FRA Milestone - 1

Project Report

by

Pinak Pani Gogoi

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 Networth 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 Set : Company_Data2015-1.xlsx,
Data Dictionary : Credit Default Data Dictionary.xlsx

#	Field Name	Description
1	Co_Code	Company Code
2	Co_Name	Company Name
3	Networth Next Year	Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities)
4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders
5	Networth	Value of a company as on 2015 - Current Year
6	Capital Employed	Total amount of capital used for the acquisition of profits by a company
7	Total Debt	The sum of money borrowed by the company and is due to be paid
8	Gross Block	Total value of all of the assets that a company owns
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).
10	Current Assets	All the assets of a company that are expected to be sold or used as a result of standard business operations over the next year.
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)
12	Total Assets/Liabilities	Ratio of total assets to liabailities of the company
13	Gross Sales	The grand total of sale transactions within the accounting period
14	Net Sales	Gross sales minus returns, allowances, and discounts
15	Other Income	Income realized from non-business activities (e.g. sale of long term asset)
16	Value Of Output	Product of physical output of goods and services produced by company and its market price
17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service
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)
19	PBIDT	Profit Before Interest, Depreciation & Taxes
20	PBDT	Profit Before Depreciation and Tax
21	PBIT	Profit before interest and taxes
22	PBT	Profit before tax
23	PAT	Profit After Tax
24	Adjusted PAT	Adjusted profit is the best estimate of the true profit
26	CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.
27	Revenue earnings in forex	Revenue earned in foreign currency
28	Revenue expenses in forex	Expenses due to foreign currency transactions
29	Capital expenses in forex	Long term investment in forex
30	Book Value (Unit Curr)	Net asset value
31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value
32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share
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
34	Cash Flow From Operating Activities	Use of cash from ongoing regular business activities
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
36	Cash Flow From Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)
37	ROG-Net Worth (%)	Rate of Growth - Networth
38	ROG-Capital Employed (%)	Rate of Growth - Capital Employed
39	ROG-Gross Block (%)	Rate of Growth - Gross Block
40	ROG-Gross Sales (%)	Rate of Growth - Gross Sales
41	ROG-Net Sales (%)	Rate of Growth - Net Sales
42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production
43	ROG-Total Assets (%)	Rate of Growth - Total Assets
44	ROG-PBIDT (%)	Rate of Growth- PBIDT
45	ROG-PBDT (%)	Rate of Growth- PBDT
46	ROG-PBIT (%)	Rate of Growth- PBIT
47	ROG-PBT (%)	Rate of Growth- PBT
48	ROG-PAT (%)	Rate of Growth- PAT
49	ROG-CP (%)	Rate of Growth- CP
50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex
51	ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex
52	ROG-Market Capitalisation (%)	Rate of Growth - Market Capitalisation
53	Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year
54	Fixed Assets Ratio[Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating
55	Inventory Ratio[Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company
56	Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company
57	Total Asset Turnover Ratio[Latest]	The value of a company's revenues relative to the value of its assets
58	Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt
59	PBIDTM (%)[Latest]	Profit before Interest Depreciation and Tax Margin
60	PBITM (%)[Latest]	Profit Before Interest Tax Margin
61	PBDTM (%)[Latest]	Profit Before Depreciation Tax Margin
62	CPM (%)[Latest]	Cost per thousand (advertising cost)
63	APATM (%)[Latest]	After tax profit margin
64	Debtors Velocity (Days)	Average days required for receiving the payments
65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers
66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales
67	Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets
68	Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block

Many of the Variables (Field names) are presented in absolute number as well as % number. These may be highly related to each other. But we are considering it as distinct variable as exact definition/calculation formula is not provided.

Exploratory Data Analysis (EDA)

EDA and model building was executed using Python in Jupyter notebook. The Jupyter notebook details has been shared below. Please refer Pinak-Predictive-Modelling-Project-ANSWER-1.ipynb for code details.

About dataset:

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	Other Income	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block
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	7.60	0	45.0	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	46.27	101	2.0	0.31	0.24
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	9.55	558	0.0	-0.03	-0.26
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	223.85	63	2.0	0.24	1.90
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	9.82	346	0.0	0.01	0.05

Key points:

- 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.

Target variable:

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” (other variables are not displayed for clarity).

	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 seen that almost 11% of the total entries in "Default" belong to category "1". The dataset has class imbalance issue.

Data type and Missing value:

All the variables are of numeric types except Target variable “Default” and “Co_Name”.
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. The following figure shows the overall data types and the variable with missing values.

Data info:

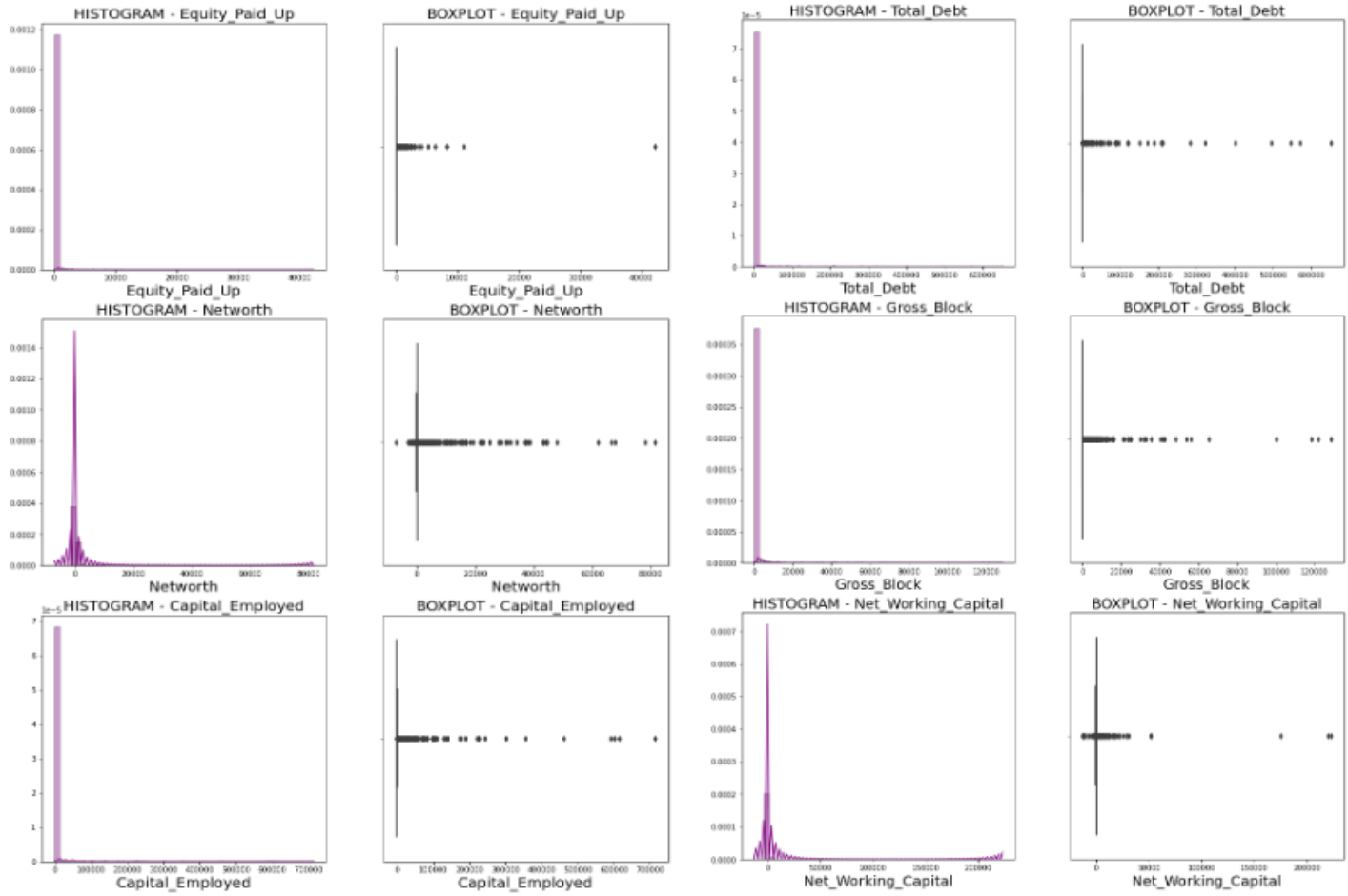
#	Column	Non-Null Count	Dtype
0	Co_Code	3586 non-null	int64
1	Co_Name	3586 non-null	object
2	Networth_Next_Year	3586 non-null	float64
3	Equity_Paid_Up	3586 non-null	float64
4	Networth	3586 non-null	float64
5	Capital_Employed	3586 non-null	float64
6	Total_Debt	3586 non-null	float64
7	Gross_Block	3586 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
67	Default	3586 non-null	int64

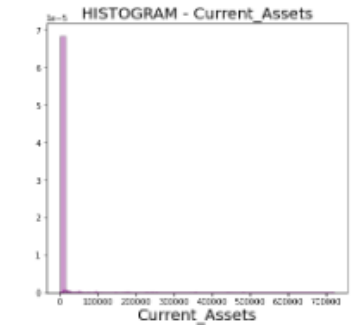
The following figure shows the missing value columns,

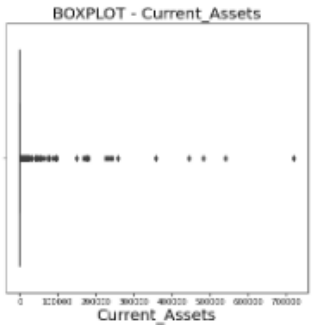
Book_Value_Adj._Unit_Curr	4
Current_Ratio_Latest	1
Fixed_Assets_Ratio_Latest	1
Inventory_Ratio_Latest	1
Debtors_Ratio_Latest	1
Total_Asset_Turnover_Ratio_Latest	1
Interest_Cover_Ratio_Latest	1
PBIDTM_perc_Latest	1
PBITM_perc_Latest	1
PBDTM_perc_Latest	1
CPM_perc_Latest	1
APATM_perc_Latest	1
Inventory_Velocity_Days	103

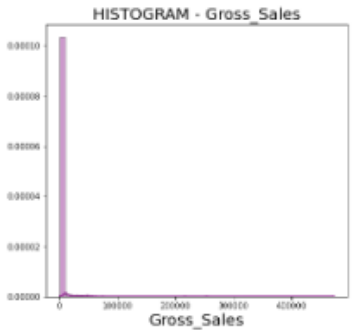
Univariate analysis:

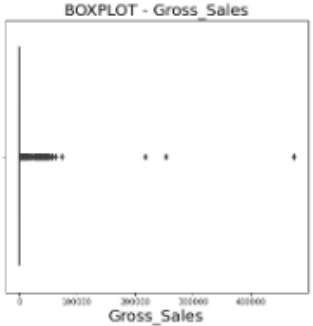
Univariate analysis involving data distribution along with outlier detection (Boxplot) plots have been shown below. Due to large number of variables, the number of plots will be high.

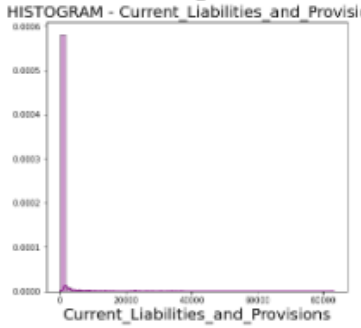



HISTOGRAM - Current_Assets
Current_Assets

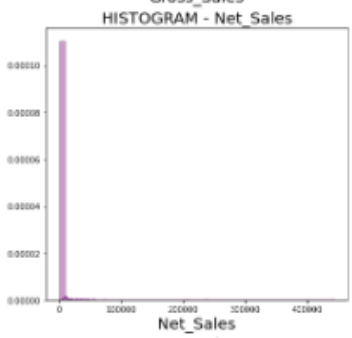

BOXPLOT - Current_Assets
Current_Assets

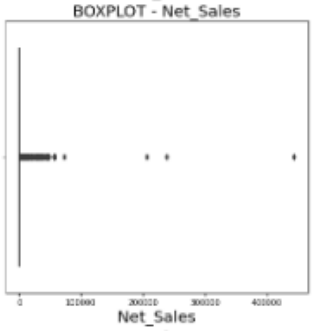

HISTOGRAM - Gross_Sales
Gross_Sales

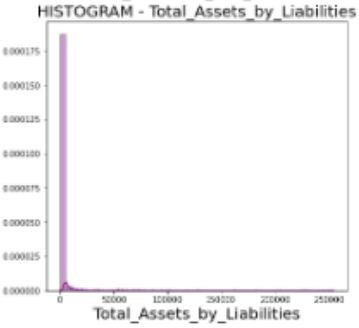

BOXPLOT - Gross_Sales
Gross_Sales

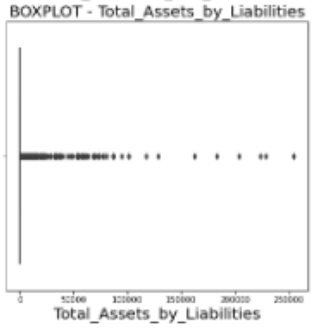

HISTOGRAM - Current_Liabilities_and_Provisions
Current_Liabilities_and_Provisions

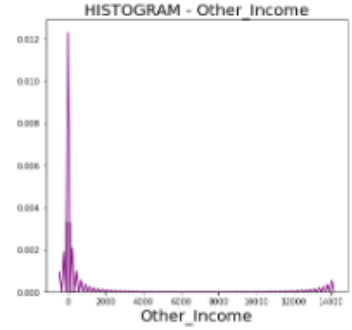

BOXPLOT - Current_Liabilities_and_Provisio
Current_Liabilities_and_Provisions

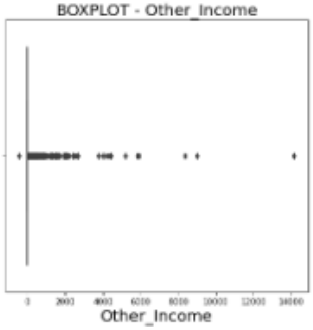

HISTOGRAM - Net_Sales
Net_Sales

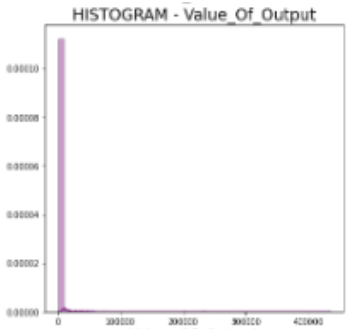

BOXPLOT - Net_Sales
Net_Sales

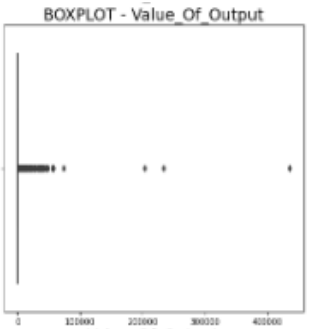

HISTOGRAM - Total_Assets_by_Liabilities
Total_Assets_by_Liabilities


BOXPLOT - Total_Assets_by_Liabilities
Total_Assets_by_Liabilities

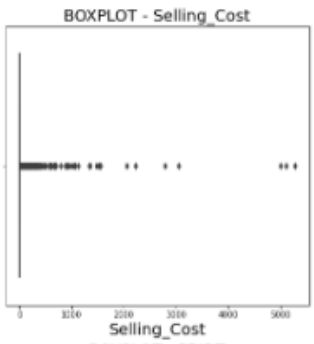

HISTOGRAM - Other_Income
Other_Income


BOXPLOT - Other_Income
Other_Income

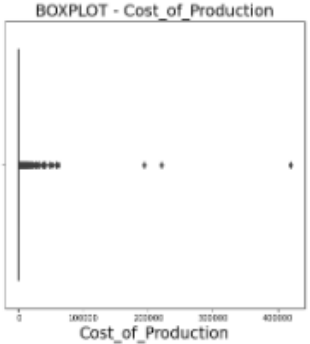

HISTOGRAM - Value_Of_Output
Value_Of_Output

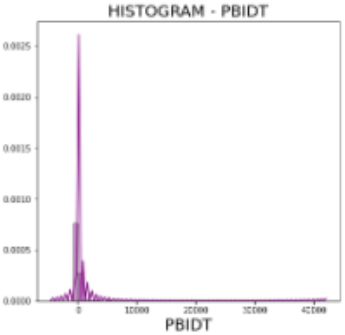

BOXPLOT - Value_Of_Output
Value_Of_Output

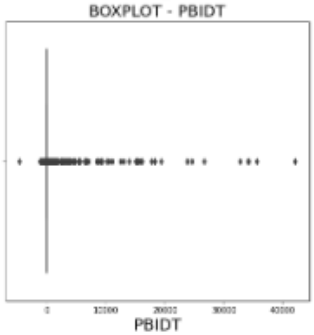

HISTOGRAM - Selling_Cost
Selling_Cost

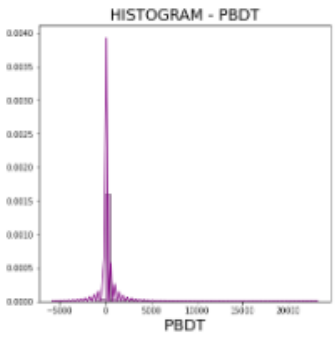

BOXPLOT - Selling_Cost
Selling_Cost

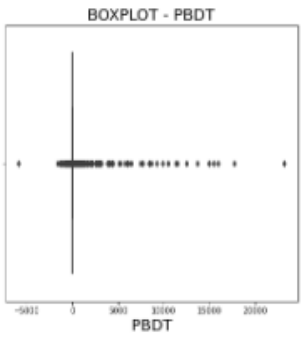

HISTOGRAM - Cost_of_Production
Cost_of_Production

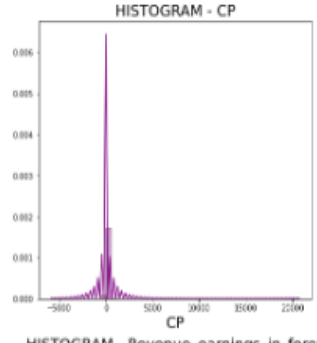

BOXPLOT - Cost_of_Production
Cost_of_Production

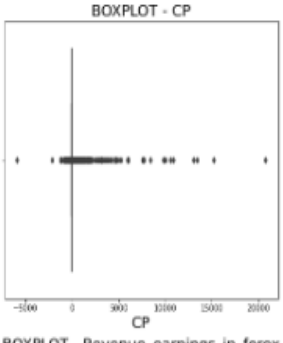

HISTOGRAM - PBIDT
PBIDT

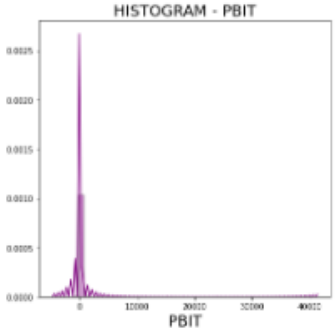

BOXPLOT - PBIDT
PBIDT

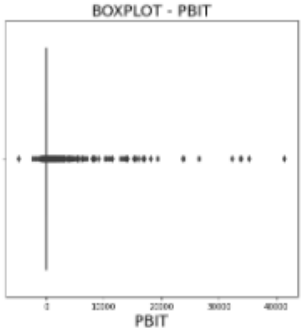

HISTOGRAM - PBDT
PBDT


BOXPLOT - PBDT
PBDT

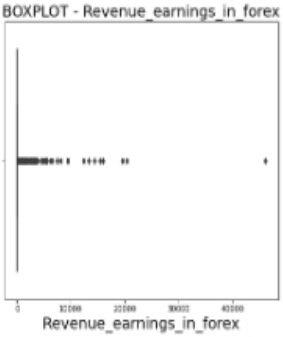

HISTOGRAM - CP
CP

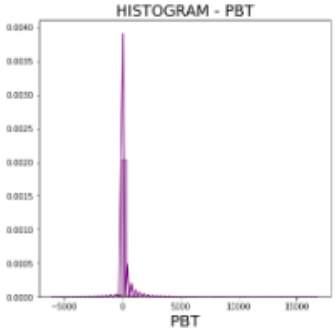

BOXPLOT - CP
CP

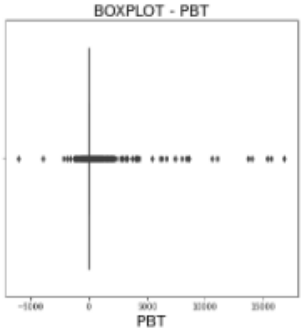

HISTOGRAM - PBIT
PBIT

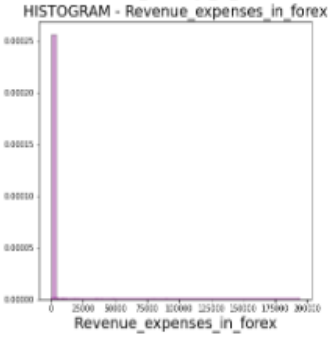

BOXPLOT - PBIT
PBIT

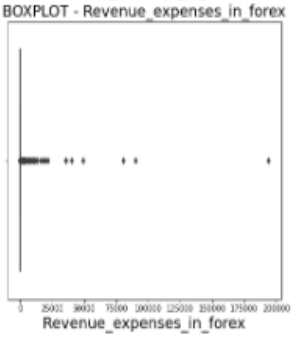

HISTOGRAM - Revenue_earnings_in_forex
Revenue_earnings_in_forex

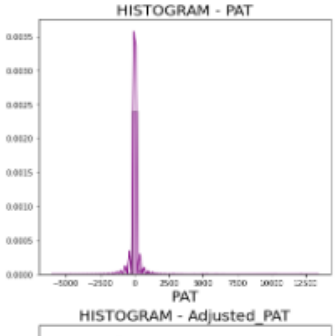

BOXPLOT - Revenue_earnings_in_forex
Revenue_earnings_in_forex

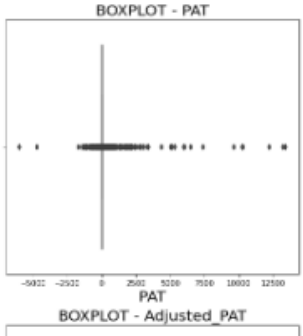

HISTOGRAM - PBT
PBT


BOXPLOT - PBT
PBT

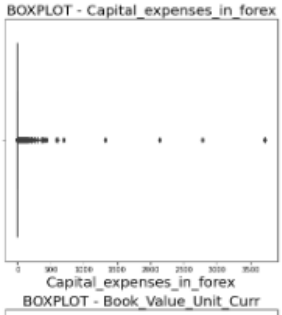

HISTOGRAM - Revenue_expenses_in_forex
Revenue_expenses_in_forex

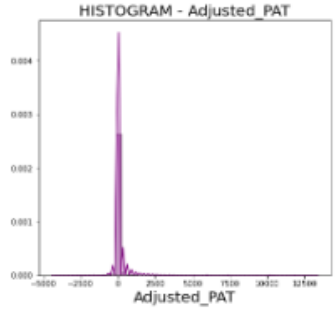

BOXPLOT - Revenue_expenses_in_forex
Revenue_expenses_in_forex

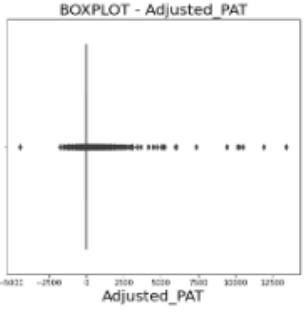

HISTOGRAM - PAT
PAT

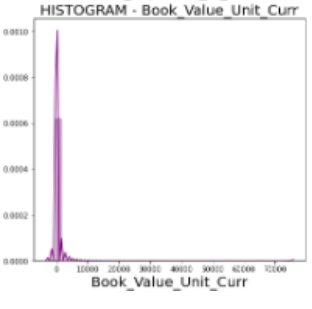

BOXPLOT - PAT
PAT

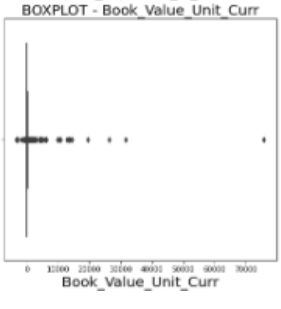

HISTOGRAM - Capital_expenses_in_forex
Capital_expenses_in_forex


BOXPLOT - Capital_expenses_in_forex
Capital_expenses_in_forex

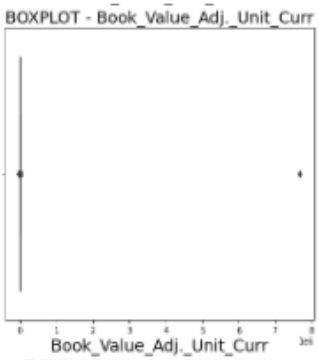

HISTOGRAM - Adjusted_PAT
Adjusted_PAT

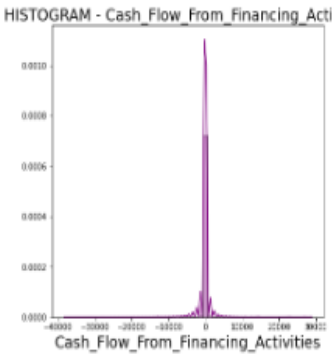

BOXPLOT - Adjusted_PAT
Adjusted_PAT


HISTOGRAM - Book_Value_Unit_Curr
Book_Value_Unit_Curr


BOXPLOT - Book_Value_Unit_Curr
Book_Value_Unit_Curr

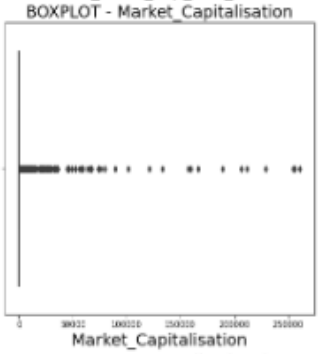

HISTOGRAM - Book_Value_Adj._Unit_Curr
Book_Value_Adj._Unit_Curr

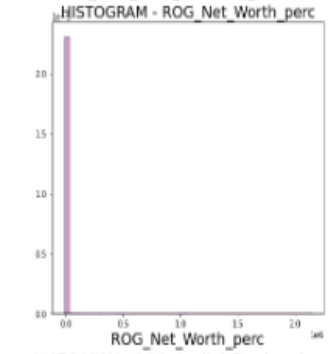

BOXPLOT - Book_Value_Adj._Unit_Curr
Book_Value_Adj._Unit_Curr

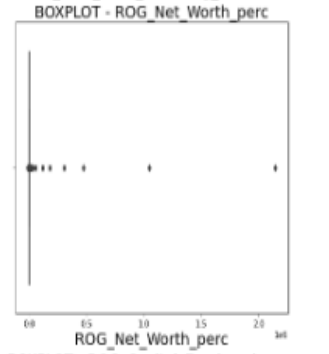

HISTOGRAM - Cash_Flow_From_Financing_Activities
Cash_Flow_From_Financing_Activities

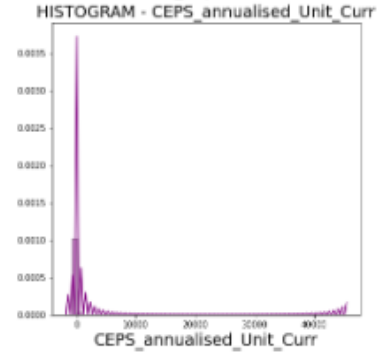

BOXPLOT - Cash_Flow_From_Financing_Activities
Cash_Flow_From_Financing_Activities

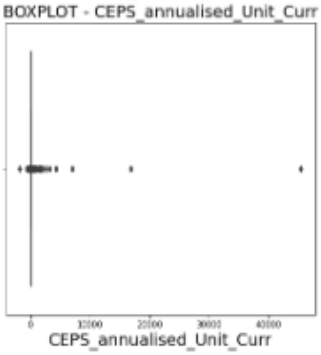

HISTOGRAM - Market_Capitalisation
Market_Capitalisation

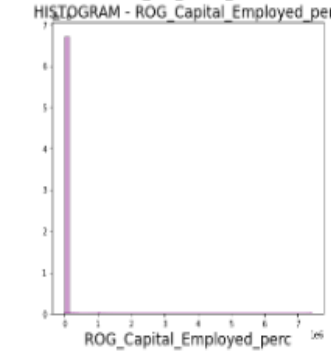

BOXPLOT - Market_Capitalisation
Market_Capitalisation

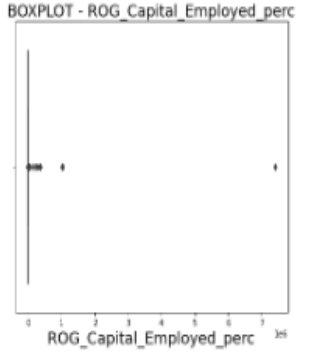

HISTOGRAM - ROG_Net_Worth_perc
ROG_Net_Worth_perc

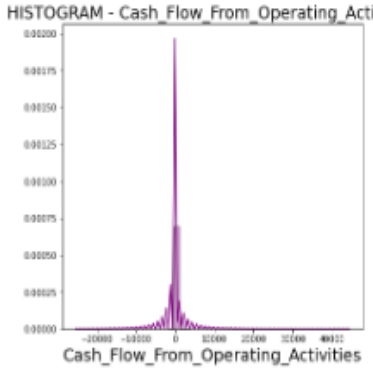

BOXPLOT - ROG_Net_Worth_perc
ROG_Net_Worth_perc


HISTOGRAM - CEPS_annualised_Unit_Curr
CEPS_annualised_Unit_Curr

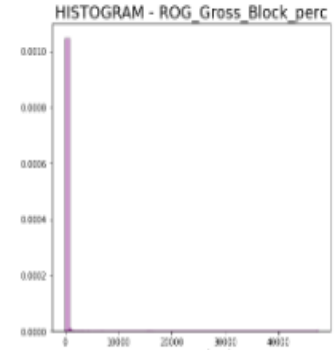

BOXPLOT - CEPS_annualised_Unit_Curr
CEPS_annualised_Unit_Curr

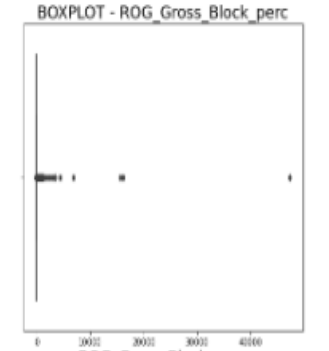

HISTOGRAM - ROG_Capital_Employed_perc
ROG_Capital_Employed_perc

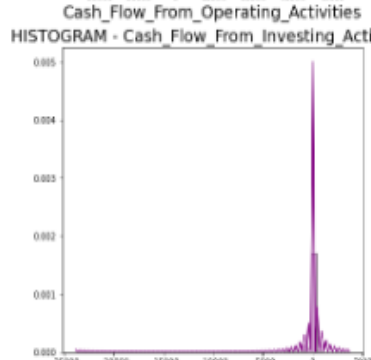

BOXPLOT - ROG_Capital_Employed_perc
ROG_Capital_Employed_perc


HISTOGRAM - Cash_Flow_From_Operating_Activities
Cash_Flow_From_Operating_Activities

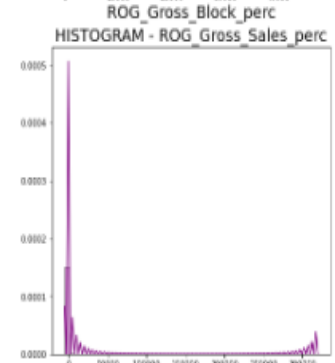

BOXPLOT - Cash_Flow_From_Operating_Activities
Cash_Flow_From_Operating_Activities


HISTOGRAM - ROG_Gross_Block_perc
ROG_Gross_Block_perc

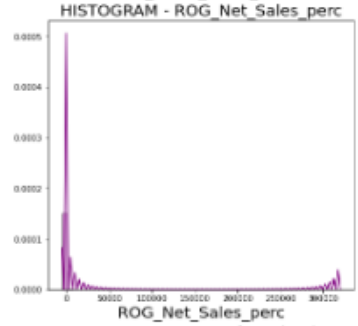

BOXPLOT - ROG_Gross_Block_perc
ROG_Gross_Block_perc

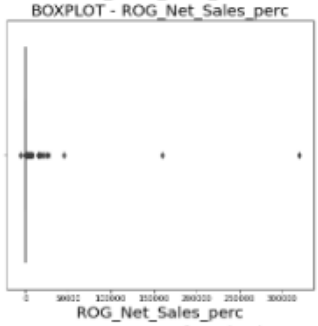

HISTOGRAM - Cash_Flow_From_Investing_Activities
Cash_Flow_From_Investing_Activities

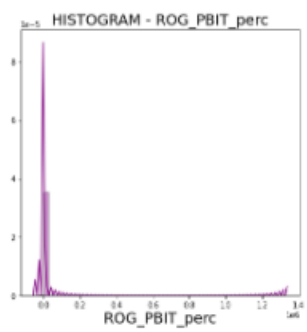

BOXPLOT - Cash_Flow_From_Investing_Activities
Cash_Flow_From_Investing_Activities

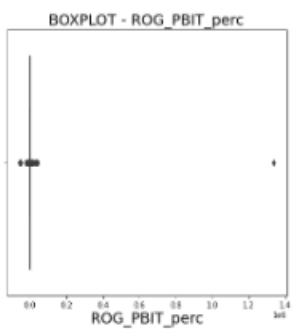

HISTOGRAM - ROG_Gross_Sales_perc
ROG_Gross_Sales_perc

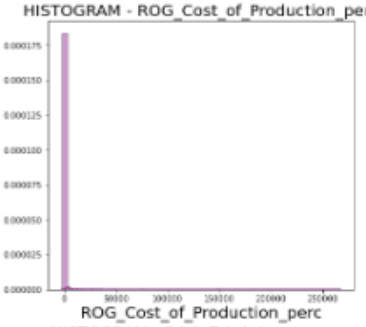

BOXPLOT - ROG_Gross_Sales_perc
ROG_Gross_Sales_perc

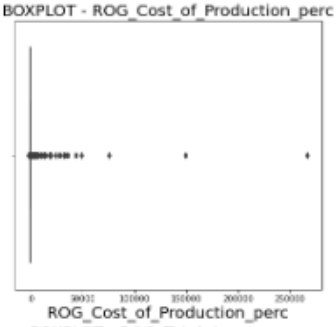

HISTOGRAM - ROG_Net_Sales_perc
ROG_Net_Sales_perc

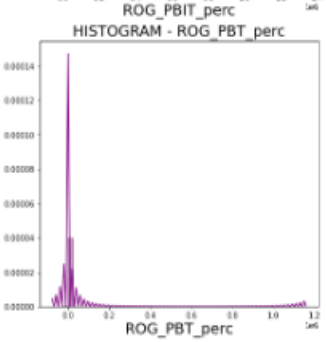

BOXPLOT - ROG_Net_Sales_perc
ROG_Net_Sales_perc

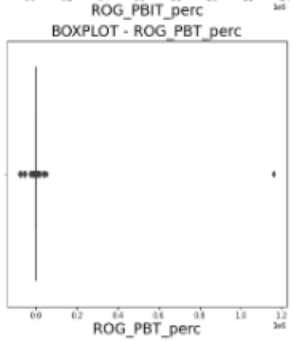

HISTOGRAM - ROG_PBIT_perc
ROG_PBIT_perc

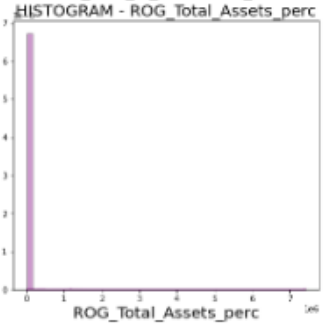

BOXPLOT - ROG_PBIT_perc
ROG_PBIT_perc

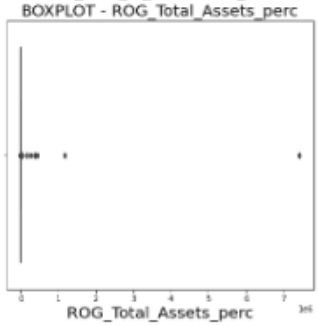

HISTOGRAM - ROG_Cost_of_Production_perc
ROG_Cost_of_Production_perc

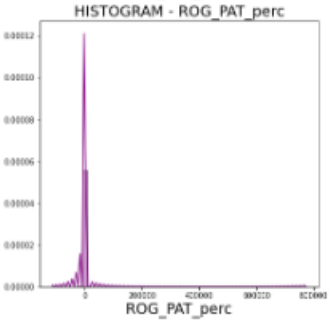

BOXPLOT - ROG_Cost_of_Production_perc
ROG_Cost_of_Production_perc

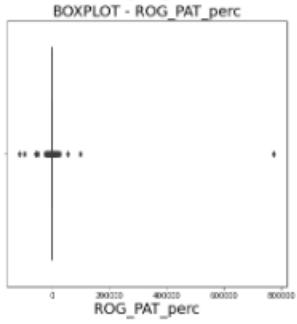

HISTOGRAM - ROG_PBT_perc
ROG_PBT_perc

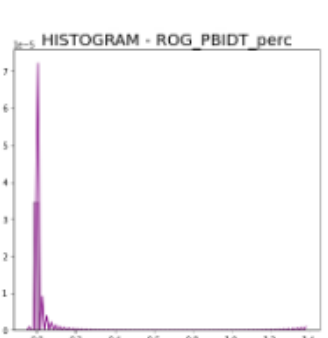

BOXPLOT - ROG_PBT_perc
ROG_PBT_perc

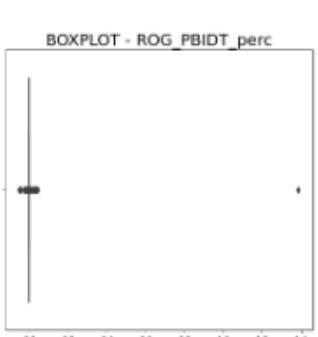

HISTOGRAM - ROG_Total_Assets_perc
ROG_Total_Assets_perc

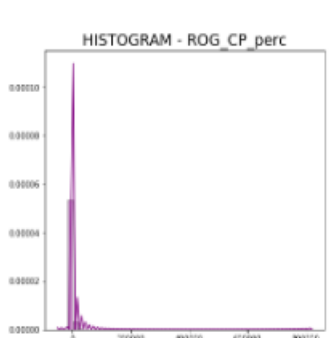

BOXPLOT - ROG_Total_Assets_perc
ROG_Total_Assets_perc

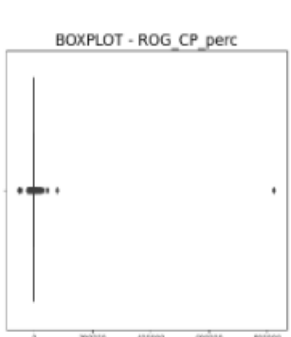

HISTOGRAM - ROG_PAT_perc
ROG_PAT_perc

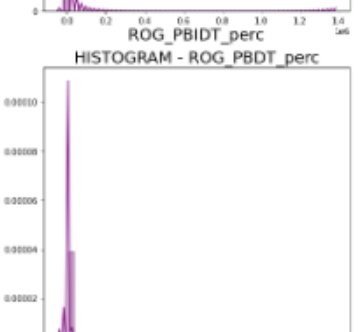

BOXPLOT - ROG_PAT_perc
ROG_PAT_perc

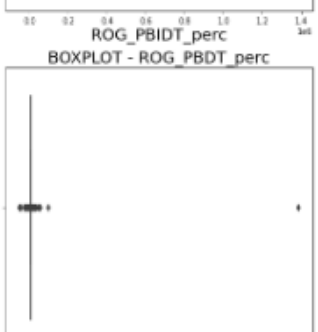

HISTOGRAM - ROG_PBIDT_perc
ROG_PBIDT_perc

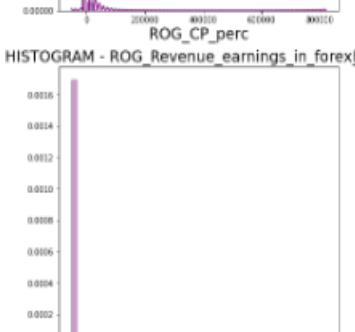

BOXPLOT - ROG_PBIDT_perc
ROG_PBIDT_perc


HISTOGRAM - ROG_CP_perc
ROG_CP_perc


BOXPLOT - ROG_CP_perc
ROG_CP_perc


HISTOGRAM - ROG_PBDT_perc
ROG_PBDT_perc

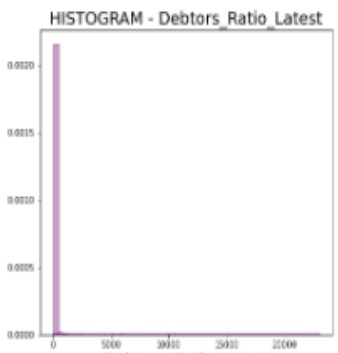

BOXPLOT - ROG_PBDT_perc
ROG_PBDT_perc

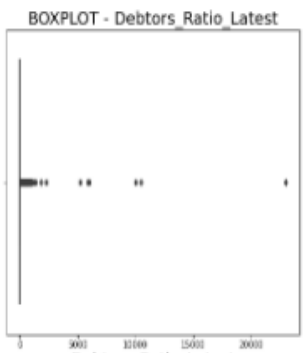

HISTOGRAM - ROG_Revenue_earnings_in_forex_perc
ROG_Revenue_earnings_in_forex_perc

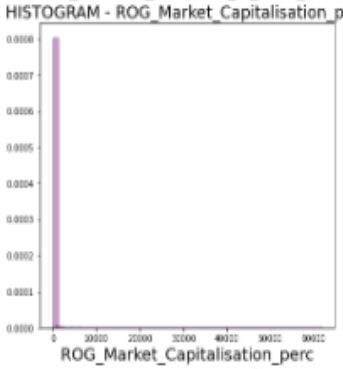

BOXPLOT - ROG_Revenue_earnings_in_forex_perc
ROG_Revenue_earnings_in_forex_perc

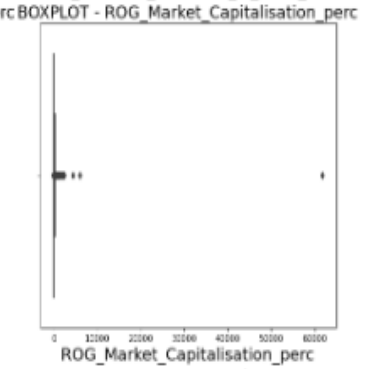

HISTOGRAM - ROG_Revenue_expenses_in_forex_perc
ROG_Revenue_expenses_in_forex_perc

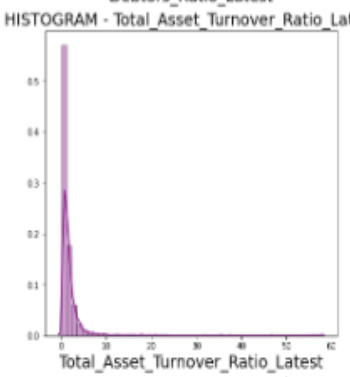

BOXPLOT - ROG_Revenue_expenses_in_forex_pe
ROG_Revenue_expenses_in_forex_perc

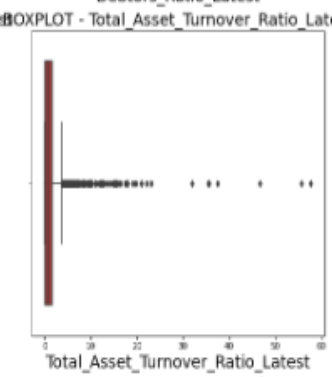

HISTOGRAM - Debtors_Ratio_Latest
Debtors_Ratio_Latest

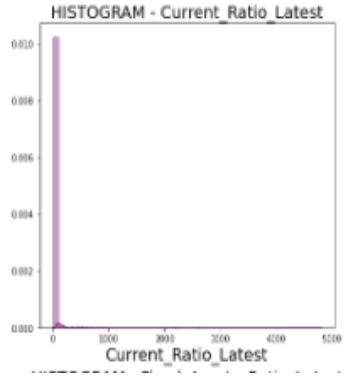

BOXPLOT - Debtors_Ratio_Latest
Debtors_Ratio_Latest

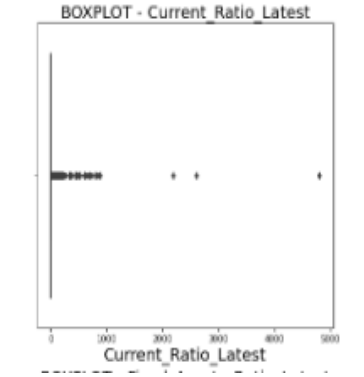

HISTOGRAM - ROG_Market_Capitalisation_perc
ROG_Market_Capitalisation_perc

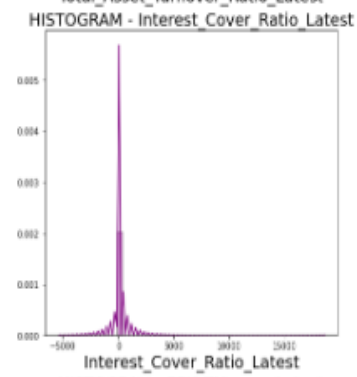

BOXPLOT - ROG_Market_Capitalisation_perc
ROG_Market_Capitalisation_perc

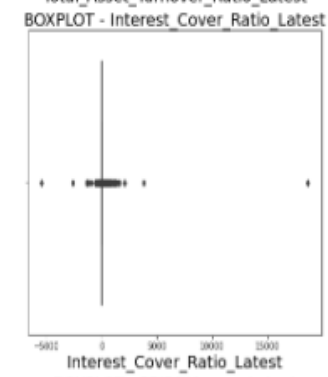

HISTOGRAM - Total_Asset_Turnover_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest

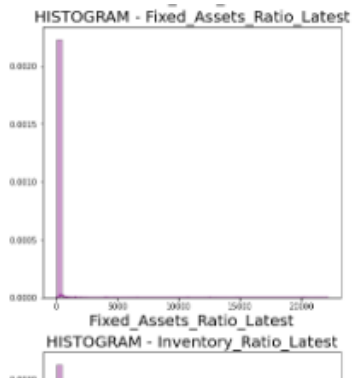

BOXPLOT - Total_Asset_Turnover_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest

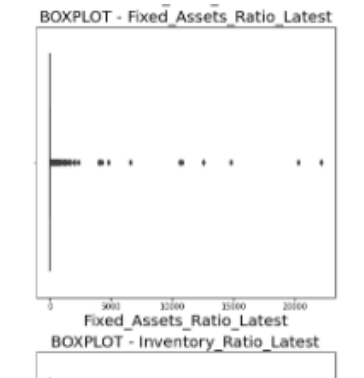

HISTOGRAM - Current_Ratio_Latest
Current_Ratio_Latest

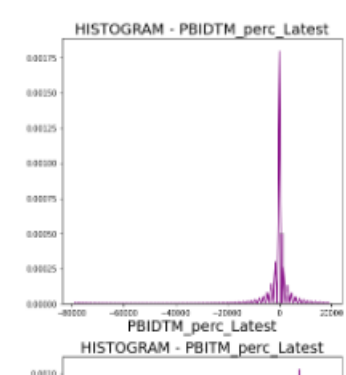

BOXPLOT - Current_Ratio_Latest
Current_Ratio_Latest

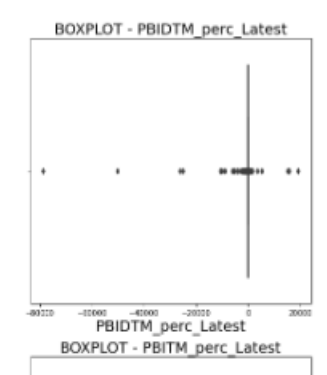

HISTOGRAM - Interest_Cover_Ratio_Latest
Interest_Cover_Ratio_Latest

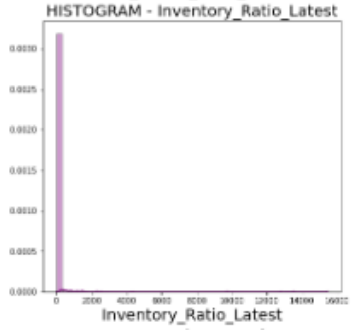

BOXPLOT - Interest_Cover_Ratio_Latest
Interest_Cover_Ratio_Latest

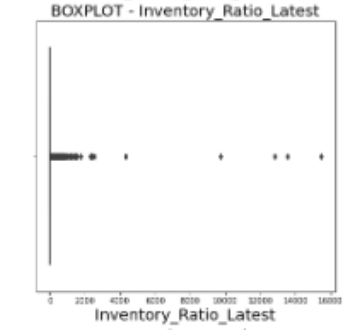

HISTOGRAM - Fixed_Assets_Ratio_Latest
Fixed_Assets_Ratio_Latest

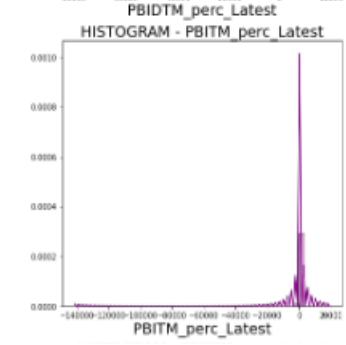

BOXPLOT - Fixed_Assets_Ratio_Latest
Fixed_Assets_Ratio_Latest

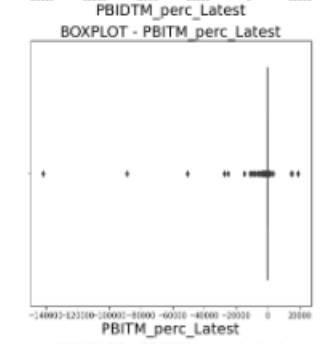

HISTOGRAM - PBIDTM_perc_Latest
PBIDTM_perc_Latest

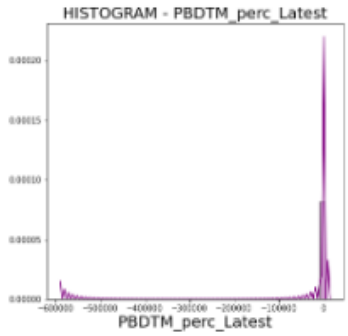

BOXPLOT - PBIDTM_perc_Latest
PBIDTM_perc_Latest

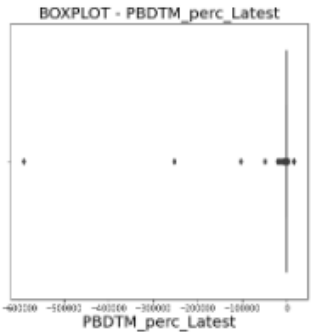

HISTOGRAM - Inventory_Ratio_Latest
Inventory_Ratio_Latest

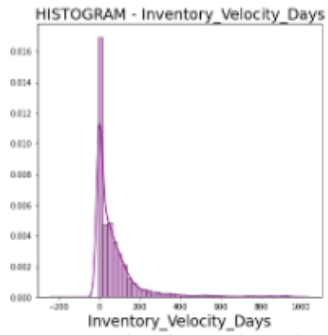

BOXPLOT - Inventory_Ratio_Latest
Inventory_Ratio_Latest

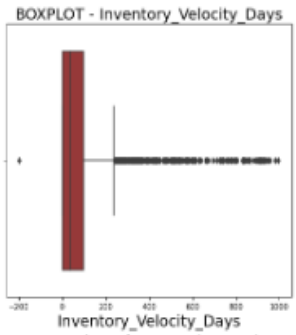

HISTOGRAM - PBITM_perc_Latest
PBITM_perc_Latest

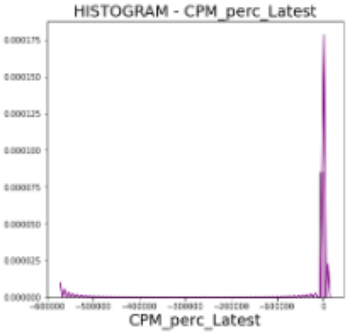

BOXPLOT - PBITM_perc_Latest
PBITM_perc_Latest

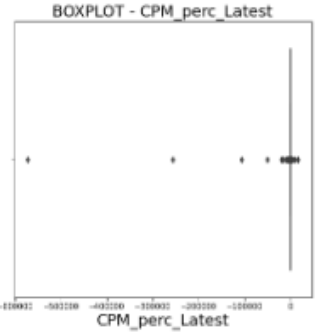


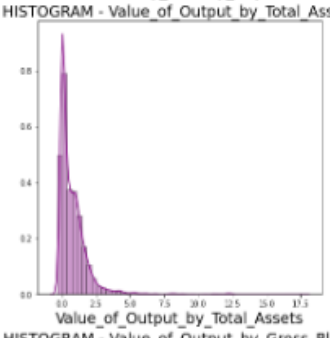


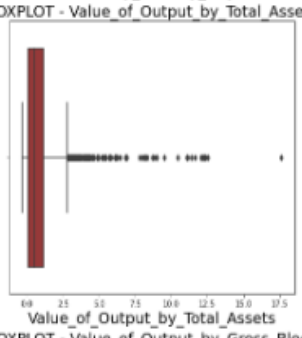


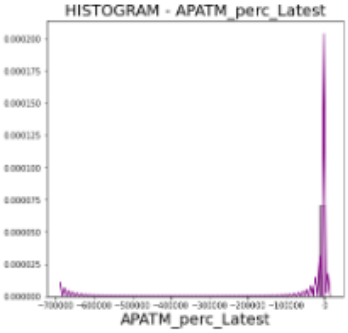


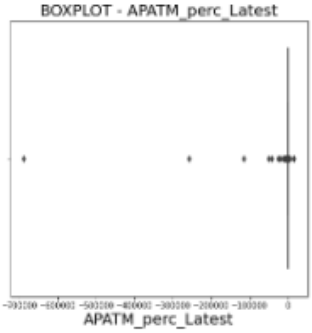


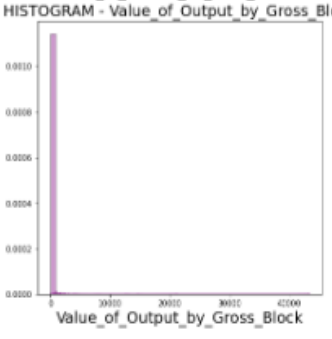


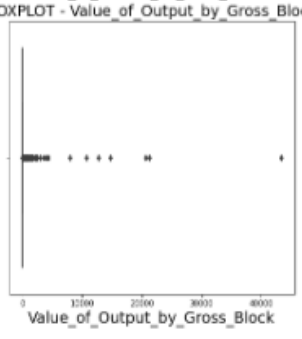


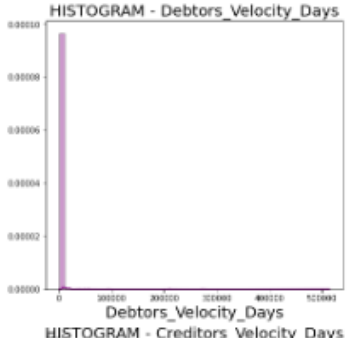


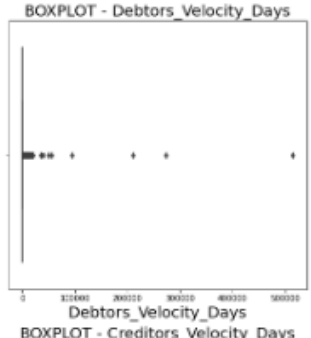


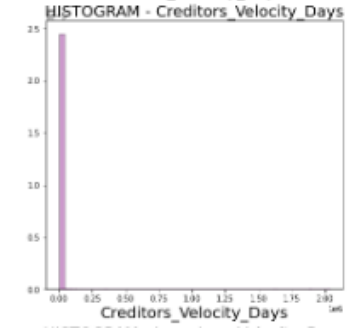


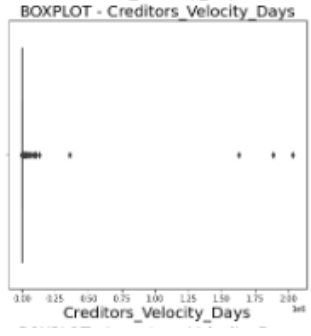










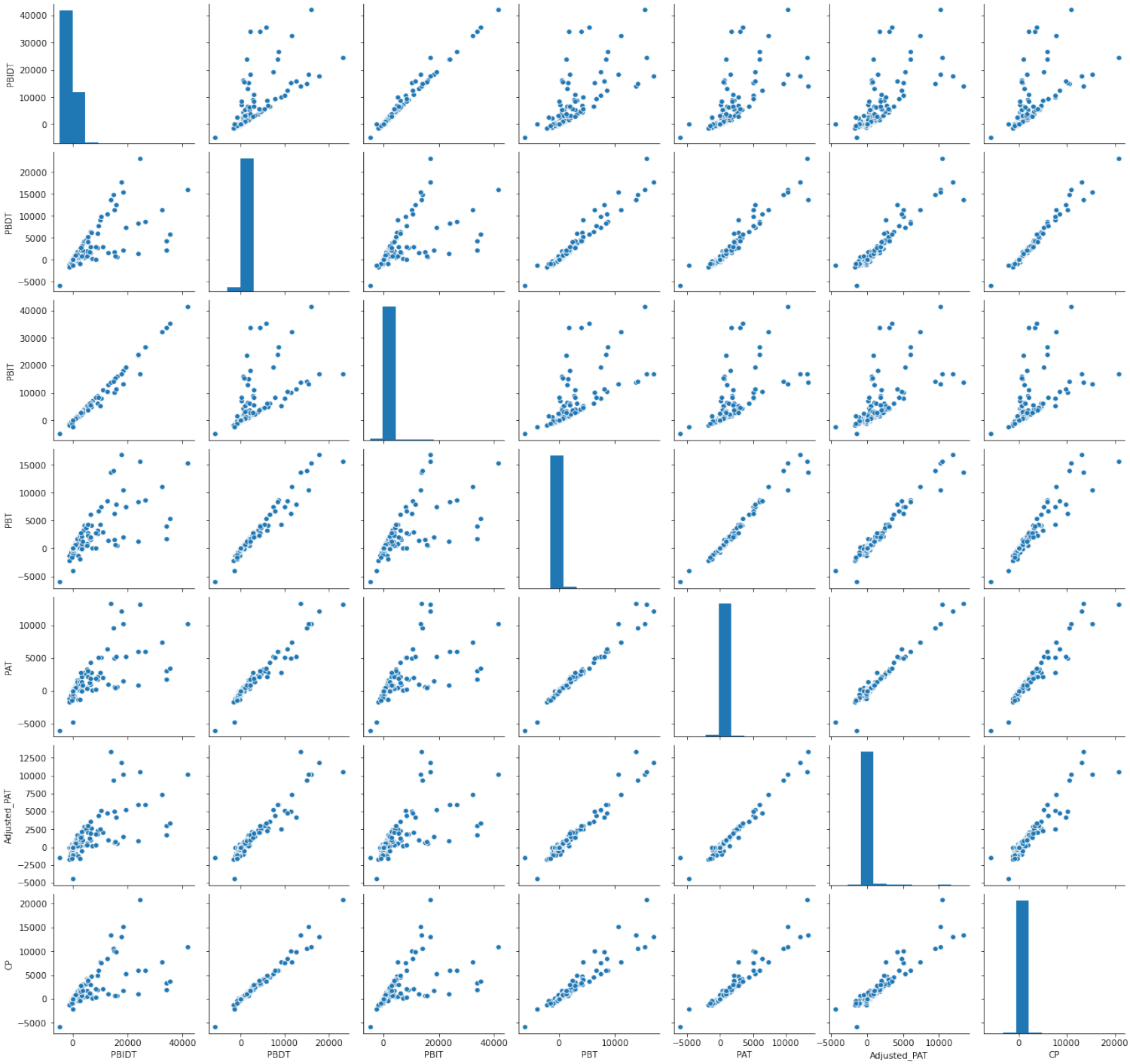


Most of the variables have skewed distribution. Also all the variables have outliers. These ouliers will be treated as we are going to apply Logistic regression to predict the outcome.

Bi-variate Analysis:

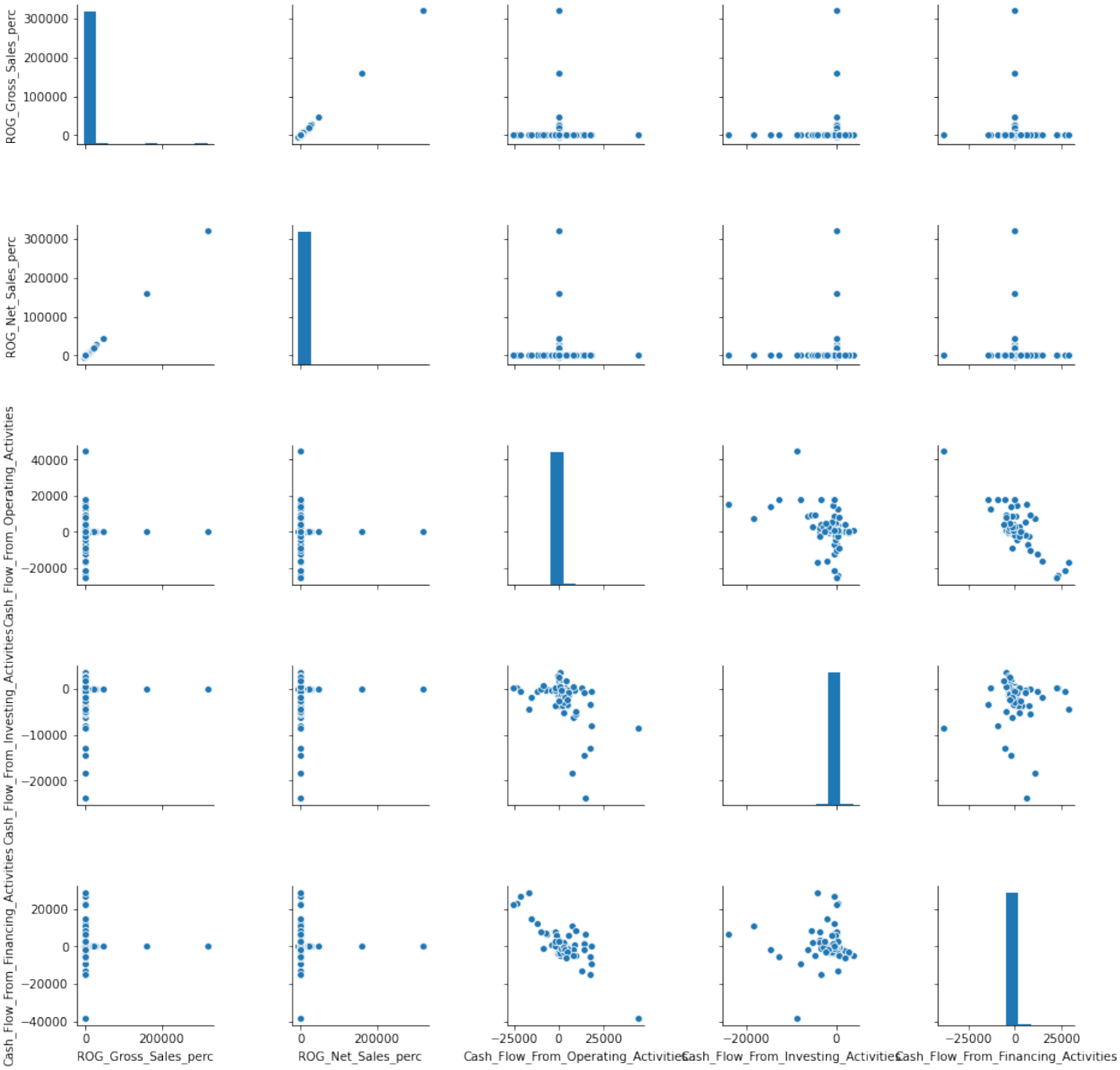
Bi-variate analysis includes pairplot and heatmap of correlation matrix. As the number of variables are high, the pairplot would not be so legible. For that resaon, the pairplots are diplayed for variables which are significant (derived using VIF score) in model prediction and which have significant corelations among each other.

Pairplot - 1



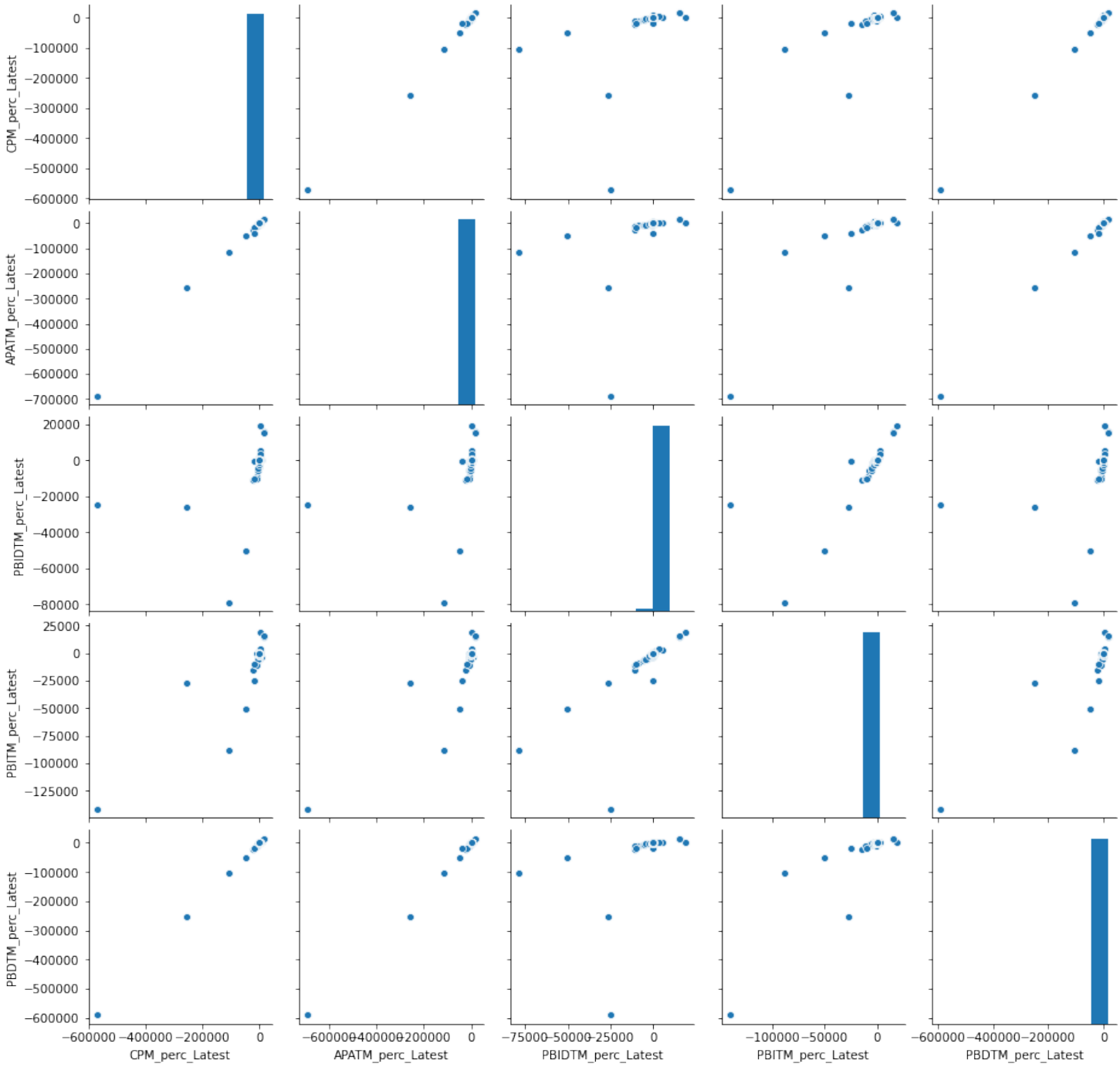
It is observed that there is high positive correlation between variables PBDT, PBIDT, PBIT, PBT, PAT, Adjusted PAT and CP.

Pairplot - 2



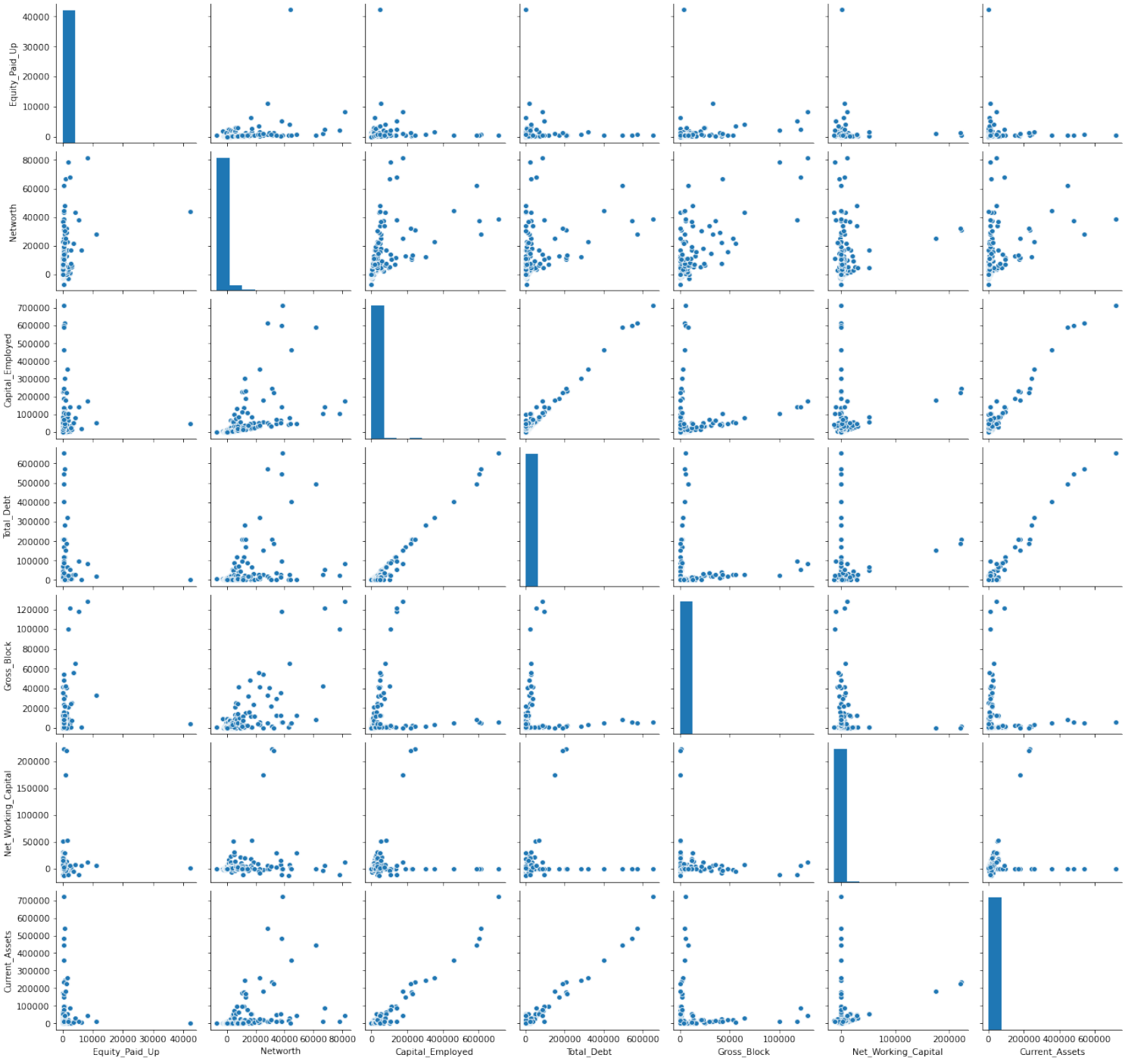
It is observed that there is negative correlation between variables Cash_Flow_From_Operating_Activities and Cash_Flow_From_Financing_Activities.

Pairplot - 3



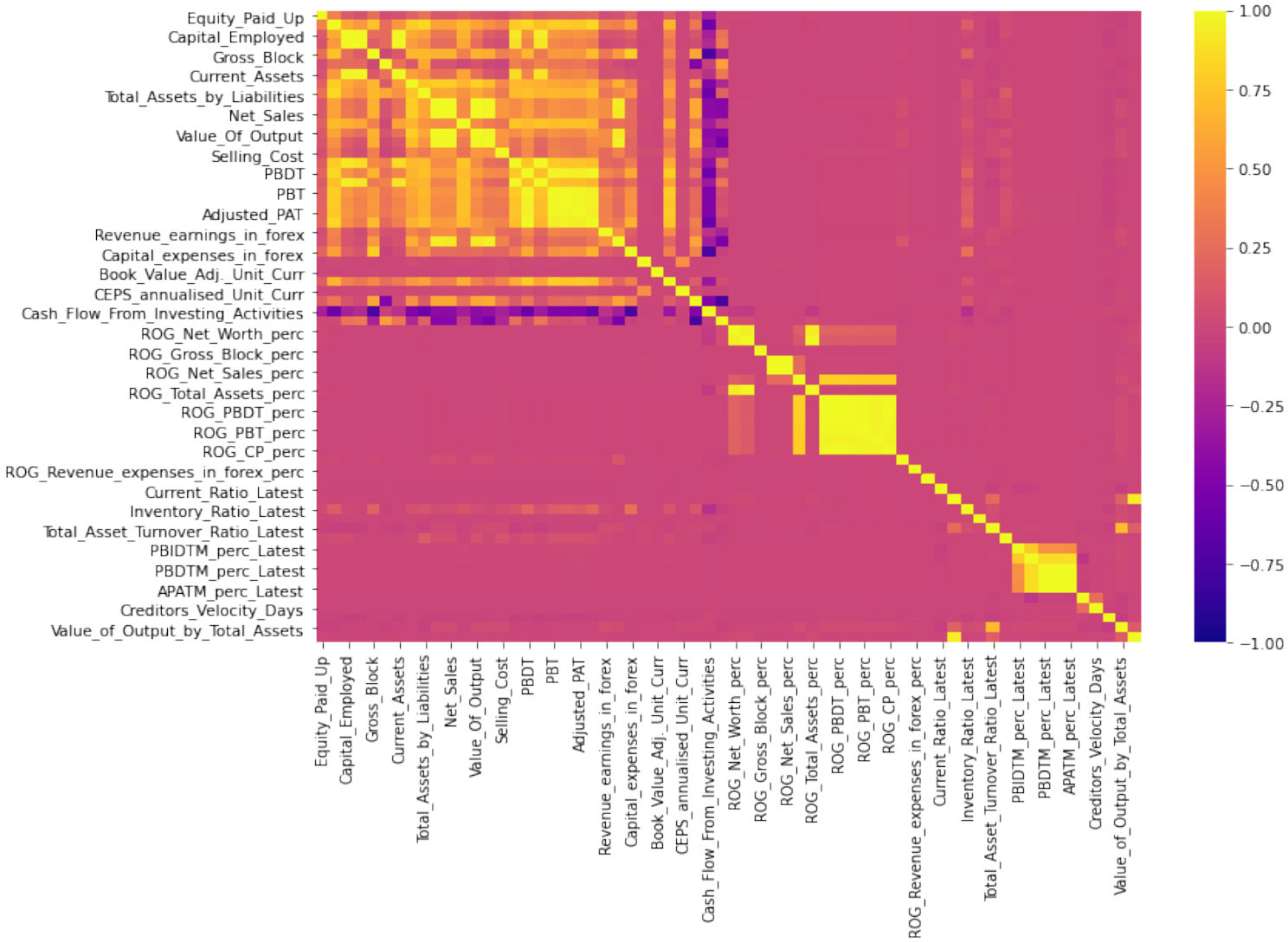
It is observed that there is positive correlation between variables CPM_perc_Latest, APATM_perc_Latest, PBIDTM_perc_Latest, PBITM_perc_Latest, PBDTM_perc_Latest.

Pairplot - 4

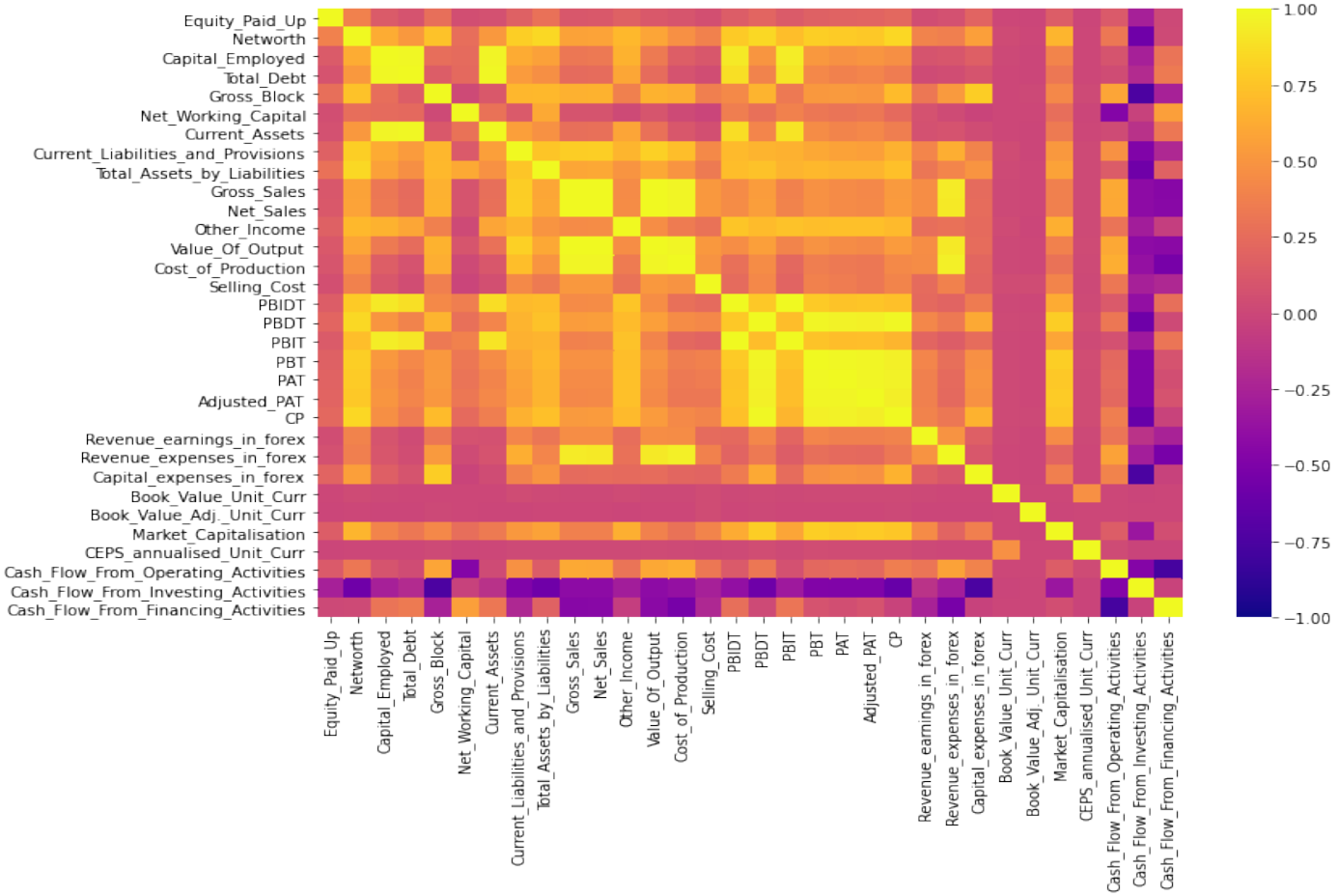


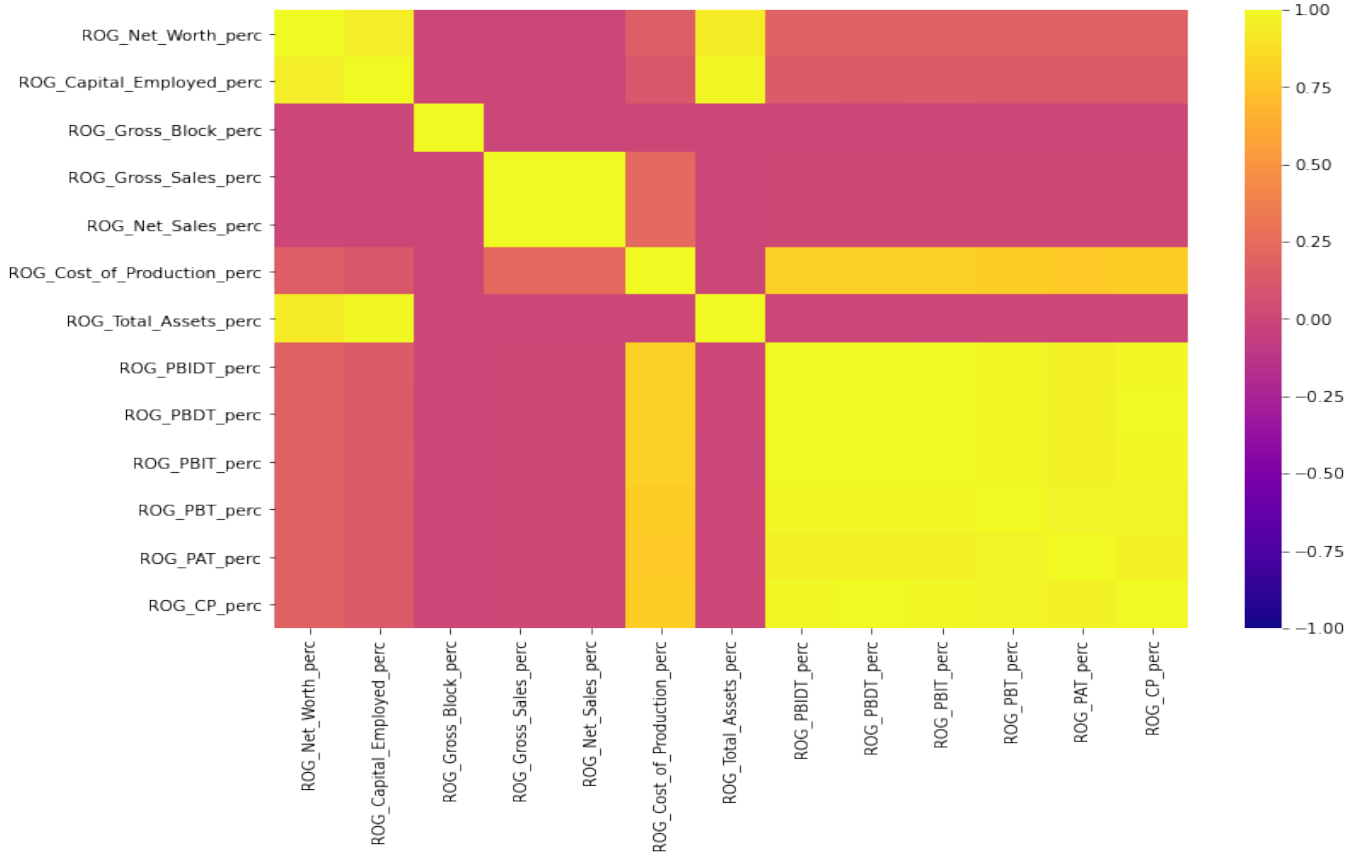
It is observed that there is positive correlation between variables Capital_Employed and Current_Assets, Current_Assets and Total_debt, Total_debt and Capital_Employed.

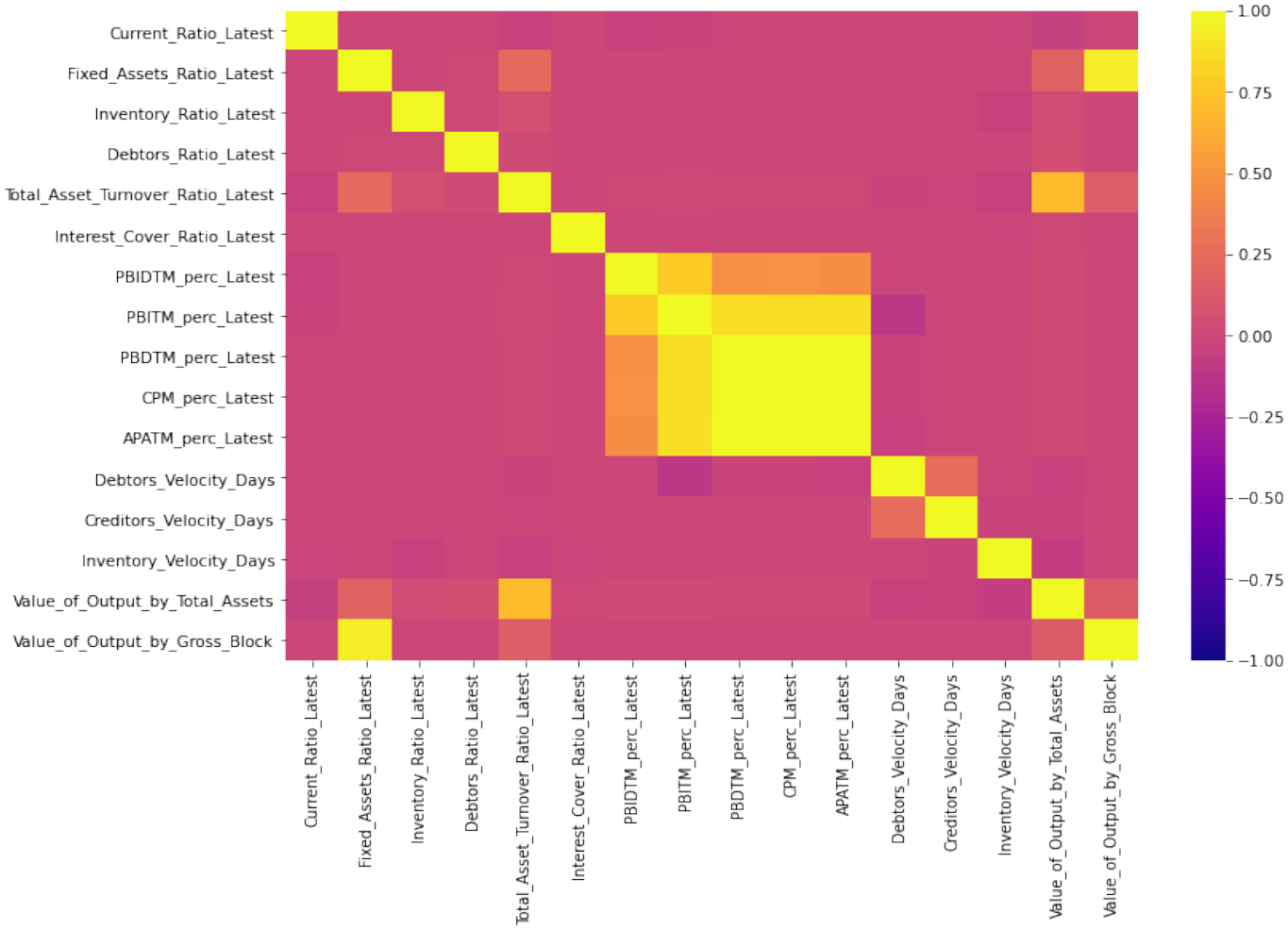
Heatmap of Correlation



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 pairplot. The above plot is disected into smaller plots for more clarity in the subsequent pages of this report.


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


ROG_Net_Worth_perc
ROG_Capital_Employed_perc
ROG_Gross_Block_perc
ROG_Gross_Sales_perc
ROG_Net_Sales_perc
ROG_Cost_of_Production_perc
ROG_Total_Assets_perc
ROG_PBIDT_perc
ROG_PBDT_perc
ROG_PBIT_perc
ROG_PBT_perc
ROG_PAT_perc
ROG_CP_perc
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



Many variables have correlation values close to 1 which denotes high collinearity among those variables.

Inferences from Univariate and Bi-variate analysis:

1. **Most of the variables have skewed distribution. But, we will not treat those distribution by any kind of transformation or new features.**
2. **All the variables have outliers. These ouliers will be treated as we are going to apply Logistic regression to predict the outcome.**
3. **Bi-variate analysis is performed on some of the important vaiables selected through VIF (discussed later in this report).**
4. **From pairplots, It is observed that there is high positive correlation between variables PBDT, PBIDT, PBIT, PBT, PAT, Adjusted PAT and CP, which is very obvious as these are the parameters to evaluate any corporates' performance.**
5. **It is observed that there is negative correlation between variables Cash_Flow_From_Operating_Activities and Cash_Flow_From_Financing_Activities.**
6. **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.

7. **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.**

Missing value treatment:

The missing values are treated with Simple Imputer Class. SimpleImputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. Here, median is used to fill up the missing value.
The following figure ensures that there is no missing values after treatment.

```
Book_Value_Adj._Unit_Curr            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
Inventory_Velocity_Days              0
```

Outlier treatment:

Ouliers are present in all of the independent variables. For our dataset, we used IQR (Inter-Quartile Range) based calculation to treat the outliers. The following is the method,
1. Arrange the data in ascending order
2. Calculate Q1 (the first Quarter)
3. Calculate Q3 (the third Quartile)
4. Find IQR = (Q3 - Q1)
5. Find the lower Range = Q1 - (1.5 * IQR)
6. Find the upper Range = Q3 + (1.5 * IQR)

Once the upperbound and lowerbound 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%.

The following figure shows the boxplot of variables after outlier treatment.


−200
0
200
400
600
800
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
ROG_Net_Worth_perc
ROG_Capital_Employed_perc
ROG_Gross_Block_perc
ROG_Gross_Sales_perc
ROG_Net_Sales_perc
ROG_Cost_of_Production_perc
ROG_Total_Assets_perc
ROG_PBIDT_perc
ROG_PBDT_perc
ROG_PBIT_perc
ROG_PBT_perc
ROG_PAT_perc
ROG_CP_perc
ROG_Revenue_earnings_in_forex_perc
ROG_Revenue_expenses_in_forex_perc
ROG_Market_Capitalisation_perc
Current_Ratio_Latest
Fixed_Assets_Ratio_Latest
Inventory_Ratio_Latest
Debtors_Ratio_Latest
Total_Asset_Turnover_Ratio_Latest
Interest_Cover_Ratio_Latest
PBIDTM_perc_Latest
PBITM_perc_Latest
PBDTM_perc_Latest
CPM_perc_Latest
APATM_perc_Latest
Debtors_Velocity_Days
Creditors_Velocity_Days
Inventory_Velocity_Days
Value_of_Output_by_Total_Assets
Value_of_Output_by_Gross_Block

Logistic Regression Model (using statsmodel library)

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: $z = \beta_0 + \sum_{i=1}^{n}(\beta_i X_1)$

In our present case, we will be using statsmodels modules for logistic regression as required by client.

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

1. **From sklearn.model_selection train_test_split for splitting the train and test set.**
2. **variance_inflation_factor module from statsmodels.stats.outliers_influence**
3. **metrics from sklearn**
4. **roc_auc_score,roc_curve,classification_report,confusion_matrix,plot_confusion_matrix from sklearn.metrics**

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

	variables	VIF
45	ROG_Revenue_earnings_in_forex_perc	1.124533
46	ROG_Revenue_expenses_in_forex_perc	1.200701
34	ROG_Gross_Block_perc	1.297988
48	Current_Ratio_Latest	1.367506
47	ROG_Market_Capitalisation_perc	1.517616
60	Creditors_Velocity_Days	1.554967
50	Inventory_Ratio_Latest	1.557235
61	Inventory_Velocity_Days	1.608086
51	Debtors_Ratio_Latest	1.655111
59	Debtors_Velocity_Days	1.665492
53	Interest_Cover_Ratio_Latest	1.674874
37	ROG_Cost_of_Production_perc	1.907790
32	ROG_Net_Worth_perc	2.164057
31	Cash_Flow_From_Financing_Activities	2.424072
22	Revenue_earnings_in_forex	2.434017

0	Equity_Paid_Up	2.547512
24	Capital_expenses_in_forex	2.652227
11	Other_Income	2.854276
23	Revenue_expenses_in_forex	2.869648
14	Selling_Cost	3.012508
30	Cash_Flow_From_Investing_Activities	3.018853
27	Market_Capitalisation	3.166142
38	ROG_Total_Assets_perc	3.241266
33	ROG_Capital_Employed_perc	3.465884
3	Total_Debt	4.005260
28	CEPS_annualised_Unit_Curr	4.052777
29	Cash_Flow_From_Operating_Activities	4.484060
5	Net_Working_Capital	4.491523

Train test split:

The original dataframe except the variables Co_Code, Co_Name, Networth_Next_Year is divided into dependent and independent variable type dataframe. Then both independent and dependent variable dataframe is splitted into 67:33 (train:test) ratio. One requirement for Statsmodel is that dependent and independent variables should be contained in same dataframe. 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
```

Model building:

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

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2373
Method:	MLE	**Df Model:**	28
Date:	Fri, 16 Jul 2021	**Pseudo R-squ.:**	0.3408
Time:	23:56:27	**Log-Likelihood:**	-542.84
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	1.507e-100

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.1878	0.146	-8.143	0.000	-1.474	-0.902
ROG_Revenue_earnings_in_forex_perc	-0.0010	0.003	-0.349	0.727	-0.006	0.005
ROG_Revenue_expenses_in_forex_perc	-0.0030	0.002	-1.924	0.054	-0.006	5.59e-05
ROG_Gross_Block_perc	-0.0107	0.005	-2.206	0.027	-0.020	-0.001
Current_Ratio_Latest	-0.0634	0.011	-5.753	0.000	-0.085	-0.042
ROG_Market_Capitalisation_perc	-0.0003	0.001	-0.227	0.821	-0.003	0.002
Creditors_Velocity_Days	0.0012	0.000	3.308	0.001	0.001	0.002
Inventory_Ratio_Latest	-0.0100	0.006	-1.731	0.084	-0.021	0.001
Inventory_Velocity_Days	-0.0014	0.001	-1.605	0.108	-0.003	0.000
Debtors_Velocity_Days	-0.0015	0.000	-4.162	0.000	-0.002	-0.001
Interest_Cover_Ratio_Latest	-0.0234	0.010	-2.437	0.015	-0.042	-0.005
Debtors_Ratio_Latest	-0.0173	0.008	-2.306	0.021	-0.032	-0.003
ROG_Cost_of_Production_perc	-0.0032	0.001	-2.739	0.006	-0.006	-0.001
ROG_Net_Worth_perc	-0.0167	0.004	-4.519	0.000	-0.024	-0.009
Cash_Flow_From_Financing_Activities	0.0030	0.003	1.078	0.281	-0.002	0.008
Revenue_earnings_in_forex	-0.0010	0.001	-1.012	0.312	-0.003	0.001
Equity_Paid_Up	0.0031	0.003	1.166	0.244	-0.002	0.008
Capital_expenses_in_forex	-0.0421	0.032	-1.310	0.190	-0.105	0.021
Selling_Cost	-0.0062	0.006	-1.072	0.284	-0.018	0.005
Cash_Flow_From_Investing_Activities	0.0047	0.003	1.476	0.140	-0.002	0.011
Other_Income	-0.0054	0.005	-1.118	0.263	-0.015	0.004
Revenue_expenses_in_forex	0.0013	0.001	1.183	0.237	-0.001	0.004
Market_Capitalisation	-0.0005	0.000	-3.416	0.001	-0.001	-0.000
ROG_Total_Assets_perc	-0.0119	0.006	-1.846	0.065	-0.024	0.001
ROG_Capital_Employed_perc	0.0010	0.005	0.179	0.858	-0.009	0.011
CEPS_annualised_Unit_Curr	-0.0852	0.022	-3.940	0.000	-0.128	-0.043
Total_Debt	0.0014	0.000	5.523	0.000	0.001	0.002
Net_Working_Capital	-0.0023	0.001	-4.146	0.000	-0.003	-0.001
Cash_Flow_From_Operating_Activities	0.0005	0.002	0.250	0.803	-0.004	0.005

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

Model 2 summary (new model):

Logit Regression Results

Dep. Variable:	Default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2389
Method:	MLE	**Df Model:**	12
Date:	Sat, 17 Jul 2021	**Pseudo R-squ.:**	0.3215
Time:	00:24:39	**Log-Likelihood:**	-558.73
converged:	True	**LL-Null:**	-823.47
Covariance Type:	nonrobust	**LLR p-value:**	1.174e-105

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3895	0.128	-10.813	0.000	-1.641	-1.138
ROG_Gross_Block_perc	-0.0139	0.005	-3.009	0.003	-0.023	-0.005
Current_Ratio_Latest	-0.0602	0.011	-5.471	0.000	-0.082	-0.039
Creditors_Velocity_Days	0.0014	0.000	3.844	0.000	0.001	0.002
Debtors_Velocity_Days	-0.0014	0.000	-3.964	0.000	-0.002	-0.001
Interest_Cover_Ratio_Latest	-0.0267	0.010	-2.713	0.007	-0.046	-0.007
Debtors_Ratio_Latest	-0.0231	0.007	-3.093	0.002	-0.038	-0.008
ROG_Cost_of_Production_perc	-0.0034	0.001	-2.896	0.004	-0.006	-0.001
ROG_Net_Worth_perc	-0.0189	0.003	-5.739	0.000	-0.025	-0.012
Market_Capitalisation	-0.0006	0.000	-4.339	0.000	-0.001	-0.000
CEPS_annualised_Unit_Curr	-0.0991	0.021	-4.761	0.000	-0.140	-0.058
Total_Debt	0.0010	0.000	5.238	0.000	0.001	0.001
Net_Working_Capital	-0.0024	0.000	-4.722	0.000	-0.003	-0.001

The new model (model 2) has all the variables with p < 0.05. This model will be considered for **Test set prediction and performance evaluation**.

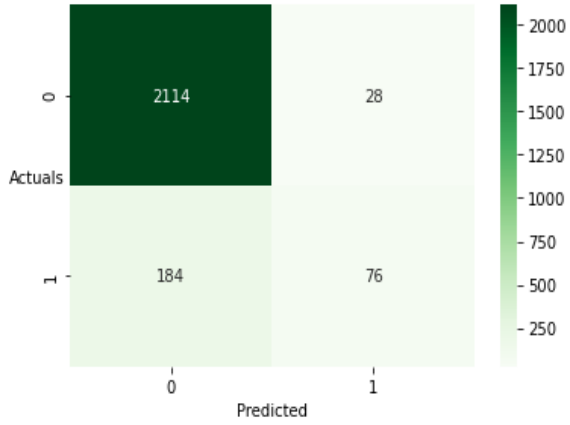
Model Evaluation on the Training Data

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

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

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

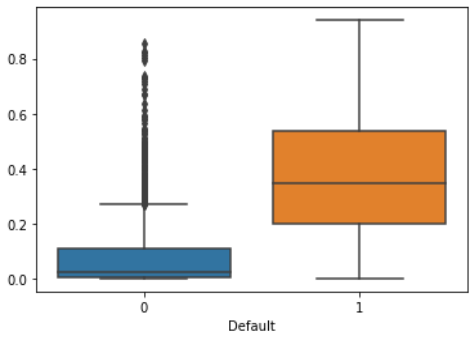
Performance of 0.5 probability cut-off:



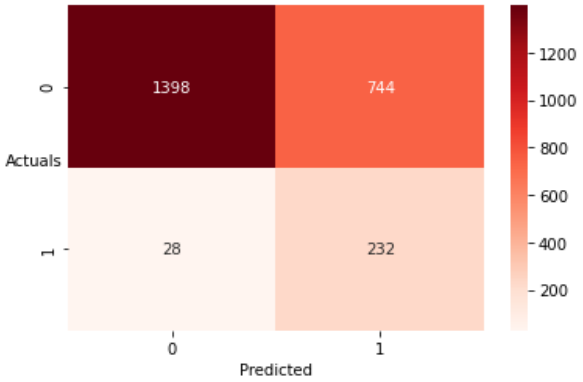
	precision	recall	f1-score	support
0	0.920	0.987	0.952	2142
1	0.731	0.292	0.418	260
accuracy			0.912	2402
macro avg	0.825	0.640	0.685	2402
weighted avg	0.899	0.912	0.894	2402

Overall 91% of correct predictions to total predictions were made by the model. 29% of those defaulted were correctly identified as defaulters by the model, which is not so good number.

So, we will change the probability cut-off to 0.07 as from the boxplot it is clear that "Default" status 0 has very low probablity median.



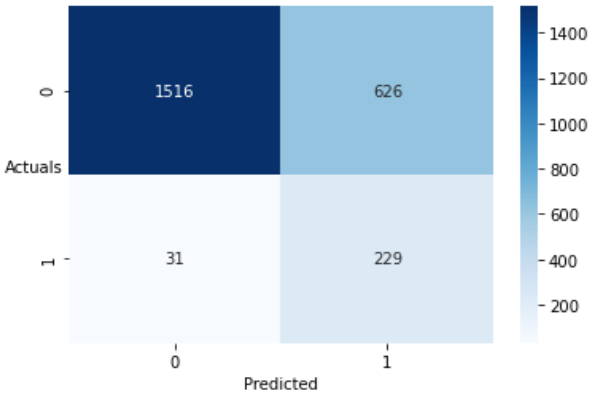
Performance of 0.07 probability cut-off:



	precision	recall	f1-score	support
0	0.980	0.653	0.784	2142
1	0.238	0.892	0.375	260
accuracy			0.679	2402
macro avg	0.609	0.772	0.580	2402
weighted avg	0.900	0.679	0.739	2402

Accuracy of the model i.e. %overall correct predictions has decreased from 91% to 68% but sensitivity of the model has increased from 29% to 89%, which is good for our prediction. But we will try with some more probability cut-off values.

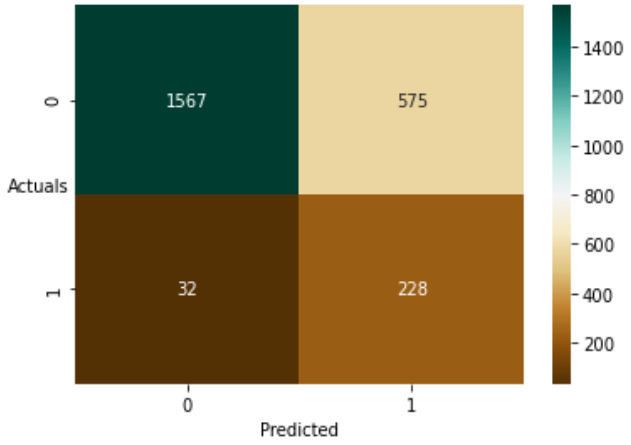
Performance of 0.09 probability cut-off:



	precision	recall	f1-score	support
0	0.980	0.708	0.822	2142
1	0.268	0.881	0.411	260
accuracy			0.726	2402
macro avg	0.624	0.794	0.616	2402
weighted avg	0.903	0.726	0.777	2402

Accuracy of the model i.e. %overall correct predictions has increased from 71% to 73% but sensitivity of the model has decresed from 89% to 88%.

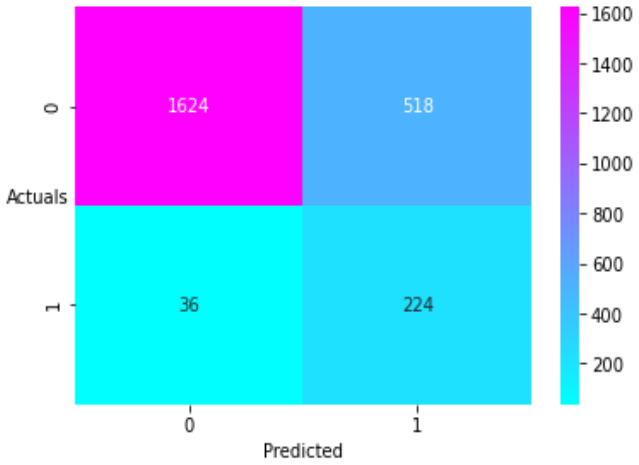
Performance of 0.1 probability cut-off:



	precision	recall	f1-score	support
0	0.980	0.732	0.838	2142
1	0.284	0.877	0.429	260
accuracy			0.747	2402
macro avg	0.632	0.804	0.633	2402
weighted avg	0.905	0.747	0.793	2402

Accuracy of the model i.e. %overall correct predictions has increased from 73% to 75% but sensitivity of the model has not decresed (88%). But we will try with some more probability cut-off values.

Performance of 0.115 probability cut-off:

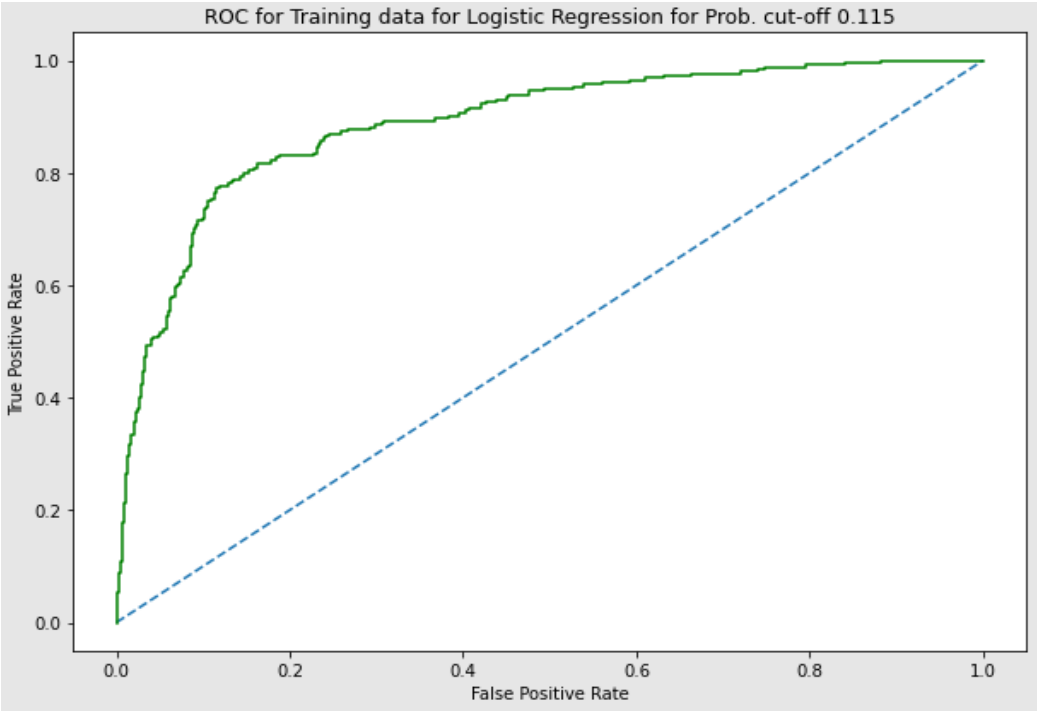


	precision	recall	f1-score	support
0	0.978	0.758	0.854	2142
1	0.302	0.862	0.447	260
accuracy			0.769	2402
macro avg	0.640	0.810	0.651	2402
weighted avg	0.905	0.769	0.810	2402

Accuracy of the model i.e. %overall correct predictions has increased from 75% to 77% but sensitivity of the model has decreased slightly from 88% to 86%.

We will keep this model (with p= 0.115 as cut-off) for further analysis as we are trying to maintain a balance between Accuracy and Recall.

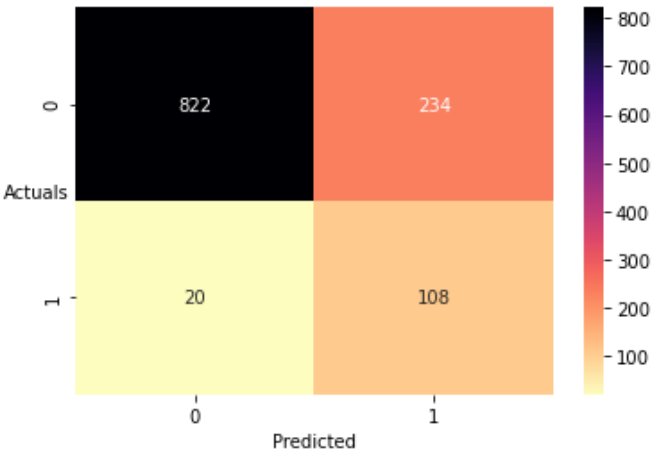
AUC-ROC Curve:



The model AUC for training set is 0.888.

Model Evaluation on the Testing Data

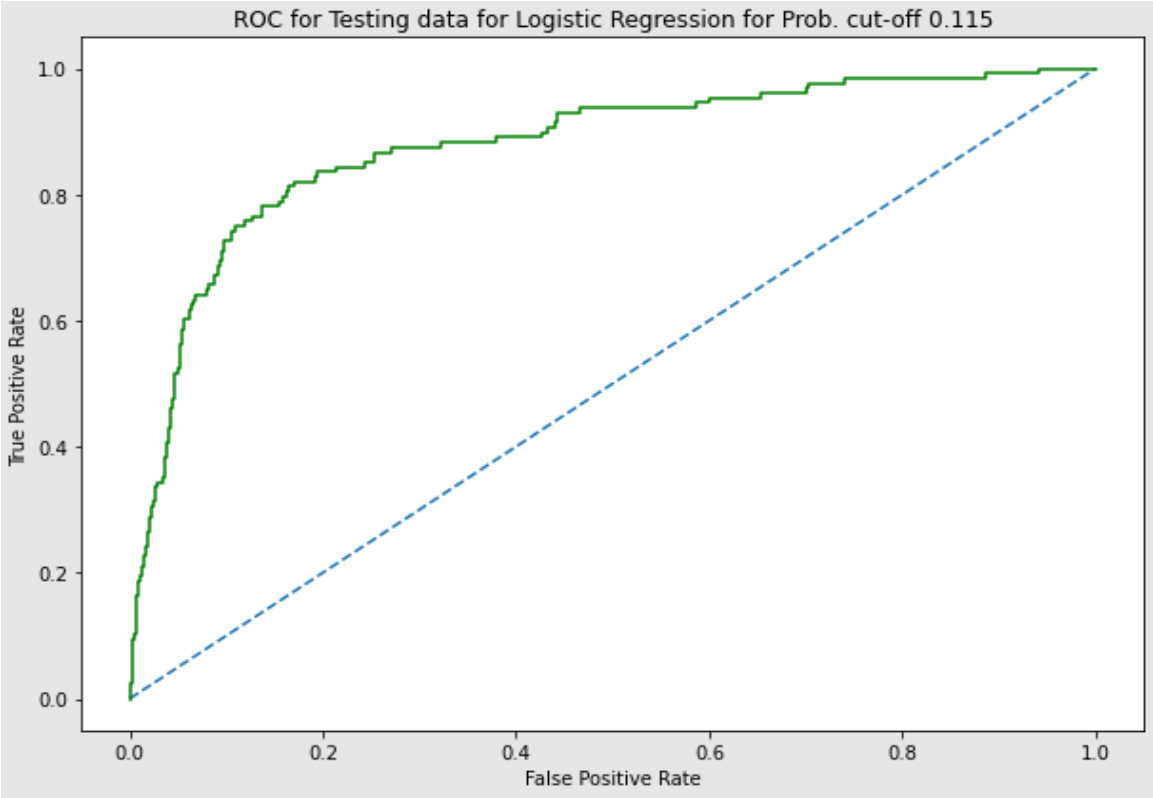
The model (with p= 0.115 as cut-off) is checked to predict on the test set. The confusion matrix and classification report is discussed below.



	precision	recall	f1-score	support
0	0.976	0.778	0.866	1056
1	0.316	0.844	0.460	128
accuracy			0.785	1184
macro avg	0.646	0.811	0.663	1184
weighted avg	0.905	0.785	0.822	1184

Accuracy of the model i.e. % overall correct prediction is 78% and sensitivity of the model is 84%. The model performs well on the test set also.

The model AUC for testing set is 0.877.



While the model results between training and test sets are similar, indicating no under or overfitting issues, overall prediction of the model is weak. There is a scope of improvement on the accuracy and recall values by using techniques like re-sampling, cross validation etc., which are not covered in the current report.