Image courtesy: https://www.freepik.com/premium-photo/risk-management-assessment-business-uds_154222385.htm

Financial and Risk Analytics Project

Business Report

November 24, 2024

Authored by: Kartik Trivedi

List of Contents

Data Dictionary……………………………………………………………………5

Executive Summary……………………………………………………………………7

Problem 1……………………………………………………………………13

1.1 Background Information……………………………………………………………………13

1.2 Business Context……………………………………………………………………13

1.3 Problem Statement……………………………………………………………………13

1.4 Methodology……………………………………………………………………13

1.5 Data Overview……………………………………………………………………15

1.6 Exploratory Data Analysis……………………………………………………………………19

1.6.1 Univariate Analysis……………………………………………………………………19

1.6.2 Bivariate Analysis……………………………………………………………………36

1.7 Outlier Treatment……………………………………………………………………37

1.8 Missing Value Treatment……………………………………………………………………39

1.9 Segregating independent and dependent variables……………………………………………………………………47

1.10 Classification Modelling……………………………………………………………………48

1.11 Model Comparison……………………………………………………………………62

1.12 Important Features……………………………………………………………………63

1.13 Conclusions……………………………………………………………………63

Problem 2……………………………………………………………………66

2.1 Background Information……………………………………………………………………66

2.2 Business Context……………………………………………………………………66

2.3 Problem Statement……………………………………………………………………66

2.4 Methodology……………………………………………………………………66

2.5 Data Overview……………………………………………………………………66

2.6 Exploratory Data Analysis……………………………………………………………………68

2.7 Analyzing Returns……………………………………………………………………71

2.8 Conclusion……………………………………………………………………74

List of Figures

Figure 1: Scatterplot return vs risk......11
Figure 2: Univariate Analysis Numeric Columns 36
Figure 3: Heatmap......37
Figure 4: Heatmap......42
Figure 5: Confusion Matrix 55
Figure 6: Confusion Matrix 56
Figure 7: AUC-ROC Curve 57
Figure 8: Confusion Matrix 58
Figure 9: Confusion Matrix......60
Figure 10: Confusion Matrix 61
Figure 11: AUC-ROC Curve 62
Figure 12: Confusion Matrix......63
Figure 13: Price trend over time for different stocks......72
Figure 14: Return vs risk 74

List of Tables

Table 1: Model Comparison......7
Table 2: Important Features......8
Table 3: Average return vs risk......10
Table 4: Sharpe ratio......11
Table 5: Target variable......14
Table 6: Dataset Shape......15
Table 7: Dataset Information......16
Table 8: Missing Values Information......18
Table 9: Data Duplicates......18
Table 10: Statistical Summary......19
Table 11: Outlier count......38
Table 12: Outlier count 39
Table 13: Missing values by columns......40

Table 14: Missing values by rows..42
Table 15: Comparison by defaulters...42
Table 16: Proportion of missing values..44
Table 17: Statistical summary...46
Table 18: Concatenated data columns...47
Table 19: Train and test data shape...47
Table 20: Data shape...48
Table 21: Data shape...48
Table 22: VIF Score..50
Table 23: VIF Score..51
Table 24: Model summary..52
Table 25: Model summary..53
Table 26: Classification Report 54
Table 27: Classification Report...55
Table 28: Classification Report 57
Table 29: Best Parameters..58
Table 30: Classification Report 59
Table 31: Classification Report 60
Table 32: Classification Report 62
Table 33: Model Comparison...63
Table 34: Important Features...63
Table 35: Dataset Shape...67
Table 36: Dataset Information..67
Table 37: Missing Values Information...67
Table 38: Data Duplicates...67
Table 39: Statistical Summary...68
Table 40: Logarithmic returns...72
Table 41: Average return vs risk...72
Table 42: Sharpe ratio..74

List of Equations

Equation 1: Sharpe ratio...73

Data Dictionary

Problem 1

Name	Description	Data Type
Networth Next Year	Net worth of the customer in the next year	Int 64
Total assets	Total assets of customer	Float 64
Net worth	Net worth of the customer of the present year	Float 64
Total income	Total income of the customer	Float 64
Change in stock	Difference between the current value of the stock and the value of stock in the last trading day	Float 64
Total expenses	Total expenses done by the customer	Float 64
Profit after tax	Profit after tax deduction	Float 64
PBDITA	Profit before depreciation, income tax, and amortization	Float 64
PBT	Profit before tax deduction	Float 64
Cash profit	Total Cash profit	Float 64
PBDITA as % of total income	PBDITA / Total income	Float 64
PBT as % of total income	PBT / Total income	Float 64
PAT as % of total income	PAT / Total income	Float 64
Cash profit as % of total income	Cash Profit / Total income	Float 64
PAT as % of net worth	PAT / Net worth	Float 64
Sales	Sales done by the customer	Float 64
Income from financial services	Income from financial services	Float 64
Other income	Income from other sources	Float 64
Total capital	Total capital of the customer	Float 64
Reserves and funds	Total reserves and funds of the customer	Float 64
Borrowings	Total amount borrowed by the customer	Float 64
Current liabilities & provisions	current liabilities of the customer	Float 64
Deferred tax liability	Future income tax customer will pay because of the current transaction	Float 64
Shareholders funds	Amount of equity in a company which belongs to shareholders	Float 64
Cumulative retained profits	Total cumulative profit retained by customer	Float 64
Capital employed	Current asset minus current liabilities	Float 64
TOL/TNW	Total liabilities of the customer divided by Total net worth	Float 64
Total term liabilities / tangible net worth	Short + long term liabilities divided by tangible net worth	Float 64

Contingent liabilities / Net worth (%)	Contingent liabilities / Net worth	Float 64
Contingent liabilities	Liabilities because of uncertain events	Float 64
Net fixed assets	The purchase price of all fixed assets	Float 64
Investments	Total invested amount	Float 64
Current assets	Assets that are expected to be converted to cash within a year	Float 64
Net working capital	Difference between the current liabilities and current assets	Float 64
Quick ratio (times)	Total cash divided by current liabilities	Float 64
Current ratio (times)	Current assets divided by current liabilities	Float 64
Debt to equity ratio (times)	Total liabilities divided by its shareholder equity	Float 64
Cash to current liabilities (times)	Total liquid cash divided by current liabilities	Float 64
Cash to average cost of sales per day	Total cash divided by the average cost of the sales	Float 64
Creditors turnover	Net credit purchase divided by average trade creditors	Float 64
Debtors turnover	Net credit sales divided by average accounts receivable	Float 64
Finished goods turnover	Annual sales divided by average inventory	Float 64
WIP turnover	The cost of goods sold for a period divided by the average inventory for that period	Float 64
Raw material turnover	Cost of goods sold is divided by the average inventory for the same period	Float 64
Shares outstanding	Number of issued shares minus the number of shares held in the company	Float 64
Equity face value	cost of the equity at the time of issuing	Float 64
EPS	Net income divided by the total number of outstanding shares	Float 64
Adjusted EPS	Adjusted net earnings divided by the weighted average number of common shares outstanding on a diluted basis during the plan year	Float 64
Total liabilities	Sum of all types of liabilities	Float 64
PE on BSE	Company's current stock price divided by its earnings per share	Float 64

Problem 2

Name	Description	Data Type
Date	Week starting date.	Object
ITC Limited	Weekly closing price for ITC Limited’s stocks.	Int 64
Bharti Airtel	Weekly closing price for Bharti Airtel's stocks.	Int 64
Tata Motors	Weekly closing price for Tata Motors's stocks.	Int 64
DLF Limited	Weekly closing price for DLF Limited's stocks.	Int 64

Executive Summary

Problem 1

Background Information

In today's financial landscape, managing debt obligations to maintain a favorable credit standing while driving sustainable growth has become increasingly challenging for businesses. As a result, investors and financial institutions must carefully evaluate companies that can effectively navigate financial complexities while maintaining stability and profitability. A company's balance sheet is a crucial tool in this assessment, offering a detailed snapshot of its assets, liabilities, and shareholders' equity. This comprehensive overview provides valuable insights into a business's financial health and operational efficiency, supporting informed decision-making and strategic planning.

Business Objective

The current financial challenges have created a unique opportunity for venture capitalists. A group of them has collaborated to develop a Financial Health Assessment Tool designed to perform Debt Management Analysis and Credit Risk Evaluation on historical financial statements. This tool aims to generate valuable insights that will support informed decision-making.

Problem Statement

The objective of this project is to analyze financial metrics data from various companies to identify potential challenges in their financial performance and develop proactive strategies for effective risk mitigation.

Model Comparison

We created 4 models using logistic regression and random forest techniques and compared each model’s performance for test and train data using key metrics and found that all the models are stable, here we will compare these models with each other to find the best model based on combination of Accuracy, Precision and Recall scores for test data.

	Model	Accuracy	Precision	Recall
0	Logit_model	0.79	0.50	0.02
1	Logit_model_optimal	0.70	0.30	0.30
2	RF_model	0.79	0.57	0.04
3	RF_model_optimal	0.77	0.44	0.27

Table 1: Model Comparison

On evaluating all the models based on combination of Accuracy, Precision and Recall scores Random Forest model optimized for threshold is performing the best as it is providing the best balance for all the three metrics wherein other models are performing significantly poorly on 1 of the 3 metrics. Moving forward we will take this model as the final model. We will check for the most important features which play crucial role in distinguishing between classes.

Important Features

	imp
TOL_to_TNW	0.15
PBT_as_perc_of_total_income	0.12
Cash_profit_as_perc_of_total_income	0.10
PAT_as_perc_of_total_income	0.08
Reserves_and_funds	0.07

Table 2: Important Features

On examining the top 5 most important features for RF_model_optimal, TOL_to_TNW emerges as the most influential, contributing 15% of the model's total importance. TOL_to_TNW reflects the proportion of total liabilities to a company's net worth, indicating the extent to which its assets are financed by debt rather than equity. A higher value signifies greater financial leverage and potentially increased financial risk, making it a crucial factor for predicting financial performance and identifying default risks.

Similarly, other significant features, such as PBT_as_perc_of_total_income, Cash_profit_as_perc_of_total_income, PAT_as_perc_of_total_income, and Reserves_and_funds, provide insights into a company's profitability and cash flow. These metrics play a vital role in assessing a company's ability to generate income, maintain liquidity, and service its liabilities effectively. Together, these features offer a comprehensive view of a company's financial health, aiding in accurate predictions and proactive risk management.

Conclusion

Key Takeaways

1. The dataset comprises over 50 attributes for each company. However, upon analysis, it was observed that nearly 50% of the companies had more than 10% of their data missing. Further investigation revealed that these companies with higher proportions of missing data exhibited a significantly higher likelihood of default.
2. For the classification models developed, the Random Forest model with an adjusted threshold emerged as the best performer, offering the most balanced trade-off between accuracy, precision, and recall—key metrics for evaluating model effectiveness. Models using the standard threshold performed poorly in terms of recall, often misclassifying nearly all defaulters as non-defaulters, which significantly undermines the model's utility. Among the models tested, the Logistic Regression model with an adjusted threshold had the weakest performance, with the lowest accuracy and precision scores. This indicates that it struggled to classify companies correctly and exhibited the highest rate of misclassification for both defaulters and non-defaulters, which could lead to negative consequences if deployed in real-world scenarios.

3. The primary goal of this project is to classify companies based on their ability to meet future financial obligations. To achieve this, key factors should include metrics that offer insights into a company's income-generating capacity and cash flow stability. Upon analyzing the most significant features in the best-performing model, Total Liabilities to Total Net Worth (TOL_to_TNW) emerged as the top contributor, indicating the degree of financial leverage and risk associated with the company. Other important features include:

- Profit Before Tax (PBT) as a Percentage of Total Income
- Profit After Tax (PAT) as a Percentage of Total Income
- Cash Profit as a Percentage of Total Income
- Reserves and Surplus

These factors collectively provide a comprehensive understanding of a company's current financial health, operational efficiency, and capacity to generate income. By incorporating these features, the model ensures a more accurate prediction of a company's ability to meet its financial obligations, thereby aiding in effective decision-making.

Key Recommendations

1. Companies with over 10% missing data have demonstrated a significantly higher probability of default. It is recommended to conduct a thorough investigation to determine whether this non-disclosure is incidental or a deliberate attempt to withhold critical information. Establishing the intent behind these gaps in data can provide valuable insights into patterns of non-compliance or potentially fraudulent activity. This investigation will not only enhance the reliability of the dataset but also help refine the model's ability to identify high-risk companies effectively.
2. We have successfully built models using logistic regression and random forest and identified the best-performing model. However, there is considerable scope for improvement, especially regarding precision and recall. To address these limitations and enhance model performance, we recommend the following:

- Approximately 8% of the dataset was missing, which is significant, given that some variables were derived from others. Furthermore, the possibility of deliberate non-disclosure raises concerns about the reliability of the data. To ensure completeness and trustworthiness, it is recommended that future datasets are sourced directly from audited financial statements of the companies. This would eliminate doubts about data integrity and provide a more robust foundation for model development.
- Logistic regression, which was a mandatory model for this project, is highly sensitive to outliers. Consequently, an outlier treatment process was applied to the dataset, affecting over 8% of the data (based on conservative thresholds at the 5th and 95th percentiles). This resulted in over 16% of the data being imputed, likely impacting model performance. Given the high prevalence of outliers and missing data, we recommend exploring alternative modeling techniques such as decision trees, bagging, and boosting methods. These models are less sensitive to outliers and better equipped to handle missing data, potentially yielding improved results.
- Features related to income generation, cash flows, and financial standing were identified as the most important predictors of default. To enhance predictive power, we recommend collecting financial records from the past few years in addition to the current year. This historical data can be used to build regression models that forecast future performance, which can then be integrated into the classification

model. This approach will likely provide a more comprehensive understanding of the company's financial trajectory and improve overall model accuracy.

Problem 2

Background Information

Investing in financial markets involves substantial risk, primarily driven by potential price fluctuations of assets. These swings often result from unforeseen economic events or geopolitical developments, which can drastically impact investor sentiment and market dynamics.

Business Context

Given the significant risks inherent in financial markets, it is crucial for investors to assess and understand the risks they are undertaking. This understanding enables them to align their investment strategies with their financial objectives, fostering informed decision-making and portfolio optimization.

Problem Statement

The objective of this is to develop a robust risk evaluation framework that leverages historical market data by quantifying and predicting potential risks, the framework aims to guide investors in selecting investment strategies that balance risk and reward effectively, ultimately supporting their financial goals.

Mean vs Standard Deviation for all stock returns

	Average	Volatility
ITC_Limited	0.0016	0.0359
Bharti_Airtel	0.0033	0.0387
DLF_Limited	0.0049	0.0578
Tata_Motors	0.0022	0.0605
Yes_Bank	-0.0047	0.0939

Table 3: Average return and risk

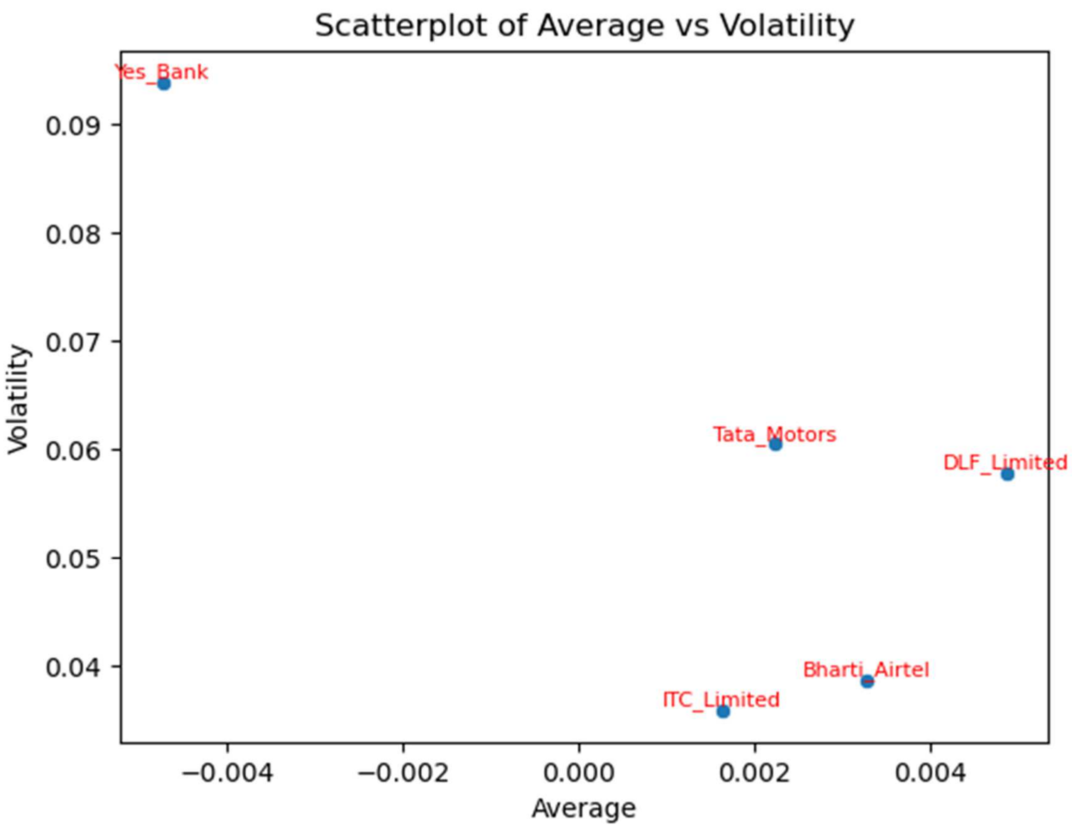


Figure 1: Scatterplot return vs risk

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:

- ITC_Limited
- Bharti_Airtel
- DLF_Limited

To identify the stocks which give the best balance between risk and return we can evaluate the Sharpe ratio. For Sharpe ratio we need risk free return which is normally considered to be rate for government bonds which currently is 5% per annum.

Sharpe Ratio

	Sharpe_Ratio
DLF_Limited	0.0675
Bharti_Airtel	0.0596
Tata_Motors	0.0210
ITC_Limited	0.0187
Yes_Bank	-0.0607

Table 4: Sharpe ratio

Evaluating stocks solely based on average return and volatility can lead to misleading conclusions. For instance, ITC Limited shows the lowest volatility, followed by Bharti Airtel, which might initially suggest they are the best-performing stocks. However, this simplistic assessment overlooks the balance between risk and return. When we incorporate Sharpe's Ratio, which evaluates performance relative to risk, a different picture emerges. DLF Limited stands out as the best-performing stock, followed by Bharti Airtel. Interestingly, despite its low volatility, ITC Limited ranks as the second-worst in terms of Sharpe's Ratio, highlighting the importance of a comprehensive evaluation that accounts for both risk and return.

Conclusion

The Market Risk Analysis provided valuable insights into the risk-return dynamics of a portfolio. By incorporating statistical measures and the Sharpe ratio, we were able to move beyond simplistic metrics like mean return and volatility, enabling a more comprehensive evaluation of portfolio performance. Key insights and actionable recommendations are as follows:

Key Insights

1. The analysis underscores the importance of considering both risk and return when evaluating stocks. Solely relying on metrics like average return or volatility can be misleading, as they fail to account for the risk-adjusted performance of investments.
2. By integrating the Sharpe Ratio, we identified that DLF Limited offers the best risk-adjusted returns, despite having higher volatility compared to other stocks like ITC Limited and Bharti Airtel. This demonstrates the necessity of incorporating comprehensive measures for informed decision-making.
3. Although ITC Limited has the lowest volatility, it performs poorly in terms of risk-adjusted returns. This highlights that low risk does not necessarily translate to high performance if returns are not proportionately higher.
4. Bharti Airtel emerges as a strong contender with a balanced performance, making it a viable choice for investors seeking moderate risk and returns.

Key Recommendations

1. Rather than relying solely on standalone metrics such as average return or volatility incorporating risk-adjusted measures like the Sharpe Ratio to gain a complete understanding of stock performance could be more beneficial.
2. DLF Limited, with the highest Sharpe Ratio, should be considered a top priority for inclusion in the portfolio, as it offers the best balance of return relative to risk.
3. ITC Limited’s lower Sharpe Ratio suggests it may not add substantial value to the portfolio. Reassess its inclusion, especially if there are other stocks offering better risk-adjusted returns.
4. While focusing on high Sharpe Ratio stocks, it recommended that the portfolio remains diversified to minimize exposure to stock-specific risks and maintain a balance of industries.
5. Continuously monitoring the portfolio performance and market conditions and adjusting stock allocations based on evolving Sharpe Ratios and changing economic scenarios could be beneficail to sustain optimal risk-adjusted returns.

Problem 1

1.1 Background Information

In today's financial landscape, managing debt obligations to maintain a favorable credit standing while driving sustainable growth has become increasingly challenging for businesses. As a result, investors and financial institutions must carefully evaluate companies that can effectively navigate financial complexities while maintaining stability and profitability. A company's balance sheet is a crucial tool in this assessment, offering a detailed snapshot of its assets, liabilities, and shareholders' equity. This comprehensive overview provides valuable insights into a business's financial health and operational efficiency, supporting informed decision-making and strategic planning.

1.2 Business Objective

The current financial challenges have created a unique opportunity for venture capitalists. A group of them has collaborated to develop a Financial Health Assessment Tool designed to perform Debt Management Analysis and Credit Risk Evaluation on historical financial statements. This tool aims to generate valuable insights that will support informed decision-making.

1.3 Problem Statement

The objective of this project is to analyze financial metrics data from various companies to identify potential challenges in their financial performance and develop proactive strategies for effective risk mitigation.

1.4 METHODOLOGY

Import the libraries – Load the data – Check the structure of the data – Check the types of the data – Check for missing values – Check the statistical summary – Check for and treat (if needed) Data Irregularities – Extract target variable – Drop irrelevant columns – Univariate Analysis – Bivariate Analysis – Check for outliers and (if needed) convert to missing values – Drop columns with over 30% missing data – Data Scaling – Missing value imputation – Data Splitting – Apply Classification Models – Predict values – Evaluate model – Compare model – Get Important Features – Conclusion

Key Points

1. **Data Collection**: Data was provided which contained information regarding the financial metrics of 4265 different countries different companies.
2. **Target Variable**: Target variable was created using column Networth Next Year where companies with negative net worth were considered defaulters assigning value 1 to them.

	default	Networth_Next_Year
0	0	395.30
1	0	36.20
2	0	84.00
3	0	2041.40
4	0	41.80
5	0	291.50
6	0	93.30
7	0	985.10
8	0	188.60
9	0	229.60

Table 5: Target variable

```
Value count for defaulters

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

```
 Proportion of defaulters

 default
 0   0.7876
 1   0.2124
 Name: proportion, dtype: float64
```

In the given data about 21% of companies are considered defaulters

3. **Data Cleaning and Pre-processing:** The dataset was thoroughly examined for column names, duplicates, missing values, bad data, and outliers. An irrelevant column, 'Num', was identified and removed. Additionally, 'Networth Next Year' was dropped as it was used to derive the target variable, and 'Equity_face_value' was removed because it did not contribute meaningful information, given that it remains constant or identical for most companies. Inconsistent column names were also standardized by renaming relevant attributes to ensure uniformity in nomenclature.
4. **Univariate Analysis:** Individual variables were analyzed using boxplot and histogram to understand distribution, central tendency and variability of variables.
5. **Bivariate Analysis:** All the variables were examined with the aim of gaining deeper insights about correlation between attributes.
6. **Visualization Techniques:** In the report we have used histograms and boxplot for univariate analysis, in bivariate analysis, to understand correlation between numeric variables heatmap is used.

7. **Tools and Software:** We have carried out the analysis using programming language python on Jupyter notebook. For this analysis Python libraries Numpy, Pandas, Matplotlib, Seaborn, Statsmodel and Scikit-learn were used.

1.5 Data Overview

1. **Data Description:** Dataset has 4256 rows and 51 columns.

```
shape of the dataset
----------------------------------------------------------------------------

(4256, 51)
```

Table 6: Dataset Shape

2. **Dataset Information:** Of the 51 columns in the dataset, 1 is int 64 type and 50 are float 64 type.

```
information of features
----------------------------------------------------------------------------
<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 % of total income                    4177 non-null   float64
 12  PBT as % of total income                       4177 non-null   float64
 13  PAT as % of total income                       4177 non-null   float64
 14  Cash profit as % of total income               4177 non-null   float64
 15  PAT as % 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/TNW                                        4256 non-null   float64
 28  Total term liabilities / tangible net worth    4256 non-null   float64
 29  Contingent liabilities / Net worth (%)         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

None
```

Table 7: Dataset Information

3. **Missing Value Check:** There were over 8% missing values in the dataset.

```
Proportion of missing values
8.19 %
```

```
missing values
----------------------------------------------------------------------------

Num                                                   0
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 % of total income                          79
PBT as % of total income                             79
PAT as % of total income                             79
Cash profit as % of total income                     79
PAT as % 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/TNW                                               0
Total term liabilities / tangible net worth           0
Contingent liabilities / Net worth (%)                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
Equity face value                                   810
EPS                                                   0
Adjusted EPS                                          0
Total liabilities                                     0
PE on BSE                                          2627
dtype: int64
```

Table 8: Missing values information

4. **Duplicate Values:** Data was checked for duplicate values and no duplicates were found

```
checking for duplicates
-----------------------------------------------------------------------------
number of dupliacte rows: 0
```

Table 9: Data Duplicates

5. **Statistical Summary:**

```
statistical summary
--------------------------------------------------------------------------------
```

	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	51398.09	-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 % of total income	4177.00	3.18	172.26	-6400.00	4.97	9.68	16.47	100.00
PBT as % of total income	4177.00	-18.20	419.91	-21340.00	0.56	3.34	8.94	100.00
PAT as % of total income	4177.00	-20.03	423.58	-21340.00	0.35	2.37	6.42	150.00
Cash profit as % of total income	4177.00	-9.02	299.96	-15020.00	2.00	5.66	10.73	100.00
PAT as % 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/TNW	4256.00	4.03	20.88	-350.48	0.60	1.42	2.83	473.00

Total term liabilities / tangible net worth	4256.00	1.85	15.88	-325.60	0.05	0.34	1.00	456.00
Contingent liabilities / Net worth (%)	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	559506.80
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	3182.03	-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

Table 10: Statistical Summary

Key observations

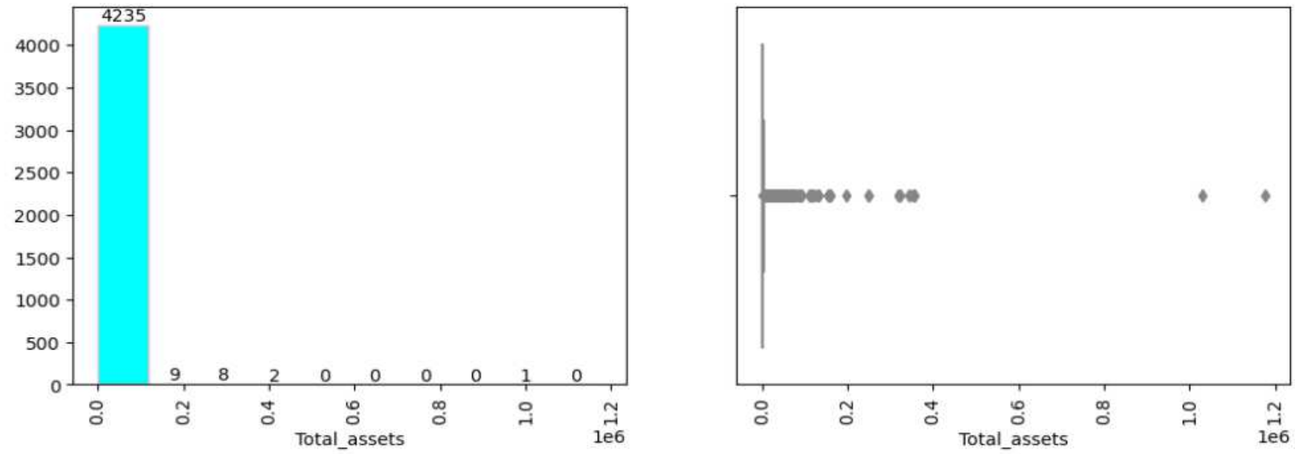
1. Column names are messy (has spaces) are inconsistent which we will have to fix.
2. There are 4256 rows and 51 columns in the dataset.
3. The dataset comprises financial data, and as expected, all columns have numeric data types (either integers or floats). This consistency indicates that the dataset is free from junk data.
4. There are missing values in the dataset, on checking more thoroughly missing values account for over 8% of the data in the dataset.
5. The dataset does not include a predefined target variable. However, given the problem's objective of identifying companies likely to face financial difficulties, we will define a company as a "defaulter" if its net worth in the following year is negative.
6. Column 'Num' contains serial numbers which are irrelevant for our analysis and equity face value remains constant which makes it irrelevant, additionally, 'Networth Next Year' will be used to extract the target variable. We drop both these columns.

1.6 Exploratory Data Analysis

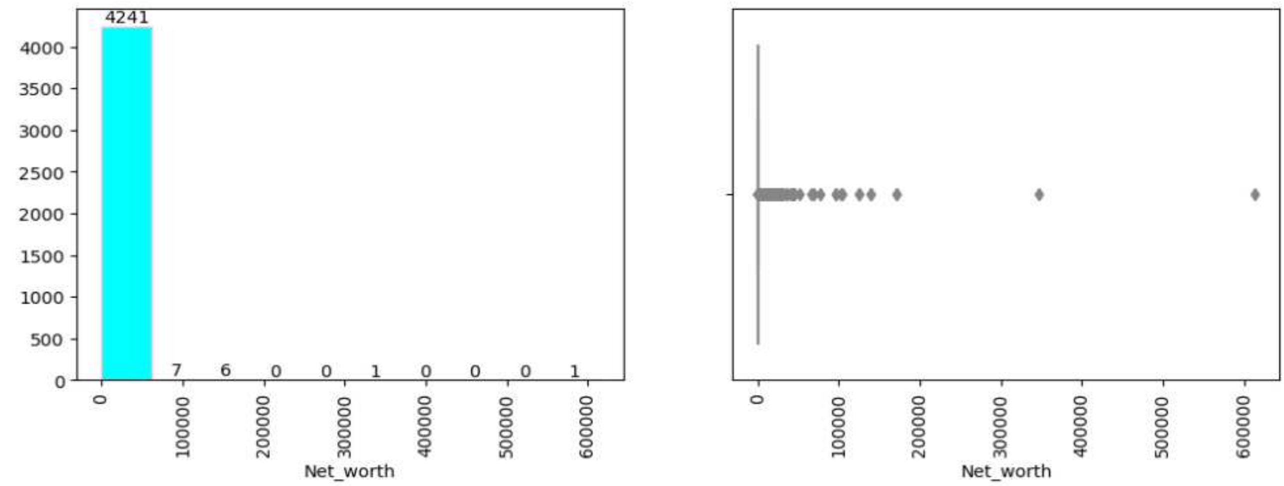
1.6.1 Univariate Analysis

For numeric columns

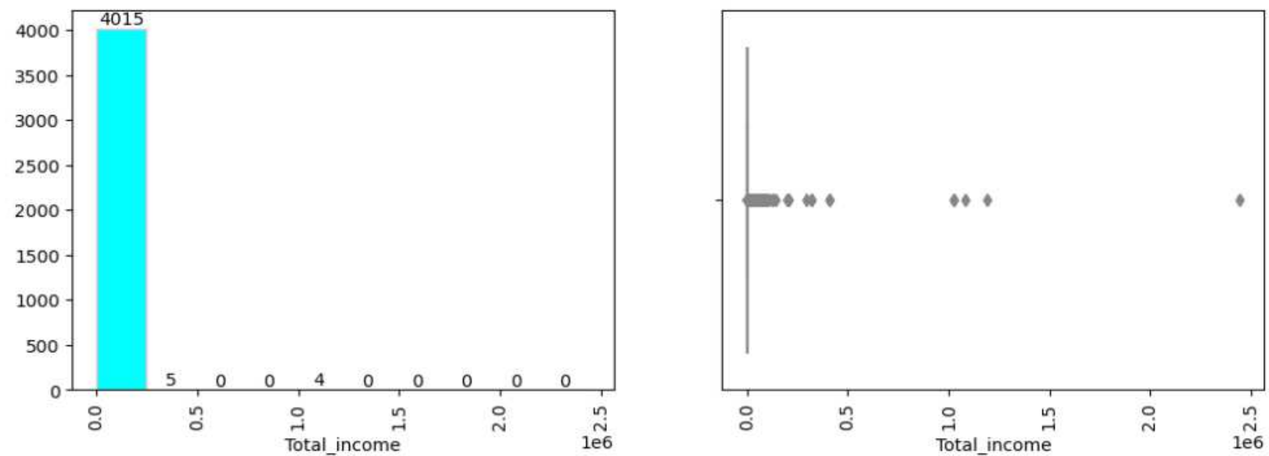
```
Skewness of Total_assets:  26.422680474857692
Distribution of Total_assets
----------------------------------------------------------------------------
```



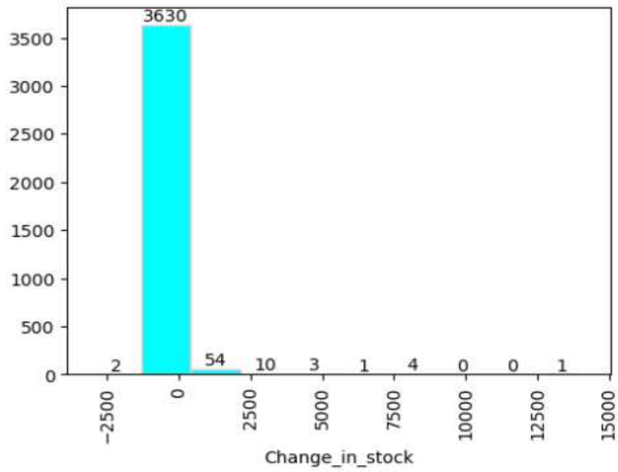

```
Skewness of Net_worth:  31.85168555023475
Distribution of Net_worth
----------------------------------------------------------------------------
```

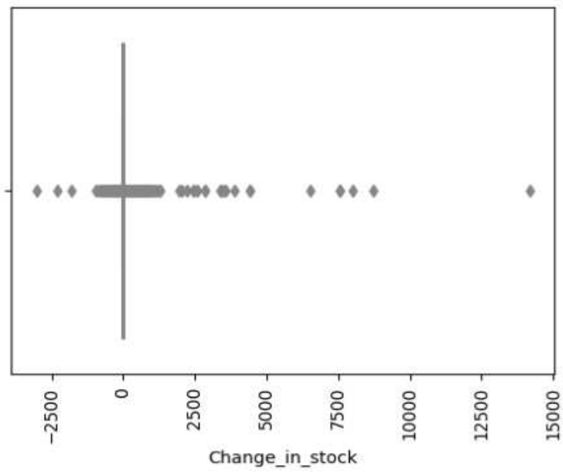



```
Skewness of Total_income:  31.443117127058954
Distribution of Total_income
----------------------------------------------------------------------------
```

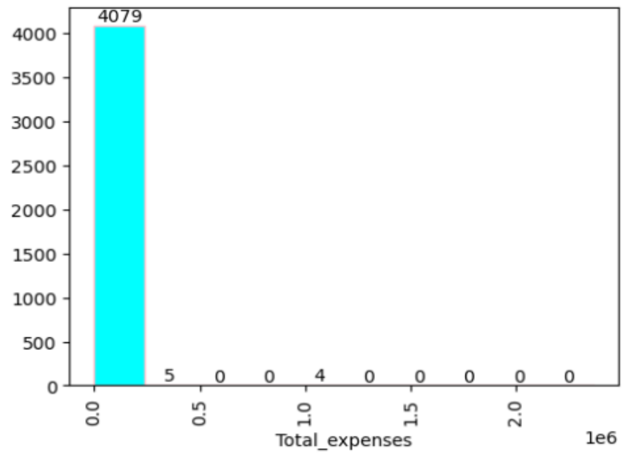


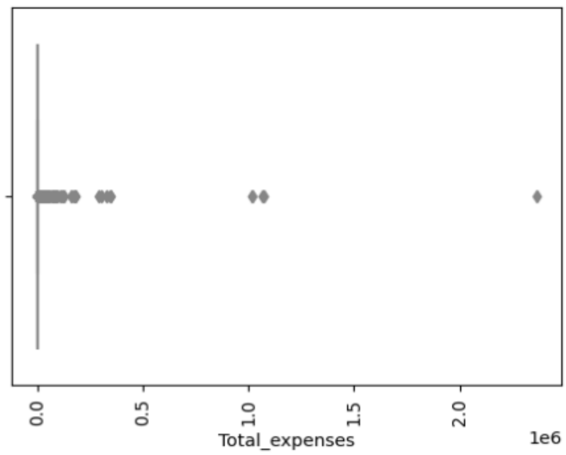
```
Skewness of Change_in_stock:  18.02425906208548
Distribution of Change_in_stock
----------------------------------------------------------------------------
```



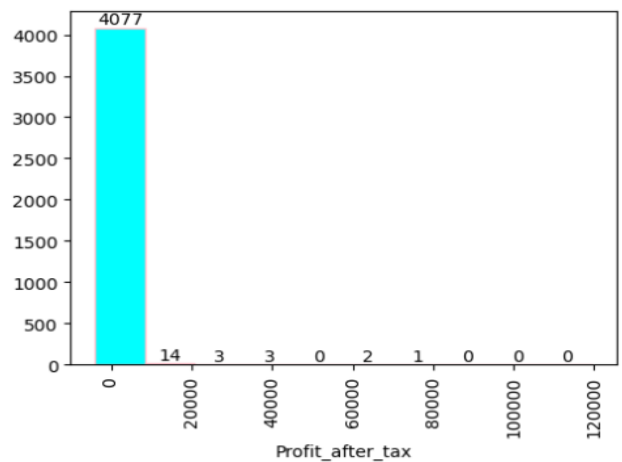


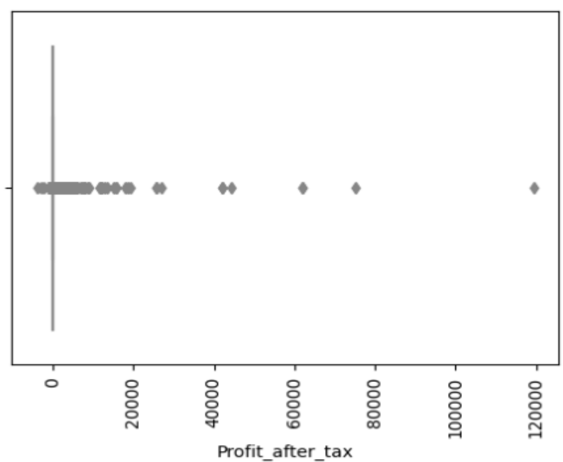

```
Skewness of Total_expenses:  32.19039096721928
Distribution of Total_expenses
----------------------------------------------------------------------------
```



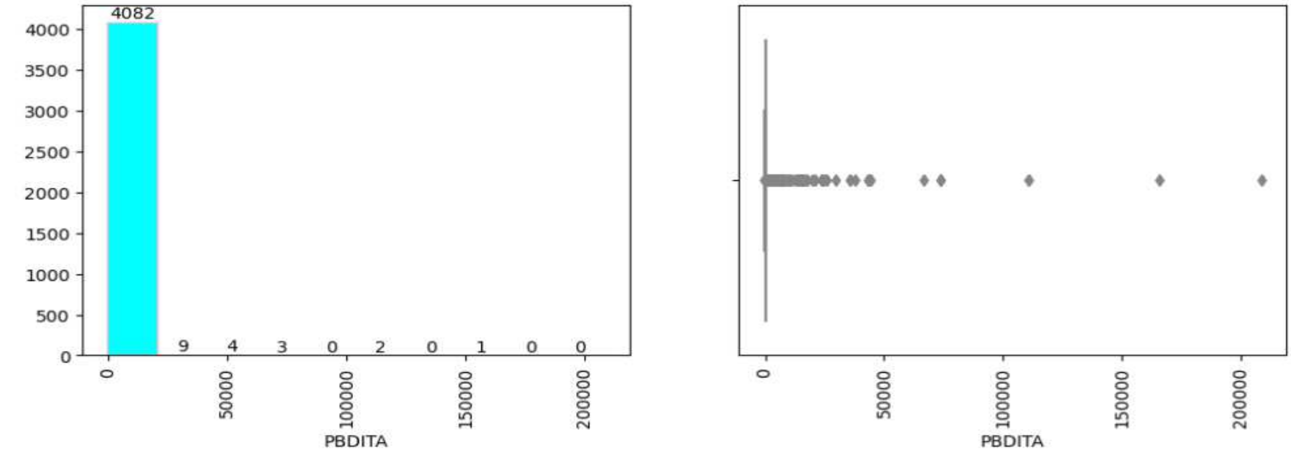



```
Skewness of Profit_after_tax:  24.290605539925448
Distribution of Profit_after_tax
----------------------------------------------------------------------------
```

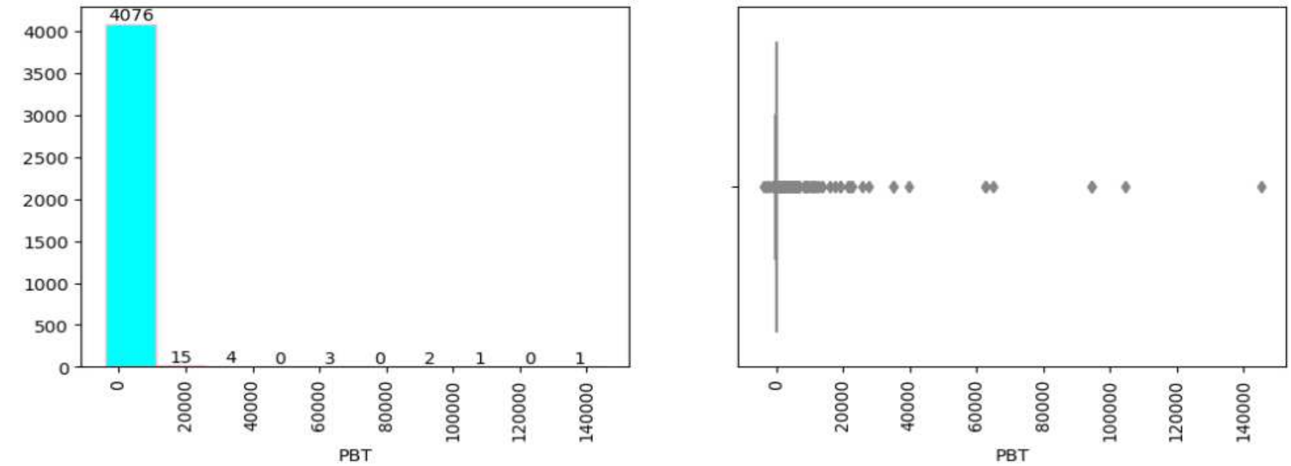




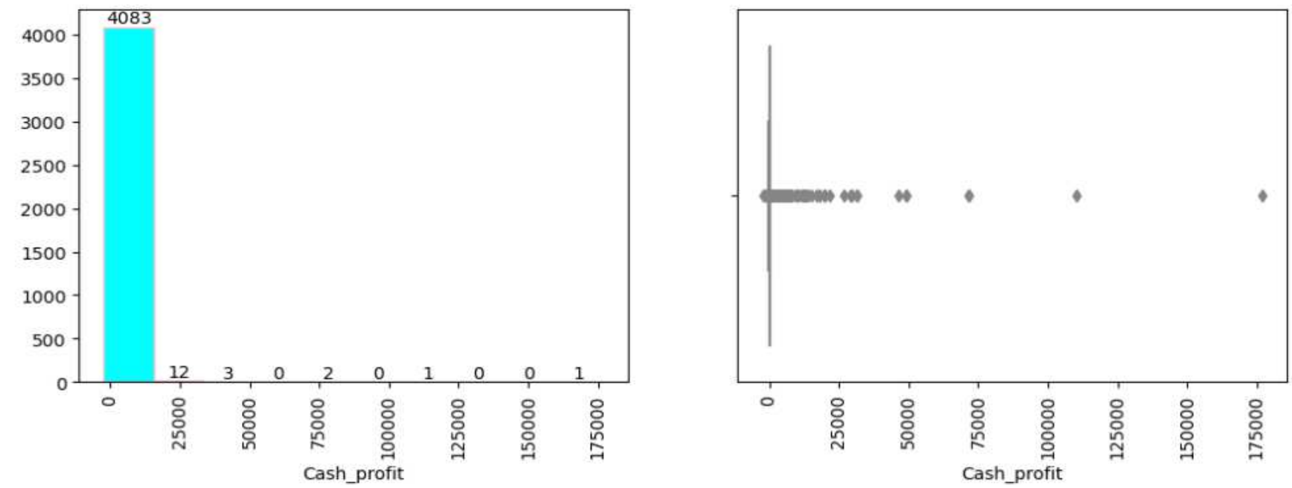
```
Skewness of PBDITA:  24.124350397794316
Distribution of PBDITA
----------------------------------------------------------------------------
```



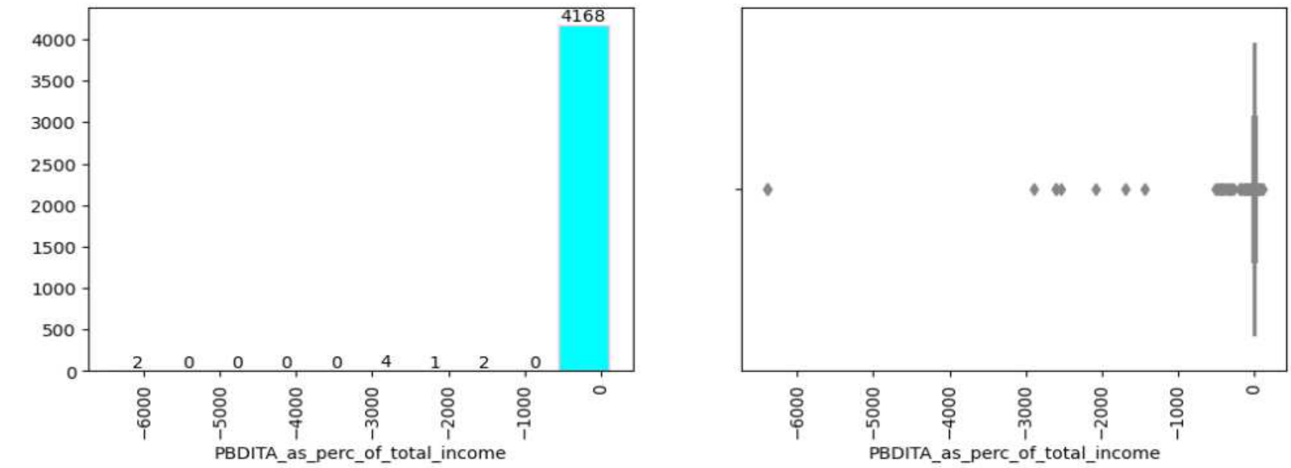

```
Skewness of PBT:  22.27588296254738
Distribution of PBT
----------------------------------------------------------------------------
```



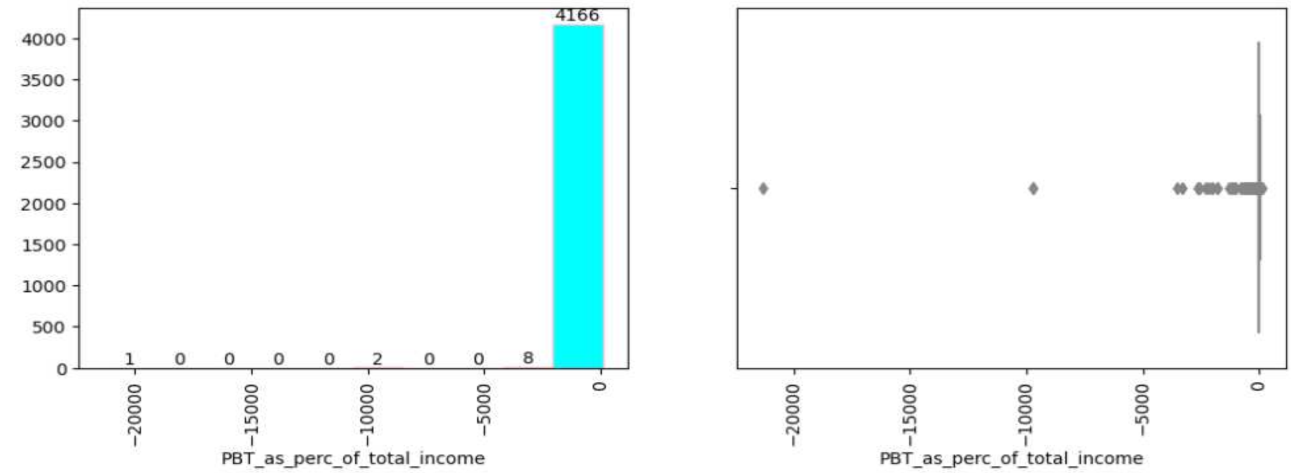

```
Skewness of Cash_profit:  27.667906279757602
Distribution of Cash_profit
----------------------------------------------------------------------------
```



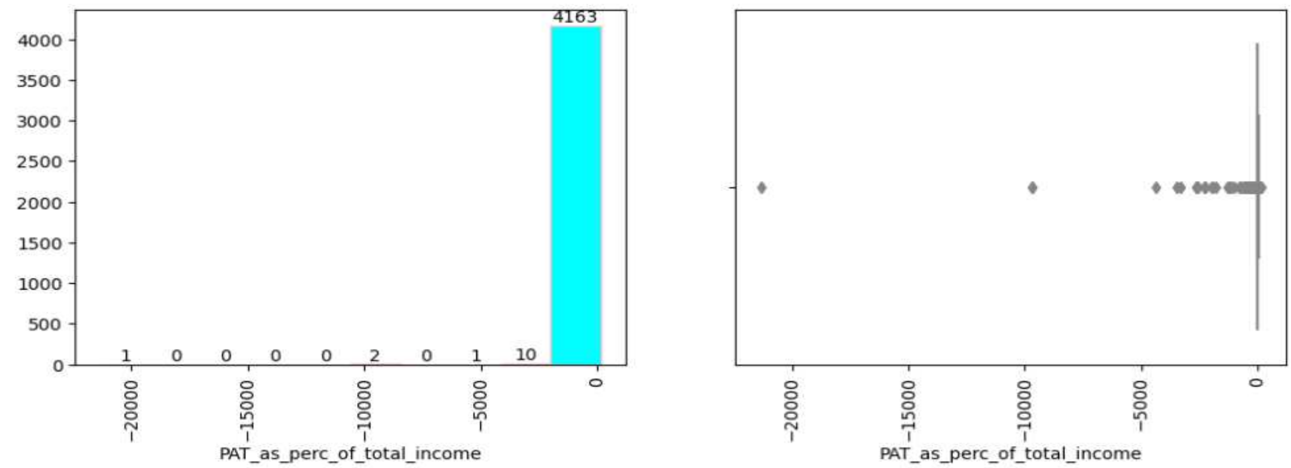
```
Skewness of PBDITA_as_perc_of_total_income:  -29.030768915099028
Distribution of PBDITA_as_perc_of_total_income
----------------------------------------------------------------------------
```



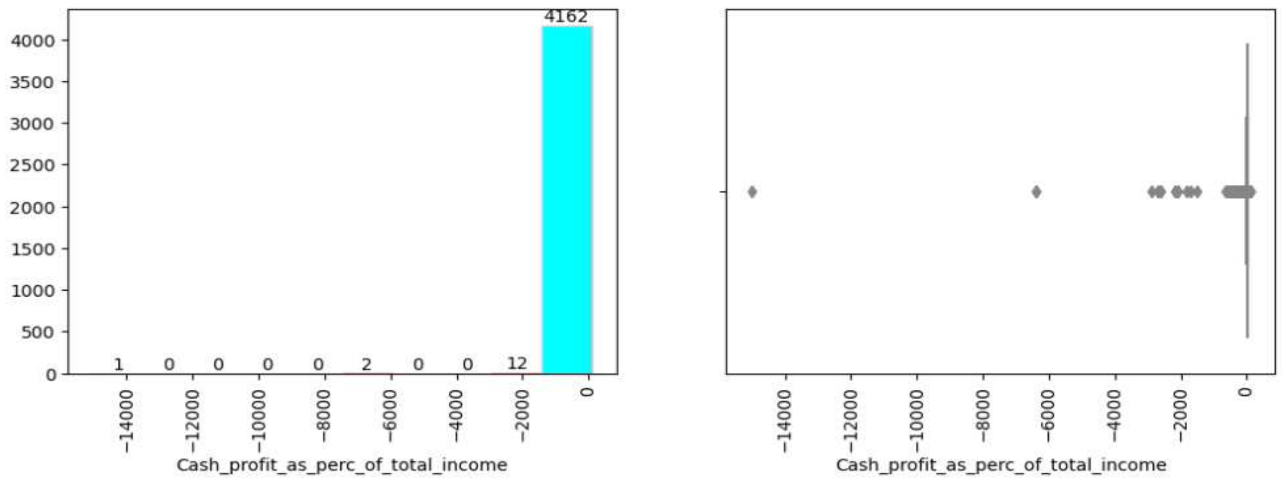

```
Skewness of PBT_as_perc_of_total_income:  -37.93698143766266
Distribution of PBT_as_perc_of_total_income
----------------------------------------------------------------------------
```



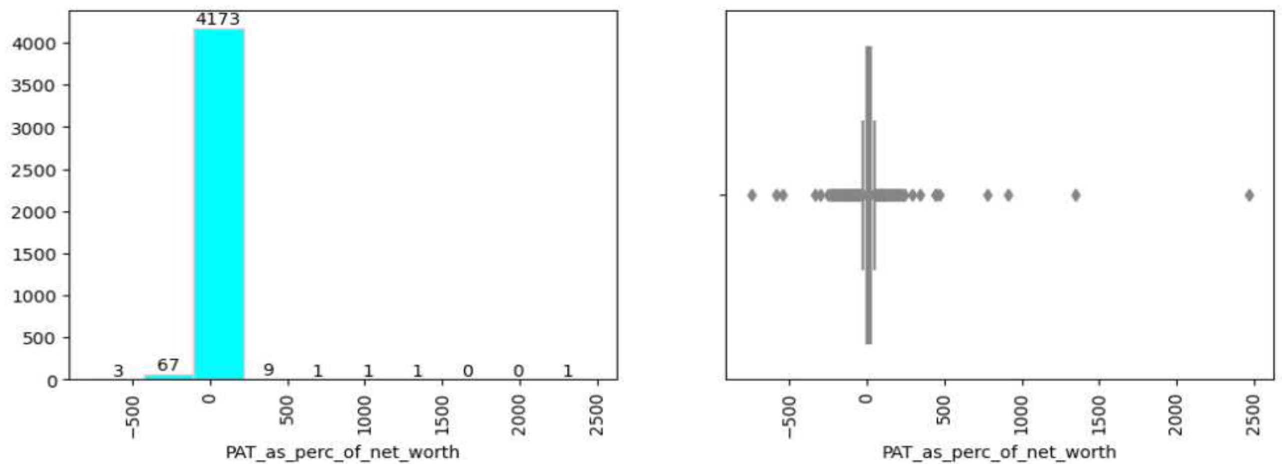

```
Skewness of PAT_as_perc_of_total_income:  -37.170127782409594
Distribution of PAT_as_perc_of_total_income
----------------------------------------------------------------------------
```



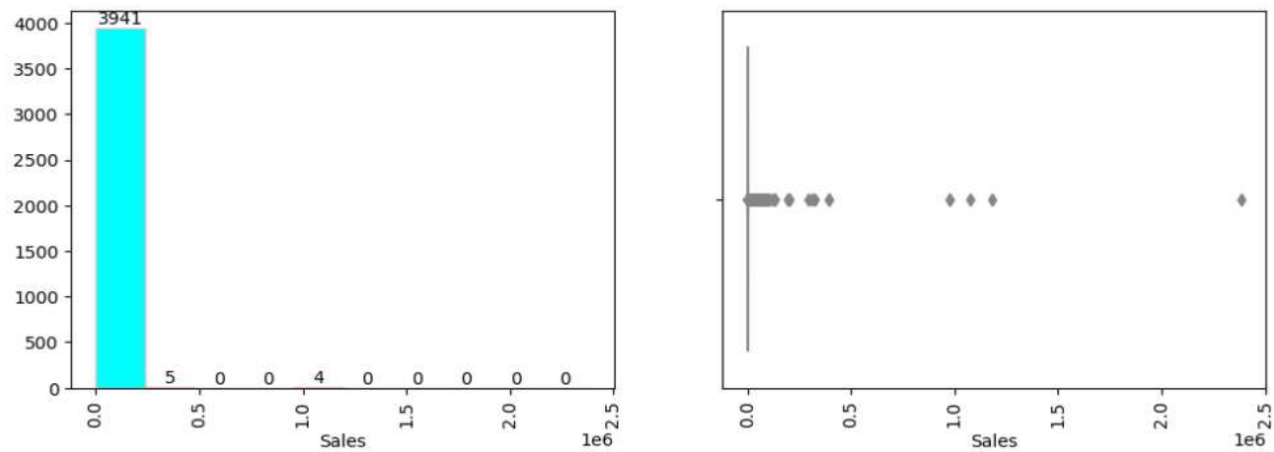
```
Skewness of Cash_profit_as_perc_of_total_income:  -36.017774923113926
Distribution of Cash_profit_as_perc_of_total_income
----------------------------------------------------------------------------
```



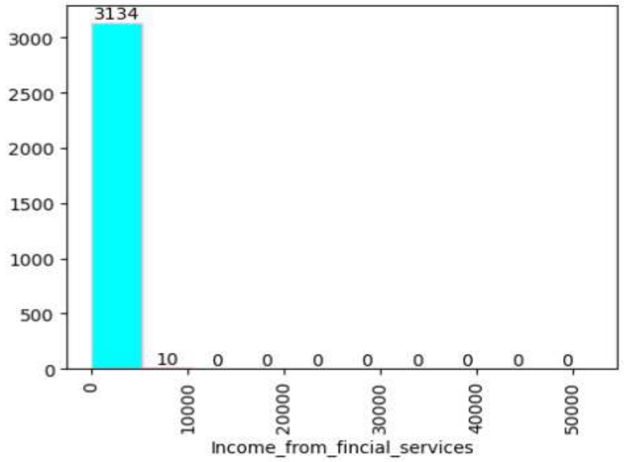
```
Skewness of PAT_as_perc_of_net_worth:  17.76197818185262
Distribution of PAT_as_perc_of_net_worth
----------------------------------------------------------------------------
```

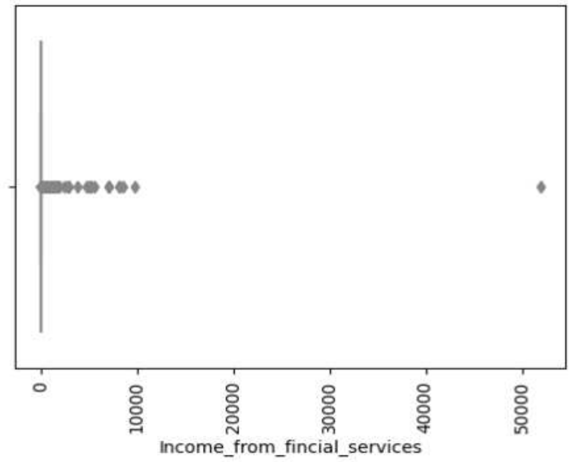


```
Skewness of Sales:  31.233586758881085
Distribution of Sales
----------------------------------------------------------------------------
```

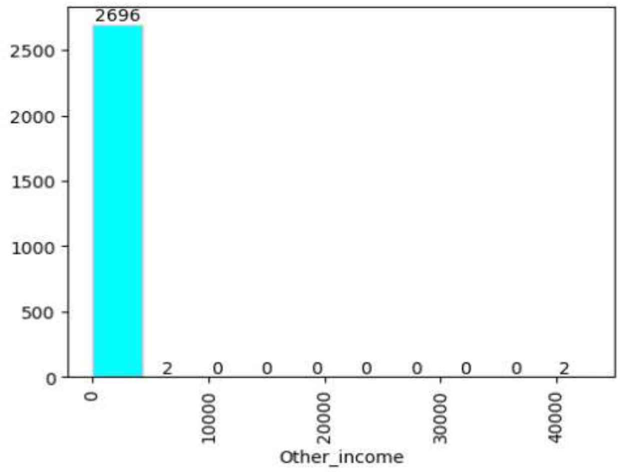


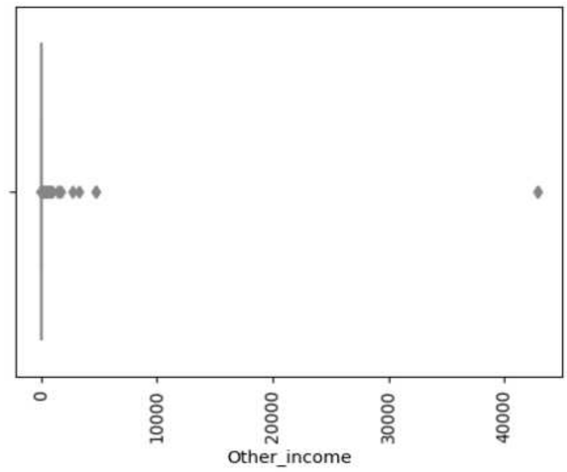
```
Skewness of Income_from_fincial_services:  40.46214235747733
Distribution of Income_from_fincial_services
----------------------------------------------------------------------------
```



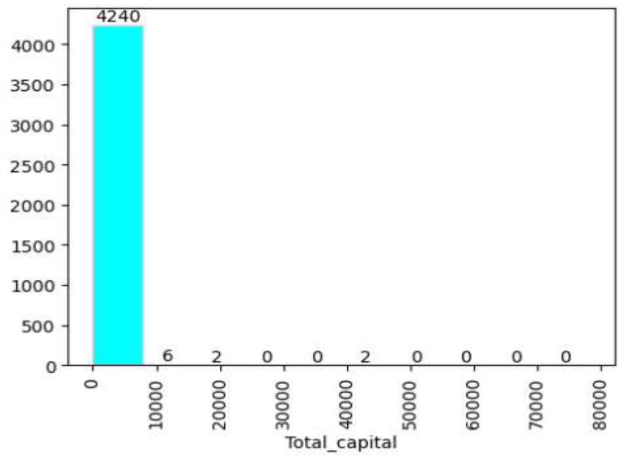


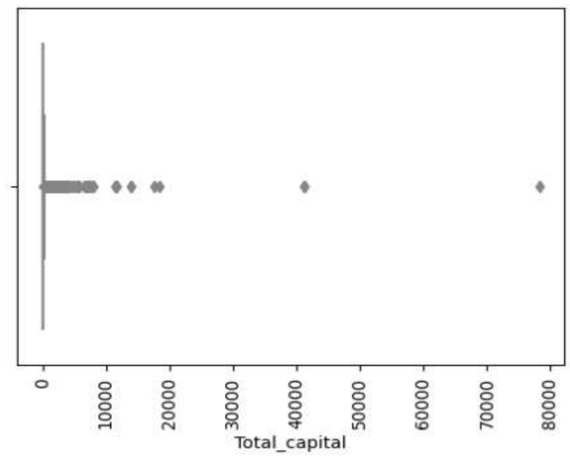

```
Skewness of Other_income:  35.59157972695797
Distribution of Other_income
----------------------------------------------------------------------------
```



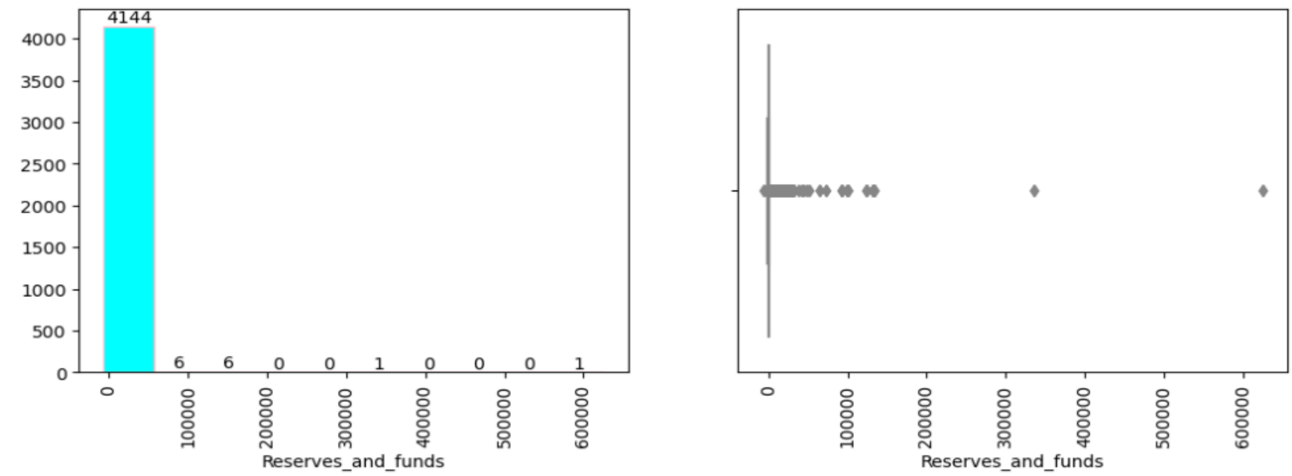



```
Skewness of Total_capital:  31.49232680482334
Distribution of Total_capital
----------------------------------------------------------------------------
```

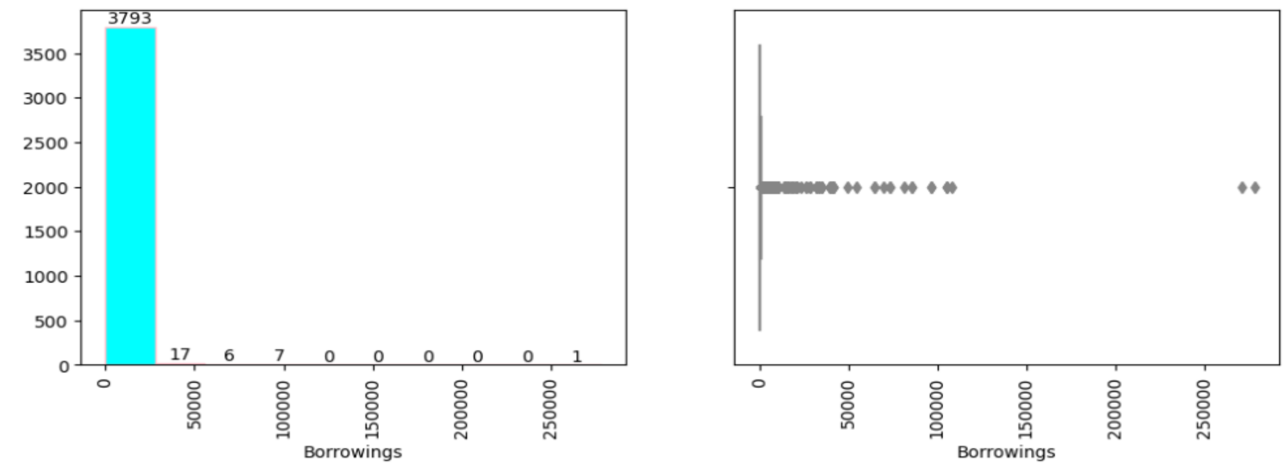




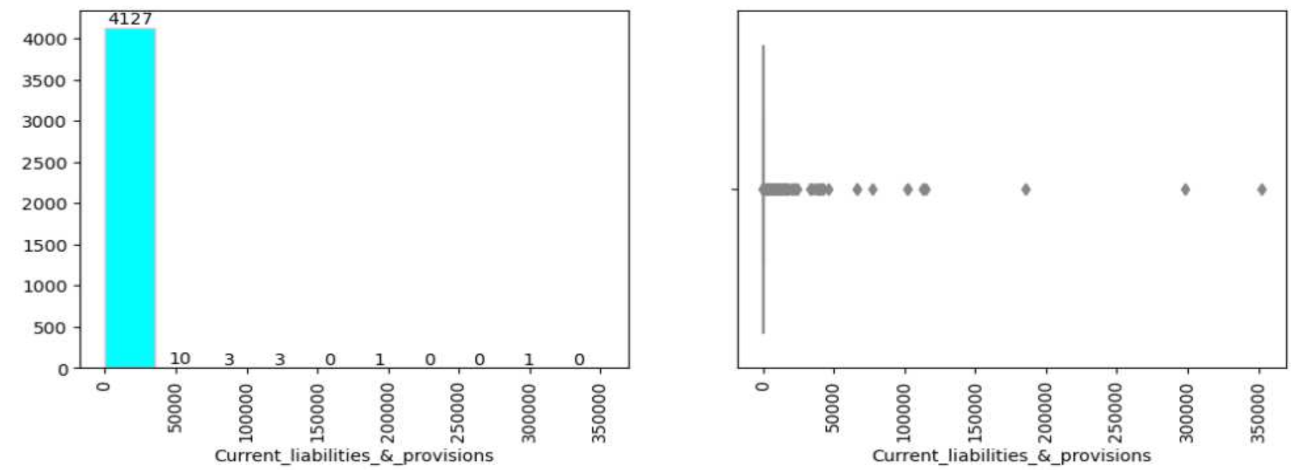
```
Skewness of Reserves_and_funds:  34.10896619433152
Distribution of Reserves_and_funds
----------------------------------------------------------------------------
```



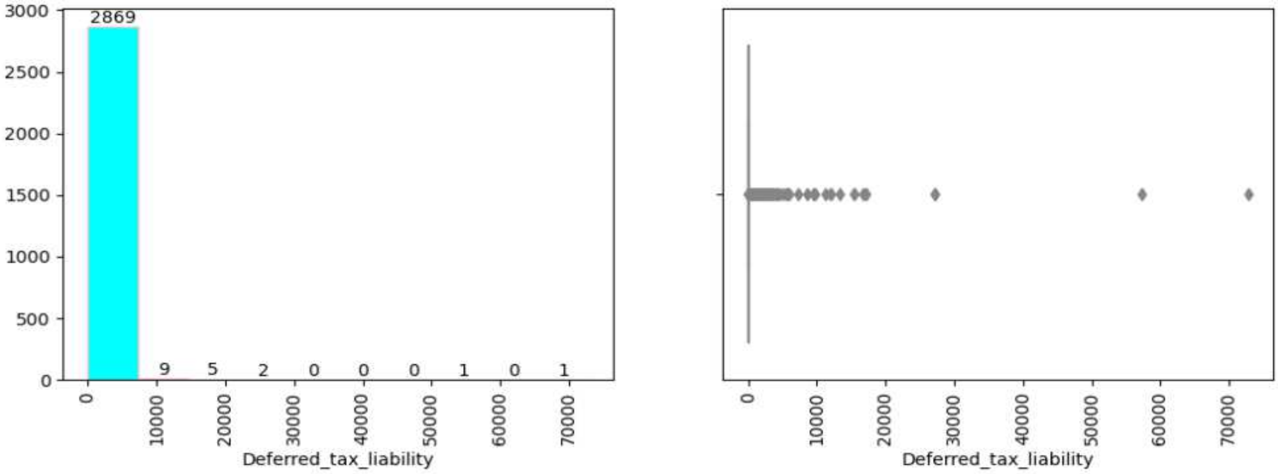

```
Skewness of Borrowings:  20.89130094122057
Distribution of Borrowings
----------------------------------------------------------------------------
```



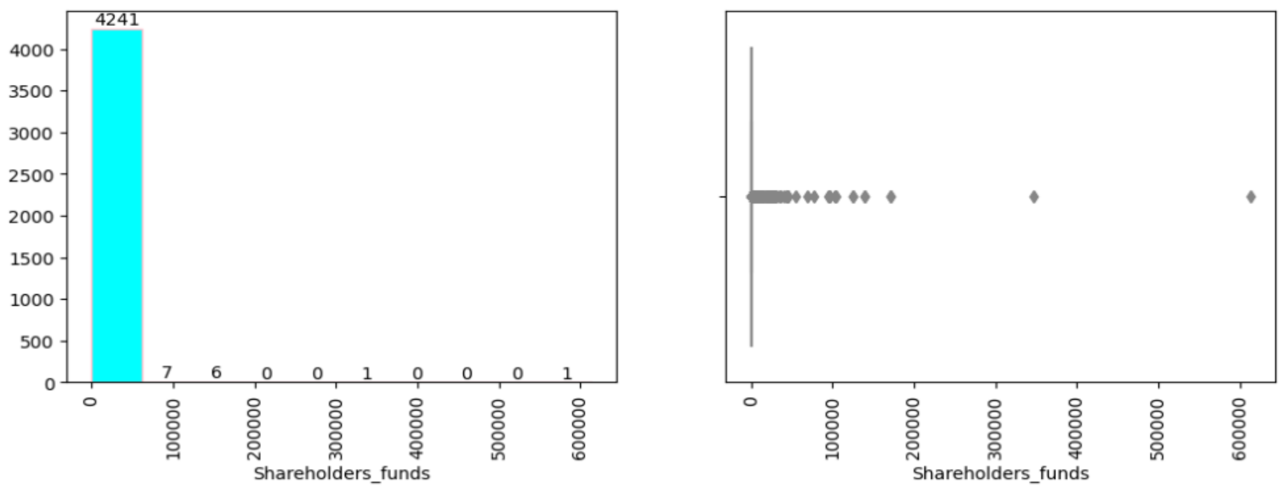

```
Skewness of Current_liabilities_&_provisions:  26.506919789566954
Distribution of Current_liabilities_&_provisions
----------------------------------------------------------------------------
```



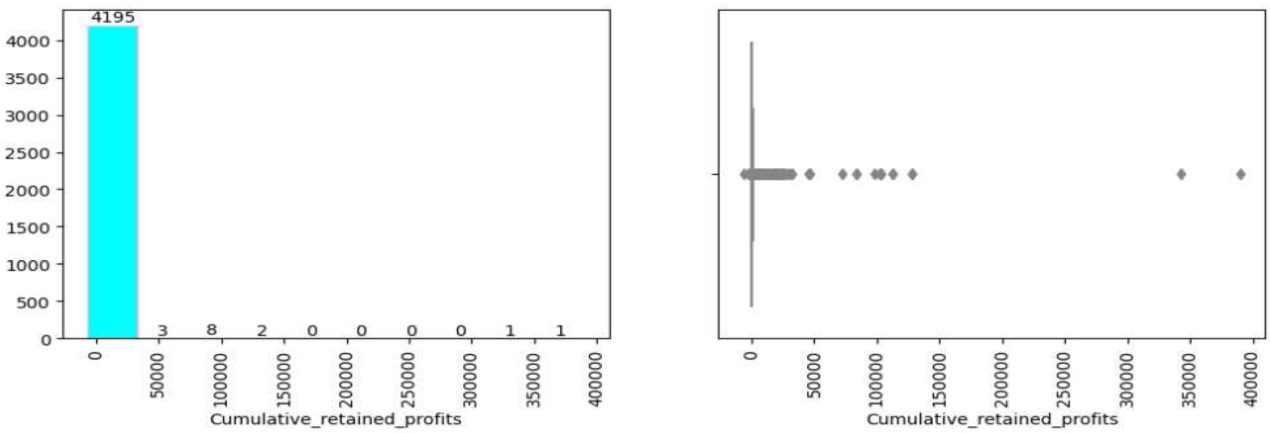
```
Skewness of Deferred_tax_liability:  23.73930173510226
Distribution of Deferred_tax_liability
----------------------------------------------------------------------------
```



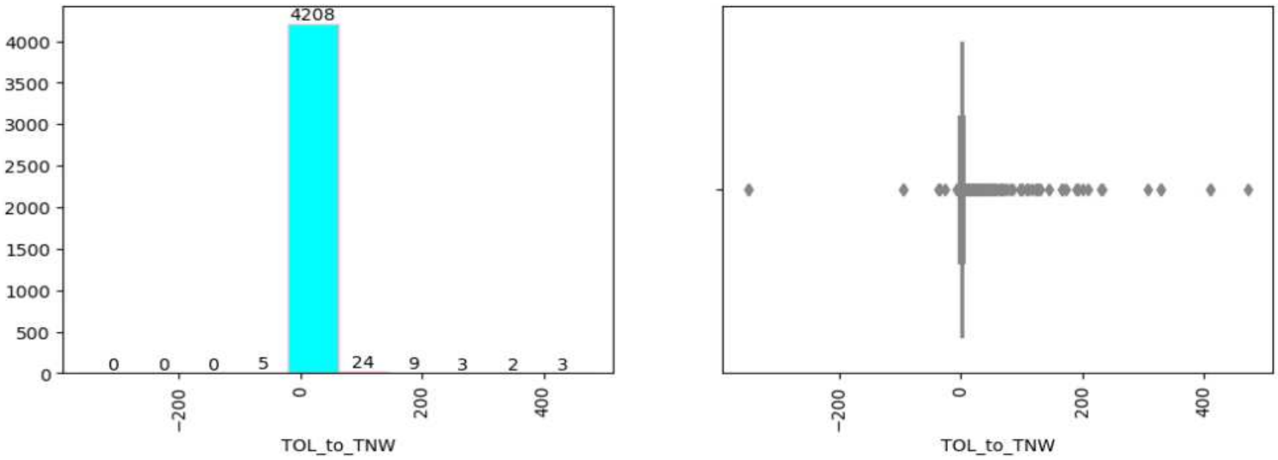

```
Skewness of Shareholders_funds:  31.549033473390544
Distribution of Shareholders_funds
----------------------------------------------------------------------------
```



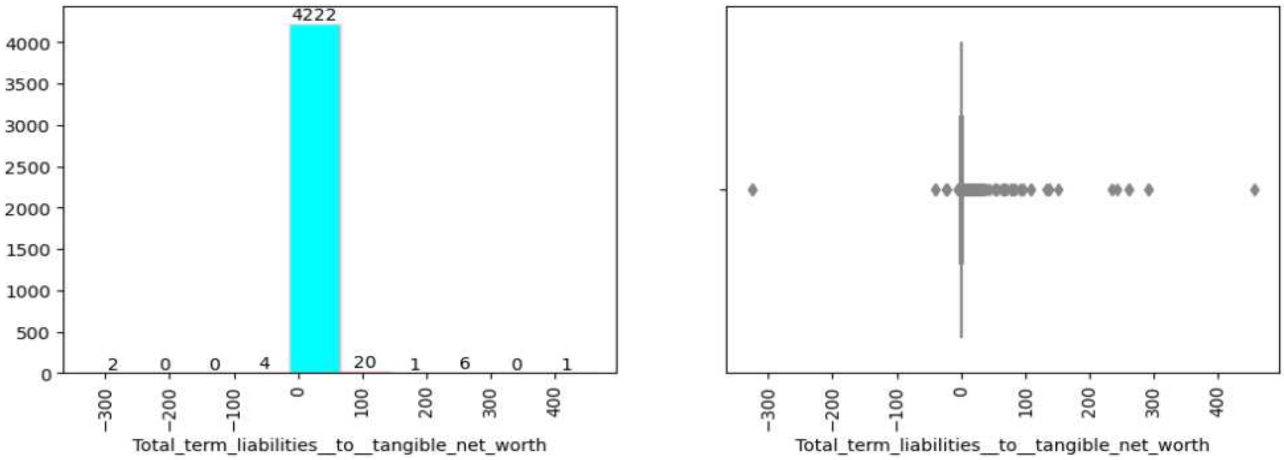

```
Skewness of Cumulative_retained_profits:  27.82460089549344
Distribution of Cumulative_retained_profits
----------------------------------------------------------------------------
```



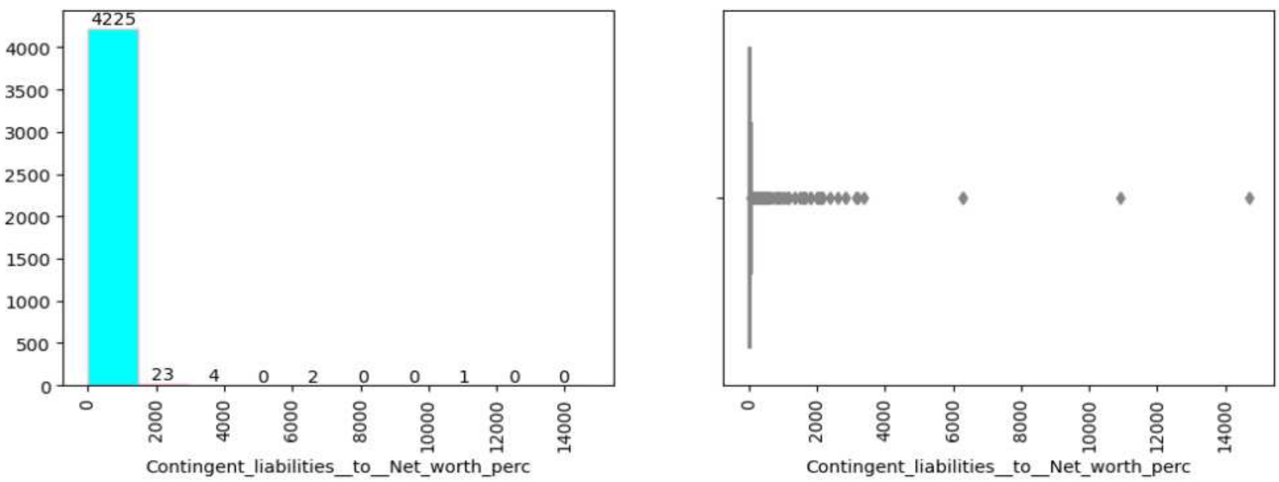
```
Skewness of TOL_to_TNW:  8.893421434492717
Distribution of TOL_to_TNW
----------------------------------------------------------------------------
```



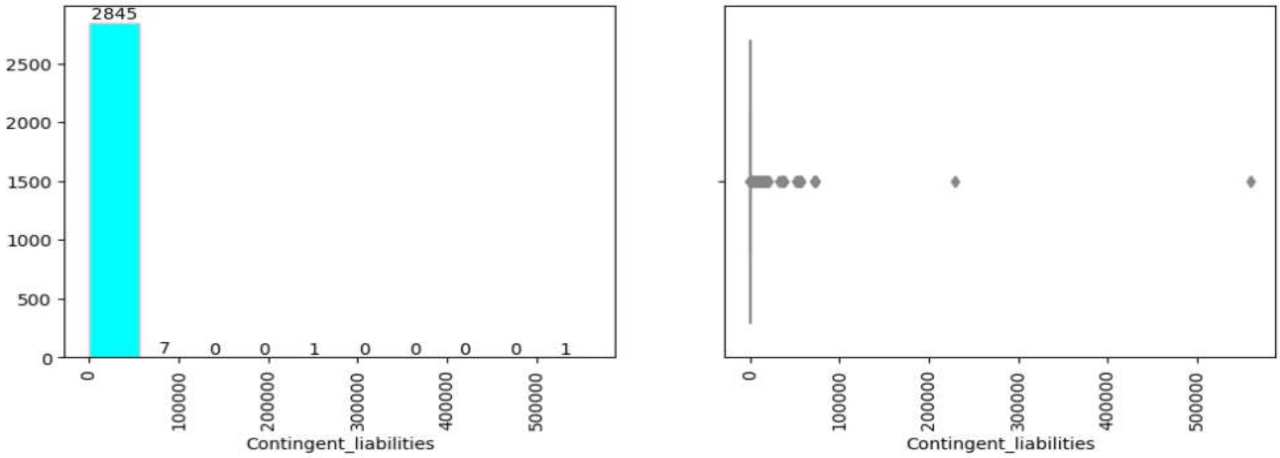

```
Skewness of Total_term_liabilities__to__tangible_net_worth:  9.033640135164498
Distribution of Total_term_liabilities__to__tangible_net_worth
----------------------------------------------------------------------------
```



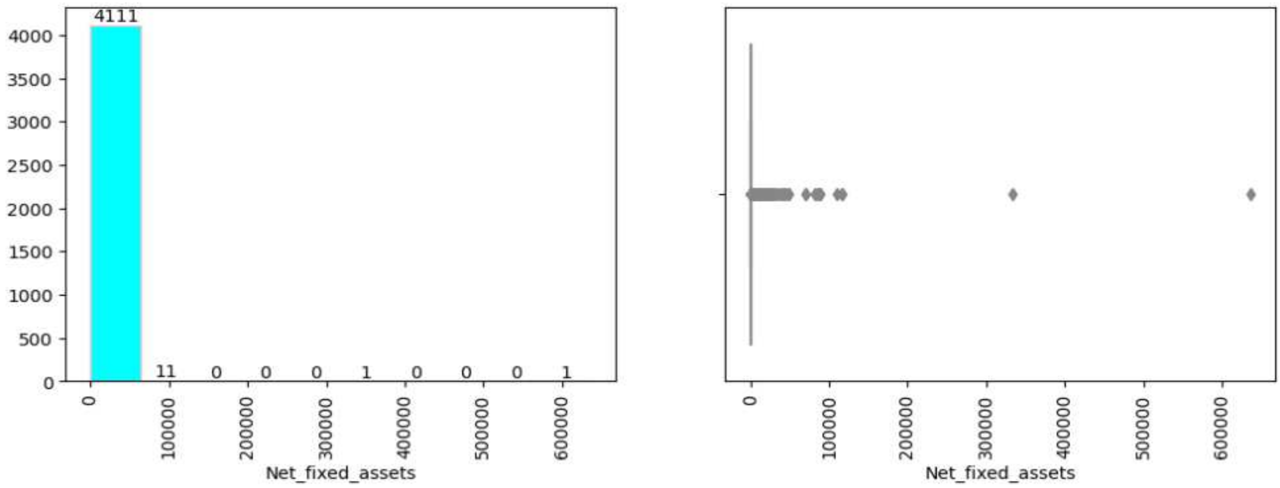

```
Skewness of Contingent_liabilities__to__Net_worth_perc:  24.542579962375754
Distribution of Contingent_liabilities__to__Net_worth_perc
----------------------------------------------------------------------------
```



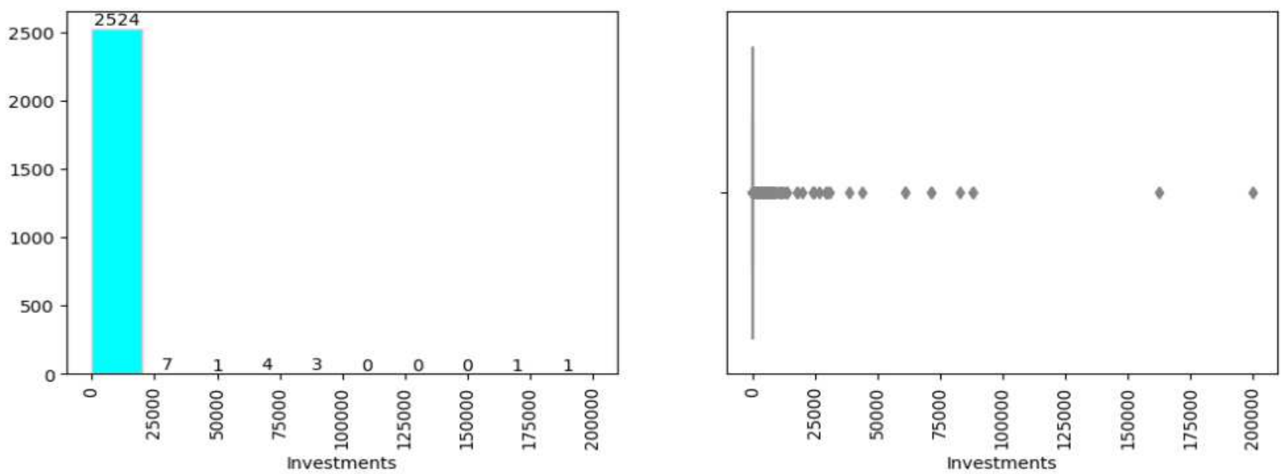
```
Skewness of Contingent_liabilities:  37.76261464043822
Distribution of Contingent_liabilities
----------------------------------------------------------------------------
```



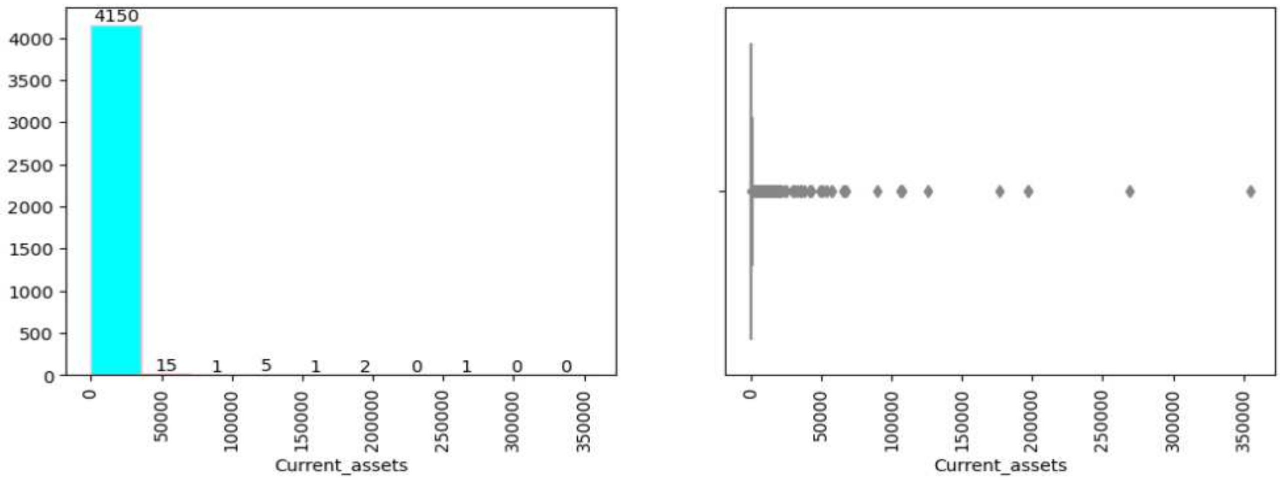

```
Skewness of Net_fixed_assets:  37.623726678314426
Distribution of Net_fixed_assets
----------------------------------------------------------------------------
```



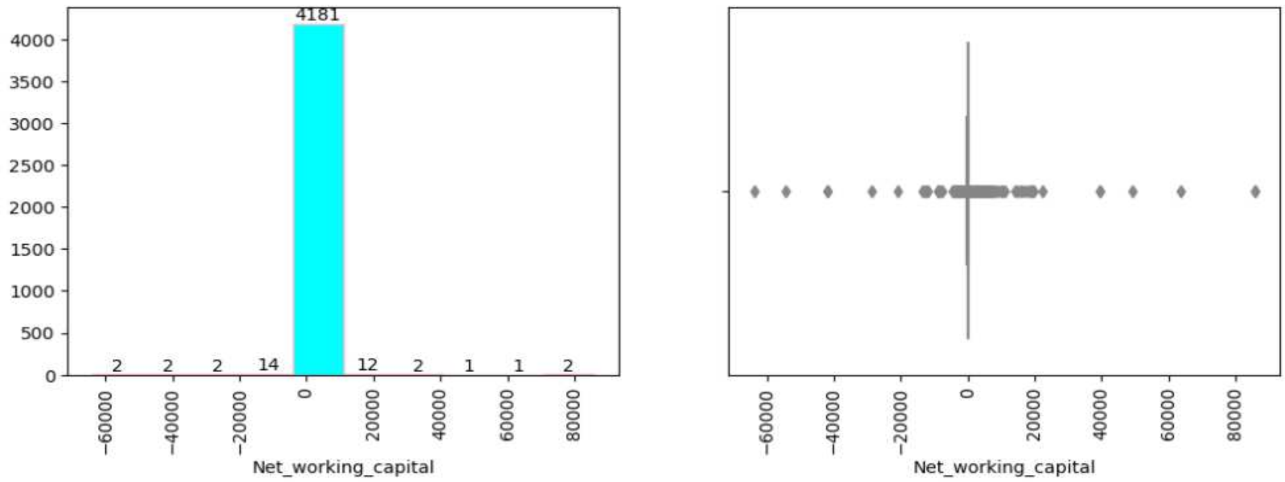

```
Skewness of Investments:  19.44284742648704
Distribution of Investments
----------------------------------------------------------------------------
```



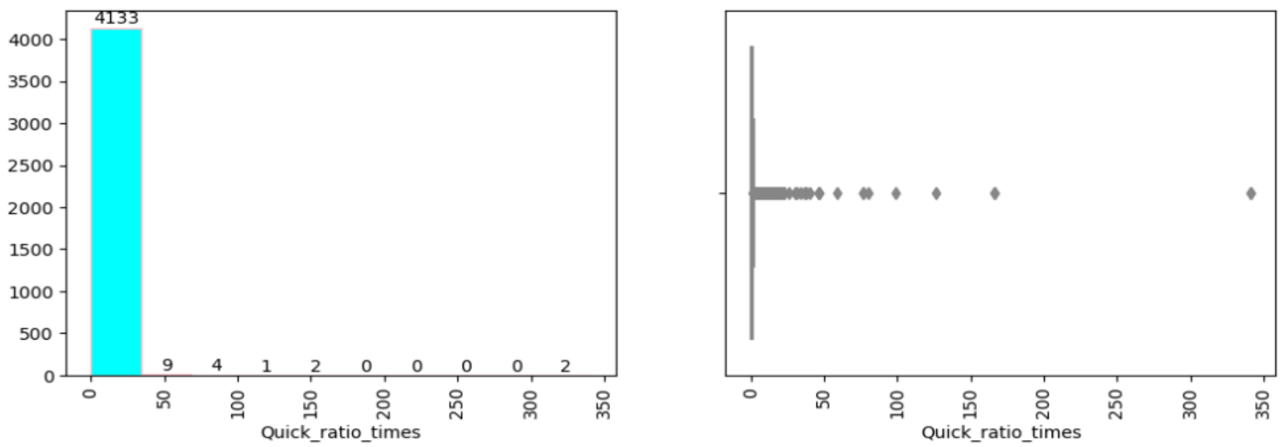
```
Skewness of Current_assets:  21.325078906073383
Distribution of Current_assets
----------------------------------------------------------------------------
```



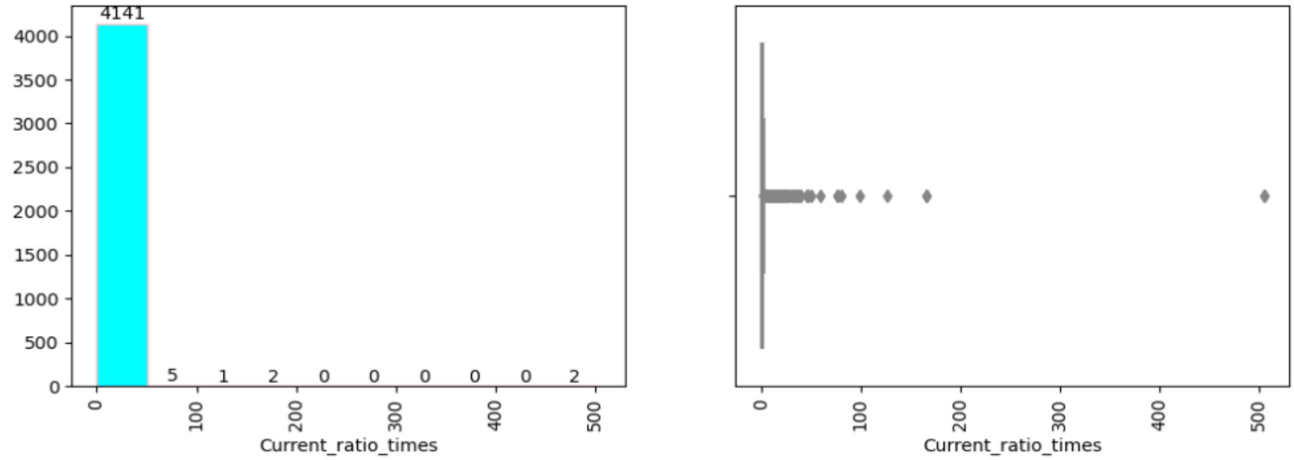
```
Skewness of Net_working_capital:  8.83680862778684
Distribution of Net_working_capital
----------------------------------------------------------------------------
```



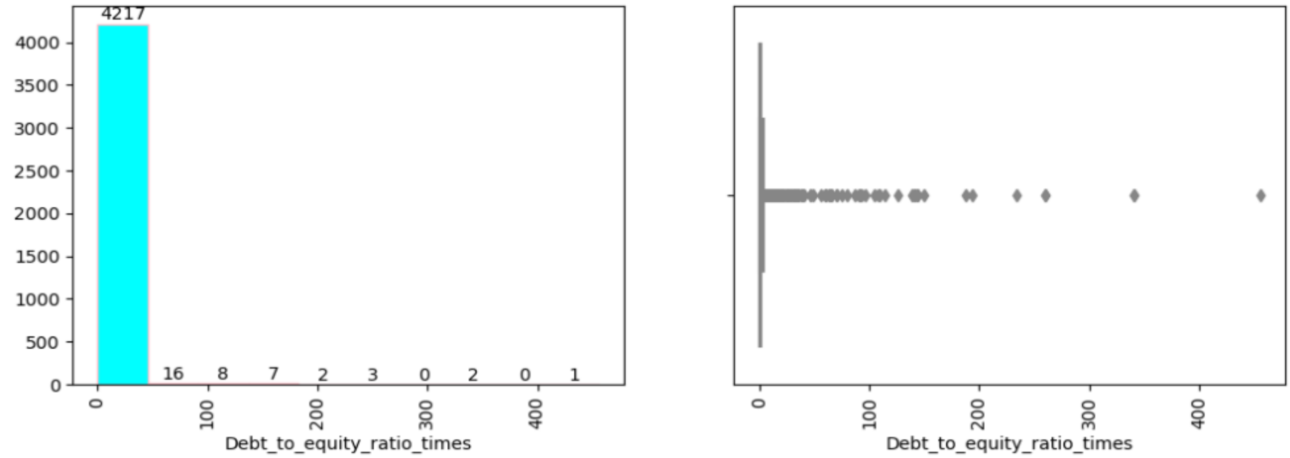
```
Skewness of Quick_ratio_times:  27.43150509863591
Distribution of Quick_ratio_times
----------------------------------------------------------------------------
```



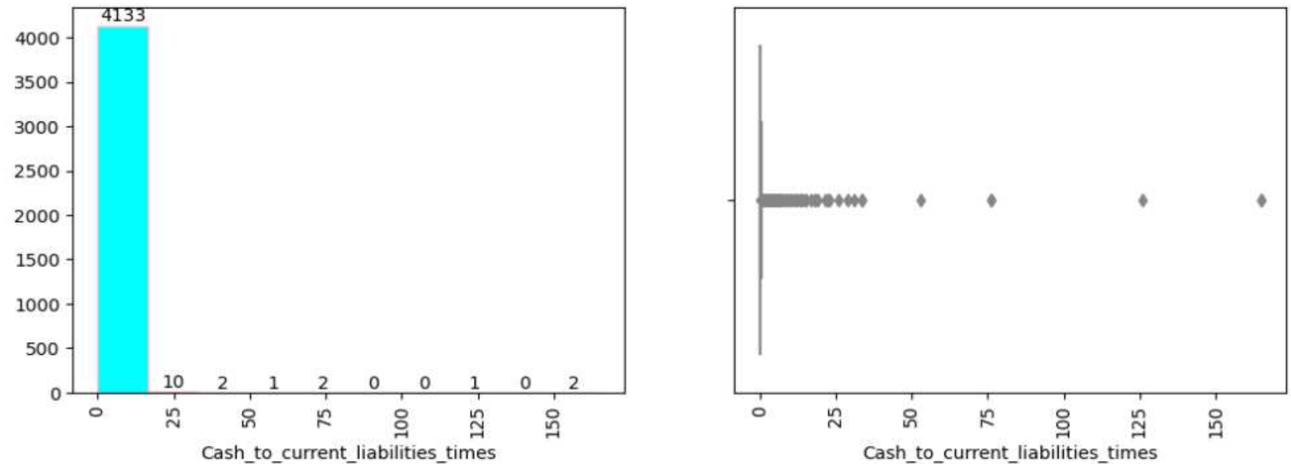
```
Skewness of Current_ratio_times:  33.284367631977865
Distribution of Current_ratio_times
----------------------------------------------------------------------------
```



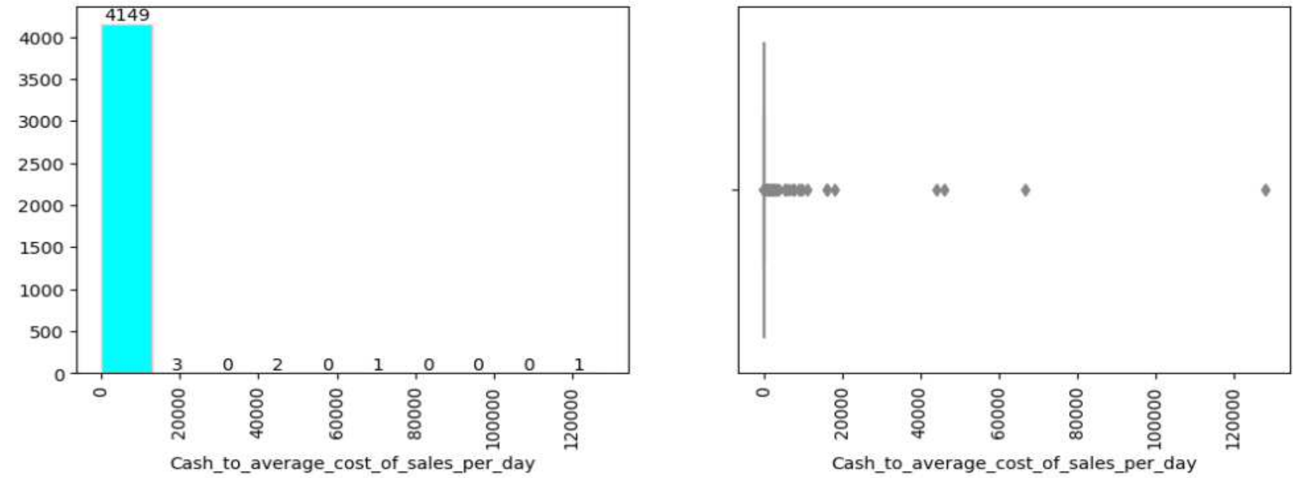

```
Skewness of Debt_to_equity_ratio_times:  16.33081181955665
Distribution of Debt_to_equity_ratio_times
----------------------------------------------------------------------------
```



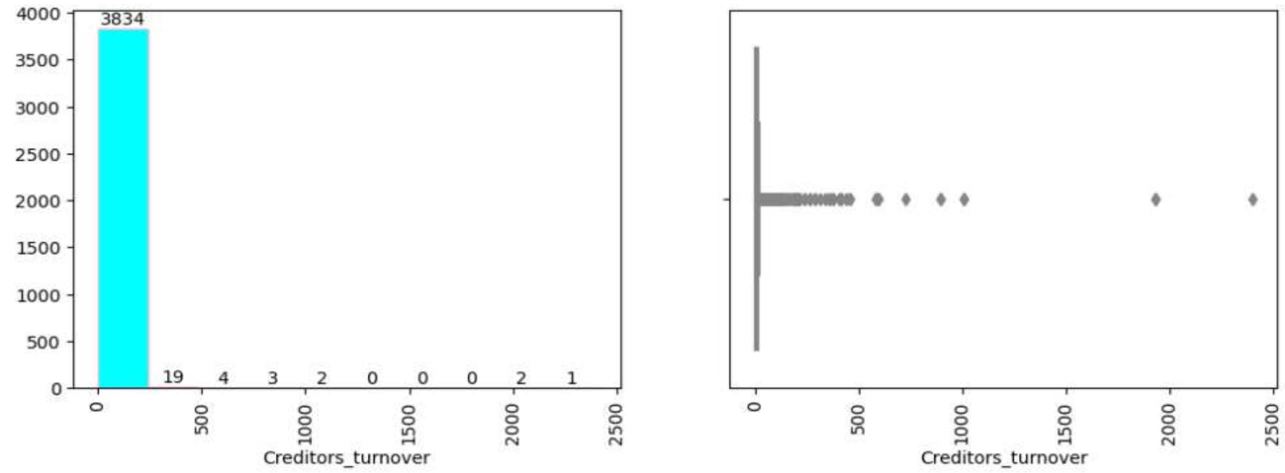

```
Skewness of Cash_to_current_liabilities_times:  26.45695782397687
Distribution of Cash_to_current_liabilities_times
----------------------------------------------------------------------------
```



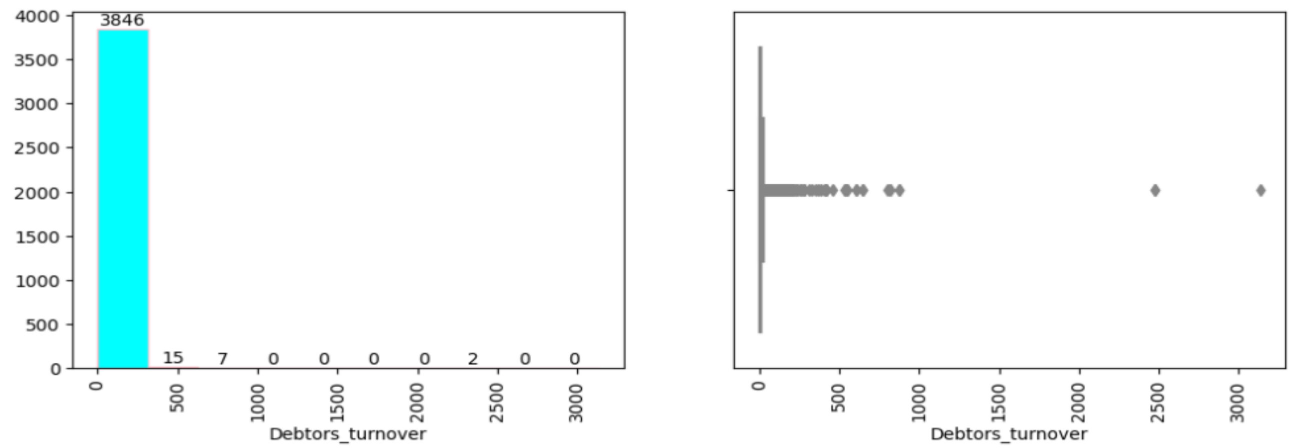
```
Skewness of Cash_to_average_cost_of_sales_per_day:  38.84093937509801
Distribution of Cash_to_average_cost_of_sales_per_day
----------------------------------------------------------------------------
```



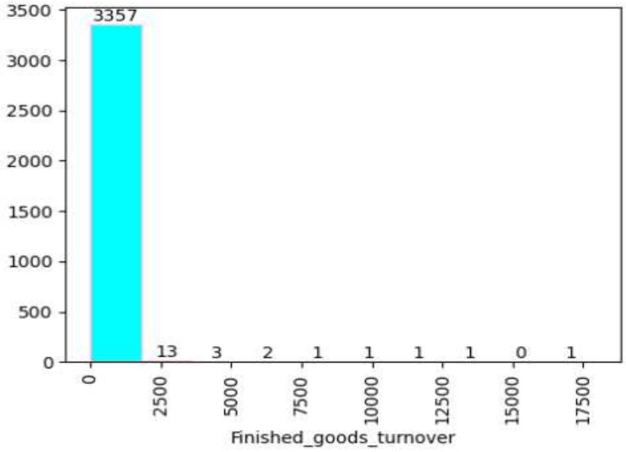

```
Skewness of Creditors_turnover:  19.719290987425236
Distribution of Creditors_turnover
----------------------------------------------------------------------------
```

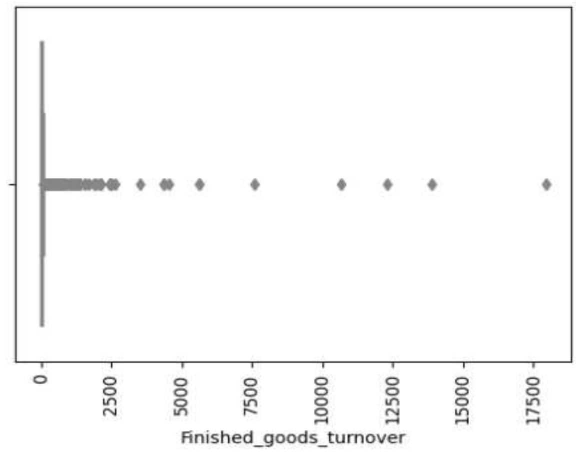



```
Skewness of Debtors_turnover:  22.907661706656093
Distribution of Debtors_turnover
----------------------------------------------------------------------------
```

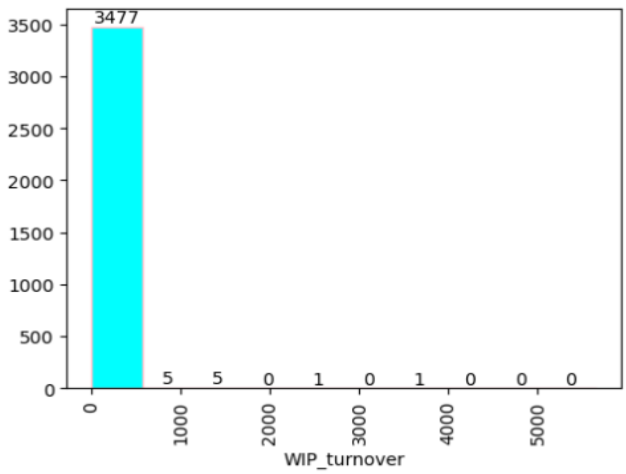


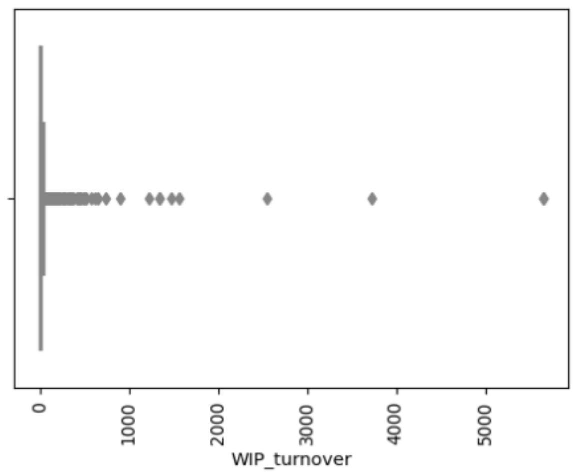
```
Skewness of Finished_goods_turnover:  20.84466000026286
Distribution of Finished_goods_turnover
----------------------------------------------------------------------------
```



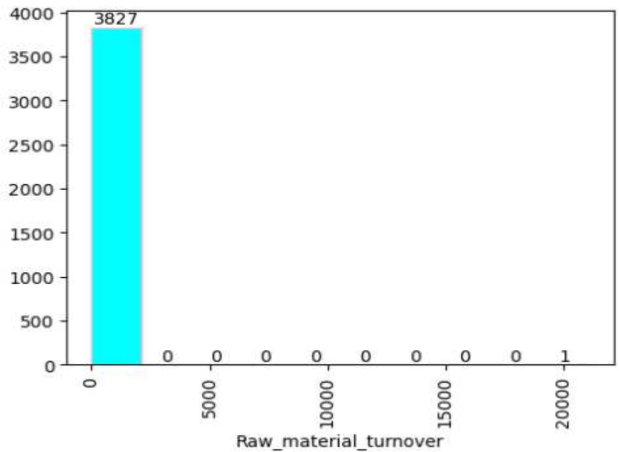


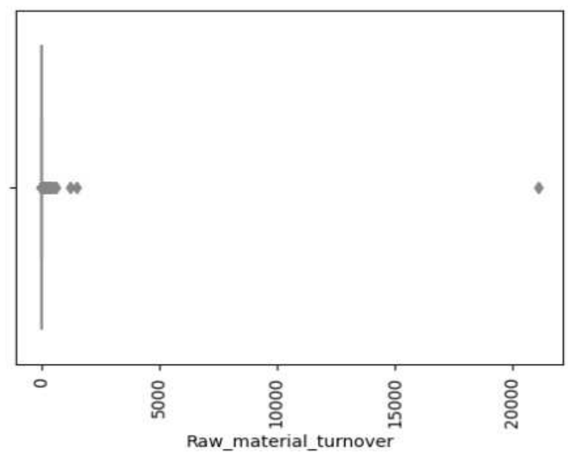

```
Skewness of WIP_turnover:  25.686670200282673
Distribution of WIP_turnover
----------------------------------------------------------------------------
```



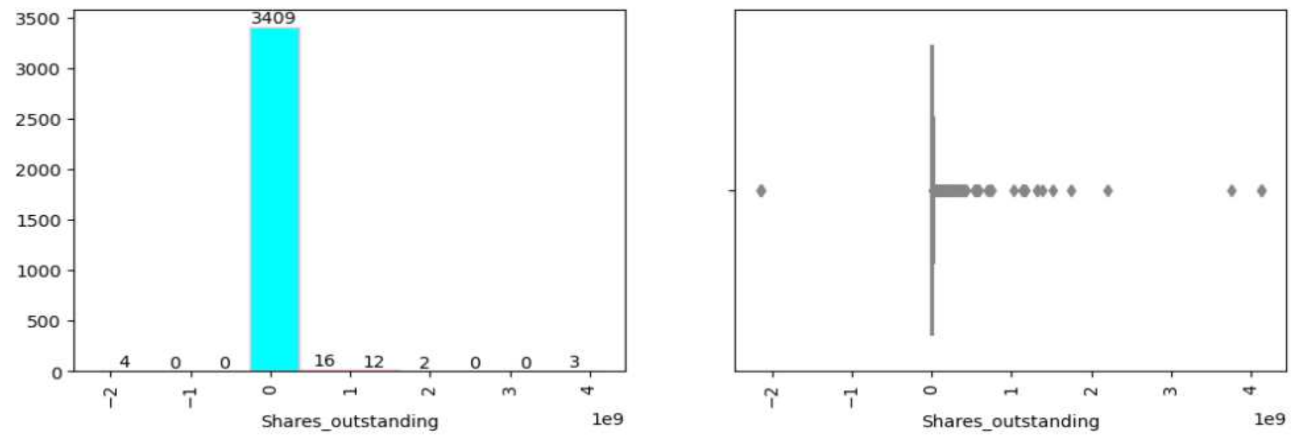



```
Skewness of Raw_material_turnover:  60.60776081295366
Distribution of Raw_material_turnover
----------------------------------------------------------------------------
```

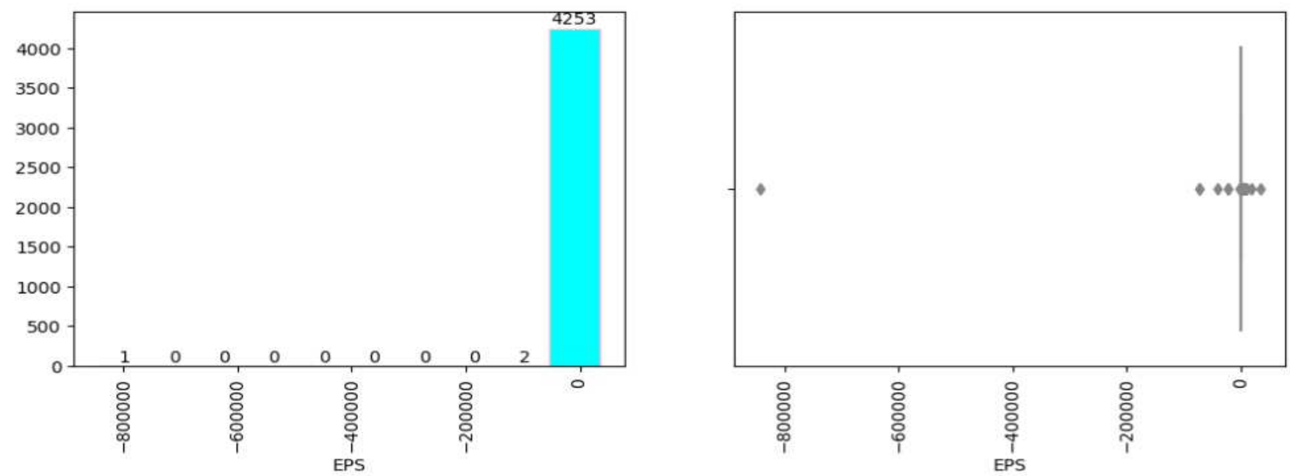




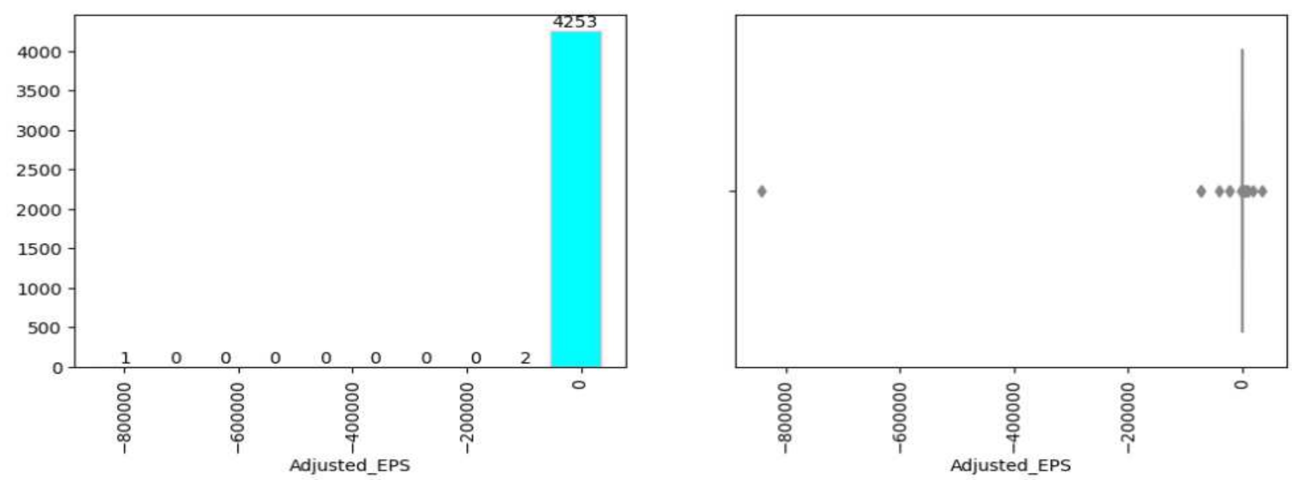
```
Skewness of Shares_outstanding:  11.034062150689422
Distribution of Shares_outstanding
----------------------------------------------------------------------------
```



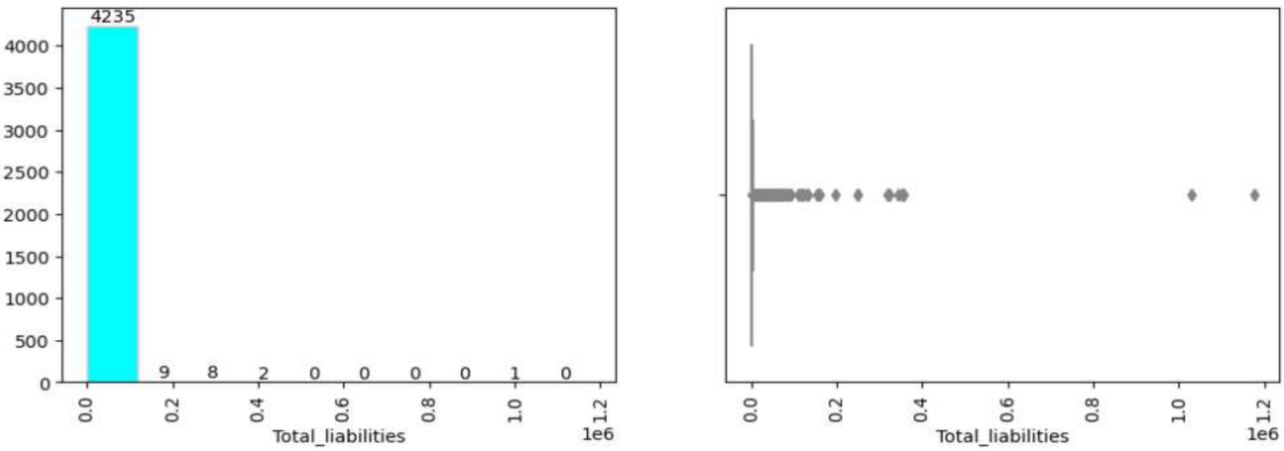

```
Skewness of EPS:  -63.28748213566746
Distribution of EPS
----------------------------------------------------------------------------
```



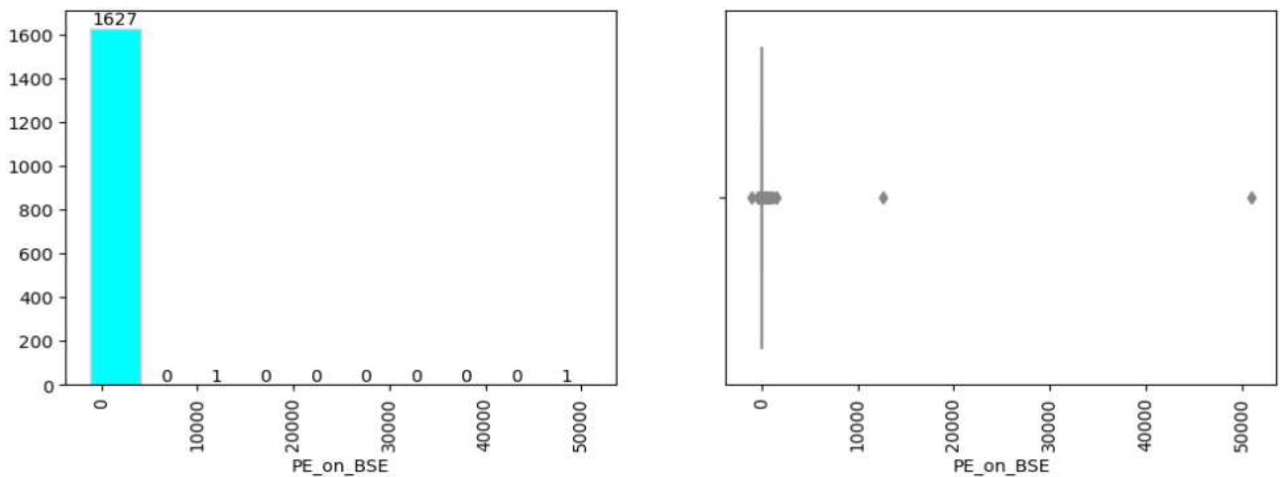

```
Skewness of Adjusted_EPS:  -63.28752879020988
Distribution of Adjusted_EPS
----------------------------------------------------------------------------
```



Skewness of Total_liabilities: 26.422680474857692
Distribution of Total_liabilities



Skewness of PE_on_BSE: 37.1968344949466
Distribution of PE_on_BSE



Skewness of default: 1.4067868482705692
Distribution of default

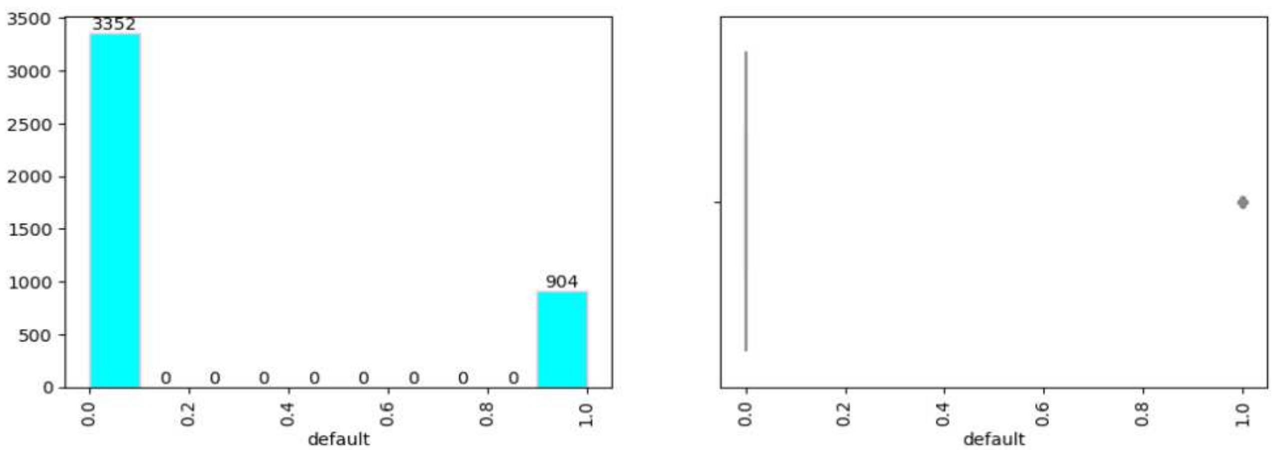


Figure 2: Univariate Analysis numeric columns

1.6.2 Bivariate Analysis

Relation between numeric columns

Figure 6: Pair plot

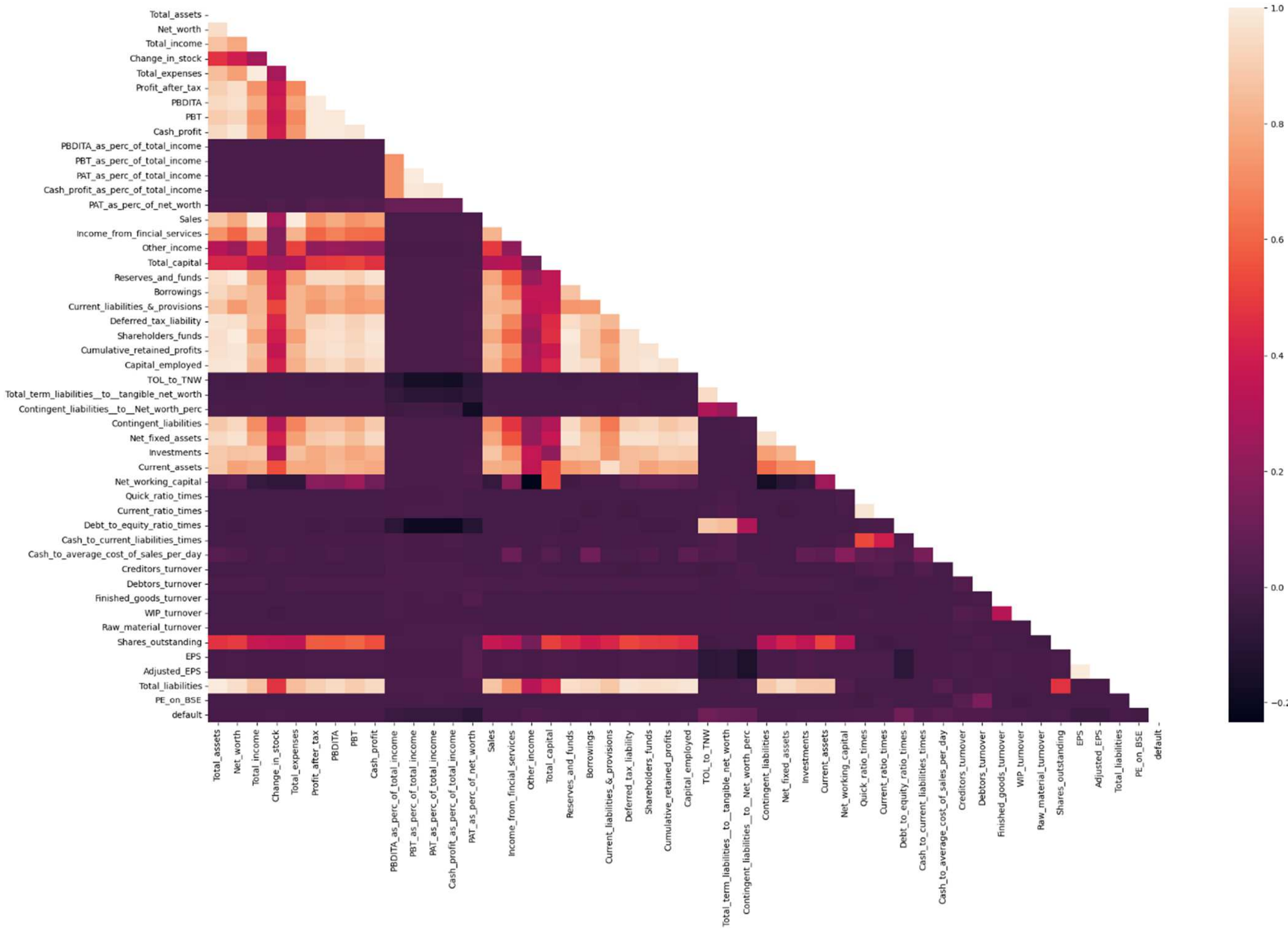


Figure 3: Heatmap

Key Observations

1. In the univariate analysis, plotting each attribute revealed that most of the data is concentrated within a narrow range, with a substantial number of extreme values falling outside this range making data heavily skewed.
2. The heatmap reveals a high correlation between multiple pairs of attributes, likely due to their interdependence or derivation from one another. To address this issue, we will employ the Variance Inflation Factor (VIF) from the statsmodels library to identify and drop attributes with high levels of multicollinearity.
3. The response variable does not show any significant correlation with any variable.

1.7 Outlier Treatment

From the univariate analysis we can clearly conclude that there are outliers in all the columns. We will check number of outliers by each column.

```
Total_assets                                   585
Net_worth                                      595
Total_income                                   508
Change_in_stock                                750
Total_expenses                                 518
Profit_after_tax                               712
PBDITA                                         584
PBT                                            704
Cash_profit                                    627
PBDITA_as_perc_of_total_income                 346
PBT_as_perc_of_total_income                    546
PAT_as_perc_of_total_income                    610
Cash_profit_as_perc_of_total_income            426
PAT_as_perc_of_net_worth                       427
Sales                                          500
Income_from_fincial_services                   517
Other_income                                   389
Total_capital                                  551
Reserves_and_funds                             643
Borrowings                                     532
Current_liabilities_&_provisions               581
Deferred_tax_liability                         406
Shareholders_funds                             588
Cumulative_retained_profits                    699
Capital_employed                               572
TOL_to_TNW                                     414
Total_term_liabilities__to__tangible_net_worth 406
Contingent_liabilities__to__Net_worth_perc     478
Contingent_liabilities                         393
Net_fixed_assets                               569
Investments                                    451
Current_assets                                 532
Net_working_capital                            806
Quick_ratio_times                              371
Current_ratio_times                            397
Debt_to_equity_ratio_times                     381
Cash_to_current_liabilities_times              539
Cash_to_average_cost_of_sales_per_day          583
Creditors_turnover                             442
Debtors_turnover                               408
Finished_goods_turnover                        399
WIP_turnover                                   378
Raw_material_turnover                          296
Shares_outstanding                             476
EPS                                            638
```

Adjusted_EPS	694
Total_liabilities	585
PE_on_BSE	237

```
dtype: int64
Outliers as a proportion of total data
12.13 %
```

Table 11: Outlier count

If we take the standard approach where we consider outliers to above 1.5 times the IQR over Q3 value or 1.5 times the IQR below Q1 value then we will have over 12% of the data as outlier adding to this the missing values which account over 8% of the data, we will have over 20% of the data as made-up data. Rather than using IQR and Q1, Q3 we will use 5 and 95 percentile as cutoff and check number of outliers based on this.

```
missing values based 5 and 95 percentile as cutoff
```

Total_assets	424
Net_worth	421
Total_income	404
Change_in_stock	371
Total_expenses	410
Profit_after_tax	412
PBDITA	407
PBT	412
Cash_profit	411
PBDITA_as_perc_of_total_income	418
PBT_as_perc_of_total_income	418
PAT_as_perc_of_total_income	418
Cash_profit_as_perc_of_total_income	416
PAT_as_perc_of_net_worth	426
Sales	396
Income_from_fincial_services	159
Other_income	138
Total_capital	420
Reserves_and_funds	416
Borrowings	377
Current_liabilities_&_provisions	411
Deferred_tax_liability	269
Shareholders_funds	421
Cumulative_retained_profits	422
Capital_employed	422
TOL_to_TNW	404
Total_term_liabilities__to__tangible_net_worth	232
Contingent_liabilities__to__Net_worth_perc	213
Contingent_liabilities	267
Net_fixed_assets	412
Investments	147

Current_assets	417
Net_working_capital	421
Quick_ratio_times	411
Current_ratio_times	413
Debt_to_equity_ratio_times	213
Cash_to_current_liabilities_times	205
Cash_to_average_cost_of_sales_per_day	415
Creditors_turnover	193
Debtors_turnover	194
Finished_goods_turnover	339
WIP_turnover	347
Raw_material_turnover	195
Shares_outstanding	346
EPS	423
Adjusted_EPS	425
Total_liabilities	424
PE_on_BSE	164

dtype: int64

Outliers as a proportion of total data
8.24 %

Table 12: Outlier count

On taking upper limit at 95 percentile and lower limit at 5 percentile we have bought the proportion of outliers to 8% from 12% thus we will be considering these value as upper limit and lower limit. Rather than assigning the upper limit and lower limit values to the outliers we will change them to null values and then treat them like missing values using KNN imputer on them also.

1.8 Missing Value Treatment

Checking for missing values by columns

Column vice null data

Total_assets	424
Net_worth	421
Total_income	635
Change_in_stock	921
Total_expenses	575
Profit_after_tax	566
PBDITA	561
PBT	566
Cash_profit	565
PBDITA_as_perc_of_total_income	497
PBT_as_perc_of_total_income	497
PAT_as_perc_of_total_income	497
Cash_profit_as_perc_of_total_income	495
PAT_as_perc_of_net_worth	426
Sales	701
Income_from_fincial_services	1270
Other_income	1694

Total_capital	425
Reserves_and_funds	514
Borrowings	808
Current_liabilities_&_provisions	521
Deferred_tax_liability	1638
Shareholders_funds	421
Cumulative_retained_profits	467
Capital_employed	422
TOL_to_TNW	404
Total_term_liabilities__to__tangible_net_worth	232
Contingent_liabilities__to__Net_worth_perc	213
Contingent_liabilities	1669
Net_fixed_assets	544
Investments	1862
Current_assets	497
Net_working_capital	458
Quick_ratio_times	516
Current_ratio_times	518
Debt_to_equity_ratio_times	213
Cash_to_current_liabilities_times	310
Cash_to_average_cost_of_sales_per_day	515
Creditors_turnover	584
Debtors_turnover	579
Finished_goods_turnover	1213
WIP_turnover	1111
Raw_material_turnover	623
Shares_outstanding	1156
EPS	423
Adjusted_EPS	425
Total_liabilities	424
PE_on_BSE	2791

dtype: int64

Total number of null values: 33807

Null values as a proportion of total data 16.55 %

Table 13: Missing values by columns

After converting outliers to null values total missing values account for 16.55% of the data, we will check missing values by columns using heatmap.

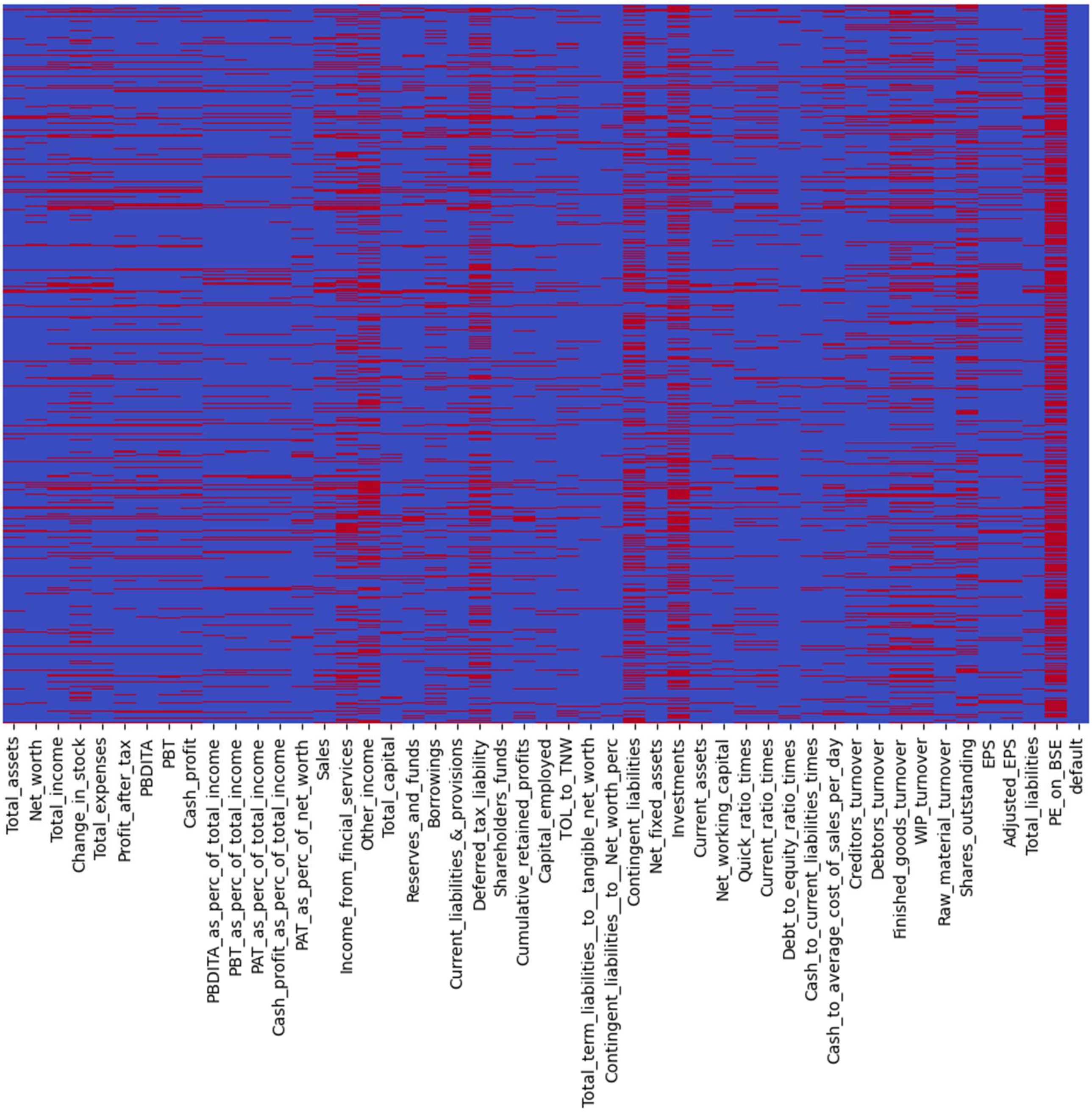


Figure 4: Heatmap

For some columns like PE_on_BSE, Investments etc. there are a lot reds in the heatmap depicting missing data meaning we have large missing data for these columns.

Checking for missing values by row

```
0        3
1        8
2        3
3        8
4        6
        ..
4251    32
4252     4
4253     2
4254     5
4255     2
Length: 4256, dtype: int64
```

Table 14: Missing values by rows

On checking for missing values by rows we can see that for some rows over 60% of the data is not present which is not an ideal condition as we have to make up over 60% information for these rows.

We will filter out data with over 10% missing values and check how much data is present with over 90% values.

```
data which is 90% or more complete at the row level

(2285, 49)
```

Approximately half of the rows in the dataset have more than 10% missing values. To address this, we can filter out these rows and build the model using the remaining data. Additionally, it is crucial to determine whether the missing information is genuine or if it indicates an attempt by companies to conceal data. To investigate this, we will analyze the proportion of defaulters in the filtered dataset, which includes companies with over 90% of their data available.

```
defaults for filtered data

default
0    0.83
1    0.17
Name: proportion, dtype: float64

default for original data

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

Table 15: Comparison of defaulters

Companies with over 90% of their data available have a default rate of 17%, compared to 21% for the entire dataset. This indicates that companies with more than 10% of their data missing tend to have a higher likelihood of defaulting. This observation highlights the potential relationship between missing data and financial instability, warranting further investigation.

Treating Missing Values

Since, we have significant missing data for some columns we will check the missing data column wise in proportion terms sorted in descending order of missing values.

```
PE_on_BSE                              0.66
Investments                            0.44
Other_income                           0.40
Contingent_liabilities                 0.39
Deferred_tax_liability                 0.38
Income_from_fincial_services           0.30
Finished_goods_turnover                0.29
Shares_outstanding                     0.27
WIP_turnover                           0.26
Change_in_stock                        0.22
Borrowings                             0.19
Sales                                  0.16
Total_income                           0.15
Raw_material_turnover                  0.15
Creditors_turnover                     0.14
Debtors_turnover                       0.14
Total_expenses                         0.14
PBT                                    0.13
Profit_after_tax                       0.13
Cash_profit                            0.13
PBDITA                                 0.13
Net_fixed_assets                       0.13
Current_liabilities_&_provisions       0.12
Current_ratio_times                    0.12
```

```
Quick_ratio_times                              0.12
Cash_to_average_cost_of_sales_per_day          0.12
Reserves_and_funds                             0.12
Current_assets                                 0.12
PBDITA_as_perc_of_total_income                 0.12
PBT_as_perc_of_total_income                    0.12
PAT_as_perc_of_total_income                    0.12
Cash_profit_as_perc_of_total_income            0.12
Cumulative_retained_profits                    0.11
Net_working_capital                            0.11
PAT_as_perc_of_net_worth                       0.10
Total_capital                                  0.10
Adjusted_EPS                                   0.10
Total_liabilities                              0.10
Total_assets                                   0.10
EPS                                            0.10
Capital_employed                               0.10
Net_worth                                      0.10
Shareholders_funds                             0.10
TOL_to_TNW                                     0.09
Cash_to_current_liabilities_times              0.07
Total_term_liabilities__to__tangible_net_worth 0.05
Debt_to_equity_ratio_times                     0.05
Contingent_liabilities__to__Net_worth_perc     0.05
default                                        0.00
dtype: float64
```

Table 16: Proportion of missing values

On checking missing values by columns there are some columns with over 30% missing values, we dropped all those columns and for the remaining data will impute values using KNN imputation for which we have to first scale the data.

Data Scaling

For scaling we used standard scaler which ensures that data for all columns have an mean of 0 and standard deviation of 1.

statistical summary of scaled data

	count	mean	std	min	25%	50%	75%	max
Total_assets	3832.00	-0.00	1.00	-0.61	-0.55	-0.41	0.03	5.49
Net_worth	3835.00	0.00	1.00	-0.59	-0.53	-0.40	0.03	5.28
Total_income	3621.00	-0.00	1.00	-0.67	-0.59	-0.40	0.12	4.92
Change_in_stock	3335.00	-0.00	1.00	-1.85	-0.42	-0.33	0.11	5.30
Total_expenses	3681.00	-0.00	1.00	-0.67	-0.60	-0.40	0.11	4.98
Profit_after_tax	3690.00	-0.00	1.00	-0.64	-0.49	-0.40	-0.05	5.58
PBDITA	3695.00	-0.00	1.00	-0.58	-0.54	-0.41	0.01	5.23
PBT	3690.00	-0.00	1.00	-0.62	-0.49	-0.41	-0.04	5.79
Cash_profit	3691.00	-0.00	1.00	-0.59	-0.53	-0.42	0.00	5.26
PBDITA_as_perc_of_total_income	3759.00	-0.00	1.00	-1.64	-0.75	-0.17	0.59	3.14
PBT_as_perc_of_total_income	3759.00	0.00	1.00	-4.32	-0.55	-0.17	0.55	2.87
PAT_as_perc_of_total_income	3759.00	0.00	1.00	-4.94	-0.48	-0.14	0.51	2.77
Cash_profit_as_perc_of_total_income	3761.00	0.00	1.00	-2.98	-0.71	-0.15	0.58	3.01
PAT_as_perc_of_net_worth	3830.00	-0.00	1.00	-3.18	-0.78	-0.19	0.59	2.93
Sales	3555.00	-0.00	1.00	-0.67	-0.59	-0.39	0.12	4.88
Income_from_fincial_services	2986.00	-0.00	1.00	-0.45	-0.44	-0.39	-0.12	6.08
Total_capital	3831.00	0.00	1.00	-0.72	-0.60	-0.35	0.11	5.15
Reserves_and_funds	3742.00	0.00	1.00	-0.62	-0.51	-0.41	-0.02	5.54
Borrowings	3448.00	0.00	1.00	-0.59	-0.54	-0.40	0.01	5.74
Current_liabilities_&_provisions	3735.00	-0.00	1.00	-0.60	-0.55	-0.41	0.04	5.26
Shareholders_funds	3835.00	-0.00	1.00	-0.59	-0.53	-0.41	0.02	5.16

Shareholders_funds	3835.00	-0.00	1.00	-0.59	-0.53	-0.41	0.02	5.16
Cumulative_retained_profits	3789.00	0.00	1.00	-0.71	-0.50	-0.40	-0.03	5.49
Capital_employed	3834.00	-0.00	1.00	-0.61	-0.55	-0.40	0.04	5.68
TOL_to_TNW	3852.00	0.00	1.00	-1.01	-0.68	-0.30	0.32	4.50
Total_term_liabilities__to__tangible_net_worth	4024.00	-0.00	1.00	-0.77	-0.72	-0.38	0.33	4.53
Contingent_liabilities__to__Net_worth_perc	4043.00	-0.00	1.00	-0.63	-0.63	-0.49	0.20	4.45
Net_fixed_assets	3712.00	0.00	1.00	-0.61	-0.55	-0.41	0.05	5.54
Current_assets	3759.00	-0.00	1.00	-0.65	-0.58	-0.41	0.09	5.07
Net_working_capital	3798.00	0.00	1.00	-1.76	-0.48	-0.34	0.11	5.09
Quick_ratio_times	3740.00	-0.00	1.00	-1.34	-0.67	-0.22	0.40	4.22
Current_ratio_times	3738.00	0.00	1.00	-1.51	-0.63	-0.24	0.33	4.07
Debt_to_equity_ratio_times	4043.00	0.00	1.00	-0.88	-0.73	-0.31	0.35	4.57
Cash_to_current_liabilities_times	3946.00	-0.00	1.00	-0.66	-0.57	-0.39	0.06	4.94
Cash_to_average_cost_of_sales_per_day	3741.00	0.00	1.00	-0.63	-0.53	-0.37	0.01	5.61
Creditors_turnover	3672.00	-0.00	1.00	-1.05	-0.60	-0.32	0.25	4.80
Debtors_turnover	3677.00	-0.00	1.00	-1.07	-0.60	-0.30	0.26	4.75
Finished_goods_turnover	3043.00	-0.00	1.00	-0.82	-0.62	-0.38	0.16	4.91
WIP_turnover	3145.00	0.00	1.00	-0.94	-0.66	-0.35	0.25	4.15
Raw_material_turnover	3633.00	-0.00	1.00	-1.13	-0.71	-0.25	0.40	3.71
Shares_outstanding	3100.00	0.00	1.00	-0.68	-0.57	-0.34	0.05	5.28
EPS	3833.00	-0.00	1.00	-0.81	-0.52	-0.42	0.03	5.48
Adjusted_EPS	3831.00	0.00	1.00	-0.82	-0.49	-0.40	-0.00	5.86
Total_liabilities	3832.00	-0.00	1.00	-0.61	-0.55	-0.41	0.03	5.49

Table 17: Statistical summary

Before applying the knn imputation we merged the independent and dependent variables.

```
Index(['Total_assets', 'Net_worth', 'Total_income', 'Change_in_stock',
       'Total_expenses', 'Profit_after_tax', 'PBDITA', 'PBT', 'Cash_profit',
       'PBDITA_as_perc_of_total_income', 'PBT_as_perc_of_total_income',
       'PAT_as_perc_of_total_income', 'Cash_profit_as_perc_of_total_income',
       'PAT_as_perc_of_net_worth', 'Sales', 'Income_from_fincial_services',
       'Total_capital', 'Reserves_and_funds', 'Borrowings',
       'Current_liabilities_&_provisions', 'Shareholders_funds',
       'Cumulative_retained_profits', 'Capital_employed', 'TOL_to_TNW',
       'Total_term_liabilities__to__tangible_net_worth',
       'Contingent_liabilities__to__Net_worth_perc', 'Net_fixed_assets',
       'Current_assets', 'Net_working_capital', 'Quick_ratio_times',
       'Current_ratio_times', 'Debt_to_equity_ratio_times',
       'Cash_to_current_liabilities_times',
       'Cash_to_average_cost_of_sales_per_day', 'Creditors_turnover',
       'Debtors_turnover', 'Finished_goods_turnover', 'WIP_turnover',
       'Raw_material_turnover', 'Shares_outstanding', 'EPS', 'Adjusted_EPS',
       'Total_liabilities', 'default'],
      dtype='object')
```

Table 18: Concatinated data columns

and split the data into train and test sets where for this problem we have taken train to test split ratio of 67:33.

```
Train data

(2851, 44)

Test data

(1405, 44)
```

Table 19: Train and test data shape

Applying KNN Imputation

We applied KNN imputation taking K value as 5 meaning the average value of 5 nearest neighbors will be imputed for missing value instances for train data and for test data we will fit the average values of 5 nearest neighbors from train set.

```
Missing values for train data
0
Missing values for test data
0
```

1.9 Segregating independent and dependent variables

Here data is divided into X_train, X_test and y_train and y_train where X contains all the independent attributes and Y has response variable.

```
Train set independent data

(2851, 43)
```

```
Train set dependent data

(2851,)
```

Table 20: Data Shape

```
Test set independent data

(1405, 43)
Test set dependent data

(1405,)
```

Table 21: Data Shape

1.10 Classification Modelling

We will build models using different classification techniques namely Logistic Regression and Random Forest and then we will try to improve the model performance by finding optimal threshold using ROC curve. We will compare different model performances using their Accuracy, Precision and Recall scores. The accuracy score measures the overall performance of the model on both training and test datasets, allowing us to assess its stability and potential bias. Precision and recall, on the other hand, are critical for evaluating the model's effectiveness in identifying positive cases while minimizing false positives and false negatives. These metrics collectively ensure a comprehensive assessment of the model's performance.

For evaluation of each model, we will additionally be using classification table and confusion matrix as a classification report provides a detailed summary of key metrics like precision, recall, F1 score, and support for each class, helping to evaluate the performance of a model comprehensively. A confusion matrix offers a visual and numerical breakdown of true positives, false positives, true negatives, and false negatives, allowing for an in-depth understanding of the model's accuracy and error types.

Logistic Regression Model

We will build the model using statsmodel library, however, before building the logistic regression model we will check for the Variance Inflation Factor score also called VIF score which quantifies how much the variance of a regression coefficient is inflated due to multicollinearity and since logistic regression technique is very sensitive towards multicollinearity it is important to the VIF scores for all the independent attributes and remove those attributes which have high VIF scores.

Checking VIF Scores

VIF scores in descending order

	Feature	VIF
43	Total_liabilities	inf
1	Total_assets	inf
3	Total_income	121.44
5	Total_expenses	92.20
21	Shareholders_funds	69.37
2	Net_worth	66.79
15	Sales	58.99
8	PBT	34.59
6	Profit_after_tax	32.36
9	Cash_profit	20.71
7	PBDITA	19.71
23	Capital_employed	17.74
11	PBT_as_perc_of_total_income	13.10
18	Reserves_and_funds	13.02
12	PAT_as_perc_of_total_income	11.70
28	Current_assets	9.76
22	Cumulative_retained_profits	8.53
41	EPS	7.44
42	Adjusted_EPS	6.72
20	Current_liabilities_&_provisions	6.71
32	Debt_to_equity_ratio_times	5.70

27	Net_fixed_assets	5.59
13	Cash_profit_as_perc_of_total_income	5.18
19	Borrowings	4.48
25	Total_term_liabilities__to__tangible_net_worth	4.15
10	PBDITA_as_perc_of_total_income	3.69
24	TOL_to_TNW	3.08
30	Quick_ratio_times	3.05
17	Total_capital	3.00
40	Shares_outstanding	2.96
31	Current_ratio_times	2.66
29	Net_working_capital	2.23
14	PAT_as_perc_of_net_worth	2.18
16	Income_from_fincial_services	2.16
33	Cash_to_current_liabilities_times	1.99
38	WIP_turnover	1.75
34	Cash_to_average_cost_of_sales_per_day	1.75
4	Change_in_stock	1.57
37	Finished_goods_turnover	1.56
35	Creditors_turnover	1.50
36	Debtors_turnover	1.49
39	Raw_material_turnover	1.39
26	Contingent_liabilities__to__Net_worth_perc	1.23
0	const	1.12

Table 22: VIF scores

There are multiple independent variables which have high VIF scores indicating strong correlation between independent variables and since logistic regression is very sensitive to correlation, we will drop those variables which have VIF score in excess of 10. For this we will drop one variable at a time and check the VIF score, repeating this process till VIF score for all the remaining variables is below 10.

Final VIF scores:

	Feature	VIF
0	const	1.09
1	Net_worth	7.60
2	Change_in_stock	1.54
3	Total_expenses	7.45
4	Profit_after_tax	5.01
5	PBDITA_as_perc_of_total_income	3.49
6	PAT_as_perc_of_total_income	3.09
7	Cash_profit_as_perc_of_total_income	4.95
8	PAT_as_perc_of_net_worth	2.13
9	Income_from_fincial_services	2.04
10	Total_capital	2.92
11	Borrowings	3.46
12	Current_liabilities_&_provisions	6.04
13	Cumulative_retained_profits	5.98
14	TOL_to_TNW	3.05
15	Total_term_liabilities__to__tangible_net_worth	4.09
16	Contingent_liabilities__to__Net_worth_perc	1.22
17	Net_fixed_assets	4.55
18	Current_assets	9.30
19	Net_working_capital	2.15
20	Quick_ratio_times	3.03
21	Current_ratio_times	2.65
22	Debt_to_equity_ratio_times	5.58
23	Cash_to_current_liabilities_times	1.97
24	Cash_to_average_cost_of_sales_per_day	1.75
25	Creditors_turnover	1.49
26	Debtors_turnover	1.48
27	Finished_goods_turnover	1.55
28	WIP_turnover	1.74
29	Raw_material_turnover	1.39
30	Shares_outstanding	2.91
31	EPS	7.36
32	Adjusted_EPS	6.66

Table 23: VIF scores

We have dropped the variables with VIF score of over 10 one at a time and will build the model using remaining variables.

Model Summary

Logit Regression Results

Dep. Variable:	default	No. Observations:	2851
Model:	Logit	Df Residuals:	2818
Method:	MLE	Df Model:	32
Date:	Sun, 24 Nov 2024	Pseudo R-squ.:	0.03199
Time:	08:34:59	Log-Likelihood:	-1427.7
converged:	True	LL-Null:	-1474.9
Covariance Type:	nonrobust	LLR p-value:	4.591e-08

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3706	0.050	-27.484	0.000	-1.468	-1.273
Net_worth	-0.0797	0.121	-0.656	0.512	-0.318	0.158
Change_in_stock	-0.0256	0.061	-0.419	0.675	-0.146	0.094
Total_expenses	-0.0103	0.120	-0.086	0.932	-0.245	0.224
Profit_after_tax	0.1178	0.095	1.242	0.214	-0.068	0.304
PBDITA_as_perc_of_total_income	-0.0106	0.082	-0.128	0.898	-0.172	0.151
PAT_as_perc_of_total_income	-0.2428	0.075	-3.233	0.001	-0.390	-0.096
Cash_profit_as_perc_of_total_income	-0.0916	0.099	-0.926	0.354	-0.286	0.102
PAT_as_perc_of_net_worth	-0.0272	0.068	-0.403	0.687	-0.160	0.105
Income_from_fincial_services	0.0839	0.069	1.215	0.224	-0.051	0.219
Total_capital	-0.0034	0.077	-0.044	0.965	-0.154	0.148
Borrowings	0.0510	0.084	0.604	0.546	-0.114	0.216
Current_liabilities_&_provisions	0.0636	0.109	0.583	0.560	-0.150	0.277
Cumulative_retained_profits	0.0828	0.109	0.761	0.446	-0.130	0.296
TOL_to_TNW	0.2529	0.070	3.609	0.000	0.116	0.390
Total_term_liabilities__to__tangible_net_worth	-0.0549	0.087	-0.634	0.526	-0.225	0.115
Contingent_liabilities__to__Net_worth_perc	0.0260	0.050	0.517	0.605	-0.072	0.124
Net_fixed_assets	-0.0404	0.096	-0.422	0.673	-0.228	0.147
Current_assets	-0.3116	0.140	-2.222	0.026	-0.586	-0.037
Net_working_capital	0.1102	0.067	1.640	0.101	-0.021	0.242
Quick_ratio_times	0.0174	0.081	0.214	0.831	-0.142	0.177
Current_ratio_times	-0.0311	0.076	-0.411	0.681	-0.180	0.117
Debt_to_equity_ratio_times	-0.0336	0.096	-0.348	0.728	-0.223	0.156
Cash_to_current_liabilities_times	-0.0095	0.067	-0.142	0.887	-0.141	0.122
Cash_to_average_cost_of_sales_per_day	-0.0205	0.055	-0.377	0.706	-0.127	0.086
Creditors_turnover	0.0090	0.059	0.153	0.879	-0.107	0.125
Debtors_turnover	0.0025	0.060	0.042	0.967	-0.116	0.121
Finished_goods_turnover	-0.0339	0.065	-0.524	0.600	-0.161	0.093
WIP_turnover	0.0507	0.067	0.761	0.447	-0.080	0.181
Raw_material_turnover	-0.0700	0.058	-1.209	0.227	-0.184	0.044
Shares_outstanding	0.1128	0.080	1.405	0.160	-0.045	0.270
EPS	0.0726	0.132	0.550	0.582	-0.186	0.331
Adjusted_EPS	-0.0700	0.127	-0.550	0.582	-0.319	0.179

Table 24: Model summary

On checking the model summary for logistic regression model there are variables with p-value of over 0.05 which means that there is not enough evidence to suggest that these variables are helpful in predicting the target variable. Thus, we dropped those variables one at a time for whom p-value is over 0.05 and then check the p-value for all the remaining variable repeating this process till the p-value for all the remaining variables is below 0.05 and rebuild the model using the remaining variables.

```
Optimization terminated successfully.
         Current function value: 0.503579
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:                 2851
Model:                          Logit   Df Residuals:                     2846
Method:                           MLE   Df Model:                            4
Date:                Sun, 24 Nov 2024   Pseudo R-squ.:                 0.02657
Time:                        08:34:59   Log-Likelihood:                -1435.7
converged:                       True   LL-Null:                       -1474.9
Covariance Type:            nonrobust   LLR p-value:                 3.845e-16
=================================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                          -1.3705      0.048    -28.592      0.000      -1.464      -1.277
Profit_after_tax                0.1737      0.074      2.343      0.019       0.028       0.319
PAT_as_perc_of_total_income    -0.3150      0.050     -6.270      0.000      -0.413      -0.216
TOL_to_TNW                      0.1932      0.043      4.487      0.000       0.109       0.278
Current_assets                 -0.1552      0.072     -2.155      0.031      -0.296      -0.014
=================================================================================================
```

Table 25: Model summary

After dropping the variables which have VIF scores and p-value above the required limit we have found that only 4 attributes are statistically significant to predict the default value amongst which PAT_as_perc_of_total_income has the highest coefficient value of -0.3150 meaning the companies which have high after-tax profit as a percentage of total income or in simple terms have high net margins such companies are less likely to default.

Model Evaluation

For model evaluation, we will utilize a confusion matrix and a classification report, focusing on metrics such as accuracy, precision, and recall. The confusion matrix provides a detailed comparison of actual versus predicted values, helping to understand the distribution of correct and incorrect predictions. The accuracy score measures the overall performance of the model on both training and test datasets, allowing us to assess its stability and potential bias. Precision and recall, on the other hand, are critical for evaluating the model's effectiveness in identifying positive cases while minimizing false positives and false negatives. These metrics collectively ensure a comprehensive assessment of the model's performance.

For Train Data

Confusion Matrix

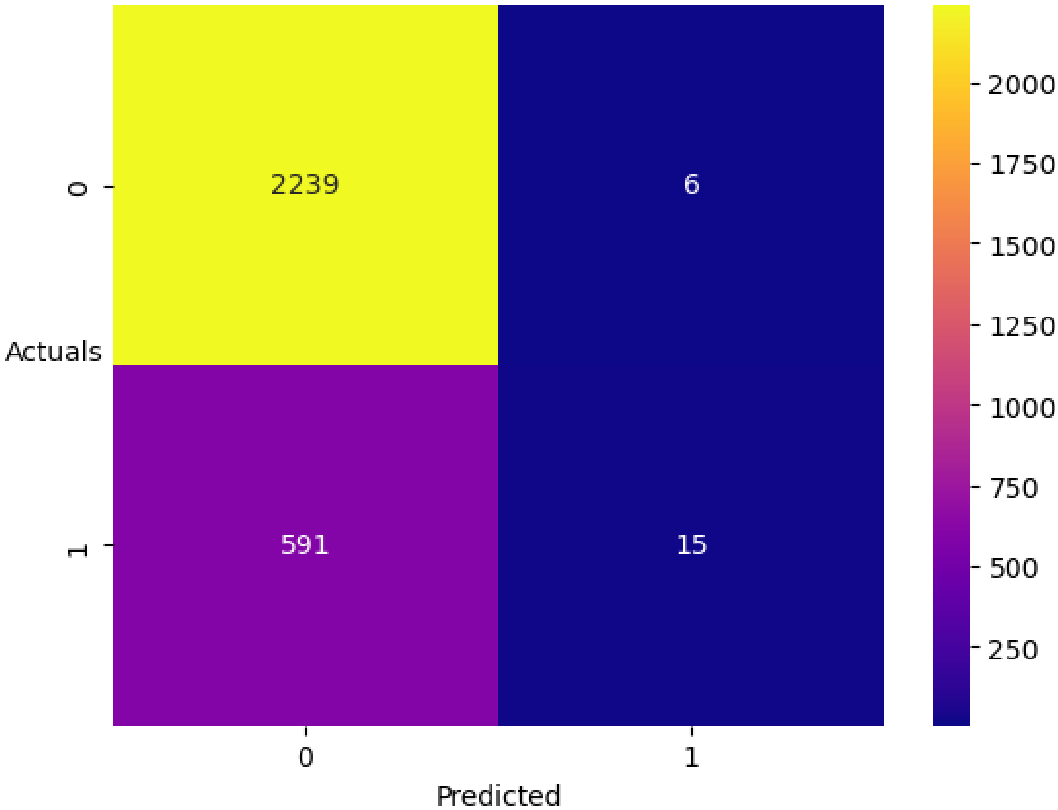


Figure 5: Confusion matrix

Classification Report

	precision	recall	f1-score	support
0.0	0.791	0.997	0.882	2245
1.0	0.714	0.025	0.048	606
accuracy			0.791	2851
macro avg	0.753	0.511	0.465	2851
weighted avg	0.775	0.791	0.705	2851

Table 26: Classification report

While the model demonstrates decent performance in terms of accuracy and precision, its recall for predicting defaults is significantly low. To further evaluate its stability and reliability, we will test the model on the test dataset and analyze its performance.

Checking on test data

Confusion Matrix

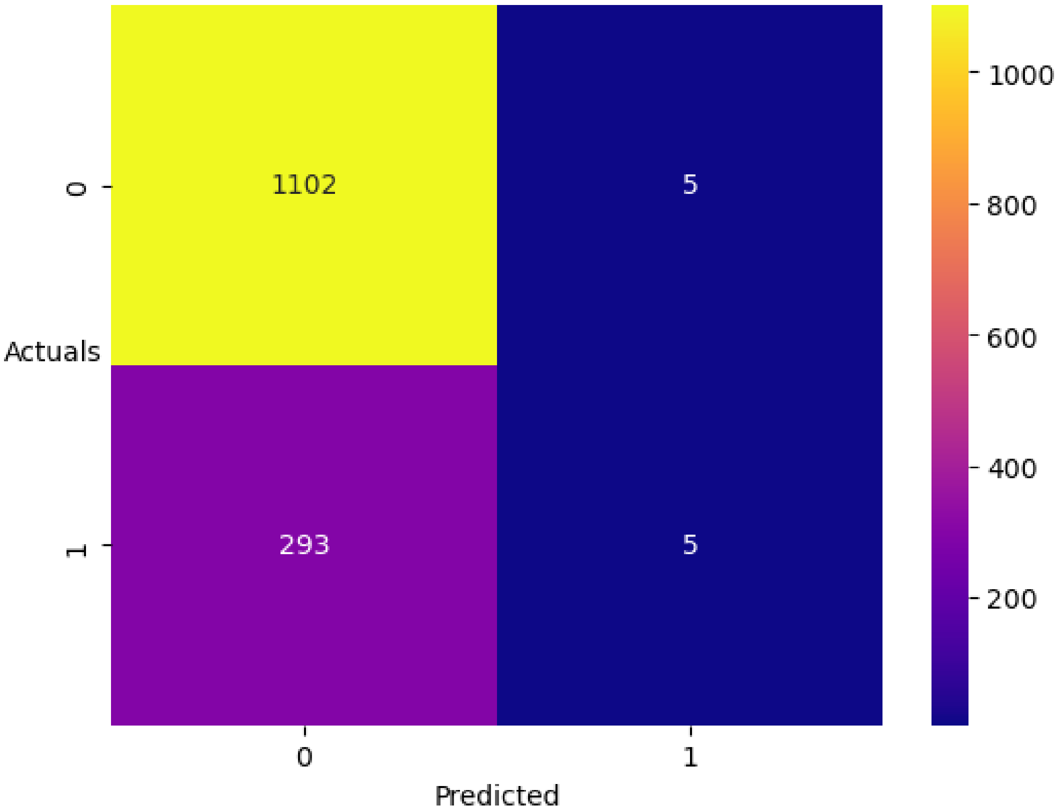


Figure 6: Confusion matrix

Classification Report

	precision	recall	f1-score	support
0.0	0.790	0.995	0.881	1107
1.0	0.500	0.017	0.032	298
accuracy			0.788	1405
macro avg	0.645	0.506	0.457	1405
weighted avg	0.728	0.788	0.701	1405

Table 27: Classification report

Model performance for both test and train data are almost identical, however, recall for default is very poor and to improve it we will use ROC curve by Youden method to find optimal threshold which could help improve the recall score.

Optimal threshold using ROC curve

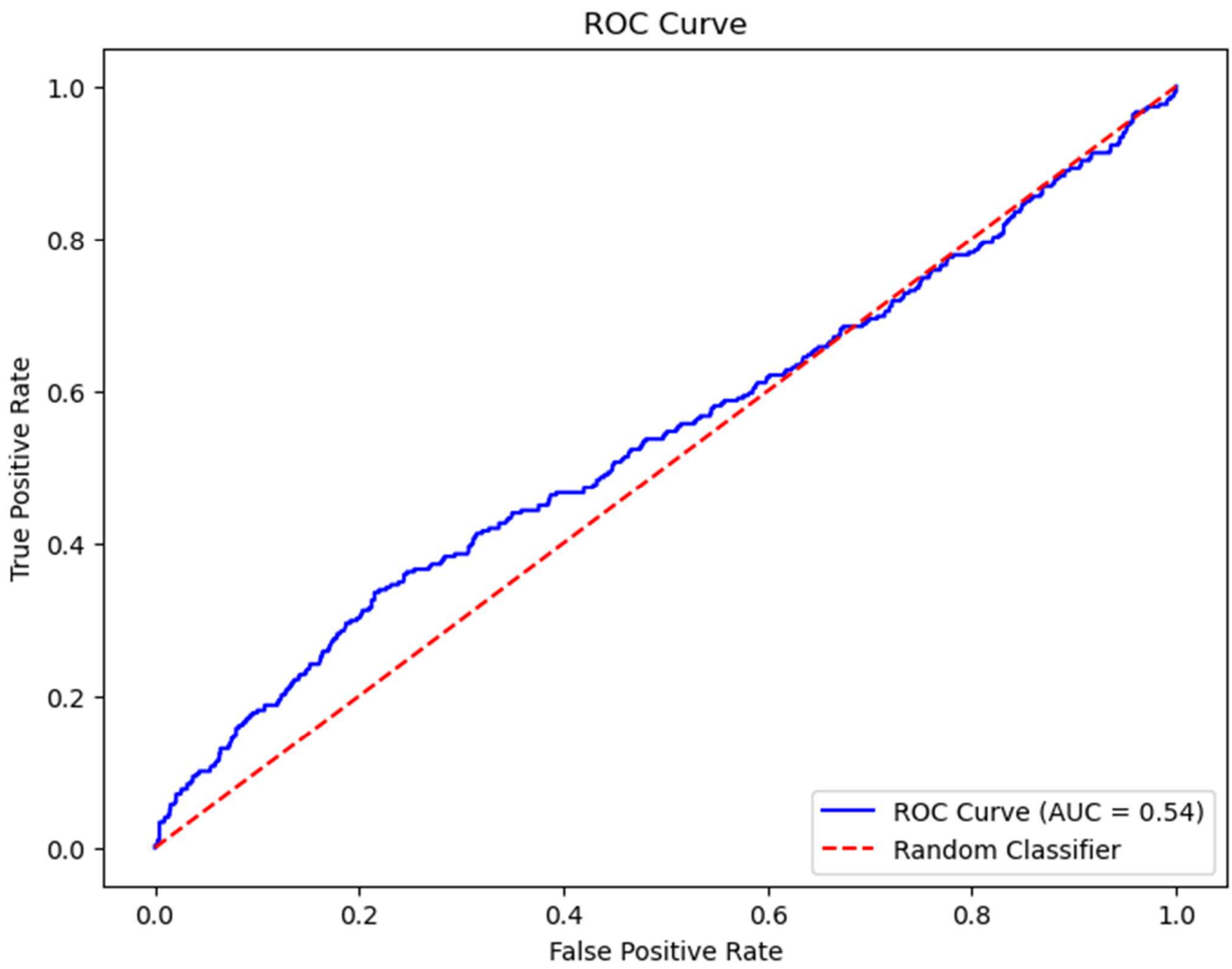


Figure 7: AUC-ROC curve

```
Optimal Threshold Value: 0.24
```

Logistic Regression Optimal Model

By taking optimal threshold at 0.24 we will predict the target class wherein if the probability is greater than the optimal threshold then the company will be predicted as defaulter.

Model Evaluation

On Test Data

Confusion Matrix

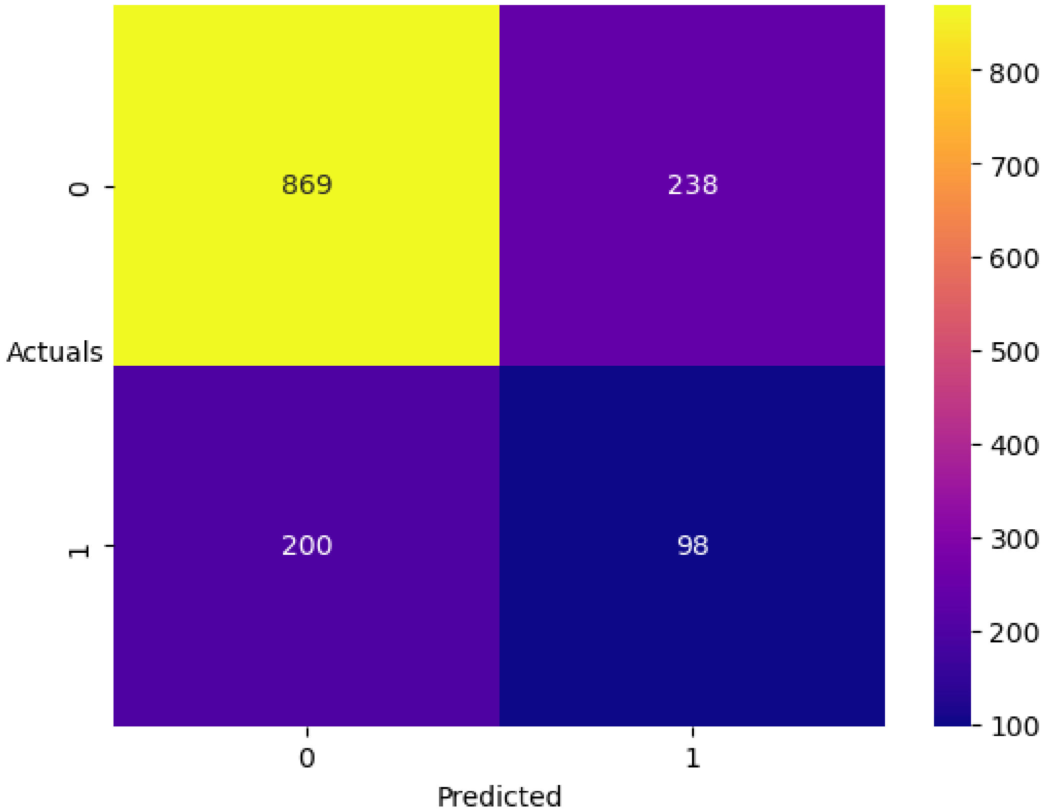


Figure 8: Confusion matrix

Classification Report

	precision	recall	f1-score	support
0.0	0.813	0.785	0.799	1107
1.0	0.292	0.329	0.309	298
accuracy			0.688	1405
macro avg	0.552	0.557	0.554	1405
weighted avg	0.702	0.688	0.695	1405

Table 28: Classification report

By adjusting the prediction threshold to 0.24, we successfully improved the recall score from 0.017 to 0.329. However, this improvement in recall comes at the cost of a slight decline in both precision and accuracy. This trade-off highlights the balance between correctly.

Building model using Random Forest

We built a classification model using Random Forest technique from ensemble module in scikit-learn library and since this technique is capable of handling multi-collinearity on its own, we can build the model straight away whose accuracy on train and test data are:

Model accuracy for train data
0.9635

Model accuracy for test data
0.7032

Accuracy score for test and train data show significant variance meaning model is not stable. We will have to tune the hyperparameters to make the model stable.

Hyperparameter Tuning

We run the model using different sets of parameters under GridSearchCV from model_selection module in scikit-learn library and best parameters came as:

```
{'max_depth': 3, 'max_features': 0.55, 'n_estimators': 125}
```

Table 29: Best parameters

Using these parameters, we built a model whose accuracy scores are:

```
Model accuracy for train data
0.8014731673097159

 Model accuracy for test data
0.7900355871886121
```

Accuracy score for test and train data are almost similar. We will evaluate the model performance using confusion matrix and classification table.

Model Evaluation

For train data

Confusion Matrix

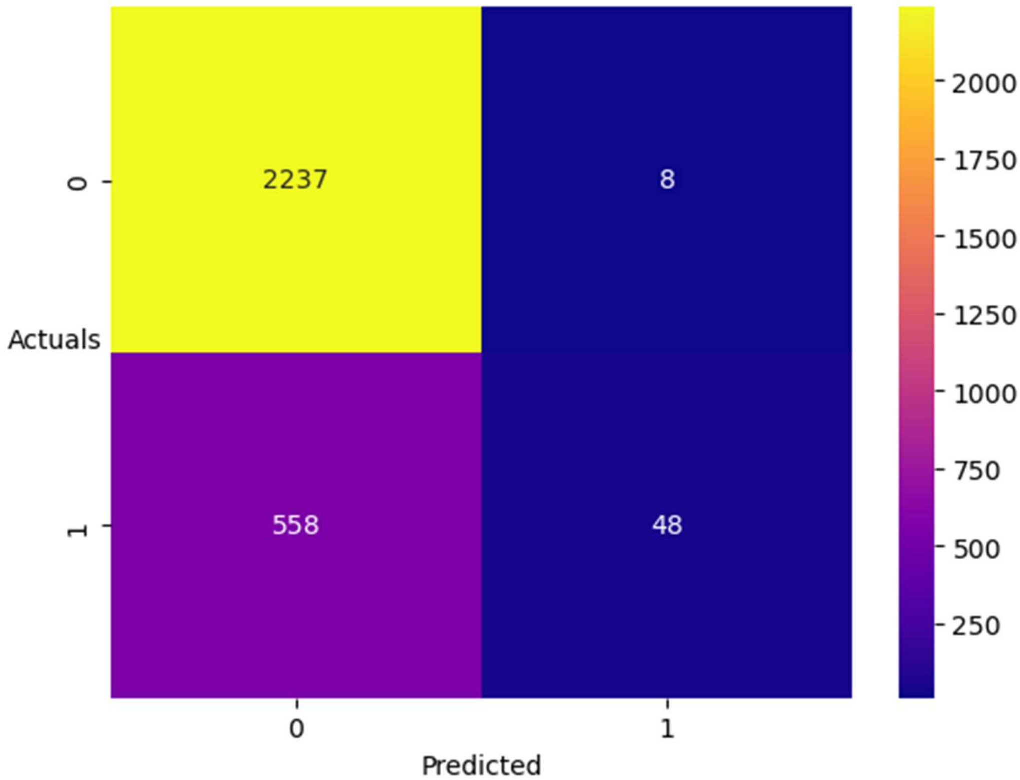


Figure 9: Confusion matrix

Classification Report

	precision	recall	f1-score	support
0.0	0.800	0.996	0.888	2245
1.0	0.857	0.079	0.145	606
accuracy			0.801	2851
macro avg	0.829	0.538	0.516	2851
weighted avg	0.812	0.801	0.730	2851

Table 30: Classification report

Checking on test data

Confusion Matrix

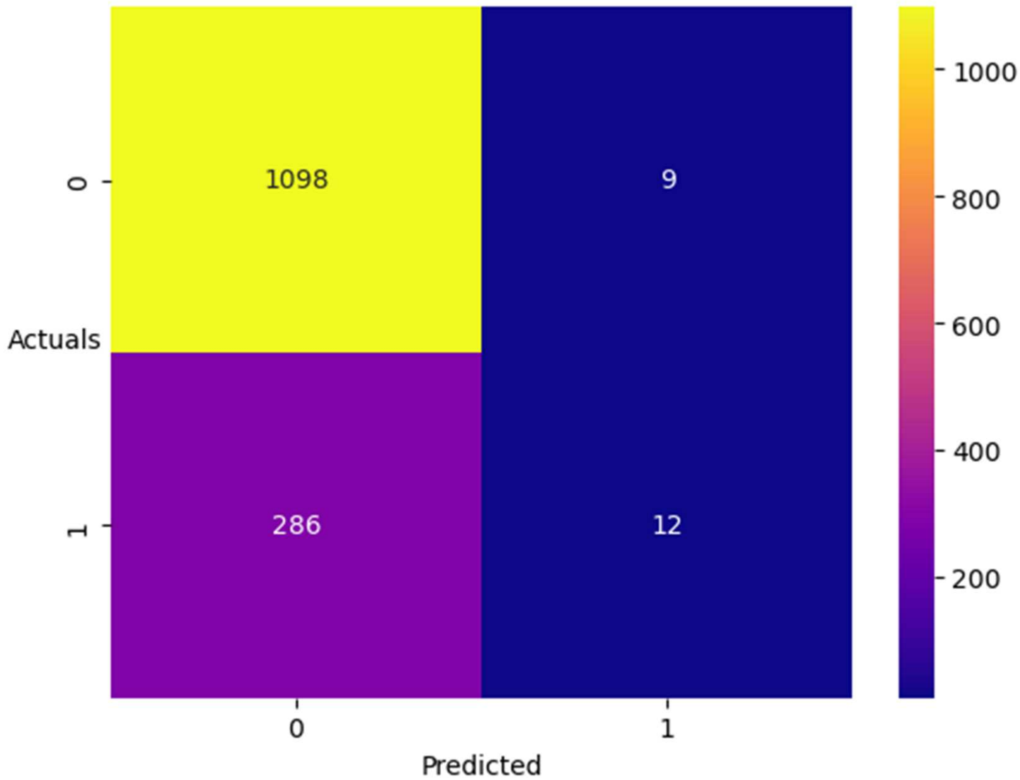


Figure 10: Confusion matrix

Classification Report

	precision	recall	f1-score	support
0.0	0.793	0.992	0.882	1107
1.0	0.571	0.040	0.075	298
accuracy			0.790	1405
macro avg	0.682	0.516	0.478	1405
weighted avg	0.746	0.790	0.711	1405

Table31: Classification report

Model performance for both test and train data are almost identical, however, recall for default is very poor and to improve it we will use ROC curve to find optimal threshold which could help improve the recall score.

Optimal threshold using ROC curve

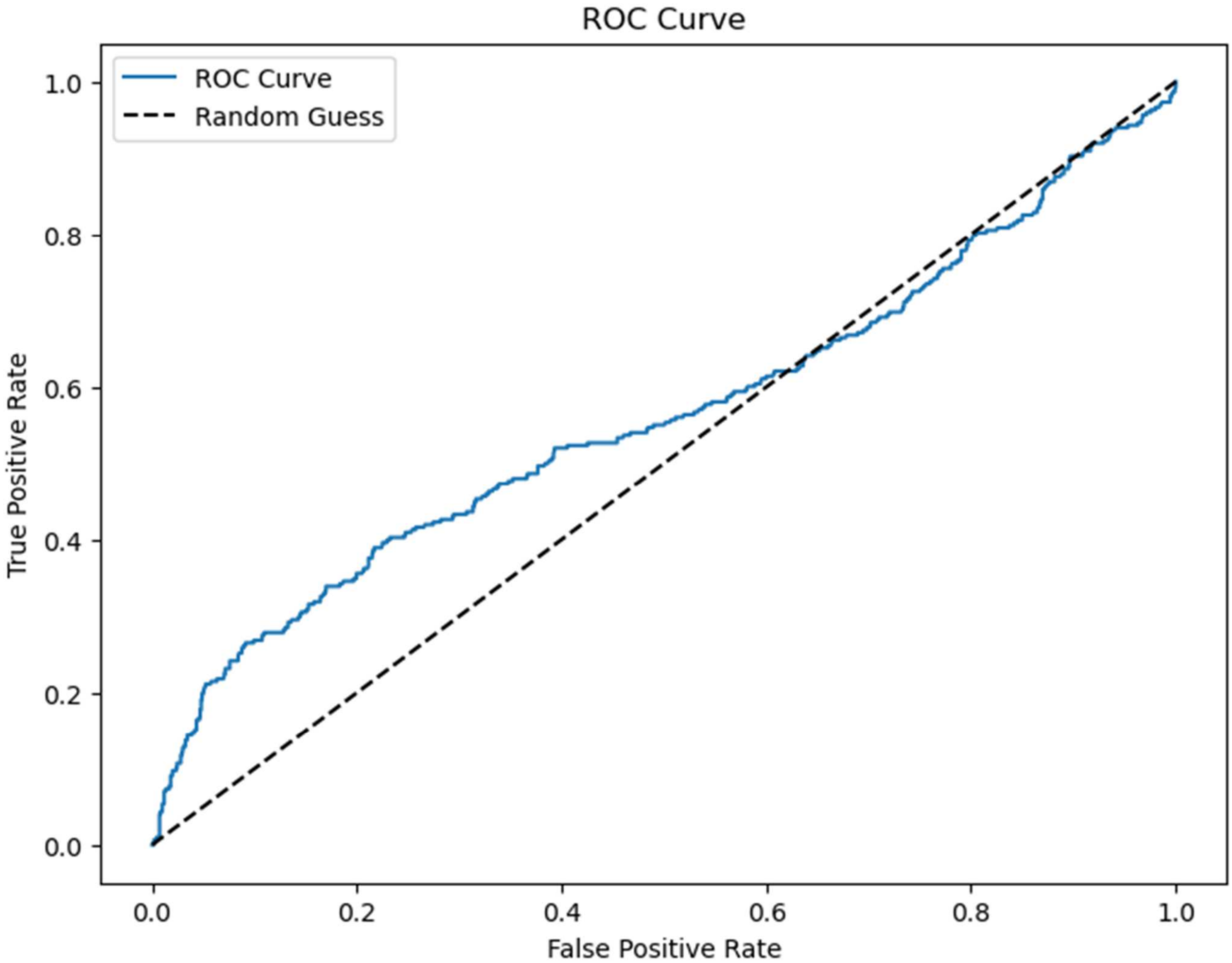


Figure 11: AUC_ROC curve

```
Optimal Threshold Value: 0.27
```

Model Evaluation

For Test Data

Confusion Matrix

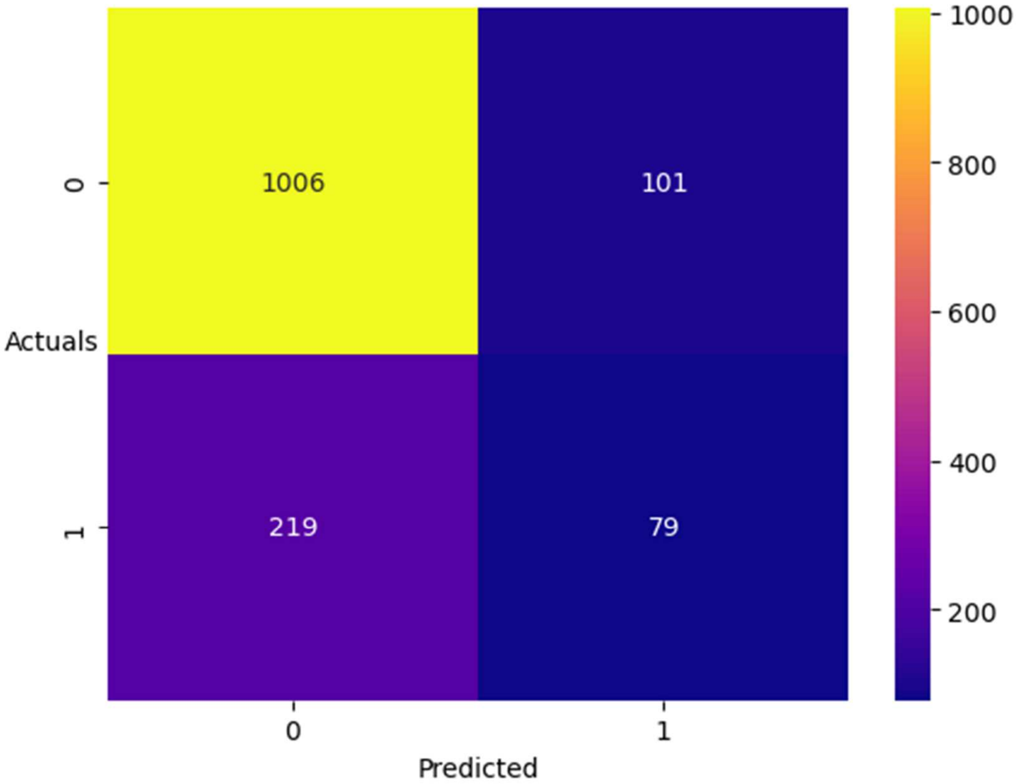


Figure 12: Confusion matrix

Classification Report

	precision	recall	f1-score	support
0.0	0.82	0.91	0.86	1107
1.0	0.44	0.27	0.33	298
accuracy			0.77	1405
macro avg	0.63	0.59	0.60	1405
weighted avg	0.74	0.77	0.75	1405

Table 32: Classification report

By adjusting the prediction threshold to 0.27, we successfully improved the recall score from 0.04 to 0.27. However, this improvement in recall comes at the cost of a slight decline in both precision and accuracy. This trade-off highlights the balance between correctly.

1.11 Model Comparison

	Model	Accuracy	Precision	Recall
0	Logit_model	0.79	0.50	0.02
1	Logit_model_optimal	0.69	0.29	0.33
2	RF_model	0.79	0.57	0.04
3	RF_model_optimal	0.77	0.44	0.27

Table 33: Model comparison

On evaluating all the models based on combination of Accuracy, Precision and Recall scores Random Forest model optimized for threshold is performing the best as it is providing the best balance for all the three metrics wherein other models are performing significantly poorly on 1 of the 3 metrics. Moving forward we will take this model as the final model.

1.12 Most Important Features

	imp
TOL_to_TNW	0.15
PBT_as_perc_of_total_income	0.12
Cash_profit_as_perc_of_total_income	0.10
PAT_as_perc_of_total_income	0.08
Reserves_and_funds	0.07

Table 34: Important Features

On examining the most important features for RF_model_optimal, TOL_to_TNW emerges as the most influential, contributing 15% of the model's total importance. TOL_to_TNW reflects the proportion of total liabilities to a company's net worth, indicating the extent to which its assets are financed by debt rather than equity. A higher value signifies greater financial leverage and potentially increased financial risk, making it a crucial factor for predicting financial performance and identifying default risks.

Similarly, other significant features, such as PBT_as_perc_of_total_income, Cash_profit_as_perc_of_total_income, PAT_as_perc_of_total_income, and Reserves_and_funds, provide insights into a company's profitability and cash flow. These metrics play a vital role in assessing a company's ability to generate income, maintain liquidity, and service its liabilities effectively. Together, these features offer a comprehensive view of a company's financial health, aiding in accurate predictions and proactive risk management.

1.13 Conclusion

Key Takeaways

1. The dataset comprises over 50 attributes for each company. However, upon analysis, it was observed that nearly 50% of the companies had more than 10% of their data missing. Further investigation revealed that these companies with higher proportions of missing data exhibited a significantly higher likelihood of default.
2. For the classification models developed, the Random Forest model with an adjusted threshold emerged as the best performer, offering the most balanced trade-off between accuracy, precision, and recall—key metrics for evaluating model effectiveness. Models using the standard threshold performed poorly in terms of recall, often misclassifying nearly all defaulters as non-defaulters, which significantly undermines the model's utility. Among the models tested, the Logistic Regression model with an adjusted threshold had the weakest performance, with the lowest accuracy and precision scores. This indicates that it struggled to classify companies correctly and exhibited the highest rate of misclassification for both defaulters and non-defaulters, which could lead to negative consequences if deployed in real-world scenarios.
3. The primary goal of this project is to classify companies based on their ability to meet future financial obligations. To achieve this, key factors should include metrics that offer insights into a company's income-generating capacity and cash flow stability. Upon analyzing the most significant features in the best-performing model, Total Liabilities to Total Net Worth (TOL_to_TNW) emerged as the top contributor, indicating the degree of financial leverage and risk associated with the company. Other important features include:

- Profit Before Tax (PBT) as a Percentage of Total Income
- Profit After Tax (PAT) as a Percentage of Total Income
- Cash Profit as a Percentage of Total Income
- Reserves and Surplus

These factors collectively provide a comprehensive understanding of a company's current financial health, operational efficiency, and capacity to generate income. By incorporating these features, the model ensures a more accurate prediction of a company's ability to meet its financial obligations, thereby aiding in effective decision-making.

Key Recommendations

1. Companies with over 10% missing data have demonstrated a significantly higher probability of default. It is recommended to conduct a thorough investigation to determine whether this non-disclosure is incidental or a deliberate attempt to withhold critical information. Establishing the intent behind these gaps in data can provide valuable insights into patterns of non-compliance or potentially fraudulent activity. This investigation will not only enhance the reliability of the dataset but also help refine the model's ability to identify high-risk companies effectively.
2. We have successfully built models using logistic regression and random forest and identified the best-performing model. However, there is considerable scope for improvement, especially regarding precision and recall. To address these limitations and enhance model performance, we recommend the following:

- Approximately 8% of the dataset was missing, which is significant, given that some variables were derived from others. Furthermore, the possibility of deliberate non-disclosure raises concerns about the reliability of the data. To ensure completeness and trustworthiness, it is recommended that future datasets are sourced directly from audited financial statements of the companies. This would eliminate doubts about data integrity and provide a more robust foundation for model development.
- Logistic regression, which was a mandatory model for this project, is highly sensitive to outliers. Consequently, an outlier treatment process was applied to the dataset, affecting over 8% of the data

(based on conservative thresholds at the 5th and 95th percentiles). This resulted in over 16% of the data being imputed, likely impacting model performance. Given the high prevalence of outliers and missing data, we recommend exploring alternative modelling techniques such as decision trees, bagging, and boosting methods. These models are less sensitive to outliers and better equipped to handle missing data, potentially yielding improved results.

- Features related to income generation, cash flows, and financial standing were identified as the most important predictors of default. To enhance predictive power, we recommend collecting financial records from the past few years in addition to the current year. This historical data can be used to build regression models that forecast future performance, which can then be integrated into the classification model. This approach will likely provide a more comprehensive understanding of the company's financial trajectory and improve overall model accuracy.

Problem 2

2.1 Background Information

Investing in financial markets involves substantial risk, primarily driven by potential price fluctuations of assets. These swings often result from unforeseen economic events or geopolitical developments, which can drastically impact investor sentiment and market dynamics.

2.2 Business Context

Given the significant risks inherent in financial markets, it is crucial for investors to assess and understand the risks they are undertaking. This understanding enables them to align their investment strategies with their financial objectives, fostering informed decision-making and portfolio optimization.

2.3 Problem Statement

The objective of this is to develop a robust risk evaluation framework that leverages historical market data by quantifying and predicting potential risks, the framework aims to guide investors in selecting investment strategies that balance risk and reward effectively, ultimately supporting their financial goals.

2.4 METHODOLOGY

Import the libraries – Load the data – Check the structure of the data – Check the types of the data – Check for missing values – Check the statistical summary – Check for and treat (if needed) Data Irregularities – Univariate Analysis – Analyzing Returns – Conclusion

Key Points

1. **Data Collection**: Historic data of stock price movement was taken from stock exchange.
2. **Data Cleaning and Pre-processing:** The dataset was thoroughly examined for column names, duplicates, missing values, bad data, and outliers. Inconsistent column names were standardized by renaming relevant attributes to ensure uniformity in nomenclature.
3. **Bivariate Analysis:** All the stock prices were examined over the period of time with aim of gaining deeper insights about price movement over time.
4. **Visualization Techniques:** In the report we have used scatter plots.
5. **Tools and Software:** We have carried out the analysis using programming language python on Jupyter notebook. For this analysis Python libraries Numpy, Pandas, Matplotlib and Seaborn were used.

2.5 Data Overview

1. **Data Description:** Dataset has 418 rows and 6 columns.

```
shape of the dataset
----------------------------------------------------------------------------

(418, 6)
```

Table 35: Dataset Shape

2. **Dataset Information:** Of the 6 columns in the dataset, 1 is object type and 5 are int 64 type.

```
information of features
----------------------------------------------------------------------------
<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
```

Table 36: Dataset Information

3. **Missing Value Check:** There are no missing values in the dataset.

```
missing values
----------------------------------------------------------------------------

Date             0
ITC Limited      0
Bharti Airtel    0
Tata Motors      0
DLF Limited      0
Yes Bank         0
dtype: int64
```

Table 37: Missing values information

4. **Duplicate Values:** Data was checked for duplicate values and no duplicates were found

```
checking for duplicates
----------------------------------------------------------------------------
number of duplicate rows: 0
```

Table 38: Data Duplicates

5. **Statistical Summary:**

```
statistical summary
--------------------------------------------------------------------------
```

	count	mean	std	min	25%	50%	75%	max
ITC Limited	418.00	278.96	75.11	156.00	224.25	265.50	304.00	493.00
Bharti Airtel	418.00	528.26	226.51	261.00	334.00	478.00	706.75	1236.00
Tata Motors	418.00	368.62	182.02	65.00	186.00	399.50	466.00	1035.00
DLF Limited	418.00	276.83	156.28	110.00	166.25	213.00	360.50	928.00
Yes Bank	418.00	124.44	130.09	11.00	16.00	30.00	249.75	397.00

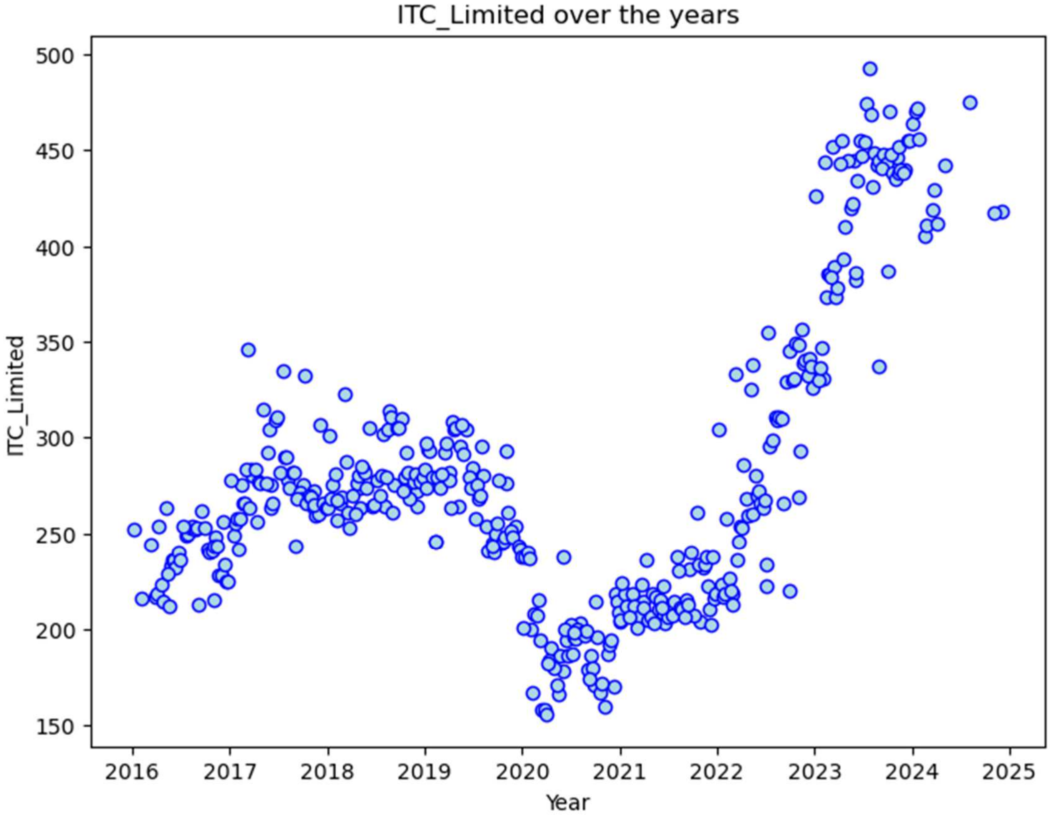
Table 39: Statistical summary

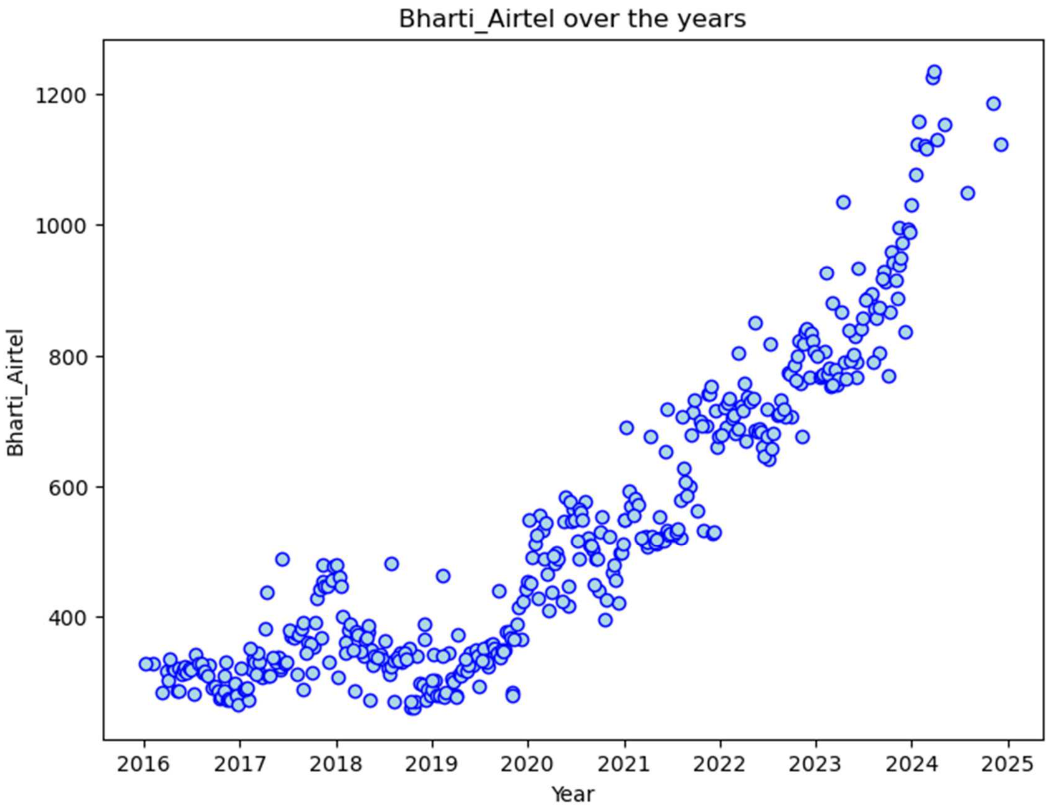
Key Observation

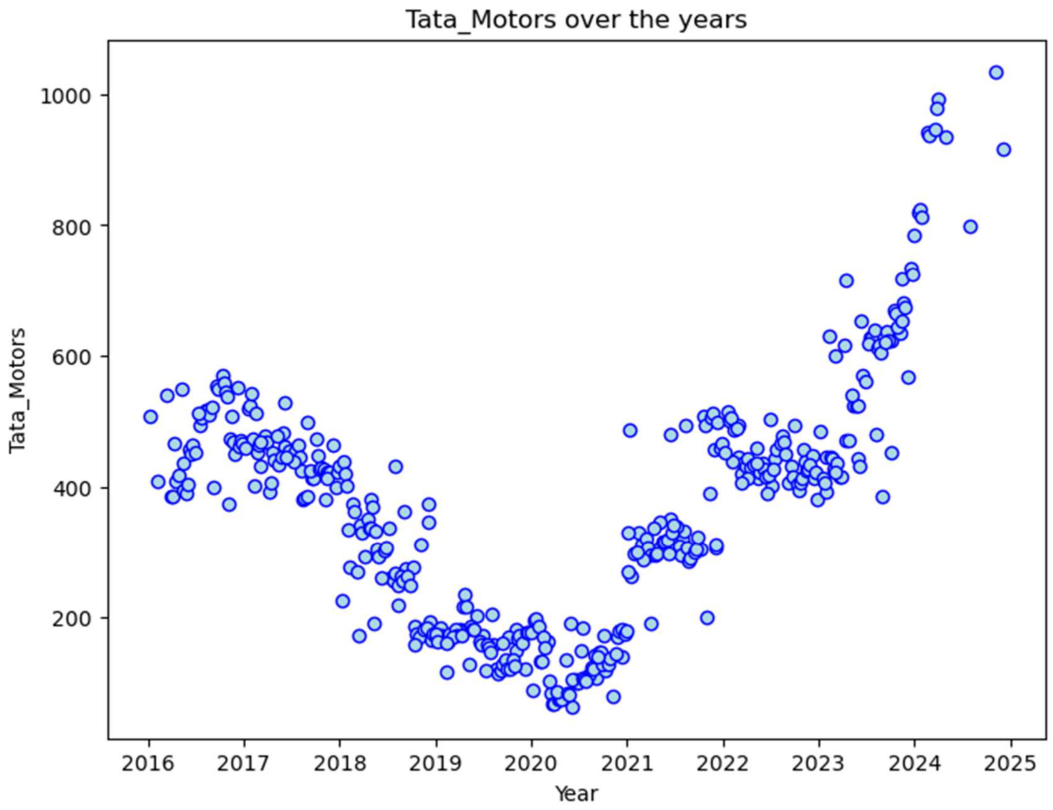
1. There are 418 rows and 6 columns in the dataset.
2. Datatype for date column is object which we will have to convert to date time format and for rest five columns datatype is integer type meaning there is no junk data in these columns.
3. On checking statistical summary there is nothing unusual in the data.
4. Column names have spaces in them which we will have to remove we will do so during data pre-processing.
5. There are no missing values or duplicates in data.

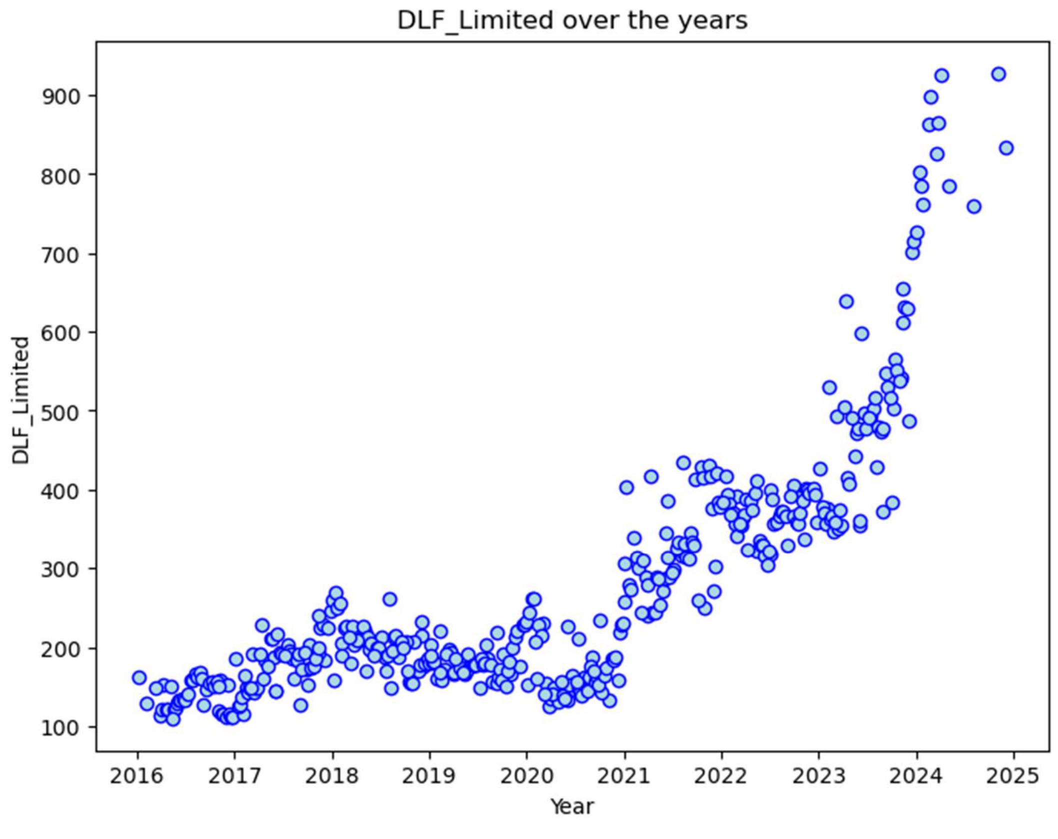
2.6 Exploratory Data Analysis

Plotting price trend over time for different companies


ITC_Limited over the years
ITC_Limited
Year
500
450
400
350
300
250
200
150
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025


Bharti_Airtel over the years
Bharti_Airtel
Year
1200
1000
800
600
400
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025


Tata_Motors over the years
Tata_Motors
Year
1000
800
600
400
200
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025


DLF_Limited over the years
DLF_Limited
Year
900
800
700
600
500
400
300
200
100
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

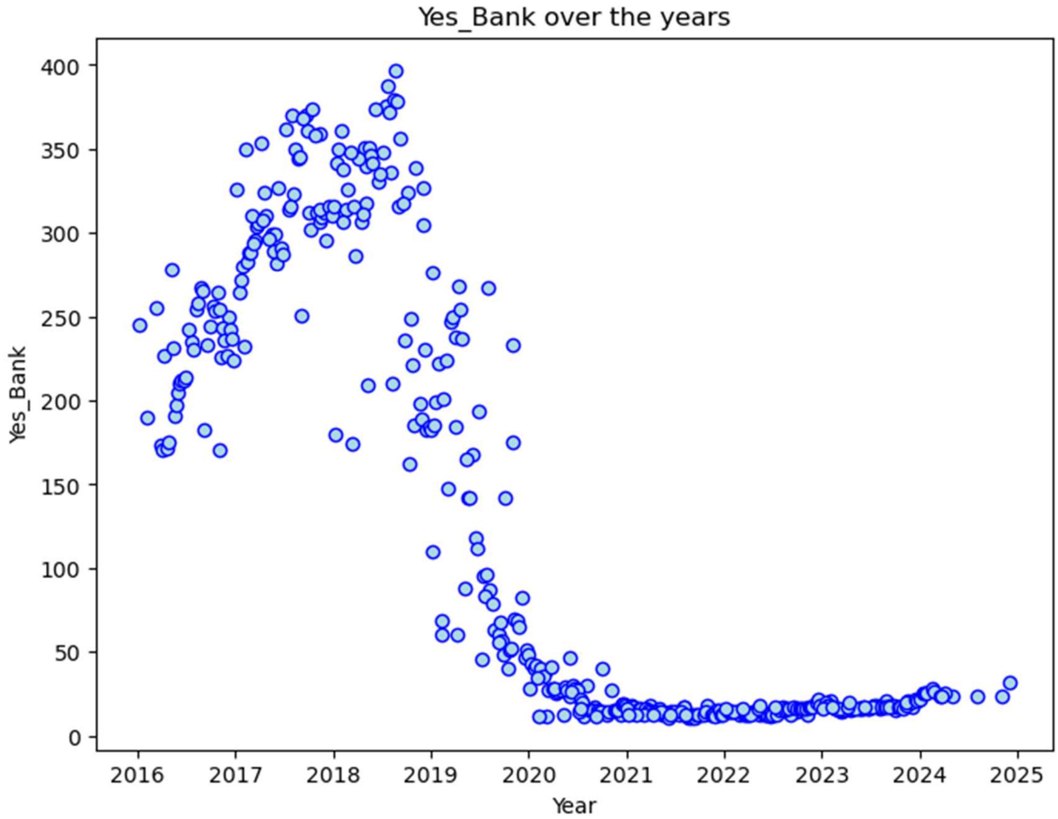


Figure 13: Price trend over time for different stocks

Key Observations

1. Amongst the five stocks the trend for all except Yes Bank is upward while Yes Bank is showing a downward trend.
2. In terms of the scattering of markers for Yes Bank markers appear most scattered followed by ITC Limited and for DLF Limited it appears to be least scattered.

2.7 Analysing Returns

Taking Logarithms and Differences

To analyse stock returns, we calculated the logarithmic returns, which provide a more accurate measure of percentage change compared to simple returns, particularly for financial data. This was achieved by taking the natural logarithm of stock prices and computing the difference between the current price and the previous price. Logarithmic returns are additive over time and help address issues of scale, making them ideal for comparing returns across different stocks and time periods.

	ITC_Limited	Bharti_Airtel	Tata_Motors	DLF_Limited	Yes_Bank
0	NaN	NaN	NaN	NaN	NaN
1	0.00	-0.05	0.00	0.06	-0.01
2	-0.01	0.02	-0.03	-0.01	0.00
3	0.04	0.04	0.09	0.02	0.01
4	-0.04	-0.00	0.02	0.00	0.02

Table 40: Logarithmic returns

Using the calculated logarithmic price changes, we determined the mean price change and standard deviation for each stock to evaluate their average performance and volatility. The results were compiled into a table, where the stocks were sorted in ascending order of volatility, providing a clear ranking from the least to the most volatile stocks. This approach helps in identifying stable investment options while analyzing risk associated with each stock.

Calculating average return and Volatility

	Average	Volatility
ITC_Limited	0.0016	0.0359
Bharti_Airtel	0.0033	0.0387
DLF_Limited	0.0049	0.0578
Tata_Motors	0.0022	0.0605
Yes_Bank	-0.0047	0.0939

Table 41: Average return and risk

To understand the relation between Volatility and average return in better way we plotted the above table in a scatter plot.

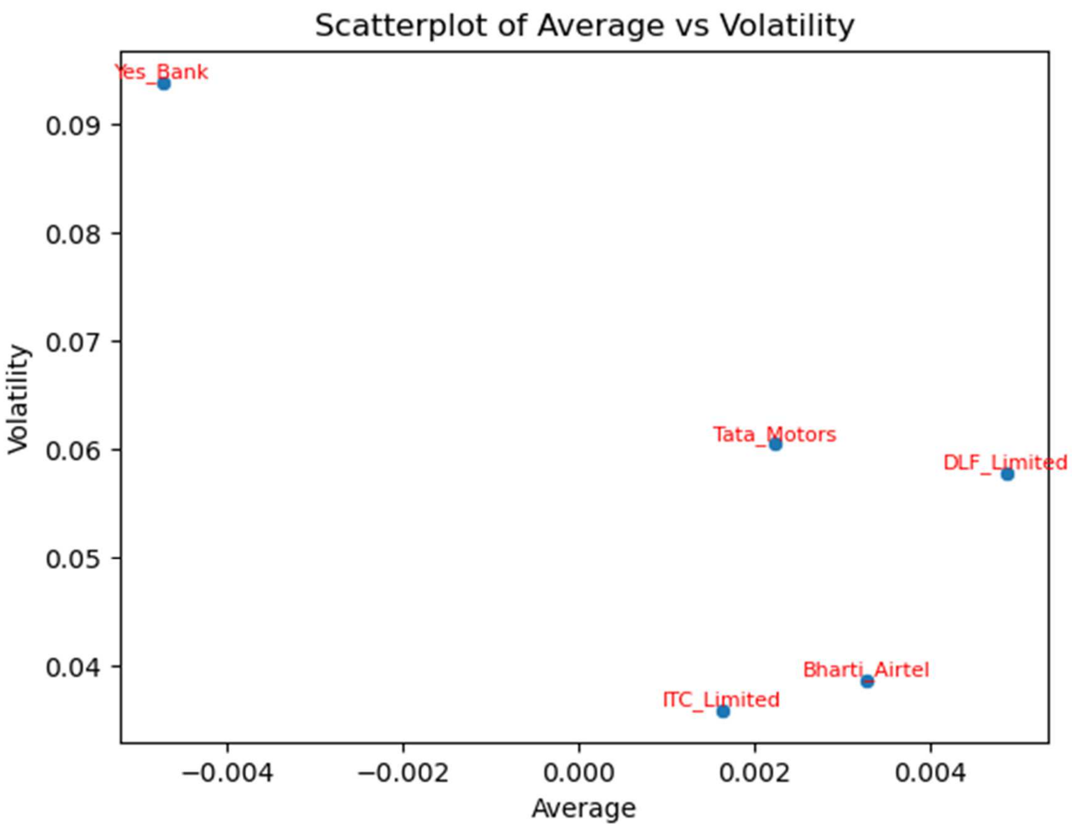


Figure 14: Return vs risk

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:

- ITC Limited
- Bharti Airtel
- DLF Limited
- Tata Motors

To identify the stocks which give the best balance between risk and return we can evaluate the Sharpe ratio.

Sharpe Ratio

The Sharpe Ratio is a measure used to evaluate the risk-adjusted return of an investment or portfolio. It helps in better assessing portfolio performance because it takes both risk and return into account.

$$\text{Sharpe Ratio} = \frac{\text{Mean Return} - \text{Risk-Free Rate}}{\text{Standard Deviation of Return}}$$

Equation 1: Sharpe ratio

For Sharpe ratio we need risk free return which is normally considered to be rate for government bonds which currently is 5% per annum.

Since, the government bond rate is per annum and our data is in weekly terms we converted the risk-free rate in weekly terms, taking natural log value and calculated the Sharpe ratio whose values came at:

	Sharpe_Ratio
DLF_Limited	0.0675
Bharti_Airtel	0.0596
Tata_Motors	0.0210
ITC_Limited	0.0187
Yes_Bank	-0.0607

Table 42: Sharpe ratio

Evaluating stocks solely based on average return and volatility can lead to misleading conclusions. For instance, ITC Limited shows the lowest volatility, followed by Bharti Airtel, which might initially suggest they are the best-performing stocks. However, this simplistic assessment overlooks the balance between risk and return. When we incorporate Sharpe's Ratio, which evaluates performance relative to risk, a different picture emerges. DLF Limited stands out as the best-performing stock, followed by Bharti Airtel. Interestingly, despite its low volatility, ITC Limited ranks as the second-worst in terms of Sharpe's Ratio, highlighting the importance of a comprehensive evaluation that accounts for both risk and return.

2.8 Conclusion

The Market Risk Analysis provided valuable insights into the risk-return dynamics of a portfolio. By incorporating statistical measures and the Sharpe ratio, we were able to move beyond simplistic metrics like mean return and volatility, enabling a more comprehensive evaluation of portfolio performance. Key insights and actionable recommendations are as follows:

Key Insights

1. The analysis underscores the importance of considering both risk and return when evaluating stocks. Solely relying on metrics like average return or volatility can be misleading, as they fail to account for the risk-adjusted performance of investments.
2. By integrating the Sharpe Ratio, we identified that DLF Limited offers the best risk-adjusted returns, despite having higher volatility compared to other stocks like ITC Limited and Bharti Airtel. This demonstrates the necessity of incorporating comprehensive measures for informed decision-making.
3. Although ITC Limited has the lowest volatility, it performs poorly in terms of risk-adjusted returns. This highlights that low risk does not necessarily translate to high performance if returns are not proportionately higher.
4. Bharti Airtel emerges as a strong contender with a balanced performance, making it a viable choice for investors seeking moderate risk and returns.

Key Recommendations

1. Rather than relying solely on standalone metrics such as average return or volatility incorporating risk-adjusted measures like the Sharpe Ratio to gain a complete understanding of stock performance could be more beneficial.
2. DLF Limited, with the highest Sharpe Ratio, should be considered a top priority for inclusion in the portfolio, as it offers the best balance of return relative to risk.

3. ITC Limited’s lower Sharpe Ratio suggests it may not add substantial value to the portfolio. Reassess its inclusion, especially if there are other stocks offering better risk-adjusted returns.
4. While focusing on high Sharpe Ratio stocks, it recommended that the portfolio remains diversified to minimize exposure to stock-specific risks and maintain a balance of industries.
5. Continuously monitoring the portfolio performance and market conditions and adjusting stock allocations based on evolving Sharpe Ratios and changing economic scenarios could be beneficail to sustain optimal risk-adjusted returns.