

Pentapalli Suneel Kumar | January 5, 2024

Finance and Risk Analytics

Contents

[List of Figures 3](#_Toc155541840)

[List of Tables 4](#_Toc155541841)

[PART A: 5](#_Toc155541842)

[Data Shape and Dictionary: 5](#_Toc155541843)

[Data Info: 9](#_Toc155541844)

[Null Check: 10](#_Toc155541845)

[Duplicate Check: 10](#_Toc155541846)

[Data Description: 10](#_Toc155541847)

[Outlier Check and Treatment: 14](#_Toc155541848)

[Missing value Treatment: 18](#_Toc155541849)

[Univariate & Bivariate analysis: 19](#_Toc155541850)

[Train Test Split: 25](#_Toc155541851)

[Logistic Regression(From Statsmodel): 26](#_Toc155541852)

[(Logistic Regression)Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model: 27](#_Toc155541853)

[Build a Random Forest Model on Train Dataset. Also showcase your model building approach: 28](#_Toc155541854)

[Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model: 28](#_Toc155541855)

[Build a LDA Model on Train Dataset. Also showcase your model building approach: 29](#_Toc155541856)

[Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model: 29](#_Toc155541857)

[Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve): 29](#_Toc155541858)

[Conclusions and Recommendations: 32](#_Toc155541859)

[PART B: 32](#_Toc155541860)

[Problem Statement: 32](#_Toc155541861)

[Exploratory Data Analysis: 33](#_Toc155541862)

[Data Info: 33](#_Toc155541863)

[Five Point summary: 33](#_Toc155541864)

[Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference: 34](#_Toc155541865)

[Calculate Returns for all stocks with inference: 35](#_Toc155541866)

[Calculate Stock Means and Standard Deviation for all stocks with inference: 36](#_Toc155541867)

[Draw a plot of Stock Means vs Standard Deviation and state your inference: 37](#_Toc155541868)

[Conclusions and Recommendations: 37](#_Toc155541869)

# List of Figures

Fig 1: Credit Risk Data Missing Values ………………………………………………………………………………………………………………. 9

Fig 2: Credit Rist Box Plots(Before Outlier Treatment) ……………………………………………………………………………………. 14

Fig 3: Credit Rist Box Plots(After Outlier Treatment) ……………………………………………………………………………………… 16

Fig 4: Credit Risk(Missing Values Before Treatment) ………………………………………………………………………………………. 18

Fig 5: Credit Risk(Box Plots after Outlier and Missing Value Treatment) ……………………………………………………….. 18

Fig 6: Credit Risk(Histograms) ………………………………………………………………………………………………………………………… 20

Fig 7: Credit Risk(Heatmap) ……………………………………………………………………………………………………………………………. 22

Fig 8: Scatter Plot 1 …………………………………………………………………………………………………………………………………………… 23

Fig 9: Scatter Plot 2 ………………………………………………………………………………………………………………………………………….. 24

Fig 10: Scatter Plot 3 …………………………………………………………………………………………………………………………………………. 24

Fig 11: Train Test Split(Data Shape) …………………………………………………………………………………………………………………. 26

Fig 12: Logit Model Summary …………………………………………………………………………………………………………………………… 26

Fig 13: Classification Report(Logit) ………………………………………………………………………………………………………………….. 28

Fig 14: Classification Report(Random Forest) …………………………………………………………………………………………………. 28

Fig 15: Classification Report(LDA) …………………………………………………………………………………………………………………… 29

Fig 16: AUC Score Interpretation ……………………………………………………………………………………………………………………… 30

Fig 17: ROC Curve Logit Model ………………………………………………………………………………………………………………………… 30

Fig 18: ROC Curve Random Forest Model ………………………………………………………………………………………………………… 31

Fig 19: ROC Curve LDA Model …………………………………………………………………………………………………………………………. 31

Fig 20: Data head Stock data ……………………………………………………………………………………………………………………………. 33

Fig 21: Data Info Stock data ……………………………………………………………………………………………………………………………… 33

Fig 22: Data Description Stock data …………………………………………………………………………………………………………………. 34

Fig 23: Price vs Time Plot 1 ………………………………………………………………………………………………………………………………. 34

Fig 24: Price vs Time Plot 2 ………………………………………………………………………………………………………………………………. 35

Fig 25: Logarithmic Returns …………………………………………………………………………………………………………………………….. 36

Fig 26: Stock Means …………………………………………………………………………………………………………………………………………. 36

Fig 27: Stock Standard Deviation …………………………………………………………………………………………………………………….. 37

Fig 28: Stock Mean vs Standard Deviation Plot ……………………………………………………………………………………………….. 37

Fig 29: Best Stocks Top to Bottom …………………………………………………………………………………………………………………… 39

# List of Tables

Table 1: Credit Rist Data Dictionary ………………………………………………………………………………………………………………….. 4

Table 2: Credit Rist 5 point summary ………………………………………………………………………………………………………………. 10

Table 3: Sum of nulls after Missing value treatment ………………………………………………………………………………………… 18

# PART A:

**Problem Statement:**

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interest on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

# Data Shape and Dictionary:

Given Credit Risk data set has 2058 entries with total 58 features. Data dictionary for each of these features is as follows.

Table 1: Credit Rist Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Column Name** | **Description** |
| 1 | Co\_Code | Company Code |
| 2 | Co\_Name | Company Name |
| 3 | \_Operating\_Expense\_Rate | Operating Expense Rate: Operating Expenses/Net Sales. The operating expense ratio (OER) is the cost to operate a piece of property compared to the income the property brings in. |
| 4 | \_Research\_and\_development\_expense\_rate | Research and development expense rate: (Research and Development Expenses)/Net Sales. Research and development (R&D) expenses are direct expenditures relating to a company's efforts to develop, design, and enhance its products, services, technologies, or processes. |
| 5 | \_Cash\_flow\_rate | Cash flow rate: Cash Flow from Operating/Current Liabilities. Cash flow is a measure of how much cash a business brought in or spent in total over a period of time. |
| 6 | \_Interest\_bearing\_debt\_interest\_rate | Interest-bearing debt interest rate: Interest-bearing Debt/Equity |
| 7 | \_Tax\_rate\_A | Tax rate (A): Effective Tax Rate. Effective tax rate represents the percentage of their taxable income that individuals pay in taxes. For corporations, the effective corporate tax rate is the rate they pay on their pre-tax profits. |
| 8 | \_Cash\_Flow\_Per\_Share | Cash Flow Per Share. It is the after-tax earnings plus depreciation on a per-share basis that functions as a measure of a firm's financial strength |
| 9 | \_Per\_Share\_Net\_profit\_before\_tax\_Yuan\_ | Per Share Net profit before tax (Yuan ¥): Pretax Income Per Share. Pretax income, also known as earnings before tax or pretax earnings, is the net income earned by a business before taxes are subtracted/accounted for. |
| 10 | \_Realized\_Sales\_Gross\_Profit\_Growth\_Rate | Realized Sales Gross Profit Growth Rate. |
| 11 | \_Operating\_Profit\_Growth\_Rate | Operating Profit Growth Rate: Operating Income Growth. It is the rate of increase in operating income over the last year. |
| 12 | \_Continuous\_Net\_Profit\_Growth\_Rate | Continuous Net Profit Growth Rate: Net Income-Excluding Disposal Gain or Loss Growth |
| 13 | \_Total\_Asset\_Growth\_Rate | Total Asset Growth Rate: Total Asset Growth. It is the rate at which how quickly the company has been growing its Assets |
| 14 | \_Net\_Value\_Growth\_Rate | Net Value Growth Rate: Total Equity Growth |
| 15 | \_Total\_Asset\_Return\_Growth\_Rate\_Ratio | Total Asset Return Growth Rate Ratio: Return on Total Asset Growth |
| 16 | \_Cash\_Reinvestment\_perc | Cash Reinvestment %: Cash Reinvestment Ratio. It is the valuation ratio that is used to measure the percentage of annual cash flow that the company invests back into the business as a new investment. |
| 17 | \_Current\_Ratio | Current Ratio. The current ratio describes the relationship between a company's assets and liabilities |
| 18 | \_Quick\_Ratio | Quick Ratio: Acid Test. Acid-test ratio (also known as quick ratio) is a measure of a company's liquidity, which is its ability to pay its short-term obligations using only its most liquid assets. |
| 19 | \_Interest\_Expense\_Ratio | Interest Expense Ratio: Interest Expenses/Total Revenue |
| 20 | \_Total\_debt\_to\_Total\_net\_worth | Total debt/Total net worth: Total Liability/Equity Ratio |
| 21 | \_Long\_term\_fund\_suitability\_ratio\_A | Long-term fund suitability ratio (A): (Long-term Liability+Equity)/Fixed Assets |
| 22 | \_Net\_profit\_before\_tax\_to\_Paid\_in\_capital | Net profit before tax/Paid-in capital: Pretax Income/Capital |
| 23 | \_Total\_Asset\_Turnover | Total Asset Turnover. Net Sales/Average Total Assets |
| 24 | \_Accounts\_Receivable\_Turnover | Accounts Receivable Turnover. The accounts receivable turnover ratio, or receivables turnover, is used in business accounting to quantify how well companies are managing the credit that they extend to their customers by evaluating how long it takes to collect the outstanding debt throughout the accounting period. |
| 25 | \_Average\_Collection\_Days | Average Collection Days: Days Receivable Outstanding |
| 26 | \_Inventory\_Turnover\_Rate\_times | Inventory Turnover Rate (times). The inventory turnover ratio is the number of times a company has sold and replenished its inventory over a specific amount of time. The formula can also be used to calculate the number of days it will take to sell the inventory on hand. |
| 27 | \_Fixed\_Assets\_Turnover\_Frequency | Fixed Assets Turnover Frequency. Fixed Asset Turnover (FAT) is an efficiency ratio that indicates how well or efficiently a business uses fixed assets to generate sales. This ratio divides net sales by net fixed assets, calculated over an annual period. |
| 28 | \_Net\_Worth\_Turnover\_Rate\_times | Net Worth Turnover Rate (times): Equity Turnover. Equity turnover is a ratio that measures the proportion of a company's sales to its stockholders' equity. The intent of the measurement is to determine the efficiency with which management is using equity to generate revenue. |
| 29 | \_Operating\_profit\_per\_person | Operating profit per person: Operation Income Per Employee |
| 30 | \_Allocation\_rate\_per\_person | Allocation rate per person: Fixed Assets Per Employee |
| 31 | \_Quick\_Assets\_to\_Total\_Assets | Quick Assets/Total Assets |
| 32 | \_Cash\_to\_Total\_Assets | Cash/Total Assets |
| 33 | \_Quick\_Assets\_to\_Current\_Liability | Quick Assets/Current Liability |
| 34 | \_Cash\_to\_Current\_Liability | Cash/Current Liability |
| 35 | \_Operating\_Funds\_to\_Liability | Operating Funds to Liability |
| 36 | \_Inventory\_to\_Working\_Capital | Inventory/Working Capital |
| 37 | \_Inventory\_to\_Current\_Liability | Inventory/Current Liability |
| 38 | \_Long\_term\_Liability\_to\_Current\_Assets | Long-term Liability to Current Assets |
| 39 | \_Retained\_Earnings\_to\_Total\_Assets | Retained Earnings to Total Assets |
| 40 | \_Total\_income\_to\_Total\_expense | Total income/Total expense |
| 41 | \_Total\_expense\_to\_Assets | Total expense/Assets |
| 42 | \_Current\_Asset\_Turnover\_Rate | Current Asset Turnover Rate: Current Assets to Sales. The current assets turnover ratio indicates how many times the current assets are turned over in the form of sales within a specific period of time. A higher asset turnover ratio means a better percentage of sales. |
| 43 | \_Quick\_Asset\_Turnover\_Rate | Quick Asset Turnover Rate: Quick Assets to Sales. The asset turnover ratio measures the efficiency of a company's assets in generating revenue or sales. |
| 44 | \_Cash\_Turnover\_Rate | Cash Turnover Rate: Cash to Sales. The cash turnover ratio is an efficiency ratio that reveals the number of times that cash is turned over in an accounting period. |
| 45 | \_Fixed\_Assets\_to\_Assets | Fixed Assets to Assets. Fixed assets are also known as non-current assets—assets that can't be easily converted into cash. |
| 46 | \_Cash\_Flow\_to\_Total\_Assets | Cash Flow to Total Assets. This ratio indicates the cash a company can generate in relation to its size. |
| 47 | \_Cash\_Flow\_to\_Liability | Cash Flow to Liability. The amount of money available to run business operations and complete transactions. This is calculated as current assets (cash or near-cash assets, like notes receivable) minus current liabilities (liabilities due during the upcoming accounting period) |
| 48 | \_CFO\_to\_Assets | CFO to Assets. Cash flow on total assets is an efficiency ratio that rates cash flows to the company assets without being affected by income recognition or income measurements. |
| 49 | \_Cash\_Flow\_to\_Equity | Cash Flow to Equity. cash flow to equity is a measure of how much cash is available to the equity shareholders of a company after all expenses, reinvestment, and debt are paid. |
| 50 | \_Current\_Liability\_to\_Current\_Assets | Current Liability to Current Assets. Current liabilities are a company's financial commitments that are due and payable within a year, Current assets are projected to be consumed, sold, or converted into cash within a year or within the operational cycle. |
| 51 | \_Liability\_Assets\_Flag | Liability-Assets Flag: 1 if Total Liability exceeds Total Assets, 0 otherwise |
| 52 | \_Total\_assets\_to\_GNP\_price | Total assets to GNP price. Gross National Product (GNP) is the total value of all finished goods and services produced by a country’s citizens in a given financial year, irrespective of their location. |
| 53 | \_No\_credit\_Interval | No-credit Interval |
| 54 | \_Degree\_of\_Financial\_Leverage\_DFL | Degree of Financial Leverage (DFL). The degree of financial leverage is a financial ratio that measures the sensitivity in fluctuations of a company's overall profitability to the volatility of its operating income caused by changes in its capital structure. |
| 55 | \_Interest\_Coverage\_Ratio\_Interest\_expense\_to\_EBIT | Interest Coverage Ratio (Interest expense to EBIT). The interest coverage ratio is a debt and profitability ratio used to determine how easily a company can pay interest on its outstanding debt. The interest coverage ratio is calculated by dividing a company's earnings before interest and taxes (EBIT) by its interest expense during a given period. |
| 56 | \_Net\_Income\_Flag | Net Income Flag: 1 if Net Income is Negative for the last two years, 0 otherwise |
| 57 | \_Equity\_to\_Liability | Equity to Liability Ratio. |
| 58 | Default | Whether the Company has Default (Bankrupted) or not? 1 - Defaulted, 0 - Not Defaulted. |

# Data Info:

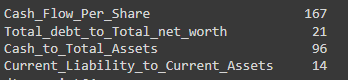
All the features of of int/float data types. Default is out target feature with values 0 and 1. 1 being identified as defaulters and 0 being non defaulters.

Given data set has data related to 1838 non defaulters and 220 defaulters. i.,e we only have 10% of records that give information about defaulters. Hence we call this imbalanced data set which needs to be handled using balancing techniques like SMOTE.

# Null Check:

From the 58 features, below features have nulls/missing data.

Fig 1: Credit Risk Data Missing Values



When we convert above null entries in each column to percentages, it will be too low, but still as part of data preparation, we will impute those nulls going further.

# Duplicate Check:

No Duplicates were found in the dataset.

# Data Description:

Five point summary of all the features is as follows,

From the below summary, it is evident that there are few features with outliers as the mean and median are not in sync.

Table 2: Credit Rist 5 point summary

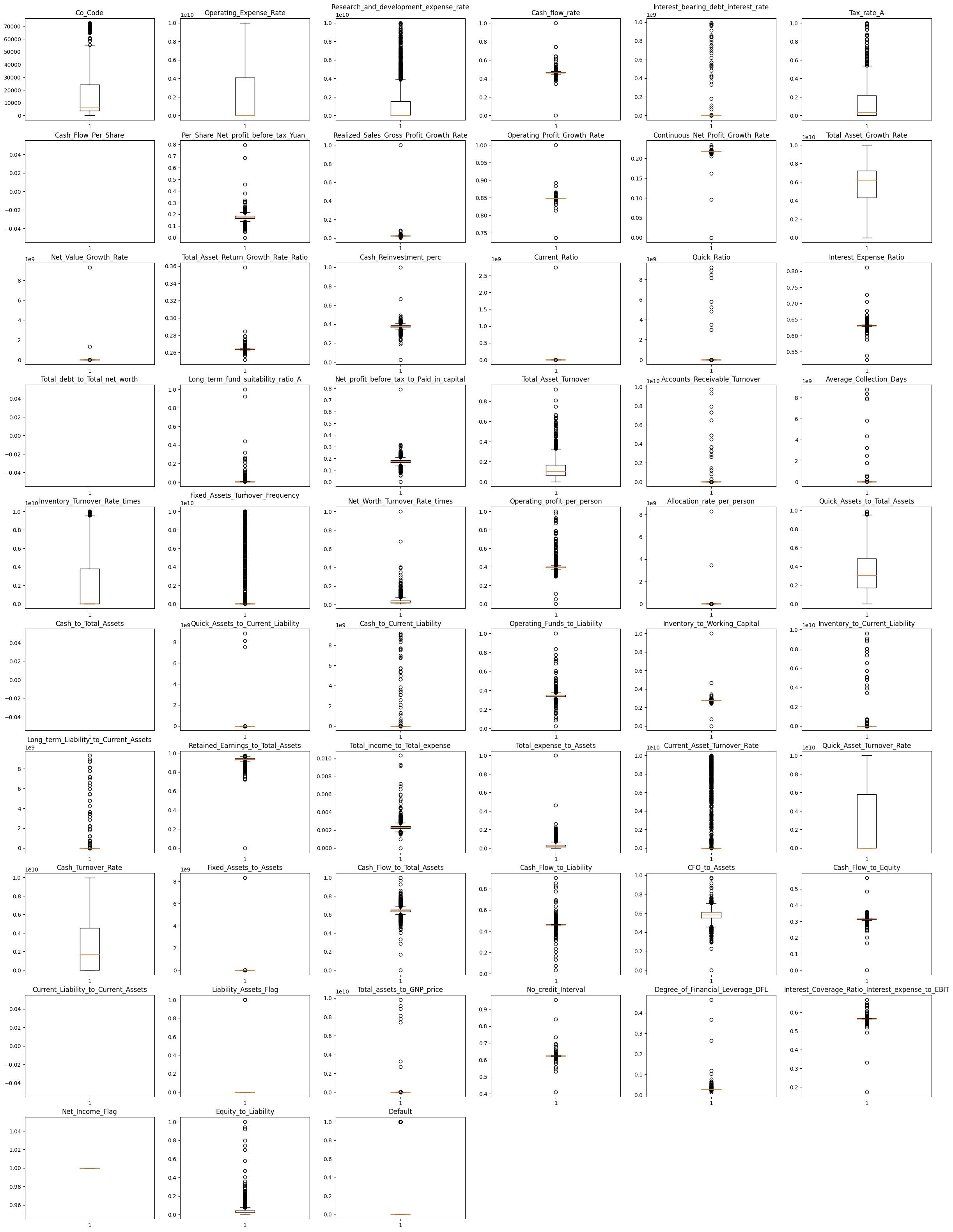
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| Co\_Code | 2058 | 1.76E+04 | 2.19E+04 | 4 | 3.67E+03 | 6.24E+03 | 2.43E+04 | 7.25E+04 |
| Operating\_Expense\_Rate | 2058 | 2.05E+09 | 3.25E+09 | 0 | 0.00E+00 | 0.00E+00 | 4.11E+09 | 9.98E+09 |
| Research\_and\_development\_expense\_rate | 2058 | 1.21E+09 | 2.14E+09 | 0 | 0.00E+00 | 0.00E+00 | 1.55E+09 | 9.98E+09 |
| Cash\_flow\_rate | 2058 | 4.70E-01 | 2.00E-02 | 0 | 4.60E-01 | 4.60E-01 | 4.70E-01 | 1.00E+00 |
| Interest\_bearing\_debt\_interest\_rate | 2058 | 1.11E+07 | 9.04E+07 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 9.90E+08 |
| Tax\_rate\_A | 2058 | 1.10E-01 | 1.50E-01 | 0 | 0.00E+00 | 4.00E-02 | 2.20E-01 | 1.00E+00 |
| Cash\_Flow\_Per\_Share | 1891 | 3.20E-01 | 2.00E-02 | 0.17 | 3.10E-01 | 3.20E-01 | 3.30E-01 | 4.60E-01 |
| Per\_Share\_Net\_profit\_before\_tax\_Yuan\_ | 2058 | 1.80E-01 | 3.00E-02 | 0 | 1.70E-01 | 1.80E-01 | 1.90E-01 | 7.90E-01 |
| Realized\_Sales\_Gross\_Profit\_Growth\_Rate | 2058 | 2.00E-02 | 2.00E-02 | 0 | 2.00E-02 | 2.00E-02 | 2.00E-02 | 1.00E+00 |
| Operating\_Profit\_Growth\_Rate | 2058 | 8.50E-01 | 0.00E+00 | 0.74 | 8.50E-01 | 8.50E-01 | 8.50E-01 | 1.00E+00 |
| Continuous\_Net\_Profit\_Growth\_Rate | 2058 | 2.20E-01 | 1.00E-02 | 0 | 2.20E-01 | 2.20E-01 | 2.20E-01 | 2.30E-01 |
| Total\_Asset\_Growth\_Rate | 2058 | 5.29E+09 | 2.91E+09 | 0 | 4.32E+09 | 6.23E+09 | 7.22E+09 | 9.98E+09 |
| Net\_Value\_Growth\_Rate | 2058 | 5.19E+06 | 2.08E+08 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 9.33E+09 |
| Total\_Asset\_Return\_Growth\_Rate\_Ratio | 2058 | 2.60E-01 | 0.00E+00 | 0.25 | 2.60E-01 | 2.60E-01 | 2.60E-01 | 3.60E-01 |
| Cash\_Reinvestment\_perc | 2058 | 3.80E-01 | 3.00E-02 | 0.03 | 3.70E-01 | 3.80E-01 | 3.90E-01 | 1.00E+00 |
| Current\_Ratio | 2058 | 1.34E+06 | 6.06E+07 | 0 | 1.00E-02 | 1.00E-02 | 1.00E-02 | 2.75E+09 |
| Quick\_Ratio | 2058 | 2.78E+07 | 4.45E+08 | 0 | 0.00E+00 | 1.00E-02 | 1.00E-02 | 9.23E+09 |
| Interest\_Expense\_Ratio | 2058 | 6.30E-01 | 1.00E-02 | 0.53 | 6.30E-01 | 6.30E-01 | 6.30E-01 | 8.10E-01 |
| Total\_debt\_to\_Total\_net\_worth | 2037 | 1.07E+07 | 2.70E+08 | 0 | 0.00E+00 | 1.00E-02 | 1.00E-02 | 9.94E+09 |
| Long\_term\_fund\_suitability\_ratio\_A | 2058 | 1.00E-02 | 3.00E-02 | 0 | 1.00E-02 | 1.00E-02 | 1.00E-02 | 1.00E+00 |
| Net\_profit\_before\_tax\_to\_Paid\_in\_capital | 2058 | 1.80E-01 | 3.00E-02 | 0 | 1.70E-01 | 1.70E-01 | 1.80E-01 | 7.90E-01 |
| Total\_Asset\_Turnover | 2058 | 1.30E-01 | 1.00E-01 | 0 | 6.00E-02 | 1.00E-01 | 1.70E-01 | 9.20E-01 |
| Accounts\_Receivable\_Turnover | 2058 | 4.16E+07 | 5.05E+08 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 9.74E+09 |
| Average\_Collection\_Days | 2058 | 2.63E+07 | 4.11E+08 | 0 | 0.00E+00 | 1.00E-02 | 1.00E-02 | 8.80E+09 |
| Inventory\_Turnover\_Rate\_times | 2058 | 2.03E+09 | 3.08E+09 | 0 | 0.00E+00 | 1.91E+07 | 3.82E+09 | 9.99E+09 |
| Fixed\_Assets\_Turnover\_Frequency | 2058 | 1.23E+09 | 2.65E+09 | 0 | 0.00E+00 | 0.00E+00 | 1.00E-02 | 9.99E+09 |
| Net\_Worth\_Turnover\_Rate\_times | 2058 | 4.00E-02 | 4.00E-02 | 0.01 | 2.00E-02 | 3.00E-02 | 4.00E-02 | 1.00E+00 |
| Operating\_profit\_per\_person | 2058 | 4.00E-01 | 5.00E-02 | 0 | 3.90E-01 | 4.00E-01 | 4.00E-01 | 1.00E+00 |
| Allocation\_rate\_per\_person | 2058 | 5.73E+06 | 1.98E+08 | 0 | 0.00E+00 | 1.00E-02 | 2.00E-02 | 8.28E+09 |
| Quick\_Assets\_to\_Total\_Assets | 2058 | 3.40E-01 | 2.10E-01 | 0 | 1.70E-01 | 3.10E-01 | 4.80E-01 | 9.90E-01 |
| Cash\_to\_Total\_Assets | 1962 | 8.00E-02 | 1.00E-01 | 0 | 2.00E-02 | 5.00E-02 | 1.00E-01 | 9.30E-01 |
| Quick\_Assets\_to\_Current\_Liability | 2058 | 1.19E+07 | 3.12E+08 | 0 | 0.00E+00 | 1.00E-02 | 1.00E-02 | 8.82E+09 |
| Cash\_to\_Current\_Liability | 2058 | 9.28E+07 | 7.85E+08 | 0 | 0.00E+00 | 0.00E+00 | 1.00E-02 | 9.17E+09 |
| Operating\_Funds\_to\_Liability | 2058 | 3.50E-01 | 4.00E-02 | 0.03 | 3.40E-01 | 3.50E-01 | 3.50E-01 | 1.00E+00 |
| Inventory\_to\_Working\_Capital | 2058 | 2.80E-01 | 2.00E-02 | 0 | 2.80E-01 | 2.80E-01 | 2.80E-01 | 1.00E+00 |
| Inventory\_to\_Current\_Liability | 2058 | 5.79E+07 | 6.28E+08 | 0 | 0.00E+00 | 1.00E-02 | 1.00E-02 | 9.60E+09 |
| Long\_term\_Liability\_to\_Current\_Assets | 2058 | 7.34E+07 | 6.69E+08 | 0 | 0.00E+00 | 0.00E+00 | 1.00E-02 | 9.31E+09 |
| Retained\_Earnings\_to\_Total\_Assets | 2058 | 9.30E-01 | 3.00E-02 | 0 | 9.30E-01 | 9.40E-01 | 9.40E-01 | 9.70E-01 |
| Total\_income\_to\_Total\_expense | 2058 | 0.00E+00 | 0.00E+00 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.00E-02 |
| Total\_expense\_to\_Assets | 2058 | 3.00E-02 | 4.00E-02 | 0 | 1.00E-02 | 2.00E-02 | 4.00E-02 | 1.00E+00 |
| Current\_Asset\_Turnover\_Rate | 2058 | 1.27E+09 | 2.84E+09 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 9.99E+09 |
| Quick\_Asset\_Turnover\_Rate | 2058 | 2.57E+09 | 3.45E+09 | 0 | 0.00E+00 | 0.00E+00 | 5.79E+09 | 1.00E+10 |
| Cash\_Turnover\_Rate | 2058 | 2.65E+09 | 2.82E+09 | 0 | 0.00E+00 | 1.73E+09 | 4.55E+09 | 9.99E+09 |
| Fixed\_Assets\_to\_Assets | 2058 | 4.04E+06 | 1.83E+08 | 0 | 1.00E-01 | 2.10E-01 | 4.20E-01 | 8.32E+09 |
| Cash\_Flow\_to\_Total\_Assets | 2058 | 6.40E-01 | 5.00E-02 | 0 | 6.30E-01 | 6.40E-01 | 6.50E-01 | 1.00E+00 |
| Cash\_Flow\_to\_Liability | 2058 | 4.60E-01 | 3.00E-02 | 0.03 | 4.60E-01 | 4.60E-01 | 4.60E-01 | 9.10E-01 |
| CFO\_to\_Assets | 2058 | 5.80E-01 | 6.00E-02 | 0 | 5.50E-01 | 5.80E-01 | 6.10E-01 | 9.80E-01 |
| Cash\_Flow\_to\_Equity | 2058 | 3.10E-01 | 1.00E-02 | 0 | 3.10E-01 | 3.10E-01 | 3.20E-01 | 5.70E-01 |
| Current\_Liability\_to\_Current\_Assets | 2044 | 4.00E-02 | 5.00E-02 | 0 | 2.00E-02 | 3.00E-02 | 4.00E-02 | 1.00E+00 |
| Liability\_Assets\_Flag | 2058 | 0.00E+00 | 6.00E-02 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.00E+00 |
| Total\_assets\_to\_GNP\_price | 2058 | 2.78E+07 | 4.72E+08 | 0 | 0.00E+00 | 0.00E+00 | 1.00E-02 | 9.82E+09 |
| No\_credit\_Interval | 2058 | 6.20E-01 | 1.00E-02 | 0.41 | 6.20E-01 | 6.20E-01 | 6.20E-01 | 9.60E-01 |
| Degree\_of\_Financial\_Leverage\_DFL | 2058 | 3.00E-02 | 1.00E-02 | 0.01 | 3.00E-02 | 3.00E-02 | 3.00E-02 | 4.60E-01 |
| Interest\_Coverage\_Ratio\_Interest\_expense\_to\_EBIT | 2058 | 5.70E-01 | 1.00E-02 | 0.17 | 5.70E-01 | 5.70E-01 | 5.70E-01 | 6.70E-01 |
| Net\_Income\_Flag | 2058 | 1.00E+00 | 0.00E+00 | 1 | 1.00E+00 | 1.00E+00 | 1.00E+00 | 1.00E+00 |
| Equity\_to\_Liability | 2058 | 4.00E-02 | 6.00E-02 | 0 | 2.00E-02 | 3.00E-02 | 4.00E-02 | 1.00E+00 |
| Default | 2058 | 1.10E-01 | 3.10E-01 | 0 | 0.00E+00 | 0.00E+00 | 0.00E+00 | 1.00E+00 |

# Outlier Check and Treatment:

As seen below, most of the features have outliers in them which can be seen as dots.

Outlier treatment is necessary as they will impact our model performance and accuracy.

Fig 2: Credit Rist Box Plots(Before Outlier Treatment)



Here we have treated outliers using inter quartile method.

Inter Quantile Range(IQR)

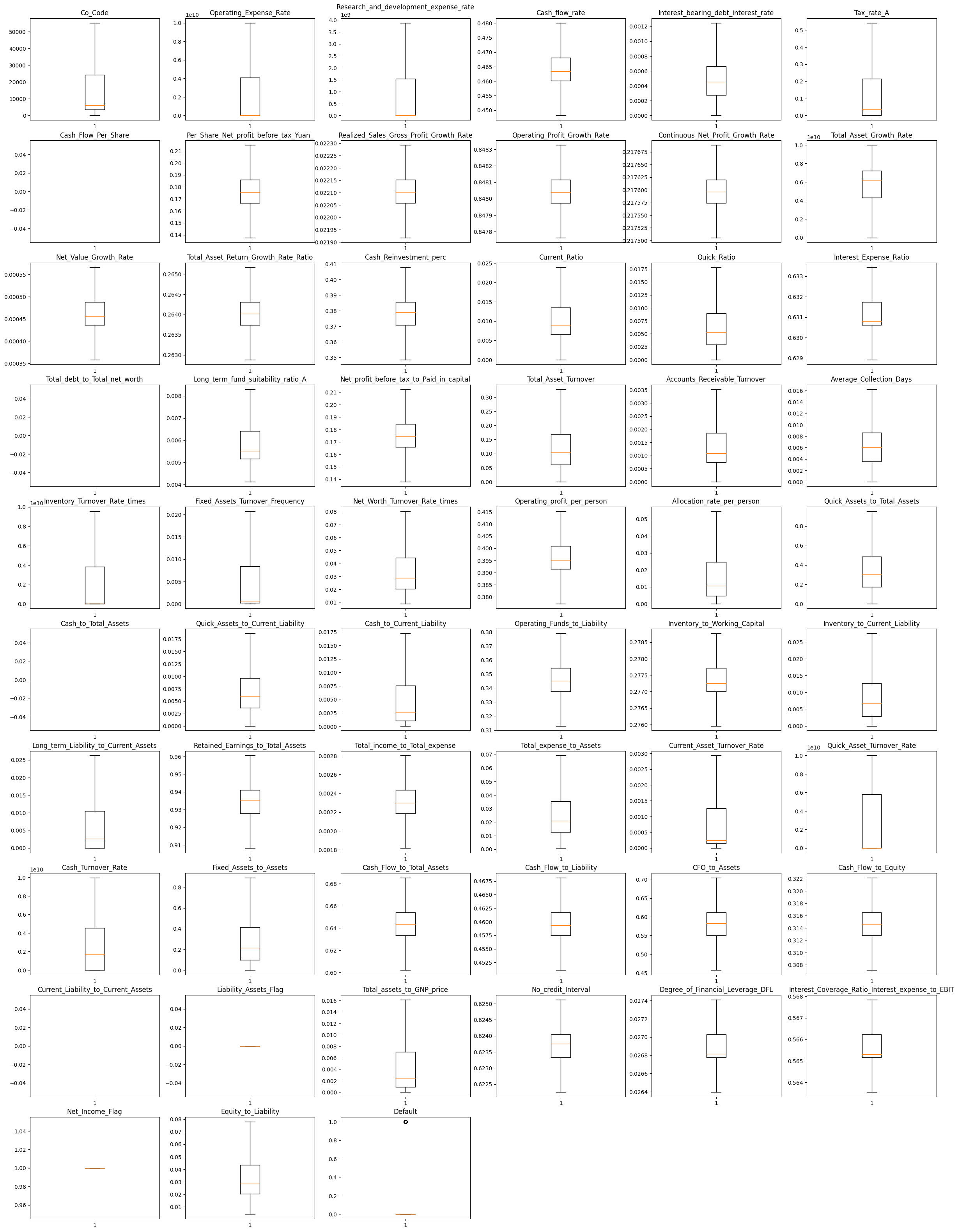
Criteria: data points that lie 1.5 times of IQR above Q3 and below Q1 are outliers. This shows in detail about outlier treatment in Python.

Steps:

* Sort the dataset in ascending order
* calculate the 1st and 3rd quartiles(Q1, Q3)
* compute IQR=Q3-Q1
* compute lower bound = (Q1–1.5\*IQR), upper bound = (Q3+1.5\*IQR)
* loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outliers

Below is the box plot matrix for all the features after outlier treatment,

Fig 3: Credit Rist Box Plots(After Outlier Treatment)



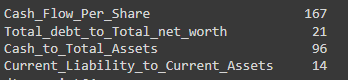
Thus, we have removed in all the features except our target feature i.e., Default

Also we can see few features being shown as empty in above matrix, that is due to null values in those features. This will be fixed after removing null values.

# Missing value Treatment:

We have imputed the nulls in below features with their respective means.

Fig 4: Credit Risk(Missing Values Before Treatment)



Here is the sum of nulls in above features after removing nulls.

Table 3: Sum of nulls after Missing value treatment

Cash\_Flow\_Per\_Share 0

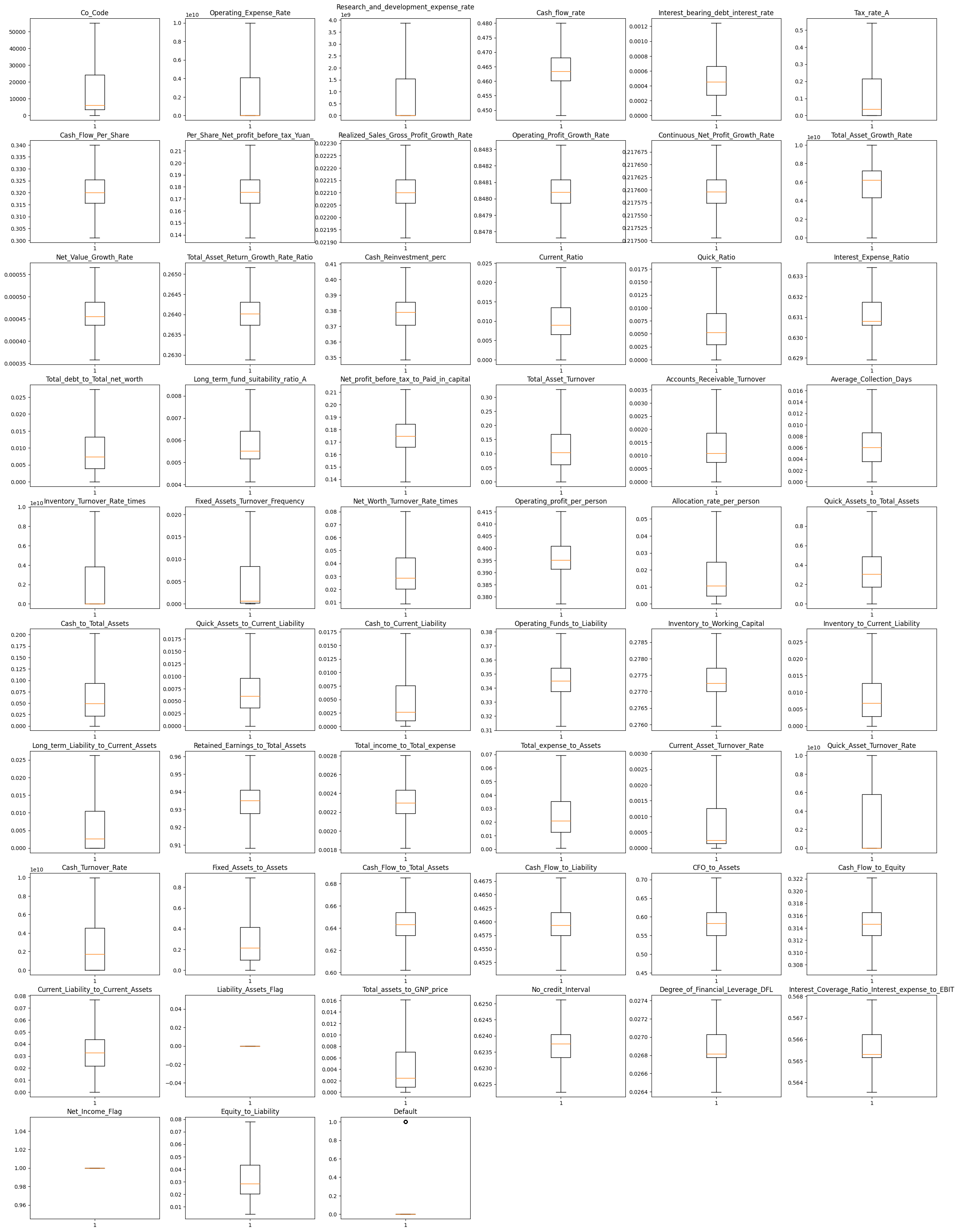
Total\_debt\_to\_Total\_net\_worth 0

Cash\_to\_Total\_Assets 0

Current\_Liability\_to\_Current\_Assets 0

Now as we have removed nulls, we can do outlier treatment on these columns, after treating outliers in these columns, box plots are as follows,

Fig 5: Credit Risk(Box Plots after Outlier and Missing Value Treatment)



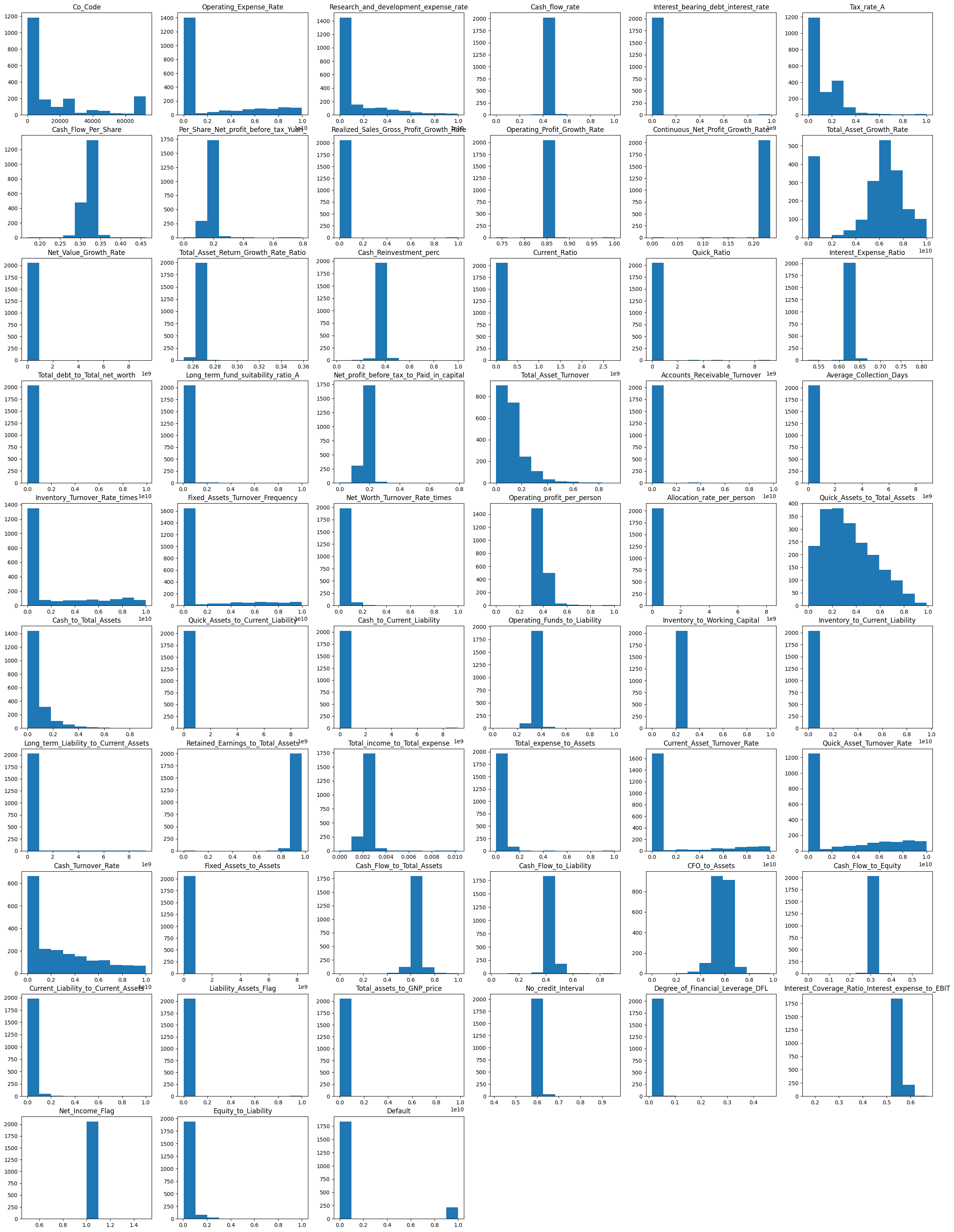
# Univariate & Bivariate analysis:

Apart from box plots which we discussed above, we can also visualize distributions using Histplot.

Histograms of all the features is as follows,

We can see that none of the feature follow any distribution pattern. Most of them have the data skewed towards right, with exceptions in Total\_Asset\_Growth\_Rate,Cash\_Flow\_Rate, Operating\_Profit\_Growth\_Rate, Continuous\_Net\_Profit\_Growth\_Rate, Cash\_Flow\_To\_Total\_Assets, Cash\_Flow\_to\_Equity and Net\_Income\_Flag.

Fig 6: Credit Risk(Histograms)



Heatmap for all the features is as follows,

Heatmap is nothing but the visual representation of correlation matrix.

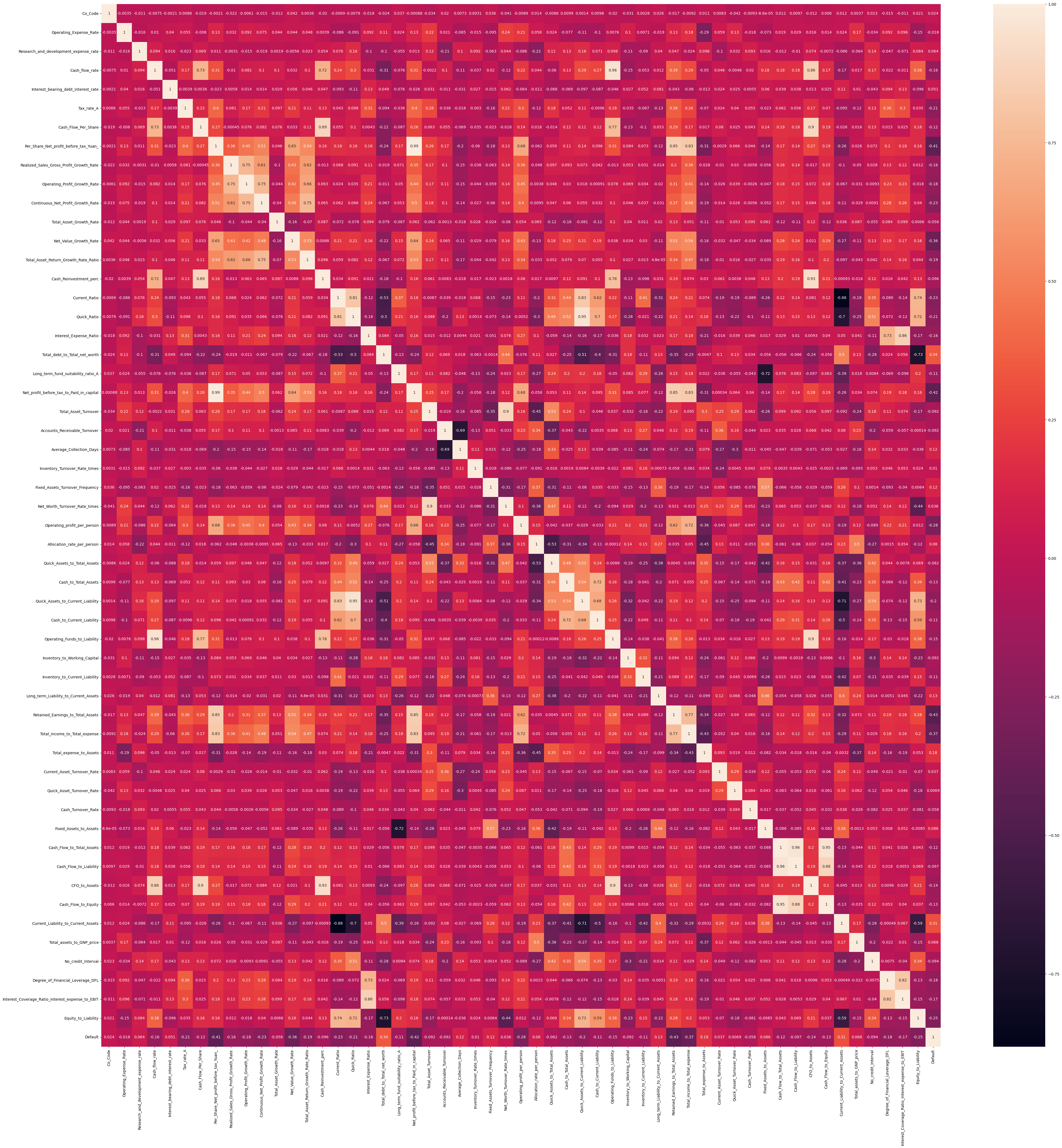
The value of the correlation coefficient can take any values from -1 to 1.

If the value is 1, it is said to be a positive correlation between two variables. This means that when one variable increases, the other variable also increases.

If the value is -1, it is said to be a negative correlation between the two variables. This means that when one variable increases, the other variable decreases.

If the value is 0, there is no correlation between the two variables. This means that the variables changes in a random manner with respect to each other.

Fig 7: Credit Risk(Heatmap)



In the above chart, the more darker the block is, the more negatively correlated are those respective features and vice-versa.

For instance, (Cash\_Reinvestment\_perc - Cash\_Flow\_Per\_Share) and (Operating\_Funds\_to\_Liability - Cash\_flow\_rate) have high correlation coefficients with values 0.89 and 0.96 respectively.

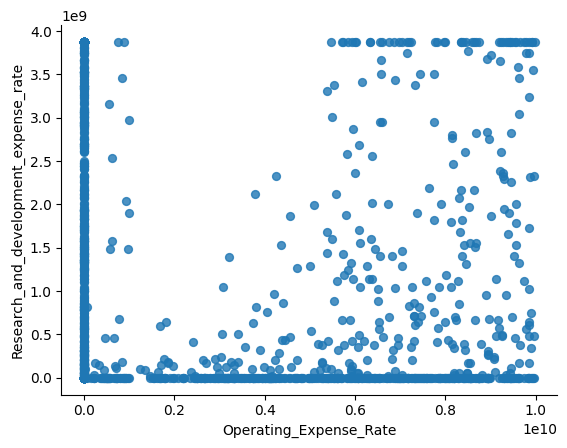
Similary, (Current\_Ratio - Current\_Liablity\_t0\_Current\_Assets) and

(Current\_Liablity\_t0\_Current\_Assets - Quick\_Assets\_to\_Current\_Liability) are highly negatively correlated with coefficients of -0.88 and -0.71 respectively.

**Scatter Plot:**

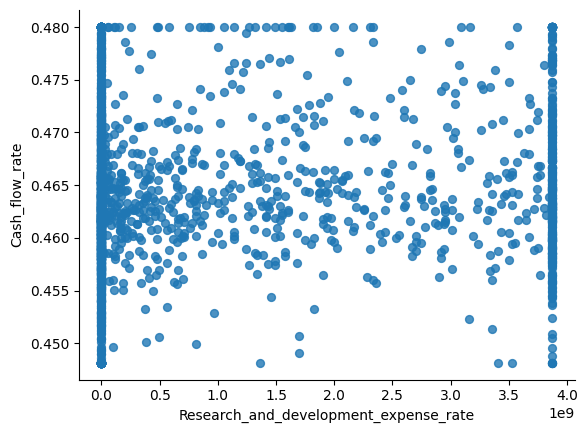
Scatter Plot between Research\_and\_development\_expense\_rate and Operating\_Expense\_Rate shows more values being concentrated near 0 for one feature and values being spread across in other feature.

Fig 8: Scatter Plot 1



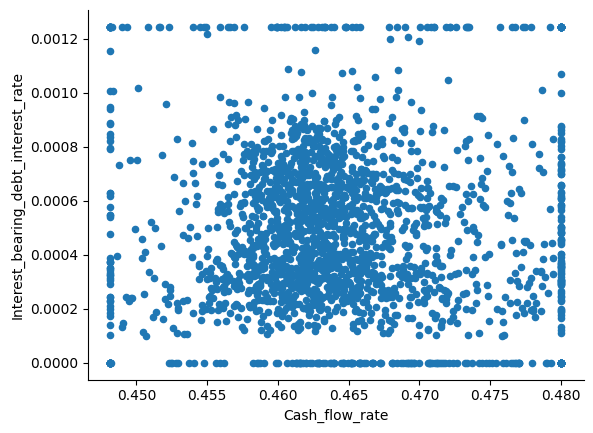
Similarly, scatter plot between Cash\_flow\_rate and Research\_and\_development\_expense\_rate shows Cash\_flow\_rate is slightly high near lower values of Research\_and\_development\_expense\_rate.

Fig 9: Scatter Plot 2



Scatter plot between Interest\_bearing\_debt\_interest\_rate and Cash\_flow\_rate shows more values are being concentrated towards center.

Fig 10: Scatter Plot 3



# Train Test Split:

As we have total 2058 rows with 58 features, now we divide features into independent and dependent sets.

We have split original dataset into X(independent variables) and Y(dependent variables).

FYI, we have removed column “Co\_Name” from the list as it doesn’t add any value to the prediction.

Now “X” will have all features except “Co\_Name” and “Default”.

“Y” contains only “Default”

After splitting data to independent and dependent sets, now we split data into train and test sets in 67:33 ratio.

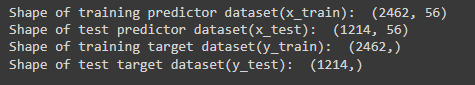
Also, as already mentioned, the target variable data is not balanced as it has only 10% of data related to defaulters,

Hence we apply SMOTE oversampling technique here to balance data,

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

Below are the counts of train and test data sets after applying SMOTE oversampling

Fig 11: Train Test Split(Data Shape)



# Logistic Regression(From Statsmodel):

Now that we have 56 independent variable, all of them may not be useful when it comes to model building and prediction. Hence we have used RFE(Recursive Feature Elimination) technique for our feature selection.

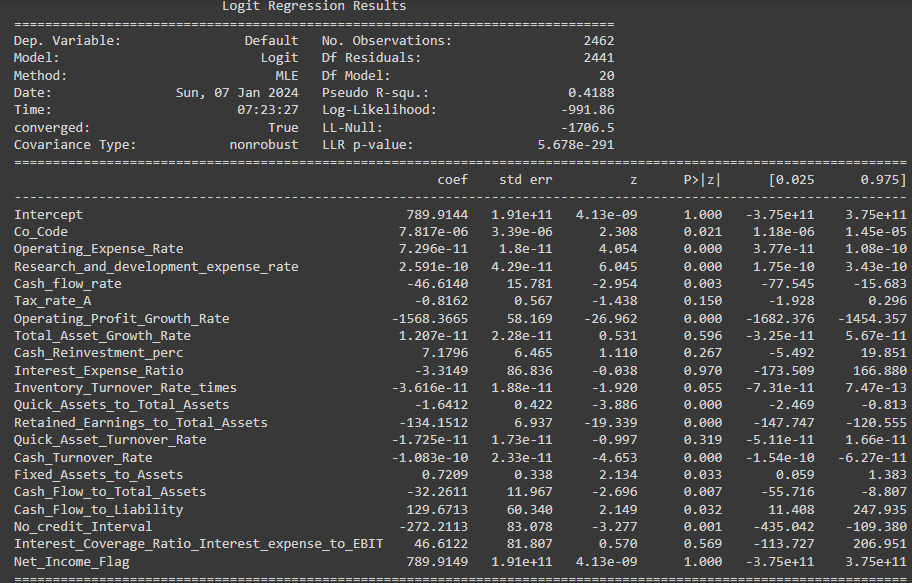
Feature selection refers to techniques that select a subset of the most relevant features (columns) for a dataset. Fewer features can allow machine learning algorithms to run more efficiently (less space or time complexity) and be more effective. Some machine learning algorithms can be misled by irrelevant input features, resulting in worse predictive performance.

After this method, we have come up with 20 best features, which are as follows

['Co\_Code', 'Operating\_Expense\_Rate', 'Research\_and\_development\_expense\_rate', 'Cash\_flow\_rate', 'Tax\_rate\_A', 'Operating\_Profit\_Growth\_Rate', 'Total\_Asset\_Growth\_Rate', 'Cash\_Reinvestment\_perc', 'Interest\_Expense\_Ratio', 'Inventory\_Turnover\_Rate\_times', 'Quick\_Assets\_to\_Total\_Assets', 'Retained\_Earnings\_to\_Total\_Assets', 'Quick\_Asset\_Turnover\_Rate', 'Cash\_Turnover\_Rate', 'Fixed\_Assets\_to\_Assets', 'Cash\_Flow\_to\_Total\_Assets', 'Cash\_Flow\_to\_Liability', 'No\_credit\_Interval', 'Interest\_Coverage\_Ratio\_Interest\_expense\_to\_EBIT', 'Net\_Income\_Flag']

Upon fitting the model on train data using above features, these are the coefficients we get from Logit Model,

Fig 12: Logit Model Summary



Interpretations:

Pseudo R-squ. : a substitute for the R-squared value in Least Squares linear regression. It is the ratio of the log-likelihood of the null model to that of the full model. This value ranges from 0 to 1. 0 being the poor model and 1 being the best model. In our case, we got this value around 0.41 which is pretty decent.

LLR p-value: Null hypothesis for this model says that the model is no significant in making predictions and alternate hypotheses says the model is significant and efficient in predicting our dependent variable. If p value is less than 0.05 then we can say this model is significant and useful. In our case its less than significant value 0.05 and hence we can reject null hypothesis.

# (Logistic Regression)Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model:

Below is the Classification report of Logistic Regression model on test data,

1. Precision: Percentage of correct positive predictions relative to total positive predictions.

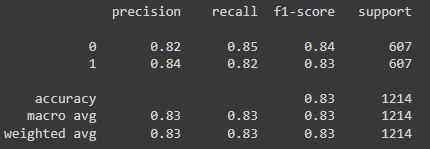
2. Recall: Percentage of correct positive predictions relative to total actual positives.

3. F1 Score: A weighted harmonic mean of precision and recall. The closer to 1, the better the model.

F1 Score: 2 \* (Precision \* Recall) / (Precision + Recall)

Using these three metrics, we can understand how well a given classification model is able to predict the outcomes for some response variable.

Fig 13: Classification Report(Logit)



In this particular case we are more interested in predicting defaulters i.e., 1’s. Hence recall would be correct parameter to rely on.

Hence our Logistic Regression model can predict 83% of total defaulters.

# Build a Random Forest Model on Train Dataset. Also showcase your model building approach:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

We can use GridSearchCV algorithm to find the best possible parameters to be used for our train data in Random Forest model.

Upon Passing Random forest model to GridSearchCV algorithm with our train data, below are the best parameters suggested, Hence the model has been built on the same hyper parameters.

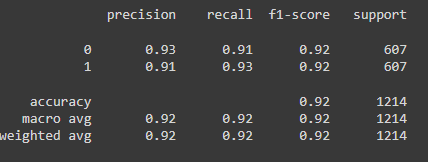
{'max\_depth': 7, 'min\_samples\_leaf': 5, 'min\_samples\_split': 15, 'n\_estimators': 50}

# Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model:

Classification report for our test data using Random Forest Model is as follows,

We can see recall value has been improved a lot when compared to Logistic regression model. Recall in this case can be interpreted as, Our model can predict 93% of total defaulters.

Fig 14: Classification Report(Random Forest)



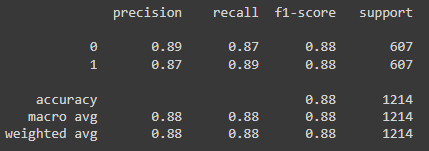
# Build a LDA Model on Train Dataset. Also showcase your model building approach:

Linear Discriminant Analysis (LDA) is a supervised learning algorithm used for classification tasks in machine learning. It is a technique used to find a linear combination of features that best separates the classes in a dataset.

# Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model:

Classification report of LDA model on test data is as follows, here based on Recall value, this model can predict 89% of our defaulters. Hence we can say this models predictability is between Logistic Regression and Random Forest Model.

Fig 15: Classification Report(LDA)



# Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve):

Now that we have built classification models using Logistic Regression, Random Forest and Linear Discriminant Analysis, let compare performance of each model using AUC Score and ROC Curve.

AUC is a common abbreviation for Area Under the Receiver Operating Characteristic Curve (ROC AUC). It’s a metric used to assess the performance of classification machine learning models.

The ROC is a graph which maps the relationship between true positive rate (TPR) and the false positive rate (FPR), showing the TPR that we can expect to receive for a given trade-off with FPR. The AUC score is the area under this ROC curve, meaning that the resulting score represents in broad terms the model's ability to predict classes correctly.

AUC score is interpreted as the probability that the model will assign a larger probability to a random positive observation than a random negative observation. More simplistically, AUC score can be interpreted as the model’s ability to accurately classify classes on a scale from 0 to 1, where 1 is best and 0.5 is as good as random choice.

Fig 16: AUC Score Interpretation



So lets see the AUC Score and ROC Curves of all our models on train and test data sets,

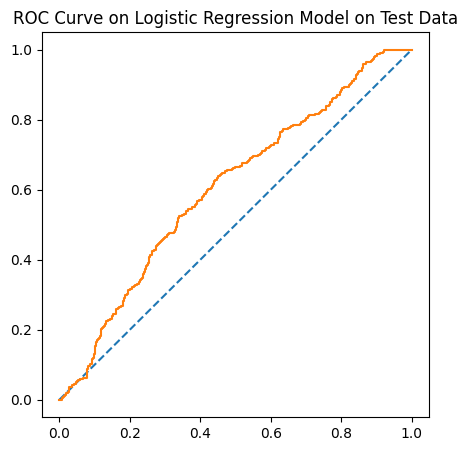
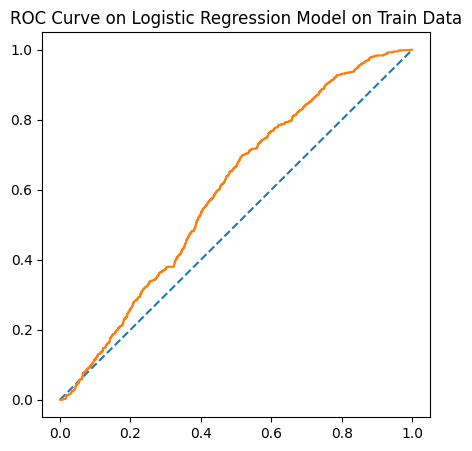
Logistic Regression:

AUC Score on Train data – 0.596

AUC Score on Test data – 0.6

ROC Curve –

Fig 17: ROC Curve Logit Model



Random Forest:

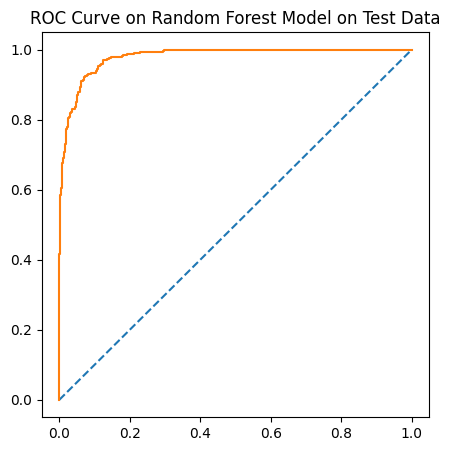
AUC Score on Train data – 0.992

AUC Score on Test data – 0.98

ROC Curve –

Fig 18: ROC Curve Random Forest Model

A graph of a function

Description automatically generated

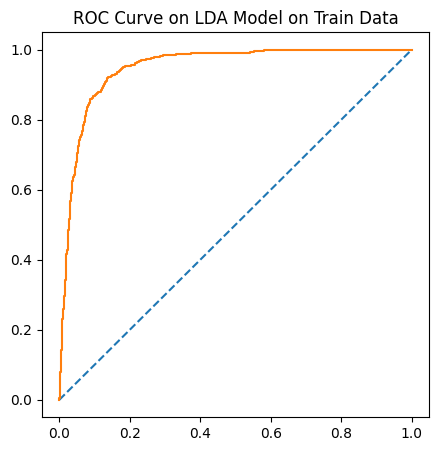
Linear Discriminant Analysis(LDA):

AUC Score on Train data – 0.948

AUC Score on Test data – 0.943

ROC Curve –

Fig 19: ROC Curve LDA Model

A graph of a curve

Description automatically generated

# Conclusions and Recommendations:

Based on prediction capabilities of all the above models on our data, random forest clearly is the best model to use which can predict 93% of total defaulters.

It is recommended that the companies that are predicted as defaulters should be given credit only after assessing company’s future plans and also we might consider different interest rates for these companies.

# PART B:

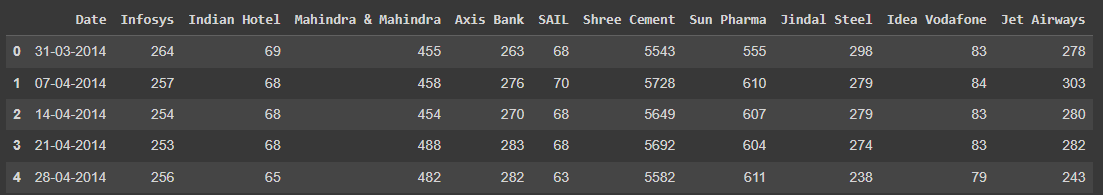
# Problem Statement:

The dataset contains 6 years of information(weekly stock information) on the stock prices of 10 different Indian Stocks. Calculate the mean and standard deviation on the stock returns and share insights. You are expected to do the Market Risk Analysis using Python.

# Exploratory Data Analysis:

Sample Data:

Fig 20: Data head Stock data



# Data Info:

Given dataset has total 314 entries with no duplicates and 11 columns or features. Non of the feature has null or missing values.

Fig 21: Data Info Stock data

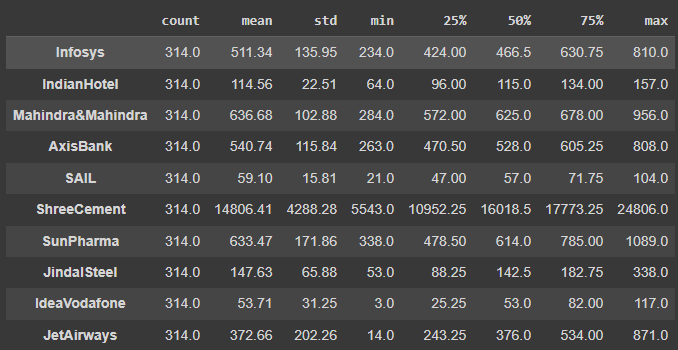
A screenshot of a computer

Description automatically generated

# Five Point summary:

Given data is related to stock prices of 10 companies across 6 years with weekly frequency. There is no much skewness in the data as mean and median values are almost close. ShreeCement is the company with highest average stock price of 14806.41 and IdeaVodafone is the company with least average price of around 53.71

Fig 22: Data Description Stock data



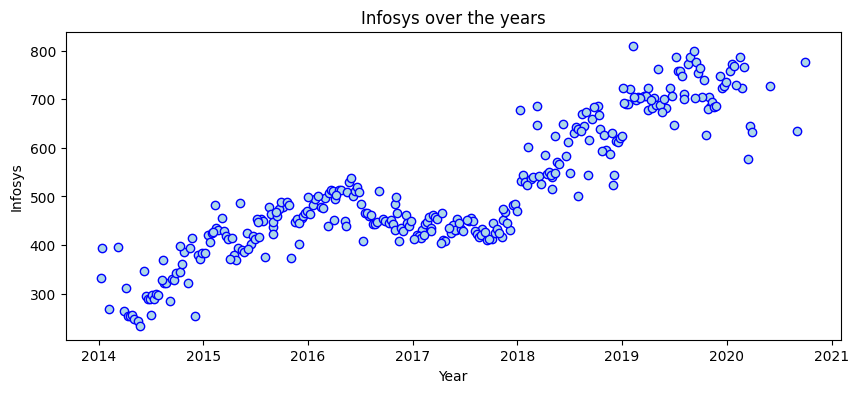
# Draw Stock Price Graph(Stock Price vs Time) for any 2 given stocks with inference:

Lets plot Stock Price vs Time graphs for Infosys and IdeaVodafone,

Infosys(Stock Price vs Time):

We can see the stock price has seen a decent upward trend across years and this stock seems pretty bullish.

Fig 23: Price vs Time Plot 1



IdeaVodafone(Stock Price vs Time):

The price of this Stock has seen a continuous downward trend from the past 6 years and it seems pretty bearish.

Fig 24: Price vs Time Plot 2

A graph showing a line of red dots

Description automatically generated

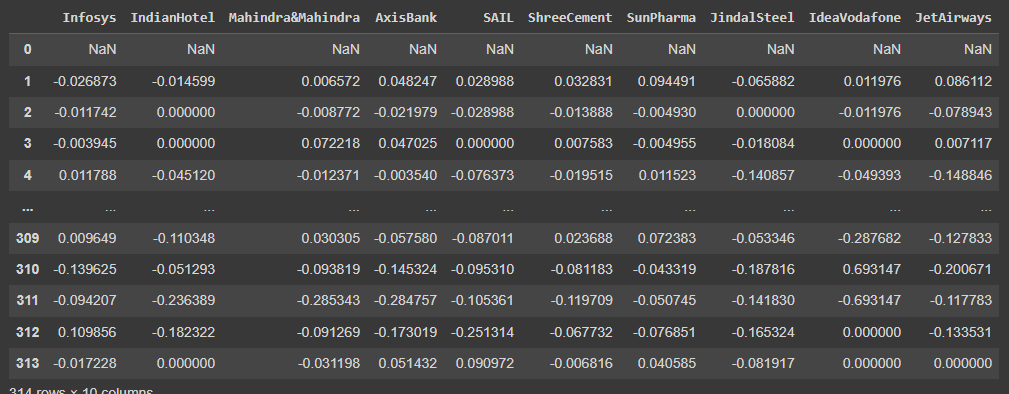
# Calculate Returns for all stocks with inference:

Steps for calculating returns from prices:

* Take logarithms
* Take differences

The logarithmic returns of all the stocks each week is as follows(Showing only first 5 and last 5 records)

Fig 25: Logarithmic Returns

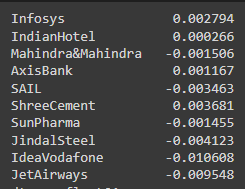


This can be interpreted as the logarithmic change of stock price of a week compared to its previous week. If it is positive then it means the stock has seen a price increase from previous and if its negative then it indicates the stock price has been reduced since last week. Higher the value, higher is the change of price compared to previous week.

# Calculate Stock Means and Standard Deviation for all stocks with inference:

Stock means here indicate the average return on logarithmic scale over six years,

Fig 26: Stock Means



As per the data, only Infosys, IndianHotel, AxisBank and ShreeCement are the companies that gave profits when stock in these companies are held for six years. IdeaVodafone Seems to be the lowest performing stock from the given list followed by JetAirways and JindalSteel.

Stock Standard Deviation indicates the volatility/risk associated with each stock. In long term,

Fig 27: Stock Standard Deviation

A screenshot of a computer

Description automatically generated

Above is the volatility of stocks in ascending order, Infosys is the stock with least stock followed by ShreeCement and Mahindra&Mahindra.

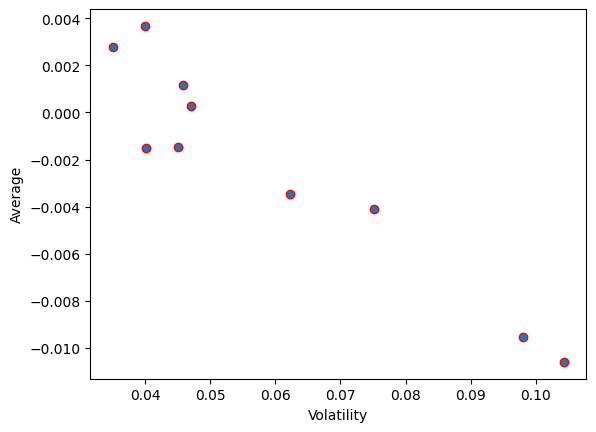
IdeaVodafone is the stock with high volatility followed by JetAirwars and JindalSteel.

# Draw a plot of Stock Means vs Standard Deviation and state your inference:

Below is the graph between Stock Means(Returns) and Standard Deviation(Volatility) for all the stocks.

As per the graph, the average returns are increasing with decrease in volatility. Stocks with high average returns are the ones with least volatility. Hence these stocks are stable and high performing stocks.

Fig 28: Stock Mean vs Standard Deviation Plot



# Conclusions and Recommendations:

Best stocks are the ones with higher average and less volatility, now if we sort the stocks by higher average and less volatility, we could pick the best stocks from the top to bottom. The top ones are the best performing stocks compared to bottom ones.

As per this analogy, if we have to pick top 5 stocks, then ShreeCement, Infosys, AxisBank, IndianHotel and SunPharms would be the best choice.

Hence when a portfolio to be built from the below 10 stocks, then pick the top 4 stocks and avoid the rest. Provided, there are no developments in the bottom stocks.

Fig 29: Best Stocks Top to Bottom

