# FINANCIAL AND RISK ANALYTICS BUSINESS REPORT

THANUSRI A

30/06/2024

**Problem 1**

**Define the problem and perform Exploratory Data Analysis**
Problem definition - Check shape, Data types, and statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

**Data Pre-processing**
Prepare the data for modeling: - Outlier Detection (treat, if needed) - Encode the data - Data split - Scale the data - Target variable creation * The target variable is default and should take the value 1 when net worth next year is negative & 0 when net worth next year is positive

**Model Building**
Metrics of Choice (Justify the evaluation metrics) - Model Building (Logistic Regression, Random Forest) - Model performance check across different metrics

**Model Performance Improvement**
Dealing with multicollinearity using VIF - Identify optimal threshold for Logistic Regression using ROC curve - Hyperparameter Tuning for Random Forest - Model performance check across different metrics

**Model Performance Comparison and Final Model Selection**
Compare all the models built - Select the final model with the proper justification - Check the most important features in the final model and draw inferences

**Actionable Insights & Recommendations**
Actionable insights and recommendations

The top five rows of the data :

| | Num | Networth Next Year | Total assets | Net worth | Total income | Change in stock | Total expenses | Profit after tax | PBDITA | PBT | ... | Debtors turnover | Finished goods turnover | WIP turnover | Raw material turnover | Shares outstanding | Equity face value | EP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 395.30 | 827.60 | 336.50 | 534.10 | 13.50 | 508.70 | 38.90 | 124.40 | 64.60 | ... | 5.65 | 3.99 | 3.37 | 14.87 | 8760056.00 | 10.00 | 4.4 |
| **1** | 2 | 36.20 | 67.70 | 24.30 | 137.90 | -3.70 | 131.00 | 3.20 | 5.50 | 1.00 | ... | NaN | NaN | NaN | NaN | NaN | NaN | 0.0 |
| **2** | 3 | 84.00 | 238.40 | 78.90 | 331.20 | -18.10 | 309.20 | 3.90 | 25.80 | 10.50 | ... | 2.51 | 17.67 | 8.76 | 8.35 | NaN | NaN | 0.0 |
| **3** | 4 | 2041.40 | 6883.50 | 1443.30 | 8448.50 | 212.20 | 8482.40 | 178.30 | 418.40 | 185.10 | ... | 1.91 | 18.14 | 18.62 | 11.11 | 10000000.00 | 10.00 | 17.6 |
| **4** | 5 | 41.80 | 90.90 | 47.00 | 388.60 | 3.40 | 392.70 | -0.70 | 7.20 | -0.60 | ... | 68.00 | 45.87 | 28.67 | 19.93 | 107315.00 | 100.00 | -6.5 |

The bottom five rows of the dataset :

| | Num | Networth Next Year | Total assets | Net worth | Total income | Change in stock | Total expenses | Profit after tax | PBDITA | PBT | ... | Debtors turnover | Finished goods turnover | WIP turnover | Raw material turnover | Shares outstanding | Equity face value | EPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4251** | 4252 | 0.2 | 0.4 | 0.2 | NaN | NaN | NaN | NaN | NaN | NaN | ... | 0.00 | NaN | NaN | 0.00 | NaN | NaN | 0.00 |
| **4252** | 4253 | 93.3 | 159.6 | 86.7 | 172.9 | 0.1 | 169.7 | 3.3 | 18.4 | 3.7 | ... | 1.80 | 11.00 | 8.28 | 9.88 | 8162700.0 | 10.0 | 0.42 |
| **4253** | 4254 | 932.2 | 833.8 | 664.6 | 2314.7 | 32.1 | 2151.6 | 195.2 | 348.4 | 303.0 | ... | 6.08 | 59.28 | 31.14 | 9.87 | 7479762.0 | 10.0 | 26.58 |
| **4254** | 4255 | 64.6 | 95.0 | 48.5 | 110.5 | 4.6 | 113.5 | 1.6 | 9.7 | 2.6 | ... | 3.71 | 78.99 | 11.51 | 14.95 | NaN | NaN | 0.00 |
| **4255** | 4256 | 0.0 | 384.6 | 111.3 | 345.8 | 11.3 | 341.7 | 15.4 | 57.6 | 20.7 | ... | 4.71 | 53.37 | 8.33 | 3.74 | 960000.0 | 10.0 | 15.63 |

5 rows × 51 columns

There are 4256 rows in the data frame and 51 columns in the dataset

Info about the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4256 entries, 0 to 4255
Data columns (total 51 columns):
 #   Column                                             Non-Null Count  Dtype
---  ------                                             --------------  -----
 0   Num                                                4256 non-null   int64
 1   Networth_Next_Year                                 4256 non-null   float64
 2   Total_assets                                       4256 non-null   float64
 3   Net_worth                                          4256 non-null   float64
 4   Total_income                                       4025 non-null   float64
 5   Change_in_stock                                    3706 non-null   float64
 6   Total_expenses                                     4091 non-null   float64
 7   Profit_after_tax                                   4102 non-null   float64
 8   PBDITA                                             4102 non-null   float64
 9   PBT                                                4102 non-null   float64
 10  Cash_profit                                        4102 non-null   float64
 11  PBDITA_as_perc_of_total_income                     4177 non-null   float64
 12  PBT_as_perc_of_total_income                        4177 non-null   float64
 13  PAT_as_perc_of_total_income                        4177 non-null   float64
 14  Cash_profit_as_perc_of_total_income                4177 non-null   float64
 15  PAT_as_perc_of_net_worth                           4256 non-null   float64
 16  Sales                                              3951 non-null   float64
 17  Income_from_fincial_services                       3145 non-null   float64
 18  Other_income                                       2700 non-null   float64
 19  Total_capital                                      4251 non-null   float64
 20  Reserves_and_funds                                 4158 non-null   float64
 21  Borrowings                                         3825 non-null   float64
 22  Current_liabilities_&_provisions                   4146 non-null   float64
 23  Deferred_tax_liability                             2887 non-null   float64
 24  Shareholders_funds                                 4256 non-null   float64
 25  Cumulative_retained_profits                        4211 non-null   float64
 26  Capital_employed                                   4256 non-null   float64
 27  TOL_to_TNW                                         4256 non-null   float64
 28  Total_term_liabilities__to__tangible_net_worth     4256 non-null   float64
 29  Contingent_liabilities__to__Net_worth_perc         4256 non-null   float64
 30  Contingent_liabilities                             2854 non-null   float64
 31  Net_fixed_assets                                   4124 non-null   float64
 32  Investments                                        2541 non-null   float64
 33  Current_assets                                     4176 non-null   float64
 34  Net_working_capital                                4219 non-null   float64
 35  Quick_ratio_times                                  4151 non-null   float64
 36  Current_ratio_times                                4151 non-null   float64
 37  Debt_to_equity_ratio_times                         4256 non-null   float64
 38  Cash_to_current_liabilities_times                  4151 non-null   float64
 39  Cash_to_average_cost_of_sales_per_day              4156 non-null   float64
 40  Creditors_turnover                                 3865 non-null   float64
 41  Debtors_turnover                                   3871 non-null   float64
 42  Finished_goods_turnover                            3382 non-null   float64
 43  WIP_turnover                                       3492 non-null   float64
 44  Raw_material_turnover                              3828 non-null   float64
 45  Shares_outstanding                                 3446 non-null   float64
 46  Equity_face_value                                  3446 non-null   float64
 47  EPS                                                4256 non-null   float64
 48  Adjusted_EPS                                       4256 non-null   float64
 49  Total_liabilities                                  4256 non-null   float64
 50  PE_on_BSE                                          1629 non-null   float64
dtypes: float64(50), int64(1)
memory usage: 1.7 MB
```

Summary of the data:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Num | 4256.00 | 2128.50 | 1228.75 | 1.00 | 1064.75 | 2128.50 | 3192.25 | 4256.00 |
| Networth_Next_Year | 4256.00 | 1344.74 | 15936.74 | -74265.60 | 3.98 | 72.10 | 330.82 | 805773.40 |
| Total_assets | 4256.00 | 3573.62 | 30074.44 | 0.10 | 91.30 | 315.50 | 1120.80 | 1176509.20 |
| Net_worth | 4256.00 | 1351.95 | 12961.31 | 0.00 | 31.48 | 104.80 | 389.85 | 613151.60 |
| Total_income | 4025.00 | 4688.19 | 53918.95 | 0.00 | 107.10 | 455.10 | 1485.00 | 2442828.20 |
| Change_in_stock | 3706.00 | 43.70 | 436.92 | -3029.40 | -1.80 | 1.60 | 18.40 | 14185.50 |
| Total_expenses | 4091.00 | 4356.30 | 51396.09 | -0.10 | 96.80 | 426.80 | 1395.70 | 2366035.30 |
| Profit_after_tax | 4102.00 | 295.05 | 3079.90 | -3908.30 | 0.50 | 9.00 | 53.30 | 119439.10 |
| PBDITA | 4102.00 | 605.94 | 5646.23 | -440.70 | 6.93 | 36.90 | 158.70 | 208576.50 |
| PBT | 4102.00 | 410.26 | 4217.42 | -3894.80 | 0.80 | 12.60 | 74.17 | 145292.60 |
| Cash_profit | 4102.00 | 408.27 | 4143.93 | -2245.70 | 2.90 | 19.40 | 96.25 | 176911.80 |
| PBDITA_as_perc_of_total_income | 4177.00 | 3.18 | 172.26 | -6400.00 | 4.97 | 9.68 | 16.47 | 100.00 |
| PBT_as_perc_of_total_income | 4177.00 | -18.20 | 419.91 | -21340.00 | 0.56 | 3.34 | 8.94 | 100.00 |
| PAT_as_perc_of_total_income | 4177.00 | -20.03 | 423.58 | -21340.00 | 0.35 | 2.37 | 6.42 | 150.00 |
| Cash_profit_as_perc_of_total_income | 4177.00 | -9.02 | 299.96 | -15020.00 | 2.00 | 5.66 | 10.73 | 100.00 |
| PAT_as_perc_of_net_worth | 4256.00 | 10.17 | 61.53 | -748.72 | 0.00 | 8.04 | 20.20 | 2466.67 |
| Sales | 3951.00 | 4645.68 | 53080.90 | 0.10 | 113.35 | 468.60 | 1481.20 | 2384984.40 |
| Income_from_fincial_services | 3145.00 | 81.36 | 1042.76 | 0.00 | 0.50 | 1.90 | 9.80 | 51938.20 |
| Other_Income | 2700.00 | 55.95 | 1178.42 | 0.00 | 0.40 | 1.50 | 6.20 | 42856.70 |
| Total_capital | 4251.00 | 224.56 | 1684.95 | 0.10 | 13.20 | 42.60 | 103.15 | 78273.20 |
| Reserves_and_funds | 4158.00 | 1210.56 | 12816.23 | -6525.90 | 5.30 | 55.15 | 282.52 | 625137.80 |
| Borrowings | 3825.00 | 1176.25 | 8581.25 | 0.10 | 24.40 | 99.80 | 358.30 | 278257.30 |
| Current_liabilities_&_provisions | 4146.00 | 960.63 | 9140.54 | 0.10 | 17.50 | 70.30 | 265.92 | 352240.30 |
| Deferred_tax_liability | 2887.00 | 234.50 | 2106.25 | 0.10 | 3.20 | 13.50 | 51.30 | 72796.60 |
| Shareholders_funds | 4256.00 | 1376.49 | 13010.69 | 0.00 | 32.30 | 107.60 | 408.90 | 613151.60 |
| Cumulative_retained_profits | 4211.00 | 937.18 | 9853.10 | -6534.30 | 1.10 | 37.40 | 206.20 | 390133.80 |
| Capital_employed | 4256.00 | 2433.62 | 20496.40 | 0.00 | 61.30 | 221.20 | 790.30 | 891408.90 |
| TOL_to_TNW | 4256.00 | 4.03 | 20.88 | -350.48 | 0.60 | 1.42 | 2.83 | 473.00 |
| Total_term_liabilities__to__tangible_net_worth | 4256.00 | 1.85 | 15.88 | -325.60 | 0.05 | 0.34 | 1.00 | 456.00 |
| Contingent_liabilities__to__Net_worth_perc | 4256.00 | 55.71 | 369.17 | 0.00 | 0.00 | 5.36 | 31.01 | 14704.27 |
| Contingent_liabilities | 2854.00 | 948.55 | 12056.74 | 0.10 | 6.00 | 37.85 | 195.32 | 559606.80 |
| Net_fixed_assets | 4124.00 | 1209.49 | 12502.40 | 0.00 | 26.20 | 93.85 | 352.82 | 636604.60 |
| Investments | 2541.00 | 721.87 | 6793.86 | 0.00 | 1.00 | 8.20 | 63.80 | 199978.60 |
| Current_assets | 4176.00 | 1350.36 | 10155.57 | 0.10 | 36.60 | 148.35 | 515.00 | 354815.20 |
| Net_working_capital | 4219.00 | 162.87 | 3162.03 | -63839.00 | -1.10 | 16.70 | 86.50 | 85782.80 |
| Quick_ratio_times | 4151.00 | 1.50 | 9.33 | 0.00 | 0.41 | 0.67 | 1.03 | 341.00 |
| Current_ratio_times | 4151.00 | 2.26 | 12.48 | 0.00 | 0.93 | 1.23 | 1.72 | 505.00 |
| Debt_to_equity_ratio_times | 4256.00 | 2.87 | 15.60 | 0.00 | 0.22 | 0.79 | 1.75 | 456.00 |
| Cash_to_current_liabilities_times | 4151.00 | 0.53 | 4.80 | 0.00 | 0.02 | 0.07 | 0.19 | 165.00 |
| Cash_to_average_cost_of_sales_per_day | 4156.00 | 145.16 | 2521.99 | 0.00 | 2.88 | 8.04 | 21.97 | 128040.76 |
| Creditors_turnover | 3865.00 | 16.81 | 75.67 | 0.00 | 3.72 | 6.17 | 11.69 | 2401.00 |
| Debtors_turnover | 3871.00 | 17.93 | 90.16 | 0.00 | 3.81 | 6.47 | 11.85 | 3135.20 |
| Finished_goods_turnover | 3382.00 | 84.37 | 562.64 | -0.09 | 8.19 | 17.32 | 40.01 | 17947.60 |
| WIP_turnover | 3492.00 | 28.68 | 169.65 | -0.18 | 5.10 | 9.86 | 20.24 | 5651.40 |
| Raw_material_turnover | 3828.00 | 17.73 | 343.13 | -2.00 | 3.02 | 6.41 | 11.82 | 21092.00 |
| Shares_outstanding | 3446.00 | 23764909.56 | 170979041.33 | -2147483647.00 | 1308382.50 | 4750000.00 | 10906020.00 | 4130400545.00 |
| Equity_face_value | 3446.00 | -1094.83 | 34101.36 | -999998.90 | 10.00 | 10.00 | 10.00 | 100000.00 |
| EPS | 4256.00 | -196.22 | 13061.95 | -843181.82 | 0.00 | 1.49 | 10.00 | 34522.53 |
| Adjusted_EPS | 4256.00 | -197.53 | 13061.93 | -843181.82 | 0.00 | 1.24 | 7.62 | 34522.53 |
| Total_liabilities | 4256.00 | 3573.62 | 30074.44 | 0.10 | 91.30 | 315.50 | 1120.80 | 1176509.20 |
| PE_on_BSE | 1629.00 | 55.46 | 1304.45 | -1116.64 | 2.97 | 8.69 | 17.00 | 51002.74 |

Target variable:

- Now we create a binary target variable using 'Networth_Next_Year':
- As required, a transformed target variable “Default” is added to the dataset based on whether the variable “Networth Next Year” is positive or negative. “Default” will take value as 0 if “Networth Next Year” is positive, otherwise “Default” is 1.
- The target variable default is taken as the value 1 when net worth next year is negative & 0 when net worth next year is positive

| | default | Networth_Next_Year |
|---|---|---|
| 0 | 0 | 395.30 |
| 1 | 0 | 36.20 |
| 2 | 0 | 84.00 |
| 3 | 0 | 821.10 |
| 4 | 0 | 41.80 |
| 5 | 0 | 291.50 |
| 6 | 0 | 93.30 |
| 7 | 0 | 821.10 |
| 8 | 0 | 188.60 |
| 9 | 0 | 229.60 |
| 10 | 0 | 292.80 |
| 11 | 0 | 821.10 |
| 12 | 0 | 821.10 |
| 13 | 1 | -5.30 |
| 14 | 0 | 26.70 |
| 15 | 0 | 560.30 |
| 16 | 0 | 283.20 |
| 17 | 0 | 28.20 |
| 18 | 0 | 1.60 |
| 19 | 0 | 3.90 |
| 20 | 1 | -121.30 |
| 21 | 0 | 57.00 |
| 22 | 0 | 65.90 |
| 23 | 0 | 664.20 |
| 24 | 0 | 388.80 |

The value counts of the new created variable ‘default’

```
default
0    3352
1     904
Name: count, dtype: int64
```

The value counts of the feature ‘default’ after normalizing

```
default
0   0.79
1   0.21
Name: proportion, dtype: float64
```

Now we create two data frames Default_X containing all the features and drop the ‘default’ feature and Default_Y containing only the feature ‘default’.

Now we replace the spaces with underscore and do some other replacement of the data for convenience of the data.

Now the dataframe looks like.

| | Num | Networth_Next_Year | Total_assets | Net_worth | Total_income | Change_in_stock | Total_expenses | Profit_after_tax | PBDITA | PBT | ... | Debtors_turnover |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 395.30 | 827.60 | 336.50 | 534.10 | 13.50 | 508.70 | 38.90 | 124.40 | 64.60 | ... | 5.65 |
| **1** | 2 | 36.20 | 67.70 | 24.30 | 137.90 | -3.70 | 131.00 | 3.20 | 5.50 | 1.00 | ... | NaN |
| **2** | 3 | 84.00 | 238.40 | 78.90 | 331.20 | -18.10 | 309.20 | 3.90 | 25.80 | 10.50 | ... | 2.51 |
| **3** | 4 | 2041.40 | 6883.50 | 1443.30 | 8448.50 | 212.20 | 8482.40 | 178.30 | 418.40 | 185.10 | ... | 1.91 |
| **4** | 5 | 41.80 | 90.90 | 47.00 | 388.60 | 3.40 | 392.70 | -0.70 | 7.20 | -0.60 | ... | 68.00 |

5 rows × 51 columns

Missing values treatment:

There are missing values present in the data

The missing values are treated with Simple Imputer Class. SimpleImputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. Here, median is used to fill up the missing value.

The following figure ensures that there is no missing values after treatment.

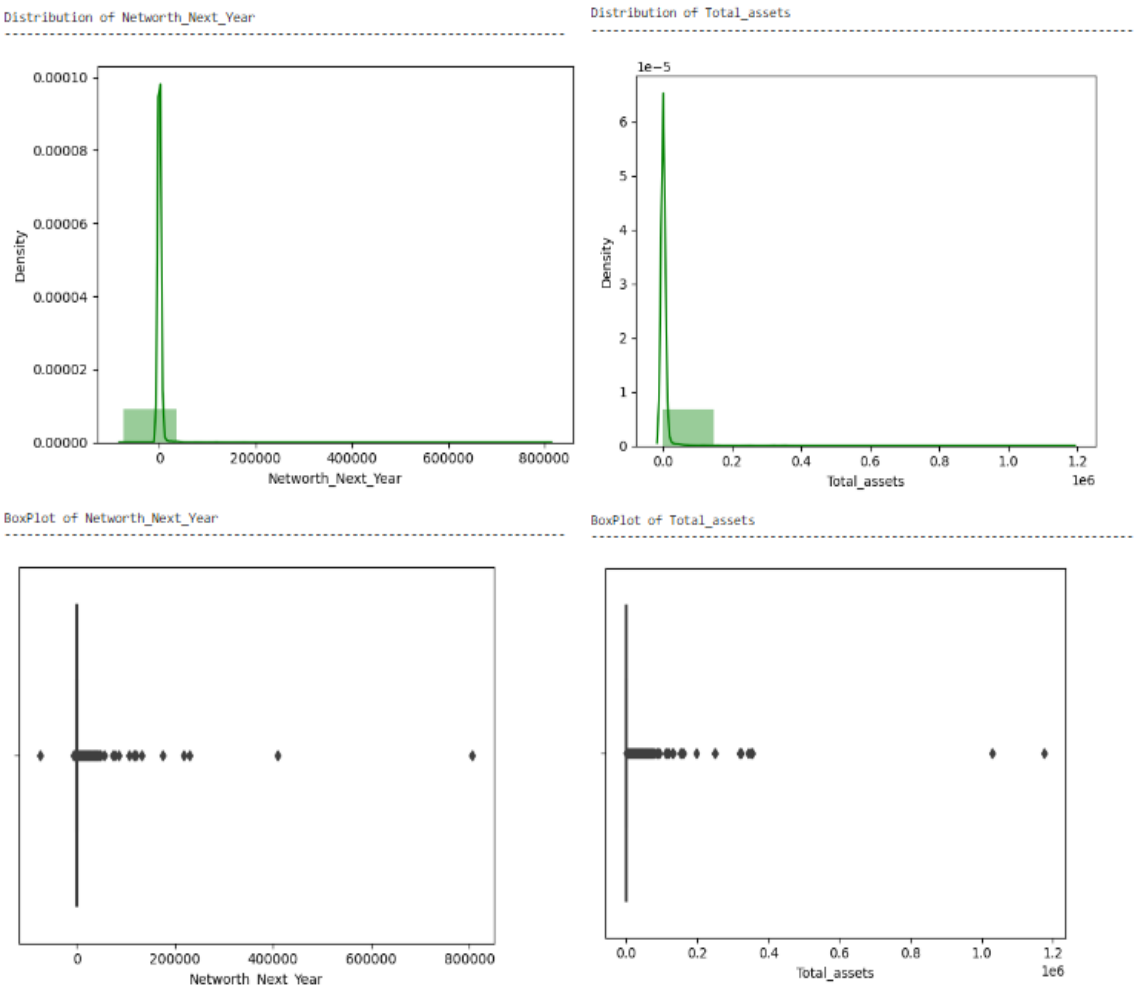
```
Networth_Next_Year                                0
Total_assets                                      0
Net_worth                                         0
Total_income                                    231
Change_in_stock                                 550
Total_expenses                                  165
Profit_after_tax                                154
PBDITA                                          154
PBT                                             154
Cash_profit                                     154
PBDITA_as_perc_of_total_income                   79
PBT_as_perc_of_total_income                      79
PAT_as_perc_of_total_income                      79
Cash_profit_as_perc_of_total_income              79
PAT_as_perc_of_net_worth                          0
Sales                                           305
Income_from_fincial_services                   1111
Other_income                                   1556
Total_capital                                     5
Reserves_and_funds                               98
Borrowings                                      431
Current_liabilities_&_provisions                110
Deferred_tax_liability                         1369
Shareholders_funds                                0
Cumulative_retained_profits                      45
Capital_employed                                  0
TOL_to_TNW                                        0
Total_term_liabilities__to__tangible_net_worth    0
Contingent_liabilities__to__Net_worth_perc        0
Contingent_liabilities                         1402
Net_fixed_assets                                132
Investments                                    1715
Current_assets                                   80
Net_working_capital                              37
Quick_ratio_times                               105
Current_ratio_times                             105
Debt_to_equity_ratio_times                        0
Cash_to_current_liabilities_times               105
Cash_to_average_cost_of_sales_per_day           100
Creditors_turnover                              391
Debtors_turnover                                385
Finished_goods_turnover                         874
WIP_turnover                                    764
Raw_material_turnover                           428
Shares_outstanding                              810
EPS                                               0
Adjusted_EPS                                      0
Total_liabilities                                 0
default                                           0
dtype: int64
```

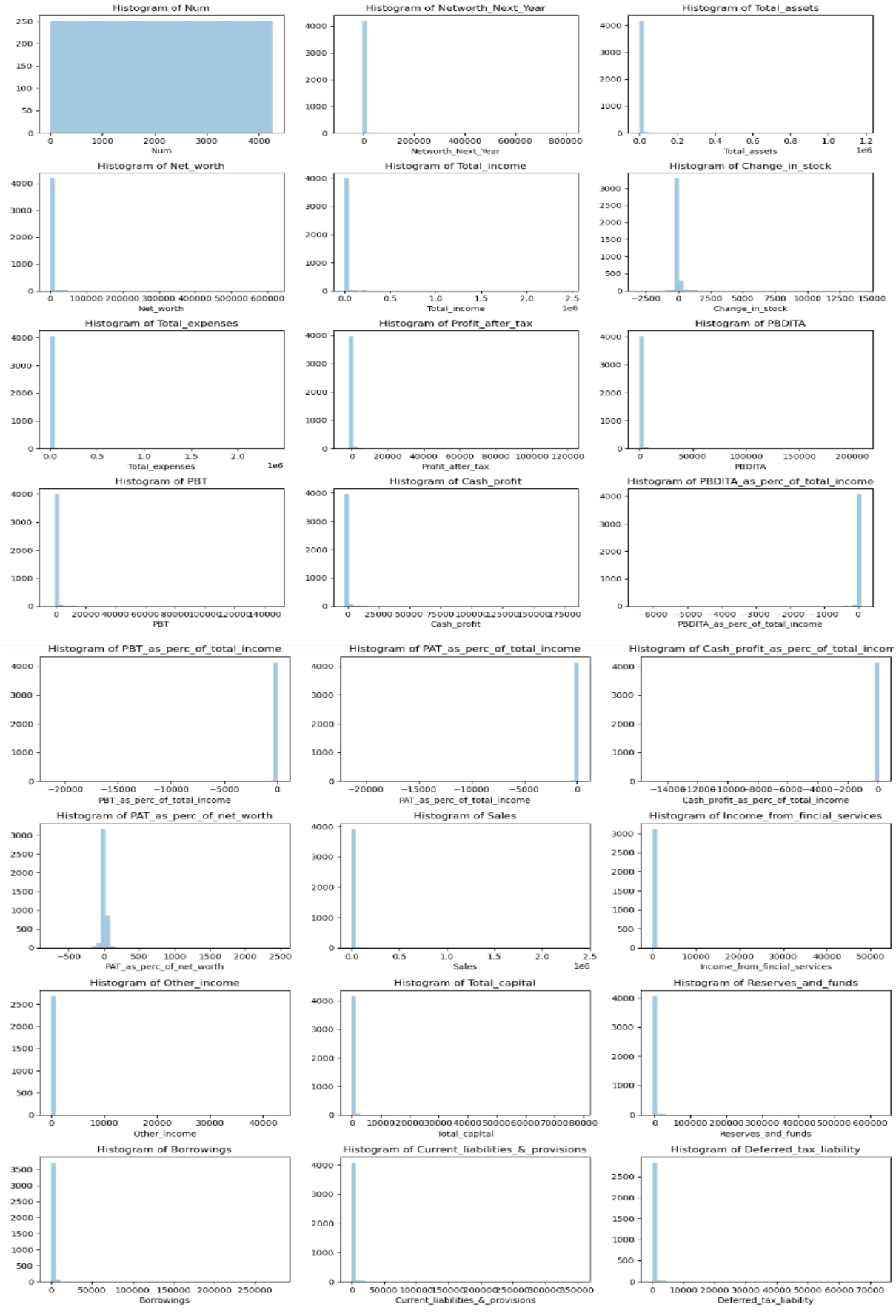
- The total number of missing values present are 14341.
- Now we check and drop the columns that have more than 30 percent missing values. These columns are ‘Investments','Other_income', 'Contingent_liabilities', 'Deferred_tax_liability'
- Now we import standard scalar and scale the values present in the dataset for the sake of irregularities.
- Now we impute the remaining missing values by KNN imputer. After imputing the missing values, the dataset looks like

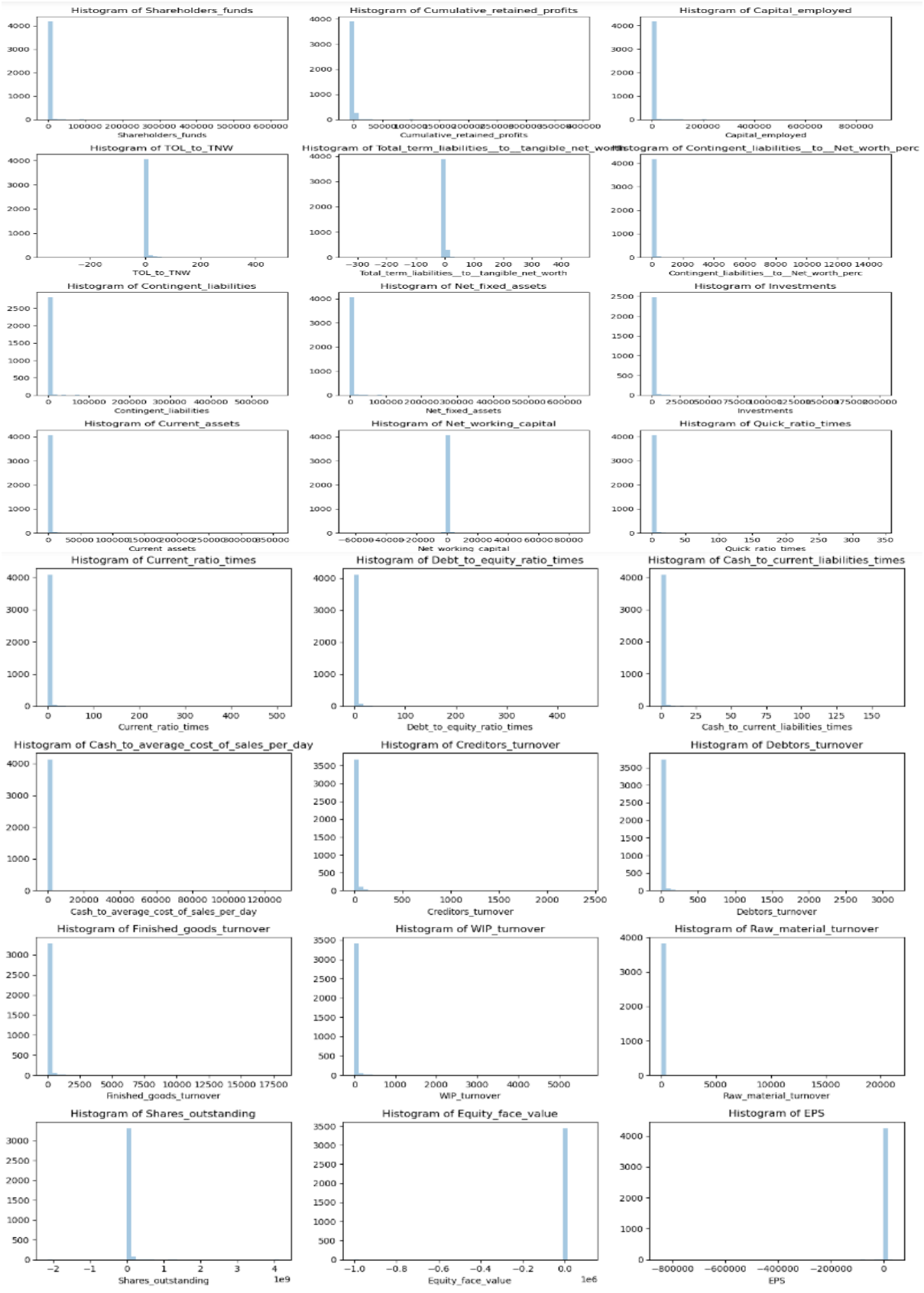
```
Networth_Next_Year                                0
Total_assets                                      0
Net_worth                                         0
Total_income                                      0
Change_in_stock                                   0
Total_expenses                                    0
Profit_after_tax                                  0
PBDITA                                            0
PBT                                               0
Cash_profit                                       0
PBDITA_as_perc_of_total_income                    0
PBT_as_perc_of_total_income                       0
PAT_as_perc_of_total_income                       0
Cash_profit_as_perc_of_total_income               0
PAT_as_perc_of_net_worth                          0
Sales                                             0
Income_from_fincial_services                      0
Total_capital                                     0
Reserves_and_funds                                0
Borrowings                                        0
Current_liabilities_&_provisions                  0
Shareholders_funds                                0
Cumulative_retained_profits                       0
Capital_employed                                  0
TOL_to_TNW                                        0
Total_term_liabilities__to__tangible_net_worth    0
Contingent_liabilities__to__Net_worth_perc        0
Net_fixed_assets                                  0
Current_assets                                    0
Net_working_capital                               0
Quick_ratio_times                                 0
Current_ratio_times                               0
Debt_to_equity_ratio_times                        0
Cash_to_current_liabilities_times                 0
Cash_to_average_cost_of_sales_per_day             0
Creditors_turnover                                0
Debtors_turnover                                  0
Finished_goods_turnover                           0
WIP_turnover                                      0
Raw_material_turnover                             0
Shares_outstanding                                0
EPS                                               0
Adjusted_EPS                                      0
Total_liabilities                                 0
default                                           0
dtype: int64
```

Univariate analysis:

- The histogram and boxplot for the features is as follows
- Univariate analysis involving data distribution along with outlier detection (Boxplot) plots have been shown below. Due to large number of variables, the number of plots will be high.

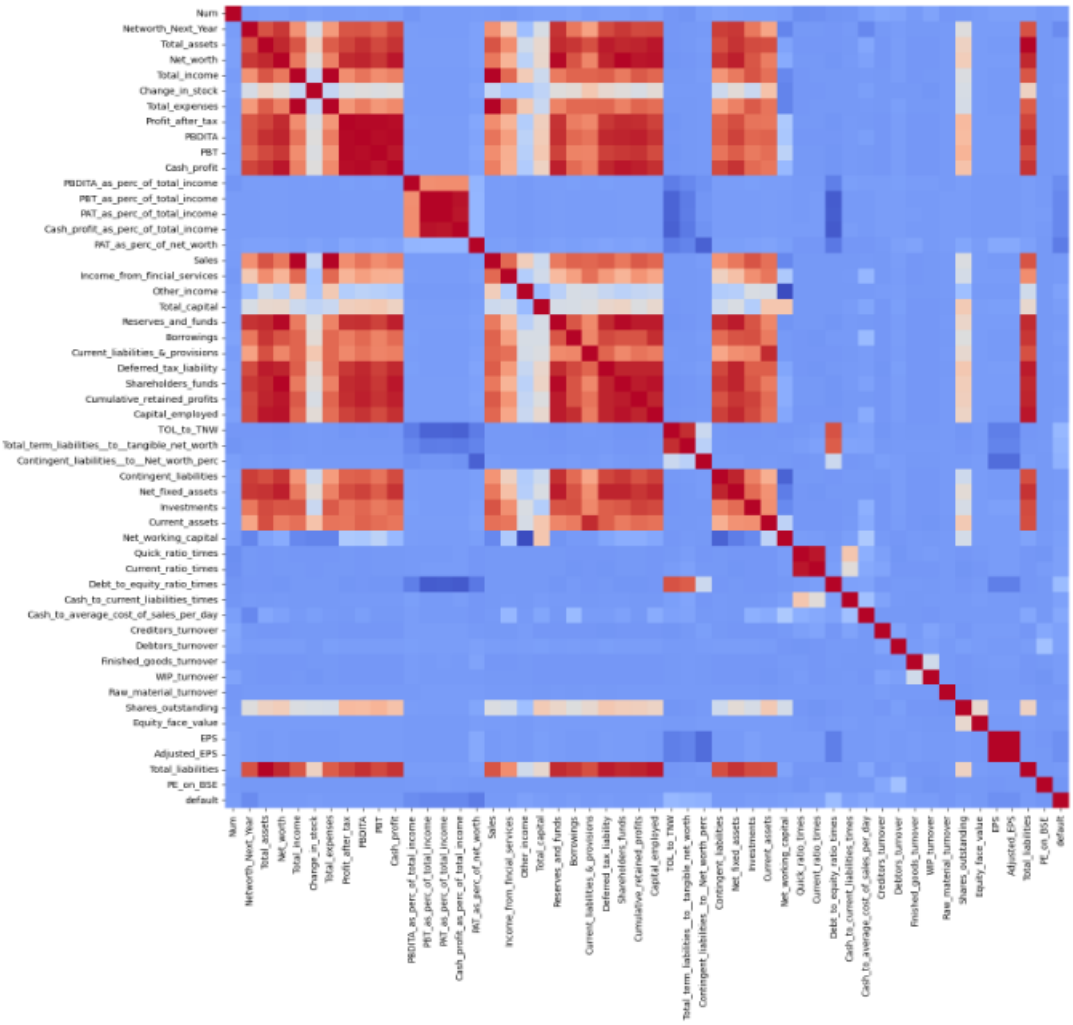



Histogram of Num
Num
Histogram of Networth_Next_Year
Networth_Next_Year
Histogram of Total_assets
Total_assets
1e6
Histogram of Net_worth
Net_worth
Histogram of Total_income
Total_income
1e6
Histogram of Change_in_stock
Change_in_stock
Histogram of Total_expenses
Total_expenses
1e6
Histogram of Profit_after_tax
Profit_after_tax
Histogram of PBDITA
PBDITA
Histogram of PBT
PBT
Histogram of Cash_profit
Cash_profit
Histogram of PBDITA_as_perc_of_total_income
PBDITA_as_perc_of_total_income
Histogram of PBT_as_perc_of_total_income
PBT_as_perc_of_total_income
Histogram of PAT_as_perc_of_total_income
PAT_as_perc_of_total_income
Histogram of Cash_profit_as_perc_of_total_incom
Cash_profit_as_perc_of_total_income
Histogram of PAT_as_perc_of_net_worth
PAT_as_perc_of_net_worth
Histogram of Sales
Sales
1e6
Histogram of Income_from_fincial_services
Income_from_fincial_services
Histogram of Other_income
Other_income
Histogram of Total_capital
Total_capital
Histogram of Reserves_and_funds
Reserves_and_funds
Histogram of Borrowings
Borrowings
Histogram of Current_liabilities_&_provisions
Current_liabilities_&_provisions
Histogram of Deferred_tax_liability
Deferred_tax_liability




Most of the variables have skewed distribution..

Multivariate analysis:

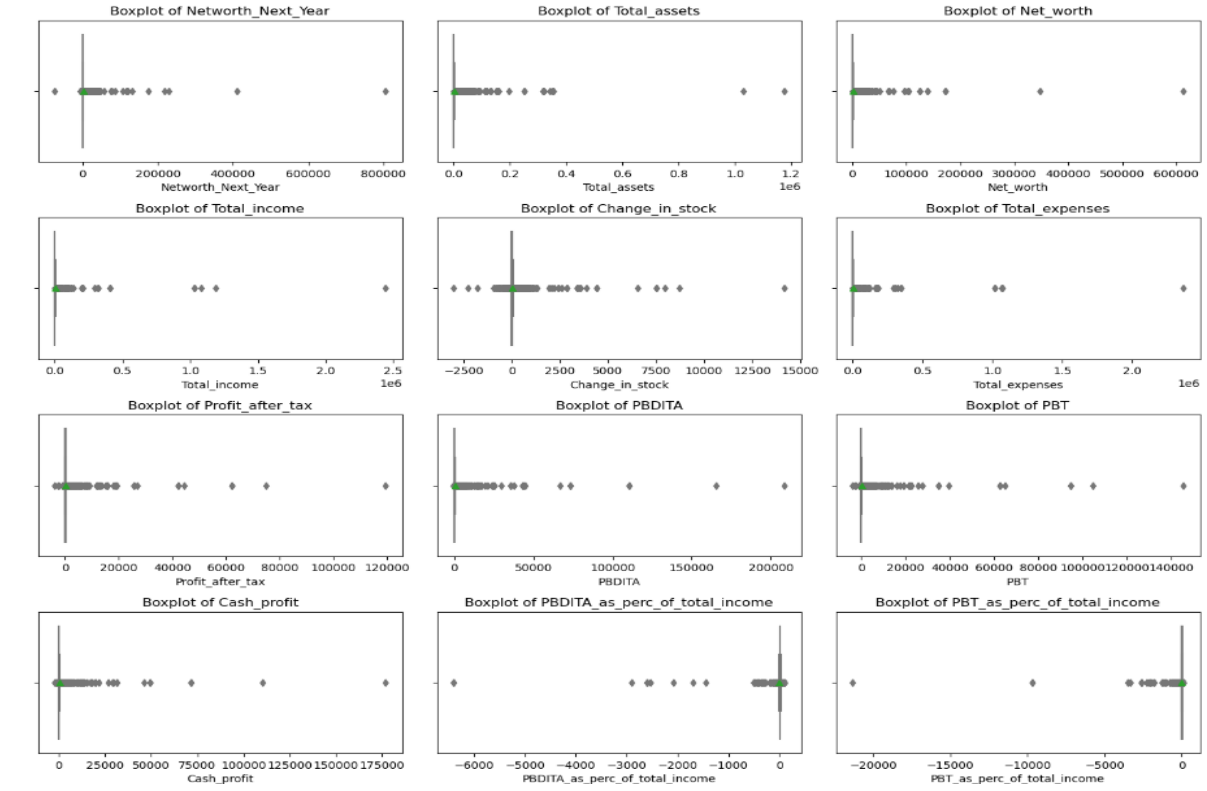


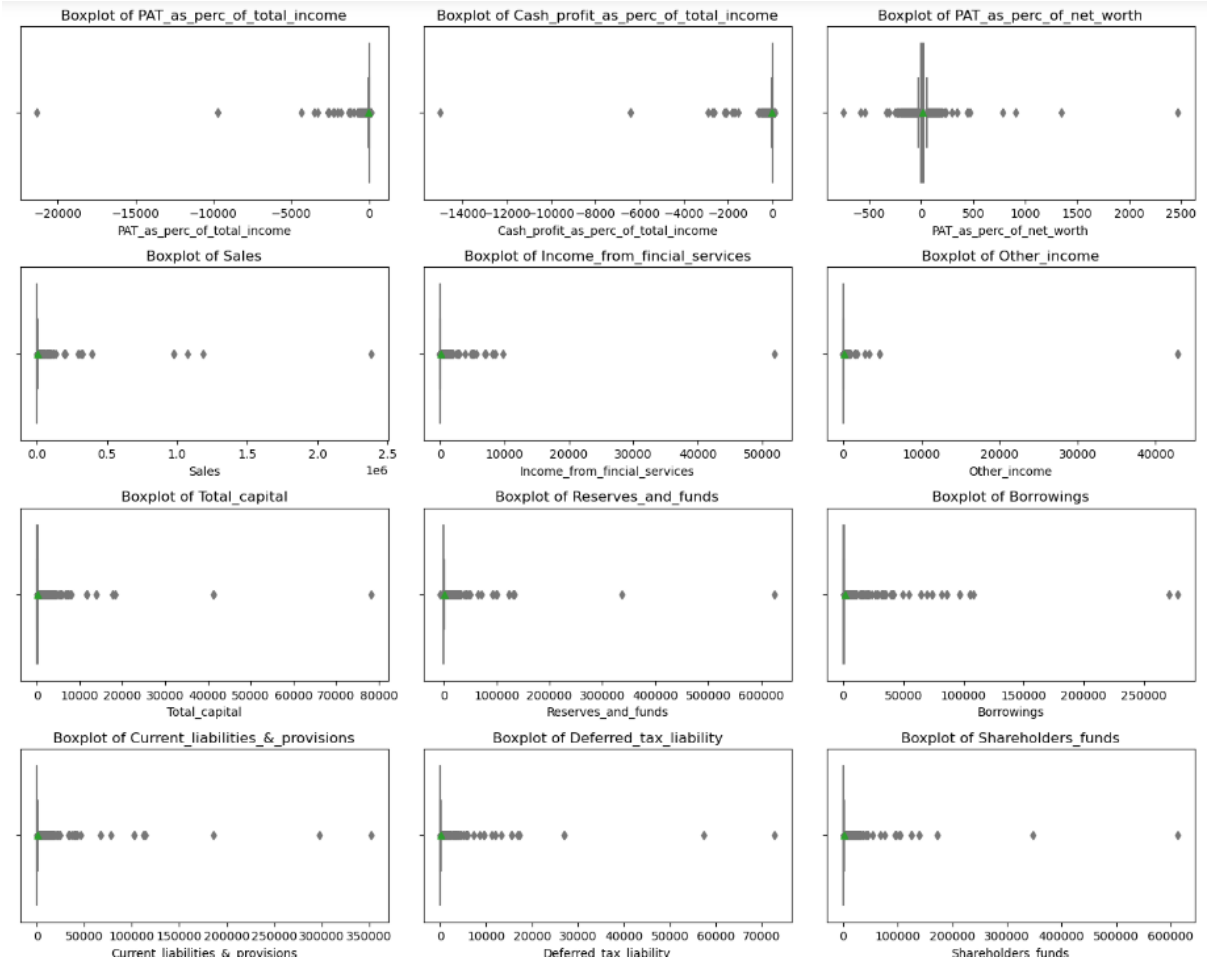

High positive and negative correlation between variables can be seen above. Majority of the variables are not correlated. Highly correlated variables are already captured in the pair plot. The above plot is dissected into smaller plots for more clarity in the subsequent pages of this report.

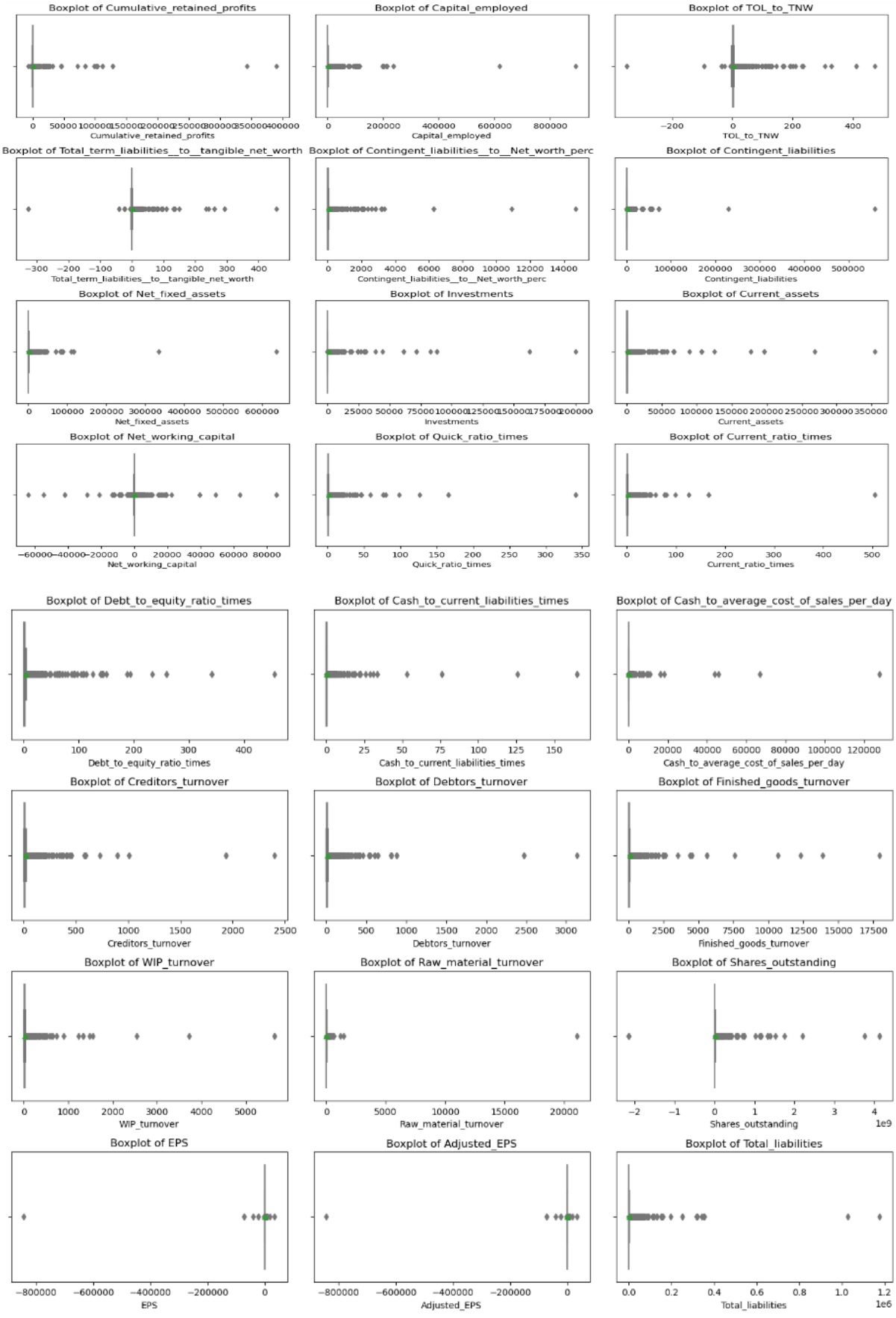
- Removing unwanted columns such as 'Num', 'Networth_Next_Year', 'Equity_face_value' because the column ‘Num’ has no value since its just the row number, and there are lot of missing values in the features 'Equity_face_value','Networth_Next_Year'.

- We also remove the other unwanted columns such as 'PE_on_BSE', 'Investments', 'Other_income', 'Contingent_liabilities', 'Deferred_tax_liability', 'Income_from_fincial_services', 'Change_in_stock', 'Shares_outstanding'

Boxplot of the features is as follows :




Boxplot of Cumulative_retained_profits
Cumulative_retained_profits
Boxplot of Capital_employed
Capital_employed
Boxplot of TOL_to_TNW
TOL_to_TNW
Boxplot of Total_term_liabilities__to__tangible_net_worth
Total_term_liabilities__to__tangible_net_worth
Boxplot of Contingent_liabilities__to__Net_worth_perc
Contingent_liabilities__to__Net_worth_perc
Boxplot of Contingent_liabilities
Contingent_liabilities
Boxplot of Net_fixed_assets
Net_fixed_assets
Boxplot of Investments
Investments
Boxplot of Current_assets
Current_assets
Boxplot of Net_working_capital
Net_working_capital
Boxplot of Quick_ratio_times
Quick_ratio_times
Boxplot of Current_ratio_times
Current_ratio_times
Boxplot of Debt_to_equity_ratio_times
Debt_to_equity_ratio_times
Boxplot of Cash_to_current_liabilities_times
Cash_to_current_liabilities_times
Boxplot of Cash_to_average_cost_of_sales_per_day
Cash_to_average_cost_of_sales_per_day
Boxplot of Creditors_turnover
Creditors_turnover
Boxplot of Debtors_turnover
Debtors_turnover
Boxplot of Finished_goods_turnover
Finished_goods_turnover
Boxplot of WIP_turnover
WIP_turnover
Boxplot of Raw_material_turnover
Raw_material_turnover
Boxplot of Shares_outstanding
Shares_outstanding
1e9
Boxplot of EPS
EPS
Boxplot of Adjusted_EPS
Adjusted_EPS
Boxplot of Total_liabilities
Total_liabilities
1e6

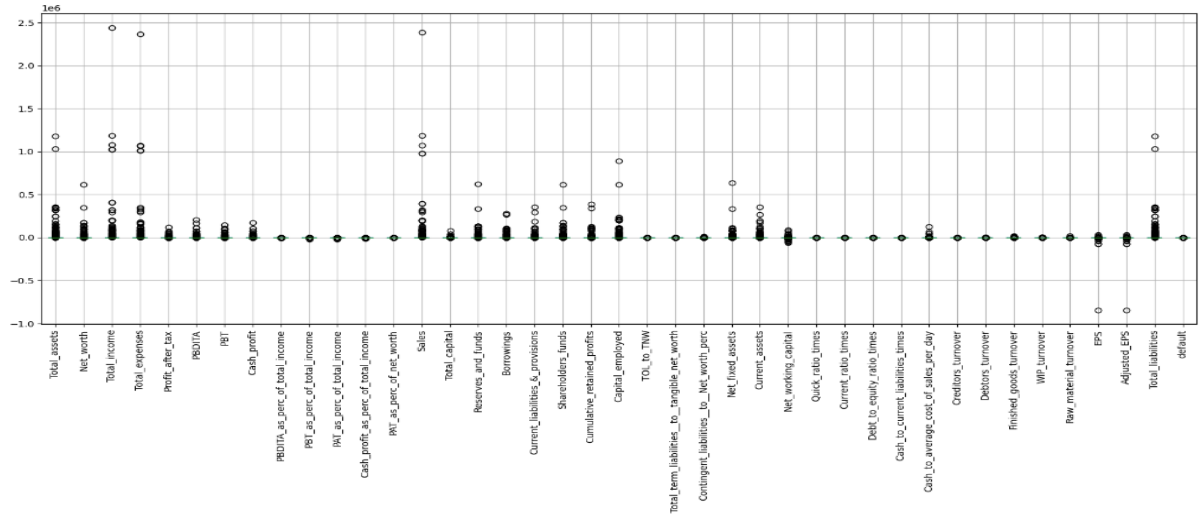
Also, all the variables have outliers. These outliers will be treated as we are going to apply Logistic regression to predict the outcome Outliers are present in all of the independent variables. For our dataset, we used IQR (Inter-Quartile Range) based calculation to treat the outliers. The following is the method,

1. Arrange the data in ascending order
2. Calculate Q1 (the first Quarter)
3. Calculate Q3 (the third Quartile)
4. Find IQR = (Q3 - Q1)
5. Find the lower Range = Q1 - (1.5 * IQR)
6. Find the upper Range = Q3 + (1.5 * IQR)

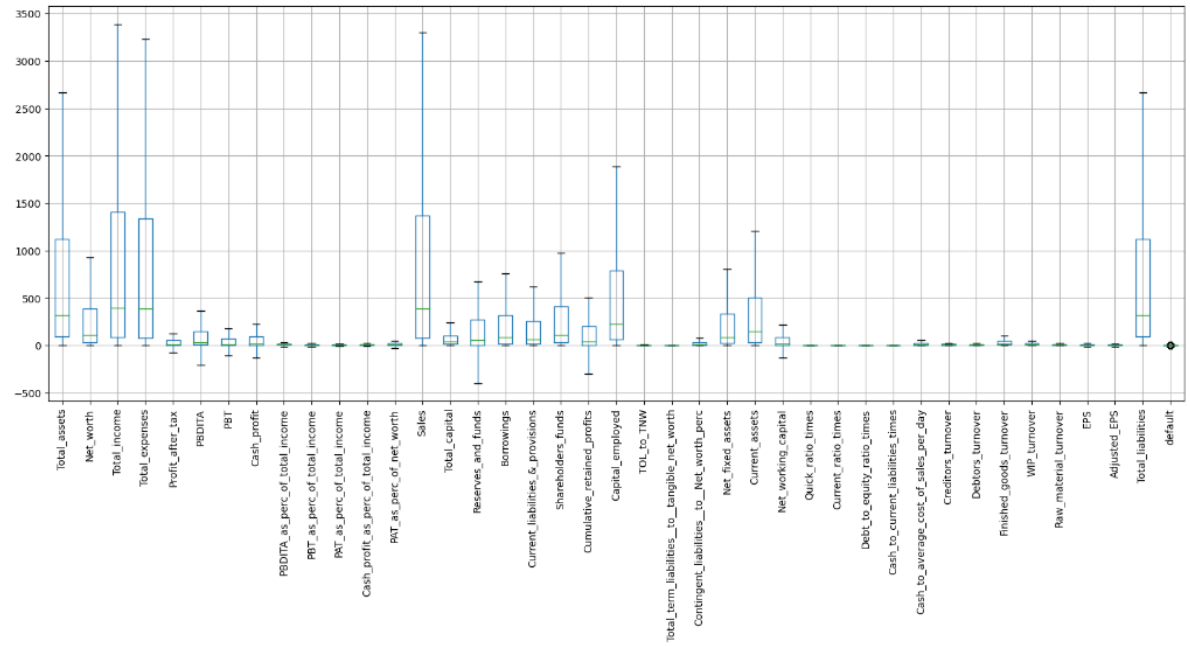
Once the upper bound and lower bound range is calculated, we snap the values above upper range and values below lower range to upper and lower range values respectively.

- There are outliers present in the data, these needs to be treated
- There are many methods in which outliers can be treated
- We choose IQR method to treat them
- So, we treat them using IQR method. In this method, any observation that is less than Q1-1.5 IQR or more than Q3 + 1.5 IQR is considered an outlier.

Before treatment:



After outlier treatment:




- Outliers have been successfully treated from the dataset now.

The top five rows of the Default_X data frame :

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Networth_Next_Year | 395.30 | 36.20 | 84.00 | 821.10 | 41.80 |
| Total_assets | 827.60 | 67.70 | 238.40 | 2665.05 | 90.90 |
| Net_worth | 336.50 | 24.30 | 78.90 | 927.41 | 47.00 |
| Total_income | 534.10 | 137.90 | 331.20 | 3551.85 | 388.60 |
| Change_in_stock | 13.50 | -3.70 | -18.10 | 48.70 | 3.40 |
| Total_expenses | 508.70 | 131.00 | 309.20 | 3344.05 | 392.70 |
| Profit_after_tax | 38.90 | 3.20 | 3.90 | 132.50 | -0.70 |
| PBDITA | 124.40 | 5.50 | 25.80 | 386.36 | 7.20 |
| PBT | 64.60 | 1.00 | 10.50 | 184.24 | -0.60 |
| Cash_profit | 95.20 | 3.80 | 9.40 | 178.00 | 3.90 |
| PBDITA_as_perc_of_total_income | 23.29 | 3.99 | 7.79 | 4.95 | 1.85 |
| PBT_as_perc_of_total_income | 12.10 | 0.73 | 3.17 | 2.19 | -0.15 |
| PAT_as_perc_of_total_income | 7.28 | 2.32 | 1.18 | 2.11 | -0.18 |
| Cash_profit_as_perc_of_total_income | 17.82 | 2.76 | 2.84 | 2.11 | 1.00 |
| PAT_as_perc_of_net_worth | 12.27 | 0.00 | 5.07 | 13.17 | -1.48 |
| Sales | 533.50 | 135.50 | 330.60 | 3532.97 | 387.60 |
| Income_from_fincial_services | 0.60 | NaN | 0.60 | 2.00 | 0.20 |
| Other_income | NaN | 0.20 | NaN | NaN | 0.80 |
| Total_capital | 87.60 | 11.90 | 25.00 | 100.00 | 10.70 |
| Reserves_and_funds | 249.00 | 4.30 | 56.70 | 696.36 | 35.80 |
| Borrowings | 390.70 | 16.60 | 44.70 | 859.15 | 25.50 |
| Current_liabilities_&_provisions | 43.90 | 23.70 | 102.20 | 638.56 | 14.10 |
| Deferred_tax_liability | 56.40 | 3.10 | 9.80 | 0.10 | 4.30 |
| Shareholders_funds | 336.50 | 24.30 | 78.90 | 973.80 | 47.00 |
| Cumulative_retained_profits | 248.90 | -8.20 | 53.10 | 513.85 | 35.80 |
| Capital_employed | 727.20 | 40.90 | 123.60 | 1883.80 | 72.50 |
| TOL_to_TNW | 1.28 | 1.53 | 1.70 | 3.69 | 0.81 |
| Total_term_liabilities__to__tangible_net_worth | 0.99 | 0.21 | 0.33 | 0.22 | 0.44 |
| Contingent_liabilities__to__Net_worth_perc | 77.53 | 47.74 | 30.42 | 10.79 | 0.00 |
| Contingent_liabilities | 479.31 | 11.60 | 24.00 | 155.70 | NaN |
| Net_fixed_assets | 461.10 | 18.50 | 56.80 | 8.60 | 36.30 |
| Investments | 18.10 | 0.20 | 0.20 | NaN | NaN |
| Current_assets | 257.60 | 39.00 | 158.30 | 1232.60 | 39.80 |
| Net_working_capital | 163.10 | 3.90 | 38.30 | 217.90 | 20.80 |
| Quick_ratio_times | 0.99 | 0.67 | 1.11 | 0.99 | 0.35 |
| Current_ratio_times | 2.52 | 1.11 | 1.31 | 1.28 | 2.09 |
| Debt_to_equity_ratio_times | 1.16 | 0.68 | 0.57 | 1.93 | 0.54 |
| Cash_to_current_liabilities_times | 0.06 | 0.02 | 0.19 | 0.07 | 0.05 |
| Cash_to_average_cost_of_sales_per_day | 5.41 | 1.62 | 26.42 | 15.93 | 0.85 |
| Creditors_turnover | 11.60 | NaN | 2.24 | 3.48 | 21.67 |
| Debtors_turnover | 5.65 | NaN | 2.51 | 1.91 | 23.91 |
| Finished_goods_turnover | 3.99 | NaN | 17.67 | 18.14 | 45.87 |
| WIP_turnover | 3.37 | NaN | 8.76 | 18.62 | 28.67 |
| Raw_material_turnover | 14.87 | NaN | 8.35 | 11.11 | 19.93 |
| Shares_outstanding | 8760056.00 | NaN | NaN | 10000000.00 | 107315.00 |
| EPS | 4.44 | 0.00 | 0.00 | 17.60 | -6.52 |
| Adjusted_EPS | 4.44 | 0.00 | 0.00 | 17.60 | -6.52 |
| Total_liabilities | 827.60 | 67.70 | 238.40 | 2665.05 | 90.90 |

The top rows of the Default_Y data frame:

```
0    0
1    0
2    0
3    0
4    0
Name: default, dtype: int32
```

Since there are larger number of variables present in the dataset and we observed that many of the variables are highly correlated, the problem of multicollinearity may occur. So, we identified those correlated variables through **VIF (variance inflation factor) calculation**. We did not consider the variables for model building whose VIF is greater than 5 (industry standard). The following variables are used for the preliminary model building after VIF calculation.

This is the primary VIF values obtained

| | variables | VIF |
|---|---|---|
| 0 | Total_assets | inf |
| 35 | Total_liabilities | inf |
| 2 | Total_income | 462.917355 |
| 13 | Sales | 431.494073 |
| 3 | Total_expenses | 305.340265 |
| 18 | Shareholders_funds | 205.492414 |
| 1 | Net_worth | 180.809313 |
| 20 | Capital_employed | 146.046874 |
| 6 | PBT | 82.190467 |
| 4 | Profit_after_tax | 77.593556 |
| 5 | PBDITA | 47.728097 |
| 9 | PBT_as_perc_of_total_income | 36.449355 |
| 10 | PAT_as_perc_of_total_income | 34.181659 |
| 7 | Cash_profit | 33.611763 |
| 25 | Current_assets | 32.218738 |
| 17 | Current_liabilities_&_provisions | 25.447561 |
| 15 | Reserves_and_funds | 19.255442 |
| 28 | Current_ratio_times | 17.727828 |
| 27 | Quick_ratio_times | 17.445973 |
| 16 | Borrowings | 14.788700 |
| 24 | Net_fixed_assets | 14.435402 |
| 37 | EPS | 14.312364 |
| 38 | Adjusted_EPS | 13.049223 |
| 29 | Debt_to_equity_ratio_times | 12.570628 |
| 11 | Cash_profit_as_perc_of_total_income | 12.238274 |
| 19 | Cumulative_retained_profits | 12.032450 |
| 8 | PBDITA_as_perc_of_total_income | 12.016900 |
| 21 | TOL_to_TNW | 10.676555 |
| 22 | Total_term_liabilities__to__tangible_net_worth | 8.451951 |
| 30 | Cash_to_current_liabilities_times | 8.067244 |
| 14 | Total_capital | 6.579530 |
| 35 | WIP_turnover | 5.805554 |
| 31 | Cash_to_average_cost_of_sales_per_day | 5.761688 |
| 33 | Debtors_turnover | 4.628263 |
| 34 | Finished_goods_turnover | 4.611512 |
| 32 | Creditors_turnover | 4.319566 |
| 12 | PAT_as_perc_of_net_worth | 3.748254 |
| 36 | Raw_material_turnover | 3.180932 |
| 26 | Net_working_capital | 2.480146 |
| 23 | Contingent_liabilities__to__Net_worth_perc | 1.952323 |

Now we one by one remove the features that have more VIF value than 5. Finally we obtain

| | variables | VIF |
|---|---|---|
| 3 | Borrowings | 4.849900 |
| 8 | Quick_ratio_times | 4.004498 |
| 10 | Creditors_turnover | 3.579060 |
| 2 | Total_capital | 3.575055 |
| 11 | Debtors_turnover | 3.486647 |
| 4 | Cumulative_retained_profits | 3.003792 |
| 13 | Raw_material_turnover | 2.927046 |
| 0 | Cash_profit_as_perc_of_total_income | 2.760219 |
| 1 | PAT_as_perc_of_net_worth | 2.511117 |
| 9 | Cash_to_average_cost_of_sales_per_day | 2.298673 |
| 14 | Adjusted_EPS | 2.226444 |
| 5 | Total_term_liabilities__to__tangible_net_worth | 2.207241 |
| 12 | Finished_goods_turnover | 2.112341 |
| 6 | Contingent_liabilities__to__Net_worth_perc | 1.850951 |
| 7 | Net_working_capital | 1.793888 |

Splitting the data into training and testing data :

- We split the independent variables X into two parts, one for training X_train and one for testing X_test
- we split the dependent variable Y into two parts, one for training Y_train and one for the testing Y_test
- Using stats model api as SM to intercept the X variable.
- Using sklearn to split the data into x_train and y_train

A preliminary logistic regression model is built on the **train set** with the variables whose VIF value is less than 5. The model output is shown below

Logit Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | default | **No. Observations:** | 2979 |
| **Model:** | Logit | **Df Residuals:** | 2963 |
| **Method:** | MLE | **Df Model:** | 15 |
| **Date:** | Sat, 29 Jun 2024 | **Pseudo R-squ.:** | 0.03676 |
| **Time:** | 23:05:01 | **Log-Likelihood:** | -1512.7 |
| **converged:** | True | **LL-Null:** | -1570.4 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 1.423e-17 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -1.3918 | 0.146 | -9.505 | 0.000 | -1.679 | -1.105 |
| **Cash_profit_as_perc_of_total_income** | -0.0199 | 0.007 | -2.713 | 0.007 | -0.034 | -0.006 |
| **PAT_as_perc_of_net_worth** | -0.0112 | 0.003 | -3.335 | 0.001 | -0.018 | -0.005 |
| **Total_capital** | 0.0010 | 0.001 | 1.280 | 0.200 | -0.001 | 0.003 |
| **Borrowings** | -0.0002 | 0.000 | -0.755 | 0.450 | -0.001 | 0.000 |
| **Cumulative_retained_profits** | -0.0001 | 0.000 | -0.297 | 0.766 | -0.001 | 0.001 |
| **Total_term_liabilities__to__tangible_net_worth** | 0.2566 | 0.064 | 4.010 | 0.000 | 0.131 | 0.382 |
| **Contingent_liabilities__to__Net_worth_perc** | 0.0019 | 0.002 | 1.068 | 0.286 | -0.002 | 0.005 |
| **Net_working_capital** | -0.0002 | 0.001 | -0.410 | 0.682 | -0.001 | 0.001 |
| **Quick_ratio_times** | -0.1915 | 0.103 | -1.866 | 0.062 | -0.393 | 0.010 |
| **Cash_to_average_cost_of_sales_per_day** | 0.0089 | 0.003 | 3.356 | 0.001 | 0.004 | 0.014 |
| **Creditors_turnover** | 0.0105 | 0.007 | 1.421 | 0.155 | -0.004 | 0.025 |
| **Debtors_turnover** | 0.0056 | 0.008 | 0.727 | 0.467 | -0.009 | 0.021 |
| **Finished_goods_turnover** | 0.0024 | 0.001 | 1.573 | 0.116 | -0.001 | 0.005 |
| **Raw_material_turnover** | -0.0146 | 0.007 | -2.029 | 0.042 | -0.029 | -0.000 |
| **Adjusted_EPS** | 0.0028 | 0.008 | 0.358 | 0.721 | -0.013 | 0.018 |

We checked the probability values for each independent variable and some of them are found to be > 0.05. So, at 95% confidence level, if $p < 0.05$, we can say that there is a relation between dependent and other independent variable. Alternately we can say that variables whose $p > 0.05$ donot have influence on the dependent variable. Therefore, ***a new model*** is prepared by discarding the variables whose $p > 0.05$.

Now we proceed with the same procedure until we get the $p<0.05$

This went until 11 models and the summary is

Logit Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | default | **No. Observations:** | 2979 |
| **Model:** | Logit | **Df Residuals:** | 2973 |
| **Method:** | MLE | **Df Model:** | 5 |
| **Date:** | Sat, 29 Jun 2024 | **Pseudo R-squ.:** | 0.03296 |
| **Time:** | 23:05:01 | **Log-Likelihood:** | -1518.7 |
| **converged:** | True | **LL-Null:** | -1570.4 |
| **Covariance Type:** | nonrobust | **LLR p-value:** | 9.482e-21 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -1.2356 | 0.103 | -12.024 | 0.000 | -1.437 | -1.034 |
| **Cash_profit_as_perc_of_total_income** | -0.0176 | 0.007 | -2.510 | 0.012 | -0.031 | -0.004 |
| **PAT_as_perc_of_net_worth** | -0.0112 | 0.003 | -3.703 | 0.000 | -0.017 | -0.005 |
| **Total_term_liabilities__to__tangible_net_worth** | 0.2651 | 0.055 | 4.792 | 0.000 | 0.157 | 0.374 |
| **Quick_ratio_times** | -0.2169 | 0.087 | -2.480 | 0.013 | -0.388 | -0.046 |
| **Cash_to_average_cost_of_sales_per_day** | 0.0089 | 0.002 | 3.552 | 0.000 | 0.004 | 0.014 |

## Model Evaluation on the Training Data

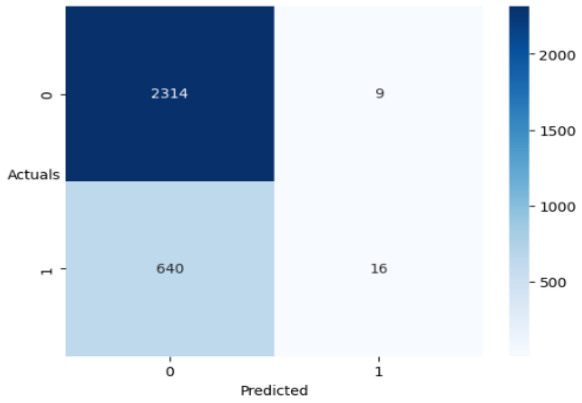
First, we will check the training set performance with predicted classes with **0.5** probability cut-off.

Different matrices were used to check the model performance, namely,

1. Confusion matrix
2. Classification report (precision, recall, accuracy)
3. ROC curve

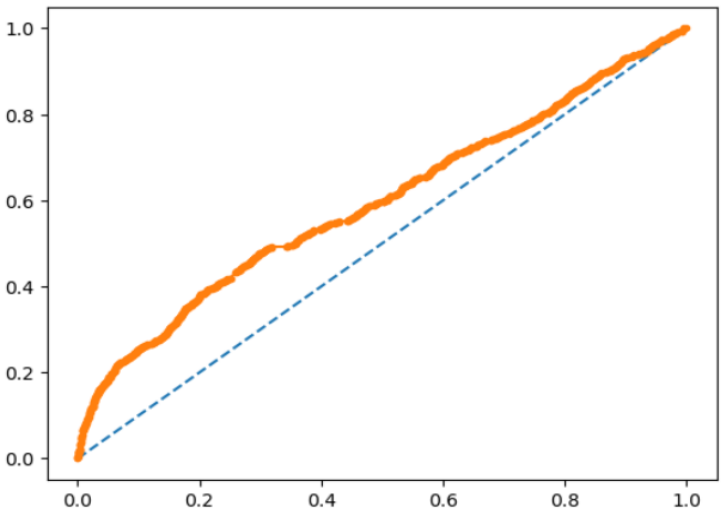
**LOGISTIC REGRESSION:**

**Optimal threshold = 0.5**

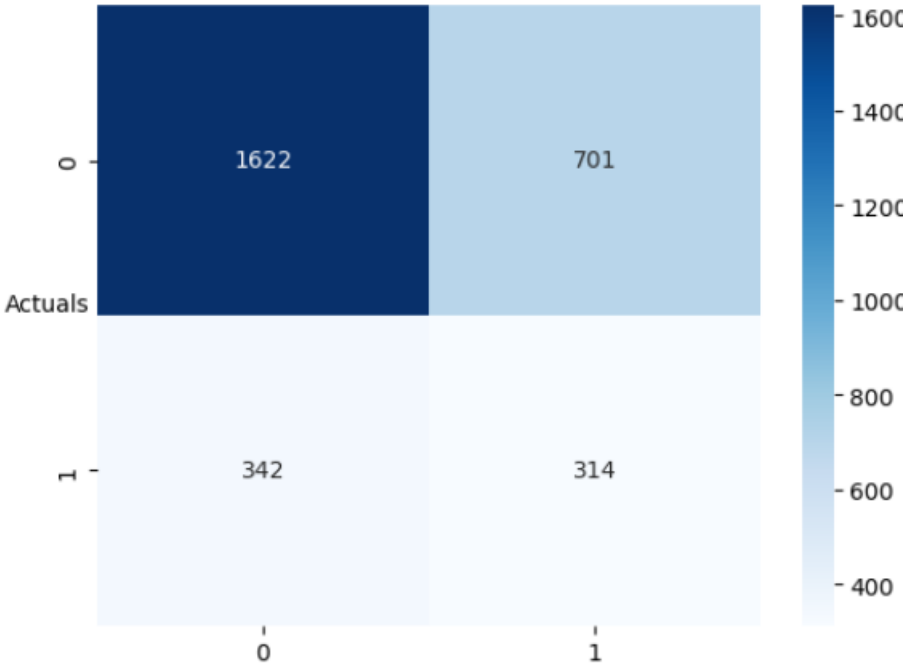



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.783 | 0.996 | 0.877 | 2323 |
| 1.0 | 0.640 | 0.024 | 0.047 | 656 |
| accuracy | | | 0.782 | 2979 |
| macro avg | 0.712 | 0.510 | 0.462 | 2979 |
| weighted avg | 0.752 | 0.782 | 0.694 | 2979 |

AUC: 0.597

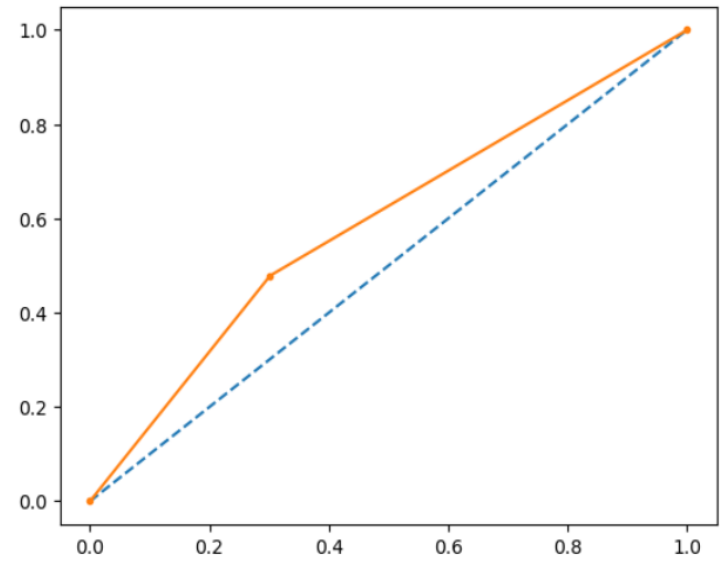


Optimal threshold =0.225

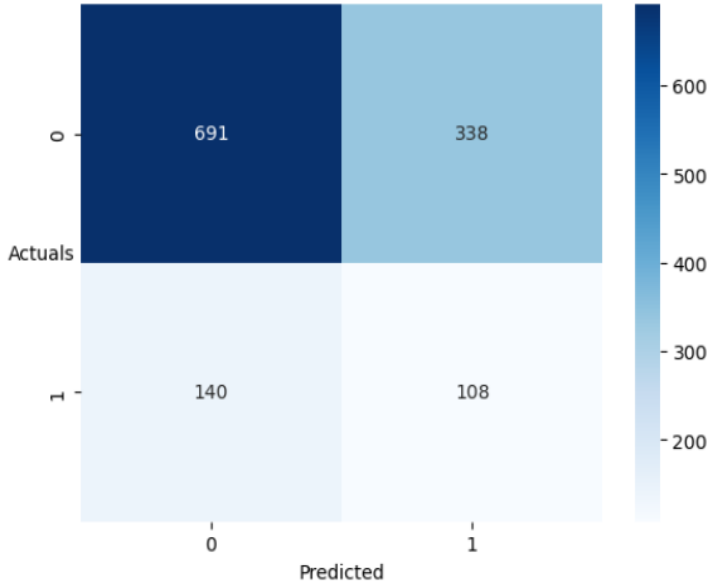



| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.826 | 0.698 | 0.757 | 2323 |
| 1.0 | 0.309 | 0.479 | 0.376 | 656 |
| accuracy | | | 0.650 | 2979 |
| macro avg | 0.568 | 0.588 | 0.566 | 2979 |
| weighted avg | 0.712 | 0.650 | 0.673 | 2979 |

AUC: 0.588

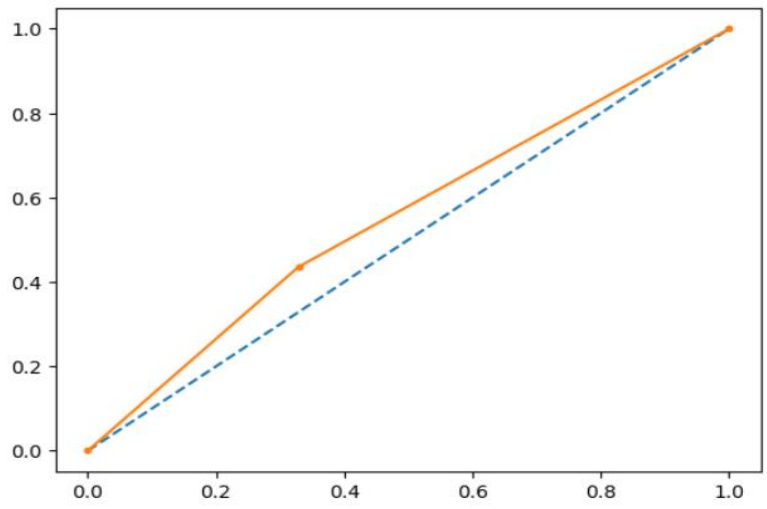


Validating logistic regression on test data




| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.832 | 0.672 | 0.743 | 1029 |
| 1.0 | 0.242 | 0.435 | 0.311 | 248 |
| accuracy | | | 0.626 | 1277 |
| macro avg | 0.537 | 0.554 | 0.527 | 1277 |
| weighted avg | 0.717 | 0.626 | 0.659 | 1277 |

AUC: 0.554




Accuracy of the model i.e., % overall correct prediction is 62% and sensitivity of the model is 43.5%. The model performs well on the test set also.

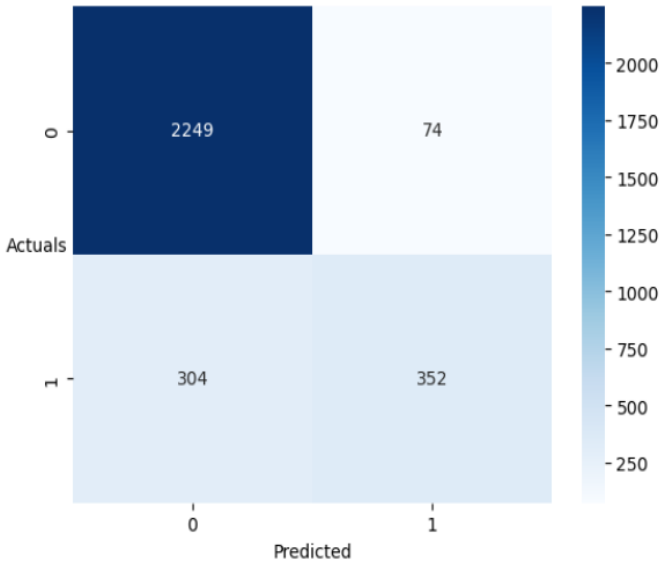
**RANDOM FOREST CLASSIFIER**:

The model score, confusion matrix and classification report of the training data respectively on applying random forest classifier is as follows :

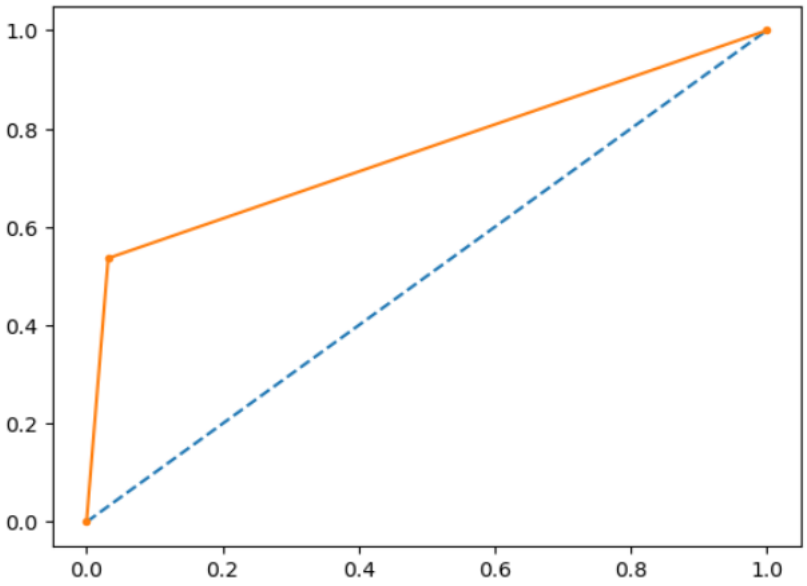
```
0.8731117824773413
[[2249   74]
 [ 304  352]]
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.92      2323
         1.0       0.83      0.54      0.65       656

    accuracy                           0.87      2979
   macro avg       0.85      0.75      0.79      2979
weighted avg       0.87      0.87      0.86      2979
```




AUC: 0.752



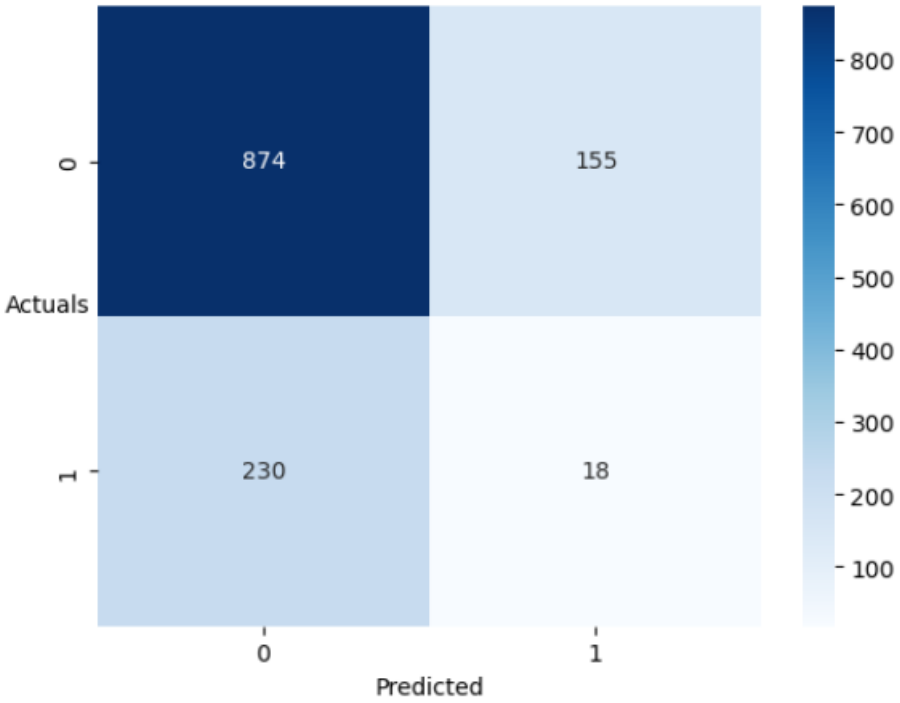
Accuracy of the model i.e., % overall correct prediction is 87% and sensitivity of the model is 54%. The model performs well on the test set also.

On test data:

```
0.6985121378230227
[[874 155]
 [230  18]]
              precision    recall  f1-score   support

         0.0       0.79      0.85      0.82      1029
         1.0       0.10      0.07      0.09       248

    accuracy                           0.70      1277
   macro avg       0.45      0.46      0.45      1277
weighted avg       0.66      0.70      0.68      1277
```




Accuracy of the model i.e., % overall correct prediction is 70% and sensitivity of the model is 7%. The model performs well on the test set also.

While the model results between training and test sets are similar, indicating no under or overfitting issues, overall prediction of the model is weak. There is a scope of improvement on the accuracy and recall values by using techniques like re-sampling, cross validation etc.,

**PROBLEM 2**

**Stock Price Graph Analysis**
Draw a Stock Price Graph (Stock Price vs Time) for the given stocks - Write observations

**Stock Returns Calculation and Analysis**
Calculate Returns for all stocks - Calculate the Mean and Standard Deviation for the returns of all stocks - Draw a plot of Mean vs Standard Deviation for all stock returns - Write observations and inferences

**Actionable Insights & Recommendations**
Actionable insights and recommendations

Top five rows of the data :

| | Date | ITC Limited | Bharti Airtel | Tata Motors | DLF Limited | Yes Bank |
|---|---|---|---|---|---|---|
| 0 | 28-03-2016 | 217 | 316 | 386 | 114 | 173 |
| 1 | 04-04-2016 | 218 | 302 | 386 | 121 | 171 |
| 2 | 11-04-2016 | 215 | 308 | 374 | 120 | 171 |
| 3 | 18-04-2016 | 223 | 320 | 408 | 122 | 172 |
| 4 | 25-04-2016 | 214 | 319 | 418 | 122 | 175 |

Info about the data set:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Date           418 non-null    object
 1   ITC Limited    418 non-null    int64
 2   Bharti Airtel  418 non-null    int64
 3   Tata Motors    418 non-null    int64
 4   DLF Limited    418 non-null    int64
 5   Yes Bank       418 non-null    int64
dtypes: int64(5), object(1)
memory usage: 19.7+ KB
```

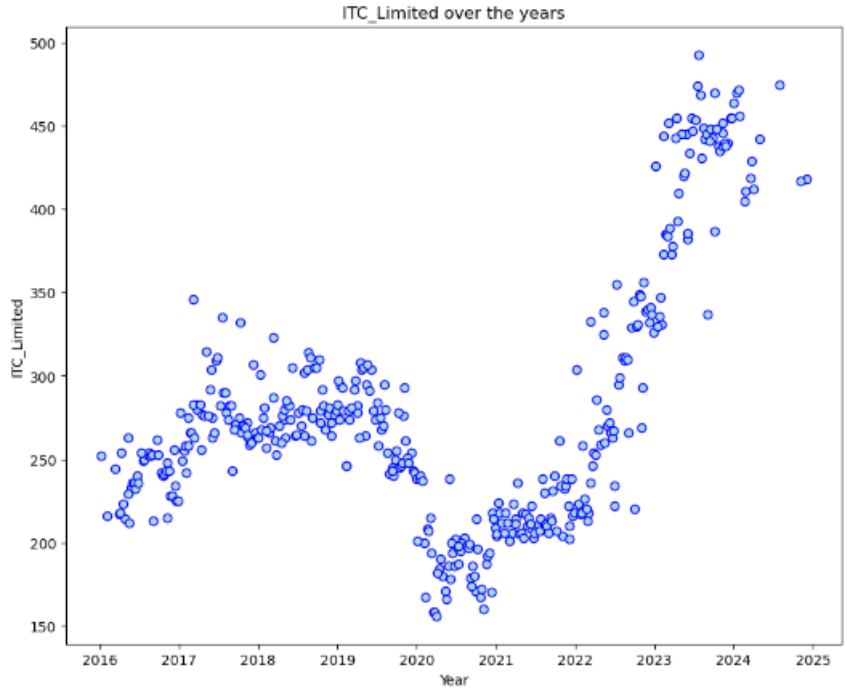
- Now we replace the spaces with underscore and do some other replacement of the data for convenience of the data.
- There are 418 rows and 6 columns in the data set.

Summary of the data :

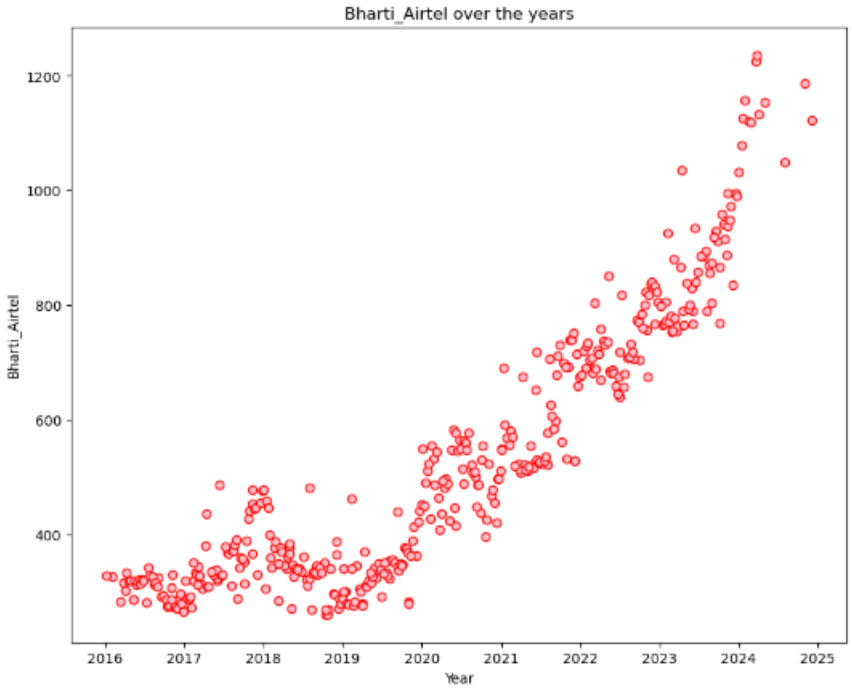
| | ITC_Limited | Bharti_Airtel | Tata_Motors | DLF_Limited | Yes_Bank |
|---|---|---|---|---|---|
| count | 418.000000 | 418.000000 | 418.000000 | 418.000000 | 418.000000 |
| mean | 278.964115 | 528.260766 | 368.617225 | 276.827751 | 124.442584 |
| std | 75.114405 | 226.507879 | 182.024419 | 156.280781 | 130.090884 |
| min | 156.000000 | 261.000000 | 65.000000 | 110.000000 | 11.000000 |
| 25% | 224.250000 | 334.000000 | 186.000000 | 166.250000 | 16.000000 |
| 50% | 265.500000 | 478.000000 | 399.500000 | 213.000000 | 30.000000 |
| 75% | 304.000000 | 706.750000 | 466.000000 | 360.500000 | 249.750000 |
| max | 493.000000 | 1236.000000 | 1035.000000 | 928.000000 | 397.000000 |

- There are no duplicate values present in the data

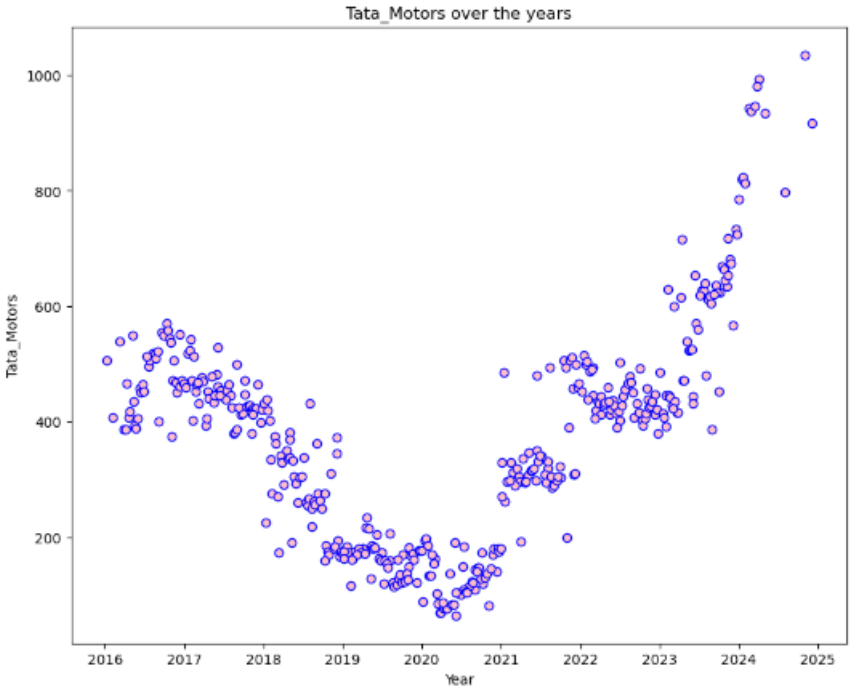
Scatterplot of the features present in the data set are as follows :



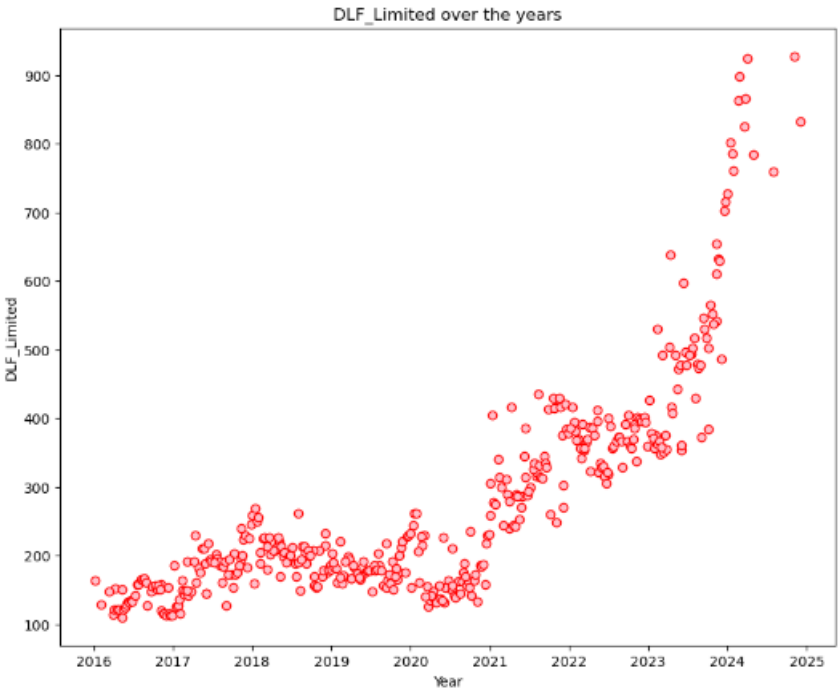

The Stock price for the ITC is on increasing trend from 2020 to 2024. There is an almost increase of 300 points within the span of 4 years.



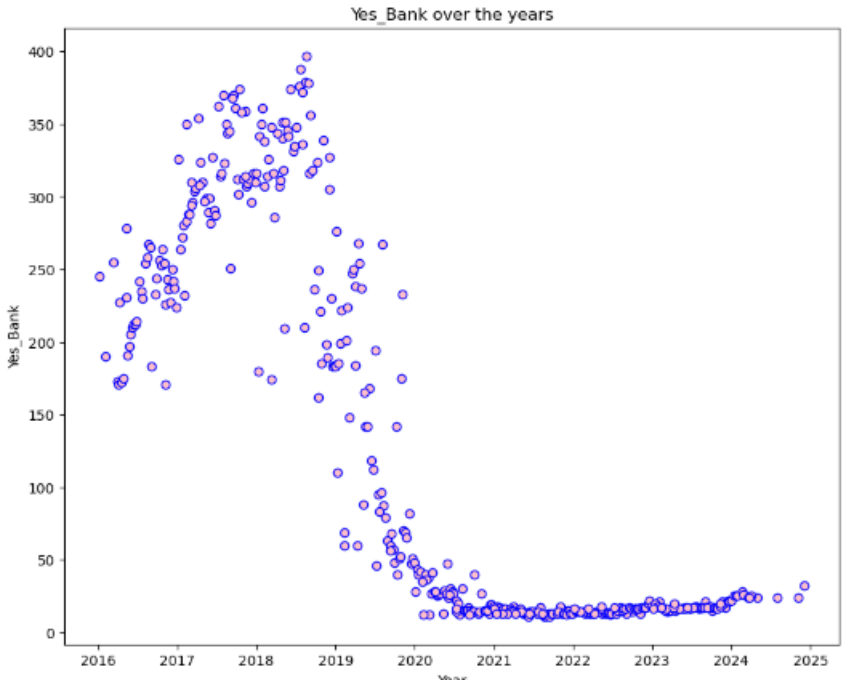

The Stock price for the Bharati_airtel is on increasing trend from 2019 to 2024. There is an almost decrease of 800 points within the span of 5 years.




The Stock price for the Tata_Motots is on decresing trend from 2016 to 2020 first and then increasing trend from 2021 to 2024. There is an almost decrease of 200 points within the span of 4 years. And then increase of 800 points in the next 4 years




The Stock price for the DLF_Limited is constant from 2013 to 2020 and then on increasing trend from 2021 to 2024. There is an almost decrease of 700 points within the span of 3 years.




The Stock price for the Yes_bank is on decreasing trend from 2018 to 2024. There is an almost decrease of 400 points within the span of 6 years.

| | ITC_Limited | Bharti_Airtel | Tata_Motors | DLF_Limited | Yes_Bank |
|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | NaN | NaN |
| 1 | 0.004598 | -0.045315 | 0.000000 | 0.059592 | -0.011628 |
| 2 | -0.013857 | 0.019673 | -0.031582 | -0.008299 | 0.000000 |
| 3 | 0.036534 | 0.038221 | 0.087011 | 0.016529 | 0.005831 |
| 4 | -0.041196 | -0.003130 | 0.024214 | 0.000000 | 0.017291 |

The negative value of Return means there is decrease in price compared to previous week and the positive value of Return means there is increase in price compared to previous week.

Now we drop the the feature ‘Date’

Average returns that the stock is making on a week-to-week basis.

```
ITC_Limited      0.001634
Bharti_Airtel    0.003271
Tata_Motors      0.002234
DLF_Limited      0.004863
Yes_Bank        -0.004737
dtype: float64
```

DLF_Limited has highest Stock Means and Yes_Bank has lowest Stock Means.

**Stock Standard Deviation:** It is a measure of volatility meaning the more a stock's returns vary from the stock's average return, the more volatile the stock.

The standard deviations are

```
ITC_Limited      0.035904
Bharti_Airtel    0.038728
Tata_Motors      0.060484
DLF_Limited      0.057785
Yes_Bank         0.093879
dtype: float64
```

**Yes_Bank** has **highest** Volatility and **ITC_Limited** has **lowest** Volatility.

Now we create a data frame namely 'Average' and 'Volatility' for all the above companies.

| | Average | Volatility |
|---|---|---|
| ITC_Limited | 0.001634 | 0.035904 |
| Bharti_Airtel | 0.003271 | 0.038728 |
| Tata_Motors | 0.002234 | 0.060484 |
| DLF_Limited | 0.004863 | 0.057785 |
| Yes_Bank | -0.004737 | 0.093879 |

Of all the above stocks, only the following stocks are having positive average means.

| | | |
|---|---|---|
| **ITC_Limited** | 0.001634 | 0.035904 |
| **Bharti_Airtel** | 0.003271 | 0.038728 |
| **Tata_Motors** | 0.002234 | 0.060484 |
| **DLF_Limited** | 0.004863 | 0.057785 |

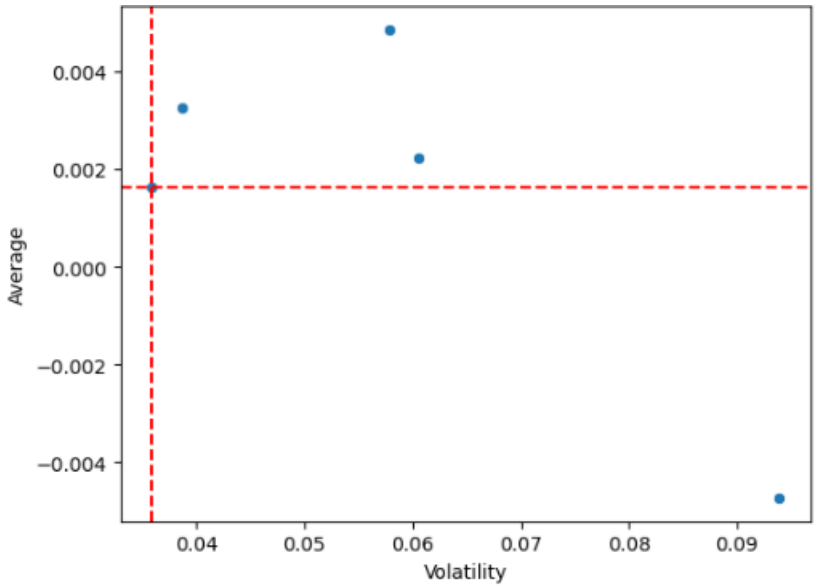
Stock with a lower mean & higher standard deviation do not play a role in a portfolio that has competing stock with more returns & less risk.

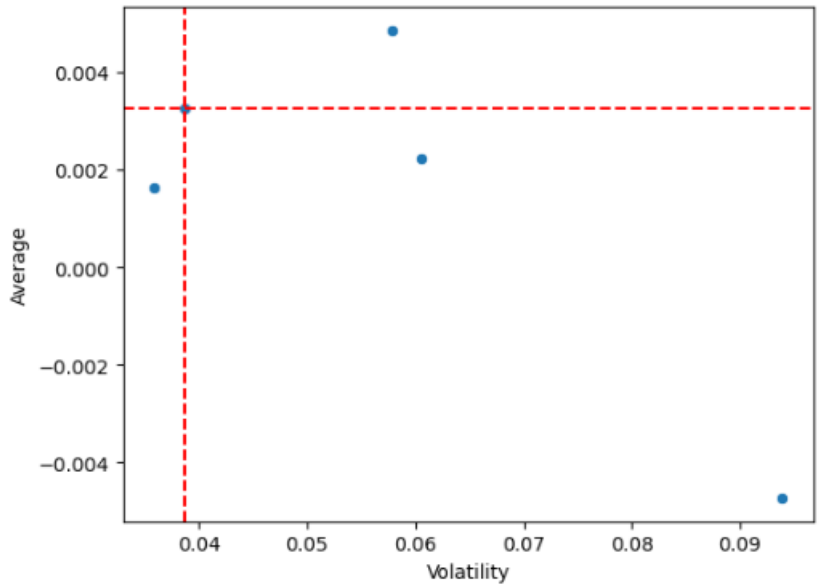
Thus, for the data we have here, we are only left few stocks

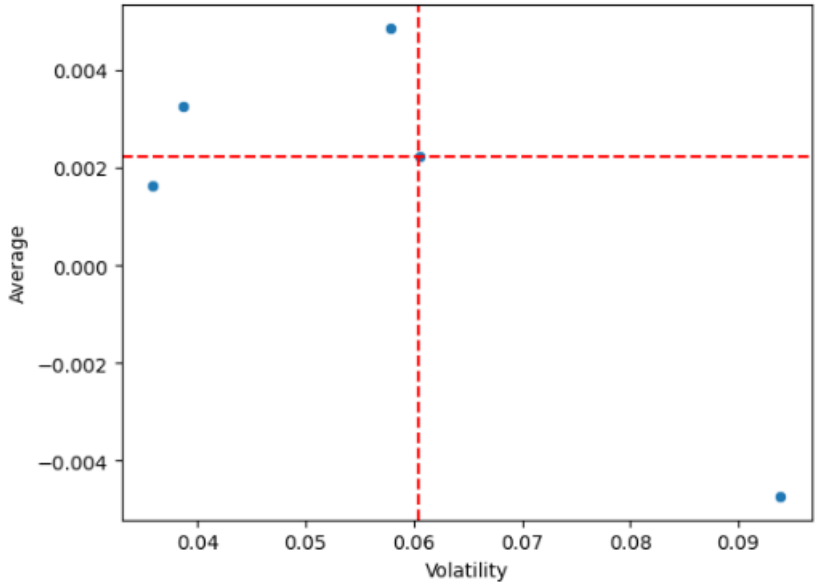
Among the above stocks, DLF_Limited and Bharati_Airtel stocks are having best average with low volatility.

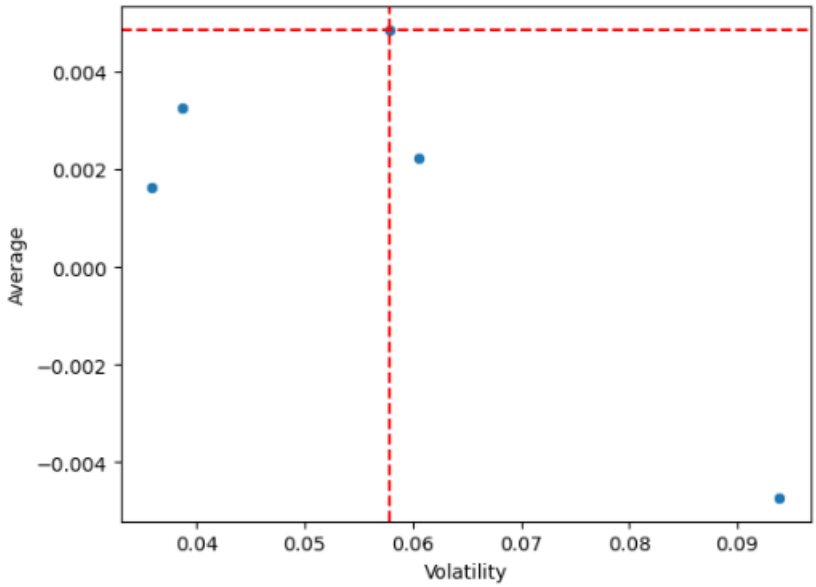
Therefore, the stocks with higher return for a comparative or lower risk are considered better among all the available stocks.

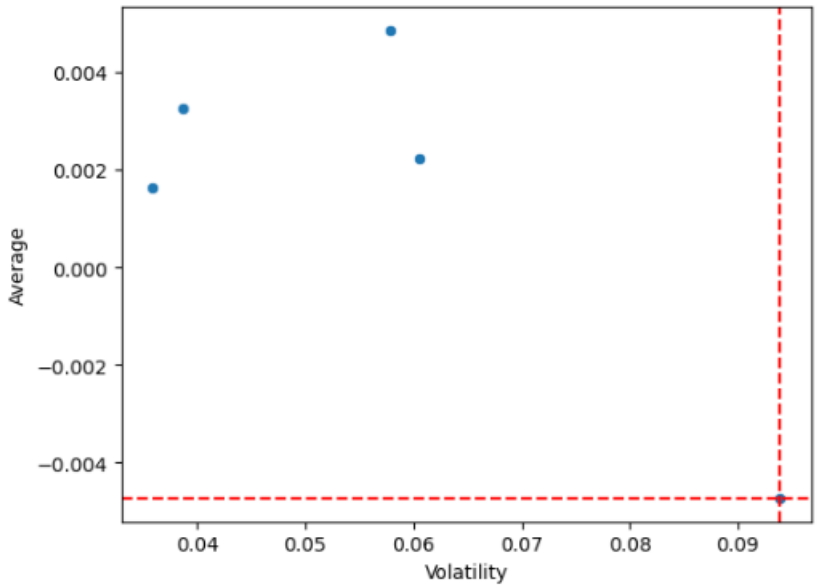
Plot between the means and standard deviations for all stock returns :




0.004
0.002
0.000
−0.002
−0.004
Average
0.04
0.05
0.06
0.07
0.08
0.09
Volatility


0.004
0.002
0.000
−0.002
−0.004
Average
0.04
0.05
0.06
0.07
0.08
0.09
Volatility


0.004
0.002
0.000
−0.002
−0.004
Average
0.04
0.05
0.06
0.07
0.08
0.09
Volatility


0.004
0.002
0.000
−0.002
−0.004
Average
0.04
0.05
0.06
0.07
0.08
0.09
Volatility