 37173.0  6833.5  3690.8  3039.7  2958.9
[449]  6213.1  3681.6  9622.2  6612.6 69760.6  5911.0 71126.0 32927.2 129320.7  5337.2 1176509.2 40612.3  4601.1  5435.2
[463]  3836.9  6505.0  4095.3 33592.2 354727.3 90428.2  9116.7 41522.2  7126.4 48113.8 196067.9  4196.3 321314.6  3995.8
```

The variables are identified for the outliers in the datasets and the outliers are treated in the exploratory analysis of the datasets.

```
> outlier1=boxplot(my_test[,-c(1,2)],plot = FALSE)$out
```

```
> print(outlier1)
```

```
> print(outlier1)
  [1] 354727.30 318611.70  15375.20  21552.10   5977.00   3193.70 118940.00   3886.10  12912.50   6505.00   3219.70   3334.50
 [13]  21411.60  12259.90  32927.20  41522.20  11807.20   3222.70   3222.60   7233.30   3354.40  47628.80 321314.60   4041.10
 [25]   3544.60   4601.10  10170.30  10768.60   8452.90  11705.40   3316.30  23779.20   4719.40   3214.90  45250.20  46698.10
 [37]   6612.60  29792.30  54022.10  12102.60  15223.40   3826.00   3410.50   9116.70   5731.80   5481.70   6894.00  58188.00
 [49]   9174.20   4295.20   8378.80   3441.50   5690.20   7447.50  20099.10   3819.90   3513.00   7363.70   3276.80  76375.30
 [61]  12087.90   5263.30   4664.60   4433.80   8183.10   4253.40   7103.30  17512.30  11638.90  23511.60   3467.30   4196.30
 [73]   6168.00  12006.60   4553.00   5164.90   3582.00   4701.00   6077.30  32654.90   5592.30   5342.60   9322.70   6357.30
 [85]  14718.20  44911.90  25936.70   6883.30   6669.40 250277.10   5682.30  13358.80 156112.70  16359.90   3537.30   4012.80
 [97]   6382.20   1683.50 171840.00  95986.50   2138.90   1094.00   8539.00   2016.20  36300.70   2585.50   1431.80   1170.50
[109]   4066.00   1157.60   4572.50   4876.70   1681.30  14473.00  24481.20   5417.80   1679.80   1901.60   2734.30  13934.40
[121] 138936.20   1831.30   1080.00   1593.70   5150.10   3825.30   2874.80  11899.50  18472.00  17637.80   2620.20   8826.10
[133]  12145.10   5365.70   3254.00   1679.30   1821.80   1352.10   3488.60   1377.30   1440.10   1153.30  43733.60   5846.10
[145]   1139.70   1786.90   5616.80   2931.90   2083.40   2632.40   2871.70   1451.60   1448.10   1394.20   2481.80  24924.90
[157]   6335.80   1354.70   1725.70   2348.10   1838.60   2687.80   7093.20   3126.80   3907.20   1493.80   1583.70   3756.80
[169]   3170.00   2836.70   3166.40   2461.30   1523.70  13652.10   1735.20   3713.40   3233.90   6567.00  28495.10   9486.00
[181]   1465.50   1369.20   5177.20   1979.20 124148.70   1350.80   1855.10  29951.30   5362.40   1273.20   1640.60   2609.40
[193] 408220.70 1028087.40  22760.70   5124.70   5455.00   8302.20   4685.70 105892.90   8172.30  21739.70   6976.90   6143.00
[205]  11386.80  12467.70  36623.30  24723.90  10263.50   4798.70   4946.00  10258.10   4119.40   7078.70  26230.40 203440.90
[217]   6501.80   5647.20  10416.40   7294.90   7519.80   4640.60  18406.60  14678.80   7081.70  86258.60  55455.00   7299.40
[229]  89589.00  42822.60   6430.20   6468.70   4628.80   4017.20   5434.90   8188.20  11097.60  43752.80   8276.30   5013.50
[241]   9350.10   6155.10   8476.30   6212.10   8941.70   4290.20  39810.70   9426.90   6424.80   5395.30   5652.00  24965.20
[253]   9123.00  30136.30   4555.50  11072.50   4660.20   4229.60   7075.10  36476.00   5590.10   5751.40  11531.60  10929.70
[265]  18080.70  28236.60   4902.90  18901.20  26362.40   8057.80 202629.00   6075.00  33862.30   4843.70   7348.00   4599.00
[277]    -28.10     56.60     64.80   2852.10   2435.50     69.50   -248.70    105.90   -228.10    -82.10   7540.00    123.60
[289]    392.90    -44.90    -50.60    -96.50     48.90     94.00     39.80    -96.10   -339.50     73.50     66.90    248.60
[301]   3405.70    414.10    -41.10    -71.50   -153.10   -441.40    309.50    -88.50   -116.40     52.90     64.50     45.10
[313]    824.70     47.70    -63.90    -42.30    124.10     64.60    -70.80     62.30     86.10     45.00     70.80    187.30
[325]     43.00    -32.80     42.10     49.00    -56.40    164.20    182.20     67.50    265.10    489.80     50.00     43.20
[337]   -150.70   -307.60     61.00    -40.30    157.80    550.30    951.70     78.80    130.00    124.90     95.20    -55.80
[349]     52.10   -467.20    -39.00   -356.20    -28.60    -32.00    -30.60   -488.10    195.40    232.40    276.00    -65.00
[361]     64.10     47.30     76.80    -38.70    201.00     54.50    -35.00    -30.20    -29.40    -60.40    -55.70    170.30
```

3.6 Variable Transformation/ Feature creation

The variables transformation are created with the categorical variables and the numeric variables are diversified for the various prediction in the customer churn.

```
> my_train$Creditors.turnover=as.numeric(my_train$Creditors.turnover)
> my_train$Debtors.turnover=as.numeric(my_train$Debtors.turnover)
> my_train$Finished.goods.turnover=as.numeric(my_train$Finished.goods.turnover)
> my_train$WIP.turnover=as.numeric(my_train$WIP.turnover)
> my_train$Raw.material.turnover=as.numeric(my_train$Raw.material.turnover)
> my_train$Shares.outstanding=as.numeric(my_train$Shares.outstanding)
> my_train$Equity.face.value=as.numeric(my_train$Equity.face.value)
> my_train$PE.on.BSE=as.numeric(my_train$PE.on.BSE)
```

The category variables are converted into continuous variables for the default status of the companies in the train datasets.

```
> my_test$Creditors.turnover=as.numeric(my_test$Creditors.turnover)
> my_test$Debtors.turnover=as.numeric(my_test$Debtors.turnover)
> my_test$Finished.goods.turnover=as.numeric(my_test$Finished.goods.turnover)
> my_test$WIP.turnover=as.numeric(my_test$WIP.turnover)
> my_test$Raw.material.turnover=as.numeric(my_test$Raw.material.turnover)
> my_test$Shares.outstanding=as.numeric(my_test$Shares.outstanding)
> my_test$Equity.face.value=as.numeric(my_test$Equity.face.value)
> my_test$PE.on.BSE=as.numeric(my_test$PE.on.BSE)
```

The category variables are converted into the continuous variables and the variables are related with the defaulted cases of company in validation datasets.

```
> my_train$Default...1=ifelse(my_train$Networth.Next.Year>0,0,1)
```

Default variable is created for the train dataset with pre-sets like net worth next year positive is non defaulters and net worth next year negative is defaulters.

```
> train_data$Gross.Margin.Ratio=train_data$Total.income/train_data$Sales
> train_data$Return.on.assets.ratio=train_data$Total.income/train_data$Total.assets
> train_data$Return.on.equity.ratio=train_data$Total.income/train_data$Shareholders.funds
> train_data$Debt.to.equity.ratio=train_data$Total.liabilities/train_data$Shareholders.funds
>train_data$Current.ratio=train_data$Current.assets/train_data$Current.liabilities...provisions
> train_data$Cash.ratio=train_data$Cash.profit/train_data$Current.liabilities...provisions
> train_data$Asset.turnover.ratio=train_data$Sales/train_data$Total.assets
```

```
> train_data$Day.sales.in.inventory.ratio=365/train_data$Finished.goods.turnover
```

The profitability ratio, leverage ratio, liquidity ratio, efficiency ratios are created for the train datasets for the analysis of ratio in the company.

```
> test_data$Gross.Margin.Ratio=test_data$Total.income/test_data$Sales
> test_data$Return.on.assets.ratio=test_data$Total.income/test_data$Total.assets
> test_data$Return.on.equity.ratio=test_data$Total.income/test_data$Shareholders.funds
> test_data$Debt.to.equity.ratio=test_data$Total.liabilities/test_data$Shareholders.funds
> test_data$Current.ratio=test_data$Current.assets/test_data$Current.liabilities...provisions
> test_data$Cash.ratio=test_data$Cash.profit/test_data$Current.liabilities...provisions
> test_data$Asset.turnover.ratio=test_data$Sales/test_data$Total.assets
> test_data$Day.sales.in.inventory.ratio=365/test_data$Finished.goods.turnover
```

The ratios are created for profitability, leverage, liquidity and efficiency of financial ratios in the validation datasets for the analysis.

4. Conclusion

The Default Status of the companies are predicted for the original datasets and the validation datasets in the variables of various ratios like profitability, leverage, liquidity and company size. Defaulters are measured for the credit lenders are gaining the profitability of the ratio of the company by 94% on the validation of the defaulters. The defaulters will require the amount of transactions by the company size and the non-defaulters will get the possible ways to decrease the defaulters by the rate of 68% for the company defaulters. The company will require the defaulters will increase and credit risk is exposed to the 94% of the companies. Credit Risk Models are better in predicting the return value of 6.5% in the returned credit to the defaulted companies.

5. Appendix

```
#====================================Set working Directory=========================================
setwd("/Users/numerp/Documents/PGP-BABI/Module 9 Financial Risk Analytics/Project 8 Financial Risk Analytics")
getwd()
#====================================Calling Dataset=============================================
library(readxl)
my_train=read_excel("raw-data.xlsx",sheet = 1)
my_test=read_excel("validation_data.xlsx",sheet = 1)
#====================================Continuous to Discrete Variable=============================
dim(my_train)
colnames(my_train)=make.names(colnames(my_train))
names(my_train)
str(my_train)
my_train$Creditors.turnover=as.numeric(my_train$Creditors.turnover)
my_train$Debtors.turnover=as.numeric(my_train$Debtors.turnover)
my_train$Finished.goods.turnover=as.numeric(my_train$Finished.goods.turnover)
my_train$WIP.turnover=as.numeric(my_train$WIP.turnover)
my_train$Raw.material.turnover=as.numeric(my_train$Raw.material.turnover)
my_train$Shares.outstanding=as.numeric(my_train$Shares.outstanding)
my_train$Equity.face.value=as.numeric(my_train$Equity.face.value)
my_train$PE.on.BSE=as.numeric(my_train$PE.on.BSE)
str(my_train)
summary(my_train)
dim(my_test)
colnames(my_test)=make.names(colnames(my_test))
names(my_test)
str(my_test)
my_test$Creditors.turnover=as.numeric(my_test$Creditors.turnover)
my_test$Debtors.turnover=as.numeric(my_test$Debtors.turnover)
my_test$Finished.goods.turnover=as.numeric(my_test$Finished.goods.turnover)
my_test$WIP.turnover=as.numeric(my_test$WIP.turnover)
my_test$Raw.material.turnover=as.numeric(my_test$Raw.material.turnover)
my_test$Shares.outstanding=as.numeric(my_test$Shares.outstanding)
my_test$Equity.face.value=as.numeric(my_test$Equity.face.value)
my_test$PE.on.BSE=as.numeric(my_test$PE.on.BSE)
str(my_test)
summary(my_test)
#====================================Missing Value Treatments====================================
colSums(is.na(my_train))
my_train=subset(my_train,select = -c(22))
class(my_train)
my_train=as.data.frame(my_train)
for(i in 1:ncol(my_train)){
```

```
  my_train[,i]=as.numeric(unlist(my_train[,i]))
  my_train[is.na(my_train[,i]),i]=median(my_train[,i],na.rm = TRUE)
}
any(is.na(my_train))
colSums(is.na(my_test))
my_test=subset(my_test,select = -c(22))
class(my_test)
my_test=as.data.frame(my_test)
for(i in 1:ncol(my_test)){
  my_test[,i]=as.numeric(unlist(my_test[,i]))
  my_test[is.na(my_test[,i]),i]=median(my_test[,i],na.rm = TRUE)
}
any(is.na(my_test))
#====================================Outlier
Treatments============================================
summary(my_train)
outlier=boxplot(my_train[,-c(1,2)],plot = FALSE)$out
print(outlier)
library(scales)
for(i in 3:ncol(my_train)){
q=quantile(my_train[,i],c(0.1,0.99))
my_train[,i]=squish(my_train[,i],q)
}
summary(my_train)
summary(my_test)
outlier1=boxplot(my_test[,-c(1,2)],plot = FALSE)$out
print(outlier1)
for(i in 3:ncol(my_test)){
  q=quantile(my_test[,i],c(0.1,0.99))
  my_test[,i]=squish(my_test[,i],q)
}
summary(my_test)
#====================================Default                                varaible
Creation=====================================
my_train$Default...1=ifelse(my_train$Networth.Next.Year>0,0,1)
summary(my_train$Default...1)
train_data=subset(my_train,select = -c(1,2))
test_data=subset(my_test,select = -c(1))
train_data=train_data[,c(50,1:49)]
colnames(train_data)
colnames(test_data)
train_data$Default...1=as.factor(train_data$Default...1)
test_data$Default...1=as.factor(test_data$Default...1)
#====================================EDA
Visualization=========================================
library(ggplot2)
library(DataExplorer)
attach(train_data)
plot_bar(data = train_data,ggtheme = theme_lares(),
         title = "Default Status in Train Data")
```

```
plot_histogram(data = train_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
               title = "Histogram for Train Dataset")
plot_boxplot(data = train_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol = 5,
             title = "Boxplot for Train Dataset")
plot_density(data = train_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
             title = "Density Plots for Train Dataset")
plot_scatterplot(data = train_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol =
5,
                 title = "Scatterplot for Train Dataset")
plot_qq(data = train_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
        title = "Quantile Chart for Train Dataset")
plot_bar(data = test_data,ggtheme = theme_lares(),
         title = "Default Status in Test Data")
plot_histogram(data = test_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
               title = "Histogram for Test Dataset")
plot_boxplot(data = test_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol = 5,
             title = "Boxplot for Test Dataset")
plot_density(data = test_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
             title = "Density Plots for Test Dataset")
plot_scatterplot(data = test_data,by = "Default...1",ggtheme = theme_lares(),nrow = 5,ncol = 5,
                 title = "Scatterplot for Test Dataset")
plot_qq(data = test_data,ggtheme = theme_lares(),nrow = 5,ncol = 5,
        title = "Quantile Chart for Test Dataset")
#====================================Correlation
Check=========================================
summary(as.factor(Default...1))
243/(3298+243) #0.06862468
correlation=cor(train_data[,-1])
correlation
library(ggcorrplot)
ggcorrplot::ggcorrplot(correlation,method = "square",ggtheme = theme_classic(),
                       title = "Correlation Chart")
library(devtools)
#devtools::install_github("laresbernardo/lares")
library(lares)
corr_cross(correlation)
#====================================Multicollinearity
Check=====================================
attach(train_data)
model1=glm(Default...1~. -Default...1,data = train_data,family = binomial(link = logit))
summary(model1)
#Removing Insignificant variables
model2=glm(Default...1~Total.assets+PBDITA+PAT.as...of.net.worth+Other.income+Reserv
es.and.funds

+Current.liabilities...provisions+Cumulative.retained.profits+Capital.employed+TOL.TNW
           +Total.term.liabilities...tangible.net.worth+Contingent.liabilities+Net.fixed.assets
           +Debt.to.equity.ratio..times.+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
           +PE.on.BSE,data = train_data,family = binomial)
summary(model2)
```

```
#Removing Insignificant variables
model3=glm(Default...1~Total.assets+PBDITA+PAT.as...of.net.worth+Reserves.and.funds

+Current.liabilities...provisions+Cumulative.retained.profits+Capital.employed+TOL.TNW
        +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
        +Debt.to.equity.ratio..times.+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
        +PE.on.BSE,data = train_data,family = binomial)
summary(model3)
#Removing Insignificant variables
model4=glm(Default...1~Total.assets+PAT.as...of.net.worth+Reserves.and.funds

+Current.liabilities...provisions+Cumulative.retained.profits+Capital.employed+TOL.TNW
        +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
        +Debt.to.equity.ratio..times.+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
        +PE.on.BSE,data = train_data,family = binomial)
summary(model4)
library(car)
car::vif(model4)
#Removing Higher Variance Inflation Rate variables -Total Assets -Capital Employed
model5=glm(Default...1~PAT.as...of.net.worth+Reserves.and.funds
        +Cumulative.retained.profits+TOL.TNW
        +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
        +Debt.to.equity.ratio..times.
        +Cash.to.average.cost.of.sales.per.day+Creditors.turnover
        +PE.on.BSE,data = train_data,family = binomial)
summary(model5)
#Removing Insignificant Variables
model6=glm(Default...1~PAT.as...of.net.worth
        +Cumulative.retained.profits+TOL.TNW
        +Total.term.liabilities...tangible.net.worth+Net.fixed.assets
        +Debt.to.equity.ratio..times.
        +Cash.to.average.cost.of.sales.per.day+Creditors.turnover
        +PE.on.BSE,data = train_data,family = binomial)
summary(model6)
car::vif(model6)
#Removing Higher Variation Inflation Rate Variables -Total Term Liabilities Tangible Net Worth -Debt Equity Ratio Times
model7=glm(Default...1~PAT.as...of.net.worth+Cumulative.retained.profits+TOL.TNW
        +Net.fixed.assets+Cash.to.average.cost.of.sales.per.day+Creditors.turnover
        +PE.on.BSE,data = train_data,family = binomial)
summary(model7)
car::vif(model7)
#=====================================New Variables Creation=====================================
dim(train_data)
#Profitability Ratio - Gross Margin Ratio
train_data$Gross.Margin.Ratio=train_data$Total.income/train_data$Sales
#Profitability Ratio - Return on Assets Ratio
train_data$Return.on.assets.ratio=train_data$Total.income/train_data$Total.assets
#Profitability - Return on Equity Ratio
```

```
train_data$Return.on.equity.ratio=train_data$Total.income/train_data$Shareholders.funds
#Leverage Financial Ratio - Debt to Equity Ratio
train_data$Debt.to.equity.ratio=train_data$Total.liabilities/train_data$Shareholders.funds
#Liquidity Ratio - Current Ratio
train_data$Current.ratio=train_data$Current.assets/train_data$Current.liabilities...provisions
#Liquidity Ratio - Cash Ratio
train_data$Cash.ratio=train_data$Cash.profit/train_data$Current.liabilities...provisions
#Efficiency Ratio - Asset Turnover Ratio
train_data$Asset.turnover.ratio=train_data$Sales/train_data$Total.assets
#Efficiency Ratio - Day Sales in Inventory Ratio
train_data$Day.sales.in.inventory.ratio=365/train_data$Finished.goods.turnover
dim(train_data)
summary(train_data)
dim(test_data)
#Profitability Ratio - Gross Margin Ratio
test_data$Gross.Margin.Ratio=test_data$Total.income/test_data$Sales
#Profitability Ratio - Return on Assets Ratio
test_data$Return.on.assets.ratio=test_data$Total.income/test_data$Total.assets
#Profitability - Retun on Equity Ratio
test_data$Return.on.equity.ratio=test_data$Total.income/test_data$Shareholders.funds
#Leverage Financial Ratio - Debt to Equity Ratio
test_data$Debt.to.equity.ratio=test_data$Total.liabilities/test_data$Shareholders.funds
#Liquidity Ratio - Current Ratio
test_data$Current.ratio=test_data$Current.assets/test_data$Current.liabilities...provisions
#Liquidity Ratio - Cash Ratio
test_data$Cash.ratio=test_data$Cash.profit/test_data$Current.liabilities...provisions
#Efficiency Ratio - Asset Turnover Ratio
test_data$Asset.turnover.ratio=test_data$Sales/test_data$Total.assets
#Efficiency Ratio - Day Sales in Inventory Ratio
test_data$Day.sales.in.inventory.ratio=365/test_data$Finished.goods.turnover
dim(test_data)
summary(test_data)
#====================================Logisitics                    Regression
Model==================================
log.reg1=glm(Default...1~.,data = train_data,family = binomial)
summary(log.reg1)
#Removing Insignificant VariablesA
log.reg2=glm(Default...1~Total.assets+Net.worth+Total.income+Change.in.stock+Total.expe
nses+
        Profit.after.tax+PBDITA+PBT+Cash.profit+PBDITA.as...of.total.income+
        PBT.as...of.total.income+PAT.as...of.total.income+Cash.profit.as...of.total.income+
        PAT.as...of.net.worth+Sales+Income.from.financial.services+Other.income+
        Total.capital+Reserves.and.funds+Borrowings+Current.liabilities...provisions+
        Deferred.tax.liability+Shareholders.funds+Cumulative.retained.profits+
        Capital.employed+TOL.TNW+Total.term.liabilities...tangible.net.worth+
        Contingent.liabilities...Net.worth....+Contingent.liabilities+Net.fixed.assets+
        Investments+Current.assets+Net.working.capital+Quick.ratio..times.+
        Current.ratio..times.+Debt.to.equity.ratio..times.+
        Cash.to.current.liabilities..times.+Cash.to.average.cost.of.sales.per.day+
        Creditors.turnover+Debtors.turnover+Finished.goods.turnover+WIP.turnover+
```

```
            Raw.material.turnover+Shares.outstanding+Equity.face.value+EPS+Adjusted.EPS+
            PE.on.BSE+Gross.Margin.Ratio+Return.on.assets.ratio+Return.on.equity.ratio+
            Debt.to.equity.ratio+Current.ratio+Cash.ratio+Asset.turnover.ratio+
            Day.sales.in.inventory.ratio,data = train_data,
          family = binomial)
summary(log.reg2)
car::vif(log.reg2)
#Removing Higher Variance Inflation Factor Variables
log.reg3=glm(Default...1~Change.in.stock+
            PBDITA+Cash.profit+PBDITA.as...of.total.income+
            PBT.as...of.total.income+PAT.as...of.total.income+Cash.profit.as...of.total.income+
            PAT.as...of.net.worth+Income.from.financial.services+Other.income+
            Total.capital+Reserves.and.funds+Borrowings+Current.liabilities...provisions+
            Deferred.tax.liability+Cumulative.retained.profits+
            TOL.TNW+Total.term.liabilities...tangible.net.worth+
            Contingent.liabilities...Net.worth....+Contingent.liabilities+Net.fixed.assets+
            Investments+Current.assets+Net.working.capital+Quick.ratio..times.+
            Current.ratio..times.+Debt.to.equity.ratio..times.+
            Cash.to.current.liabilities..times.+Cash.to.average.cost.of.sales.per.day+
            Creditors.turnover+Debtors.turnover+Finished.goods.turnover+WIP.turnover+
            Raw.material.turnover+Shares.outstanding+Equity.face.value+
            PE.on.BSE+Gross.Margin.Ratio+Return.on.assets.ratio+Return.on.equity.ratio+
            Debt.to.equity.ratio+Current.ratio+Cash.ratio+Asset.turnover.ratio+
            Day.sales.in.inventory.ratio,data = train_data,
          family = binomial)
summary(log.reg3)
#Removing Insignificant Variables
log.reg4=glm(Default...1~PAT.as...of.net.worth+Current.liabilities...provisions+
            Cumulative.retained.profits+TOL.TNW+Total.term.liabilities...tangible.net.worth+
            Contingent.liabilities...Net.worth....+Contingent.liabilities+Net.fixed.assets+
            Current.assets+Debt.to.equity.ratio..times.+PE.on.BSE+Gross.Margin.Ratio+
            Debt.to.equity.ratio+Day.sales.in.inventory.ratio,data = train_data,
          family = binomial)
summary(log.reg4)
#Removing Insignificant Variables
log.reg5=glm(Default...1~PAT.as...of.net.worth+Current.liabilities...provisions+
            Cumulative.retained.profits+TOL.TNW+Total.term.liabilities...tangible.net.worth+
            Contingent.liabilities+Net.fixed.assets+
            Current.assets+Debt.to.equity.ratio..times.+PE.on.BSE+Gross.Margin.Ratio+
            Debt.to.equity.ratio+Day.sales.in.inventory.ratio,data = train_data,
          family = binomial)
summary(log.reg5)
car::vif(log.reg5)
#Removing Higher Variance Inflation Factor
log.reg6=glm(Default...1~PAT.as...of.net.worth+
            Cumulative.retained.profits+TOL.TNW+
            Contingent.liabilities+Net.fixed.assets+
            PE.on.BSE+Gross.Margin.Ratio+
            Debt.to.equity.ratio+Day.sales.in.inventory.ratio,data = train_data,
          family = binomial)
```

```
summary(log.reg6)
#Removing Insignificance Variables
log.reg7=glm(Default...1~PAT.as...of.net.worth+
               Cumulative.retained.profits+TOL.TNW+
               Net.fixed.assets+
               PE.on.BSE+
               Day.sales.in.inventory.ratio,data = train_data,
             family = binomial)
summary(log.reg7)
car::vif(log.reg7)
exp(coef(log.reg7))
exp(coef(log.reg7))/(1+exp(coef(log.reg7)))
library(rms)
library(pscl)
#McFadden 0to0.10 - Bad,0.10to0.15 - Average,0.15to0.3 - Moderate,0.3to0.5 - Good,>0.5
Excellent
pscl::pR2(log.reg7)["McFadden"]
logLik(log.reg7)
#=====================================Prediction on Train
Dataset==================================
log.reg=predict.glm(log.reg7,train_data,type = "response")
log.reg
tab.train=table(train_data$Default...1,log.reg>0.3)
tab.train
sum(diag(tab.train))/sum(tab.train)
pred1=ifelse(log.reg>0.68,0,1)
pred1
actual1=train_data$Default...1
library(caret)
cm_log_train=confusionMatrix(as.factor(pred1),actual1,positive = "1")
cm_log_train
#=====================================Prediction on Test
Dataset==================================
log.reg.pred=predict.glm(log.reg7,test_data,type = "response")
log.reg.pred
table.log=table(test_data$Default...1,log.reg.pred>0.3)
table.log
sum(diag(table.log))/sum(table.log)
pred=ifelse(log.reg.pred>0.68,0,1)
pred
actual=test_data$Default...1
actual
cm_log=confusionMatrix(as.factor(pred),actual,positive = "1")
cm_log
#=====================================Decile Based on Train
Data==================================
train_data$pred=predict(log.reg7,train_data,type = "response")
decile = function(x){
  deciles = vector(length=10)
  for (i in seq(0.1,1,.1)){
```

```
  deciles[i*10] = quantile(x, i, na.rm=T)
 }
 return (
  ifelse(x<deciles[1], 1,
       ifelse(x<deciles[2], 2,
            ifelse(x<deciles[3], 3,
                 ifelse(x<deciles[4], 4,
                      ifelse(x<deciles[5], 5,
                           ifelse(x<deciles[6], 6,
                                ifelse(x<deciles[7], 7,
                                     ifelse(x<deciles[8], 8,
                                          ifelse(x<deciles[9], 9, 10
                                          ))))))))))
}
train_data$deciles=decile(train_data$pred)
m=data.table::data.table(train_data)
rank.default = m[, list(cnt=length(Default...1),
                cnt_resp=sum(Default...1==1),
                cnt_non_resp=sum(Default...1==0)
                ), by=deciles][order(-deciles)]
rank.default$rrate=round(rank.default$cnt_resp/rank.default$cnt,4)
rank.default$cum_resp=cumsum(rank.default$cnt_resp)
rank.default$cum_non_resp=cumsum(rank.default$cnt_non_resp)
rank.default$cum_rel_resp=round(rank.default$cum_resp/sum(rank.default$cnt_non_resp),4)
rank.default$cum_rel_non_resp=round(rank.default$cum_non_resp/sum(rank.default$cnt_no
n_resp),4)
rank.default$ks=abs(rank.default$cum_rel_resp - rank.default$cum_rel_non_resp)*100
rank.default$rrate=scales::percent(rank.default$rrate)
rank.default$cum_rel_resp=scales::percent(rank.default$cum_rel_resp)
rank.default$cum_rel_non_resp=scales::percent(rank.default$cum_rel_non_resp)
train_data_rank=rank.default
View(train_data_rank)
print(train_data_rank)
plot(train_data_rank$ks)
lines(train_data_rank$ks)
#====================================Decile Based on Test
Data==================================
test_data$pred=predict(log.reg7,test_data,type = "response")
decile = function(x){
 deciles = vector(length=10)
 for (i in seq(0.1,1,.1)){
  deciles[i*10] = quantile(x, i, na.rm=T)
 }
 return (
  ifelse(x<deciles[1], 1,
       ifelse(x<deciles[2], 2,
            ifelse(x<deciles[3], 3,
                 ifelse(x<deciles[4], 4,
                      ifelse(x<deciles[5], 5,
                           ifelse(x<deciles[6], 6,
```

```
                                    ifelse(x<deciles[7], 7,
                                        ifelse(x<deciles[8], 8,
                                            ifelse(x<deciles[9], 9, 10
                                            ))))))))))
}
test_data$deciles=decile(test_data$pred)
n=data.table::data.table(test_data)
rank.test.default = n[, list(cnt=length(Default...1),
                    cnt_resp=sum(Default...1==1),
                    cnt_non_resp=sum(Default...1==0)
), by=deciles][order(-deciles)]
rank.test.default$rrate=round(rank.test.default$cnt_resp/rank.test.default$cnt,4)
rank.test.default$cum_resp=cumsum(rank.test.default$cnt_resp)
rank.test.default$cum_non_resp=cumsum(rank.test.default$cnt_non_resp)
rank.test.default$cum_rel_resp=round(rank.test.default$cum_resp/
                                sum(rank.test.default$cnt_non_resp),4)
rank.test.default$cum_rel_non_resp=round(rank.test.default$cum_non_resp/
                                    sum(rank.test.default$cnt_non_resp),4)
rank.test.default$ks=abs(rank.test.default$cum_rel_resp                                    -
rank.test.default$cum_rel_non_resp)*100
rank.test.default$rrate=percent(rank.test.default$rrate)
rank.test.default$cum_rel_resp=percent(rank.test.default$cum_rel_resp)
rank.test.default$cum_rel_non_resp=percent(rank.test.default$cum_rel_non_resp)
test_data_rank=rank.test.default
View(test_data_rank)
print(test_data_rank)
plot(test_data_rank$ks)
lines(test_data_rank$ks)
#====================================Model                              Performance
Measures==================================
library(ROCR)
library(pROC)
rocpred=prediction(log.reg.pred,test_data$Default...1)
#ROC Curve
ROC=pROC::roc(Default...1,log.reg7$fitted.values)
ROC
plot.roc(Default...1,log.reg7$fitted.values)
#Area Under Curve
AUC=as.numeric(performance(rocpred, "auc")@y.values)
AUC
#KS Chart
perf = performance(rocpred, "tpr","fpr")
KS=max(attr(perf, 'y.values')[[1]]-attr(perf, 'x.values')[[1]])
KS
plot(perf, colorize =T, print.cutoffs.at= seq(0, 1, .1), text.adj = c(-.2, 1.7))
#Gini Coefficient
library(ineq)
gini=ineq(log.reg.pred, type="Gini")
gini
```

```
#**************************************************************************
************************
```