Great Learning

Financial Risk Analytics Project - Business Report

Submitted by:

N. Aishwarya

PGP-DSBA

October 2022

Business Report Outline

List of Tables

List of Formulas

List of Figures

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 interests 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 that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Net worth of the company in the following year (2016) is provided which can be used to drive the labeled field.

We need to create a default variable that should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

Data dictionary as described below:

S. No	Field Name	Description	New Field Name
1	Co_Code	Company Code	Co_Code
2	Co_Name	Company Name	Co_Name
3	Networth Next Year	Value of a company as on 2016 - Next Year (difference between the value of total assets and total liabilities)	Networth_Next_Year
4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders	Equity_Paid_Up
5	Networth	Value of a company as on 2015 - Current Year	Networth
6	Capital Employed	Total amount of capital used for the acquisition of profits by a company	Capital_Employed
7	Total Debt	The sum of money borrowed by the company and is due to be paid	Total Debt
8	Gross Block	Total value of all the assets that a company owns	Gross Block
9	Net Working Capital	The difference between a company's current assets (cash, accounts receivable, inventories of raw materials and finished goods) and its current liabilities (accounts payable).	Net_Working_Capital
10	Current Assets	All the assets of a company that are expected to be sold or used because of standard business operations over the next year.	Curr_Assets
11	Current Liabilities and Provisions	Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)	Curr_Liab_and_Prov
12	Total Assets/Liabilities	Ratio of total assets to liabilities of the company	Total_Assets_to_Liab
13	Gross Sales	The grand total of sale transactions within the accounting period	Gross_Sales
14	Net Sales	Gross sales minus returns, allowances, and discounts	Net_Sales
15	Other Income	Income realized from non-business activities (e.g., sale of long-term asset)	Other Income
16	Value Of Output	Product of physical output of goods and services produced by company and its market price	Value_Of_Output

17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service	Cost_of_Prod
18	Selling Cost	Costs which are made to create the demand for the product (advertising expenditures, packaging and styling, salaries, commissions and travelling expenses of sales personnel, and the cost of shops and showrooms)	Selling_Cost
19	PBIDT	Profit Before Interest, Depreciation & Taxes	PBIDT
20	PBDT	Profit Before Depreciation and Tax	PBDT
21	PBIT	Profit before interest and taxes	PBIT
22	PBT	Profit before tax	PBT
23	PAT	Profit After Tax	PAT
24	Adjusted PAT	Adjusted profit is the best estimate of the true profit	Adjusted PAT
26	CP	Commercial paper, a short-term debt instrument to meet short-term liabilities.	CP
27	Revenue earnings in forex	Revenue earned in foreign currency	Rev_earn_in_forex
28	Revenue expenses in forex	Expenses due to foreign currency transactions	Rev_exp_in_forex
29	Capital expenses in forex	Long term investment in forex	Capital_exp_in_forex
30	Book Value (Unit Curr)	Net asset value	Book_Value_Unit_Curr
31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value	Book_Value_Adj_Unit_Curr
32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share	Market_Capitalisation
33	CEPS (annualised) (Unit Curr)	Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis	CEPS_annualised_Unit_Curr
34	Cash Flow from Operating Activities	Use of cash from ongoing regular business activities	Cash_Flow_From_Opr
35	Cash Flow from Investing Activities	Cash used in the purchase of non-current assets–or long-term assets– that will deliver value in the future	Cash_Flow_From_Inv
36	Cash Flow from Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)	Cash_Flow_From_Fin
37	ROG-Net Worth (%)	Rate of Growth - Networth	ROG_Net_Worth_perc
38	ROG-Capital Employed (%)	Rate of Growth - Capital Employed	ROG_Capital_Employed_perc
39	ROG-Gross Block (%)	Rate of Growth - Gross Block	ROG_Gross_Block_perc
40	ROG-Gross Sales (%)	Rate of Growth - Gross Sales	ROG_Gross_Sales_perc
41	ROG-Net Sales (%)	Rate of Growth - Net Sales	ROG_Net_Sales_perc
42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production	ROG_Cost_of_Prod_perc
43	ROG-Total Assets (%)	Rate of Growth - Total Assets	ROG_Total_Assets_perc
44	ROG-PBIDT (%)	Rate of Growth- PBIDT	ROG_PBIDT_perc
45	ROG-PBDT (%)	Rate of Growth- PBDT	ROG_PBDT_perc
46	ROG-PBIT (%)	Rate of Growth- PBIT	ROG_PBIT_perc
47	ROG-PBT (%)	Rate of Growth- PBT	ROG_PBT_perc
48	ROG-PAT (%)	Rate of Growth- PAT	ROG_PAT_perc
49	ROG-CP (%)	Rate of Growth- CP	ROG_CP_perc
50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex	ROG_Rev_earn_in_forex_perc
51	ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex	ROG_Rev_exp_in_forex_perc

52	ROG-Market Capitalisation (%)	Rate of Growth - Market Capitalisation	ROG_Market_Capitalisation_perc
53	Current Ratio [Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year	Curr_Ratio_Latest
54	Fixed Assets Ratio [Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating	Fixed_Assets_Ratio_Latest
55	Inventory Ratio [Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company	Inventory_Ratio_Latest
56	Debtors Ratio [Latest]	Measures how quickly cash debtors are paying back to the company	Debtors_Ratio_Latest
57	Total Asset Turnover Ratio [Latest]	The value of a company's revenues relative to the value of its assets	Total_Asset_Turnover_Ratio_Latest
58	Interest Cover Ratio [Latest]	Determines how easily a company can pay interest on its outstanding debt	Interest_Cover_Ratio_Latest
59	PBIDTM (%) [Latest]	Profit before Interest Depreciation and Tax Margin	PBIDTM_perc_Latest
60	PBITM (%) [Latest]	Profit Before Interest Tax Margin	PBITM_perc_Latest
61	PBDTM (%) [Latest]	Profit Before Depreciation Tax Margin	PBDTM_perc_Latest
62	CPM (%) [Latest]	Cost per thousand (advertising cost)	CPM_perc_Latest
63	APATM (%) [Latest]	After tax profit margin	APATM_perc_Latest
64	Debtors Velocity (Days)	Average days required for receiving the payments	Debtors_Vel_Days
65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers	Creditors_Vel_Days
66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales	Inventory_Vel_Days
67	Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets	Value_of_Output_to_Total_Assets
68	Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block	Value_of_Output_to_Gross_Block

Table 1: Data dictionary for the dataset

There is total 68 variables in this dataset. It contains various measures related to company business.

Exploratory Data Analysis:

FRA dataset data is loaded using pandas and the dataset has 3,586 observations (rows) and 67 variables (columns). A quick glimpse of the data is shown below:

	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	Current Liabilities and Provisions	Total Assets/Liabilities	Gross Sales	Net Sales
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	1116.85	109.60	0.00	0.00
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	1585.74	6043.94	2892.73	2892.73
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	4601.39	12316.07	392.13	392.13
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	3646.54	6000.42	1354.39	1354.39
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	2849.58	7524.91	38.72	38.72

Figure 1: Dataset overview

Description of variables are as below, to understand the data better:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3586 entries, 0 to 3585
Data columns (total 67 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Co_Code                               3586 non-null   int64
 1   Co_Name                               3586 non-null   object
 2   Networth_Next_Year                    3586 non-null   float64
 3   Equity_Paid_Up                        3586 non-null   float64
 4   Networth                              3586 non-null   float64
 5   Capital_Employed                      3586 non-null   float64
 6   Total_Debt                            3586 non-null   float64
 7   Gross_Block                           3586 non-null   float64
 8   Net_Working_Capital                   3586 non-null   float64
 9   Current_Assets                        3586 non-null   float64
 10  Current_Liabilities_and_Provisions    3586 non-null   float64
 11  Total_Assets_by_Liabilities           3586 non-null   float64
 12  Gross_Sales                           3586 non-null   float64
 13  Net_Sales                             3586 non-null   float64
 14  Other_Income                          3586 non-null   float64
 15  Value_Of_Output                       3586 non-null   float64
 16  Cost_of_Production                    3586 non-null   float64
 17  Selling_Cost                          3586 non-null   float64
 18  PBIDT                                 3586 non-null   float64
 19  PBDT                                  3586 non-null   float64
 20  PBIT                                  3586 non-null   float64
 21  PBT                                   3586 non-null   float64
 22  PAT                                   3586 non-null   float64
 23  Adjusted_PAT                          3586 non-null   float64
 24  CP                                    3586 non-null   float64
 25  Revenue_earnings_in_forex             3586 non-null   float64
 26  Revenue_expenses_in_forex             3586 non-null   float64
 27  Capital_expenses_in_forex             3586 non-null   float64
 28  Book_Value_Unit_Curr                  3586 non-null   float64
 29  Book_Value_Adj._Unit_Curr             3582 non-null   float64
 30  Market_Capitalisation                 3586 non-null   float64
 31  CEPS_annualised_Unit_Curr             3586 non-null   float64
 32  Cash_Flow_From_Operating_Activities   3586 non-null   float64
 33  Cash_Flow_From_Investing_Activities   3586 non-null   float64
 34  Cash_Flow_From_Financing_Activities   3586 non-null   float64
 35  ROG_Net_Worth_perc                    3586 non-null   float64
 36  ROG_Capital_Employed_perc             3586 non-null   float64
 37  ROG_Gross_Block_perc                  3586 non-null   float64
 38  ROG_Gross_Sales_perc                  3586 non-null   float64
 39  ROG_Net_Sales_perc                    3586 non-null   float64
 40  ROG_Cost_of_Production_perc           3586 non-null   float64
 41  ROG_Total_Assets_perc                 3586 non-null   float64
 42  ROG_PBIDT_perc                        3586 non-null   float64
 43  ROG_PBDT_perc                         3586 non-null   float64
 44  ROG_PBIT_perc                         3586 non-null   float64
 45  ROG_PBT_perc                          3586 non-null   float64
 46  ROG_PAT_perc                          3586 non-null   float64
 47  ROG_CP_perc                           3586 non-null   float64
 48  ROG_Revenue_earnings_in_forex_perc    3586 non-null   float64
 49  ROG_Revenue_expenses_in_forex_perc    3586 non-null   float64
 50  ROG_Market_Capitalisation_perc        3586 non-null   float64
 51  Current_Ratio_Latest                  3585 non-null   float64
 52  Fixed_Assets_Ratio_Latest             3585 non-null   float64
 53  Inventory_Ratio_Latest                3585 non-null   float64
 54  Debtors_Ratio_Latest                  3585 non-null   float64
 55  Total_Asset_Turnover_Ratio_Latest     3585 non-null   float64
 56  Interest_Cover_Ratio_Latest           3585 non-null   float64
 57  PBIDTM_perc_Latest                    3585 non-null   float64
 58  PBITM_perc_Latest                     3585 non-null   float64
 59  PBDTM_perc_Latest                     3585 non-null   float64
 60  CPM_perc_Latest                       3585 non-null   float64
 61  APATM_perc_Latest                     3585 non-null   float64
 62  Debtors_Velocity_Days                 3586 non-null   int64
 63  Creditors_Velocity_Days               3586 non-null   int64
 64  Inventory_Velocity_Days               3483 non-null   float64
 65  Value_of_Output_by_Total_Assets       3586 non-null   float64
 66  Value_of_Output_by_Gross_Block        3586 non-null   float64
dtypes: float64(63), int64(3), object(1)
```

Figure 2: Description of variables

Observations:

- The special characters in the variable names (Field names) have been replaced to get to the suggested variable names mentioned in data dictionary.
- There are 3586 rows and 67 columns (variables).
- All the variables are numeric type except one variable (Co_Name) which is object type.
- For our analysis, Co_Code and Co_Name are dropped.
- There is no duplicate entry in the dataset.
- The problem statement requires to predict “default” status of the company where the “Networth Next Year” of the company is used to drive the “default” field. The “default” is 1 when “Networth Next Year” is negative, and it is 0 when “Networth Next Year” is positive. The “Default” field is created and added to the dataset based on the condition mentioned above. Subsequently “Networth Next Year” is not considered further as it became redundant.
- There are missing values in 13 of the variables. Missing values will be treated with either mean or median values of corresponding variables.
- There are outliers in the dataset. It will be treated for our analysis.

Statistical description of the dataset:

	count	mean	std	min	25%	50%	75%	max
Co_Code	3586.0	16065.388734	19776.817379	4.00	3029.2500	6077.500	24269.5000	72493.00
Networth_Next_Year	3586.0	725.045251	4769.681004	-8021.60	3.9850	19.015	123.8025	111729.10
Equity_Paid_Up	3586.0	62.966584	778.761744	0.00	3.7500	8.290	19.5175	42263.46
Networth	3586.0	649.746299	4091.988792	-7027.48	3.8925	18.580	117.2975	81657.35
Capital_Employed	3586.0	2799.611054	26975.135385	-1824.75	7.6025	39.090	226.6050	714001.25
Total_Debt	3586.0	1994.823779	23652.842746	-0.72	0.0300	7.490	72.3500	652823.81
Gross_Block	3586.0	594.178829	4871.547802	-41.19	0.5700	15.870	131.8950	128477.59
Net_Working_Capital	3586.0	410.809665	6301.218546	-13162.42	0.9425	10.145	61.1750	223257.56
Current_Assets	3586.0	1960.349172	22577.570829	-0.91	4.0000	24.540	135.2775	721166.00
Current_Liabilities_and_Provisions	3586.0	391.992078	2675.001631	-0.23	0.7325	9.225	65.6500	83232.98
Total_Assets_by_Liabilities	3586.0	1778.453751	11437.574690	-4.51	10.5550	52.010	310.5400	254737.22
Gross_Sales	3586.0	1123.738985	10603.703837	-62.59	1.4425	31.210	242.2500	474182.94
Net_Sales	3586.0	1079.702579	9996.574173	-62.59	1.4400	30.440	234.4400	443775.16
Other_Income	3586.0	48.729824	426.040665	-448.72	0.0200	0.450	3.6350	14143.40
Value_Of_Output	3586.0	1077.187292	9843.880293	-119.10	1.4125	30.895	235.8375	435559.09
Cost_of_Production	3586.0	798.544621	9076.702982	-22.65	0.9400	25.990	189.5500	419913.50
Selling_Cost	3586.0	25.554997	194.244466	0.00	0.0000	0.160	3.8825	5283.91
PBIDT	3586.0	248.175282	1949.593350	-4655.14	0.0400	2.045	23.5250	42059.26

Figure 3: Statistical description of the dataset

The values of mean, standard deviation, minimum and maximum, 25th, 50th and 75th percentile is mentioned in the above tables.

The statistical summary shows that there are many outliers present in the dataset in almost all the variables. The median Networth is 18.58 units whereas minimum is -7027 units and maximum are 81657 units. This shows that how much deviation is there which is also proved by the difference in mean and std. deviation.

ROG-Revenue earnings in forex and expense variable shows that the 25th & 75th percentile are 0 which means they have a large chunk of data having zero which might not be contributing enough towards the output and can be eliminated. We will look further into the outliers to see how bad the situation is, whether any treatment is required or not

Checking for Duplicate values in dataset:

Number of duplicate rows = 0

Checking for NULL value:

Co_Code	0
Co_Name	0
Networth_Next_Year	0
Equity_Paid_Up	0
Networth	0
Capital_Employed	0
Total_Debt	0
Gross_Block	0
Net_Working_Capital	0
Current_Assets	0
Current_Liabilities_and_Provisions	0
Total_Assets_by_Liabilities	0
Gross_Sales	0
Net_Sales	0
Other_Income	0
Value_Of_Output	0
Cost_of_Production	0
Selling_Cost	0
PBIDT	0
PBDT	0
PBIT	0
PBT	0
PAT	0
Adjusted_PAT	0
CP	0
Revenue_earnings_in_forex	0
Revenue_expenses_in_forex	0
Capital_expenses_in_forex	0
Book_Value_Unit_Curr	0
Book_Value_Adj._Unit_Curr	4

ROG_Gross_Sales_perc	0
ROG_Net_Sales_perc	0
ROG_Cost_of_Production_perc	0
ROG_Total_Assets_perc	0
ROG_PBIDT_perc	0
ROG_PBDT_perc	0
ROG_PBIT_perc	0
ROG_PBT_perc	0
ROG_PAT_perc	0
ROG_CP_perc	0
ROG_Revenue_earnings_in_forex_perc	0
ROG_Revenue_expenses_in_forex_perc	0
ROG_Market_Capitalisation_perc	0
Current_Ratio_Latest	1
Fixed_Assets_Ratio_Latest	1
Inventory_Ratio_Latest	1
Debtors_Ratio_Latest	1
Total_Asset_Turnover_Ratio_Latest	1
Interest_Cover_Ratio_Latest	1
PBIDTM_perc_Latest	1
PBITM_perc_Latest	1
PBDTM_perc_Latest	1
CPM_perc_Latest	1
APATM_perc_Latest	1
Debtors_Velocity_Days	0
Creditors_Velocity_Days	0
Inventory_Velocity_Days	103
Value_of_Output_by_Total_Assets	0
Value_of_Output_by_Gross_Block	0
Default	0

Table 2: Null value check

There are null values in 13 of the variables. These null values are imputed with median values as mean may not be correct one as the data variations are more and skewed.

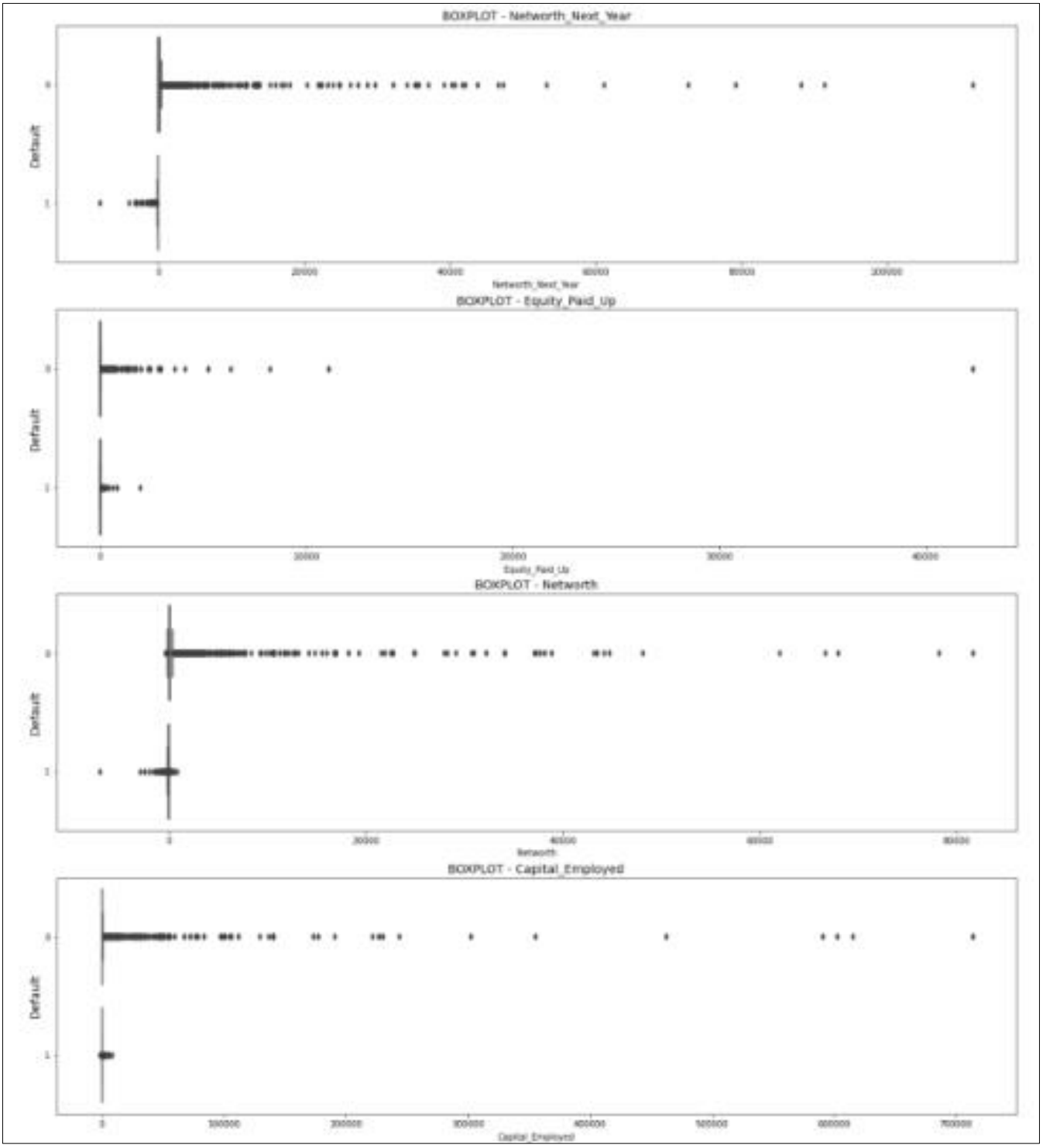
1.1 Outlier Treatment

Outliers are present in all the independent variables. For our dataset, we used IQR (Inter-Quartile Range) based calculation to treat the outliers. The following is the method,

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

Once the upper bound and lower bound range is calculated, we snap the values above upper range and values below lower range to upper and lower range values respectively. It was observed that maximum of 45% of the total rows are outliers for a particular variable in the dataset. And the mean numbers of outliers above and below the specified band is around 18%.

Below are boxplots which were plotted to analyse this data:



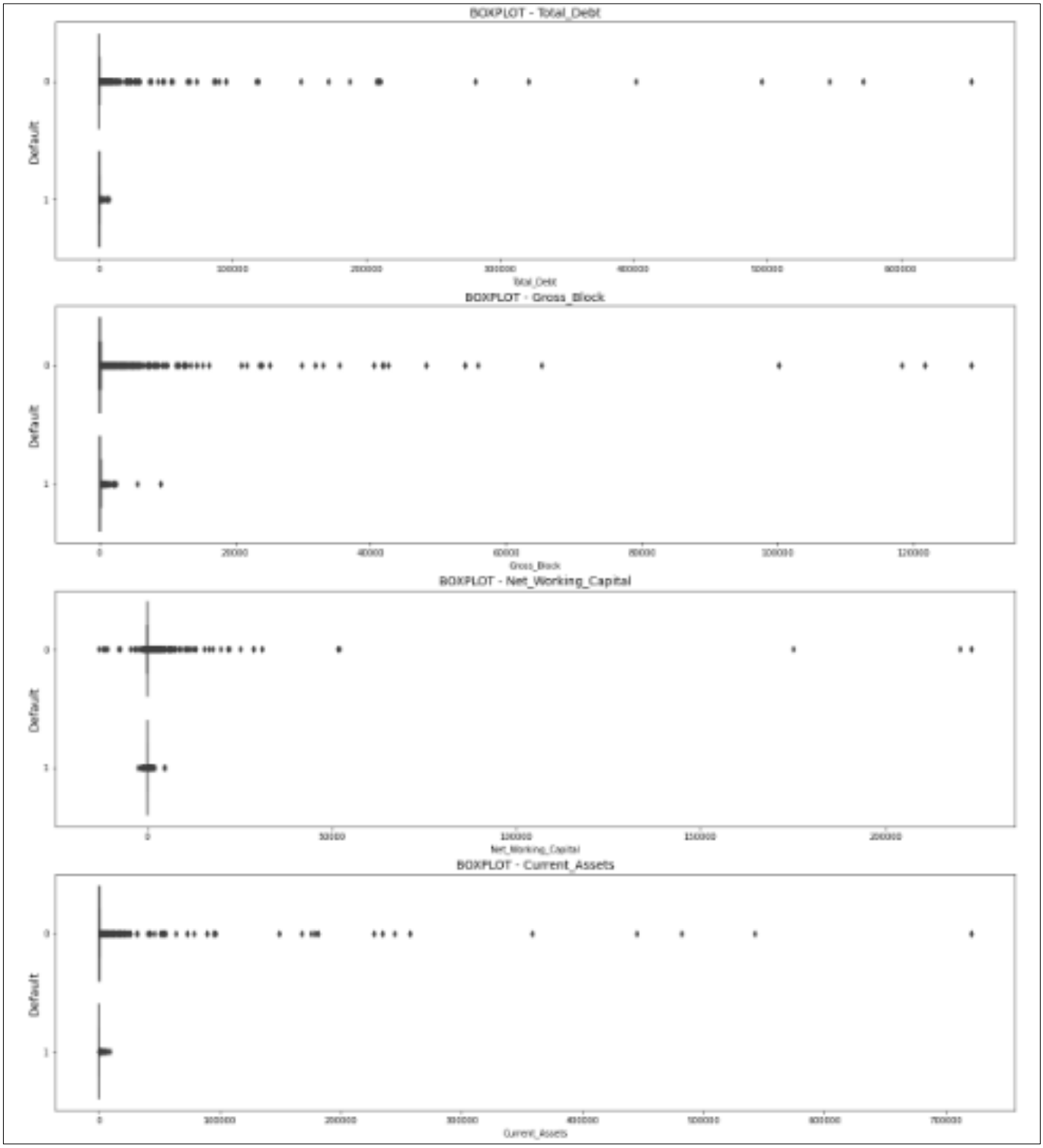


Figure 4: Boxplot prior to outlier treatment

Post treatment of outliers:

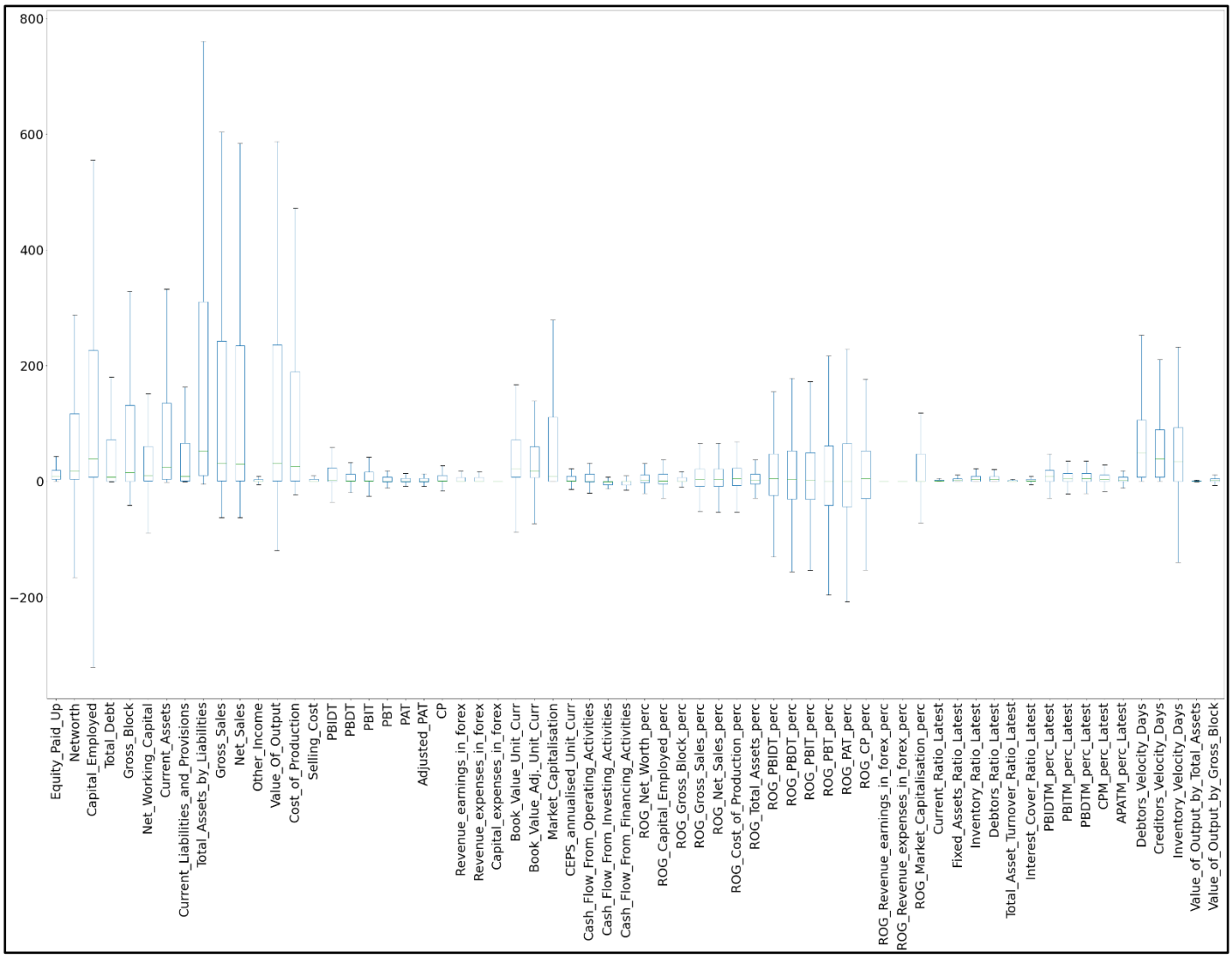


Figure 5: Plot, post outlier treatment

1.2 Missing Value Treatment

Missing value imputation:

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

The below figure proves that there are no missing values, post this median imputation.

Co_Code	0
Co_Name	0
Networth_Next_Year	0
Equity_Paid_Up	0
Networth	0
Capital_Employed	0
Total_Debt	0
Gross_Block	0
Net_Working_Capital	0
Current_Assets	0
Current_Liabilities_and_Provisions	0
Total_Assets_by_Liabilities	0
Gross_Sales	0
Net_Sales	0
Other_Income	0
Value_Of_Output	0
Cost_of_Production	0
Selling_Cost	0
PBIDT	0
PBDT	0
PBIT	0
PBT	0
PAT	0
Adjusted_PAT	0
CP	0
Revenue_earnings_in_forex	0
Revenue_expenses_in_forex	0
Capital_expenses_in_forex	0
Book_Value_Unit_Curr	0
Book_Value_Adj._Unit_Curr	0
Market_Capitalisation	0
CEPS_annualised_Unit_Curr	0
Cash_Flow_From_Operating_Activities	0
Cash_Flow_From_Investing_Activities	0
Cash_Flow_From_Financing_Activities	0
ROG_Net_Worth_perc	0
ROG_Capital_Employed_perc	0
ROG_Gross_Block_perc	0
ROG_Gross_Sales_perc	0
ROG_Net_Sales_perc	0
ROG_Cost_of_Production_perc	0
ROG_Total_Assets_perc	0
ROG_PBIDT_perc	0
ROG_PBDT_perc	0
ROG_PBIT_perc	0
ROG_PBT_perc	0
ROG_PAT_perc	0
ROG_CP_perc	0
ROG_Revenue_earnings_in_forex_perc	0
ROG_Revenue_expenses_in_forex_perc	0
ROG_Market_Capitalisation_perc	0
Current_Ratio_Latest	0
Fixed_Assets_Ratio_Latest	0
Inventory_Ratio_Latest	0
Debtors_Ratio_Latest	0
Total_Asset_Turnover_Ratio_Latest	0
Interest_Cover_Ratio_Latest	0
PBIDTM_perc_Latest	0
PBITM_perc_Latest	0
PBDTM_perc_Latest	0
CPM_perc_Latest	0
APATM_perc_Latest	0
Debtors_Velocity_Days	0
Creditors_Velocity_Days	0
Inventory_Velocity_Days	0
Value_of_Output_by_Total_Assets	0
Value_of_Output_by_Gross_Block	0
Default	0

Figure 6: Missing value treatment

1.3 Transform Target variable into 0 and 1

A new dependent variable named "Default" was created based on the criteria given in the project notes.

Criteria –

1 - If the Net Worth Next Year is negative for the company

0 - If the Net Worth Next Year is positive for the company

As required, a transformed target variable “Default” is added to the dataset based on whether the variable “Networth Next Year” is positive or negative. “Default” will take value as 0 if “Networth Next Year” is positive, otherwise “Default” is 1.

The below picture captures the new variable “Default”.

	Networth_Next_Year	Default
0	-8021.60	1
1	-3986.19	1
2	-3192.58	1
3	-3054.51	1
4	-2967.36	1
...	...	...
3581	72677.77	0
3582	79162.19	0
3583	88134.31	0
3584	91293.70	0
3585	111729.10	0

Also, the target variable “Default” is checked for counts.

```
0    3198
1     388
Name: Default, dtype: int64
```

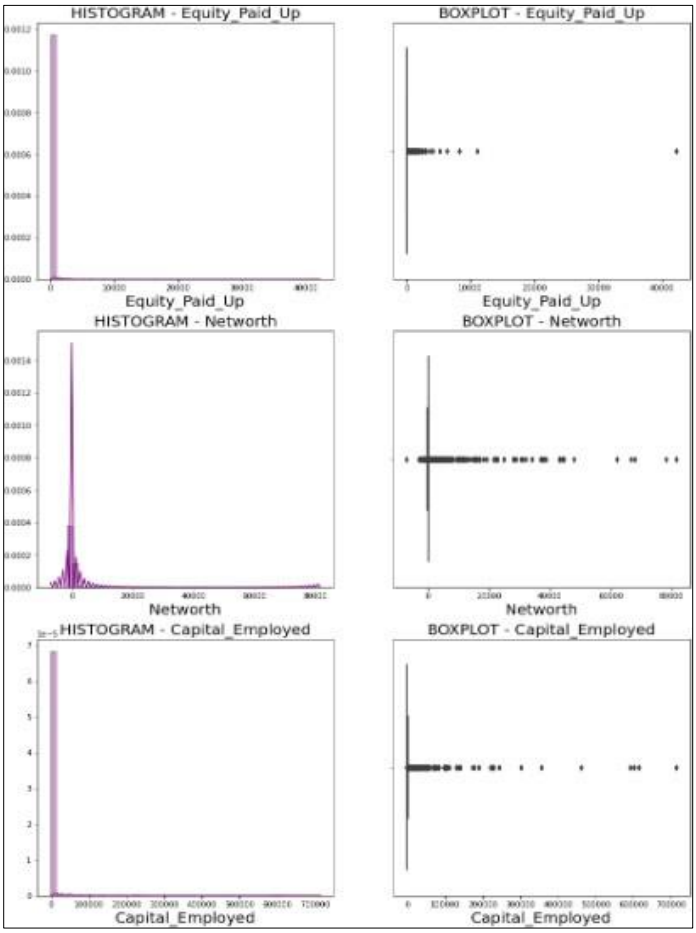
```
0    0.891801
1    0.108199
Name: Default, dtype: float64
```

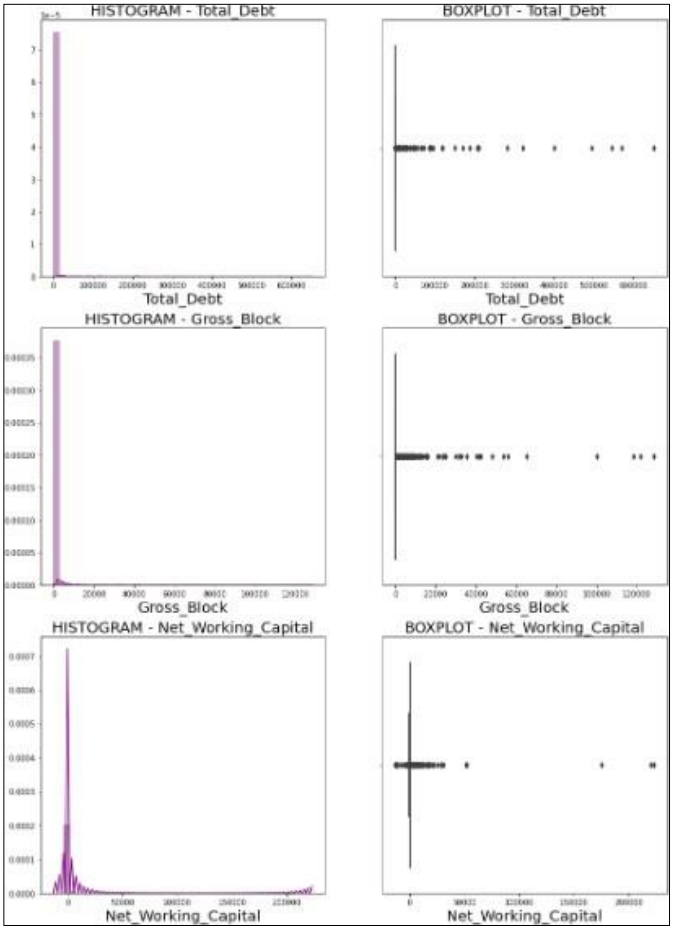
It is observed that almost 11% of the total entries in "Default" belong to category "1".

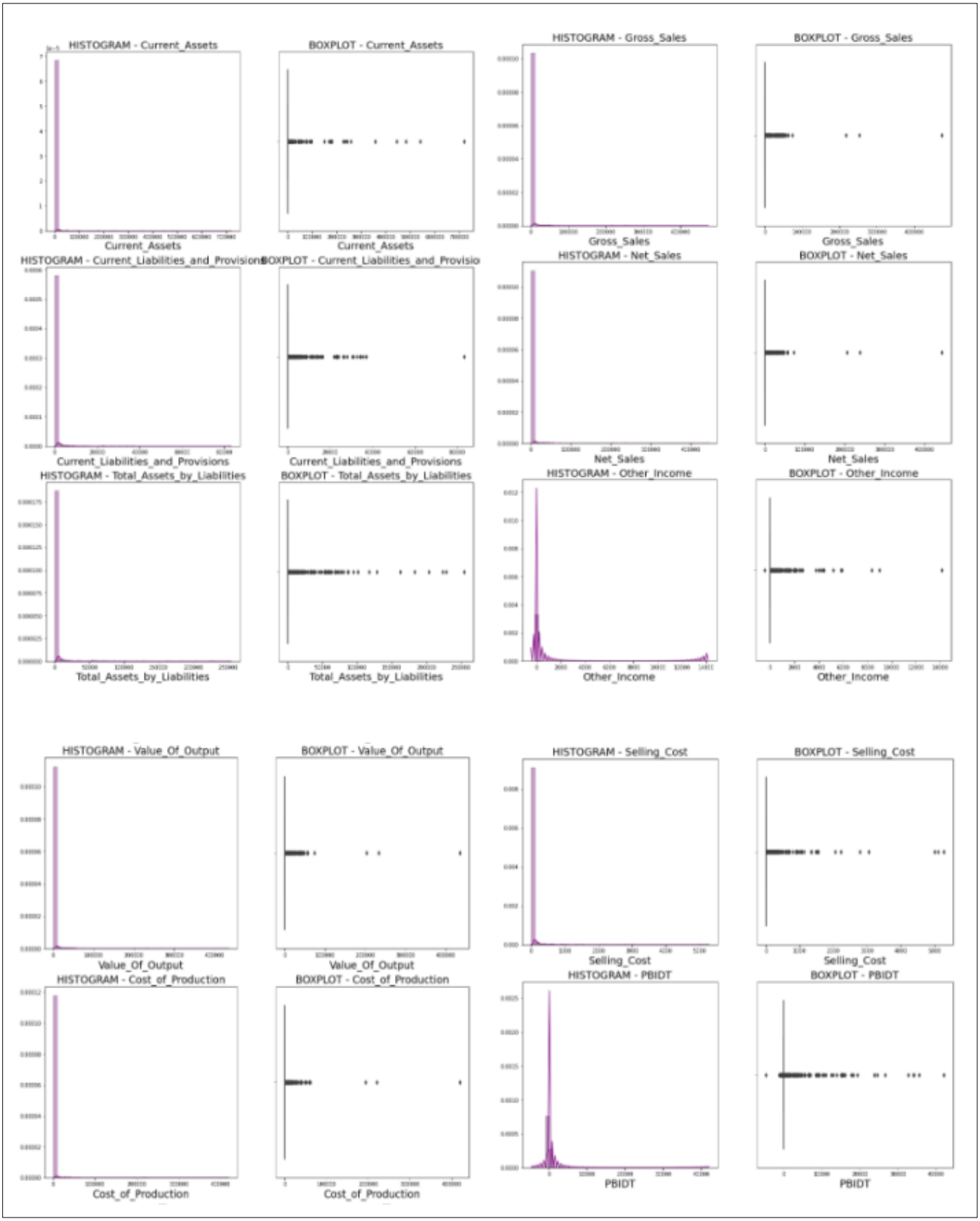
1.4 Univariate & Bivariate analysis with proper interpretation.

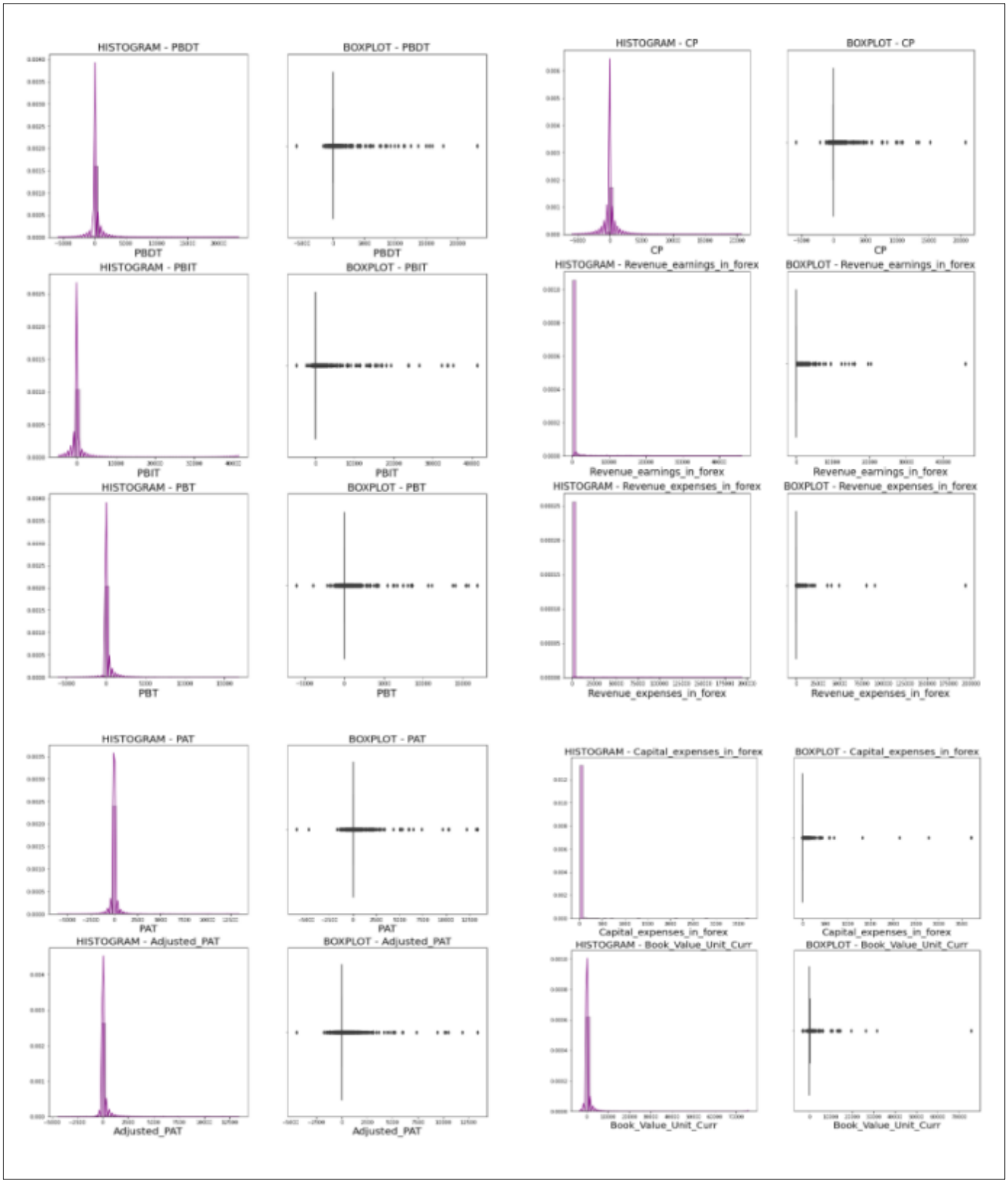
Univariate analysis:

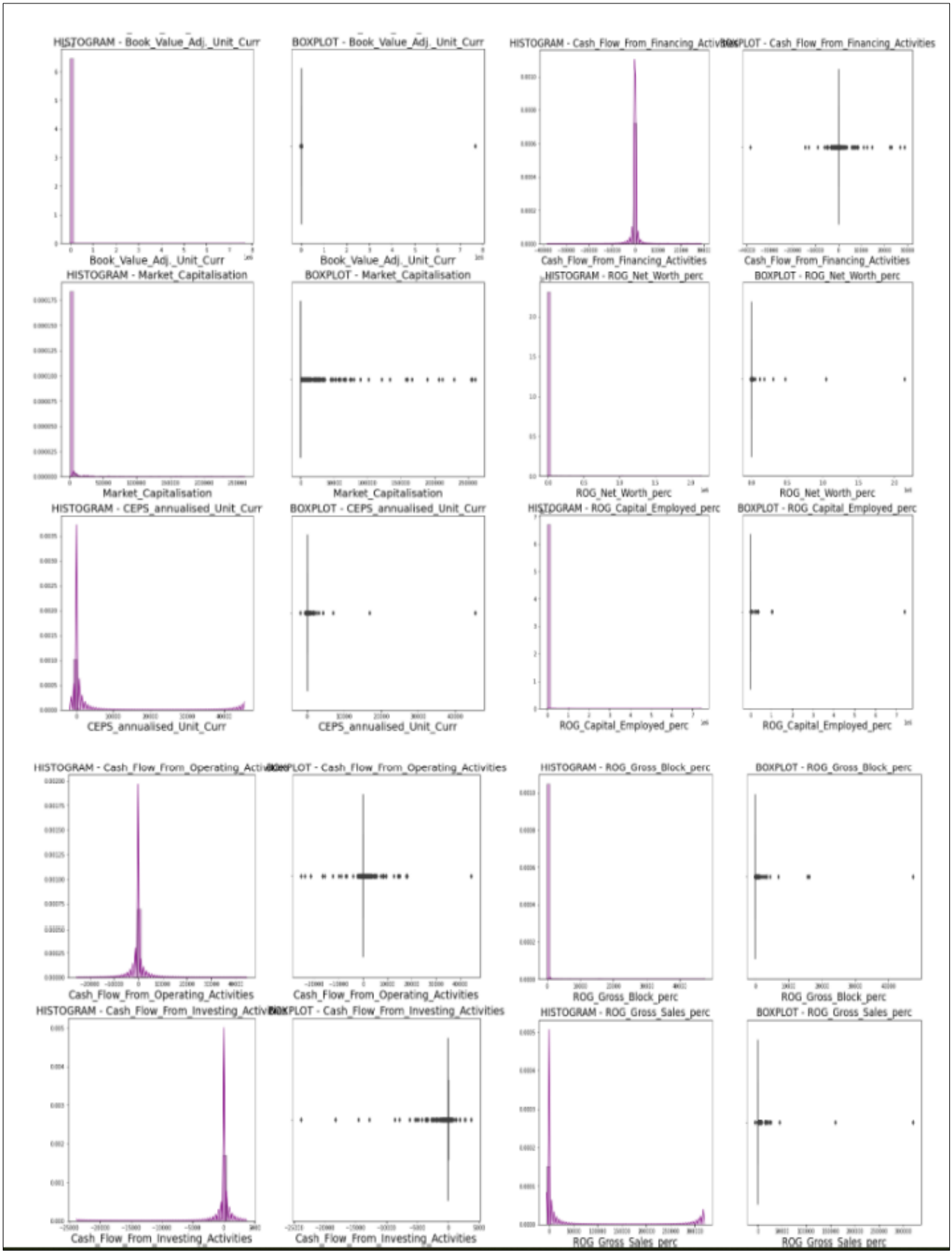
Univariate analysis involving data distribution along with outlier detection (Boxplot) plots have been shown below. As the number of variables are high, Univariate analysis involves high number of plots.

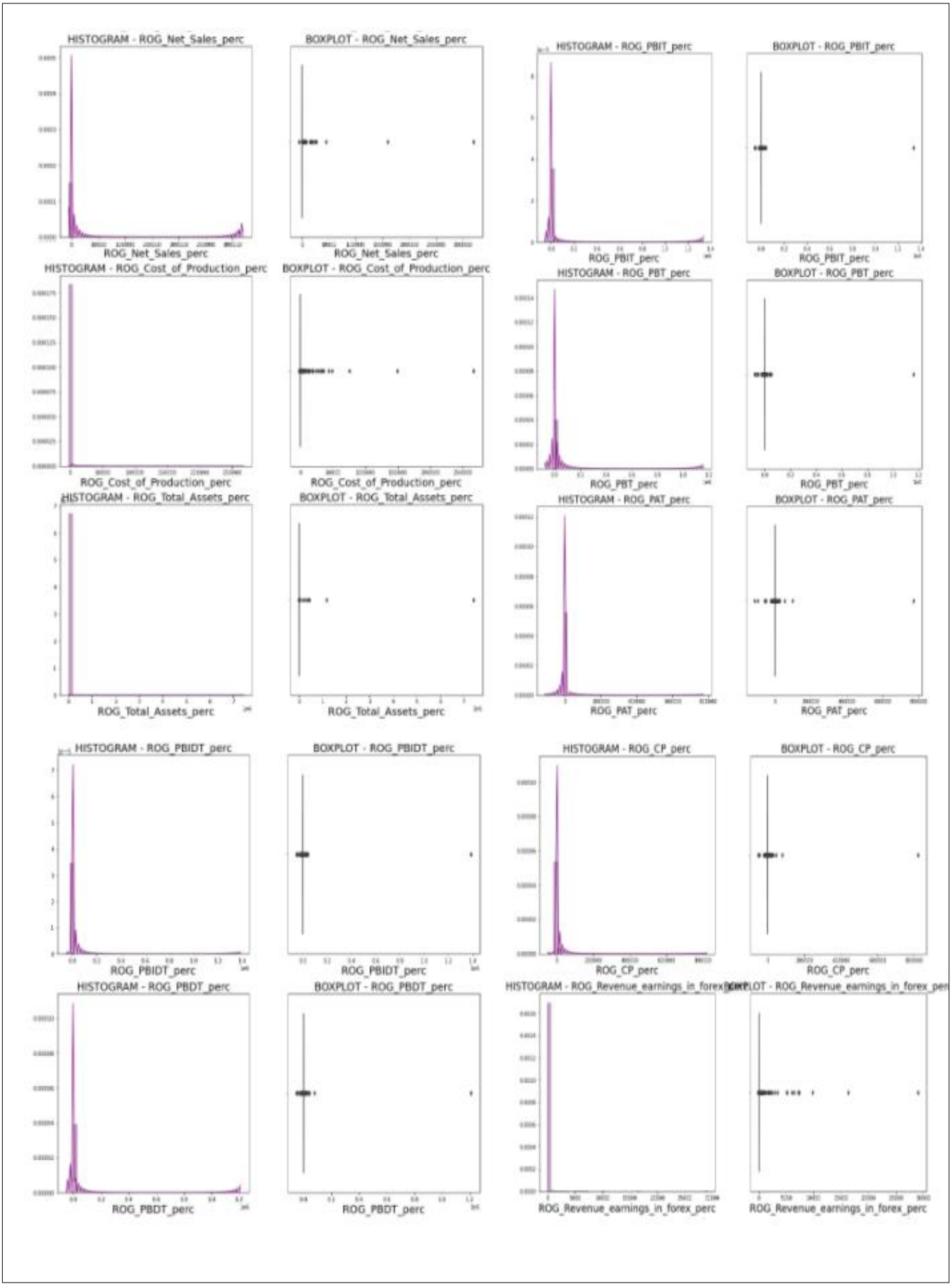


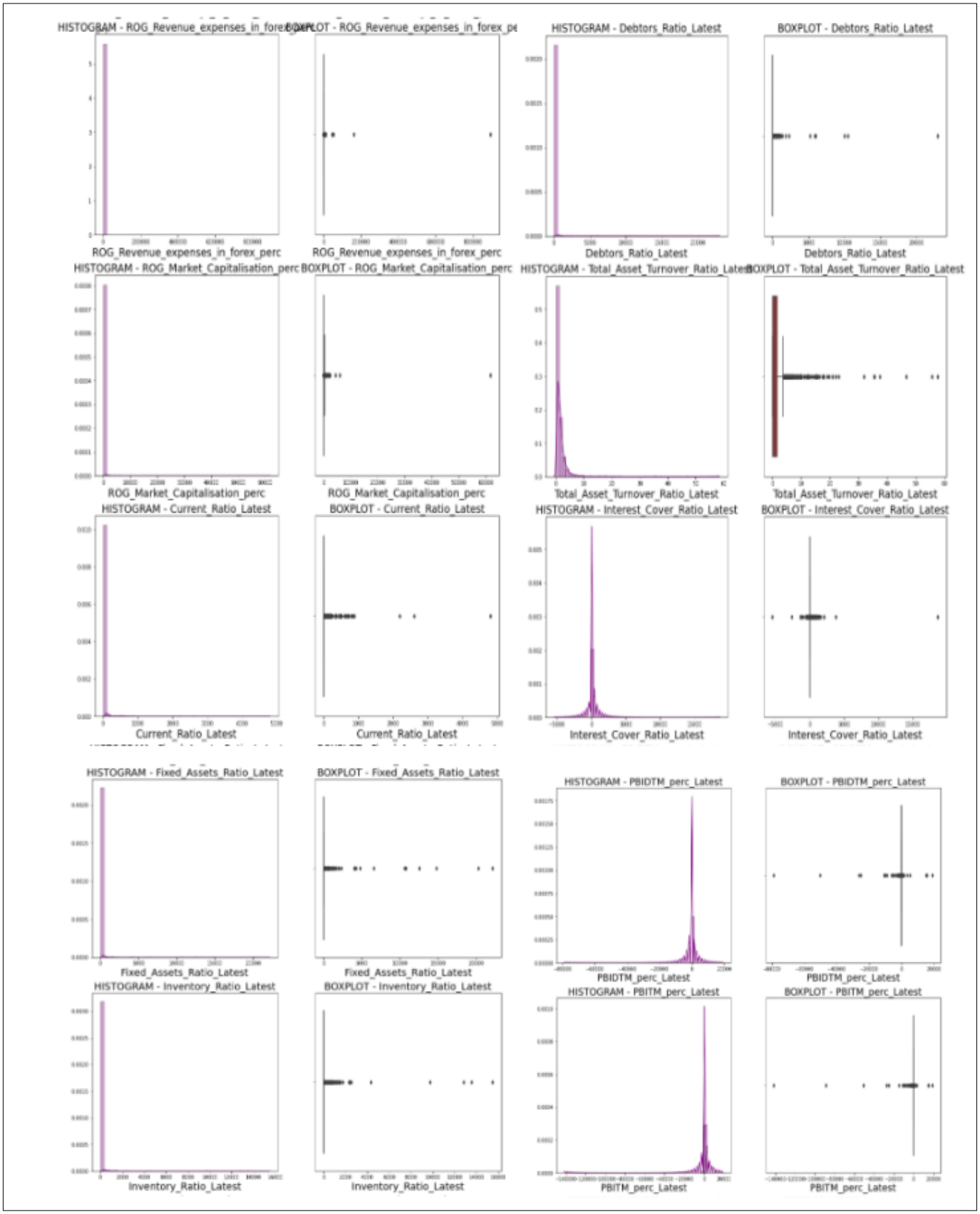



HISTOGRAM - Current_Assets
Current_Assets
BOXPLOT - Current_Assets
Current_Assets
HISTOGRAM - Gross_Sales
Gross_Sales
BOXPLOT - Gross_Sales
Gross_Sales
HISTOGRAM - Current_Liabilities_and_Provisions
Current_Liabilities_and_Provisions
BOXPLOT - Current_Liabilities_and_Provisio
Current_Liabilities_and_Provisions
HISTOGRAM - Net_Sales
Net_Sales
BOXPLOT - Net_Sales
Net_Sales
HISTOGRAM - Total_Assets_by_Liabilities
Total_Assets_by_Liabilities
BOXPLOT - Total_Assets_by_Liabilities
Total_Assets_by_Liabilities
HISTOGRAM - Other_Income
Other_Income
BOXPLOT - Other_Income
Other_Income
HISTOGRAM - Value_Of_Output
Value_Of_Output
BOXPLOT - Value_Of_Output
Value_Of_Output
HISTOGRAM - Selling_Cost
Selling_Cost
BOXPLOT - Selling_Cost
Selling_Cost
HISTOGRAM - Cost_of_Production
Cost_of_Production
BOXPLOT - Cost_of_Production
Cost_of_Production
HISTOGRAM - PBIDT
PBIDT
BOXPLOT - PBIDT
PBIDT


HISTOGRAM - PBDT
PBDT
BOXPLOT - PBDT
PBDT
HISTOGRAM - CP
CP
BOXPLOT - CP
CP
HISTOGRAM - PBIT
PBIT
BOXPLOT - PBIT
PBIT
HISTOGRAM - Revenue_earnings_in_forex
Revenue_earnings_in_forex
BOXPLOT - Revenue_earnings_in_forex
Revenue_earnings_in_forex
HISTOGRAM - PBT
PBT
BOXPLOT - PBT
PBT
HISTOGRAM - Revenue_expenses_in_forex
Revenue_expenses_in_forex
BOXPLOT - Revenue_expenses_in_forex
Revenue_expenses_in_forex
HISTOGRAM - PAT
PAT
BOXPLOT - PAT
PAT
HISTOGRAM - Capital_expenses_in_forex
Capital_expenses_in_forex
BOXPLOT - Capital_expenses_in_forex
Capital_expenses_in_forex
HISTOGRAM - Adjusted_PAT
Adjusted_PAT
BOXPLOT - Adjusted_PAT
Adjusted_PAT
HISTOGRAM - Book_Value_Unit_Curr
Book_Value_Unit_Curr
BOXPLOT - Book_Value_Unit_Curr
Book_Value_Unit_Curr


HISTOGRAM - Book_Value_Adj._Unit_Curr
Book_Value_Adj._Unit_Curr
BOXPLOT - Book_Value_Adj._Unit_Curr
Book_Value_Adj._Unit_Curr
HISTOGRAM - Cash_Flow_From_Financing_Activities
Cash_Flow_From_Financing_Activities
BOXPLOT - Cash_Flow_From_Financing_Activities
Cash_Flow_From_Financing_Activities
HISTOGRAM - Market_Capitalisation
Market_Capitalisation
BOXPLOT - Market_Capitalisation
Market_Capitalisation
HISTOGRAM - ROG_Net_Worth_perc
ROG_Net_Worth_perc
BOXPLOT - ROG_Net_Worth_perc
ROG_Net_Worth_perc
HISTOGRAM - CEPS_annualised_Unit_Curr
CEPS_annualised_Unit_Curr
BOXPLOT - CEPS_annualised_Unit_Curr
CEPS_annualised_Unit_Curr
HISTOGRAM - ROG_Capital_Employed_perc
ROG_Capital_Employed_perc
BOXPLOT - ROG_Capital_Employed_perc
ROG_Capital_Employed_perc
HISTOGRAM - Cash_Flow_From_Operating_Activities
Cash_Flow_From_Operating_Activities
BOXPLOT - Cash_Flow_From_Operating_Activities
Cash_Flow_From_Operating_Activities
HISTOGRAM - ROG_Gross_Block_perc
ROG_Gross_Block_perc
BOXPLOT - ROG_Gross_Block_perc
ROG_Gross_Block_perc
HISTOGRAM - Cash_Flow_From_Investing_Activities
Cash_Flow_From_Investing_Activities
BOXPLOT - Cash_Flow_From_Investing_Activities
Cash_Flow_From_Investing_Activities
HISTOGRAM - ROG_Gross_Sales_perc
ROG_Gross_Sales_perc
BOXPLOT - ROG_Gross_Sales_perc
ROG_Gross_Sales_perc


HISTOGRAM - ROG_Net_Sales_perc
ROG_Net_Sales_perc
BOXPLOT - ROG_Net_Sales_perc
ROG_Net_Sales_perc
HISTOGRAM - ROG_PBIT_perc
ROG_PBIT_perc
BOXPLOT - ROG_PBIT_perc
ROG_PBIT_perc
HISTOGRAM - ROG_Cost_of_Production_perc
ROG_Cost_of_Production_perc
BOXPLOT - ROG_Cost_of_Production_perc
ROG_Cost_of_Production_perc
HISTOGRAM - ROG_PBT_perc
ROG_PBT_perc
BOXPLOT - ROG_PBT_perc
ROG_PBT_perc
HISTOGRAM - ROG_Total_Assets_perc
ROG_Total_Assets_perc
BOXPLOT - ROG_Total_Assets_perc
ROG_Total_Assets_perc
HISTOGRAM - ROG_PAT_perc
ROG_PAT_perc
BOXPLOT - ROG_PAT_perc
ROG_PAT_perc
HISTOGRAM - ROG_PBIDT_perc
ROG_PBIDT_perc
BOXPLOT - ROG_PBIDT_perc
ROG_PBIDT_perc
HISTOGRAM - ROG_CP_perc
ROG_CP_perc
BOXPLOT - ROG_CP_perc
ROG_CP_perc
HISTOGRAM - ROG_PBDT_perc
ROG_PBDT_perc
BOXPLOT - ROG_PBDT_perc
ROG_PBDT_perc
HISTOGRAM - ROG_Revenue_earnings_in_forex_perc
ROG_Revenue_earnings_in_forex_perc
BOXPLOT - ROG_Revenue_earnings_in_forex_perc
ROG_Revenue_earnings_in_forex_perc


HISTOGRAM - ROG_Revenue_expenses_in_forex_perc
BOXPLOT - ROG_Revenue_expenses_in_forex_perc
ROG_Revenue_expenses_in_forex_perc
ROG_Revenue_expenses_in_forex_perc
HISTOGRAM - Debtors_Ratio_Latest
BOXPLOT - Debtors_Ratio_Latest
Debtors_Ratio_Latest
Debtors_Ratio_Latest
HISTOGRAM - ROG_Market_Capitalisation_perc
BOXPLOT - ROG_Market_Capitalisation_perc
ROG_Market_Capitalisation_perc
ROG_Market_Capitalisation_perc
HISTOGRAM - Total_Asset_Turnover_Ratio_Latest
BOXPLOT - Total_Asset_Turnover_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
HISTOGRAM - Current_Ratio_Latest
BOXPLOT - Current_Ratio_Latest
Current_Ratio_Latest
Current_Ratio_Latest
HISTOGRAM - Interest_Cover_Ratio_Latest
BOXPLOT - Interest_Cover_Ratio_Latest
Interest_Cover_Ratio_Latest
Interest_Cover_Ratio_Latest
HISTOGRAM - Fixed_Assets_Ratio_Latest
BOXPLOT - Fixed_Assets_Ratio_Latest
Fixed_Assets_Ratio_Latest
Fixed_Assets_Ratio_Latest
HISTOGRAM - PBIDTM_perc_Latest
BOXPLOT - PBIDTM_perc_Latest
PBIDTM_perc_Latest
PBIDTM_perc_Latest
HISTOGRAM - Inventory_Ratio_Latest
BOXPLOT - Inventory_Ratio_Latest
Inventory_Ratio_Latest
Inventory_Ratio_Latest
HISTOGRAM - PBITM_perc_Latest
BOXPLOT - PBITM_perc_Latest
PBITM_perc_Latest
PBITM_perc_Latest

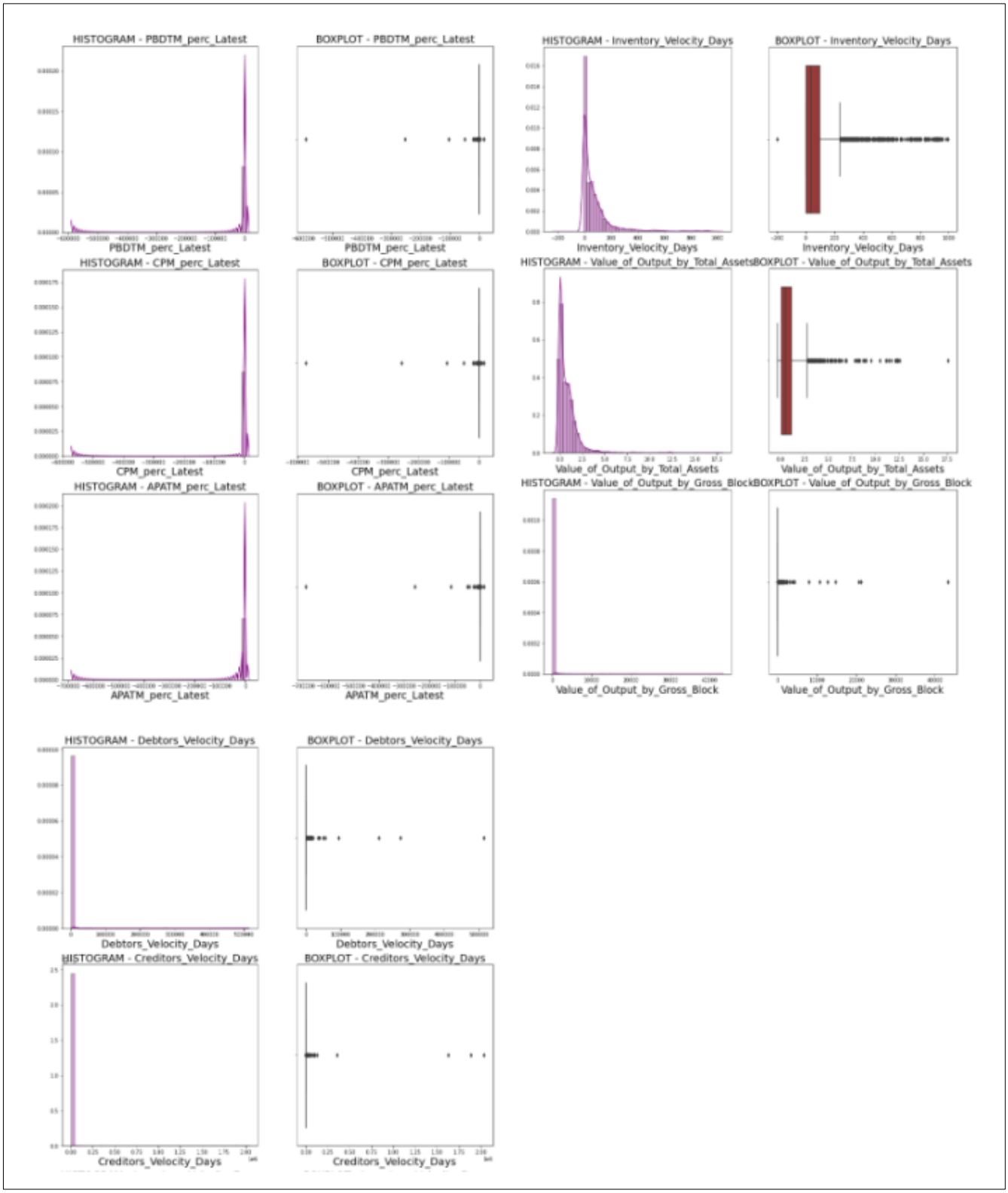


Figure 7: Univariate analysis

It is observed that all the variables have Outliers. Histogram analysis shows that all the variables are highly skewed, right or left. These outliers will be treated as we are going to apply Logistic regression to predict the outcome.

Skewness was observed in almost all the variables. Most of the variables were right skewed while a few were also found to be left skewed.

Co_Code	1.604115
Networth_Next_Year	13.041264
Equity_Paid_Up	45.928921
Networth	11.738799
Capital_Employed	18.073683
Total_Debt	19.417622
Gross_Block	18.528589
Net_Working_Capital	30.580553
Current_Assets	20.779473
Current_Liabilities_and_Provisions	15.291405
Total_Assets_by_Liabilities	13.367863
Gross_Sales	31.560200
Net_Sales	31.085039
Other_Income	18.805640
Value_Of_Output	30.812223
Cost_of_Production	34.588562
Selling_Cost	18.879055
PBIDT	13.179047
PBDT	13.555030
PBIT	14.009481
PBT	13.127556
PAT	13.067832
Adjusted_PAT	13.875662
CP	14.345253
Revenue_earnings_in_forex	24.176282
Revenue_expenses_in_forex	34.841543
Capital_expenses_in_forex	27.610257
Book_Value_Unit_Curr	32.984463
Book_Value_Adj._Unit_Curr	59.843813
Market_Capitalisation	14.391109
CEPS_annualised_Unit_Curr	48.533480
Cash_Flow_From_Operating_Activities	6.634856
Cash_Flow_From_Investing_Activities	-21.565103
Cash_Flow_From_Financing_Activities	1.703710
ROG_Net_Worth_perc	44.831986
ROG_Capital_Employed_perc	56.436450
ROG_Gross_Block_perc	44.870942
ROG_Gross_Sales_perc	45.404694
ROG_Net_Sales_perc	45.405621
ROG_Cost_of_Production_perc	37.269426

ROG_Total_Assets_perc	57.304521
ROG_PBIDT_perc	58.880737
ROG_PBDT_perc	58.407667
ROG_PBIT_perc	58.925536
ROG_PBT_perc	57.330567
ROG_PAT_perc	52.640553
ROG_CP_perc	56.788937
ROG_Revenue_earnings_in_forex_perc	31.052213
ROG_Revenue_expenses_in_forex_perc	56.807388
ROG_Market_Capitalisation_perc	57.329984
Current_Ratio_Latest	31.251406
Fixed_Assets_Ratio_Latest	24.123112
Inventory_Ratio_Latest	27.002844
Debtors_Ratio_Latest	35.256499
Total_Asset_Turnover_Ratio_Latest	10.358866
Interest_Cover_Ratio_Latest	40.823960
PBIDTM_perc_Latest	-30.931573
PBITM_perc_Latest	-35.997867
PBDTM_perc_Latest	-47.750324
CPM_perc_Latest	-47.011631
APATM_perc_Latest	-49.277483
Debtors_Velocity_Days	38.660834
Creditors_Velocity_Days	34.120441
Inventory_Velocity_Days	3.494365
Value_of_Output_by_Total_Assets	4.704950
Value_of_Output_by_Gross_Block	31.998522
Default	2.523672
dtype: float64	

Bivariate analysis:

Bi-variate analysis includes pair plot and heatmap of correlation matrix. As the number of variables are high, the pair plot would not be so legible. For that reason, the pair plots are displayed for variables which are significant (derived using VIF score) in model prediction, and which have significant correlations among each other.

Pair plot:

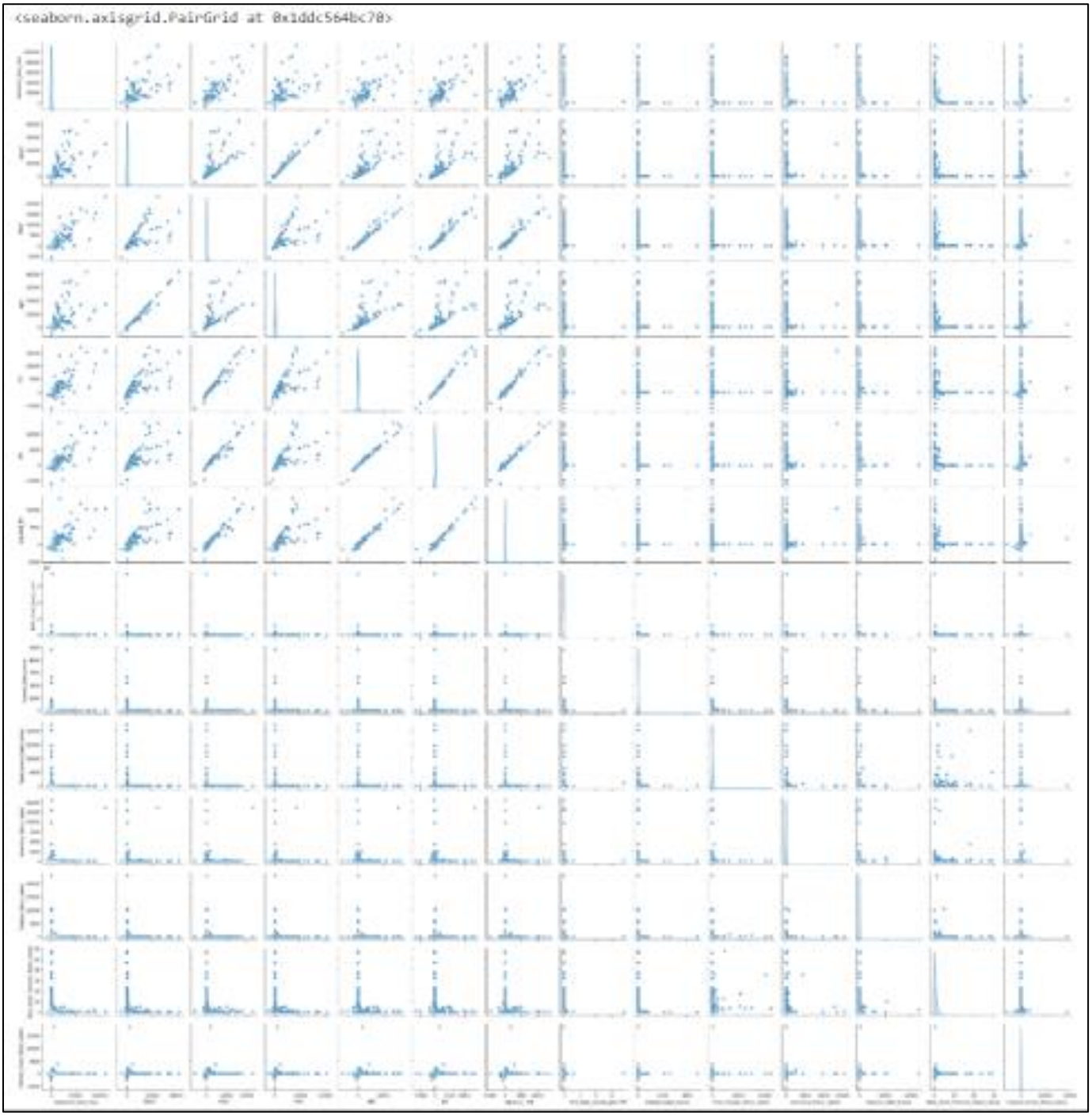
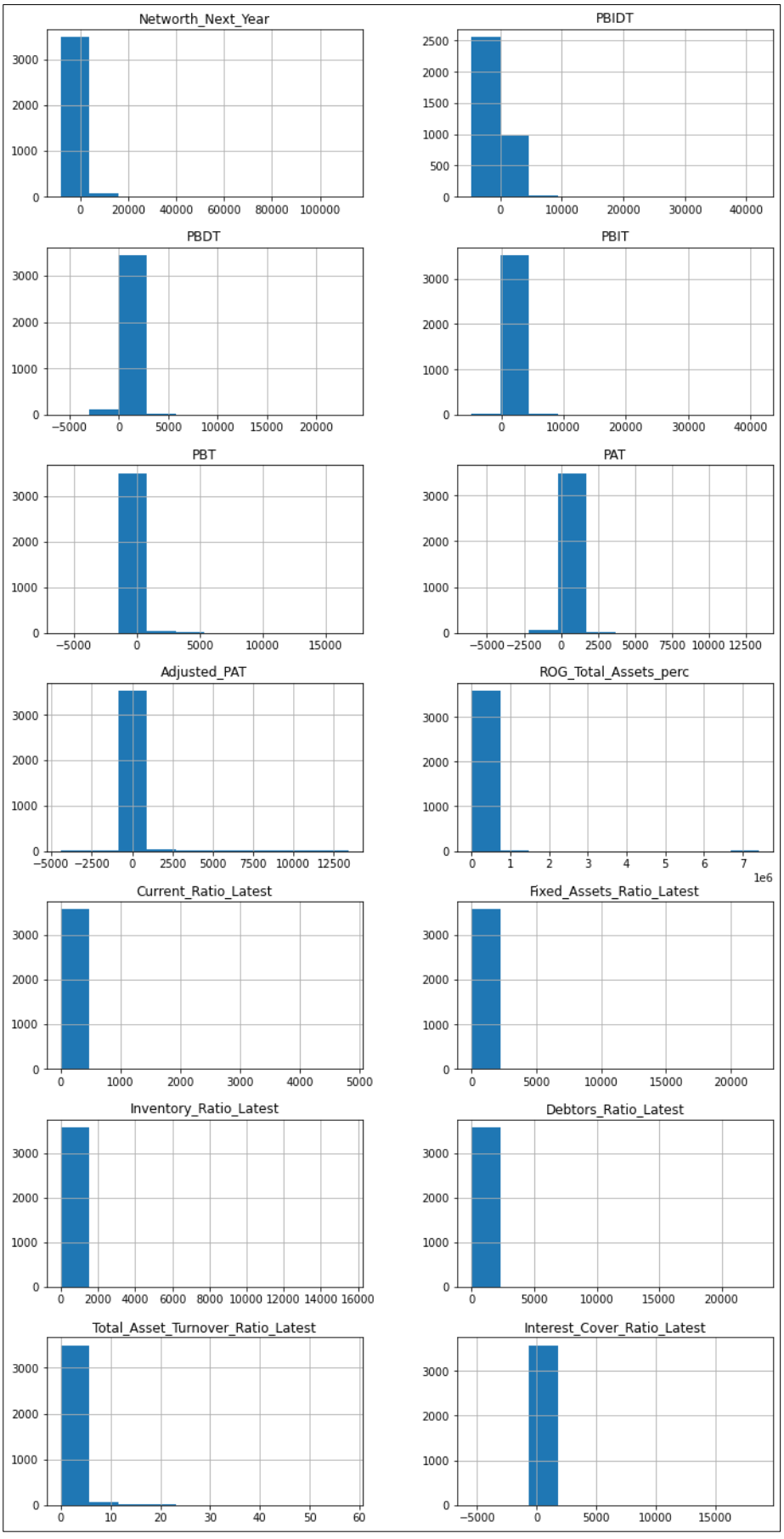


Figure 8: Bivariate analysis

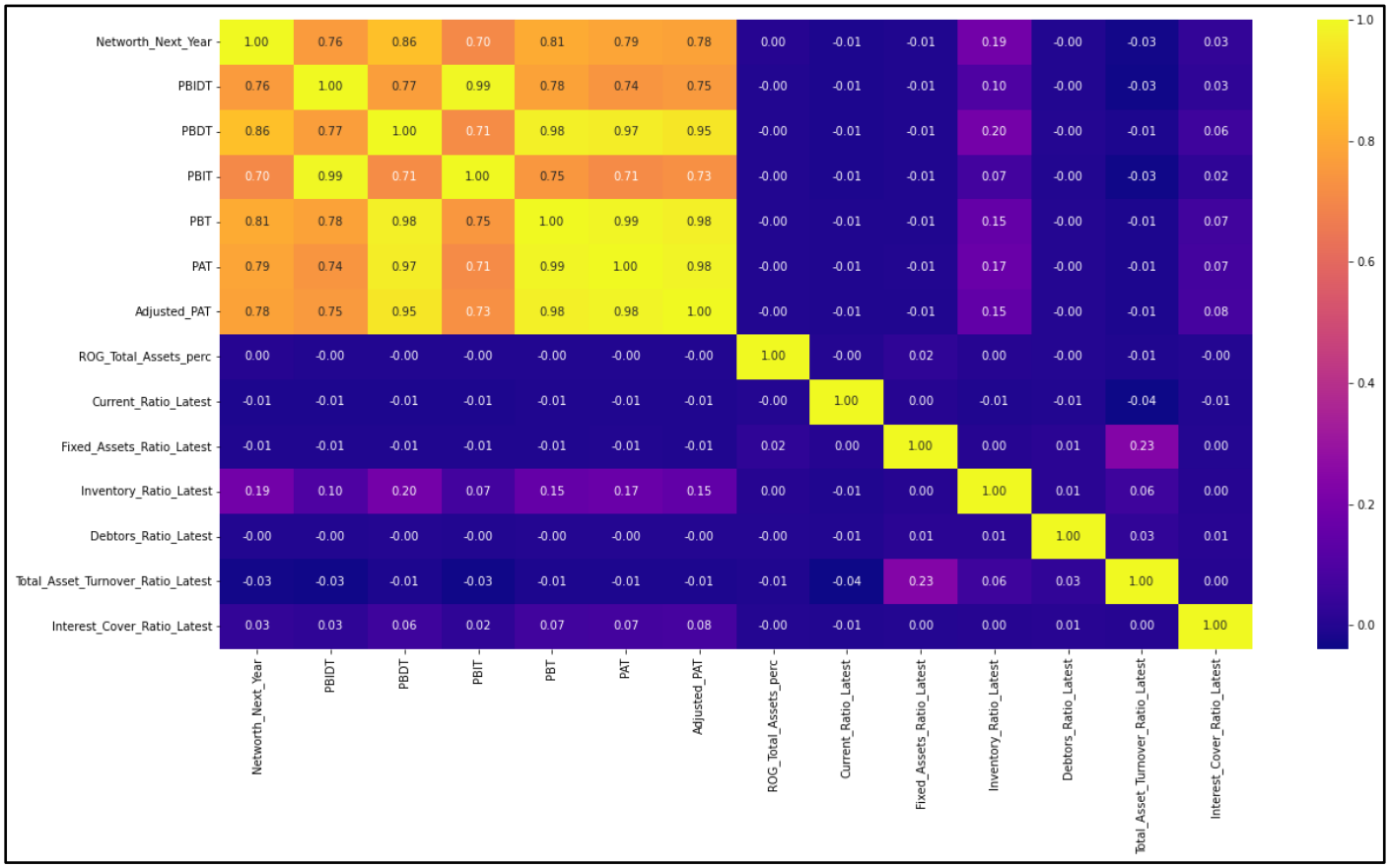
Histogram:


Networth_Next_Year
PBIDT
PBDT
PBIT
PBT
PAT
Adjusted_PAT
ROG_Total_Assets_perc
1e6
Current_Ratio_Latest
Fixed_Assets_Ratio_Latest
Inventory_Ratio_Latest
Debtors_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
Interest_Cover_Ratio_Latest

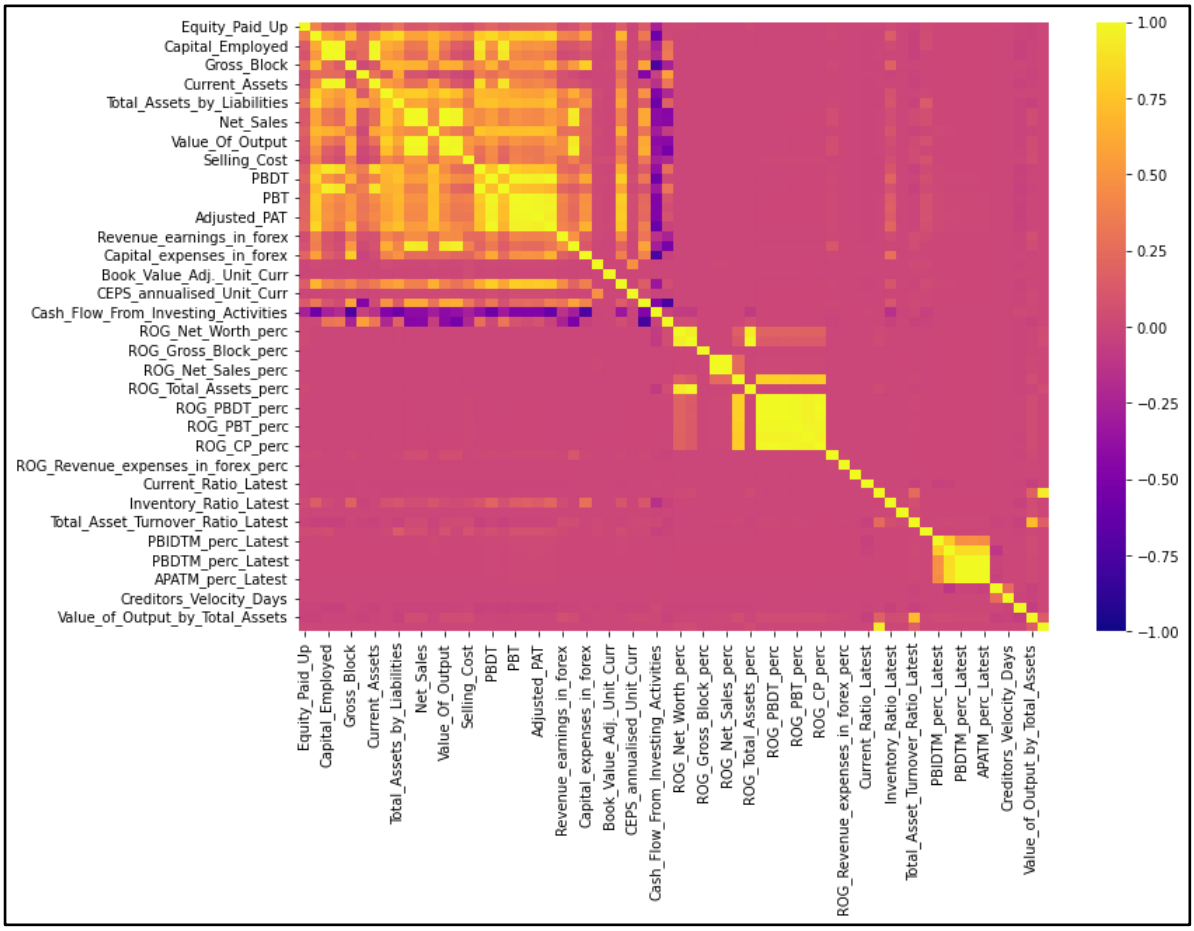
Observations:

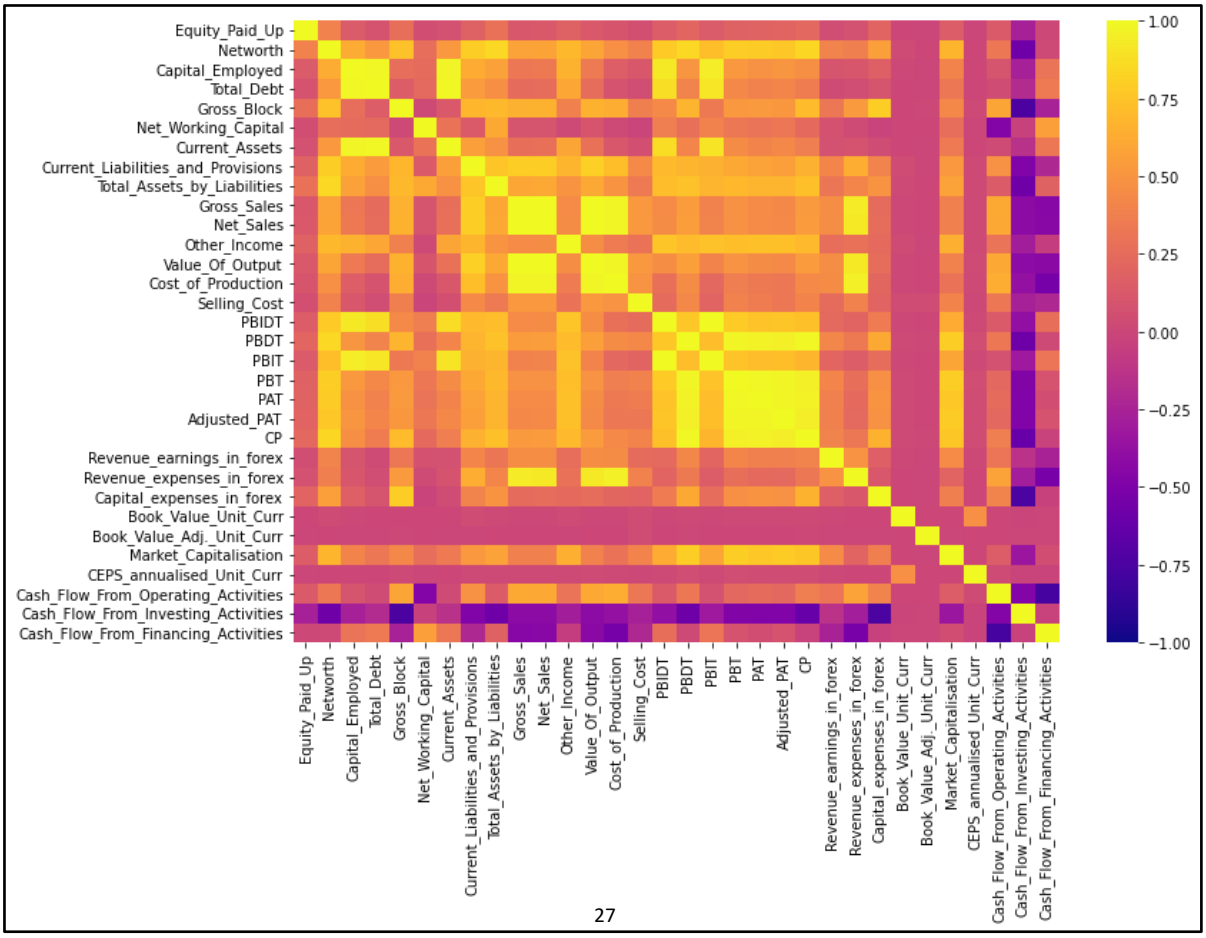
- It is observed that there is high positive correlation between variables PBDT, PBIDT, PBIT, PBT, PAT, Adjusted PAT, and CP.
- It is observed that there is negative correlation between variables Cash_Flow_From_Operating_Activities and Cash_Flow_From_Financing_Activities.
- It is observed that there is positive correlation between variables CPM_perc_Latest, APATM_perc_Latest, PBIDTM_perc_Latest, PBITM_perc_Latest, PBDTM_perc_Latest.
- It is observed that there is positive correlation between variables Capital_Employed and Current_Assets, Current_Assets and Total_debt, Total_debt and Capital_Employed.

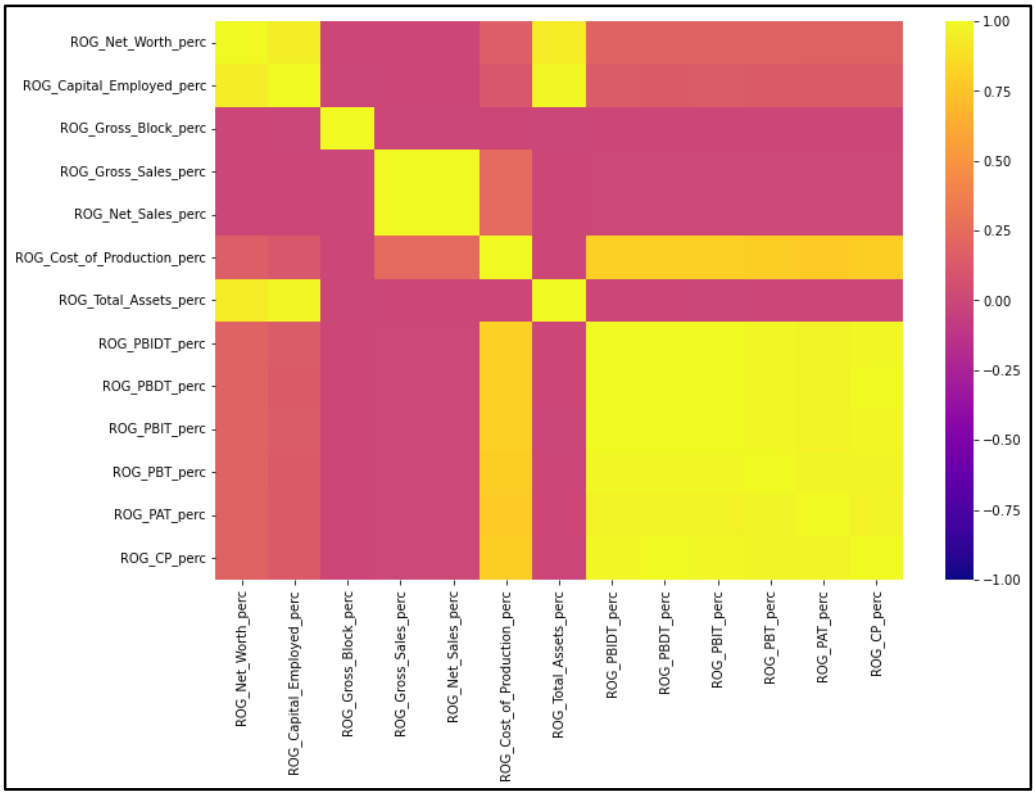
Correlation Heatmap:



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


Equity_Paid_Up
Capital_Employed
Gross_Block
Current_Assets
Total_Assets_by_Liabilities
Net_Sales
Value_Of_Output
Selling_Cost
PBDT
PBT
Adjusted_PAT
Revenue_earnings_in_forex
Capital_expenses_in_forex
Book_Value_Adj._Unit_Curr
CEPS_annualised_Unit_Curr
Cash_Flow_From_Investing_Activities
ROG_Net_Worth_perc
ROG_Gross_Block_perc
ROG_Net_Sales_perc
ROG_Total_Assets_perc
ROG_PBDT_perc
ROG_PBT_perc
ROG_CP_perc
ROG_Revenue_expenses_in_forex_perc
Current_Ratio_Latest
Inventory_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
PBIDTM_perc_Latest
PBDTM_perc_Latest
APATM_perc_Latest
Creditors_Velocity_Days
Value_of_Output_by_Total_Assets
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Equity_Paid_Up
Networth
Capital_Employed
Total_Debt
Gross_Block
Net_Working_Capital
Current_Assets
Current_Liabilities_and_Provisions
Total_Assets_by_Liabilities
Gross_Sales
Net_Sales
Other_Income
Value_Of_Output
Cost_of_Production
Selling_Cost
PBIDT
PBDT
PBIT
PBT
PAT
Adjusted_PAT
CP
Revenue_earnings_in_forex
Revenue_expenses_in_forex
Capital_expenses_in_forex
Book_Value_Unit_Curr
Book_Value_Adj._Unit_Curr
Market_Capitalisation
CEPS_annualised_Unit_Curr
Cash_Flow_From_Operating_Activities
Cash_Flow_From_Investing_Activities
Cash_Flow_From_Financing_Activities
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



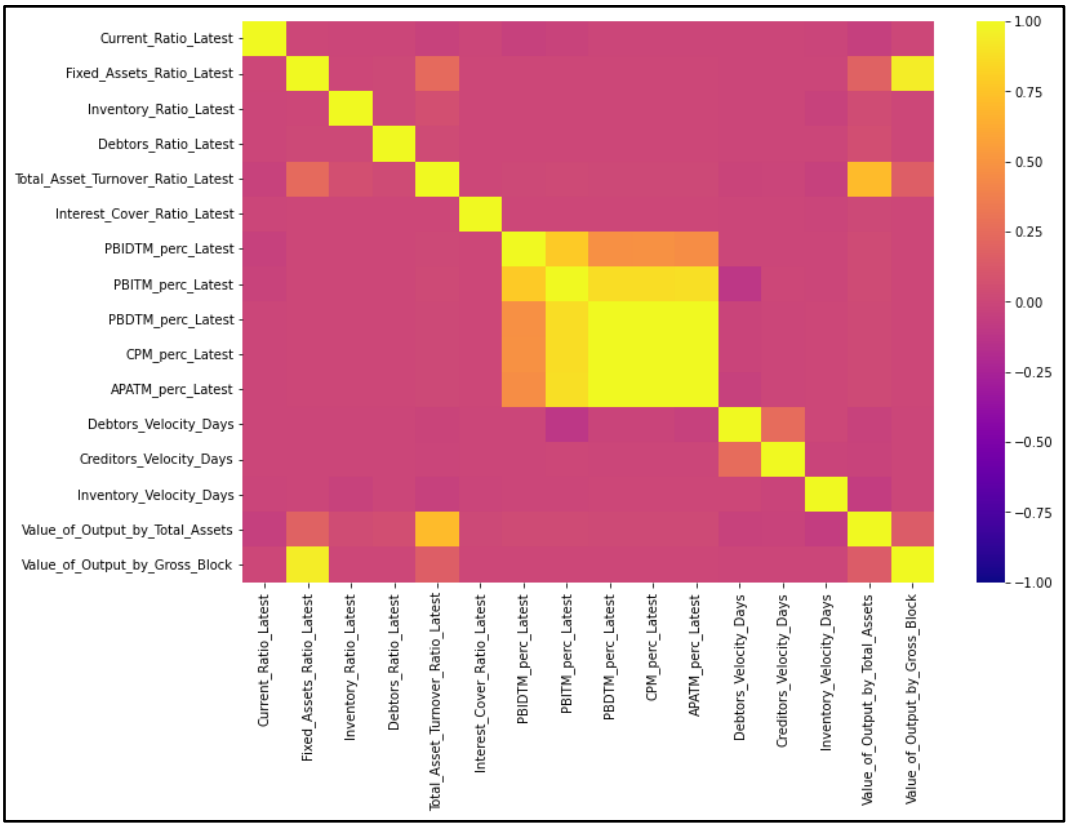


Figure 9: Correlation heatmaps

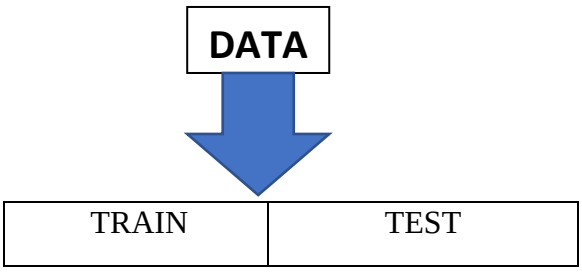
From the above heat maps, it can be observed that many variables have correlation values close to 1. This denotes high collinearity among these variables.

Inferences from Univariate and Bi-variate analysis:

- Most of the variables have skewed distribution. But we will not treat those distribution by any kind of transformation or new features.
- All the variables have outliers. These outliers will be treated as we are going to apply Logistic regression to predict the outcome.
- Bi-variate analysis is performed on some of the important variables selected through VIF (discussed later in this report).
- From pair plots, it is observed that there is high positive correlation between variables PBDT, PBIDT, PBIT, PBT, PAT, Adjusted PAT, and CP, which is obvious as these are the parameters to evaluate any corporates' performance.
- It is observed that there is negative correlation between variables Cash_Flow_From_Operating_Activities and Cash_Flow_From_Financing_Activities.
- It is observed that there is positive correlation between variables CPM_perc_Latest, APATM_perc_Latest, PBIDTM_perc_Latest, PBITM_perc_Latest, PBDTM_perc_Latest.
- It is observed that there is positive correlation between variables Capital_Employed and Current_Assets, Current_Assets and Total_debt, Total_debt and Capital_Employed. 8. Overall, high positive and negative correlation between variables can be seen above. Majority of the variables are not correlated.

1.5 Train Test Split

The original data frame except the variables Co_Code, Co_Name, Networth_Next_Year is divided into dependent and independent variable type data frame. Then both independent and dependent variable data frame is split into 67:33 (train: test) ratio. One requirement for Stats model is that dependent and independent variables should be contained in same data frame. So, concatenation was performed to combine dependent and independent variables arrays.




```
The number of rows (observations) in TRAIN set is 2402
The number of columns (variables) in TRAIN set is 65

The number of rows (observations) in TEST set is 1184
The number of columns (variables) in TEST set is 65
```

1.6 Build Logistic Regression Model (using statsmodels library) on most important variables on Train Dataset and choose the optimum cut-off. Also showcase your model building approach

Logistic Regression Model:

The equation of the Logistic Regression by which we predict the corresponding probabilities and then go on predict a discrete target variable is

$$y = \frac{1}{1+e^{-z}}$$

Note: $z = \beta_0 + \sum_{i=1}^{n}(\beta_i X_i)$

Some of the libraries we will be using are as follows:

- From sklearn. model_selection train_test_split - for splitting the train and test set.
- Variance_inflation_factor module from statsmodels. stats. outliers_influence metrics - from sklearn
- roc_auc_score, roc_curve - from sklearn.metrics
- classification_report, confusion_matrix, plot_confusion_matrix - from sklearn.metrics

Model Building for FRA dataset:

Feature Engineering approach:

The optimal machine learning problem approach is to perform extensive EDA on dataset and understand properties of the predictors before even getting into training models on these variables. However, this is not always possible. Sometimes the dataset has lot many variables; sometimes even hundreds or even thousands of variables, which can quickly outrun human comprehension. Feature selection is the process of tuning down the number of predictor variables used by the models you build.

Model building approach:

Before starting model building, let's look at the problem of multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. First, variance inflation factor (VIF) is used as criteria to eliminate some of the variables.

	variables	VIF
34	ROG_Gross_Block_perc	1.524386
47	ROG_Market_Capitalisation_perc	1.676561
61	Inventory_Velocity_Days	1.924207
37	ROG_Cost_of_Production_perc	2.029579
60	Creditors_Velocity_Days	2.305668
50	Inventory_Ratio_Latest	2.343545
48	Current_Ratio_Latest	2.397185
59	Debtors_Velocity_Days	2.474369
51	Debtors_Ratio_Latest	2.531709
53	Interest_Cover_Ratio_Latest	2.541902
30	Cash_Flow_From_Investing_Activities	2.606063
32	ROG_Net_Worth_perc	2.965792
31	Cash_Flow_From_Financing_Activities	2.991789
22	Revenue_earnings_in_forex	3.092275
38	ROG_Total_Assets_perc	3.259153
23	Revenue_expenses_in_forex	3.719369
33	ROG_Capital_Employed_perc	3.842840
29	Cash_Flow_From_Operating_Activities	4.047336
0	Equity_Paid_Up	4.568407
27	Market_Capitalisation	4.586333
11	Other_Income	4.591128
14	Selling_Cost	4.907195
5	Net_Working_Capital	5.847854
28	CEPS_annualised_Unit_Curr	6.663665
3	Total_Debt	7.596788
49	Fixed_Assets_Ratio_Latest	9.239632
63	Value_of_Output_by_Gross_Block	9.518160
52	Total_Asset_Turnover_Ratio_Latest	10.466825
62	Value_of_Output_by_Total_Assets	11.837468
4	Gross_Block	12.498251
26	Book_Value_Adj._Unit_Curr	12.588579
1	Networth	12.898845
43	ROG_PAT_perc	13.630985
41	ROG_PBIT_perc	14.097670
25	Book_Value_Unit_Curr	16.348994
44	ROG_CP_perc	16.866172
39	ROG_PBIDT_perc	17.076875
20	Adjusted_PAT	17.368941
42	ROG_PBT_perc	17.986538
58	APATM_perc_Latest	19.795721
40	ROG_PBDT_perc	23.462276
7	Current_Liabilities_and_Provisions	25.456194
56	PBDTM_perc_Latest	29.054765
55	PBITM_perc_Latest	30.919216
57	CPM_perc_Latest	32.685661
17	PBIT	33.326853
54	PBIDTM_perc_Latest	33.908813
6	Current_Assets	34.804508
15	PBIDT	40.250350
13	Cost_of_Production	49.236214
2	Capital_Employed	70.510704
19	PAT	79.517043
18	PBT	83.010367
8	Total_Assets_by_Liabilities	102.417724
21	CP	129.681878
16	PBDT	136.859867
36	ROG_Net_Sales_perc	528.691292
35	ROG_Gross_Sales_perc	529.543274
12	Value_Of_Output	652.419803
9	Gross_Sales	763.156319
10	Net_Sales	1419.303792
24	Capital_expenses_in_forex	NaN
45	ROG_Revenue_earnings_in_forex_perc	NaN
46	ROG_Revenue_expenses_in_forex_perc	NaN

Table 3: VIF analysis

Based on the above analysis, it can be observed that the value of VIF is high for many variables. Hence, we may drop variables with VIF more than 5 (very high correlation) & build the model.

Considering only the variables with VIF less than equal to 5.

Fitting the logistic regression model

Model 1:

Below are the results of the logistic regression:

Dep. Variable:	Default	No. Observations:	2402
Model:	Logit	Df Residuals:	2379
Method:	MLE	Df Model:	22
Date:	Sat, 08 Oct 2022	Pseudo R-squ.:	0.3988
Time:	20:26:26	Log-Likelihood:	-495.05
converged:	True	LL-Null:	-823.47
Covariance Type:	nonrobust	LLR p-value:	9.720e-125

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6458	0.184	-3.518	0.000	-1.005	-0.286
ROG_Gross_Block_perc	-0.0324	0.014	-2.264	0.024	-0.060	-0.004
ROG_Market_Capitalisation_perc	-0.0004	0.002	-0.178	0.859	-0.004	0.004
Inventory_Velocity_Days	-0.0013	0.001	-1.087	0.277	-0.004	0.001
ROG_Cost_of_Production_perc	-0.0093	0.003	-3.561	0.000	-0.014	-0.004
Creditors_Velocity_Days	0.0040	0.001	3.495	0.000	0.002	0.006
Inventory_Ratio_Latest	-0.0119	0.014	-0.869	0.385	-0.039	0.015
Current_Ratio_Latest	-0.7525	0.086	-8.871	0.000	-0.919	-0.586
Debtors_Velocity_Days	-0.0040	0.001	-3.796	0.000	-0.006	-0.002
Debtors_Ratio_Latest	-0.0200	0.015	-1.343	0.179	-0.049	0.009
Interest_Cover_Ratio_Latest	-0.1628	0.031	-5.192	0.000	-0.224	-0.101
Cash_Flow_From_Investing_Activities	0.0144	0.025	0.573	0.567	-0.035	0.064
ROG_Net_Worth_perc	-0.0502	0.008	-5.911	0.000	-0.067	-0.034
Cash_Flow_From_Financing_Activities	0.0116	0.023	0.503	0.615	-0.034	0.057
Revenue_earnings_in_forex	-0.0266	0.019	-1.369	0.171	-0.065	0.011
ROG_Total_Assets_perc	-0.0117	0.008	-1.489	0.136	-0.027	0.004
Revenue_expenses_in_forex	0.0365	0.020	1.804	0.071	-0.003	0.076
ROG_Capital_Employed_perc	-0.0013	0.008	-0.157	0.876	-0.017	0.014
Cash_Flow_From_Operating_Activities	-0.0059	0.013	-0.457	0.648	-0.031	0.019
Equity_Paid_Up	0.0176	0.007	2.354	0.019	0.003	0.032
Market_Capitalisation	-0.0091	0.002	-4.936	0.000	-0.013	-0.005
Other_Income	0.0198	0.037	0.534	0.593	-0.053	0.093
Selling_Cost	-0.0279	0.044	-0.634	0.526	-0.114	0.058

We can see that few variables are insignificant & may not be useful to discriminate cases of default. We will try & remove variables whose p value is greater than 0.05 & rebuild our model

Table 4: Model 1 results

Model 2:

Dep. Variable:	**Default**	**No. Observations:**	**2402**
Model:	Logit	**Df Residuals:**	2391
Method:	MLE	**Df Model:**	10
Date:	Fri, 07 Oct 2022	**Pseudo R-squ.:**	0.3868
Time:	19:43:36	**Log-Likelihood:**	-504.98
converged:	True	**LL-Null:**	-823.47
Covariance Type:	no robust	**LLR p-value:**	2.082e-130

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.8746	0.167	-5.236	0.000	-1.202	-0.547
ROG_Gross_Block_perc	-0.0442	0.013	-3.282	0.001	-0.071	-0.018
ROG_Cost_of_Production_perc	-0.0097	0.003	-3.796	0.000	-0.015	-0.005
Creditors_Velocity_Days	0.0043	0.001	3.751	0.000	0.002	0.006
Current_Ratio_Latest	-0.7667	0.088	-8.692	0.000	-0.940	-0.594
Debtors_Velocity_Days	-0.0040	0.001	-3.959	0.000	-0.006	-0.002
Interest_Cover_Ratio_Latest	-0.1707	0.031	-5.546	0.000	-0.231	-0.110
ROG_Net_Worth_perc	-0.0559	0.007	-7.982	0.000	-0.070	-0.042
Revenue_expenses_in_forex	0.0014	0.016	0.088	0.930	-0.030	0.032
Equity_Paid_Up	0.0171	0.007	2.428	0.015	0.003	0.031
Market_Capitalization	-0.0101	0.002	-6.004	0.000	-0.013	-0.007

Table 5: Model 2 results

P-value of Revenue_expenses_in_forex is the highest and is insignificant. Hence the variable can be dropped

Model 3:

Dep. Variable:	**Default**	No. Observations:	**2402**
Model:	Logit	**Df Residuals:**	2393
Method:	MLE	**Df Model:**	8
Date:	Sun, 09 Oct 2022	**Pseudo R-squ.:**	0.3830
Time:	13:46:47	**Log-Likelihood:**	-508.05
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	5.464e-13

	coef	**std err**	**z**	**P>|z|**	**[0.025]**	**[0.975]**
Intercept	-0.7265	0.155	-4.690	0.000	-1.030	-0.423
ROG_Gross_Block_perc	-0.0445	0.013	-3.322	0.001	-0.071	-0.018
ROG_Cost_of_Production_perc	-0.0101	0.003	-3.967	0.000	-0.015	-0.005
Creditors_Velocity_Days	0.0043	0.001	3.852	0.000	0.002	0.007
Current_Ratio_Latest	-0.7759	0.088	-8.789	0.000	-0.949	-0.603
Debtors_Velocity_Days	-0.0037	0.001	-3.774	0.000	-0.006	-0.002
Interest_Cover_Ratio_Latest	-0.1740	0.031	-5.702	0.000	-0.234	-0.114
ROG_Net_Worth_perc	-0.0580	0.007	-8.316	0.000	-0.072	-0.044
Market_Capitalisation	-0.0083	0.001	-5.871	0.000	-0.011	-0.00

Table 6: Model 3 results

The new model (model 3) has all the variables with $p < 0.05$. This model will be considered for **Test set prediction and performance evaluation**. **The adjusted pseudo-R-square value is 37.33**

Let us also check the multicollinearity of the model using Variance Inflation Factor (VIF) for the predictor variables

	variables_1	VIF_new
1	ROG_Cost_of_Production_perc	1.135404
0	ROG_Gross_Block_perc	1.292666
6	ROG_Net_Worth_perc	1.370568
5	Interest_Cover_Ratio_Latest	1.553236
3	Current_Ratio_Latest	1.612578
2	Creditors_Velocity_Days	2.007755
4	Debtors_Velocity_Days	2.082292
8	Market_Capitalisation	2.440010
7	Equity_Paid_Up	2.683910

We can see that multicollinearity still exists, but we will keep those variables as VIFs are not very high (<5)

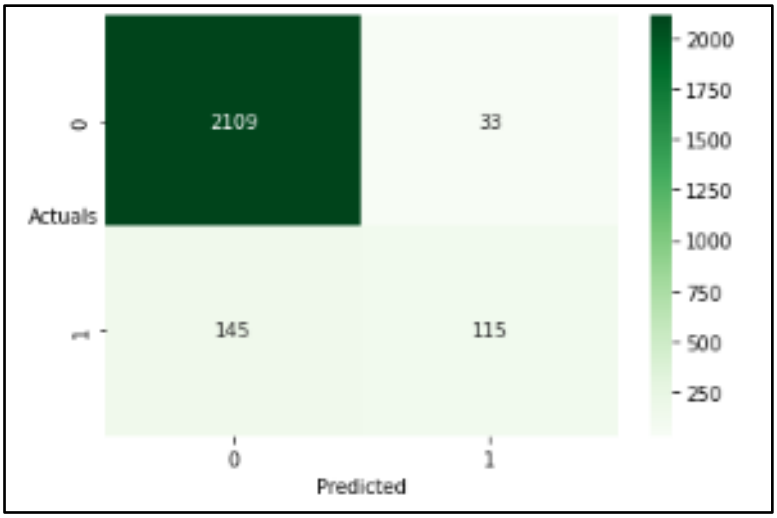
1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model.

Checking the predicted probability values:

```
842     0.005688
1057    0.000729
1595    0.000453
100     0.448392
1191    0.033383
          ...
1815    0.003182
2852    0.151420
1505    0.001055
375     0.516472
3428    0.000920
Length: 2402, dtype: float64
```

Model Evaluation on the Training Data:

Performance of 0.5 probability cut-off:

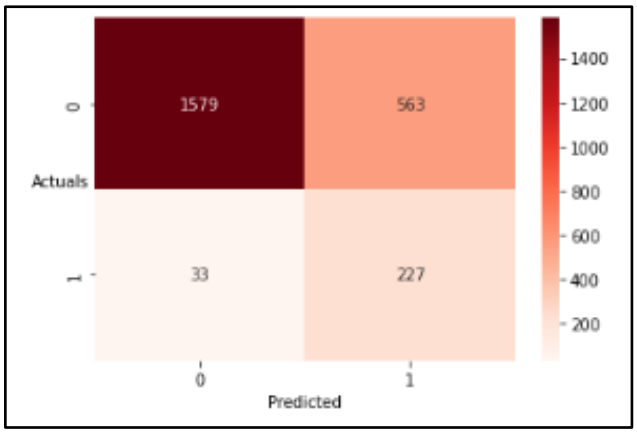


	precision	recall	f1-score	support
0	0.936	0.985	0.960	2142
1	0.777	0.442	0.564	260
accuracy			0.926	2402
macro avg	0.856	0.713	0.762	2402
weighted avg	0.918	0.926	0.917	2402

Overall, 93% of correct predictions to total predictions were made by the model. 44% of those defaulted were correctly identified as defaulters by the model, which is not so good number.

We will change the probability cut-offs and check if our predictions have improved.

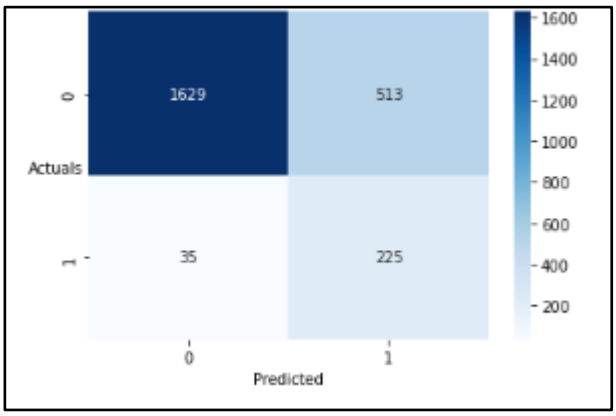
Let us take a cut-off of 0.07 and check our predictions



	precision	recall	f1-score	support
0	0.980	0.737	0.841	2142
1	0.287	0.873	0.432	260
accuracy			0.752	2402
macro avg	0.633	0.805	0.637	2402
weighted avg	0.905	0.752	0.797	2402

- Accuracy of the model i.e., %overall correct predictions has decreased from 93% to 75% but sensitivity of the model has increased from 44% to 87%, which is good for our prediction. But we will try with some more probability cut-off values.

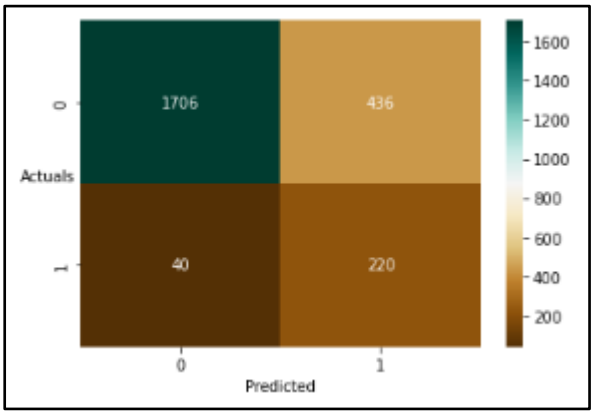
Let us take a cut-off of 0.08 and check our predictions



	precision	recall	f1-score	support
0	0.979	0.761	0.856	2142
1	0.305	0.865	0.451	260
accuracy			0.772	2402
macro avg	0.642	0.813	0.653	2402
weighted avg	0.906	0.772	0.812	2402

- Accuracy of the model i.e., %overall correct predictions has increased from 75% to 77% but sensitivity of the model remains same, which is good for our prediction. But we will try with some more probability cut-off values.

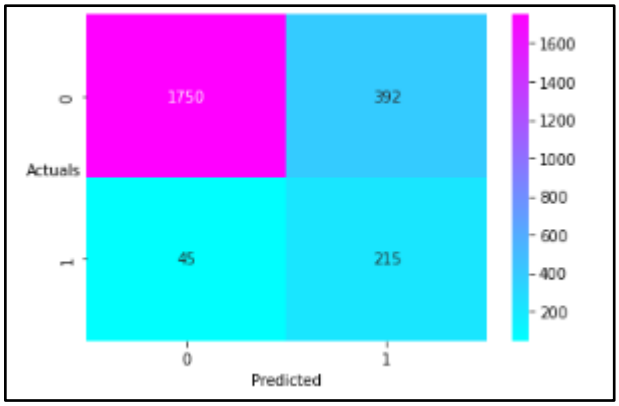
Let us take a cut-off of 0.10 and check our predictions



	precision	recall	f1-score	support
0	0.977	0.796	0.878	2142
1	0.335	0.846	0.480	260
accuracy			0.802	2402
macro avg	0.656	0.821	0.679	2402
weighted avg	0.908	0.802	0.835	2402

- Accuracy of the model i.e., %overall correct predictions has increased from 79% to 80% but sensitivity of the model remains same 85%. But we will try with some more probability cut-off values.

Let us take a cut-off of 0.115 and check our predictions



	precision	recall	f1-score	support
0	0.975	0.817	0.889	2142
1	0.354	0.827	0.496	260
accuracy			0.818	2402
macro avg	0.665	0.822	0.692	2402
weighted avg	0.908	0.818	0.846	2402

AUC and ROC for training data:

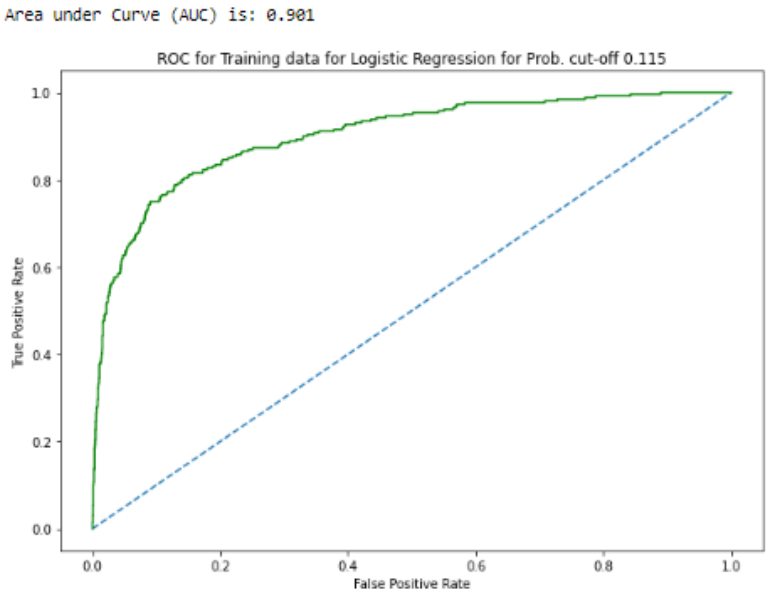
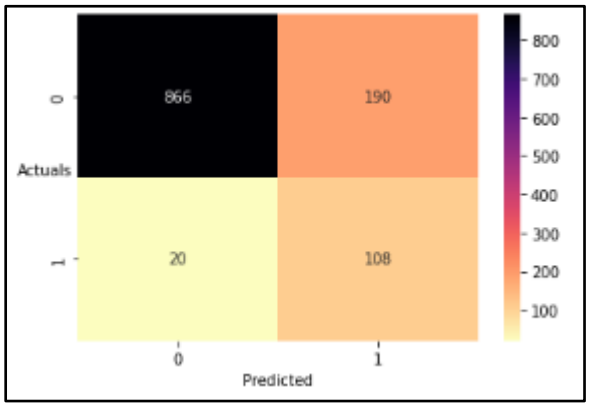


Figure 10: AUC and ROC for training data

Accuracy of the model i.e., %overall correct predictions has increased from 80% to 82% but sensitivity of the model has decreased slightly from 85% to 83%.

Model Evaluation on the Test Data:



	precision	recall	f1-score	support
0	0.977	0.820	0.892	1056
1	0.362	0.844	0.507	128
accuracy			0.823	1184
macro avg	0.670	0.832	0.699	1184
weighted avg	0.911	0.823	0.850	1184

AUC and ROC for test data:

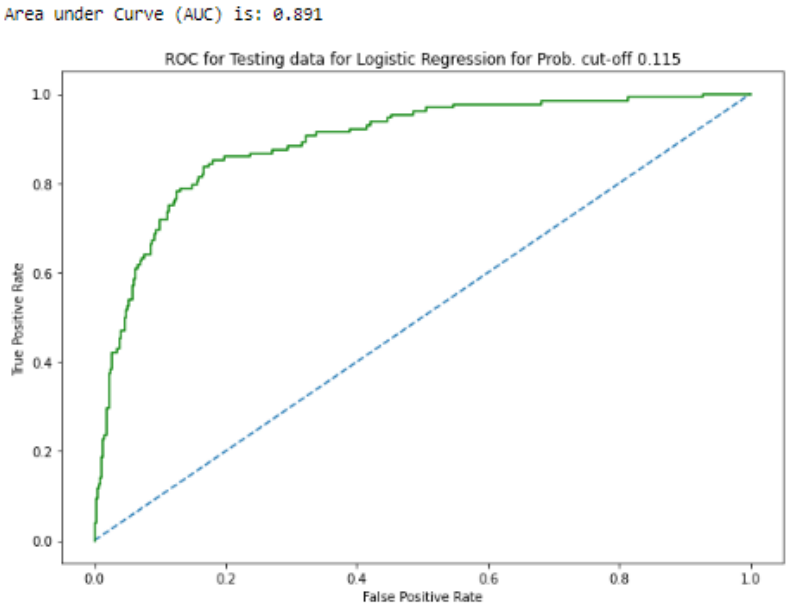


Figure 11: AUC and ROC for test data

Accuracy of the model i.e., % overall correct prediction is 82% and sensitivity of the model is 84%.

Hence, model performs well with both train and test set.

1.8. Conclusion

- The dataset was huge, hence clean and processed into a neat dataset.
- The outliers were treated.
- Univariate and Bivariate analysis was done to check the data behaviour and any early sign of multi-collinearity.
- We used various feature selection and elimination method to remove huge number of redundant variables.
- We used Logistic Regression for modelling and taken two different threshold values to predict the output.
- One threshold was manually set, and another was calculated.
- Predictions were made based on the final model.
- For reference, Jupyter Notebook file has been included.