FINANCIAL RISK ANALYSIS- LOAN DEFAULT

SUBMITTED BY: KANUPRIYA MITTAL

CONTENT:

1. Project Objective
2. Data Analysis
   - 2.1 Descriptive Data Analysis
   - 2.2 Exploratory Data Analysis
     - 2.2.1 Univariate Analysis
     - 2.2.2 Bivariate Analysis
3. Check for Integrity of Data
   - 3.1 Outlier Detection
   - 3.2 Missing Value Detection
4. Creating The Default variable
5. Missing Value Treatment
   - 4.1 Imputation using MICE
6. Creation Of New Ratios for Raw Data set
   - 6.1 Ratios for profitability
   - 6.2 Ratios for Size
   - 6.3 Ratios for Liquidity
   - 6.4 Ratios for Leverage
7. Outliers Treatment
   - 7.1 Capping the Outliers
8. Multicollinearity Treatment
   - 8.1 Creating a Logistic Model
   - 8.2 Using the VIF to remove highly correlated variables
   - 8.3 Using StepAIC to give the lowest AIC from specific variables
9. Selection of the Significant Variables For Model Building
   - 9.1 Using STEPAIC
10. Balancing the dataset
    - 10.1 Smote Method
11. Creating the Final Model using SMOTE dataset
    - 11.1 Applying Logistic regression
12. Analysis Of the Coefficients of Logistic Model
13. Predictions + Confusion Matrix on the Developed dataset
14. Preparing the Validation Dataset
    - 14.1 Importing the validation set
    - 14.2 Descriptive analysis
    - 14.3 Imputing the Missing Values
    - 14.4 Creation of New ratios(variables same as the Developed dataset)
    - 14.5 Selecting the Final Significant variables for validation set
15. Predictions + Confusion Matrix on the Validation Dataset
16. Sorting and Deciling of the Validation Set
    - 16.1 Adding The probability of default in our validation set
    - 16.2 Sorting in descending order by Probability of Default
    - 16.3 Creating deciles
      - 16.3.1 In the range form
      - 16.3.2 In the Ranking Form

1. PROJECT OBJECTIVE

Loan Defaults has a huge importance for Banks. Providing loans is the source of income for banks and if the Borrower fails to return back the loan money with interest, it's a loss to the Bank. Banks want to minimize the loss this risk. Hence, we need a model that could predict it for the banks about the chances of Default. Default on loans is a crucial problem for any bank and managing this risk would come through a solution of being able to predict. So, we are supposed to create a predictive model which would model the past data provided by a bank and predict on an unseen future set of data.

Financial Risk Analysis is done in this project to alleviate the risk of losses to Banks.

2. DATA ANALYSIS

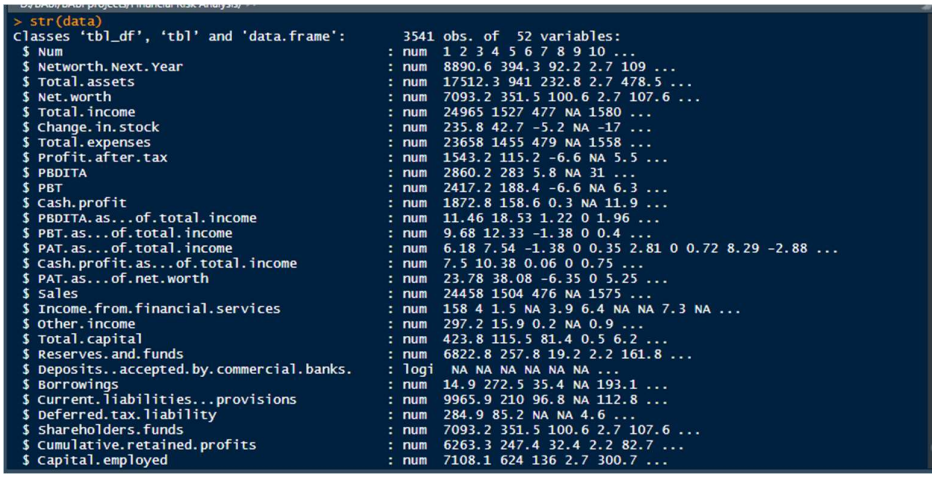
It is a culmination of Descriptive and Exploratory Data Analysis. It is done to understand the basic data structure as well as to visualize the dataset before performing any modelling. It checks for the relationship between the variables through the use of graphs. And also gives the summary and the class of each variable.

2.1 DESCRIPTIVE DATA ANALYSIS

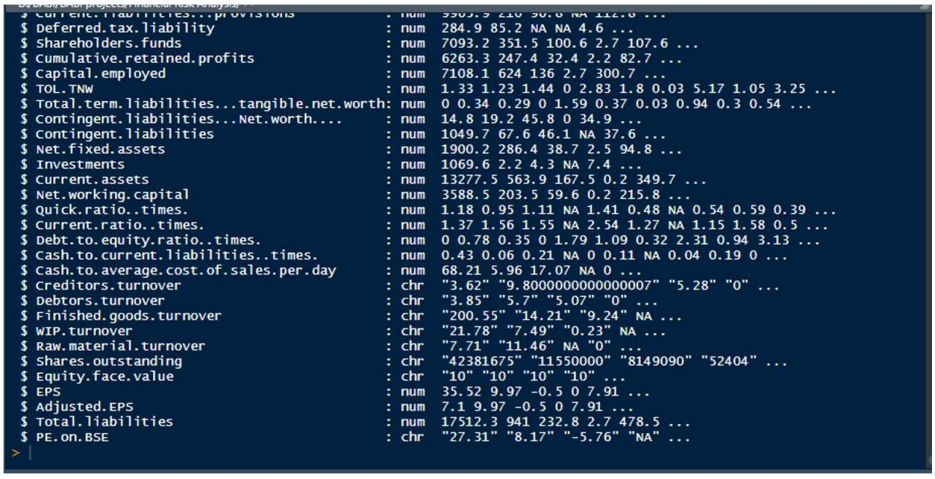
The descriptive analysis tells about the basic structure of our dataset.

Structure of dataset:

```
str(data)
```



```
> str(data)
Classes 'tbl_df', 'tbl' and 'data.frame':       3541 obs. of  52 variables:
 $ Num                                          : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Networth.Next.Year                           : num  8890.6 394.3 92.2 2.7 109 ...
 $ Total.assets                                 : num  17512.3 941 232.8 2.7 478.5 ...
 $ Net.worth                                    : num  7093.2 351.5 100.6 2.7 107.6 ...
 $ Total.income                                 : num  24965 1527 477 NA 1580 ...
 $ Change.in.stock                              : num  235.8 42.7 -5.2 NA -17 ...
 $ Total.expenses                               : num  23658 1455 479 NA 1558 ...
 $ Profit.after.tax                             : num  1543.2 115.2 -6.6 NA 5.5 ...
 $ PBDITA                                       : num  2860.2 283 5.8 NA 31 ...
 $ PBT                                          : num  2417.2 188.4 -6.6 NA 6.3 ...
 $ Cash.profit                                  : num  1872.8 158.6 0.3 NA 11.9 ...
 $ PBDITA.as...of.total.income                  : num  11.46 18.53 1.22 0 1.96 ...
 $ PBT.as...of.total.income                     : num  9.68 12.33 -1.38 0 0.4 ...
 $ PAT.as...of.total.income                     : num  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  7.5 10.38 0.06 0 0.75 ...
 $ PAT.as...of.net.worth                        : num  23.78 38.08 -6.35 0 5.25 ...
 $ Sales                                        : num  24458 1504 476 NA 1575 ...
 $ Income.from.financial.services               : num  158 4 1.5 NA 3.9 6.4 NA NA 7.3 NA ...
 $ Other.income                                 : num  297.2 15.9 0.2 NA 0.9 ...
 $ Total.capital                                : num  423.8 115.5 81.4 0.5 6.2 ...
 $ Reserves.and.funds                           : num  6822.8 257.8 19.2 2.2 161.8 ...
 $ Deposits..accepted.by.commercial.banks.      : logi  NA NA NA NA NA NA ...
 $ Borrowings                                   : num  14.9 272.5 35.4 NA 193.1 ...
 $ Current.liabilities...provisions             : num  9965.9 210 96.8 NA 112.8 ...
 $ Deferred.tax.liability                       : num  284.9 85.2 NA NA 4.6 ...
 $ Shareholders.funds                           : num  7093.2 351.5 100.6 2.7 107.6 ...
 $ Cumulative.retained.profits                  : num  6263.3 247.4 32.4 2.2 82.7 ...
 $ Capital.employed                             : num  7108.1 624 136 2.7 300.7 ...
```



```
 $ Deferred.tax.liability                       : num  284.9 85.2 NA NA 4.6 ...
 $ Shareholders.funds                           : num  7093.2 351.5 100.6 2.7 107.6 ...
 $ Cumulative.retained.profits                  : num  6263.3 247.4 32.4 2.2 82.7 ...
 $ Capital.employed                             : num  7108.1 624 136 2.7 300.7 ...
 $ TOL.TNW                                      : num  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  0 0.34 0.29 0 1.59 0.37 0.03 0.94 0.3 0.54 ...
 $ Contingent.liabilities...Net.worth....       : num  14.8 19.2 45.8 0 34.9 ...
 $ Contingent.liabilities                       : num  1049.7 67.6 46.1 NA 37.6 ...
 $ Net.fixed.assets                             : num  1900.2 286.4 38.7 2.5 94.8 ...
 $ Investments                                  : num  1069.6 2.2 4.3 NA 7.4 ...
 $ Current.assets                               : num  13277.5 563.9 167.5 0.2 349.7 ...
 $ Net.working.capital                          : num  3588.5 203.5 59.6 0.2 215.8 ...
 $ Quick.ratio..times.                          : num  1.18 0.95 1.11 NA 1.41 0.48 NA 0.54 0.59 0.39 ...
 $ Current.ratio..times.                        : num  1.37 1.56 1.55 NA 2.54 1.27 NA 1.15 1.58 0.5 ...
 $ Debt.to.equity.ratio..times.                 : num  0 0.78 0.35 0 1.79 1.09 0.32 2.31 0.94 3.13 ...
 $ Cash.to.current.liabilities..times.          : num  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  68.21 5.96 17.07 NA 0 ...
 $ Creditors.turnover                           : chr  "3.62" "9.8000000000000007" "5.28" "0" ...
 $ Debtors.turnover                             : chr  "3.85" "5.7" "5.07" "0" ...
 $ Finished.goods.turnover                      : chr  "200.55" "14.21" "9.24" NA ...
 $ WIP.turnover                                 : chr  "21.78" "7.49" "0.23" NA ...
 $ Raw.material.turnover                        : chr  "7.71" "11.46" NA "0" ...
 $ Shares.outstanding                           : chr  "42381675" "11550000" "8149090" "52404" ...
 $ Equity.face.value                            : chr  "10" "10" "10" "10" ...
 $ EPS                                          : num  35.52 9.97 -0.5 0 7.91 ...
 $ Adjusted.EPS                                 : num  7.1 9.97 -0.5 0 7.91 ...
 $ Total.liabilities                            : num  17512.3 941 232.8 2.7 478.5 ...
 $ PE.on.BSE                                    : chr  "27.31" "8.17" "-5.76" "NA" ...
>
```

Dim: Gives the dimensions of the dataset

```
dim(data)
```

```
[1] 3541   52
```

Head: Give the first 10 records from the data

head(data,10)

```
#   EPS <dbl>, Adjusted.EPS <dbl>, Total.liabilities <dbl>, PE.on.BSE <chr>
> head(data,10)
# A tibble: 10 x 52
     Num Networth.Next.Y~ Total.assets Net.worth Total.income Change.in.stock Total.expenses Profit.after.tax PBDITA   PBT
   <dbl>           <dbl>        <dbl>     <dbl>        <dbl>           <dbl>          <dbl>            <dbl>  <dbl> <dbl>
 1     1            8891.       17512.     7093.       24965.            236.          23658.           1543.  2860. 2417.
 2     2             394.         941       352.        1527.            42.7           1455.            115.    283  188.
 3     3             92.2         233.      101.         477.            -5.2            479.            -6.6    5.8 -6.6
 4     4              2.7          2.7       2.7          NA              NA              NA              NA     NA    NA
 5     5             109          478.      108.        1580.            -17            1558             5.5    31    6.3
 6     6             689.        2434.      676.        2649.            62.3           2636.            74.5  200.  74.5
 7     7             246          327.      245.         NA              NA              NA              NA     NA    NA
 8     8              13.7         80        12.7        154.            -8.5            144             1.1    9.7   2
 9     9             292.         574.      239.         583.            31             565.             48.3  110.  68.5
10    10              -7.3         88.6      19.6         83.4            -6.7            79.1            -2.4    0.3 -14.4
# ... with 42 more variables: Cash.profit <dbl>, PBDITA.as...of.total.income <dbl>, PBT.as...of.total.income <dbl>,
```

Tail: Gives the last 10 records from the data

tail(data,10)

```
> tail(data,10)
# A tibble: 10 x 52
     Num Networth.Next.Y~ Total.assets Net.worth Total.income Change.in.stock Total.expenses Profit.after.tax PBDITA   PBT
   <dbl>           <dbl>        <dbl>     <dbl>        <dbl>           <dbl>          <dbl>            <dbl>  <dbl> <dbl>
 1  3536            1647.       1707.      996.        4903.            76.9           4584.            396.   700.  630.
 2  3537              21          48.2       18          134.             0.9            131.              3.7    7.6   3.7
 3  3538             256          878.      256         1154.           -22.2           1019.            113.   209.  186.
 4  3539             271.        1404.      439         3706.           -82.4           3593.             30.2  196.   27.2
 5  3540               1.2         17.8       1.2         15.5            -1.2            14.2             0.1    1.8   0.2
 6  3541             226.         450.      172.         565             30.5           581.             14.4   76.7  41.1
 7  3542              89.4         97.6       82          75.8            -4              66.5             5.3    11.1   6.2
 8  3543             246.         903.      209.        1005.             5.6           966.             44.2  120.   70
 9  3544             147.         177       137.         371              3.9           349.             26     50.5  40.8
10  3545              -0.2         0.6       0.3          NA              NA              17.4            -17.4  -17.4 -17.4
# ... with 42 more variables: Cash.profit <dbl>, PBDITA.as...of.total.income <dbl>, PBT.as...of.total.income <dbl>,
```

Names: Gives the names of all the variables.

names(data)

```
> names(data)
 [1] "Num"                                  "Networth.Next.Year"
 [3] "Total.assets"                         "Net.worth"
 [5] "Total.income"                         "Change.in.stock"
 [7] "Total.expenses"                       "Profit.after.tax"
 [9] "PBDITA"                               "PBT"
[11] "Cash.profit"                          "PBDITA.as...of.total.income"
[13] "PBT.as...of.total.income"             "PAT.as...of.total.income"
[15] "Cash.profit.as...of.total.income"     "PAT.as...of.net.worth"
[17] "Sales"                                "Income.from.financial.services"
[19] "Other.income"                         "Total.capital"
[21] "Reserves.and.funds"                   "Deposits..accepted.by.commercial.banks."
[23] "Borrowings"                           "Current.liabilities...provisions"
[25] "Deferred.tax.liability"               "Shareholders.funds"
[27] "Cumulative.retained.profits"          "Capital.employed"
[29] "TOL.TNW"                              "Total.term.liabilities...tangible.net.worth"
[31] "Contingent.liabilities...Net.worth...." "Contingent.liabilities"
[33] "Net.fixed.assets"                     "Investments"
[35] "Current.assets"                       "Net.working.capital"
[37] "Quick.ratio..times."                  "Current.ratio..times."
[39] "Debt.to.equity.ratio..times."         "Cash.to.current.liabilities..times."
[41] "Cash.to.average.cost.of.sales.per.day" "Creditors.turnover"
[43] "Debtors.turnover"                     "Finished.goods.turnover"
[45] "WIP.turnover"                         "Raw.material.turnover"
[47] "Shares.outstanding"                   "Equity.face.value"
[49] "EPS"                                  "Adjusted.EPS"
[51] "Total.liabilities"                    "PE.on.BSE"
```

2.2 EXPLORATORY DATA ANALYSIS

Here we are going to explore the Variables by two methods: By CENTRAL TENDENCIES and by GRAPHS.

➔ Through Central Measure(MEAN, MEDIANS, MODE)

Summary of Dataset:

summary(data)

```
> dim(data)
[1] 3541   52
> summary(data)
      Num         Networth.Next.Year  Total.assets         Net.worth
 Min.   :   1   Min.   :-74265.6   Min.   :      0.1   Min.   :     0.0
 1st Qu.: 886   1st Qu.:    31.7   1st Qu.:     91.3   1st Qu.:    31.3
 Median :1773   Median :   116.3   Median :    309.7   Median :   102.3
 Mean   :1772   Mean   :  1616.3   Mean   :   3443.4   Mean   :  1295.9
 3rd Qu.:2658   3rd Qu.:   456.1   3rd Qu.:   1098.7   3rd Qu.:   377.3
 Max.   :3545   Max.   :805773.4   Max.   :1176509.2   Max.   :613151.6

  Total.income       Change.in.stock    Total.expenses
 Min.   :      0.0   Min.   :-3029.40   Min.   :     -0.1
 1st Qu.:    106.5   1st Qu.:   -1.80   1st Qu.:     95.8
 Median :    444.9   Median :    1.60   Median :    407.7
 Mean   :   4582.8   Mean   :   41.49   Mean   :   4262.9
 3rd Qu.:   1440.9   3rd Qu.:   18.05   3rd Qu.:   1359.8
 Max.   :2442828.2   Max.   :14185.50   Max.   :2366035.3
 NA's   :198         NA's   :458        NA's   :139
 Profit.after.tax        PBDITA              PBT
 Min.   : -3908.30   Min.   :  -440.7   Min.   : -3894.80
 1st Qu.:     0.50   1st Qu.:     6.9   1st Qu.:     0.70
 Median :     8.80   Median :    35.4   Median :    12.40
 Mean   :   277.36   Mean   :   578.1   Mean   :   383.81
 3rd Qu.:    52.27   3rd Qu.:   150.2   3rd Qu.:    71.97
 Max.   :119439.10   Max.   :208576.5   Max.   :145292.60
 NA's   :131         NA's   :131        NA's   :131

  Cash.profit        PBDITA.as...of.total.income PBT.as...of.total.income
 Min.   : -2245.70   Min.   :-6400.000           Min.   :-21340.00
 1st Qu.:     2.90   1st Qu.:    5.000           1st Qu.:     0.55
 Median :    18.85   Median :    9.660           Median :     3.31
 Mean   :   392.07   Mean   :    4.571           Mean   :   -17.28
 3rd Qu.:    93.20   3rd Qu.:   16.390           3rd Qu.:     8.80
 Max.   :176911.80   Max.   :  100.000           Max.   :   100.00
 NA's   :131         NA's   :68                  NA's   :68
 PAT.as...of.total.income Cash.profit.as...of.total.income
 Min.   :-21340.00        Min.   :-15020.000
 1st Qu.:     0.35        1st Qu.:     2.020
 Median :     2.34        Median :     5.640
 Mean   :   -19.20        Mean   :    -8.229
 3rd Qu.:     6.34        3rd Qu.:    10.700
 Max.   :   150.00        Max.   :   100.000
 NA's   :68               NA's   :68
 PAT.as...of.net.worth     Sales           Income.from.financial.services
 Min.   :-748.72       Min.   :      0.1   Min.   :    0.00
 1st Qu.:   0.00       1st Qu.:    112.7   1st Qu.:    0.40
 Median :   7.92       Median :    453.1   Median :    1.80
 Mean   :  10.27       Mean   :   4549.5   Mean   :   80.84
 3rd Qu.:  20.19       3rd Qu.:   1433.5   3rd Qu.:    9.68
 Max.   :2466.67       Max.   :2384984.4   Max.   :51938.20
                       NA's   :259         NA's   :935
  Other.income      Total.capital     Reserves.and.funds
 Min.   :    0.00   Min.   :    0.1   Min.   : -6525.9
 1st Qu.:    0.40   1st Qu.:   13.1   1st Qu.:     5.0
 Median :    1.40   Median :   42.1   Median :    54.8
 Mean   :   41.36   Mean   :  216.6   Mean   :  1163.8
 3rd Qu.:    5.97   3rd Qu.:  100.3   3rd Qu.:   277.3
 Max.   :42856.70   Max.   :78273.2   Max.   :625137.8
 NA's   :1295       NA's   :4         NA's   :85
```

```
 Deposits..accepted.by.commercial.banks.   Borrowings
 Mode:logical                            Min.    :     0.10
 NA's:3541                               1st Qu.:    23.95
                                         Median :    99.20
                                         Mean   :  1122.28
                                         3rd Qu.:   352.60
                                         Max.   :278257.30
                                         NA's   :366
 Current.liabilities...provisions Deferred.tax.liability Shareholders.funds
 Min.    :     0.1                Min.    :    0.1        Min.    :     0.0
 1st Qu.:    17.8                1st Qu.:    3.2        1st Qu.:    32.0
 Median :    69.4                Median :   13.4        Median :   105.6
 Mean   :   940.6                Mean   :  227.2        Mean   :  1322.1
 3rd Qu.:   261.7                3rd Qu.:   50.0        3rd Qu.:   393.2
 Max.   :352240.3                Max.   :72796.6        Max.   :613151.6
 NA's   :96                      NA's   :1140
 Cumulative.retained.profits Capital.employed      TOL.TNW
 Min.    : -6534.3           Min.    :     0.0   Min.    :-350.480
 1st Qu.:     1.1            1st Qu.:    60.8   1st Qu.:   0.600
 Median :    37.1            Median :   214.7   Median :   1.430
 Mean   :   890.5            Mean   :  2328.3   Mean   :   3.994
 3rd Qu.:   202.3            3rd Qu.:   767.3   3rd Qu.:   2.830
 Max.   :390133.8            Max.   :891408.9   Max.   : 473.000
 NA's   :38
```

```
 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.... Contingent.liabilities
 Min.    :    0.00                      Min.    :     0.1
 1st Qu.:    0.00                       1st Qu.:     6.3
 Median :    5.33                       Median :    38.0
 Mean   :   53.94                       Mean   :   932.9
 3rd Qu.:   30.76                       3rd Qu.:   192.7
 Max.   :14704.27                       Max.   :559506.8
                                        NA's   :1188
 Net.fixed.assets    Investments        Current.assets     Net.working.capital
 Min.    :     0.0   Min.    :     0.00   Min.    :     0.1   Min.    :-63839.0
 1st Qu.:    26.0   1st Qu.:     1.00   1st Qu.:    36.2   1st Qu.:    -1.1
 Median :    93.5   Median :     8.35   Median :   145.1   Median :    16.2
 Mean   :  1189.7   Mean   :   694.73   Mean   :  1293.4   Mean   :   138.6
 3rd Qu.:   344.9   3rd Qu.:    64.30   3rd Qu.:   502.2   3rd Qu.:    84.2
 Max.   :636604.6   Max.   :199978.60   Max.   :354815.2   Max.   : 85782.8
 NA's   :118        NA's   :1435        NA's   :66         NA's   :32
```

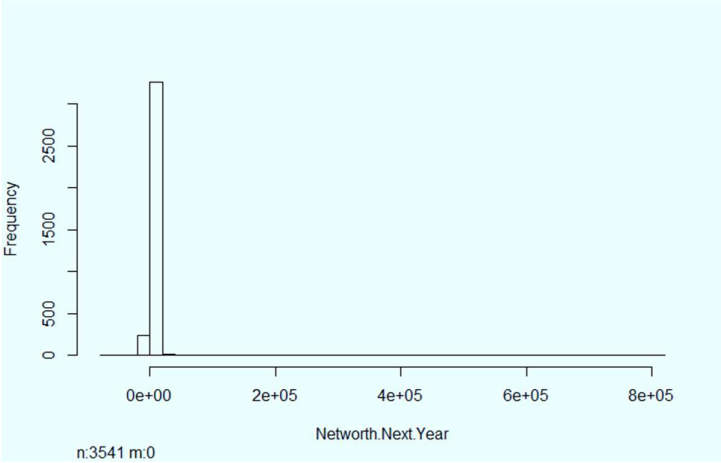
```
 Quick.ratio..times. Current.ratio..times. Debt.to.equity.ratio..times.
 Min.    :  0.000    Min.    :  0.00      Min.    :  0.00
 1st Qu.:  0.410     1st Qu.:  0.93       1st Qu.:  0.22
 Median :  0.670     Median :  1.23       Median :  0.79
 Mean   :  1.401     Mean   :  2.13       Mean   :  2.78
 3rd Qu.:  1.030     3rd Qu.:  1.71       3rd Qu.:  1.75
 Max.   :341.000     Max.   :505.00       Max.   :456.00
 NA's   :93          NA's   :93
 Cash.to.current.liabilities..times. Cash.to.average.cost.of.sales.per.day
 Min.    :  0.0000                   Min.    :     0.00
 1st Qu.:  0.0200                    1st Qu.:     2.79
 Median :  0.0700                    Median :     8.03
 Mean   :  0.4904                    Mean   :   158.44
 3rd Qu.:  0.1900                    3rd Qu.:    21.79
 Max.   :165.0000                    Max.   :128040.76
 NA's   :93                          NA's   :85
 Creditors.turnover Debtors.turnover   Finished.goods.turnover
 Length:3541        Length:3541        Length:3541
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

➔ Through Graphical Representaion

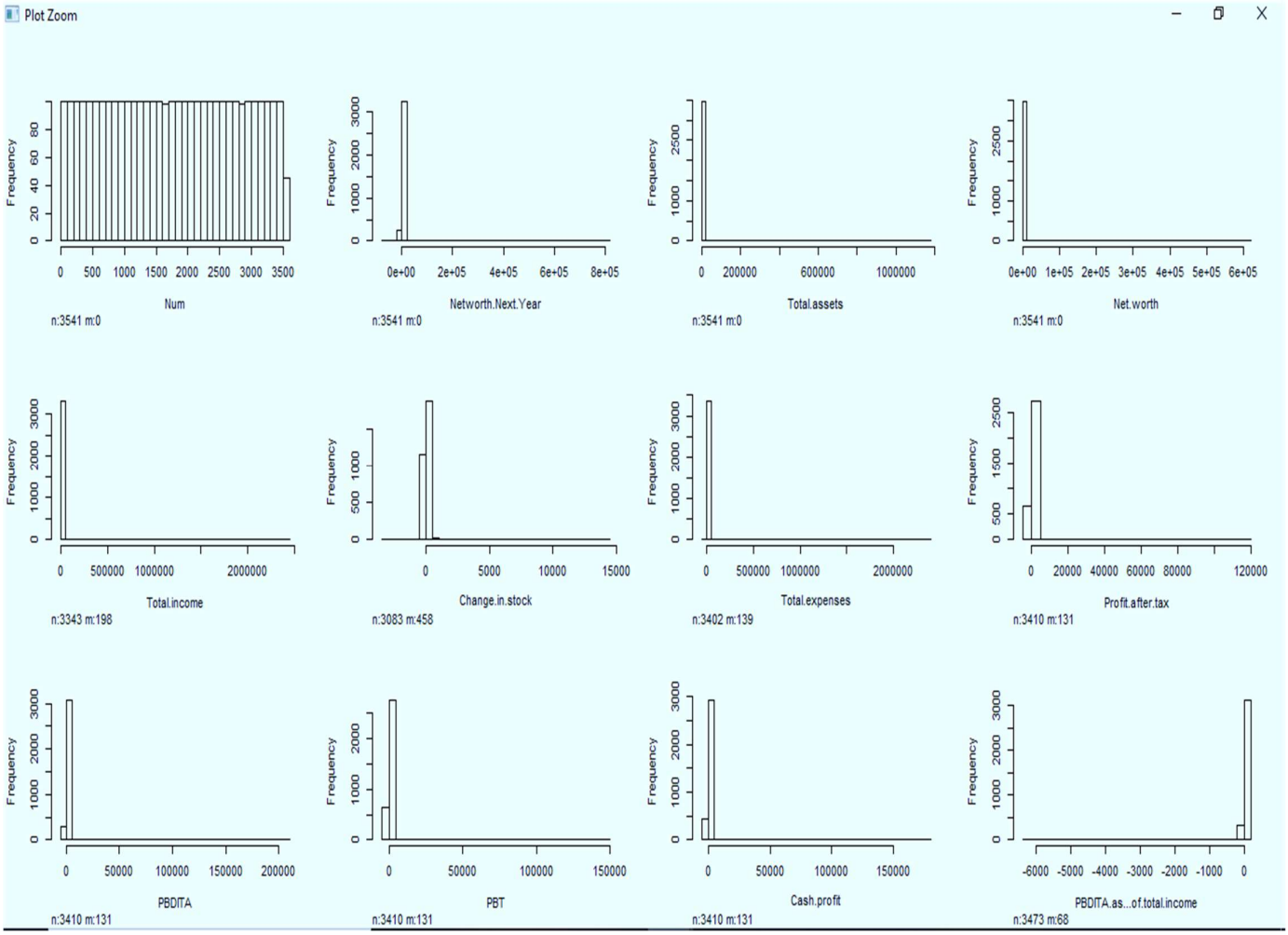
2.2.1 UNIVARIATE ANALYSIS

Done for both categorical and continuous variables. Must be numeric for histogram plotting.

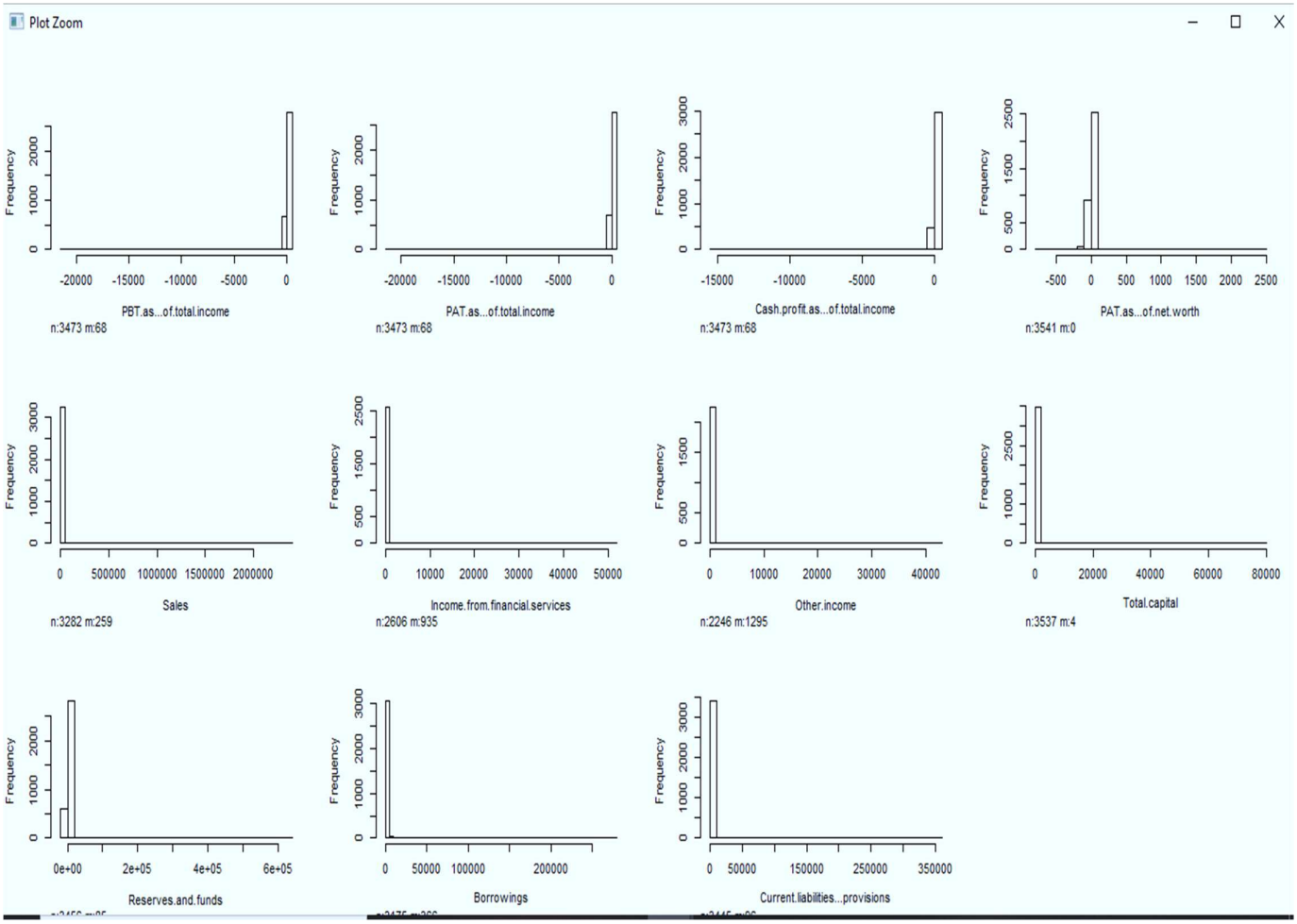
hist(data[2])



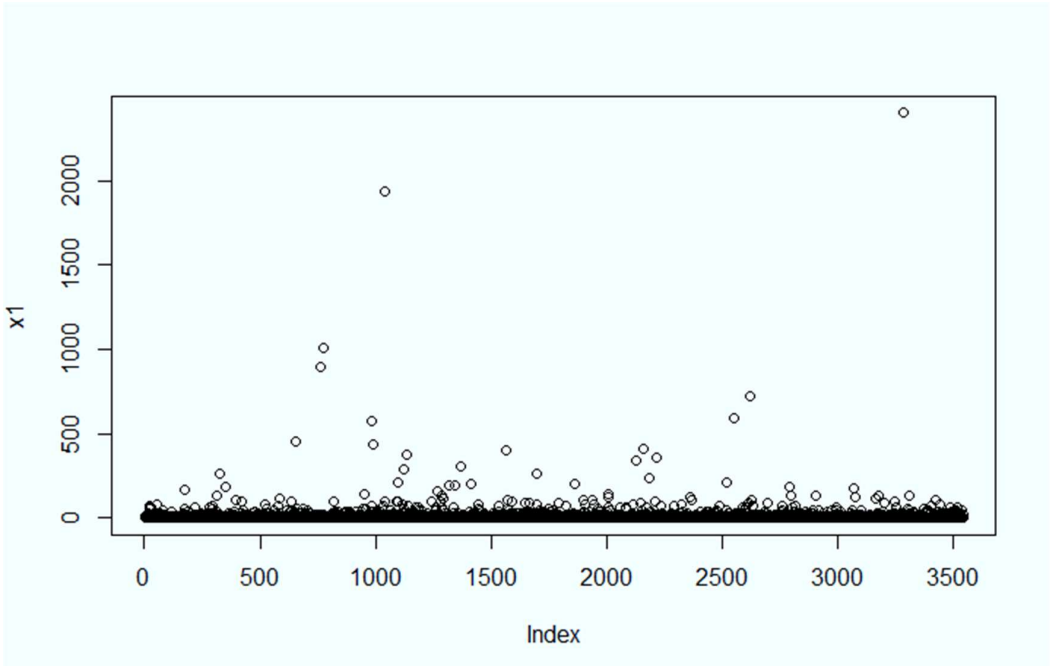
hist(data[1:12])



hist(data[13:24])



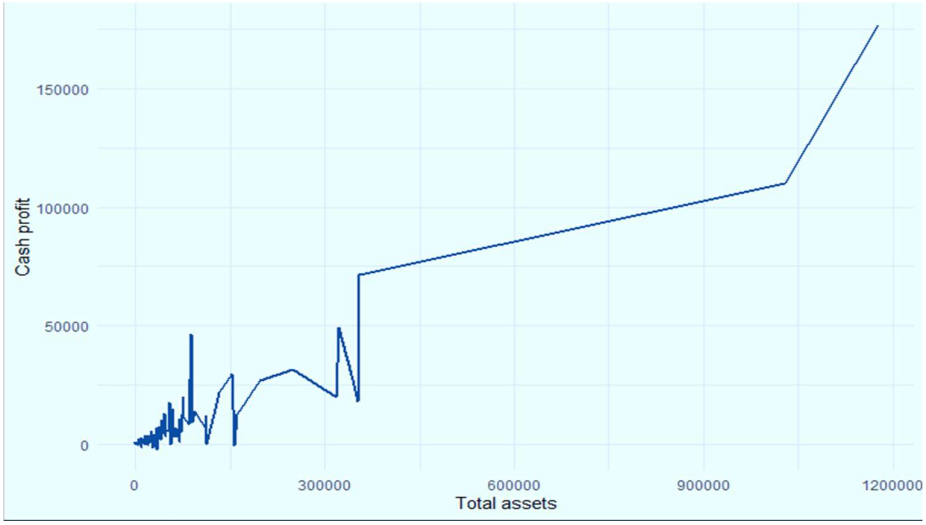
plot(data[,42]) : Creditors Turnover



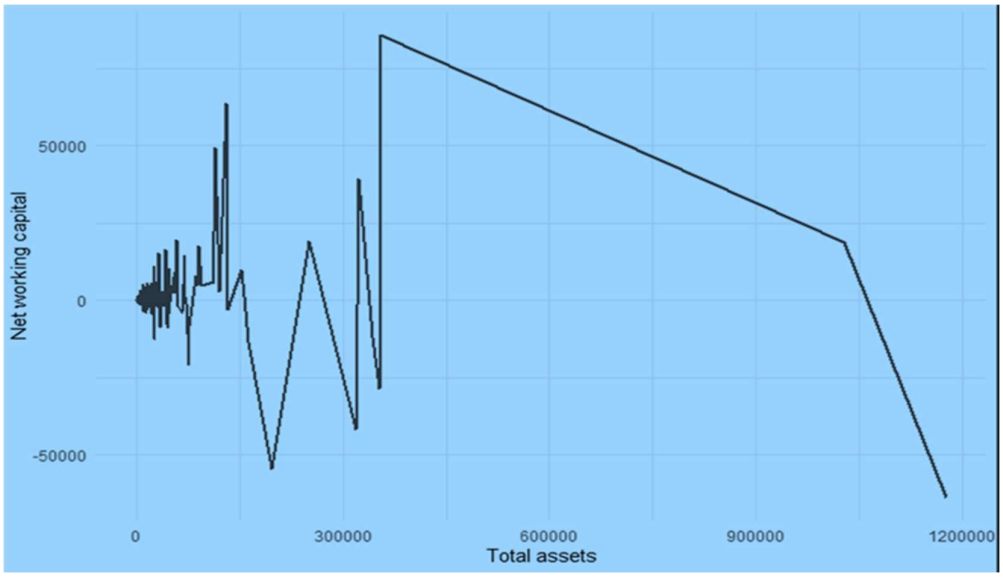
2.2.2 BIVARIATE ANALYSIS

We can use both LINE and SCATTER Plots for Bivariate A

```
# Total Assets vs Cash Profit
ggplot(data) +
  aes(x = `Total assets`, y = `Cash profit`) +
  geom_line(size = 1L, colour = "#0c4c8a") +
  theme_minimal()
```

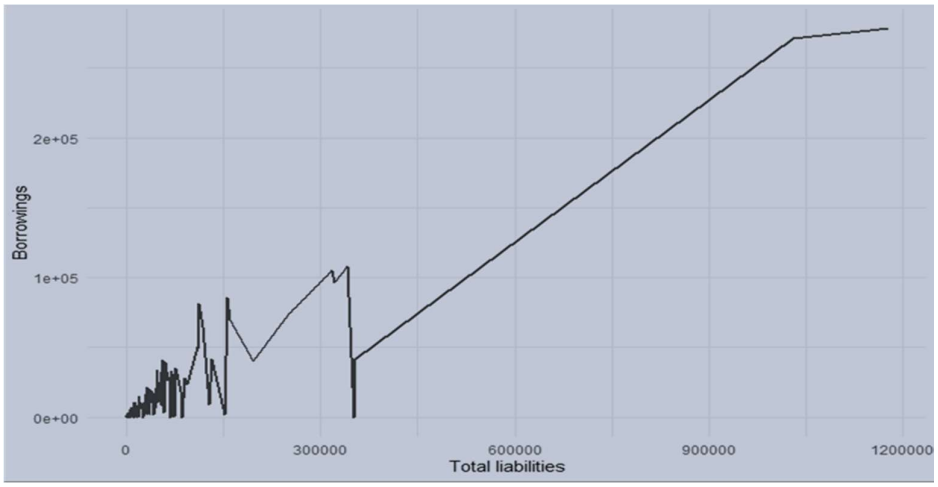



```
# Total Assets vs Net Working Capital
ggplot(data) +
  aes(x = `Total assets`, y = `Net working capital`) +
  geom_line(size = 1L, colour = "#0c4c8a") +
  theme_minimal()
```



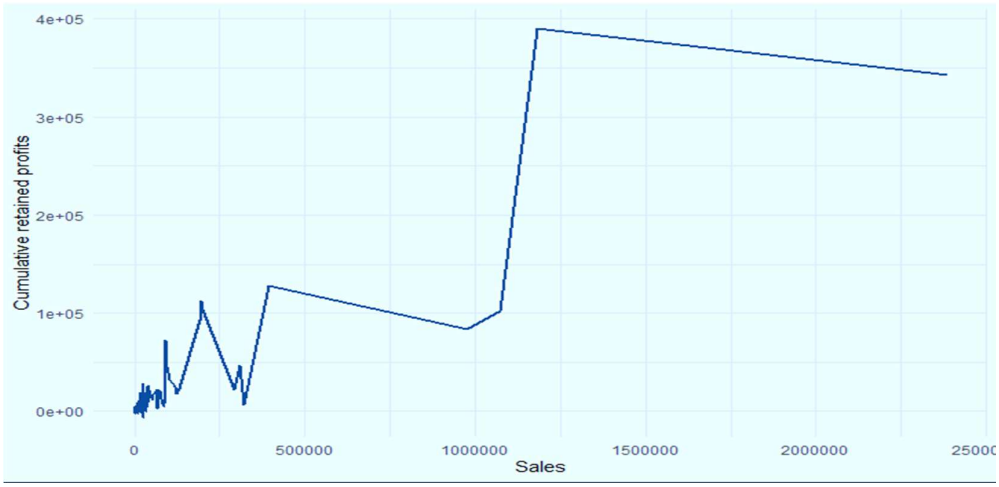
Total Liabilities vs Borrowings

```
ggplot(data) +
  aes(x = `Total liabilities`, y = Borrowings) +
  geom_line(size = 1L, colour = "#0c4c8a") +
  theme_minimal()
```



Sales Vs Cumulative Retained Profits

```
ggplot(data) +
  aes(x = Sales, y = `Cumulative retained profits`) +
  geom_line(size = 1L, colour = "#0c4c8a") +
  theme_minimal()
```

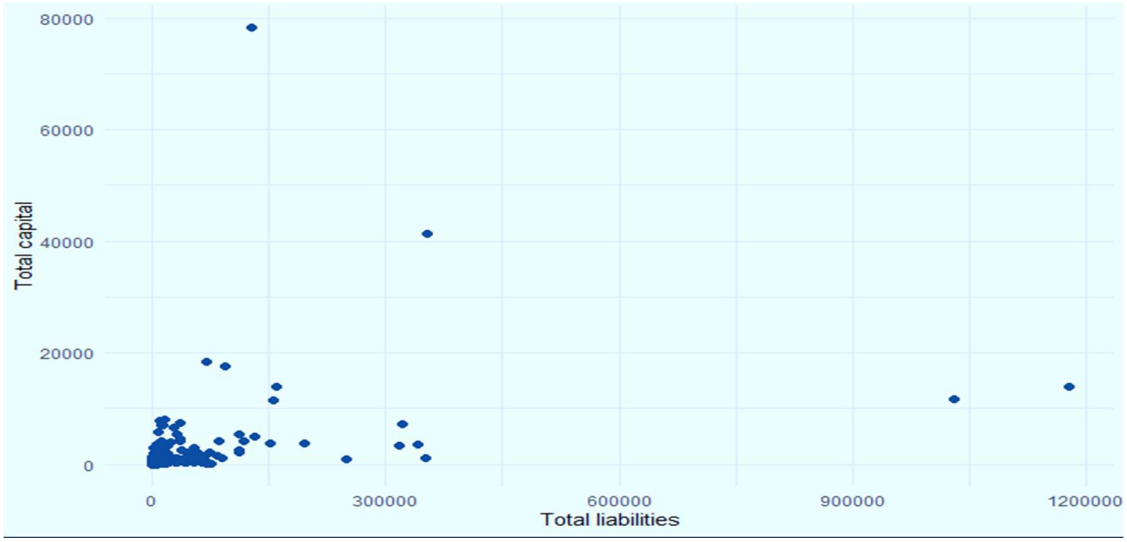


Cash Profit vs Current Liabilities and provisions

```
ggplot(data) +
  aes(x = `Cash profit`, y = `Current liabilities & provisions`) +
  geom_line(size = 1.76, colour = "#0c4c8a") +
  theme_minimal()
```

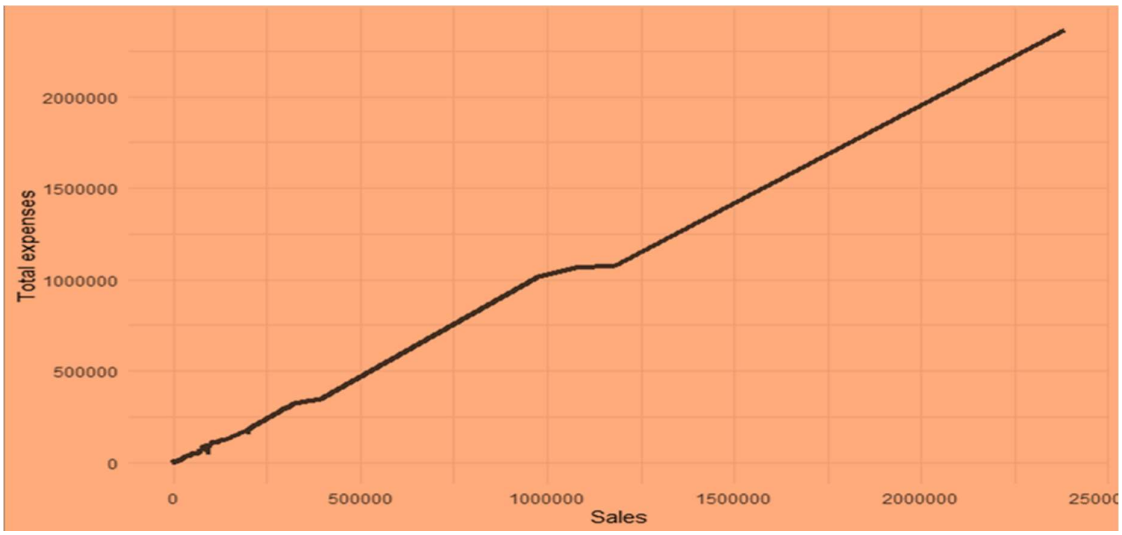
Total Capital vs Total Liabilities

```
ggplot(data) +
  aes(x = `Total liabilities`, y = `Total capital`) +
  geom_point(size = 2L, colour = "#0c4c8a") +
  theme_minimal()
```



Total Assests vs Total Liabilities

```
ggplot(data) +
  aes(x = Sales, y = `Total expenses`) +
  geom_line(size = 1.62, colour = "#0c4c8a") +
  theme_minimal()
plot(data$`Total assets`,data$`Total liabilities`)
```

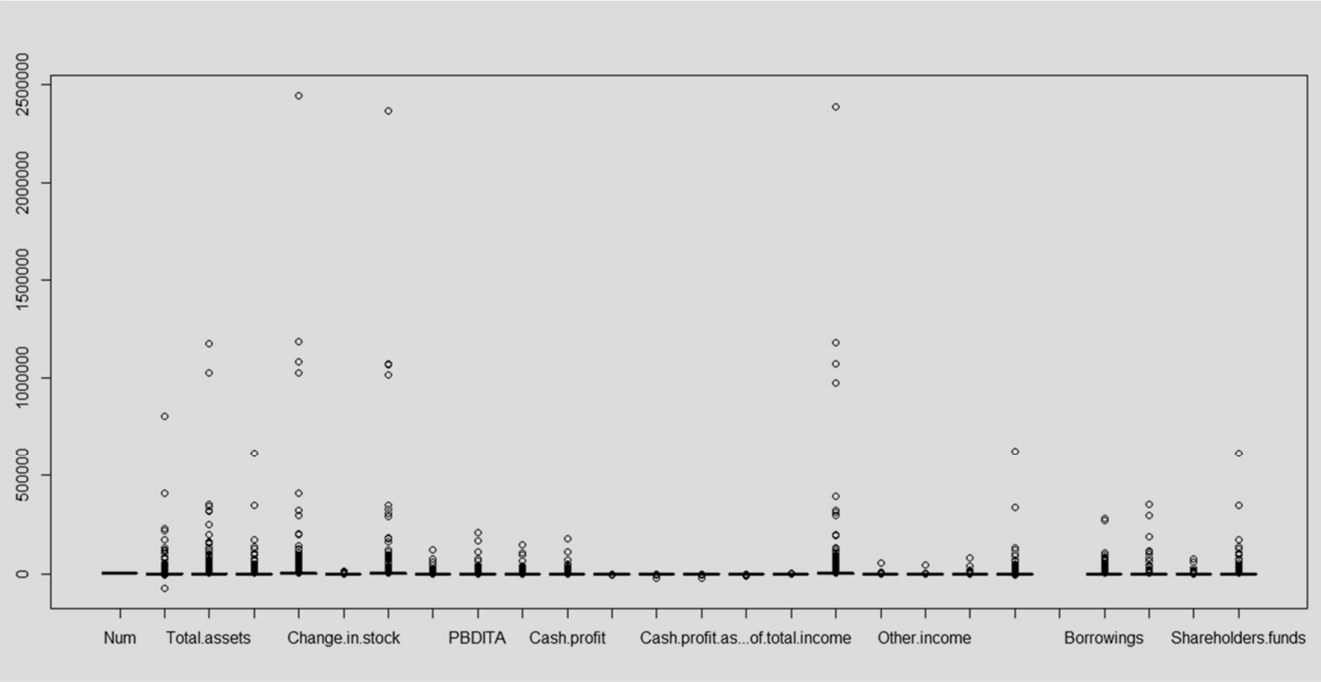


3. CHECK FOR INTEGRITY OF DATA

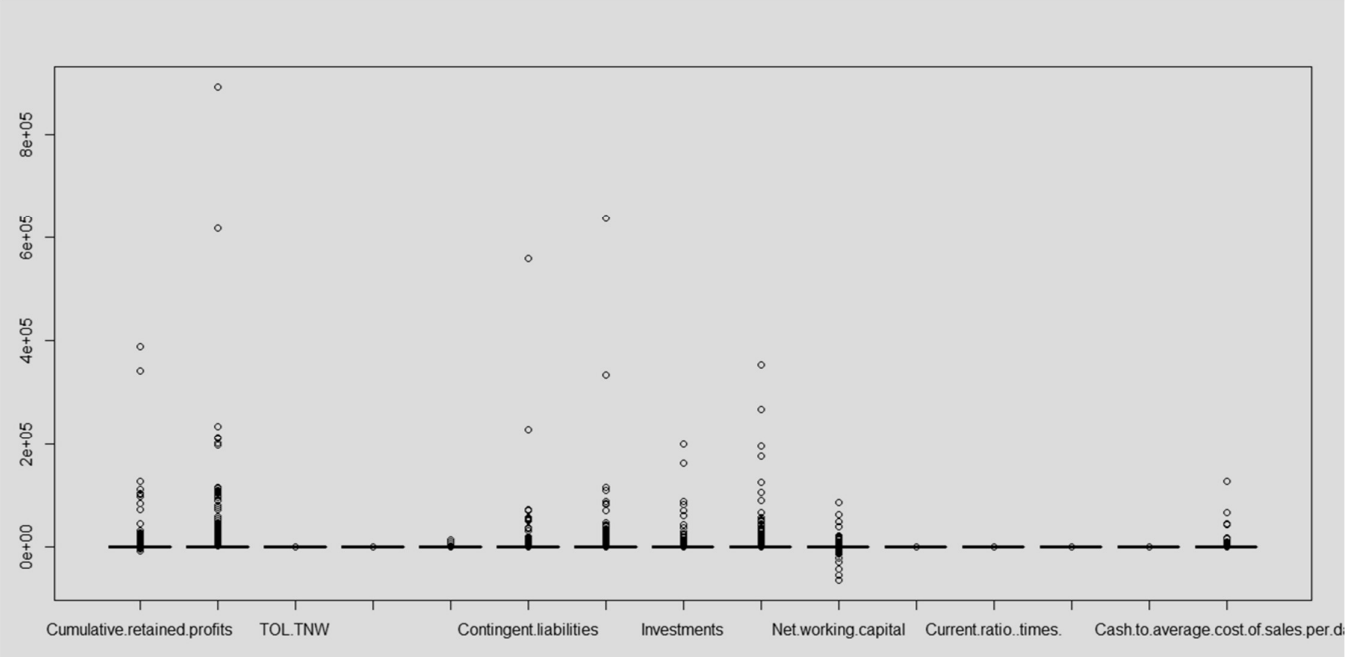
3.1 OUTLIER DETECTION

Depicts that there are outliers in our data. These plots are being made after seeing the correlation matrix and thus, figuring out which variables are to an extent related to each other and then we can visualize their relationship.

boxplot(data[,(1:26)])



boxplot(data[,(27:41)])



3.2 MISSING VALUE DETECTION

In R the missing values are coded by the symbol NA . To identify missings in your dataset the function is is.na()

colSums(is.na(data))

```
> colSums(is.na(data))
                                     Num                        Networth.Next.Year
                                       0                                         0
                            Total.assets                                Net.worth
                                       0                                         0
                            Total.income                          Change.in.stock
                                     198                                       458
                          Total.expenses                          Profit.after.tax
                                     139                                       131
                                  PBDITA                                      PBT
                                     131                                       131
                             Cash.profit                 PBDITA.as...of.total.income
                                     131                                        68
                  PBT.as...of.total.income                  PAT.as...of.total.income
                                      68                                        68
         Cash.profit.as...of.total.income                   PAT.as...of.net.worth
                                      68                                         0
                                   Sales             Income.from.financial.services
                                     259                                       935
                           Other.income                           Total.capital
                                    1295                                         4
                     Reserves.and.funds       Deposits..accepted.by.commercial.banks.
                                      85                                      3541
                             Borrowings          Current.liabilities...provisions
                                     366                                        96
                  Deferred.tax.liability                        Shareholders.funds
                                    1140                                         0
              Cumulative.retained.profits                          Capital.employed
                                      38                                         0
```

```
                                  TOL.TNW Total.term.liabilities...tangible.net.worth
                                        0                                           0
      Contingent.liabilities...Net.worth....                      Contingent.liabilities
                                        0                                        1188
                          Net.fixed.assets                                 Investments
                                      118                                        1435
                           Current.assets                           Net.working.capital
                                       66                                          32
                     Quick.ratio..times.                          Current.ratio..times.
                                       93                                          93
             Debt.to.equity.ratio..times.          Cash.to.current.liabilities..times.
                                        0                                          93
       Cash.to.average.cost.of.sales.per.day                          Creditors.turnover
                                       85                                          47
                         Debtors.turnover                       Finished.goods.turnover
                                       42                                         454
                           WIP.turnover                           Raw.material.turnover
                                      354                                          75
                      Shares.outstanding                              Equity.face.value
                                        0                                           0
                                      EPS                                  Adjusted.EPS
                                        0                                           0
                    Total.liabilities                                       PE.on.BSE
                                        0                                          23
```

4. CREATING THE DEFAULT VARIABLE

The default variable (dependent Variable) has to be created by using the Networth Next Year variable of the dataset.

```
data$default= ifelse(`Networth Next Year`<0,1,0)
names(data)
```

```
> names(data)
 [1] "Num"                                      "Networth.Next.Year"
 [3] "Total.assets"                             "Net.worth"
 [5] "Total.income"                             "Change.in.stock"
 [7] "Total.expenses"                           "Profit.after.tax"
 [9] "PBDITA"                                   "PBT"
[11] "Cash.profit"                              "PBDITA.as...of.total.income"
[13] "PBT.as...of.total.income"                 "PAT.as...of.total.income"
[15] "Cash.profit.as...of.total.income"         "PAT.as...of.net.worth"
[17] "Sales"                                    "Income.from.financial.services"
[19] "Other.income"                             "Total.capital"
[21] "Reserves.and.funds"                       "Deposits..accepted.by.commercial.banks."
[23] "Borrowings"                               "Current.liabilities...provisions"
[25] "Deferred.tax.liability"                   "Shareholders.funds"
[27] "Cumulative.retained.profits"              "Capital.employed"
[29] "TOL.TNW"                                  "Total.term.liabilities...tangible.net.worth"
[31] "Contingent.liabilities...Net.worth...."   "Contingent.liabilities"
[33] "Net.fixed.assets"                         "Investments"
[35] "Current.assets"                           "Net.working.capital"
[37] "Quick.ratio..times."                      "Current.ratio..times."
[39] "Debt.to.equity.ratio..times."             "Cash.to.current.liabilities..times."
[41] "Cash.to.average.cost.of.sales.per.day"    "Creditors.turnover"
[43] "Debtors.turnover"                         "Finished.goods.turnover"
[45] "WIP.turnover"                             "Raw.material.turnover"
[47] "Shares.outstanding"                       "Equity.face.value"
[49] "EPS"                                      "Adjusted.EPS"
[51] "Total.liabilities"                        "PE.on.BSE"
[53] "default"
>
```

➔ REMOVING THE VARIABLES WITH HIGH MISSING VALUES AND SOME WHICH AREN'T AS IMPORTANT BY BANK'S PERSPECTIVE

VARIABLES REMOVED INITIALLY :

'Num', 'Networth Next Year', 'Change in Stock','Income from Financial Services',

'Other income','Deposits accepted by commercial banks','Investments',

'Creditors Turnover','Debtors Turnover','Finished Goods Turnover',

'Wip Turover' 'Raw Material Turnover','Equity Face Value','PE on BSE'

```
dataM= data[,-c(1,2,6,18,19,22,25,32,34,42:48,52)]
names(dataM)
```

```
> names(dataM)
 [1] "Total.assets"                                "Net.worth"
 [3] "Total.income"                                "Total.expenses"
 [5] "Profit.after.tax"                            "PBDITA"
 [7] "PBT"                                         "Cash.profit"
 [9] "PBDITA.as...of.total.income"                 "PBT.as...of.total.income"
[11] "PAT.as...of.total.income"                    "Cash.profit.as...of.total.income"
[13] "PAT.as...of.net.worth"                       "Sales"
[15] "Total.capital"                               "Reserves.and.funds"
[17] "Borrowings"                                  "Current.liabilities...provisions"
[19] "Shareholders.funds"                          "Cumulative.retained.profits"
[21] "Capital.employed"                            "TOL.TNW"
[23] "Total.term.liabilities...tangible.net.worth" "Contingent.liabilities...Net.worth...."
[25] "Net.fixed.assets"                            "Current.assets"
[27] "Net.working.capital"                         "Quick.ratio..times."
[29] "Current.ratio..times."                       "Debt.to.equity.ratio..times."
[31] "Cash.to.current.liabilities..times."         "Cash.to.average.cost.of.sales.per.day"
[33] "EPS"                                         "Adjusted.EPS"
[35] "Total.liabilities"                           "default"
>
```

- TWO RATIOS CREATED IN ORDER FOR MICE TO WORK

Before Imputation using MICE we will have to remove the super correlated variables. On way is to create ratios with some other variables, before removing them so as to retain their importance in our dataset .

Sales and Total Income were super correlated with other Variables in our dataset. So, we made these 2 into ratios with other variables to reduce the collinearity.

1. Sales.per.TotalExpenses
2. Totalincome.per.TotalAssets

➔ Variables removed after these ratios creation are :
Total income and Total expenses

```
dataM=dataM[,-c(3,4)]
```

5. MISSING VALUE TREATMENT

5.1 IMPUTATION USING MICE

```
NAdata= mice(dataM,meth="pmm",seed=848,maxit = 5,m=5)

NAdata$imp

NAdata$loggedEvents

NAdata <- complete(NAdata)


colSums(is.na(NAdata))
```

```
> colSums(is.na(NAdata))
                                Total.assets                                   Net.worth
                                           0                                           0
                              Profit.after.tax                                      PBDITA
                                           0                                           0
                                         PBT                                 Cash.profit
                                           0                                           0
                  PBDITA.as...of.total.income                 PBT.as...of.total.income
                                           0                                           0
                     PAT.as...of.total.income           Cash.profit.as...of.total.income
                                           0                                           0
                        PAT.as...of.net.worth                                       Sales
                                           0                                           0
                                Total.capital                       Reserves.and.funds
                                           0                                           0
                                   Borrowings           Current.liabilities...provisions
                                           0                                           0
                         Shareholders.funds                Cumulative.retained.profits
                                           0                                           0
                           Capital.employed                                     TOL.TNW
                                           0                                           0
Total.term.liabilities...tangible.net.worth     Contingent.liabilities...Net.worth....
                                           0                                           0
                               Net.fixed.assets                             Current.assets
                                           0                                           0
                          Net.working.capital                          Quick.ratio..times.
                                           0                                           0
                        Current.ratio..times.               Debt.to.equity.ratio..times.
                                           0                                           0
       Cash.to.current.liabilities..times.      Cash.to.average.cost.of.sales.per.day
                                           0                                           0
                                         EPS                                 Adjusted.EPS
                                           0                                           0
                          Total.liabilities                                     default
                                           0                                           0
                      Sales.per.TotalExpenses                Totalincome.per.TotalAssets
                                           0                                           0
```

6. CREATION OF NEW RATIOS FOR RAW DATASET

6.1 Ratios For Profitability

CashProfit.per.TotalAssets

Cum.Profit.per.Sales

Sales.per.TotalAssets

Sales.per.CurentAssets

PBT.per.TotalCaptital

PAT.per.TotalAssets

PAT.per.Sales

PAT.per.TotalCapital

PAT.per.TotalAssets

PBT.per.TotalAssests

PBT.per.Sales

PBT.per.TotalCapital

PBDITA.per.Sales

6.2 Ratios For Size

ShareholderFunds.per.TotalAssets

Captialemployed.per.Networth

Reserves.per.TotalCapital

ShareholderFunds.per.TotalLiability

ShareholderFunds.per.TotalCapital

capitalEmp..per.ShareholderFunds

NetFixedAssets.per.CurrentAssets

NetFixedAssets.per.TotalAssets

6.3 Ratios For Leverage

Borrowings.per.TotalAssets

Borrowings.per.TotalLiabilities

Borrowings.per.TotalCapital

CurrentLiability.per.CurrentAssets

CurrentLiability.per.TotalAssets

6.4 Ratios For Liquidity

NetFixedAssets.per.TotalLiabilities

Cum.RetainedProfits.per.CurrentAssets

Networth.per.TotalLiabilities

Cum.RetainedProfit.per.TotalLiabilitie

PAT.per.CurrentLiability

CashProfit.per.CurrentLiabilities

TotalCapital.per.TotalLiabilities

TotalCapital.per.CurrentLiabilities

7. OUTLIERS TREATMENT

In this case, the outliers are not wrong values, so we cannot discard the outliers because they actually are true and correct values, however extreme they might be. So we will have to treat the outliers.

We will cap the outliers using flooring and ceiling technique.

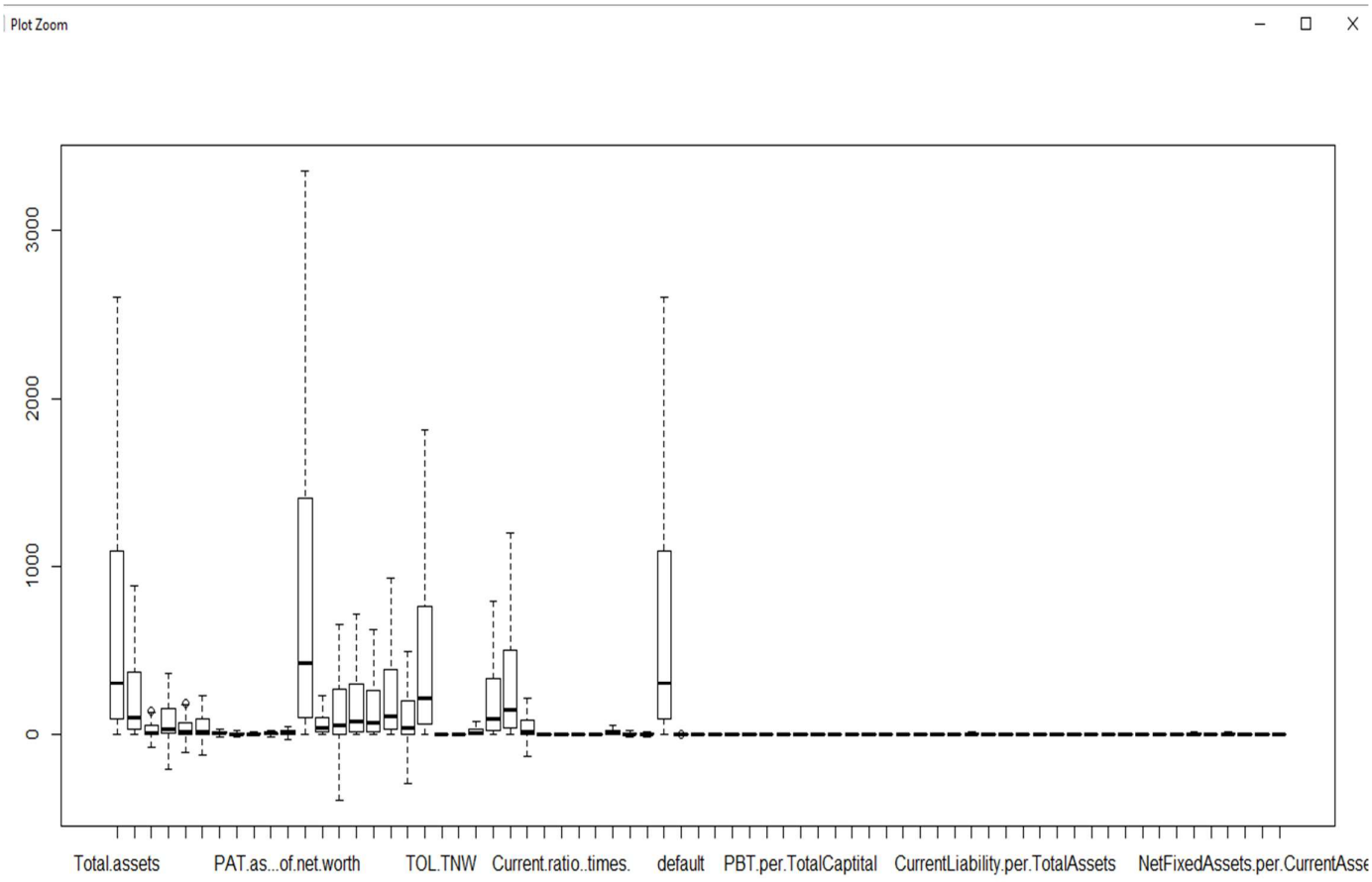
7.1 CAPPING THE OUTLIERS

```
a=c(1:33,35:69)

for(val in a){
  qnt<- quantile(outdata[,val],probs = c(0.25,0.75))
  cap<- quantile(outdata[,val],probs = c(0.10,0.85))

  h= 1.5*IQR(outdata[,val])
  outdata[,val][outdata[,val]>(qnt[2]+h)]<- cap[2]
  outdata[,val][outdata[,val]<(qnt[1]-h)]<- cap[1]
}
```

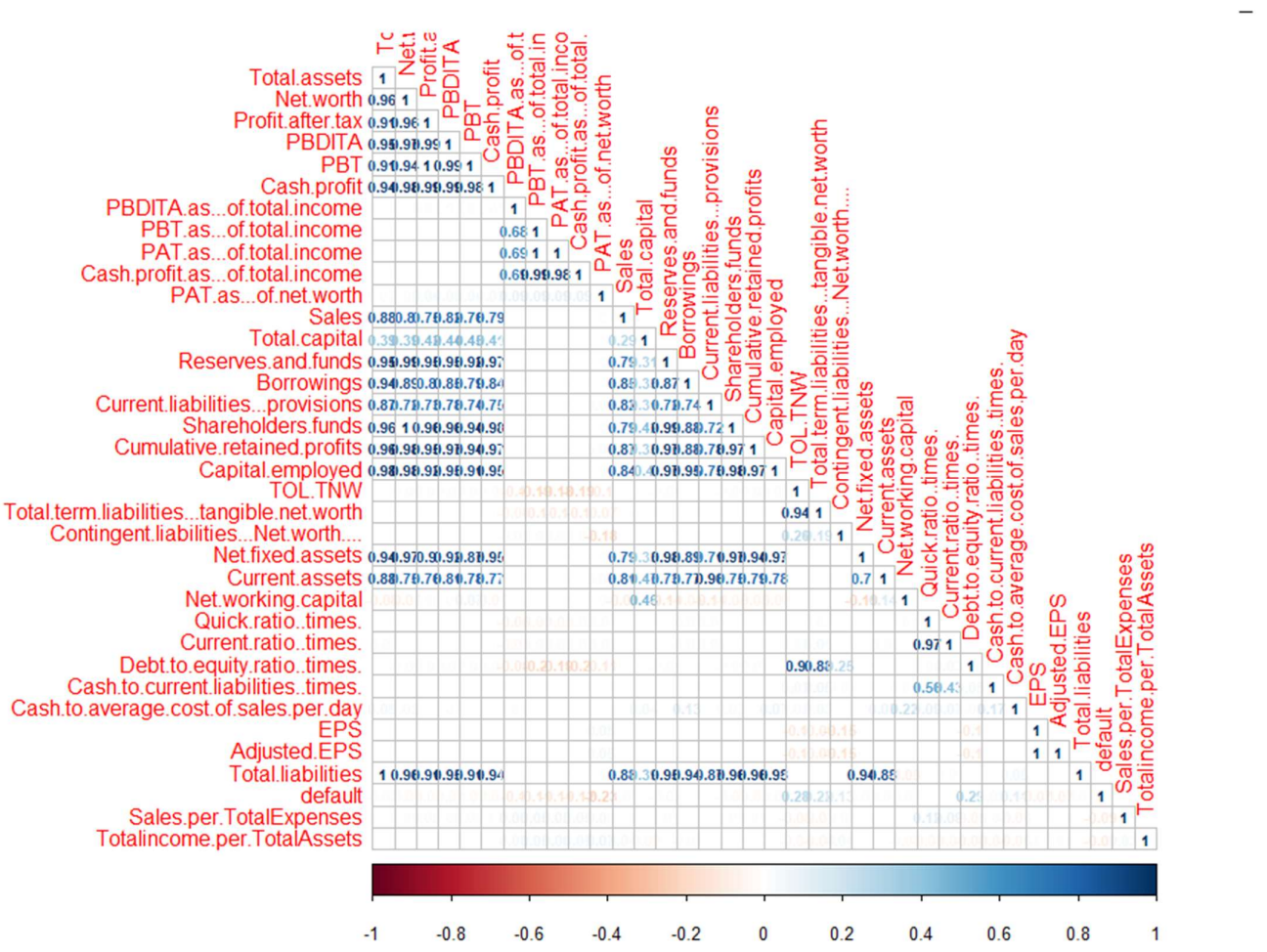
```
boxplot(outdata)
```



*We can observe that the outliers have been treated well.

8. MULTICOLLINEARITY TREATMENT

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when we fit the model and interpret the results.



➔STEPS TO REMOVING MULTICOLLINEARITY IN THIS SITUATION

8.1 Creating A Logistic Model

Firstly we will create a regression model so that we can find the Variance Inflation Factor for each of the variables.

```
m1=  glm(dataset1$default~.,data=dataset1,family = binomial())
```

8.2 Using VIF To Remove Highly Correlated Variables

Variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis

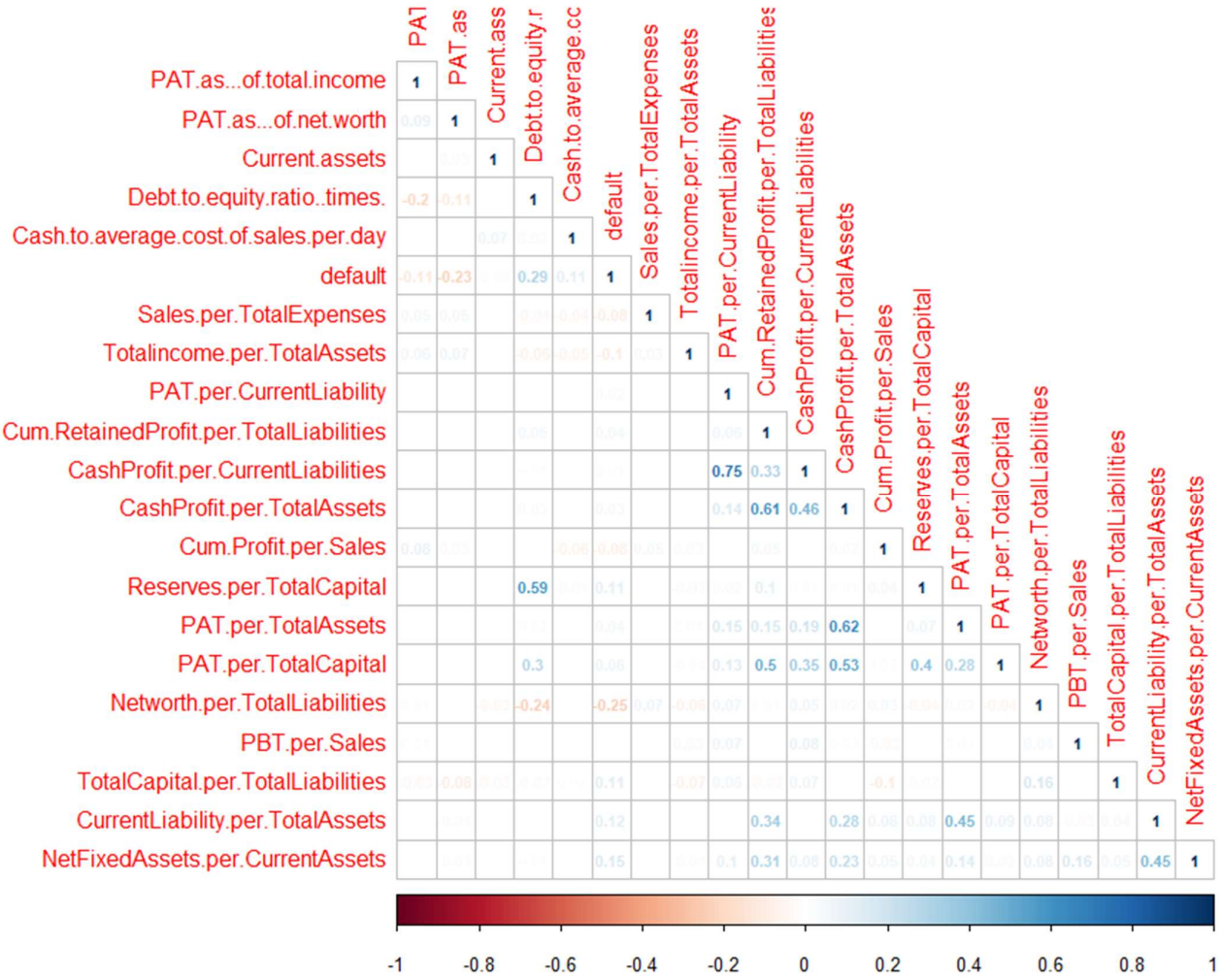
Given Below are the variable which we will remove after RECURSIVELY checking the VIF and building the model over again.

```
[1] "Total.assets"
 [2] "Net.worth"
 [3] "Profit.after.tax"
 [4] "PBDITA"
 [5] "PBT"
 [6] "Cash.profit"
 [7] "Sales"
 [8] "Total.capital"
 [9] "Reserves.and.funds"
[10] "Borrowings"
[11] "Current.liabilities...provisions"
[12] "Shareholders.funds"
[13] "Cumulative.retained.profits"
[14] "Capital.employed"
[15] "EPS"
[16] "Adjusted.EPS"
[17] "Borrowings.per.TotalLiabilities"
[18] "Captialemployed.per.Networth"
[19] "PBT.per.TotalCaptital"
[20] "ShareholderFunds.per.TotalLiability"
[21] "capitalEmp..per.ShareholderFunds"
[22] "NetFixedAssets.per.TotalAssets"
```

8.3 Using StepAIC To Remove The Variables In Such A Way As To Give The Lowest Value Of AIC

Step AIC will build a model for us that is the best i.e., has the lowest AIC value. Therefore, after executing the StepAIC function for our model we'll get the significant variables which will build the most accurate model for us.

Using the Variables which are shown significant by AIC are used as final variables. We will check for the presence of multicollinearity between those variables.



*we can observe that the Multicollinearity has been treated properly.

9. SELECTION OF SIGNIFICANT VARIABLES FOR MODEL BUILDING

9.1 USING STEPAIC

In stepwise regression, the selection procedure is automatically performed by statistical packages. The criteria for variable selection include adjusted R-square, Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows's Cp, PRESS, or false discovery rate (1,2).

```
stepAIC(m1,direction = "backward")


train= outdata[,c(9,11,24,28,30,34,35,36,38,44,45,46,47,49,
                  51,53,54,56,58,61,66)]
```

Given below are the variables which are significant for our model and are finally selected

```
[1] "PAT.as...of.total.income"
 [2] "PAT.as...of.net.worth"
 [3] "Current.assets"
 [4] "Debt.to.equity.ratio..times."
 [5] "Cash.to.average.cost.of.sales.per.day"
 [6] "default"
 [7] "Sales.per.TotalExpenses"
 [8] "Totalincome.per.TotalAssets"
 [9] "PAT.per.CurrentLiability"
[10] "Cum.RetainedProfit.per.TotalLiabilities"
[11] "CashProfit.per.CurrentLiabilities"
[12] "CashProfit.per.TotalAssets"
[13] "Cum.Profit.per.Sales"
[14] "Reserves.per.TotalCapital"
[15] "PAT.per.TotalAssets"
[16] "PAT.per.TotalCapital"
[17] "Networth.per.TotalLiabilities"
[18] "PBT.per.Sales"
[19] "TotalCapital.per.TotalLiabilities"
[20] "CurrentLiability.per.TotalAssets"
[21] "NetFixedAssets.per.CurrentAssets"


>
```

10. BALANCING THE DATASET

The dataset containes the default is unbalanced proportion. The default constitutes only 6.7% of the total dataset . We cannot build a model on this unbalanced dataset. The dependent variable's proportion in the dataset has to levelled.

10.1 SMOTE Method

It is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.

Applying smote on our Final Dataset using Default as an argument so that balancing happens according to the Depemdent variable.

```
smote.train= SMOTE(train$default~., train)
```

```
dim(smote.train)
```

```
[1] 1638   21
```

*Checking the proportion of Default in the Smote Dataset

The dependent variable i.e., is now balanced.

```
table(smote.train$default)
```

```
prop.table(table(smote.train$default))
```

```
> table(smote.train$default)

  0   1
936 702
> prop.table(table(smote.train$default))

        0         1
0.5714286 0.4285714
```

11. CREATING THE FINAL MODEL USING SMOTE DATASET

11.1 APPLYING LOGISTIC REGRESSION

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.
In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

```
FinalModel=  glm(smote.train$default~.,data=smote.train,family = binomial())
summary(FinalModel)
```

```
Call:
glm(formula = smote.train$default ~ ., family = binomial(), data = smote.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7918  -0.4297  -0.0592   0.4327   2.5511

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          6.2665430  1.4991924   4.180 2.92e-05 ***
PAT.as...of.total.income            -0.1046997  0.0375532  -2.788 0.005303 **
PAT.as...of.net.worth               -0.0598362  0.0105202  -5.688 1.29e-08 ***
Current.assets                      -0.0010733  0.0003194  -3.360 0.000779 ***
Debt.to.equity.ratio..times.        -0.4418666  0.1551694  -2.848 0.004404 **
Cash.to.average.cost.of.sales.per.day  0.0068377  0.0056206   1.217 0.223777
Sales.per.TotalExpenses             -7.6625502  1.4684969  -5.218 1.81e-07 ***
Totalincome.per.TotalAssets         -0.1461148  0.1464947  -0.997 0.318567
Cum.Profit.per.Sales                -2.4602507  0.5972662  -4.119 3.80e-05 ***
PAT.per.TotalCapital                -0.1244917  0.3460960  -0.360 0.719069
PBT.per.TotalAssests                -2.9025425  3.0526663  -0.951 0.341695
PBT.per.Sales                       -0.3684769  3.3586633  -0.110 0.912640
PBT.per.TotalCapital                 0.4790525  0.2629374   1.822 0.068466 .
Borrowings.per.TotalAssets           2.1311519  0.6258141   3.405 0.000661 ***
Borrowings.per.TotalCapital          0.2324643  0.0419003   5.548 2.89e-08 ***
CurrentLiability.per.TotalAssets     5.1476327  0.8407515   6.123 9.20e-10 ***
NetFixedAssets.per.TotalLiabilities  0.4213287  0.4231280   0.996 0.319373
Networth.per.TotalLiabilities       -3.5882468  0.6153106  -5.832 5.49e-09 ***
PAT.per.CurrentLiability             1.5074505  0.6993848   2.155 0.031131 *
TotalCapital.per.CurrentLiabilities  0.3878704  0.1292906   3.000 0.002700 **
ShareholderFunds.per.TotalCapital   -0.3276440  0.0500890  -6.541 6.10e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2237.2  on 1637  degrees of freedom
Residual deviance: 1036.3  on 1617  degrees of freedom
```

➔ vif(FinalModel)

```
> vif(FinalModel)
                PAT.as...of.total.income                  PAT.as...of.net.worth
                                2.658539                               1.912812
                           Current.assets          Debt.to.equity.ratio..times.
                                1.266561                               3.964886
  Cash.to.average.cost.of.sales.per.day               Sales.per.TotalExpenses
                                1.382232                               1.465924
           Totalincome.per.TotalAssets                  Cum.Profit.per.Sales
                                1.753987                               1.819487
                  PAT.per.TotalCapital                  PBT.per.TotalAssests
                                4.904557                               5.841236
                       PBT.per.Sales                  PBT.per.TotalCapital
                                4.866950                               5.131567
           Borrowings.per.TotalAssets             Borrowings.per.TotalCapital
                                2.859814                               3.134703
      CurrentLiability.per.TotalAssets    NetFixedAssets.per.TotalLiabilities
                                3.300038                               1.242574
         Networth.per.TotalLiabilities                PAT.per.CurrentLiability
                                3.014064                               3.986185
    TotalCapital.per.CurrentLiabilities      ShareholderFunds.per.TotalCapital
                                4.206676                               1.966321
```

*We can observe that the VIF values are all below 5

12.ANALYSIS OF THE COEFFICIENTS OF LOGISTIC MODEL

```
Call:
glm(formula = smote.train$default ~ ., family = binomial(), data = smote.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7918  -0.4297  -0.0592   0.4327   2.5511

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          6.2665430  1.4991924   4.180 2.92e-05 ***
PAT.as...of.total.income            -0.1046997  0.0375532  -2.788 0.005303 **
PAT.as...of.net.worth               -0.0598362  0.0105202  -5.688 1.29e-08 ***
Current.assets                      -0.0010733  0.0003194  -3.360 0.000779 ***
Debt.to.equity.ratio..times.        -0.4418666  0.1551694  -2.848 0.004404 **
Cash.to.average.cost.of.sales.per.day  0.0068377  0.0056206   1.217 0.223777
Sales.per.TotalExpenses             -7.6625502  1.4684969  -5.218 1.81e-07 ***
Totalincome.per.TotalAssets         -0.1461148  0.1464947  -0.997 0.318567
Cum.Profit.per.Sales                -2.4602507  0.5972662  -4.119 3.80e-05 ***
PAT.per.TotalCapital                -0.1244917  0.3460960  -0.360 0.719069
PBT.per.TotalAssests                -2.9025425  3.0526663  -0.951 0.341695
PBT.per.Sales                       -0.3684769  3.3586633  -0.110 0.912640
PBT.per.TotalCapital                 0.4790525  0.2629374   1.822 0.068466 .
Borrowings.per.TotalAssets           2.1311519  0.6258141   3.405 0.000661 ***
Borrowings.per.TotalCapital          0.2324643  0.0419003   5.548 2.89e-08 ***
CurrentLiability.per.TotalAssets     5.1476327  0.8407515   6.123 9.20e-10 ***
NetFixedAssets.per.TotalLiabilities  0.4213287  0.4231280   0.996 0.319373
Networth.per.TotalLiabilities       -3.5882468  0.6153106  -5.832 5.49e-09 ***
PAT.per.CurrentLiability             1.5074505  0.6993848   2.155 0.031131 *
TotalCapital.per.CurrentLiabilities  0.3878704  0.1292906   3.000 0.002700 **
ShareholderFunds.per.TotalCapital   -0.3276440  0.0500890  -6.541 6.10e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2237.2  on 1637  degrees of freedom
Residual deviance: 1036.3  on 1617  degrees of freedom
```

1. For every one unit change in PAT as of Total Income, the log odds of Default(versus non-Default) decreases by `0.104` since there is a –ve sign linked with it.
2. For every one unit change in PAT as of Net Worth, the log odds of Default(versus non-Default) decreases by `0.059` since there is a –ve sign linked with it.
3. For every one unit change in Current Assets, the log odds of Default(versus non-Default) decreases by `0.001` since there is a –ve sign linked with it.
4. For every one unit change in Debt to equity Ratio , the log odds of Default(versus non-Default) decreases by `0.441` since there is a –ve sign linked with it.
5. For every one unit change in Cash to average cost of sale per day, the log odds of Default(versus non-Default) increases by `0.006` .
6. For every one unit change in Sales per Total Expenses, the log odds of Default(versus non-Default) decreases by `7.662` since there is a –ve sign linked with it.
7. For every one unit change in Total Income per Total Assets, the log odds of Default(versus non-Default) decreases by `0.146` since there is a –ve sign linked with it.
8. For every one unit change in Cum.Profit per Sales, the log odds of Default(versus non-Default) decreases by `2.460` since there is a –ve sign linked with it.
9. For every one unit change in PAT per Total Capital, the log odds of Default(versus non-Default) decreases by `.124` since there is a –ve sign linked with it.
10. For every one unit change in PBT per Total Assets, the log odds of Default(versus non-Default) decreases by `2.902` since there is a –ve sign linked with it.
11. For every one unit change in PBT per Sales, the log odds of Default(versus non-Default) decreases by `0.368` since there is a –ve sign linked with it.

12. For every one unit change in PBT per Total Capital, the log odds of Default(versus non-Default) decreases by `.479` since there is a –ve sign linked with it.
13. For every one unit change in Borrowings per Total assets, the log odds of Default(versus non-Default) increases by `2.131` .
14. For every one unit change in Borrowings per Total Capital, the log odds of Default(versus non-Default) increases by `0.232` .
15. For every one unit change in Current Liability per Total Assets, the log odds of Default(versus non-Default) increases by `5.147` .
16. For every one unit change in Net Fixed Assets per Total Liability, the log odds of Default(versus non-Default) increases by `0.421`
17. For every one unit change in Net worth per Total Liability, the log odds of Default(versus non-Default) decreases by `3.588` since there is a –ve sign linked with it.
18. For every one unit change in PAT per Current Liability, the log odds of Default(versus non-Default) increases by `1.507` .
19. For every one unit change in Total Capital per Current Liability, the log odds of Default(versus non-Default) increases by `0.387` .
20. For every one unit change in Shareholder Funds per Total capital, the log odds of Default(versus non-Default) decreases by `.327` since there is a –ve sign linked with it.

13. PREDICTIONS + CONFUSION MATRIX ON THE DEVELOPED DATASET

```
pred= predict(FinalModel,smote.train)
pred.train= ifelse(pred<0.3,0,1)
```

Converting the Default and the pred.train into Factors

```
smote.train$default=as.factor(smote.train$default)
pred.train= as.factor(pred.train)
```

➔ Creating a Confusion Matrix for the predictions on the Train Data

```
caret::confusionMatrix(smote.train$default,pred.train)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 851  85
         1 127 575

               Accuracy : 0.8706
                 95% CI : (0.8533, 0.8865)
    No Information Rate : 0.5971
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7338

 Mcnemar's Test P-Value : 0.004864

            Sensitivity : 0.8701
            Specificity : 0.8712
         Pos Pred Value : 0.9092
         Neg Pred Value : 0.8191
             Prevalence : 0.5971
         Detection Rate : 0.5195
   Detection Prevalence : 0.5714
      Balanced Accuracy : 0.8707

       'Positive' Class : 0
```

Accuracy : 87.06%

Sensitivity : 87.01%

Specificity : 87.12%

14. PREPARING THE VALIDATION DATASET

The validation dataset has to be of the same structure, with the exactly same columns as in the Train dataset.

14.1 Importing the Validation set

```
test= read_xlsx("validation_data.xlsx")
```

- Removing the basic variables which are majoritarily missing

Because they won't serve purpose and if we impute them with missing values(through machine learning algorithms) , they would become sort of Artificial Data.

```
test= test[,-c(1,6,18,19,22,25,32,34,42:48,52)]
dim(test)
```

```
[1] 715  36
```

➔ Creating Two new ratios(same as we did for the Raw dataset)

1. Sales.per.TotalExpenses
2. Totalincome.per.TotalAssets

- Removing Total Income and Total expenses (like we did in Developer Data set)

```
test=test[,-c(5,4)]
```

14.2 DESCRIPTIVE ANALYSIS

Analysing the validation dataset's columns. Because these have to be made according to the training dataset.

names(test)

```
> names(test)
 [1] "Num"                                  "Default - 1"
 [3] "Total assets"                         "Net worth"
 [5] "Total income"                         "Change in stock"
 [7] "Total expenses"                       "Profit after tax"
 [9] "PBDITA"                               "PBT"
[11] "Cash profit"                          "PBDITA as % of total income"
[13] "PBT as % of total income"             "PAT as % of total income"
[15] "Cash profit as % of total income"     "PAT as % of net worth"
[17] "Sales"                                "Income from financial services"
[19] "Other income"                         "Total capital"
[21] "Reserves and funds"                   "Deposits (accepted by commercial banks)"
[23] "Borrowings"                           "Current liabilities & provisions"
[25] "Deferred tax liability"               "Shareholders funds"
[27] "Cumulative retained profits"          "Capital employed"
[29] "TOL/TNW"                              "Total term liabilities / tangible net worth"
[31] "Contingent liabilities / Net worth (%)" "Contingent liabilities"
[33] "Net fixed assets"                     "Investments"
[35] "Current assets"                       "Net working capital"
[37] "Quick ratio (times)"                  "Current ratio (times)"
[39] "Debt to equity ratio (times)"         "Cash to current liabilities (times)"
[41] "Cash to average cost of sales per day" "Creditors turnover"
[43] "Debtors turnover"                     "Finished goods turnover"
[45] "WIP turnover"                         "Raw material turnover"
[47] "Shares outstanding"                   "Equity face value"
[49] "EPS"                                  "Adjusted EPS"
[51] "Total liabilities"                    "PE on BSE"
>
```

colSums(is.na(test))

```
                                Default...1                                Total.assets
                                          0                                           0
                                 Net.worth                                Profit.after.tax
                                          0                                          23
                                     PBDITA                                         PBT
                                         23                                          23
                                Cash.profit                 PBDITA.as...of.total.income
                                         23                                          11
                  PBT.as...of.total.income                    PAT.as...of.total.income
                                         11                                          11
          Cash.profit.as...of.total.income                       PAT.as...of.net.worth
                                         11                                           0
                                      Sales                              Total.capital
                                         46                                           1
                          Reserves.and.funds                                 Borrowings
                                         13                                          65
          Current.liabilities...provisions                           Shareholders.funds
                                         14                                           0
               Cumulative.retained.profits                             Capital.employed
                                          7                                           0
                                    TOL.TNW Total.term.liabilities...tangible.net.worth
                                          0                                           0
     Contingent.liabilities...Net.worth....                           Net.fixed.assets
                                          0                                          14
                            Current.assets                          Net.working.capital
                                         14                                           5
                       Quick.ratio..times.                         Current.ratio..times.
                                         12                                          12
              Debt.to.equity.ratio..times.               Cash.to.current.liabilities..times.
                                          0                                          12
      Cash.to.average.cost.of.sales.per.day                                         EPS
                                         15                                           0
                                Adjusted.EPS                           Total.liabilities
                                          0                                           0
                   Sales.per.TotalExpenses                    Totalincome.per.TotalAssets
                                         46                                          33
>
```

14.3 IMPUTING THE MISSING VALUES

There are missing values in our test data set. We are imputing these missing values by MICE machine learning algorithm so as to avoid more missing data after we create ratios from the original variables.

```
testmice=mice(test,method = "pmm",m=5,maxit=5,seed=8488)

test=complete(testmice)

anyNA(test)
```

```
> False
```

14.4 CREATION OF NEW RATIOS FOR TEST DATASET(variables same as TRAIN Dataset)

Ratios For Profitability

CashProfit.per.TotalAssets

Cum.Profit.per.Sales

Sales.per.TotalAssets

Sales.per.CurentAssets

PBT.per.TotalCaptital

PAT.per.TotalAssets

PAT.per.Sales

PAT.per.TotalCapital

PAT.per.TotalAssets

PBT.per.TotalAssests

PBT.per.Sales

PBT.per.TotalCapital

PBDITA.per.Sales

Ratios For Size

ShareholderFunds.per.TotalAssets

Captialemployed.per.Networth

Reserves.per.TotalCapital

ShareholderFunds.per.TotalLiability

ShareholderFunds.per.TotalCapital

capitalEmp..per.ShareholderFunds

NetFixedAssets.per.CurrentAssets

NetFixedAssets.per.TotalAssets

Ratios For Leverage

Borrowings.per.TotalAssets

Borrowings.per.TotalLiabilities

Borrowings.per.TotalCapital

CurrentLiability.per.CurrentAssets

CurrentLiability.per.TotalAssets

Ratios For Liquidity

NetFixedAssets.per.TotalLiabilities

Cum.RetainedProfits.per.CurrentAssets

Networth.per.TotalLiabilities

Cum.RetainedProfit.per.TotalLiabilitie

PAT.per.CurrentLiability

CashProfit.per.CurrentLiabilities

TotalCapital.per.TotalLiabilities

TotalCapital.per.CurrentLiabilities

14.5 SELECTING THE FINAL SIGNIFICANT VARIABLES FOR VALIDATION DATASET

The final variable to be selected in the Validation dataset are those variables which came out as significant variables in out Train Dataset.

We have to make sure the columns of both the TRAIN and the VALIDATION dataset match. There cannot be any discrepancy in that.

Therefore, including the significant final variable into our validation set.

```
valdata= test[,c(10,12,25,29,31,35,36,38,44,45,46,47,49,51,53,54,56,58,61,66)]
names(valdata)
```

```
[1] "PAT.as...of.total.income"
 [2] "PAT.as...of.net.worth"
 [3] "Current.assets"
 [4] "Debt.to.equity.ratio..times."
 [5] "Cash.to.average.cost.of.sales.per.day"
 [6] "Sales.per.TotalExpenses"
 [7] "Totalincome.per.TotalAssets"
 [8] "PAT.per.CurrentLiability"
 [9] "Cum.RetainedProfit.per.TotalLiabilities"
[10] "CashProfit.per.CurrentLiabilities"
[11] "CashProfit.per.TotalAssets"
[12] "Cum.Profit.per.Sales"
[13] "Reserves.per.TotalCapital"
[14] "PAT.per.TotalAssets"
[15] "PAT.per.TotalCapital"
[16] "Networth.per.TotalLiabilities"
[17] "PBT.per.Sales"
[18] "TotalCapital.per.TotalLiabilities"
[19] "CurrentLiability.per.TotalAssets"
[20] "NetFixedAssets.per.CurrentAssets"

>
```

➔ Adding Default To Validation Test

valdata$default= as.factor(test$Default...1)

*Checking That The Structure Of The Train And The Test Is The Same

dim(valdata)
dim(smote.train)

```
> dim(valdata)
[1] 715  21
> dim(smote.train)
[1] 1638   21
>
```

15. PREDICTIONS + CONFUSION MATRIX FOR THE VALIDATION DATASET

```
pred= predict(FinalModel,valdata)
predicted= ifelse(pred<0.4,0,1)
```

Converting into factor

```
predicted=as.factor(predicted)
actual= valdata$default
```

➔ Making a Confusion Matrix for the Validation set

```
caret::confusionMatrix(predicted,actual,positive='1')
```

```
> caret::confusionMatrix(predicted,actual,positive='1')
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 546   3
         1  77  35

               Accuracy : 0.879
                 95% CI : (0.8516, 0.9029)
    No Information Rate : 0.9425
    P-Value [Acc > NIR] : 1

                  Kappa : 0.4166

 Mcnemar's Test P-Value : 3.305e-16

            Sensitivity : 0.92105
            Specificity : 0.87640
         Pos Pred Value : 0.31250
         Neg Pred Value : 0.99454
             Prevalence : 0.05749
         Detection Rate : 0.05295
   Detection Prevalence : 0.16944
      Balanced Accuracy : 0.89873

       'Positive' Class : 1

> |
```

Accuracy : 87.9%

Sensitivity : 92.10%

Specificity : 87.64%

16. SORTING AND DECILING OF THE VALIDATION SET

We will group the dataset into 10 groups. In this scenario, the groups will be created based on the Probability of Default that was predicted by our Logistic Regression model.

16.1 ADDING THE PROBABILITY OF DEFAULT IN OUR VALIDATION SET

```
valdata$Probability.of.Default=pred
```

16.2 SORTING THE VALIDATION SET IN DESCENDING ORDER BY PROBABILITY OF DEFAULT

```
valdata=valdata %>% arrange(desc(valdata$Probability.of.Default))
```

16.3 CREATING DECILES

We have two approaches for it here, we will use both.

16.3.1 Making Deciles In The Range Form

```
probs= seq(0,1 ,length=11)
dec= quantile(valdata$Probability.of.Default,probs,na.rm = TRUE)
valdata$Deciles= cut(valdata$Probability.of.Default, unique(dec),
                     lowest.include=TRUE,right = FALSE)
tables(valdata$Deciles)
```

```
> table(valdata$Deciles)

[-1.06e+16,-1.94e+15) [-1.94e+15,-1.66e+15)  [-1.66e+15,-1.5e+15)
                   71                    70                   70
 [-1.5e+15,-1.35e+15) [-1.35e+15,-1.21e+15)  [-1.21e+15,-1.1e+15)
                   71                    70                   70
 [-1.1e+15,-9.81e+14) [-9.81e+14,-7.96e+14) [-7.96e+14,-2.41e+14)
                   71                    70                   70
 [-2.41e+14,9.08e+15)
                   70
>
```

16.3.2 In The Ranking Form

```
valdata <- valdata %>% mutate(quartile = ntile(-valdata$Probability.of.Default,10))
# Viewing how our Validation Dataset has been divided into 10 Deciles.
View(valdata)

valdata$Deciles=valdata$quartile
defaulter<- data.table::data.table(valdata)
defaulter
valdata_decile <- defaulter[,list(`# Defaulter` <- sum(default==1),
                                  Total  <-length(default)) , by = Deciles][order(Deciles)]
valdata_decile
```

```
> valdata_decile
    Deciles V1 V2
 1:       1 31 71
 2:       2 15 70
 3:       3  0 71
 4:       4  1 70
 5:       5  1 70
 6:       6  1 71
 7:       7  0 70
 8:       8  0 71
 9:       9  0 70
10:      10  2 70
11:      NA  3 11
>
```

➔ Thus, we have created Deciles , and grouped the validation dataset's records into 10.

The Column V1 =The number of companies which defaulted (Default=1)

The Column V2= The Total number of companies to which loan was given

*Decile 1 has the most number of Defaults