FINANCE AND RISK ANALYTICS BUSINESS REPORT

APOORVA P

CONTENT

LIST OF FIGURES

LIST OF TABLES

PART A: Define the problem and perform Exploratory Data Analysis

CONTEXT

In the realm of modern finance, businesses encounter the perpetual challenge of managing debt obligations effectively to maintain a favourable credit standing and foster sustainable growth. Investors keenly scrutinize companies capable of navigating financial complexities while ensuring stability and profitability. A pivotal instrument in this evaluation process is the balance sheet, which provides a comprehensive overview of a company's assets, liabilities, and shareholder equity, offering insights into its financial health and operational efficiency. In this context, leveraging available financial data, particularly from preceding fiscal periods, becomes imperative for informed decision-making and strategic planning.

OBJECTIVE

A renowned credit rating organization wants to develop a Financial Health Assessment Tool. With the help of the tool, it endeavors to empower businesses and investors with a robust mechanism for evaluating the financial well-being and creditworthiness of companies. By harnessing machine learning techniques, the organization aims to analyze historical financial statements and extract pertinent insights to facilitate informed decision-making via the tool. Specifically, the organization foresees facilitating the following with the help of the tool:

1. Debt Management Analysis: Identify patterns and trends in debt management practices to assess the ability of businesses to fulfil financial obligations promptly and efficiently, and identify potential cases of default.

2. Credit Risk Evaluation: Evaluate credit risk exposure by analysing liquidity ratios, debt-to-equity ratios, and other key financial indicators to ascertain the likelihood of default and inform investment decisions.

As a part of the data science team in the organization, you have been provided with the financial metrics of different companies. The task is to analyse the data provided and develop a predictive model leveraging machine learning techniques to identify whether a given company will default on its debt repayments in the next two quarters. The predictive model will help the organization anticipate potential challenges with the financial performance of the companies and enable proactive risk mitigation strategies.

EXPLORATORY DATA ANALYSIS (EDA)

A quick glimpse of the data is shown below.

```
df.head()
```

	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_
0	16974	Hind.Cables	8820000000.00	0.00	0.46	0.00	0.(
1	21214	Tata Tele. Mah.	9380000000.00	4230000000.00	0.46	0.00	0.(
2	14852	ABG Shipyard	3800000000.00	815000000.00	0.45	0.00	0.(
3	2439	GTL	6440000000.00	0.00	0.46	0.00	0.(
4	23505	Bharati Defence	3680000000.00	0.00	0.46	0.00	0.4

5 rows × 58 columns

```
df.tail()
```

	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_
2053	2743	Kothari Ferment.	0.00	6490000000.00	0.48	0.00	
2054	21216	Firstobj.Tech.	0.00	0.00	0.47	0.00	
2055	142	Diamines & Chem.	0.00	8370000000.00	0.48	0.00	
2056	18014	IL&FS Engg.	3750000000.00	0.00	0.47	0.00	
2057	43229	Channel Nine	0.00	0.00	0.47	0.00	

5 rows × 58 columns

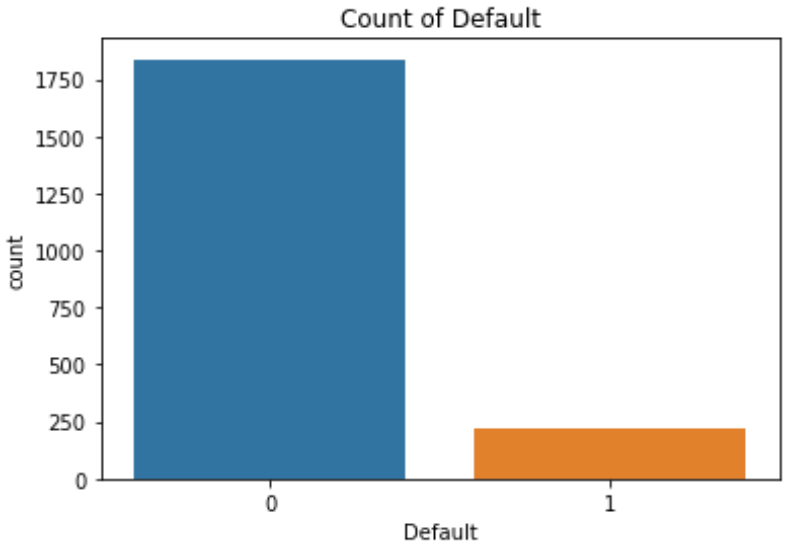
- DATASET: CompData.xlsx
- The number of rows (observations) is 2058
- The number of columns (variables) is 58
- There are no duplicate values
- There are 53 float variables I int variable and one object variable.
- For analysis we are dropping columns Co_Code and Co_Name
- The column names are modified to not start with '_'

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 58 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Co_Code                                                 2058 non-null   int64
 1   Co_Name                                                 2058 non-null   object
 2   _Operating_Expense_Rate                                 2058 non-null   float64
 3   _Research_and_development_expense_rate                  2058 non-null   float64
 4   _Cash_flow_rate                                         2058 non-null   float64
 5   _Interest_bearing_debt_interest_rate                    2058 non-null   float64
 6   _Tax_rate_A                                             2058 non-null   float64
 7   _Cash_Flow_Per_Share                                    1891 non-null   float64
 8   _Per_Share_Net_profit_before_tax_Yuan_                  2058 non-null   float64
 9   _Realized_Sales_Gross_Profit_Growth_Rate                2058 non-null   float64
 10  _Operating_Profit_Growth_Rate                           2058 non-null   float64
 11  _Continuous_Net_Profit_Growth_Rate                      2058 non-null   float64
 12  _Total_Asset_Growth_Rate                                2058 non-null   float64
 13  _Net_Value_Growth_Rate                                  2058 non-null   float64
 14  _Total_Asset_Return_Growth_Rate_Ratio                   2058 non-null   float64
 15  _Cash_Reinvestment_perc                                 2058 non-null   float64
 16  _Current_Ratio                                          2058 non-null   float64
 17  _Quick_Ratio                                            2058 non-null   float64
 18  _Interest_Expense_Ratio                                 2058 non-null   float64
 19  _Total_debt_to_Total_net_worth                          2037 non-null   float64
 20  _Long_term_fund_suitability_ratio_A                     2058 non-null   float64
 21  _Net_profit_before_tax_to_Paid_in_capital               2058 non-null   float64
 22  _Total_Asset_Turnover                                   2058 non-null   float64
 23  _Accounts_Receivable_Turnover                           2058 non-null   float64
 24  _Average_Collection_Days                                2058 non-null   float64
 25  _Inventory_Turnover_Rate_times                          2058 non-null   float64
 26  _Fixed_Assets_Turnover_Frequency                        2058 non-null   float64
 27  _Net_Worth_Turnover_Rate_times                          2058 non-null   float64
 28  _Operating_profit_per_person                            2058 non-null   float64
 29  _Allocation_rate_per_person                             2058 non-null   float64
 30  _Quick_Assets_to_Total_Assets                           2058 non-null   float64
 31  _Cash_to_Total_Assets                                   1962 non-null   float64
 32  _Quick_Assets_to_Current_Liability                      2058 non-null   float64
 33  _Cash_to_Current_Liability                              2058 non-null   float64
 34  _Operating_Funds_to_Liability                           2058 non-null   float64
 35  _Inventory_to_Working_Capital                           2058 non-null   float64
 36  _Inventory_to_Current_Liability                         2058 non-null   float64
 37  _Long_term_Liability_to_Current_Assets                  2058 non-null   float64
 38  _Retained_Earnings_to_Total_Assets                      2058 non-null   float64
 39  _Total_income_to_Total_expense                          2058 non-null   float64
 40  _Total_expense_to_Assets                                2058 non-null   float64
 41  _Current_Asset_Turnover_Rate                            2058 non-null   float64
 42  _Quick_Asset_Turnover_Rate                              2058 non-null   float64
 43  _Cash_Turnover_Rate                                     2058 non-null   float64
 44  _Fixed_Assets_to_Assets                                 2058 non-null   float64
 45  _Cash_Flow_to_Total_Assets                              2058 non-null   float64
 46  _Cash_Flow_to_Liability                                 2058 non-null   float64
 47  _CFO_to_Assets                                          2058 non-null   float64
 48  _Cash_Flow_to_Equity                                    2058 non-null   float64
 49  _Current_Liability_to_Current_Assets                    2044 non-null   float64
 50  _Liability_Assets_Flag                                  2058 non-null   int64
 51  _Total_assets_to_GNP_price                              2058 non-null   float64
 52  _No_credit_Interval                                     2058 non-null   float64
 53  _Degree_of_Financial_Leverage_DFL                       2058 non-null   float64
 54  _Interest_Coverage_Ratio_Interest_expense_to_EBIT       2058 non-null   float64
 55  _Net_Income_Flag                                        2058 non-null   int64
 56  _Equity_to_Liability                                    2058 non-null   float64
 57  Default                                                 2058 non-null   int64
dtypes: float64(53), int64(4), object(1)
memory usage: 932.7+ KB
```

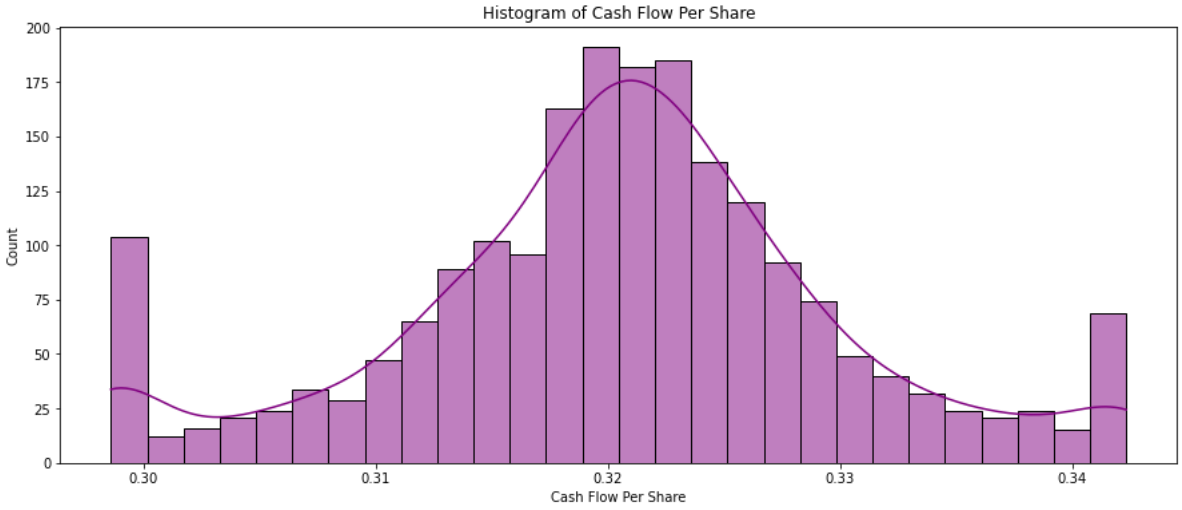
	count	mean	std	min	25%	50%	75%	max
Co_Code	2058.00	17572.11	21892.89	4.00	3674.00	6240.00	24280.75	72493.00
Operating_Expense_Rate	2058.00	2052388835.76	3252623690.29	0.00	0.00	0.00	4110000000.00	9980000000.00
Research_and_development_expense_rate	2058.00	1208634256.56	2144568158.08	0.00	0.00	0.00	1550000000.00	9980000000.00
Cash_flow_rate	2058.00	0.47	0.02	0.00	0.46	0.46	0.47	1.00
Interest_bearing_debt_interest_rate	2058.00	11130223.52	90425949.04	0.00	0.00	0.00	0.00	990000000.00
Tax_rate_A	2058.00	0.11	0.15	0.00	0.00	0.04	0.22	1.00
Cash_Flow_Per_Share	1891.00	0.32	0.02	0.17	0.31	0.32	0.33	0.46
Per_Share_Net_profit_before_tax_Yuan_	2058.00	0.18	0.03	0.00	0.17	0.18	0.19	0.79
Realized_Sales_Gross_Profit_Growth_Rate	2058.00	0.02	0.02	0.00	0.02	0.02	0.02	1.00
Operating_Profit_Growth_Rate	2058.00	0.85	0.00	0.74	0.85	0.85	0.85	1.00
Continuous_Net_Profit_Growth_Rate	2058.00	0.22	0.01	0.00	0.22	0.22	0.22	0.23
Total_Asset_Growth_Rate	2058.00	5287663257.05	2912614769.58	0.00	4315000000.00	6225000000.00	7220000000.00	9980000000.00
Net_Value_Growth_Rate	2058.00	5189504.37	207791797.86	0.00	0.00	0.00	0.00	9330000000.00
Total_Asset_Return_Growth_Rate_Ratio	2058.00	0.26	0.00	0.25	0.26	0.26	0.26	0.36
Cash_Reinvestment_perc	2058.00	0.38	0.03	0.03	0.37	0.38	0.39	1.00
Current_Ratio	2058.00	1336248.80	60619173.20	0.00	0.01	0.01	0.01	2750000000.00
Quick_Ratio	2058.00	27755102.05	444866390.47	0.00	0.00	0.01	0.01	9230000000.00
Interest_Expense_Ratio	2058.00	0.63	0.01	0.53	0.63	0.63	0.63	0.81
Total_debt_to_Total_net_worth	2037.00	107142285.73	2696960017.59	0.00	0.00	0.01	0.01	9940000000.00
Long_term_fund_suitability_ratio_A	2058.00	0.01	0.03	0.00	0.01	0.01	0.01	1.00
Net_profit_before_tax_to_Paid_in_capital	2058.00	0.18	0.03	0.00	0.17	0.17	0.18	0.79
Total_Asset_Turnover	2058.00	0.13	0.10	0.00	0.06	0.10	0.17	0.92
Accounts_Receivable_Turnover	2058.00	41598639.46	504767266.59	0.00	0.00	0.00	0.00	9740000000.00
Average_Collection_Days	2058.00	26297862.01	410996733.83	0.00	0.00	0.01	0.01	8800000000.00
Inventory_Turnover_Rate_times	2058.00	2030227259.48	3077250265.27	0.00	0.00	19100000.00	3815000000.00	9990000000.00
Fixed_Assets_Turnover_Frequency	2058.00	1230897959.18	2649288936.44	0.00	0.00	0.00	0.01	9990000000.00
Net_Worth_Turnover_Rate_times	2058.00	0.04	0.04	0.01	0.02	0.03	0.04	1.00
Operating_profit_per_person	2058.00	0.40	0.05	0.00	0.39	0.40	0.40	1.00
Allocation_rate_per_person	2058.00	5725558.82	197949961.06	0.00	0.00	0.01	0.02	8280000000.00
Quick_Assets_to_Total_Assets	2058.00	0.34	0.21	0.00	0.17	0.31	0.48	0.99
Cash_to_Total_Assets	1962.00	0.08	0.10	0.00	0.02	0.05	0.10	0.93
Quick_Assets_to_Current_Liability	2058.00	11904761.91	312292270.93	0.00	0.00	0.01	0.01	8820000000.00
Cash_to_Current_Liability	2058.00	92825072.90	785189881.95	0.00	0.00	0.00	0.01	9170000000.00
Operating_Funds_to_Liability	2058.00	0.35	0.04	0.03	0.34	0.35	0.35	1.00
Inventory_to_Working_Capital	2058.00	0.28	0.02	0.00	0.28	0.28	0.28	1.00
Inventory_to_Current_Liability	2058.00	57863459.68	627879536.23	0.00	0.00	0.01	0.01	9600000000.00
Long_term_Liability_to_Current_Assets	2058.00	7340106901	669352618.01	0.00	0.00	0.00	0.01	9310000000.00
Retained_Earnings_to_Total_Assets	2058.00	0.93	0.03	0.00	0.93	0.94	0.94	0.97
Total_income_to_Total_expense	2058.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Total_expense_to_Assets	2058.00	0.03	0.04	0.00	0.01	0.02	0.04	1.00
Current_Asset_Turnover_Rate	2058.00	1273303377.07	2839740987.63	0.00	0.00	0.00	0.00	9990000000.00
Quick_Asset_Turnover_Rate	2058.00	2571767687.08	3453544121.67	0.00	0.00	0.00	5790000000.00	10000000000.00
Cash_Turnover_Rate	2058.00	2653695544.22	2821244732.19	0.00	0.00	1730000000.00	4550000000.00	9990000000.00
Fixed_Assets_to_Assets	2058.00	4042760.23	183400553.09	0.00	0.10	0.21	0.42	8320000000.00
Cash_Flow_to_Total_Assets	2058.00	0.64	0.05	0.00	0.63	0.64	0.65	1.00
Cash_Flow_to_Liability	2058.00	0.46	0.03	0.03	0.46	0.46	0.46	0.91
CFO_to_Assets	2058.00	0.58	0.06	0.00	0.55	0.58	0.61	0.98
Cash_Flow_to_Equity	2058.00	0.31	0.01	0.00	0.31	0.31	0.32	0.57
Current_Liability_to_Current_Assets	2044.00	0.04	0.05	0.00	0.02	0.03	0.04	1.00
Liability_Assets_Flag	2058.00	0.00	0.06	0.00	0.00	0.00	0.00	1.00
Total_assets_to_GNP_price	2058.00	27793974.74	471771444.55	0.00	0.00	0.00	0.01	9820000000.00
No_credit_Interval	2058.00	0.62	0.01	0.41	0.62	0.62	0.62	0.96
Degree_of_Financial_Leverage_DFL	2058.00	0.03	0.01	0.01	0.03	0.03	0.03	0.46
Interest_Coverage_Ratio_Interest_expense_to_EBIT	2058.00	0.57	0.01	0.17	0.57	0.57	0.57	0.67
Net_Income_Flag	2058.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Equity_to_Liability	2058.00	0.04	0.06	0.00	0.02	0.03	0.04	1.00
Default	2058.00	0.11	0.31	0.00	0.00	0.00	0.00	1.00

- The above Table 4 shows the summary of the descriptive statistics of the columns in the dataset, and it also depicts the dataset's mean, median, min, max and lower and upper quartile values.

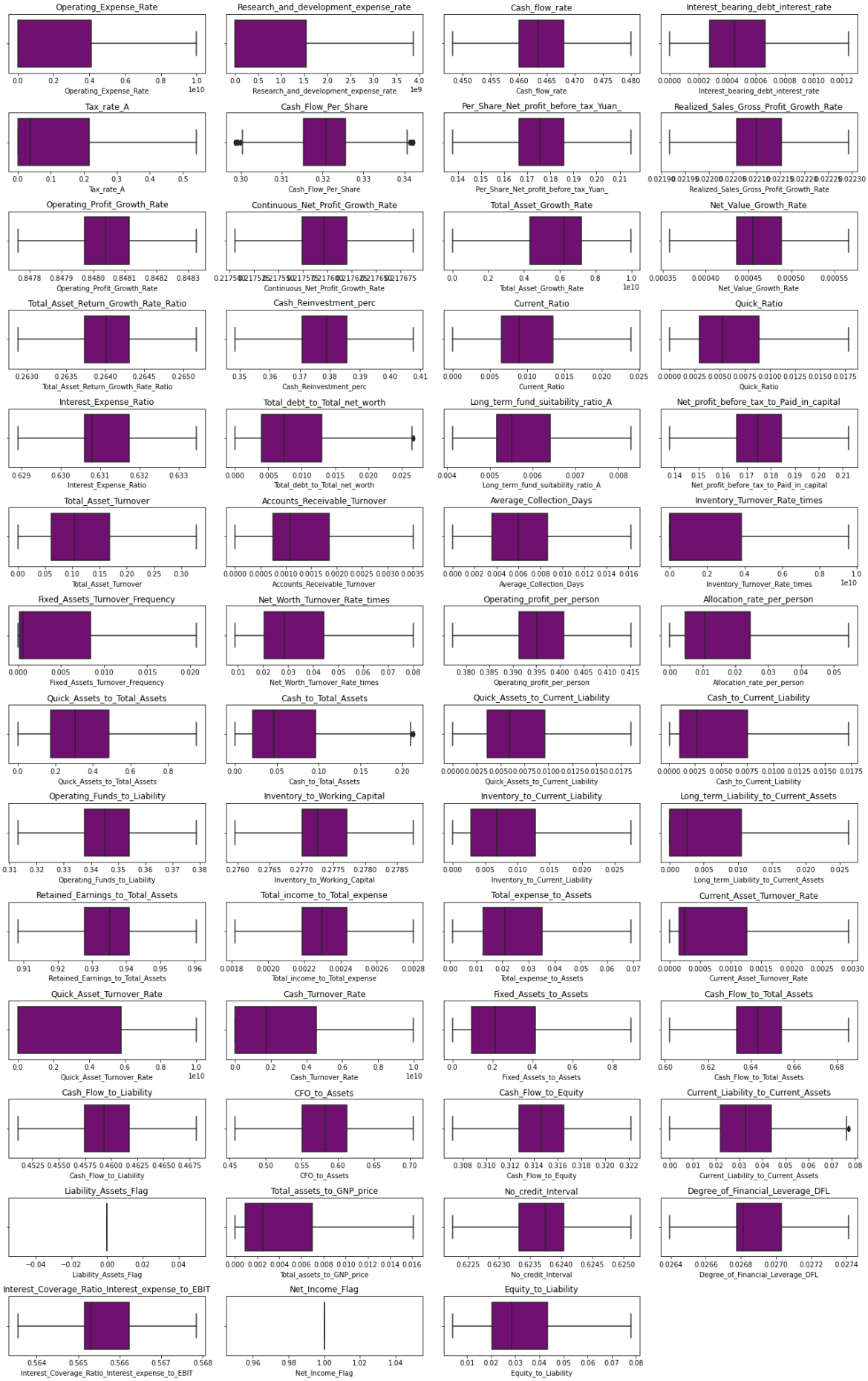
- The dataset includes a column named "Default," which is the target variable. This column contains binary values (0 and 1) indicating whether a company has defaulted or not. This variable could be used for building a predictive model or conducting further analysis related to company defaults.
- The target variable Default the value count of 0's -1838 and 1's -220

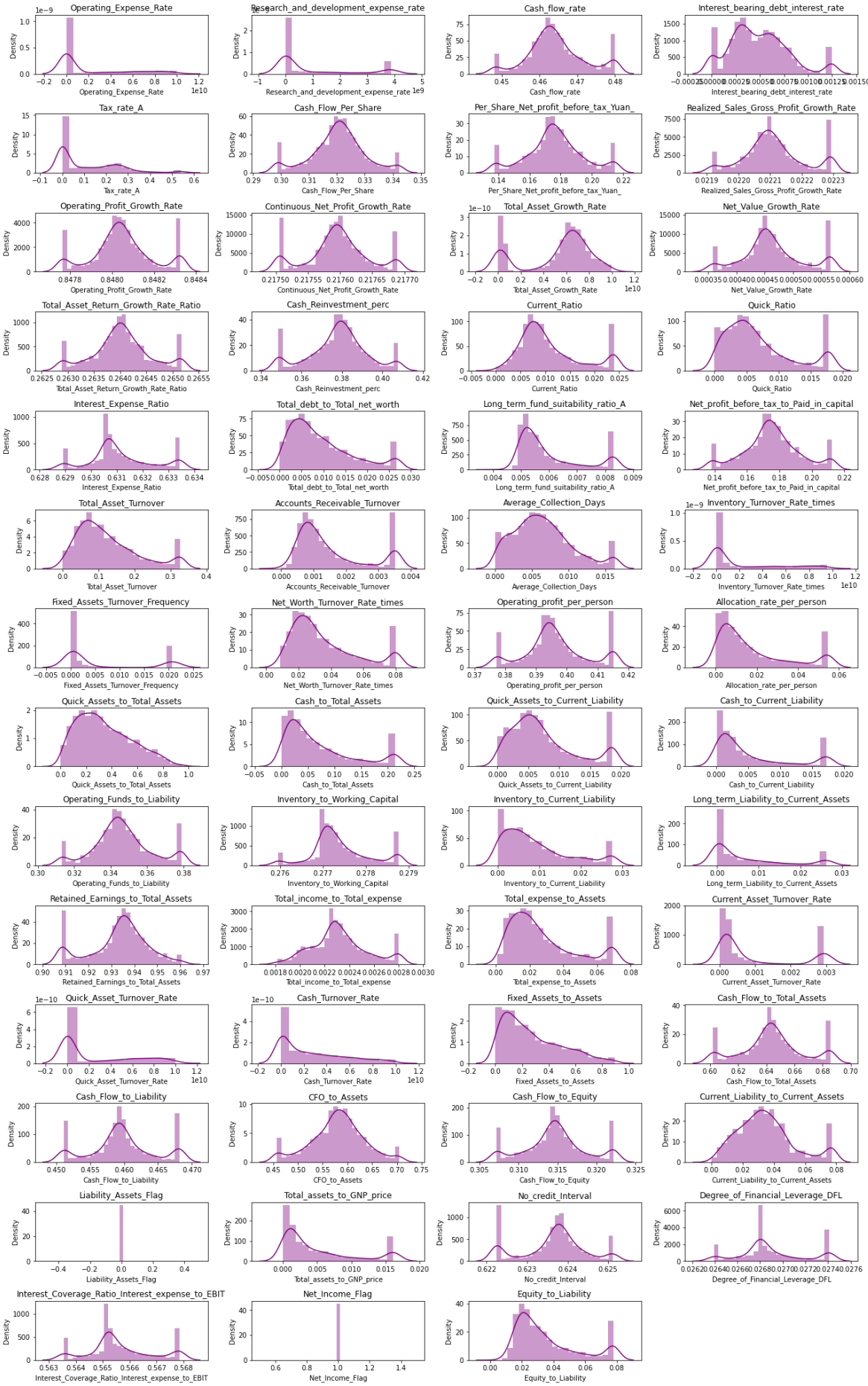


- It is seen that almost 11% of the total entries in "Default" belong to category "1". The dataset has class imbalance issue.
- From the above count plot, it is obvious that the majority of companies, totalling a large number, fall into the non-default category, while only a smaller subset of 220 companies is classified as defaulters.

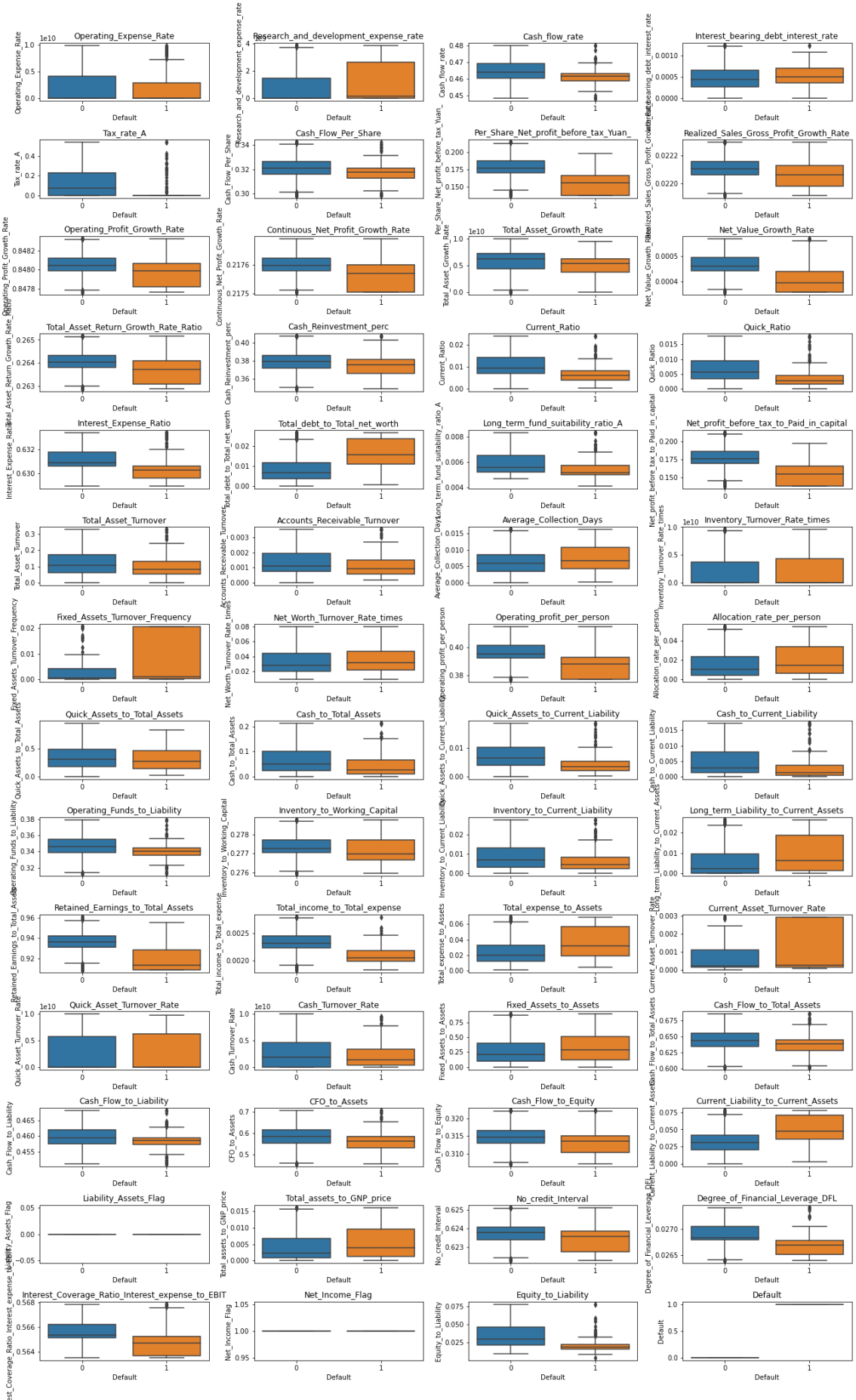


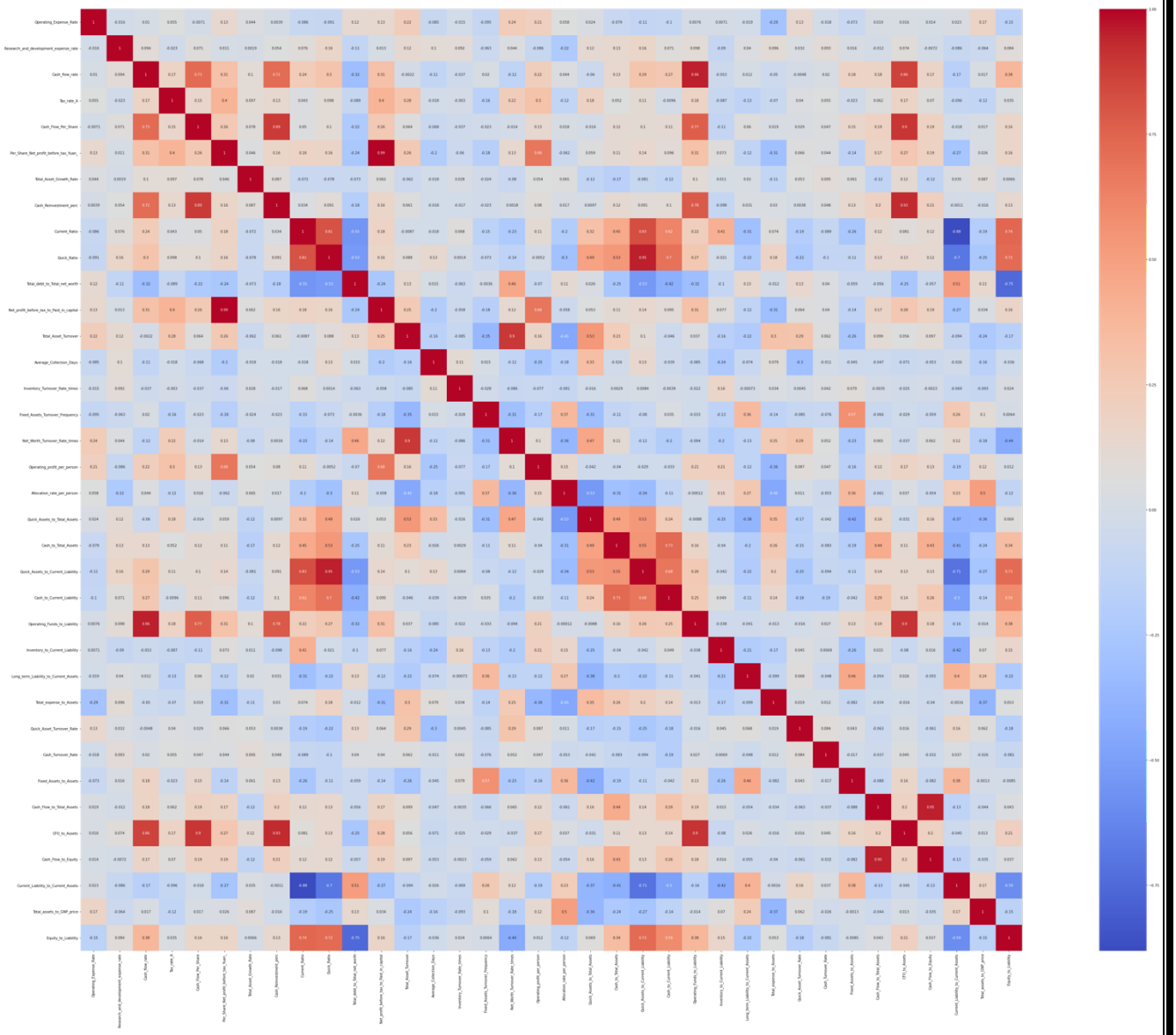
- The histogram reveals a prominent peak at a cash low per share value of 0.32, indicating that a significant proportion of the data points in the dataset are clustered around this value.


Operating_Expense_Rate
Research_and_development_expense_rate
Cash_flow_rate
Interest_bearing_debt_interest_rate
Tax_rate_A
Cash_Flow_Per_Share
Per_Share_Net_profit_before_tax_Yuan_
Realized_Sales_Gross_Profit_Growth_Rate
Operating_Profit_Growth_Rate
Continuous_Net_Profit_Growth_Rate
Total_Asset_Growth_Rate
Net_Value_Growth_Rate
Total_Asset_Return_Growth_Rate_Ratio
Cash_Reinvestment_perc
Current_Ratio
Quick_Ratio
Interest_Expense_Ratio
Total_debt_to_Total_net_worth
Long_term_fund_suitability_ratio_A
Net_profit_before_tax_to_Paid_in_capital
Total_Asset_Turnover
Accounts_Receivable_Turnover
Average_Collection_Days
Inventory_Turnover_Rate_times
Fixed_Assets_Turnover_Frequency
Net_Worth_Turnover_Rate_times
Operating_profit_per_person
Allocation_rate_per_person
Quick_Assets_to_Total_Assets
Cash_to_Total_Assets
Quick_Assets_to_Current_Liability
Cash_to_Current_Liability
Operating_Funds_to_Liability
Inventory_to_Working_Capital
Inventory_to_Current_Liability
Long_term_Liability_to_Current_Assets
Retained_Earnings_to_Total_Assets
Total_income_to_Total_expense
Total_expense_to_Assets
Current_Asset_Turnover_Rate
Quick_Asset_Turnover_Rate
Cash_Turnover_Rate
Fixed_Assets_to_Assets
Cash_Flow_to_Total_Assets
Cash_Flow_to_Liability
CFO_to_Assets
Cash_Flow_to_Equity
Current_Liability_to_Current_Assets
Liability_Assets_Flag
Total_assets_to_GNP_price
No_credit_Interval
Degree_of_Financial_Leverage_DFL
Interest_Coverage_Ratio_Interest_expense_to_EBIT
Net_Income_Flag
Equity_to_Liability


Operating_Expense_Rate
Research_and_development_expense_rate
Cash_flow_rate
Interest_bearing_debt_interest_rate
Tax_rate_A
Cash_Flow_Per_Share
Per_Share_Net_profit_before_tax_Yuan_
Realized_Sales_Gross_Profit_Growth_Rate
Operating_Profit_Growth_Rate
Continuous_Net_Profit_Growth_Rate
Total_Asset_Growth_Rate
Net_Value_Growth_Rate
Total_Asset_Return_Growth_Rate_Ratio
Cash_Reinvestment_perc
Current_Ratio
Quick_Ratio
Interest_Expense_Ratio
Total_debt_to_Total_net_worth
Long_term_fund_suitability_ratio_A
Net_profit_before_tax_to_Paid_in_capital
Total_Asset_Turnover
Accounts_Receivable_Turnover
Average_Collection_Days
Inventory_Turnover_Rate_times
Fixed_Assets_Turnover_Frequency
Net_Worth_Turnover_Rate_times
Operating_profit_per_person
Allocation_rate_per_person
Quick_Assets_to_Total_Assets
Cash_to_Total_Assets
Quick_Assets_to_Current_Liability
Cash_to_Current_Liability
Operating_Funds_to_Liability
Inventory_to_Working_Capital
Inventory_to_Current_Liability
Long_term_Liability_to_Current_Assets
Retained_Earnings_to_Total_Assets
Total_income_to_Total_expense
Total_expense_to_Assets
Current_Asset_Turnover_Rate
Quick_Asset_Turnover_Rate
Cash_Turnover_Rate
Fixed_Assets_to_Assets
Cash_Flow_to_Total_Assets
Cash_Flow_to_Liability
CFO_to_Assets
Cash_Flow_to_Equity
Current_Liability_to_Current_Assets
Liability_Assets_Flag
Total_assets_to_GNP_price
No_credit_Interval
Degree_of_Financial_Leverage_DFL
Interest_Coverage_Ratio_Interest_expense_to_EBIT
Net_Income_Flag
Equity_to_Liability
Density

- The boxplot illustrates that the middle 50% of the data is concentrated within the range of 0.0 to 1.5 for the research and development expense rate. Furthermore, the maximum value for this variable is approximately 4.0, indicating values beyond the upper range of the boxplot.
- The boxplot for "Average Collection Days" reveals a distribution with a minimum value of 0.000 and a maximum value of 0.016**.** The median (50th percentile) falls at 0.006, indicating the central tendency of the data.
- The "Quick Assets to Total Assets" boxplot analysis by default status shows distinct distributions between organisations that have defaulted (1) and those that have not (0). While the general trends might seem comparable, it is important to remember that defaulted businesses typically show somewhat lower values for "Quick Assets to Total Assets" when compared to non-defaulted businesses. According to this finding, the fast asset to total asset ratio may be a useful tool for assessing default risk, with lower levels possibly reflecting a higher likelihood of default.
- The boxplot analysis for "Total_Asset_Growth_Rate" by default status (0 – Not Defaulted, 1 - Defaulted) indicates similar distributions for defaulted and non-defaulted companies. Both groups have comparable median values, ranging from 0.5 to 0.6. This suggests that the growth rate of total assets may not be a sole distinguishing factor for default risk in this dataset.
- • Interesting insights are shown by the boxplot analysis of "Total Asset Growth Rate" by default status (0 - Not Defaulted, 1 - Defaulted). Non-defaulted businesses (0) exhibit extremely little growth, with the majority of values centred around 0.0 and a small number of outliers reaching as high as 3.75. Defaulted companies (1), on the other hand, show a greater variation in growth rates, with a median of 0.25 and some values reaching 2.75. This shows that when compared to non-defaulted enterprises, defaulted companies typically have more variable asset growth rates.
- The boxplot analysis for "Current_Liability_to_Current_Assets" based on default status (0 - Not Defaulted, 1 - Defaulted) reveals distinct distributions. Non-defaulted companies (0) show a relatively lower ratio of current liabilities to current assets, with a median value of 0.03 and a range from 0.0 to 0.04. Defaulted companies (1) display higher values, with a median of 0.05 and a wider range from 0.0 to 0.08. This suggests that a higher ratio of current liabilities to current assets may indicate a higher risk of default.
- Interesting trends may be seen in the boxplot analysis of "Operating_profit_per_person" by default status (0 - Not Defaulted, 1 - Defaulted). Operating profit per person is comparatively concentrated in non-defaulted enterprises (0), ranging from 0.393 to 0.400. Defaulted companies (1), on the other hand, show a significantly broader range, with values ranging up to 0.415 and a median of 0.390. This shows that, while the overall difference is not great, there might be some variation in operational profit per person between defaulted and non-defaulted enterprises.

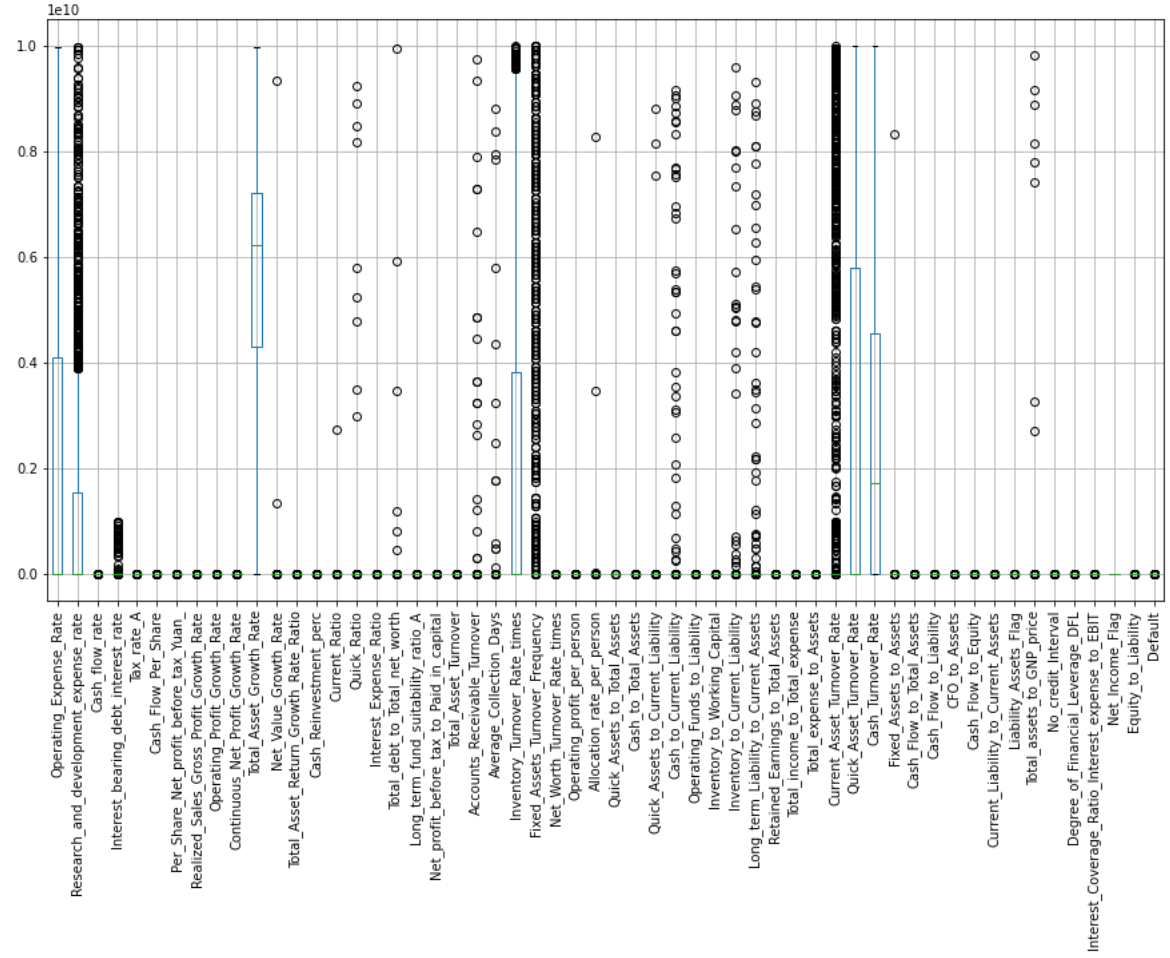

Operating_Expense_Rate
Research_and_development_expense_rate
Cash_flow_rate
Interest_bearing_debt_interest_rate
Tax_rate_A
Cash_Flow_Per_Share
Per_Share_Net_profit_before_tax_Yuan_
Realized_Sales_Gross_Profit_Growth_Rate
Operating_Profit_Growth_Rate
Continuous_Net_Profit_Growth_Rate
Total_Asset_Growth_Rate
Net_Value_Growth_Rate
Total_Asset_Return_Growth_Rate_Ratio
Cash_Reinvestment_perc
Current_Ratio
Quick_Ratio
Interest_Expense_Ratio
Total_debt_to_Total_net_worth
Long_term_fund_suitability_ratio_A
Net_profit_before_tax_to_Paid_in_capital
Total_Asset_Turnover
Accounts_Receivable_Turnover
Average_Collection_Days
Inventory_Turnover_Rate_times
Fixed_Assets_Turnover_Frequency
Net_Worth_Turnover_Rate_times
Operating_profit_per_person
Allocation_rate_per_person
Quick_Assets_to_Total_Assets
Cash_to_Total_Assets
Quick_Assets_to_Current_Liability
Cash_to_Current_Liability
Operating_Funds_to_Liability
Inventory_to_Working_Capital
Inventory_to_Current_Liability
Long_term_Liability_to_Current_Assets
Retained_Earnings_to_Total_Assets
Total_income_to_Total_expense
Total_expense_to_Assets
Current_Asset_Turnover_Rate
Quick_Asset_Turnover_Rate
Cash_Turnover_Rate
Fixed_Assets_to_Assets
Cash_Flow_to_Total_Assets
Cash_Flow_to_Liability
CFO_to_Assets
Cash_Flow_to_Equity
Current_Liability_to_Current_Assets
Liability_Assets_Flag
Total_assets_to_GNP_price
No_credit_Interval
Degree_of_Financial_Leverage_DFL
Interest_Coverage_Ratio_Interest_expense_to_EBIT
Net_Income_Flag
Equity_to_Liability
Default
Default



We created a heatmap to explore their correlations. The heatmap displayed strong correlations between certain variables, indicating a potential relationship.

This process helps us streamline the dataset by eliminating irrelevant variables and highlights the interdependencies among the remaining variables, enabling us to focus on the most meaningful analysis.

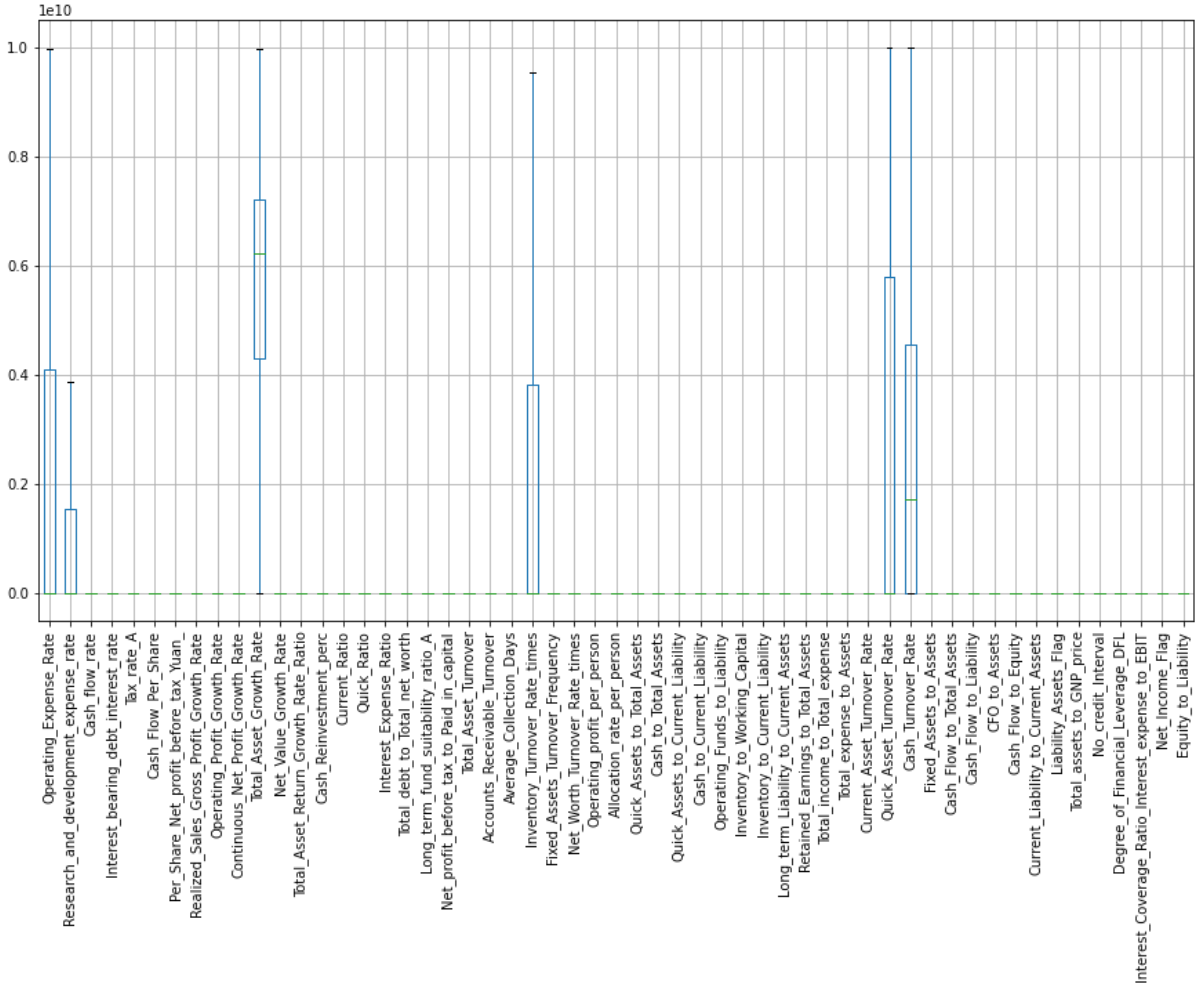
PART A: Data Pre-processing



Before performing outlier treatment, it is important to drop the target variable because outliers in the target variable can significantly affect the analysis and model training.

	Column	No. of outliers
0	Operating_Expense_Rate	0
1	Research_and_development_expense_rate	264
2	Cash_flow_rate	206
3	Interest_bearing_debt_interest_rate	94
4	Tax_rate_A	42
5	Cash_Flow_Per_Share	146
6	Per_Share_Net_profit_before_tax_Yuan_	186
7	Realized_Sales_Gross_Profit_Growth_Rate	283
8	Operating_Profit_Growth_Rate	317
9	Continuous_Net_Profit_Growth_Rate	340
10	Total_Asset_Growth_Rate	0
11	Net_Value_Growth_Rate	304
12	Total_Asset_Return_Growth_Rate_Ratio	226
13	Cash_Reinvestment_perc	220
14	Current_Ratio	193
15	Quick_Ratio	190
16	Interest_Expense_Ratio	328
17	Total_debt_to_Total_net_worth	105
18	Long_term_fund_suitability_ratio_A	234
19	Net_profit_before_tax_to_Paid_in_capital	173
20	Total_Asset_Turnover	101
21	Accounts_Receivable_Turnover	281
22	Average_Collection_Days	77
23	Inventory_Turnover_Rate_times	29
24	Fixed_Assets_Turnover_Frequency	501
25	Net_Worth_Turnover_Rate_times	165
26	Operating_profit_per_person	357
27	Allocation_rate_per_person	200
28	Quick_Assets_to_Total_Assets	4
29	Cash_to_Total_Assets	163
30	Quick_Assets_to_Current_Liability	185
31	Cash_to_Current_Liability	253
32	Operating_Funds_to_Liability	219
33	Inventory_to_Working_Capital	247
34	Inventory_to_Current_Liability	129
35	Long_term_Liability_to_Current_Assets	213
36	Retained_Earnings_to_Total_Assets	208
37	Total_income_to_Total_expense	136
38	Total_expense_to_Assets	168
39	Current_Asset_Turnover_Rate	464
40	Quick_Asset_Turnover_Rate	0
41	Cash_Turnover_Rate	0
42	Fixed_Assets_to_Assets	10
43	Cash_Flow_to_Total_Assets	317
44	Cash_Flow_to_Liability	407
45	CFO_to_Assets	110
46	Cash_Flow_to_Equity	306
47	Current_Liability_to_Current_Assets	121
48	Liability_Assets_Flag	7
49	Total_assets_to_GNP_price	235
50	No_credit_Interval	396
51	Degree_of_Financial_Leverage_DFL	438
52	Interest_Coverage_Ratio_Interest_expense_to_EBIT	376
53	Net_Income_Flag	0
54	Equity_to_Liability	190
55	Default	220

- The table above provides the count of outliers for each individual variable in the dataset. Overall, there are a total of 10,864 outliers identi?ied across the dataset. These outliers can potentially have an impact on the analysis, and it require treatment as part of the data analysis process.



- Outlier removal is a crucial step in data pre-processing to ensure accurate analysis and modelling. The Interquartile Range (IQR) method is used to identify and address outliers in the dataset. This method determines the acceptable range of values in the data by calculating lower and upper limits based on the spread of the data. Overall, by addressing outliers, we ensure that these extreme values do not negatively impact the accuracy and reliability of our analysis.

```
Operating_Expense_Rate                                  0
Research_and_development_expense_rate                   0
Cash_flow_rate                                          0
Interest_bearing_debt_interest_rate                     0
Tax_rate_A                                              0
Cash_Flow_Per_Share                                   167
Per_Share_Net_profit_before_tax_Yuan_                   0
Realized_Sales_Gross_Profit_Growth_Rate                 0
Operating_Profit_Growth_Rate                            0
Continuous_Net_Profit_Growth_Rate                       0
Total_Asset_Growth_Rate                                 0
Net_Value_Growth_Rate                                   0
Total_Asset_Return_Growth_Rate_Ratio                    0
Cash_Reinvestment_perc                                  0
Current_Ratio                                           0
Quick_Ratio                                             0
Interest_Expense_Ratio                                  0
Total_debt_to_Total_net_worth                          21
Long_term_fund_suitability_ratio_A                      0
Net_profit_before_tax_to_Paid_in_capital                0
Total_Asset_Turnover                                    0
Accounts_Receivable_Turnover                            0
Average_Collection_Days                                 0
Inventory_Turnover_Rate_times                           0
Fixed_Assets_Turnover_Frequency                         0
Net_Worth_Turnover_Rate_times                           0
Operating_profit_per_person                             0
Allocation_rate_per_person                              0
Quick_Assets_to_Total_Assets                            0
Cash_to_Total_Assets                                   96
Quick_Assets_to_Current_Liability                       0
Cash_to_Current_Liability                               0
Operating_Funds_to_Liability                            0
Inventory_to_Working_Capital                            0
Inventory_to_Current_Liability                          0
Long_term_Liability_to_Current_Assets                   0
Retained_Earnings_to_Total_Assets                       0
Total_income_to_Total_expense                           0
Total_expense_to_Assets                                 0
Current_Asset_Turnover_Rate                             0
Quick_Asset_Turnover_Rate                               0
Cash_Turnover_Rate                                      0
Fixed_Assets_to_Assets                                  0
Cash_Flow_to_Total_Assets                               0
Cash_Flow_to_Liability                                  0
CFO_to_Assets                                           0
Cash_Flow_to_Equity                                     0
Current_Liability_to_Current_Assets                    14
Liability_Assets_Flag                                   0
Total_assets_to_GNP_price                               0
No_credit_Interval                                      0
Degree_of_Financial_Leverage_DFL                        0
Interest_Coverage_Ratio_Interest_expense_to_EBIT        0
Net_Income_Flag                                         0
Equity_to_Liability                                     0
dtype: int64
```

- Treating Missing Values
- we observe a total of 298 missing values in the dataset. While this number may seem small in comparison to the overall dataset size, it is still important to address these missing values. To impute the missing values, we are utilizing KNN imputation method with a parameter value of n-Neighbour = 5.
- KNN (K-Nearest Neighbors) imputation is a technique used to ?ill in missing values in a dataset based on the values of its nearest neighbors. In this approach, each missing value is replaced with the average value of its k nearest neighbors. The value of k, in this case, is set to 5.
- Therefore, by employing KNN imputation with n-Neighbor = 5, we aim to ?ill in the missing values in the dataset and ensure a more comprehensive and robust analysis. And importantly, KNN imputation helps to minimize bias in the imputed values.

```
X_train.head()
```

	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Share	Pe
0	0.00	357000000.00	0.46	0.00	0.00	0.32	
1	0.00	0.00	0.45	0.00	0.00	0.30	
2	0.00	2000000000.00	0.46	0.00	0.12	0.31	
3	0.00	0.00	0.46	0.00	0.00	0.32	
4	0.00	0.00	0.46	0.00	0.00	0.32	

5 rows × 53 columns

```
X_test.head()
```

	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Share	Pe
0	0.00	3875000000.00	0.47	0.00	0.00	0.32	
1	8650000000.00	1510000000.00	0.47	0.00	0.00	0.32	
2	0.00	1600000000.00	0.46	0.00	0.13	0.32	
3	0.00	3875000000.00	0.45	0.00	0.00	0.31	
4	4560000000.00	1870000000.00	0.46	0.00	0.22	0.33	

5 rows × 53 columns

```
#Scaling of features is done to bring all the features to the same scale.
sc = StandardScaler()

X_train_scaled = pd.DataFrame(sc.fit_transform(X_train), columns=X_train.columns)
X_test_scaled = pd.DataFrame(sc.transform(X_test), columns=X_test.columns)
```

```
X_train_scaled.head()
```

	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Share	Pe
0	-0.63	-0.40	-0.23	-0.55	-0.82	0.09	
1	-0.63	-0.64	-2.19	-1.27	-0.82	-1.85	
2	-0.63	0.74	-0.64	0.31	0.10	-0.65	
3	-0.63	-0.64	-0.35	-0.39	-0.82	-0.16	
4	-0.63	-0.64	-0.21	-0.78	-0.82	-0.01	

5 rows × 53 columns

```
X_test_scaled.head()
```

	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Share	Pe
0	-0.63	2.03	0.42	0.94	-0.82	0.03	
1	2.02	0.40	1.25	0.59	-0.82	0.20	
2	-0.63	0.46	-0.26	-0.84	0.13	0.12	
3	-0.63	2.03	-1.61	-0.67	-0.82	-1.50	
4	0.76	0.65	-0.02	0.69	0.84	0.81	

5 rows × 53 columns

- The data is spilt to ratio of 75:25 and scaled accordingly
- we will be applying data scaling before the train-test split. This step is particularly important because we have observed that some variables in the dataset exhibit large data values. By scaling the data, we can bring all features to a similar scale, preventing any single variable from dominating the learning process and ensuring compatibility across different algorithms. This approach will help us maintain the integrity of our analysis and improve the accuracy and efficiency of our models.
- The data was scaled using the StandardScaler method. This was done to ensure that all variables are brought to a similar scale, removing any potential bias caused by variables with larger data values.
- To evaluate the performance of machine learning models, the dataset was divided into training and testing sets using the train_test_split function. This split ensures that the models are trained on a portion of the data and then tested on unseen data. The test size parameter was set to **0.33**, indicating that **33%** of the data is for testing. The random state parameter was set to **42.** This process allows us to evaluate how well the model performs on new and unseen data.

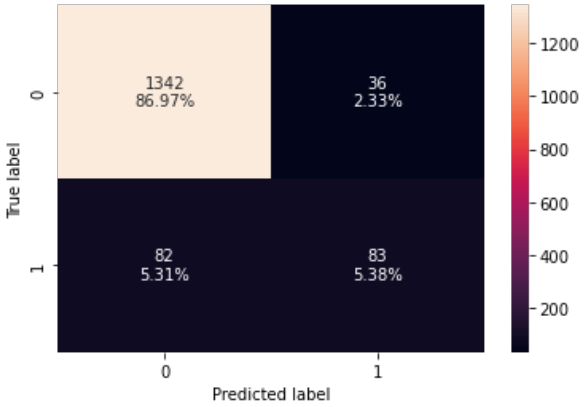
PART A: MODEL BUILDING

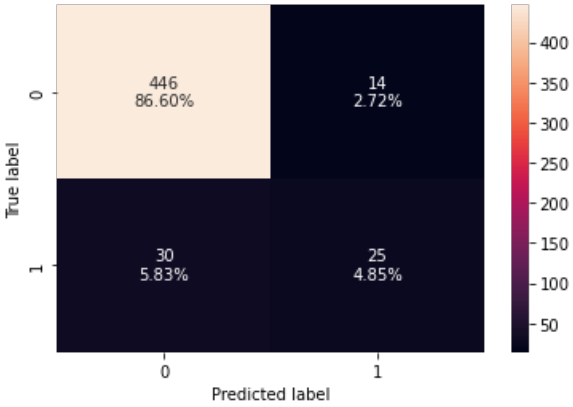
LOGISTIC REGRESSION MODEL

```
Optimization terminated successfully.
         Current function value: 0.179973
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1543
Model:                          Logit   Df Residuals:                     1489
Method:                          MLE   Df Model:                           53
Date:               Sat, 01 Jun 2024   Pseudo R-squ.:                 0.4708
Time:                       22:43:01   Log-Likelihood:                -277.70
converged:                     True   LL-Null:                       -524.71
Covariance Type:           nonrobust   LLR p-value:                7.760e-73
==========================================================================================================
                                             coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                     -4.0801      0.262    -15.561      0.000      -4.594      -3.566
Operating_Expense_Rate                     0.2270      0.136      1.668      0.095      -0.040       0.494
Research_and_development_expense_rate      0.4302      0.121      3.541      0.000       0.192       0.668
Cash_flow_rate                             0.2130      0.512      0.416      0.678      -0.791       1.217
Interest_bearing_debt_interest_rate        0.3661      0.147      2.499      0.012       0.079       0.653
Tax_rate_A                                -0.1344      0.174     -0.772      0.440      -0.476       0.207
Cash_Flow_Per_Share                       -0.0744      0.326     -0.229      0.819      -0.713       0.564
Per_Share_Net_profit_before_tax_Yuan_      0.4946      1.271      0.389      0.697      -1.996       2.985
Realized_Sales_Gross_Profit_Growth_Rate    0.0482      0.154      0.314      0.754      -0.253       0.349
Operating_Profit_Growth_Rate               0.0304      0.186      0.163      0.870      -0.334       0.394
Continuous_Net_Profit_Growth_Rate         -0.3620      0.203     -1.788      0.074      -0.759       0.035
Total_Asset_Growth_Rate                   -0.0820      0.134     -0.613      0.540      -0.344       0.180
Net_Value_Growth_Rate                     -0.0607      0.173     -0.351      0.725      -0.399       0.278
Total_Asset_Return_Growth_Rate_Ratio       0.3462      0.193      1.796      0.072      -0.032       0.724
Cash_Reinvestment_perc                     0.0750      0.451      0.166      0.868      -0.808       0.958
Current_Ratio                              0.4702      0.599      0.785      0.432      -0.703       1.643
Quick_Ratio                               -1.9551      0.574     -3.406      0.001      -3.080      -0.830
Interest_Expense_Ratio                    -0.0106      0.195     -0.054      0.957      -0.393       0.372
Total_debt_to_Total_net_worth              0.6133      0.227      2.703      0.007       0.169       1.058
Long_term_fund_suitability_ratio_A         0.2680      0.192      1.398      0.162      -0.108       0.644
Net_profit_before_tax_to_Paid_in_capital  -0.8533      1.322     -0.645      0.519      -3.444       1.738
Total_Asset_Turnover                      -0.2119      0.415     -0.511      0.609      -1.024       0.601
Accounts_Receivable_Turnover              -0.7083      0.213     -3.320      0.001      -1.126      -0.290
Average_Collection_Days                    0.1029      0.186      0.553      0.580      -0.262       0.467
Inventory_Turnover_Rate_times              0.0954      0.127      0.753      0.452      -0.153       0.344
Fixed_Assets_Turnover_Frequency            0.1249      0.154      0.809      0.419      -0.178       0.427
Net_Worth_Turnover_Rate_times             -0.0553      0.376     -0.147      0.883      -0.793       0.682
Operating_profit_per_person                0.1827      0.203      0.901      0.368      -0.215       0.580
Allocation_rate_per_person                 0.4790      0.192      2.497      0.013       0.103       0.855
Quick_Assets_to_Total_Assets              -0.6060      0.329     -1.843      0.065      -1.250       0.038
Cash_to_Total_Assets                       0.2166      0.202      1.075      0.283      -0.178       0.612
Quick_Assets_to_Current_Liability          1.3531      0.516      2.624      0.009       0.343       2.364
Cash_to_Current_Liability                  0.2084      0.185      1.126      0.260      -0.154       0.571
Operating_Funds_to_Liability               0.6072      0.712      0.853      0.394      -0.788       2.002
Inventory_to_Working_Capital              -0.1157      0.122     -0.945      0.344      -0.356       0.124
Inventory_to_Current_Liability            -0.2383      0.261     -0.912      0.362      -0.750       0.274
Long_term_Liability_to_Current_Assets     -0.2814      0.148     -1.902      0.057      -0.571       0.009
Retained_Earnings_to_Total_Assets         -0.5706      0.259     -2.200      0.028      -1.079      -0.062
Total_income_to_Total_expense             -0.6157      0.348     -1.767      0.077      -1.299       0.067
Total_expense_to_Assets                    0.4192      0.193      2.177      0.030       0.042       0.797
Current_Asset_Turnover_Rate                0.0232      0.143      0.162      0.872      -0.258       0.304
Quick_Asset_Turnover_Rate                  0.0222      0.140      0.159      0.874      -0.252       0.297
Cash_Turnover_Rate                        -0.3822      0.137     -2.787      0.005      -0.651      -0.113
Fixed_Assets_to_Assets                    -0.0261      0.221     -0.118      0.906      -0.460       0.408
Cash_Flow_to_Total_Assets                 -0.5500      1.036     -0.531      0.596      -2.581       1.481
Cash_Flow_to_Liability                     0.0074      0.806      0.009      0.993      -1.573       1.588
CFO_to_Assets                             -0.7657      0.731     -1.047      0.295      -2.199       0.668
Cash_Flow_to_Equity                        0.2887      0.448      0.644      0.519      -0.590       1.167
Current_Liability_to_Current_Assets        0.0472      0.324      0.146      0.884      -0.588       0.683
Total_assets_to_GNP_price                  0.0176      0.151      0.116      0.907      -0.278       0.314
No_credit_Interval                         0.1117      0.128      0.870      0.384      -0.140       0.363
Degree_of_Financial_Leverage_DFL           0.1159      0.165      0.701      0.484      -0.208       0.440
Interest_Coverage_Ratio_Interest_expense_to_EBIT  -0.0475  0.192  -0.247      0.805      -0.423       0.329
Equity_to_Liability                       -1.0116      0.387     -2.613      0.009      -1.771      -0.253
==========================================================================================================
```

Logistic regression is a statistical technique used to predict binary outcomes by estimating the probability of an event happening. It helps analyse factors that influence the occurrence of the outcome. By calculating odds based on independent variables, the model makes predictions and provides insights into the relationship between predictors and the outcome. It is useful when the outcome is categorical, offering valuable information for decision-making and risk assessment. In the logistic regression analysis, we will iteratively remove variables with p-values greater than 0.005 to refine the model. The p-value represents the probability of a random relationship between an independent variable and the dependent variable. By eliminating variables with higher p-values, we focus on the variables that have a significant impact on the likelihood of default. This step ensures that our model includes only the most influential variables, improving its accuracy and interpretability.

After training the logistic regression model on the training dataset, we validated its performance on the test dataset.





Train data

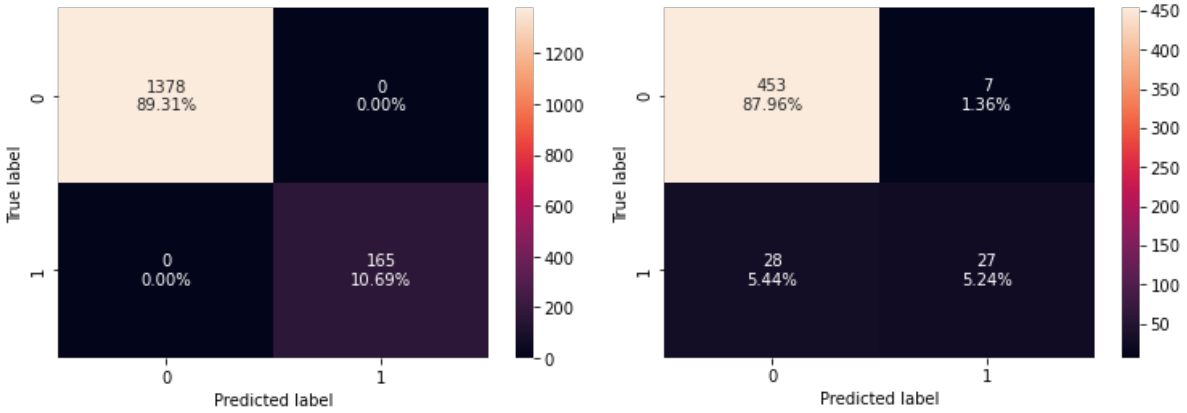
Accuracy	Recall	Precision	F1
0.92	0.50	0.70	0.58

Test data

Accuracy	Recall	Precision	F1
0.91	0.45	0.64	0.53

- The logistic regression model achieved an accuracy of 84% on the training data and 83.5% on the testing data.
- For the training data, the precision for class 0 (non-default) was 98%, indicating that 98% of the instances predicted as non-default were truly non-default. The precision for class 1 (default) was 37.1%, meaning that only 37.1% of the instances predicted as default were actually default.
- The recall for class 0 was 84.3%, indicating that the model correctly identified 84.3% of the true non-default cases. The recall for class 1 (default) was 76.7%, meaning that the model captured 76.7% of the true default cases.
- The confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives.

RANDOM FOREST MODEL



TRAIN DATA

Accuracy	Recall	Precision	F1
1.00	1.00	1.00	1.00

TEST DATA

Accuracy	Recall	Precision	F1
0.93	0.49	0.79	0.61

Random Forest Classifier is a popular machine learning algorithm that combines multiple decision trees to make predictions. By averaging the predictions of multiple trees, Random Forest Classifier provides accurate and reliable results.

The classification report shows that the model has a high precision of 0.94 for class 0 and 0.92 for class 1, indicating a good ability to correctly identify non-default and default cases, respectively. The recall score is 1.00 for class 0, indicating that the model accurately captures all non-default cases, while the recall score is 0.44 for class 1, suggesting that the model struggles to capture all default cases. The model's overall accuracy is 0.94, and the AUC score for the training data is 0.969.

PART A: Model Performance Improvement

VIF (Variance In?lation Factor) is a measure that tells us if there is a problem of multicollinearity in a regression model. Multicollinearity occurs when predictor variables are highly correlated with each other.

Firstly, to assess multicollinearity in the dataset, we calculated the VIF (Variance In?lation Factor) using the statsmodels library. VIF measures the extent to which predictor variables are correlated with each other. Higher VIF values indicate a stronger correlation, potentially leading to instability in the regression model's coef?icients. The calculated VIF values for each variable can help identify any multicollinearity issues present in the dataset.

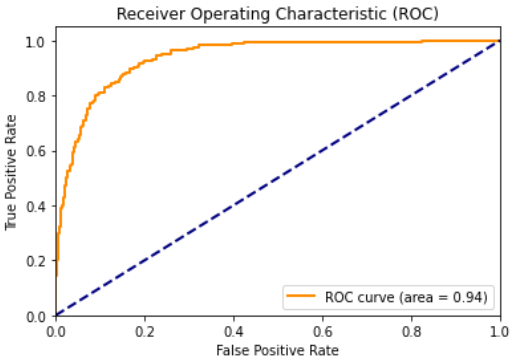
	Variable	VIF
0	Operating_Expense_Rate	1.37
1	Research_and_development_expense_rate	1.22
2	Cash_flow_rate	16.55
3	Interest_bearing_debt_interest_rate	1.14
4	Tax_rate_A	1.64
5	Cash_Flow_Per_Share	6.33
6	Per_Share_Net_profit_before_tax_Yuan_	88.13
7	Realized_Sales_Gross_Profit_Growth_Rate	2.92
8	Operating_Profit_Growth_Rate	3.82
9	Continuous_Net_Profit_Growth_Rate	3.67
10	Total_Asset_Growth_Rate	1.22
11	Net_Value_Growth_Rate	2.76
12	Total_Asset_Return_Growth_Rate_Ratio	3.39
13	Cash_Reinvestment_perc	14.43
14	Current_Ratio	16.59
15	Quick_Ratio	12.95
16	Interest_Expense_Ratio	4.63
17	Total_debt_to_Total_net_worth	5.05
18	Long_term_fund_suitability_ratio_A	2.90
19	Net_profit_before_tax_to_Paid_in_capital	87.77
20	Total_Asset_Turnover	14.80
21	Accounts_Receivable_Turnover	2.83
22	Average_Collection_Days	2.71
23	Inventory_Turnover_Rate_times	1.23
24	Fixed_Assets_Turnover_Frequency	1.97
25	Net_Worth_Turnover_Rate_times	16.02
26	Operating_profit_per_person	3.23
27	Allocation_rate_per_person	2.91
28	Quick_Assets_to_Total_Assets	6.69
29	Cash_to_Total_Assets	3.69
30	Quick_Assets_to_Current_Liability	25.36
31	Cash_to_Current_Liability	4.22
32	Operating_Funds_to_Liability	22.52
33	Inventory_to_Working_Capital	1.67
34	Inventory_to_Current_Liability	3.44
35	Long_term_Liability_to_Current_Assets	1.81
36	Retained_Earnings_to_Total_Assets	5.17
37	Total_income_to_Total_expense	5.18
38	Total_expense_to_Assets	2.31
39	Current_Asset_Turnover_Rate	1.65
40	Quick_Asset_Turnover_Rate	1.44
41	Cash_Turnover_Rate	1.13
42	Fixed_Assets_to_Assets	4.84
43	Cash_Flow_to_Total_Assets	47.94
44	Cash_Flow_to_Liability	18.73
45	CFO_to_Assets	31.68
46	Cash_Flow_to_Equity	16.22
47	Current_Liability_to_Current_Assets	8.50
48	Total_assets_to_GNP_price	1.77
49	No_credit_Interval	1.72
50	Degree_of_Financial_Leverage_DFL	3.65
51	Interest_Coverage_Ratio_Interest_expense_to_EBIT	5.91
52	Equity_to_Liability	6.27

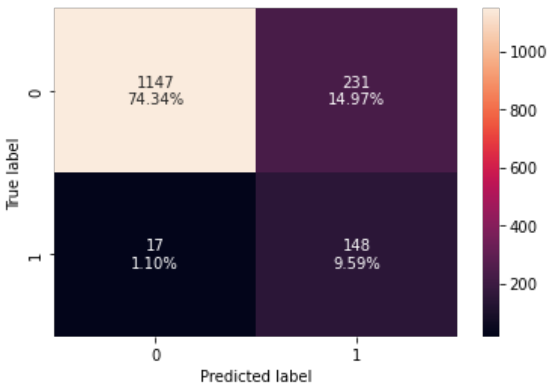
We calculated the VIF (Variance Inflation Factor) values to identify variables that have a strong correlation with each other. Higher VIF values indicate a higher degree of correlation. To ensure the stability of the regression model, we removed variables with VIF values above 5, as they could introduce multicollinearity issues. This selection process resulted in a set of variables with minimal correlation, enhancing the reliability of our analysis.

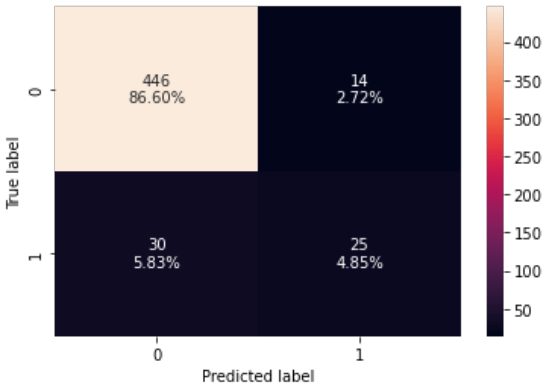
LOGISTIC REGRESSION

```
Optimization terminated successfully.
         Current function value: 0.179973
         Iterations 9
                         Logit Regression Results
==============================================================================
Dep. Variable:                Default   No. Observations:                 1543
Model:                          Logit   Df Residuals:                     1489
Method:                          MLE   Df Model:                           53
Date:               Sat, 01 Jun 2024   Pseudo R-squ.:                  0.4708
Time:                       22:22:29   Log-Likelihood:                -277.70
converged:                     True   LL-Null:                       -524.71
Covariance Type:           nonrobust   LLR p-value:                 7.760e-73
==========================================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
const                                                -4.0801      0.262    -15.561      0.000      -4.594      -3.566
Operating_Expense_Rate                                0.2270      0.136      1.668      0.095      -0.040       0.494
Research_and_development_expense_rate                 0.4302      0.121      3.541      0.000       0.192       0.668
Cash_flow_rate                                        0.2130      0.512      0.416      0.678      -0.791       1.217
Interest_bearing_debt_interest_rate                   0.3661      0.147      2.499      0.012       0.079       0.653
Tax_rate_A                                           -0.1344      0.174     -0.772      0.440      -0.476       0.207
Cash_Flow_Per_Share                                  -0.0744      0.326     -0.229      0.819      -0.713       0.564
Per_Share_Net_profit_before_tax_Yuan_                 0.4946      1.271      0.389      0.697      -1.996       2.985
Realized_Sales_Gross_Profit_Growth_Rate               0.0482      0.154      0.314      0.754      -0.253       0.349
Operating_Profit_Growth_Rate                          0.0304      0.186      0.163      0.870      -0.334       0.394
Continuous_Net_Profit_Growth_Rate                    -0.3620      0.203     -1.788      0.074      -0.759       0.035
Total_Asset_Growth_Rate                              -0.0820      0.134     -0.613      0.540      -0.344       0.180
Net_Value_Growth_Rate                                -0.0607      0.173     -0.351      0.725      -0.399       0.278
Total_Asset_Return_Growth_Rate_Ratio                  0.3462      0.193      1.796      0.072      -0.032       0.724
Cash_Reinvestment_perc                                0.0750      0.451      0.166      0.868      -0.808       0.958
Current_Ratio                                         0.4702      0.599      0.785      0.432      -0.703       1.643
Quick_Ratio                                          -1.9551      0.574     -3.406      0.001      -3.080      -0.830
Interest_Expense_Ratio                               -0.0106      0.195     -0.054      0.957      -0.393       0.372
Total_debt_to_Total_net_worth                         0.6133      0.227      2.703      0.007       0.169       1.058
Long_term_fund_suitability_ratio_A                    0.2680      0.192      1.398      0.162      -0.108       0.644
Net_profit_before_tax_to_Paid_in_capital             -0.8533      1.322     -0.645      0.519      -3.444       1.738
Total_Asset_Turnover                                 -0.2119      0.415     -0.511      0.609      -1.024       0.601
Accounts_Receivable_Turnover                         -0.7083      0.213     -3.320      0.001      -1.126      -0.290
Average_Collection_Days                               0.1029      0.186      0.553      0.580      -0.262       0.467
Inventory_Turnover_Rate_times                         0.0954      0.127      0.753      0.452      -0.153       0.344
Fixed_Assets_Turnover_Frequency                       0.1249      0.154      0.809      0.419      -0.178       0.427
Net_Worth_Turnover_Rate_times                        -0.0553      0.376     -0.147      0.883      -0.793       0.682
Operating_profit_per_person                           0.1827      0.203      0.901      0.368      -0.215       0.580
Allocation_rate_per_person                            0.4790      0.192      2.497      0.013       0.103       0.855
Quick_Assets_to_Total_Assets                         -0.6060      0.329     -1.843      0.065      -1.250       0.038
Cash_to_Total_Assets                                  0.2166      0.202      1.075      0.283      -0.178       0.612
Quick_Assets_to_Current_Liability                     1.3531      0.516      2.624      0.009       0.343       2.364
Cash_to_Current_Liability                             0.2084      0.185      1.126      0.260      -0.154       0.571
Operating_Funds_to_Liability                          0.6072      0.712      0.853      0.394      -0.788       2.002
Inventory_to_Working_Capital                         -0.1157      0.122     -0.945      0.344      -0.356       0.124
Inventory_to_Current_Liability                       -0.2383      0.261     -0.912      0.362      -0.750       0.274
Long_term_Liability_to_Current_Assets                -0.2814      0.148     -1.902      0.057      -0.571       0.009
Retained_Earnings_to_Total_Assets                    -0.5706      0.259     -2.200      0.028      -1.079      -0.062
Total_income_to_Total_expense                        -0.6157      0.348     -1.767      0.077      -1.299       0.067
Total_expense_to_Assets                               0.4192      0.193      2.177      0.030       0.042       0.797
Current_Asset_Turnover_Rate                           0.0232      0.143      0.162      0.872      -0.258       0.304
Quick_Asset_Turnover_Rate                             0.0222      0.140      0.159      0.874      -0.252       0.297
Cash_Turnover_Rate                                   -0.3822      0.137     -2.787      0.005      -0.651      -0.113
Fixed_Assets_to_Assets                               -0.0261      0.221     -0.118      0.906      -0.460       0.408
Cash_Flow_to_Total_Assets                            -0.5500      1.036     -0.531      0.596      -2.581       1.481
Cash_Flow_to_Liability                                0.0074      0.806      0.009      0.993      -1.573       1.588
CFO_to_Assets                                        -0.7657      0.731     -1.047      0.295      -2.199       0.668
Cash_Flow_to_Equity                                   0.2887      0.448      0.644      0.519      -0.590       1.167
Current_Liability_to_Current_Assets                   0.0472      0.324      0.146      0.884      -0.588       0.683
Total_assets_to_GNP_price                             0.0176      0.151      0.116      0.907      -0.278       0.314
No_credit_Interval                                    0.1117      0.128      0.870      0.384      -0.140       0.363
Degree_of_Financial_Leverage_DFL                      0.1159      0.165      0.701      0.484      -0.208       0.440
Interest_Coverage_Ratio_Interest_expense_to_EBIT     -0.0475      0.192     -0.247      0.805      -0.423       0.329
Equity_to_Liability                                  -1.0116      0.387     -2.613      0.009      -1.771      -0.253
==========================================================================================================
```

The optimal logit threshold = `0.093`







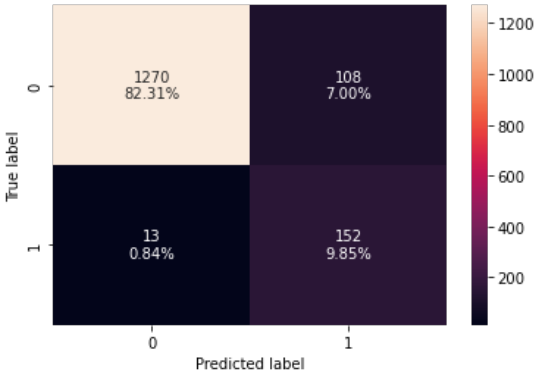
TRAIN DATA

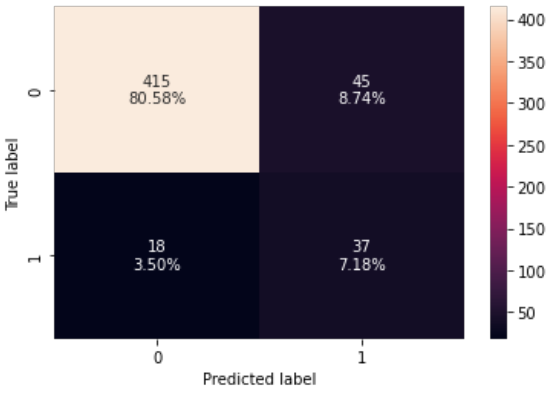
Accuracy	Recall	Precision	F1
0.84	0.90	0.39	0.54

TEST DATA

Accuracy	Recall	Precision	F1
0.91	0.45	0.64	0.53

RANDOM FOREST





TRAIN DATA

Accuracy	Recall	Precision	F1
0.92	0.92	0.58	0.72

TEST DATA

Accuracy	Recall	Precision	F1
0.88	0.67	0.45	0.54

PART A: Model Performance Comparison and Final Model Selection

Training performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.92	0.84	1.00	0.92
Recall	0.50	0.90	1.00	0.92
Precision	0.70	0.39	1.00	0.58
F1	0.58	0.54	1.00	0.72

Testing performance comparison:

	Logistic Regression	Tuned Logistic Regression	Random Forest	Tuned Random Forest
Accuracy	0.91	0.91	0.93	0.88
Recall	0.45	0.45	0.49	0.67
Precision	0.64	0.64	0.79	0.45
F1	0.53	0.53	0.61	0.54

1. Logistic Regression Model:

The model achieved an accuracy of 83.5% on the test data, with a precision of 96% for non-default cases and 37.1% for default cases.
The recall was 84.3% for non-default cases and 76.7% for default cases.
The AUC score was 0.893, indicating reasonable discriminative power.

The model shows stable performance, with no significant signs of overfitting or underfitting.

2. Random Forest Model:

The model achieved an accuracy of 91% on the test data, with high precision and recall for non-default cases but lower values for default cases.
The AUC score was 0.913, suggesting good discriminatory ability.
There is a slight indication of overfitting, as the model performed better on the training data compared to the test data.

In summary, the logistic regression model showed good performance in predicting non-default cases but had limitations in identifying default cases. The random forest model exhibited high precision and recall for non-default cases but faced challenges in accurately identifying default cases and showed slight overfitting. The LDA model demonstrated reasonably good performance in predicting both non-default and default cases, with no clear signs of overfitting or underfitting. Further improvements may be needed to enhance the models' performance in accurately identifying default cases.

PART A: Actionable Insights & Recommendations

CONCLUSION

These findings indicate that the Logistic Regression model performs steadily at an ideal threshold of 0.1076. It consistently makes predictions on both datasets, achieving an accuracy of 84% on the training data and 83.5% on the testing data. Reliability in detecting instances from both groups is suggested by the consistent precision and recall values for both default and non-default cases. The model's consistent ability to discriminate between default and non-default scenarios is further supported by the AUC ratings. In terms of default prediction, the Logistic Regression model with the ideal threshold performs steadily and consistently. To improve the models' ability to correctly detect default instances, more advancements could be required.

RECOMMENDATIONS:

• Collect more data, especially on defaults, to increase the model's exposure to defaults and improve its forecasting performance.
• Further exploration of the model using feature engineering can help improve its performance in detecting default cases.
• Include techniques such as boosting or compression to improve the performance of the model in detecting defaults..

PART B: Define the problem and perform Exploratory Data Analysis

CONTEXT

Investors face market risk, arising from asset price fluctuations due to economic events, geopolitical developments, and investor sentiment changes. Understanding and analysing this risk is crucial for informed decision-making and optimizing investment strategies.

OBJECTIVE

The objective of this analysis is to conduct Market Risk Analysis on a portfolio of Indian stocks using Python. It uses historical stock price data to understand market volatility and riskiness. Using statistical measures like mean and standard deviation, investors gain a deeper understanding of individual stocks' performance and portfolio variability.

Through this analysis, investors can aim to achieve the following objectives:

- Risk Assessment: Analyse the historical volatility of individual stocks and the overall portfolio.
- Portfolio Optimization: Use Market Risk Analysis insights to enhance risk-adjusted returns.
- Performance Evaluation: Assess portfolio management strategies' effectiveness in mitigating market risk.
- Portfolio Performance Monitoring: Monitor portfolio performance over time and adjust as market conditions and risk preferences change.

EXPLORATORY DATA ANALYSIS (EDA)

- Based on the information above, it is clear that the dataset contains 418 rows with 6 features, which include stock information of different companies.
- The dataset contains 5 numerical data and 1 categorical data, and it is also evident that there are no missing values present in the dataset.
- There are no duplicate values present in the dataset.

	Date	Dish TV	Infosys	Hindustan Unilever	Vodafone Idea	Cipla
0	28-03-2016	86	608	867	67	514
1	04-04-2016	86	607	863	65	519
2	11-04-2016	85	583	853	66	506
3	18-04-2016	87	625	900	69	515
4	25-04-2016	89	606	880	71	532

```
df.tail()
```

	Date	Dish TV	Infosys	Hindustan Unilever	Vodafone Idea	Cipla
413	2024-02-26	24	1675	2391	18	1474
414	2024-04-03	22	1658	2416	15	1483
415	2024-11-03	20	1608	2409	14	1493
416	2024-03-18	18	1630	2321	13	1489
417	2024-03-25	18	1509	2257	13	1481

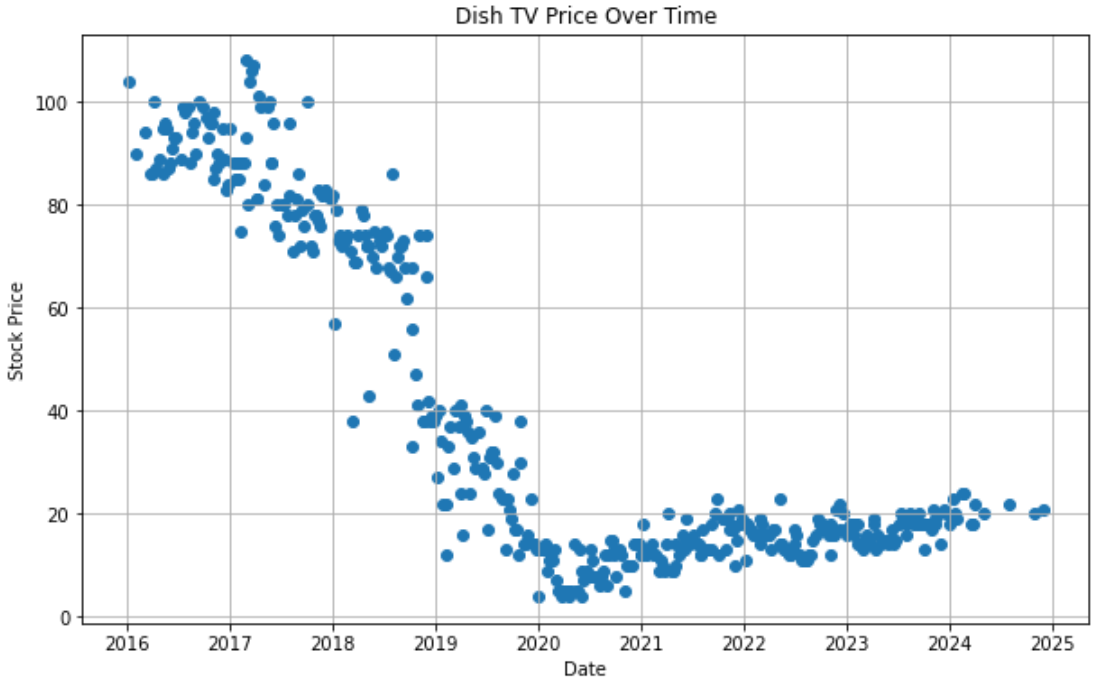
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Date                418 non-null    object
 1   Dish TV             418 non-null    int64
 2   Infosys             418 non-null    int64
 3   Hindustan Unilever  418 non-null    int64
 4   Vodafone Idea       418 non-null    int64
 5   Cipla               418 non-null    int64
dtypes: int64(5), object(1)
memory usage: 19.7+ KB
```

	Dish TV	Infosys	Hindustan Unilever	Vodafone Idea	Cipla
count	418.000000	418.000000	418.000000	418.000000	418.000000
mean	38.648325	1007.210526	1906.344498	23.234450	756.614833
std	31.944620	455.089501	597.800173	20.264854	252.969619
min	4.000000	445.000000	788.000000	3.000000	370.000000
25%	14.000000	591.250000	1368.500000	9.000000	556.000000
50%	19.500000	777.500000	2083.000000	12.000000	637.000000
75%	73.000000	1454.000000	2419.000000	43.000000	946.000000
max	108.000000	1939.000000	2798.000000	71.000000	1493.000000

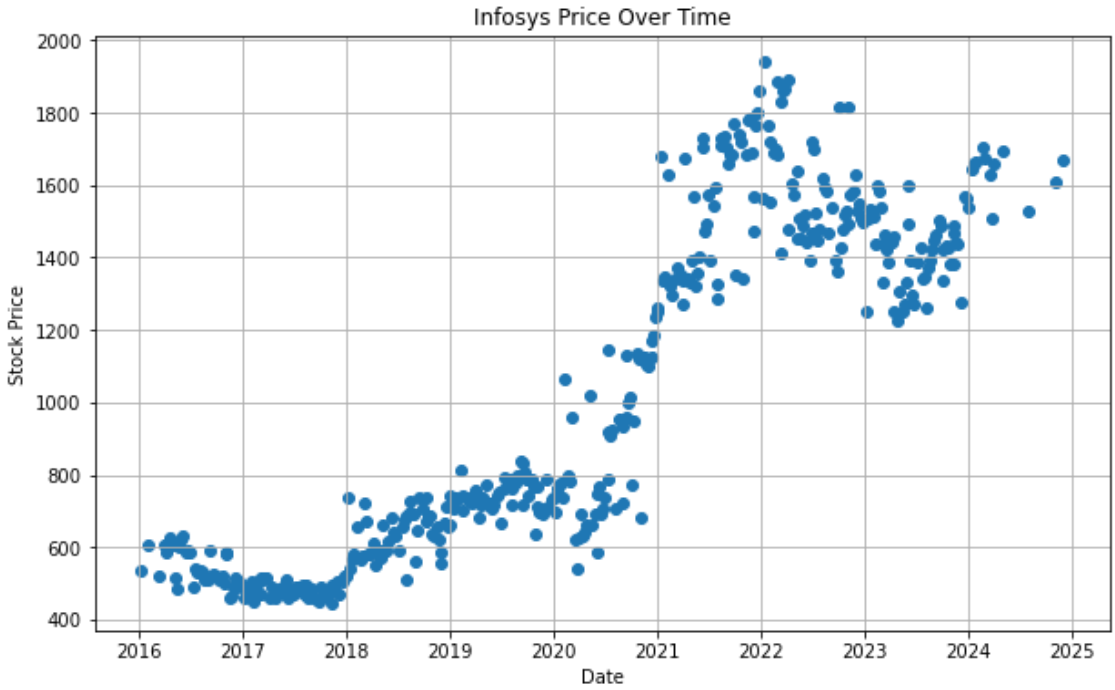
The summary of stock prices for ten companies is as follows:

- The average stock price ranges from 23.23 to 1906.34 for Vodafone to Hindustan Unilever respectively.
- The standard deviation indicates variability in stock prices, ranging from 135.95 to 202.26 for Infosys to Jet Airways, respectively.
- The minimum and maximum stock prices vary widely, with Infosys having the lowest at 234 and Jet Airways having the highest at 871.

PART B: Stock Price Graph Analysis

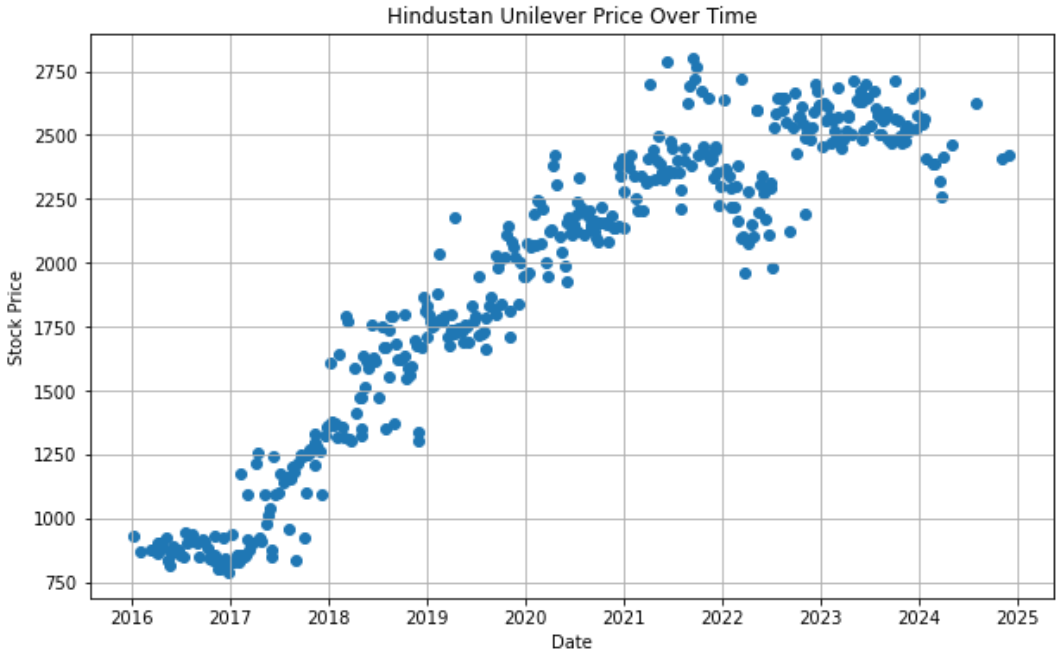


The scatter plot provides the trend in Dish Tv prices over the years. Observing the plotted data points, it becomes evident that there have been fluctuations in the stock prices throughout the observed period.
There's a gradual decline from 2016 to 2020 till it reaches a low point and then a more slight increase but mostly a constant over the next few years.

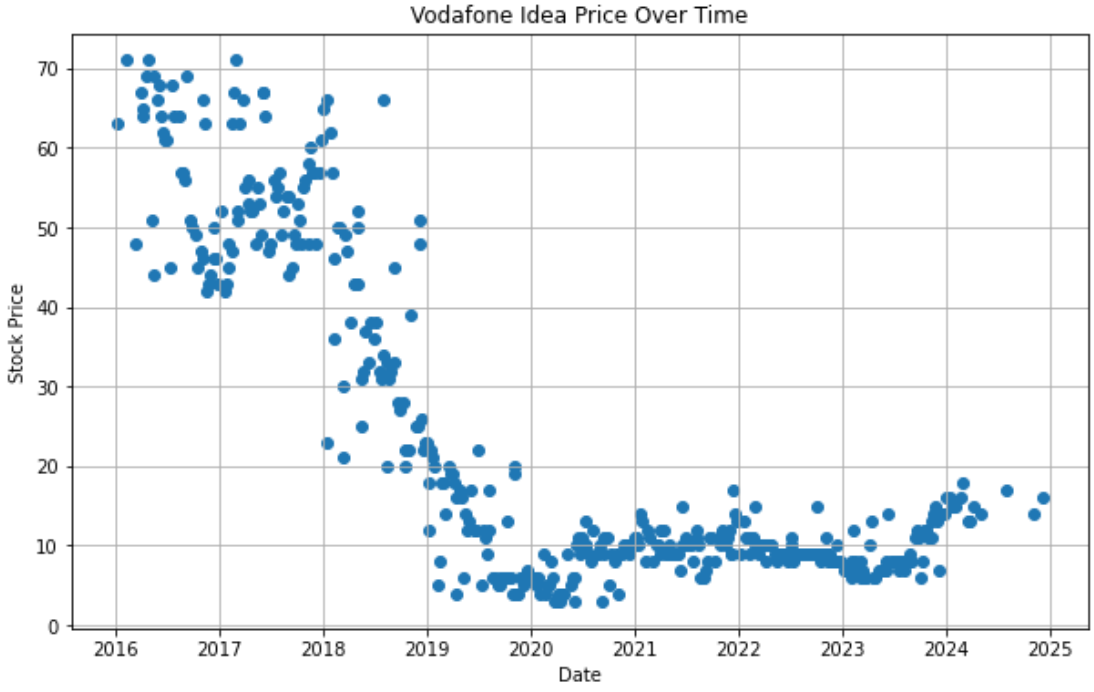


The scatter plot provides the trend in Infosys stock prices over the years. Observing the plotted data points, it becomes evident that there have been fluctuations in the stock prices throughout the observed period. From 2016 to around 2018, Infosys

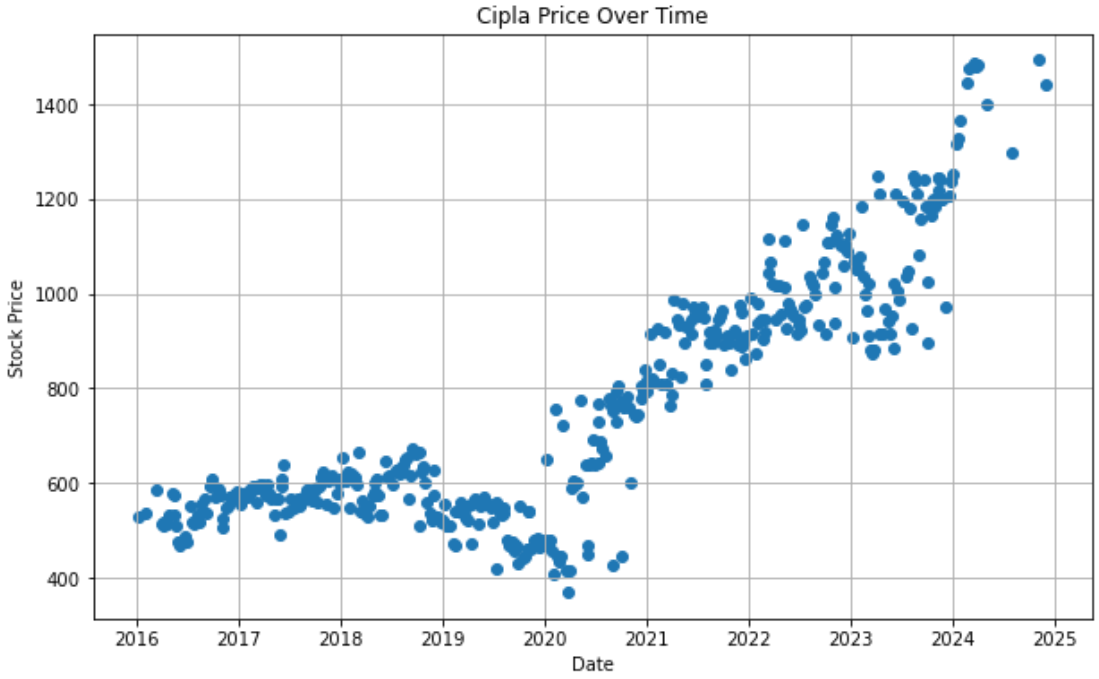
experienced a gradual increase in its stock value. However, after 2018, there was a decline in stock prices until it reached a low point. Subsequently, from 2019 onwards, the stock prices started to show a gradual upward trend, with a consistent increase until 2021.



The scatter plot provides the trend in Hindustan Unilever prices over the years. Observing the plotted data points, it becomes evident that there have been fluctuations in the stock prices throughout the observed period.
There's a gradual increase from 2017 till 2023 after which it seems to fall



The scatter plot provides the trend in Vodaphone Idea prices over the years. Observing the plotted data points, it becomes evident that there have been fluctuations in the stock prices throughout the observed period.
There's a drop from 2016 till 2020 after which there a very slow growth



The scatter plot provides the trend in Cipla prices over the years. Observing the plotted data points, it becomes evident that there have been fluctuations in the stock prices throughout the observed period.
There's a small increase from 2016 to 2018 and then dips back after 2020 there an gradual increase.

PART B: Stock Returns Calculation and Analysis

	Dish TV	Infosys	Hindustan Unilever	Vodafone Idea	Cipla
0	NaN	NaN	NaN	NaN	NaN
1	0.000000	-0.001646	-0.004624	-0.030305	0.009681
2	-0.011696	-0.040342	-0.011655	0.015267	-0.025367
3	0.023257	0.069564	0.053635	0.044452	0.017630
4	0.022728	-0.030872	-0.022473	0.028573	0.032477
5	0.011173	-0.001652	-0.014883	0.000000	0.009355
6	0.000000	-0.023412	-0.018627	-0.028573	0.001860
7	0.064539	0.023412	-0.021378	0.000000	-0.007463
8	-0.010471	-0.004971	-0.019395	-0.044452	-0.045985
9	-0.087969	0.031074	0.055934	0.029853	-0.066894
10	0.011429	0.017558	0.027399	-0.060625	-0.016914
11	0.033523	-0.072162	-0.023933	-0.031749	0.012712
12	0.021740	0.003396	0.009185	-0.016261	0.024949
13	0.000000	-0.006803	-0.019620	0.000000	-0.024949
14	0.072571	0.003407	0.047791	0.048009	0.073055
15	-0.020203	-0.006826	0.029559	-0.015748	0.027029
16	0.010152	-0.080185	0.018173	0.076373	-0.015355
17	-0.010152	-0.013072	-0.047731	-0.060625	-0.005820
18	0.059423	0.007491	0.032790	-0.015748	0.028765
19	-0.049271	-0.003738	-0.003231	0.015748	0.009407
20	-0.117783	-0.005634	0.008593	-0.115832	-0.034289
21	0.065958	-0.042314	-0.024907	0.000000	0.076458
22	0.021053	0.000000	-0.006601	-0.017700	0.019556
23	-0.010471	0.013659	0.016421	-0.093526	0.017452
24	0.000000	-0.005831	0.004334	-0.019803	-0.010435
25	0.051293	0.019306	-0.010870	0.019803	0.039422
26	-0.010050	-0.003831	-0.006579	-0.019803	0.023257

The above stock head shows the log returns of the stock prices for 5 different companies over the period. The dataset contains 417 rows and 5 columns, representing the stock returns for each company.

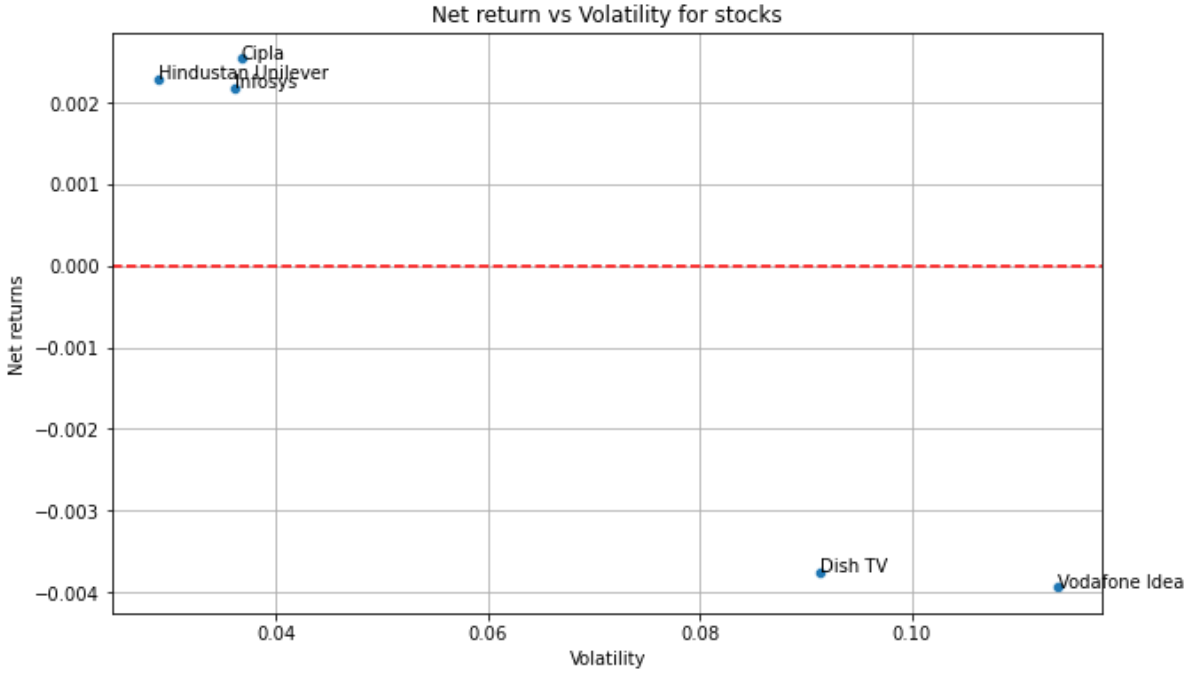
The log returns provide insights into the daily percentage change in stock prices for each company. Negative log returns indicate a decrease in stock prices, while positive log returns suggest an increase. And Important to note that the log return for the first day is NaN, as there is no previous day's data to calculate the return.

```
Vodafone Idea        -0.003932
Dish TV              -0.003751
Infosys               0.002180
Hindustan Unilever    0.002294
Cipla                 0.002538
dtype: float64
```

The mean returns provide insights into the average daily performance of each stock, with some showing positive returns and others showing negative returns. The standard deviations reflect the level of volatility or risk associated with each stock, with higher standard deviations indicating higher price fluctuations.

```
Hindustan Unilever    0.028845
Infosys               0.036102
Cipla                 0.036759
Dish TV               0.091333
Vodafone Idea         0.113747
dtype: float64
```

	Mean	Volatility
Dish TV	-0.003751	0.091333
Infosys	0.002180	0.036102
Hindustan Unilever	0.002294	0.028845
Vodafone Idea	-0.003932	0.113747
Cipla	0.002538	0.036759



The scatter plot presents mean returns and volatility for various companies. The two companies with the highest mean returns are Cipla and Hindustan Unilever.

The two companies with the lowest mean returns are Vodafone Idea and Dish TV. Among the two highest mean return companies, Hindustan Unilever has the lower volatility compared to Cipla , making it a more stable investment choice. Similarly, among the two lowest mean return companies, Dish TV has a slightly lower volatility than Vodafone Idea.

PART B: Actionable Insights & Recommendations

Conclusion:

Based on the analysis of stock data for different companies, several insights can be drawn. Cipla and Hindustan Unilever stand out as companies with the highest mean returns, indicating potentially favourable investment opportunities. On the other hand, Idea Vodafone and Dish TV have the lowest mean returns, suggesting caution while considering these companies for investment.

Recommendations:

- For investors looking to earn higher profits, considering companies like Cipla and Hindustan Unilever could be a good idea. These companies have consistently shown higher returns over the period, making them appealing choices for investors who want to maximize their investment gains.
- For investors who prefer to play it safe and avoid taking too much risk, it is recommended to be cautious when investing in companies like Idea Vodafone and Dish TV , which have shown lower mean returns.
- Short-term fluctuations are common, and holding onto investments for an extended period can help ride out market volatility and potentially yield higher returns.
- Investors should regularly keep an eye on how their investments are doing and stay informed about the latest trends in the market.