PGP - Data Science and Business Analytics

FINANCE AND RISK ANALYTICS PROJECT REPORT 1

Akshaya Nallathambi

13th February, 2022

Table Of Contents

Problem 1

List of Figures

List of Graphs

Problem

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.

Data Description-

Field Name	Description
Co_Code	Company Code
Co_Name	Company Name
Networth Next Year	Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities)
Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders
Networth	Value of a company as on 2015 - Current Year
Capital Employed	Total amount of capital used for the acquisition of profits by a company
Total Debt	The sum of money borrowed by the company and is due to be paid
Gross Block	Total value of all of the assets that a company owns
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).
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.
Current Liabilities and Provisions	Short-term financial obligations that are due within one year (includes amount that is set aside cover a future liability)
Total Assets/Liabilities	Ratio of total assets to liabailities of the company
Gross Sales	The grand total of sale transactions within the accounting period
Net Sales	Gross sales minus returns, allowances, and discounts
Other Income	Income realized from non-business activities (e.g. sale of long term asset)
Value Of Output	Product of physical output of goods and services produced by company and its market price
Cost of Production	Costs incurred by a business from manufacturing a product or providing a service
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)

PBIDT	Profit Before Interest, Depreciation & Taxes
PBDT	Profit Before Depreciation and Tax
PBIT	Profit before interest and taxes
PBT	Profit before tax
PAT	Profit After Tax
Adjusted PAT	Adjusted profit is the best estimate of the true profit
CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.
Revenue earnings in forex	Revenue earned in foreign currency
Revenue expenses in forex	Expenses due to foreign currency transactions
Capital expenses in forex	Long term investment in forex
Book Value (Unit Curr)	Net asset value
Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value
Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share
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
Cash Flow From Operating Activities	Use of cash from ongoing regular business activities
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 Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)
ROG-Net Worth (%)	Rate of Growth - Networth
ROG-Capital Employed (%)	Rate of Growth - Capital Employed
ROG-Gross Block (%)	Rate of Growth - Gross Block
ROG-Gross Sales (%)	Rate of Growth - Gross Sales
ROG-Net Sales (%)	Rate of Growth - Net Sales
ROG-Cost of Production (%)	Rate of Growth - Cost of Production
ROG-Total Assets (%)	Rate of Growth - Total Assets
ROG-PBIDT (%)	Rate of Growth- PBIDT

ROG-PBDT (%)	Rate of Growth- PBDT
ROG-PBIT (%)	Rate of Growth- PBIT
ROG-PBT (%)	Rate of Growth- PBT
ROG-PAT (%)	Rate of Growth- PAT
ROG-CP (%)	Rate of Growth- CP
ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex
ROG-Revenue expenses in forex (%)	Rate of Growth - Revenue expenses in forex
ROG-Market Capitalisation (%)	Rate of Growth - Market Capitalisation
Current Ratio[Latest]	Liquidity ratio, company's ability to pay short-term obligations or those due within one year
Fixed Assets Ratio[Latest]	Solvency ratio, the capacity of a company to discharge its obligations towards long-term lenders indicating
Inventory Ratio[Latest]	Activity ratio, specifies the number of times the stock or inventory has been replaced and sold by the company
Debtors Ratio[Latest]	Measures how quickly cash debtors are paying back to the company
Total Asset Turnover Ratio[Latest]	The value of a company's revenues relative to the value of its assets
Interest Cover Ratio[Latest]	Determines how easily a company can pay interest on its outstanding debt
PBIDTM (%)[Latest]	Profit before Interest Depreciation and Tax Margin
PBITM (%)[Latest]	Profit Before Interest Tax Margin
PBDTM (%)[Latest]	Profit Before Depreciation Tax Margin
CPM (%)[Latest]	Cost per thousand (advertising cost)
APATM (%)[Latest]	After tax profit margin
Debtors Velocity (Days)	Average days required for receiving the payments
Creditors Velocity (Days)	Average number of days company takes to pay suppliers
Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales
Value of Output/Total Assets	Ratio of Value of Output (market value) to Total Assets
Value of Output/Gross Block	Ratio of Value of Output (market value) to Gross Block

Sample of the dataset-

Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]	APATM (%) [Latest]
16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	...	0.00	0.00	0.00	0.00	0.00
21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	...	-10.30	-39.74	-57.74	-57.74	-87.18
14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	...	-5279.14	-5516.98	-7780.25	-7723.67	-7961.51
2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	...	-3.33	-7.21	-48.13	-47.70	-51.58
23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	...	-295.55	-400.55	-845.88	379.79	274.79

FIGURE 1

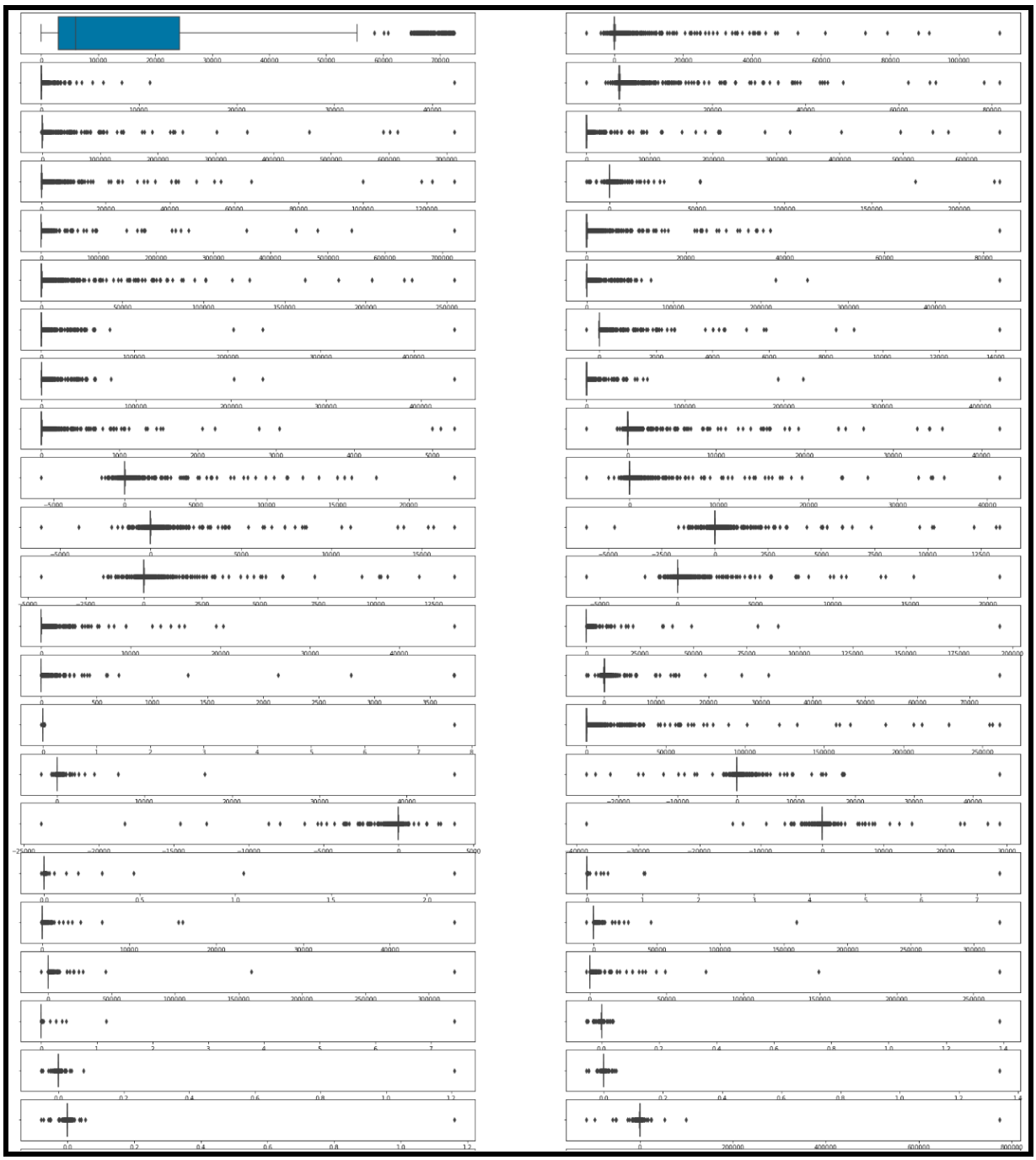
Debtors Velocity (Days)	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block
0	0	45.0	0.00	0.00
29	101	2.0	0.31	0.24
97	558	0.0	-0.03	-0.26
93	63	2.0	0.24	1.90
3887	346	0.0	0.01	0.05

FIGURE 2

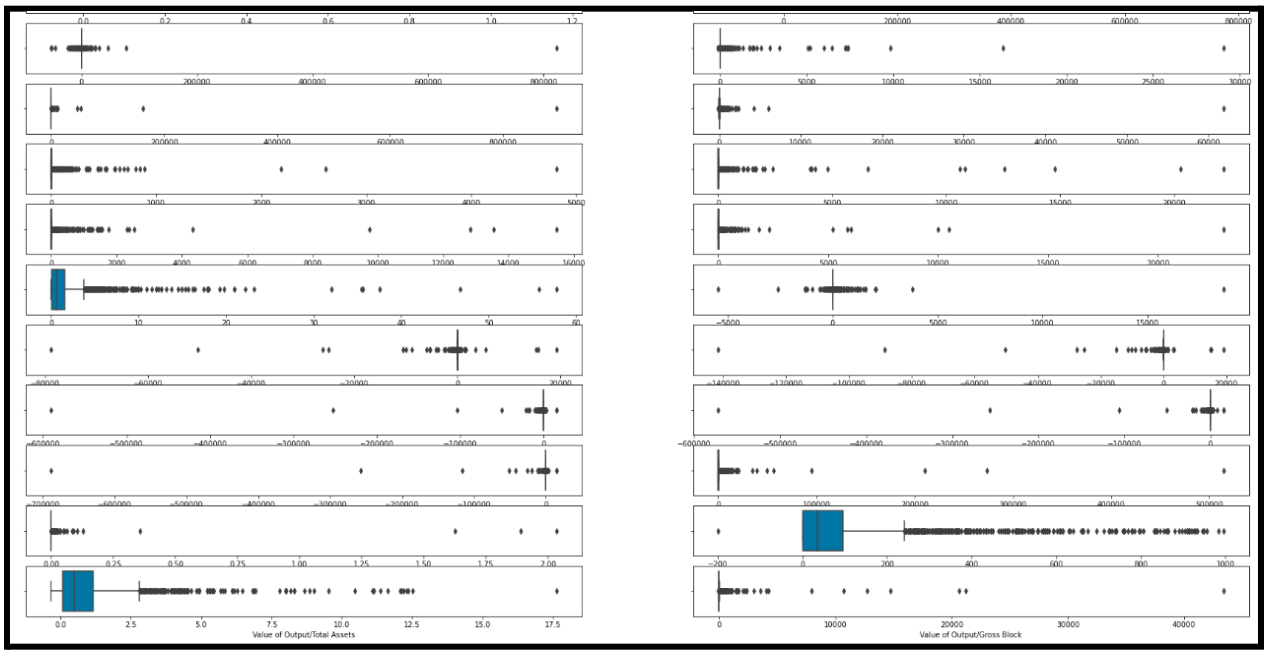
There are 67 variables, out of which 1 variable is categorical and 66 are continuous.. The data given is for 3586 individual companies. There are null values that require processing.

1.1 Outlier Treatment.

There are outliers in all the continuous variables. This is evident from the box plots below,



GRAPH 1



GRAPH 2

The outlier in the data set is treated using the IQR method.

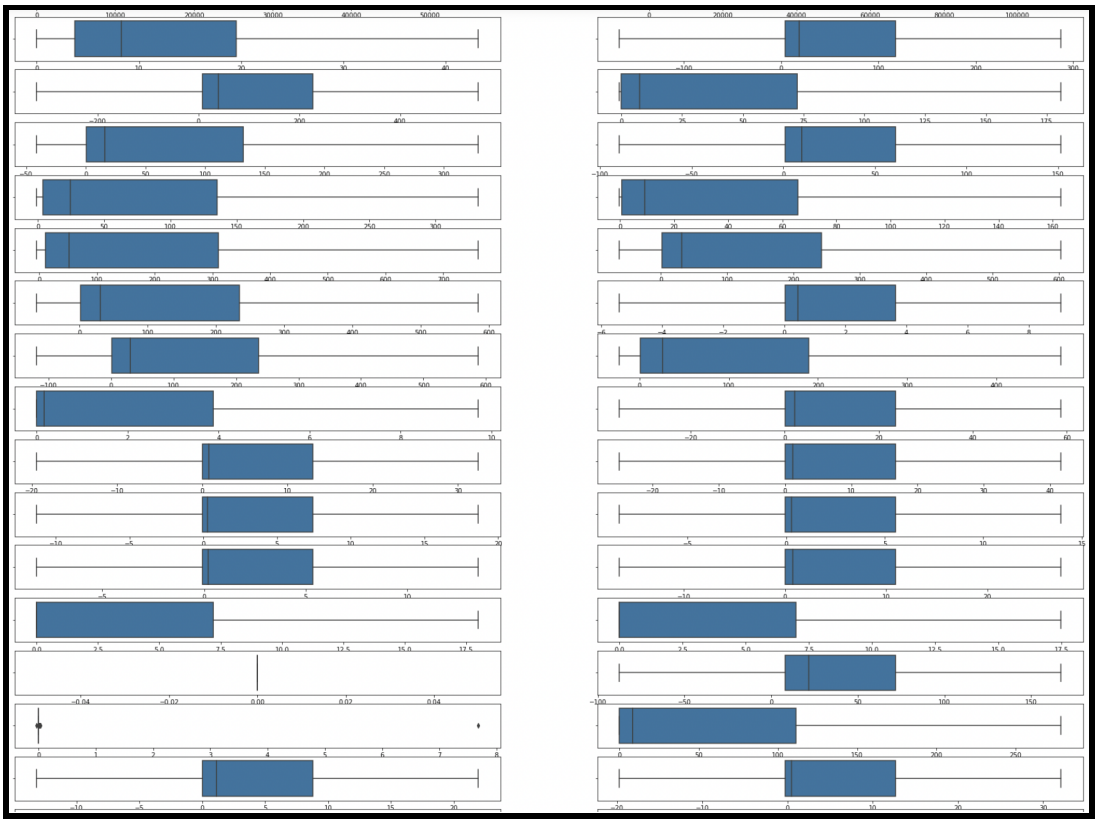
Inter quartile range (IQR) method -

Each dataset can be divided into quartiles. The first quartile point indicates that 25% of the data points are below that value whereas the second quartile is considered as the median point of the dataset. The inter quartile method finds the outliers on numerical datasets by following the procedure below,

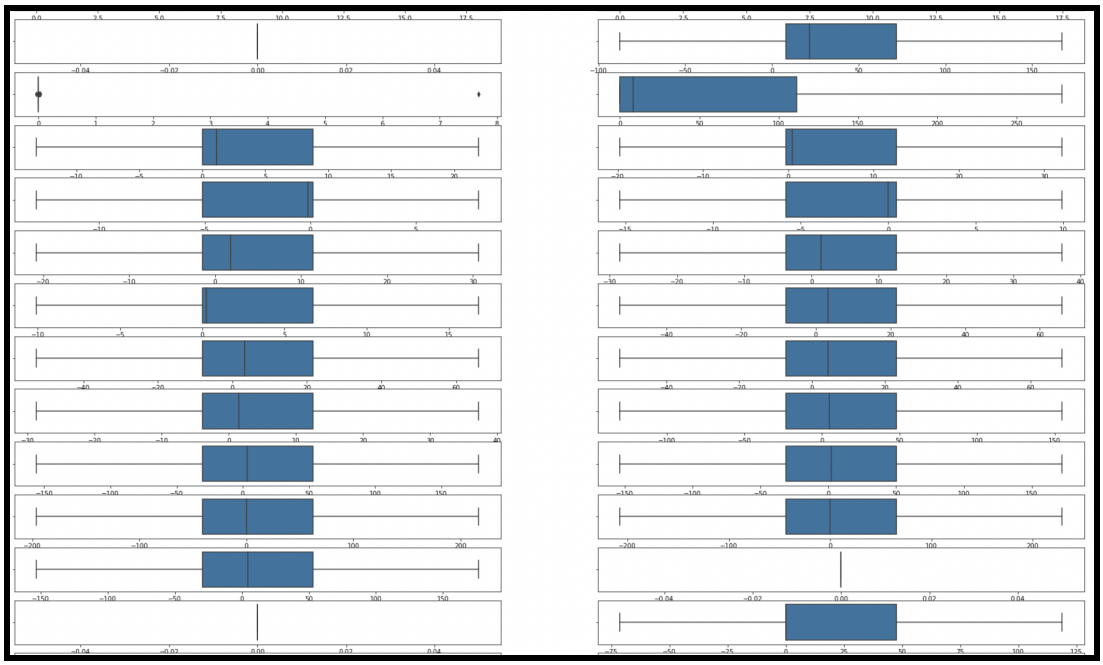
Find the first quartile, Q1.
Find the third quartile, Q3.
Calculate the IQR. IQR= Q3-Q1.
Define the normal data range with lower limit as Q1–1.5*IQR and upper limit as Q3+1.5*IQR.
Any data point outside this range is considered an outlier and should be removed for further analysis.

The concept of quartiles and IQR can best be visualized from the boxplot. It has the minimum and maximum point defined as Q1–1.5*IQR and Q3+1.5*IQR respectively. Any point outside this range is outlier. [1]

After outlier treatment-



GRAPH 3



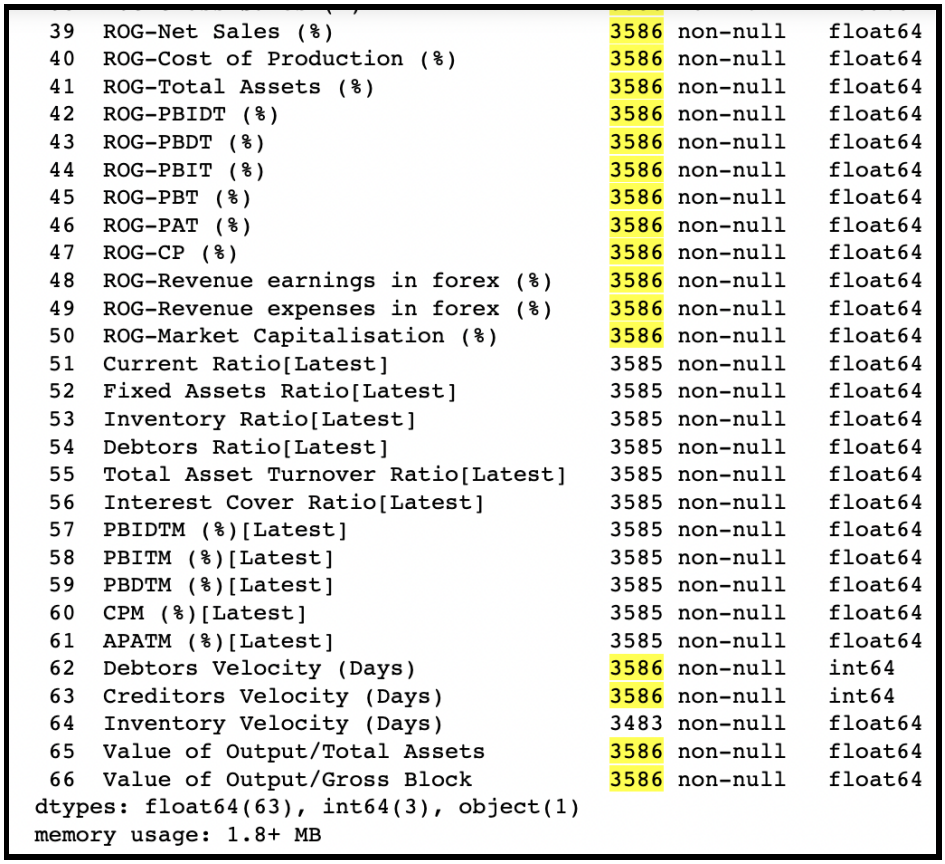
GRAPH 4

1.2 Missing Value Treatment.

There are missing values in the given data set. This is evident from the below image (unhighlighted variables contain null values).

```
<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/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 (%)                     3586 non-null   float64
 36  ROG-Capital Employed (%)              3586 non-null   float64
 37  ROG-Gross Block (%)                   3586 non-null   float64
 38  ROG-Gross Sales (%)                   3586 non-null   float64
```

FIGURE 3



```
 39  ROG-Net Sales (%)                        3586 non-null   float64
 40  ROG-Cost of Production (%)               3586 non-null   float64
 41  ROG-Total Assets (%)                     3586 non-null   float64
 42  ROG-PBIDT (%)                            3586 non-null   float64
 43  ROG-PBDT (%)                             3586 non-null   float64
 44  ROG-PBIT (%)                             3586 non-null   float64
 45  ROG-PBT (%)                              3586 non-null   float64
 46  ROG-PAT (%)                              3586 non-null   float64
 47  ROG-CP (%)                               3586 non-null   float64
 48  ROG-Revenue earnings in forex (%)        3586 non-null   float64
 49  ROG-Revenue expenses in forex (%)        3586 non-null   float64
 50  ROG-Market Capitalisation (%)            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 (%)[Latest]                       3585 non-null   float64
 58  PBITM (%)[Latest]                        3585 non-null   float64
 59  PBDTM (%)[Latest]                        3585 non-null   float64
 60  CPM (%)[Latest]                          3585 non-null   float64
 61  APATM (%)[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/Total Assets             3586 non-null   float64
 66  Value of Output/Gross Block              3586 non-null   float64
dtypes: float64(63), int64(3), object(1)
memory usage: 1.8+ MB
```

FIGURE 4

The missing values are treated using the median value of the data set.

The median imputation is a technique in which the missing values are replaced with the median value of the entire feature column. When the data is skewed, it is good to consider using the median value for replacing the missing values. Note that imputing missing data with median value can only be done with numerical data. [2]

1.3 Transform Target variable into 0 and 1.

The target variable is defined as default. As mentioned in the problem statement I have created this variable based on the condition to create a default variable that should take the value of 1 when net worth next year is negative and 0 when net worth next year is positive.

Co_Name	Networth Next Year	Default
Hind.Cables	-8021.6	1
Tata Tele. Mah.	-3986.19	1
ABG Shipyard	-3192.58	1
GTL	-3054.51	1
Bharati Defence	-2967.36	1
Usha Ispat	-2519.4	1
Hanung Toys	-2125.05	1
K S Oils	-2100.56	1
Quadrant Tele.	-1695.75	1

FIGURE 5

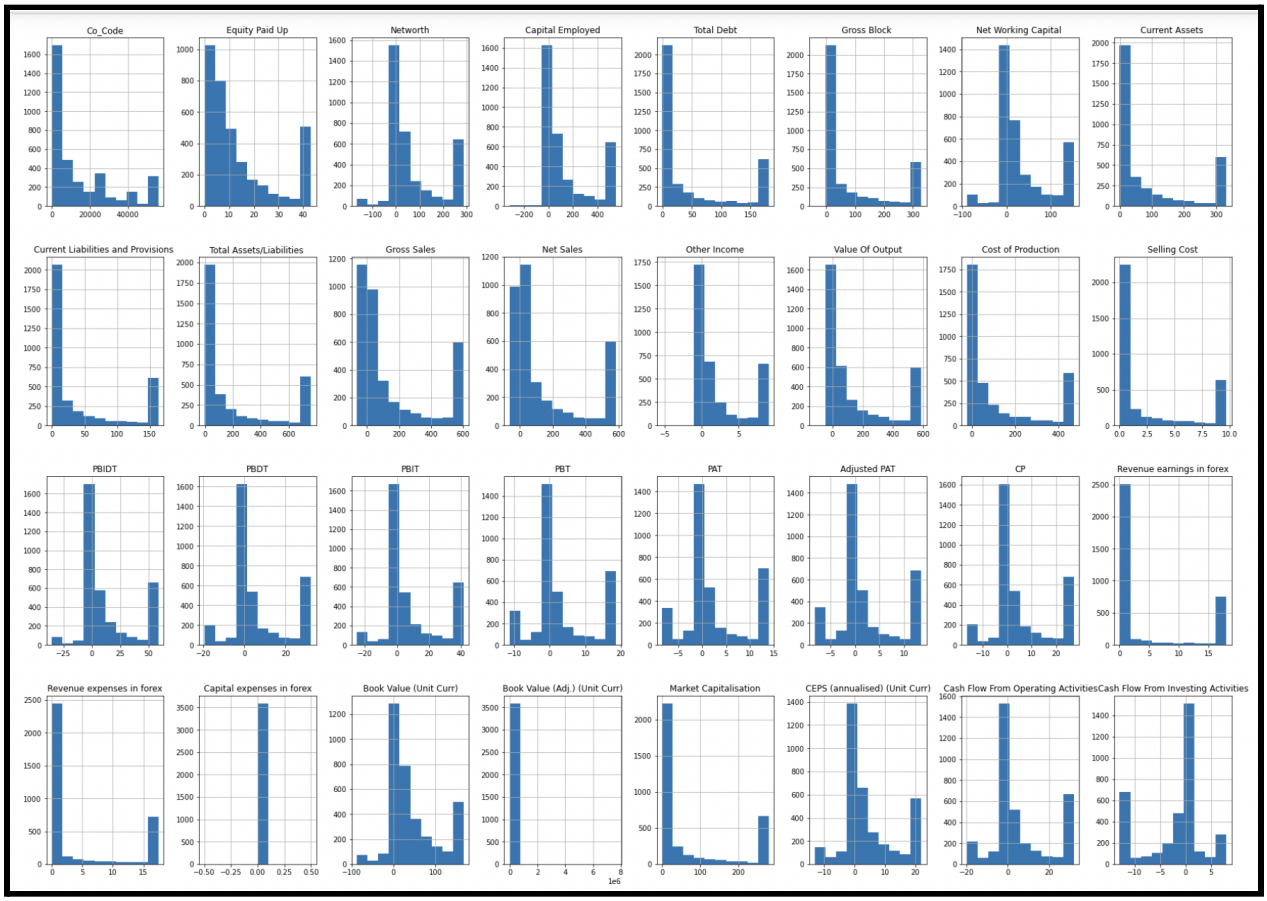
1.4 Univariate & Bivariate analysis with proper interpretation.

Univariate analysis:

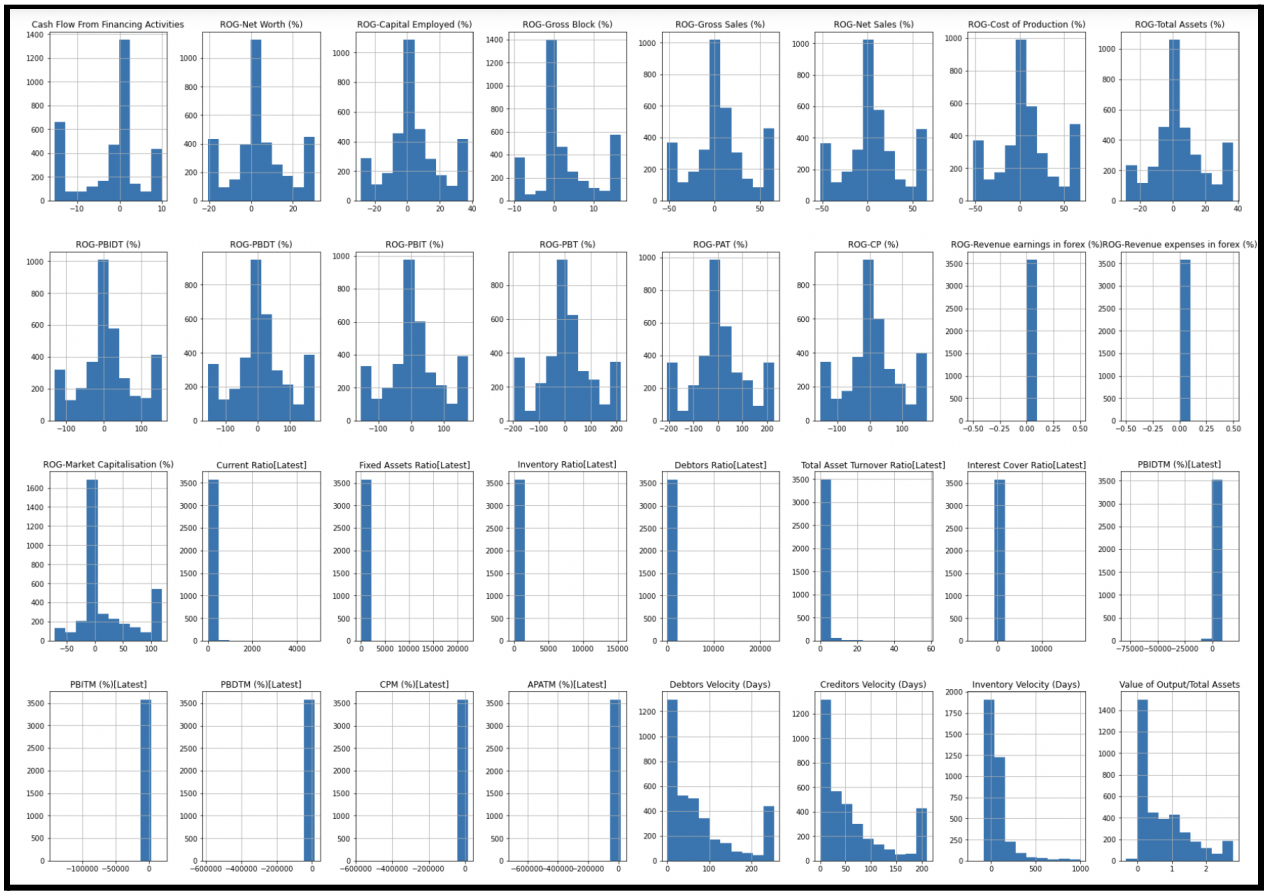
Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. [3]

The histograms are used for numerical variables to perform univariate analysis.

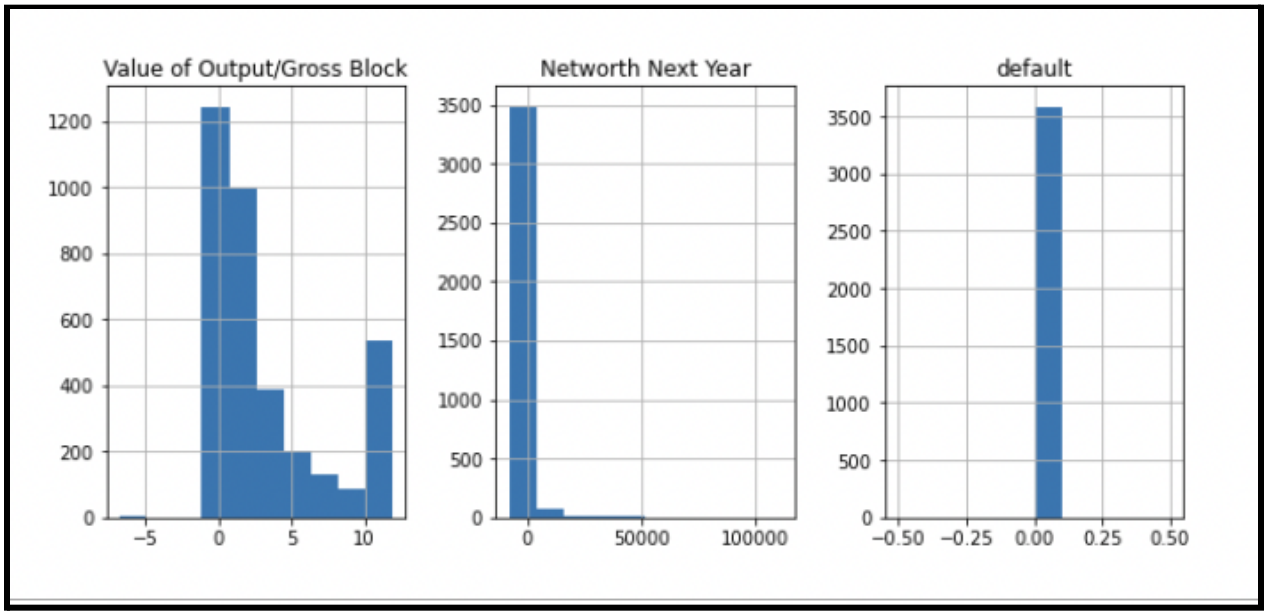
It is clear from the graph (Graph) that most of the numerical variables are rightly skewed.



GRAPH 5



GRAPH 6



GRAPH 7

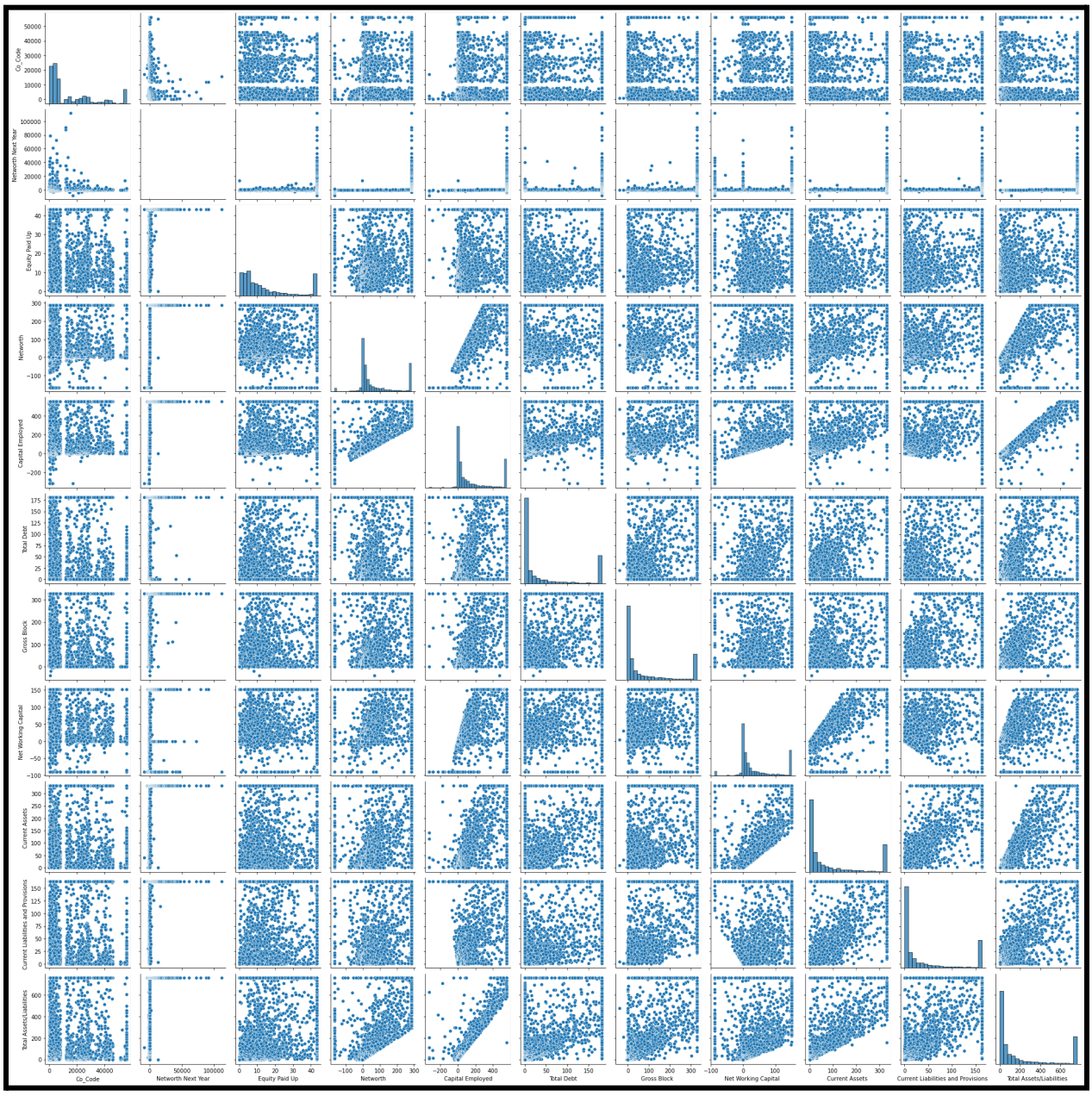
Bivariate analysis:

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. Bivariate analysis can be helpful in testing simple hypotheses of association. [4]

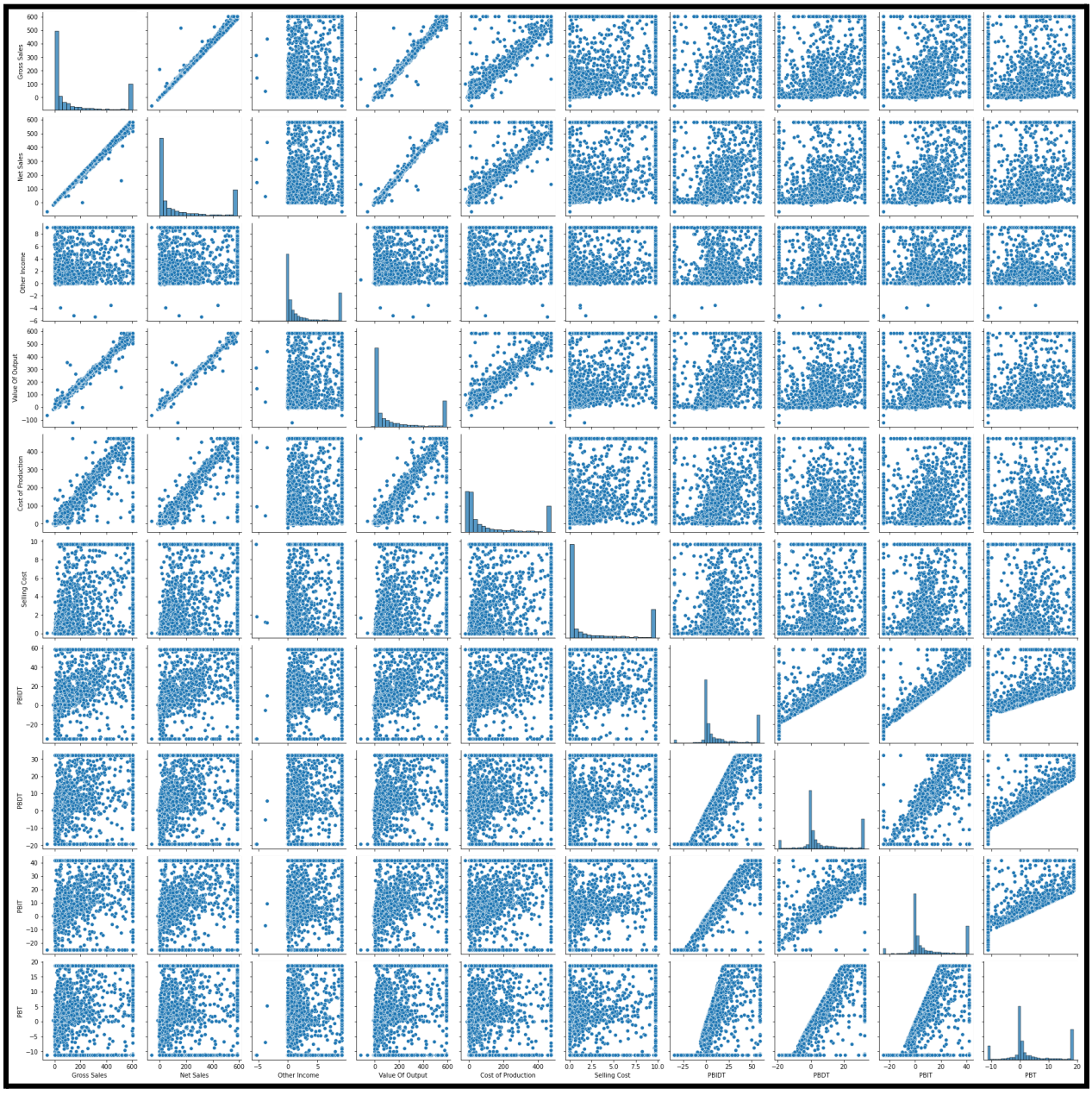
The pairplot is generally used for numerical variables to perform bivariate analysis.

A pairplot plots a pairwise relationship in a dataset. The pairplot function creates a grid of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. [5]

As there are 66 numerical columns, I have generated the pair plots for a few of them.


Co_Code
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

GRAPH 8



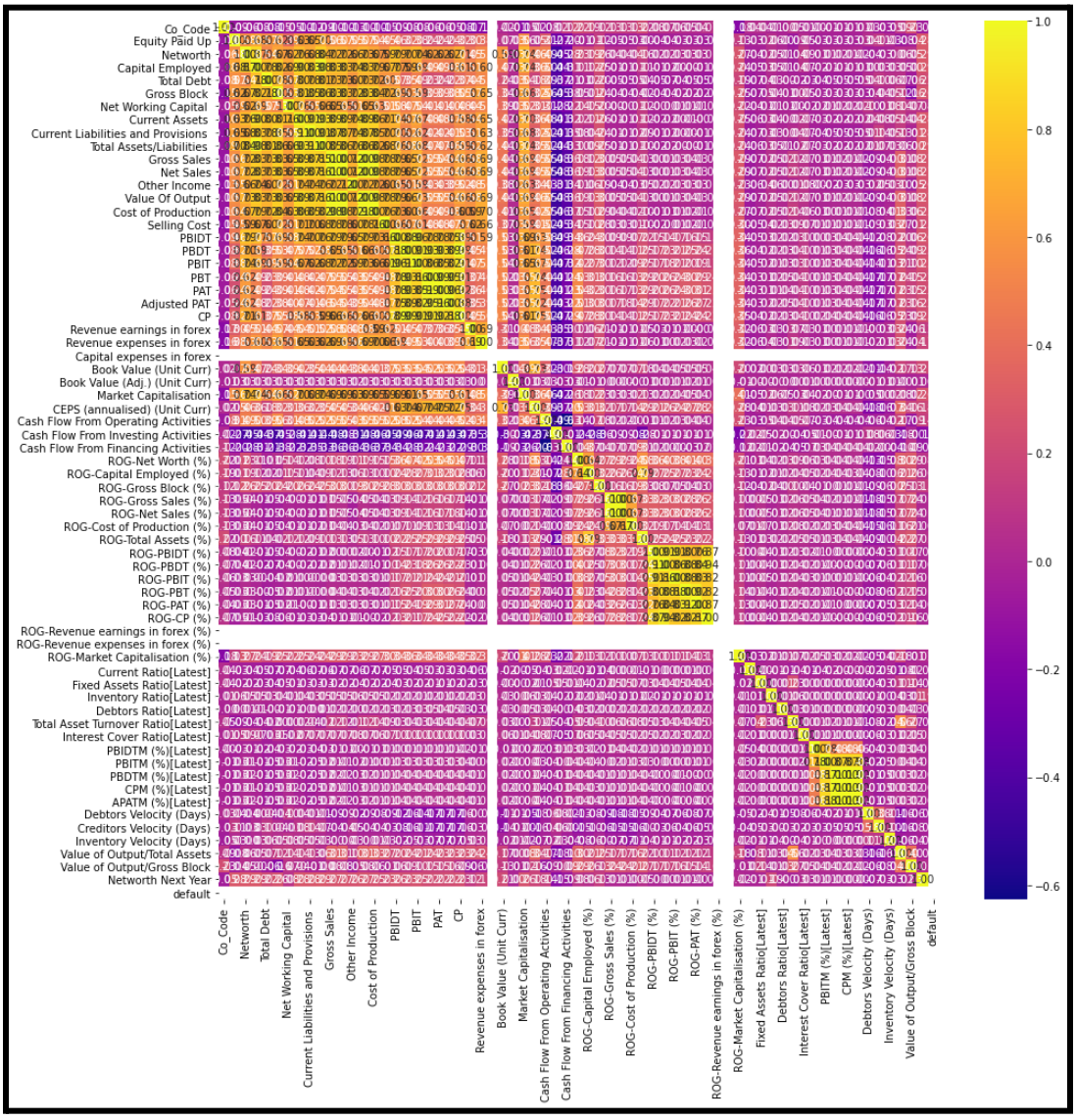
GRAPH 9

The below are the findings from the pairplot generated -

- The variable Networth is highly correlated with the variables Capital Employed, and Total Assets/Liabilities.
- The variable Capital Employed is highly correlated with the variable Total Assets/Liabilities.

- The variable Net Working Capital is highly correlated with the variables Capital Employed, Current Assets and Total Assets/Liabilities.
- The variable Current Assets is highly correlated with the variable Total Assets/Liabilities.
- The variable Gross Sales is highly correlated with the variables Net Sales, Value Of Output and Cost of Production.
- The variable Net Sales is highly correlated with the variables Value Of Output and Cost of Production.
- The variable Value Of Output is highly correlated with the variable Cost of Production.
- The variable PBIDT is highly correlated with the variables PBDT, PBIT and PBT.
- The variable PBDT is highly correlated with the variables PBIT and PBT.
- The variable PBIT is highly correlated with the variable PBT.

The heat map can also be used to check the association between two variables. All the boxes with a value higher than 0.8 are highly correlated. The heat map for all the numerical variable is below,



GRAPH 10

1.5 Train Test Split

I have split the data into Train and Test dataset in a ratio of 67:33 and used random_state =42. The same number of columns are retained after the split.

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

I have built a model using Logistic Regression for 'Probability at default'. The equation of the Logistic Regression by which we predict the corresponding probabilities and then go on predict a discrete target variable is,

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

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

The data is split into random train and test subsets. Models will be fitted on the train set and predictions will be made on the test set.

MODEL 1:

Before starting the model building, I checked the problem of multicollinearity. Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.

	variables	VIF
46	ROG-Revenue_earnings_in_forex_perc	1.134669
47	ROG-Revenue_expenses_in_forex_perc	1.167677
35	ROG-Gross_Block_perc	1.354494
49	Current_Ratio_Latest	1.407926
48	ROG-Market_Capitalisation_perc	1.557739
61	Creditors_Velocity_Days	1.560351
51	Inventory_Ratio_Latest	1.586679
62	Inventory_Velocity_Days	1.621947
52	Debtors_Ratio_Latest	1.658731
60	Debtors_Velocity_Days	1.669786
54	Interest_Cover_Ratio_Latest	1.697136
0	Co_Code	1.901165
38	ROG-Cost_of_Production_perc	1.912137
33	ROG-Net_Worth_perc	2.262686
32	Cash_Flow_From_Financing_Activities	2.365146
23	Revenue_earnings_in_forex	2.446955
25	Capital_expenses_in_forex	2.645710
1	Equity_Paid_Up	2.718180
15	Selling_Cost	2.889210
12	Other_Income	2.909667

FIGURE 6

Here, we see that the value of VIF is high for many variables. Here, I have dropped variables with VIF more than 5 (very high correlation) & build our model.

Model 1 results

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	3586
Model:	Logit	**Df Residuals:**	3564
Method:	MLE	**Df Model:**	21
Date:	Sun, 13 Feb 2022	**Pseudo R-squ.:**	0.3178
Time:	13:30:49	**Log-Likelihood:**	-837.01
converged:	True	**LL-Null:**	-1226.9
Covariance Type:	nonrobust	**LLR p-value:**	1.708e-151

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.7475	0.121	-6.191	0.000	-0.984	-0.511
Current_Ratio_Latest	-0.0716	0.010	-6.905	0.000	-0.092	-0.051
Creditors_Velocity_Days	0.0011	0.000	3.986	0.000	0.001	0.002
Inventory_Ratio_Latest	-0.0094	0.004	-2.099	0.036	-0.018	-0.001
Inventory_Velocity_Days	-0.0013	0.001	-1.888	0.059	-0.003	5.03e-05
Debtors_Ratio_Latest	-0.0138	0.006	-2.365	0.018	-0.025	-0.002
Debtors_Velocity_Days	-0.0010	0.000	-3.554	0.000	-0.002	-0.000
Interest_Cover_Ratio_Latest	-0.0254	0.007	-3.634	0.000	-0.039	-0.012
Co_Code	-3.908e-05	5.48e-06	-7.125	0.000	-4.98e-05	-2.83e-05
Cash_Flow_From_Financing_Activities	0.0005	0.002	0.268	0.789	-0.003	0.004
Revenue_earnings_in_forex	-0.0012	0.001	-1.615	0.106	-0.003	0.000
Capital_expenses_in_forex	-0.0213	0.024	-0.897	0.370	-0.068	0.025
Equity_Paid_Up	0.0040	0.002	1.921	0.055	-8.07e-05	0.008
Selling_Cost	-0.0012	0.004	-0.281	0.778	-0.010	0.007
Other_Income	0.0017	0.003	0.485	0.627	-0.005	0.009
Revenue_expenses_in_forex	0.0007	0.001	0.729	0.466	-0.001	0.002
Cash_Flow_From_Investing_Activities	0.0052	0.002	2.245	0.025	0.001	0.010
Market_Capitalisation	-0.0005	0.000	-5.237	0.000	-0.001	-0.000
CEPS_annualised_Unit_Curr	-0.1524	0.017	-9.195	0.000	-0.185	-0.120
Total_Debt	0.0012	0.000	6.099	0.000	0.001	0.002
Net_Working_Capital	-0.0022	0.000	-5.203	0.000	-0.003	-0.001
Cash_Flow_From_Operating_Activities	-0.0005	0.002	-0.277	0.782	-0.004	0.003

Possibly complete quasi-separation: A fraction 0.14 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

FIGURE 7

Most of the ratio variables are insignificant.

MODEL 2

Logit Regression Results

Dep. Variable:	default	**No. Observations:**	2402
Model:	Logit	**Df Residuals:**	2392
Method:	MLE	**Df Model:**	9
Date:	Sun, 13 Feb 2022	**Pseudo R-squ.:**	0.1655
Time:	13:36:32	**Log-Likelihood:**	-685.39
converged:	True	**LL-Null:**	-821.36
Covariance Type:	nonrobust	**LLR p-value:**	2.295e-53

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.0909	0.132	-8.245	0.000	-1.350	-0.832
Current_Ratio_Latest	-0.0617	0.011	-5.755	0.000	-0.083	-0.041
Creditors_Velocity_Days	0.0017	0.000	5.083	0.000	0.001	0.002
Inventory_Ratio_Latest	-0.0099	0.005	-1.977	0.048	-0.020	-8.38e-05
Inventory_Velocity_Days	-0.0013	0.001	-1.556	0.120	-0.003	0.000
Debtors_Ratio_Latest	-0.0119	0.007	-1.812	0.070	-0.025	0.001
Debtors_Velocity_Days	-0.0006	0.000	-1.653	0.098	-0.001	0.000
Interest_Cover_Ratio_Latest	-0.0830	0.018	-4.542	0.000	-0.119	-0.047
Co_Code	-3.366e-05	5.73e-06	-5.873	0.000	-4.49e-05	-2.24e-05
Cash_Flow_From_Financing_Activities	0.0022	0.001	1.554	0.120	-0.001	0.005

FIGURE 8

	variables	VIF
8	Cash_Flow_From_Financing_Activities	1.124181
0	Current_Ratio_Latest	1.193746
6	Interest_Cover_Ratio_Latest	1.235255
2	Inventory_Ratio_Latest	1.353612
3	Inventory_Velocity_Days	1.363517
4	Debtors_Ratio_Latest	1.401455
7	Co_Code	1.413544
1	Creditors_Velocity_Days	1.459169
5	Debtors_Velocity_Days	1.541698

FIGURE 9

The variables are now significant. The cut off is set at 0.08.

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

The performance matrices of the train data-

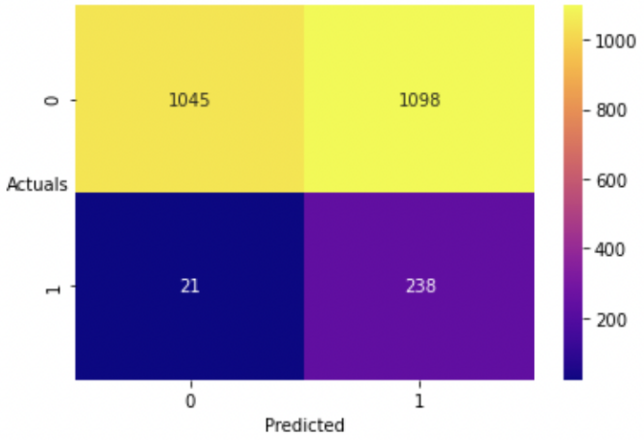


FIGURE 10

```
True Negative: 1045
False Positives: 1098
False Negatives: 21
True Positives: 238
```

FIGURE 11

	precision	recall	f1-score	support
0.0	0.980	0.488	0.651	2143
1.0	0.178	0.919	0.298	259
accuracy			0.534	2402
macro avg	0.579	0.703	0.475	2402
weighted avg	0.894	0.534	0.613	2402

FIGURE 12

The train data has an accuracy of 53.4%.

The performance matrices of the test data-

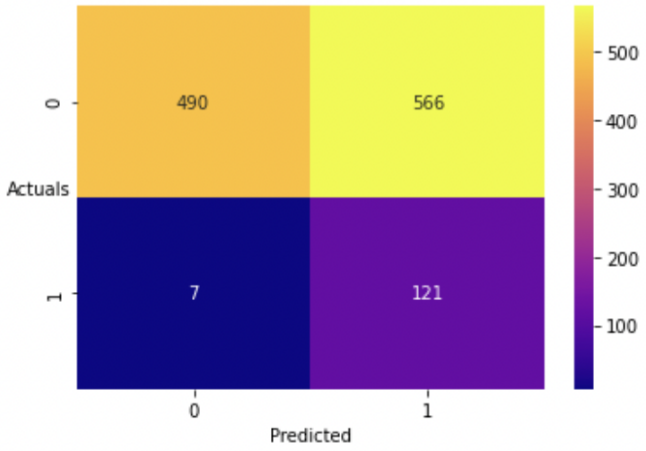


FIGURE 13

```
True Negative: 490
False Positives: 566
False Negatives: 7
True Positives: 121
```

FIGURE 14

	precision	recall	f1-score	support
0.0	0.986	0.464	0.631	1056
1.0	0.176	0.945	0.297	128
accuracy			0.516	1184
macro avg	0.581	0.705	0.464	1184
weighted avg	0.898	0.516	0.595	1184

FIGURE 15

The test data has an accuracy of 51.6%.

Websites-

[1] https://www.geeksforgeeks.org/interquartile-range-to-detect-outliers-in-data/

[2] https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

[3] https://www.statisticshowto.com/univariate/

[4] https://en.wikipedia.org/wiki/Bivariate_analysis

[5] https://pythonbasics.org/seaborn-pairplot/

End of Project