POST GRADUATE PROGRAM

DATA SCIENCE & BUSINESS ANALYTICS

BATCH II

AUGUST '21

Finance & Risk Analytics

Milestone 1

ZIANNA LYNGDOH

TABLE OF CONTENTS

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 Dictionary:

#	Field Name	Description	New Field Name
1	Co_Code	Company Code	Co_Code
2	Co_Name	Company Name	Co_Name
3	Networth Next Year	Value of a company as on 2016 - Next Year(difference between the value of total assets and total liabilities)	Networth_Next_Year
4	Equity Paid Up	Amount that has been received by the company through the issue of shares to the shareholders	Equity_Paid_Up
5	Networth	Value of a company as on 2015 - Current Year	Networth
6	Capital Employed	Total amount of capital used for the acquisition of profits by a company	Capital_Employed
7	Total Debt	The sum of money borrowed by the company and is due to be paid	Total_Debt
8	Gross Block	Total value of all of the assets that a company owns	Gross_Block
9	Net Working Capital	The difference between a company's current assets (cash, accounts receivable, inventories of raw materials and finished goods) and its current liabilities (accounts payable).	Net_Working_Capital
10	Current Assets	All the assets of a company that are expected to be sold or used as	Curr_Assets

		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)	Curr_Liab_and_Prov
12	Total Assets/Liabilities	Ratio of total assets to liabailities of the company	Total_Assets_to_Liab
13	Gross Sales	The grand total of sale transactions within the accounting period	Gross_Sales
14	Net Sales	Gross sales minus returns, allowances, and discounts	Net_Sales
15	Other Income	Income realized from non-business activities (e.g. sale of long term asset)	Other_Income
16	Value Of Output	Product of physical output of goods and services produced by company and its market price	Value_Of_Output
17	Cost of Production	Costs incurred by a business from manufacturing a product or providing a service	Cost_of_Prod
18	Selling Cost	Costs which are made to create the demand for the product (advertising expenditures, packaging and styling, salaries, commissions and travelling expenses of sales personnel, and the cost of shops and showrooms)	Selling_Cost
19	PBIDT	Profit Before Interest, Depreciation & Taxes	PBIDT
20	PBDT	Profit Before Depreciation and Tax	PBDT
21	PBIT	Profit before interest and taxes	PBIT
22	PBT	Profit before tax	PBT
23	PAT	Profit After Tax	PAT
24	Adjusted PAT	Adjusted profit is the best estimate of the true profit	Adjusted_PAT
26	CP	Commercial paper , a short-term debt instrument to meet short-term liabilities.	CP
27	Revenue earnings in forex	Revenue earned in foreign currency	Rev_earn_in_forex
28	Revenue expenses in forex	Expenses due to foreign currency transactions	Rev_exp_in_forex
29	Capital expenses in forex	Long term investment in forex	Capital_exp_in_forex

30	Book Value (Unit Curr)	Net asset value	Book_Value_Unit_Curr
31	Book Value (Adj.) (Unit Curr)	Book value adjusted to reflect asset's true fair market value	Book_Value_Adj_Unit_Curr
32	Market Capitalisation	Product of the total number of a company's outstanding shares and the current market price of one share	Market_Capitalisation
33	CEPS (annualised) (Unit Curr)	Cash Earnings per Share, profitability ratio that measures the financial performance of a company by calculating cash flows on a per share basis	CEPS_annualised_Unit_Curr
34	Cash Flow From Operating Activities	Use of cash from ongoing regular business activities	Cash_Flow_From_Opr
35	Cash Flow From Investing Activities	Cash used in the purchase of non-current assets–or long-term assets– that will deliver value in the future	Cash_Flow_From_Inv
36	Cash Flow From Financing Activities	Net flows of cash that are used to fund the company (transactions involving debt, equity, and dividends)	Cash_Flow_From_Fin
37	ROG-Net Worth (%)	Rate of Growth - Networth	ROG_Net_Worth_perc
38	ROG-Capital Employed (%)	Rate of Growth - Capital Employed	ROG_Capital_Employed_perc
39	ROG-Gross Block (%)	Rate of Growth - Gross Block	ROG_Gross_Block_perc
40	ROG-Gross Sales (%)	Rate of Growth - Gross Sales	ROG_Gross_Sales_perc
41	ROG-Net Sales (%)	Rate of Growth - Net Sales	ROG_Net_Sales_perc
42	ROG-Cost of Production (%)	Rate of Growth - Cost of Production	ROG_Cost_of_Prod_perc
43	ROG-Total Assets (%)	Rate of Growth - Total Assets	ROG_Total_Assets_perc
44	ROG-PBIDT (%)	Rate of Growth- PBIDT	ROG_PBIDT_perc
45	ROG-PBDT (%)	Rate of Growth- PBDT	ROG_PBDT_perc
46	ROG-PBIT (%)	Rate of Growth- PBIT	ROG_PBIT_perc
47	ROG-PBT (%)	Rate of Growth- PBT	ROG_PBT_perc
48	ROG-PAT (%)	Rate of Growth- PAT	ROG_PAT_perc
49	ROG-CP (%)	Rate of Growth- CP	ROG_CP_perc
50	ROG-Revenue earnings in forex (%)	Rate of Growth - Revenue earnings in forex	ROG_Rev_earn_in_forex_perc

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

Table: Data Dictionary

Introduction to Data:

Information related to 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]
0	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
1	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
2	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
3	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
4	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

1. DATASET SAMPLE

```
The number of rows in this dataset is  3586
 The number of columns in this dataset is  67
```

2. DATASET SHAPE

```
#   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
 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
```

3. INFO of DATASET

	Co_Code	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	Current Liabilities and Provisions	...	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]	APATM (%) [Latest]	Debtors Velocity (Days)	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block
count	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	...	3585.00	3585.00	3585.00	3585.00	3585.00	3586.00	3586.00	3483.00	3586.00	3586.00
mean	16065.39	725.05	62.97	649.75	2799.61	1994.82	594.18	410.81	1960.35	391.99	...	-51.16	-109.21	-311.57	-307.01	-365.06	603.89	2057.85	79.64	0.82	61.88
std	19776.82	4769.68	778.76	4091.99	26975.14	23652.84	4871.55	6301.22	22577.57	2675.00	...	1795.13	3057.64	10921.59	10676.15	12500.05	10636.76	54169.48	137.85	1.20	976.82
min	4.00	-8021.60	0.00	-7027.48	-1824.75	-0.72	-41.19	-13162.42	-0.91	-0.23	...	-78870.45	-141600.00	-590500.00	-572000.00	-688600.00	0.00	0.00	-199.00	-0.33	-61.00
25%	3029.25	3.98	3.75	3.89	7.60	0.03	0.57	0.94	4.00	0.73	...	0.00	0.00	0.00	0.00	0.00	8.00	8.00	0.00	0.07	0.27
50%	6077.50	19.02	8.29	18.58	39.09	7.49	15.87	10.14	24.54	9.23	...	8.07	5.23	4.69	3.89	1.59	49.00	39.00	35.00	0.48	1.53
75%	24269.50	123.80	19.52	117.30	226.60	72.35	131.90	61.17	135.28	65.65	...	18.99	14.29	14.11	11.39	7.41	106.00	89.00	96.00	1.16	4.91
max	72493.00	111729.10	42263.46	81657.35	714001.25	652823.81	128477.59	223257.56	721166.00	83232.98	...	19233.33	19195.70	15640.00	15640.00	15266.67	514721.00	2034145.00	996.00	17.63	43404.00

4. DESCRIPTION of DATASET

Inference:

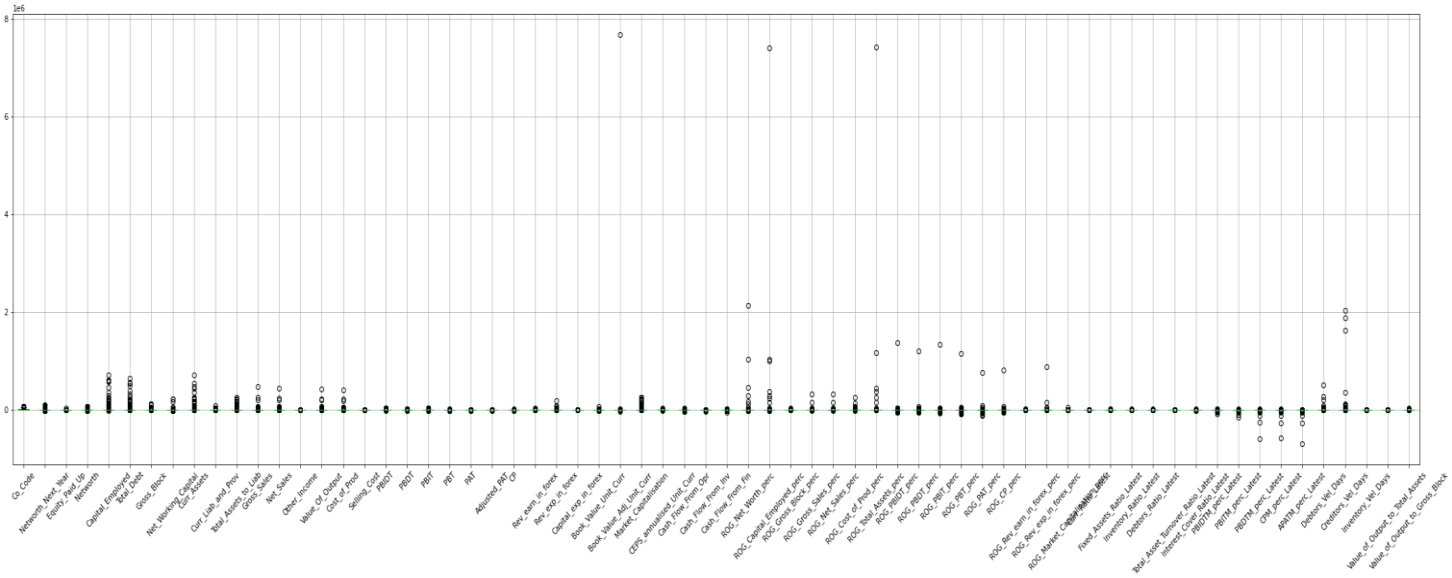
- The dataset has names of various companies of different sizes. It contains information related to the financial statement of the companies for the previous year (2015). Also given is the information about the 'Networth' of the company in the following year (2016) This information will help us derive the target column in the later stages of building the model.

- The data has 67 columns and 3586 rows.

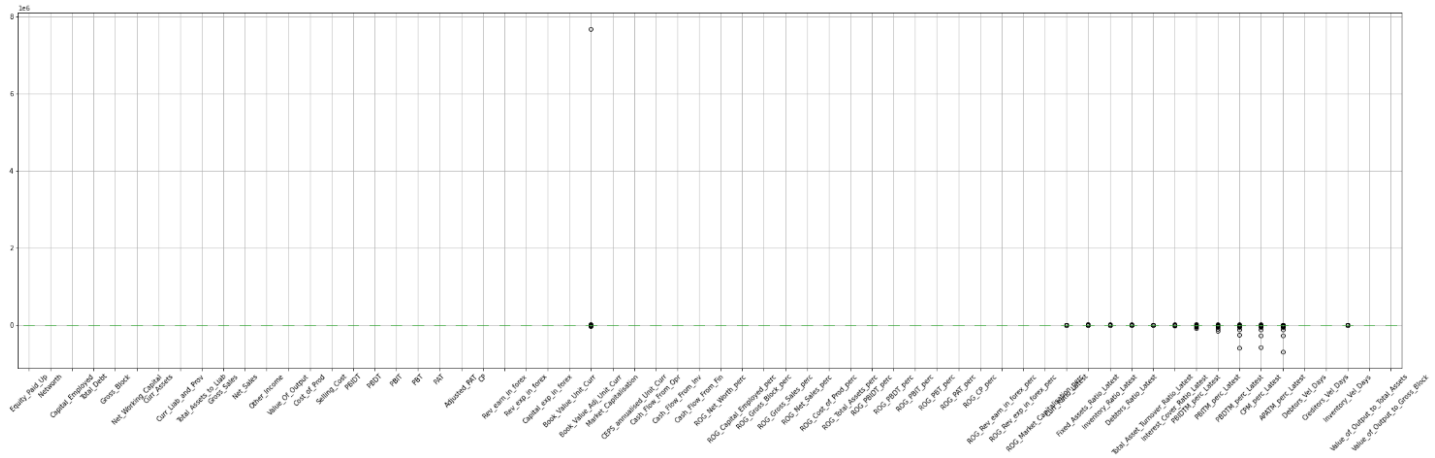
- On inspecting the information of the dataset it is found that there are 63 columns of data type 'float64', 3 columns of data type 'int64', and 1 column 'Co_name' of datatype 'object'.

- The description of the dataset gives us values of mean, min, max for the numerical columns present in the dataset. It shows that there are columns with min and max of vast differences, this is because the above dataset is not scaled yet.

1.1 OUTLIER TREATMENT



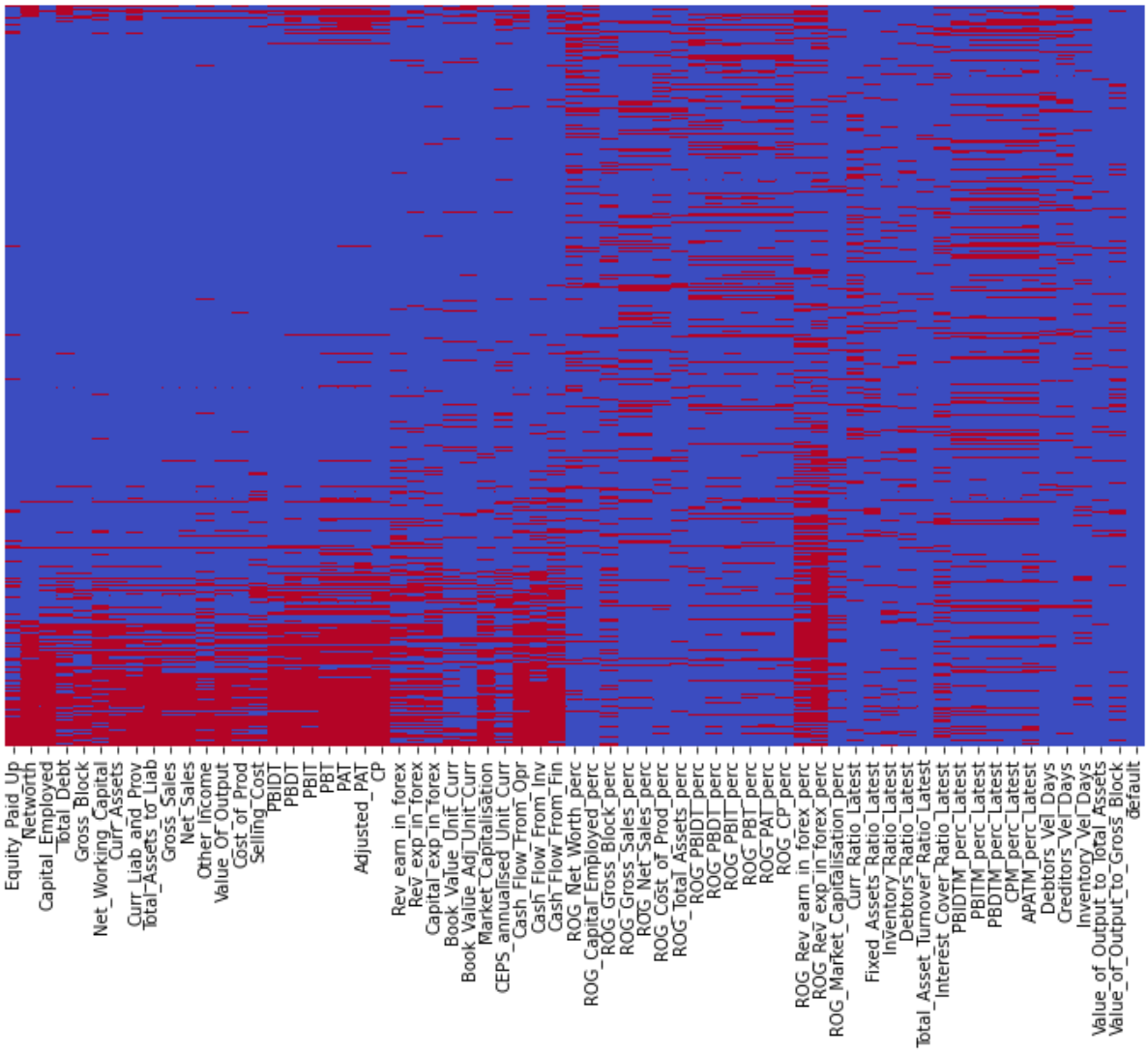
5. OUTLIERS in the DATASET



6. DATASET after OUTLIER TREATMENT

From the above figure we can see that most of the variables in the dataset contains outliers. Certain columns such as 'Co_Name', 'Networth_Next_Year' and 'Co_Code' are dropped before treating the outliers as they are not essential for model building and can make the data too noisy. Treating the outlier may affect the data. Since, the data has financial inputs and is captured from different companies of different sizes, the outliers have information which is important in nature and should be treated accordingly. The treatment of the data is done using the Inter Quantile Range method while describing the upper and lower limit within which the values should lie.

1.2 MISSING VALUE TREATMENT



7. Visual Representation of Missing values in the dataset

```
0        19
1        34
2        43
3        36
4        35
        ..
3581     30
3582     36
3583     34
3584     30
3585     36
Length: 3586, dtype: int64
```

8.Total missing variables Row wise

```
ROG_Rev_exp_in_forex_perc           0.45
ROG_Rev_earn_in_forex_perc          0.37
Cash_Flow_From_Fin                  0.28
PAT                                 0.27
Adjusted_PAT                        0.27
                                    ...
Debtors_Ratio_Latest                0.10
Inventory_Vel_Days                  0.10
Total_Asset_Turnover_Ratio_Latest   0.06
Value_of_Output_to_Total_Assets     0.04
default                             0.00
Length: 65, dtype: float64
```

9.Percentage of missing variable in each column

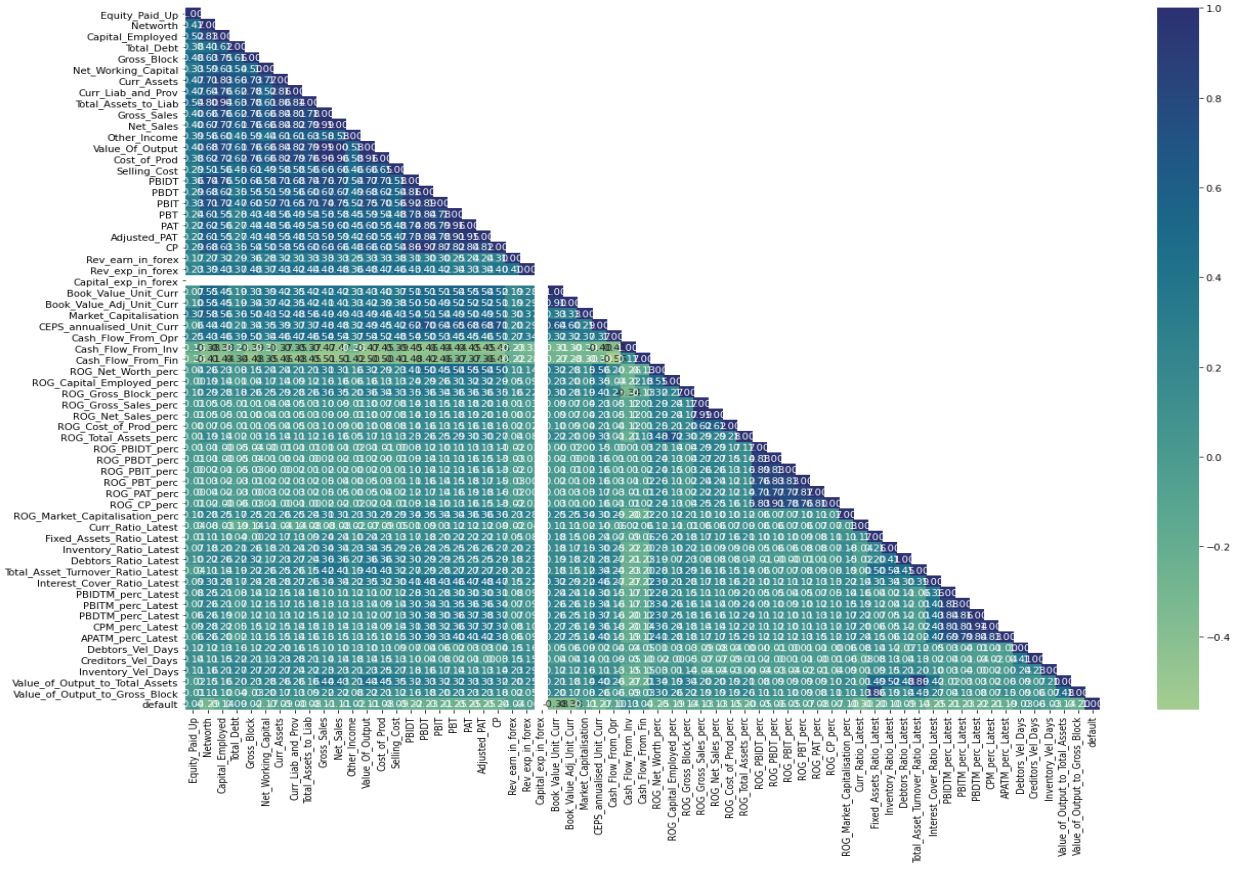
As we can see there are a lot of missing variables in many rows and columns. We drop the columns with missing values exceeding 30% and then scale the Predictor variables in the data before imputing the missing values.

For imputing the missing values, KNN Imputer is being used, it takes the average of the nearest neighbours and fills the missing values accordingly.

```
Equity_Paid_Up                  0
Networth                        0
Capital_Employed                0
Total_Debt                      0
Gross_Block                     0
                               ..
Creditors_Vel_Days              0
Inventory_Vel_Days              0
Value_of_Output_to_Total_Assets 0
Value_of_Output_to_Gross_Block  0
default                         0
Length: 63, dtype: int64
```

10.Variables after using KNN-Imputer for missing variables

CORRELATION IN THE DATASET

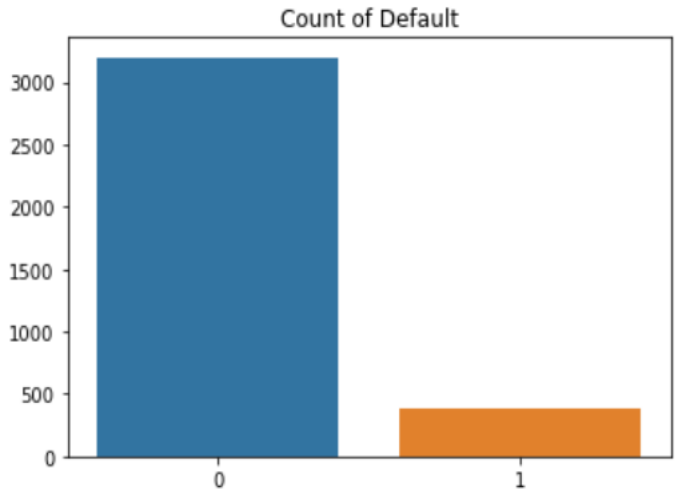


11.Correlation plot among variables after imputation

We can see that there are many variables that are highly positively and highly negatively correlated to each other. High correlation among the variables is not essential for model building.

1.3 TRANSFORMING TARGET VARIABLE TO 0 & 1

	default	Networth_Next_Year
0	1	-8021.60
1	1	-3986.19
2	1	-3192.58
3	1	-3054.51
4	1	-2967.36
5	1	-2519.40
6	1	-2125.05
7	1	-2100.56
8	1	-1695.75
9	1	-1677.18



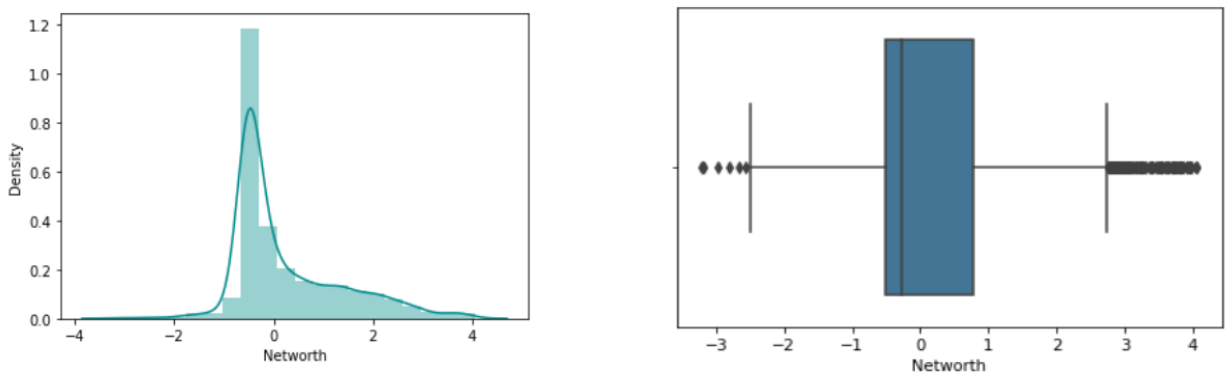

```
0   0.89
1   0.11
Name: default, dtype: float64
```

12. Converting Target Variable to 0 & 1

The target variable is converted to 0 when net worth next year is positive & 1 for when net worth next year is negative. The target variable gives us a predictio of whether the company will or will not default in the coming year taking into consideration all the variables.

1.4 UNIVARIATE & BIVARIATE ANALYSIS

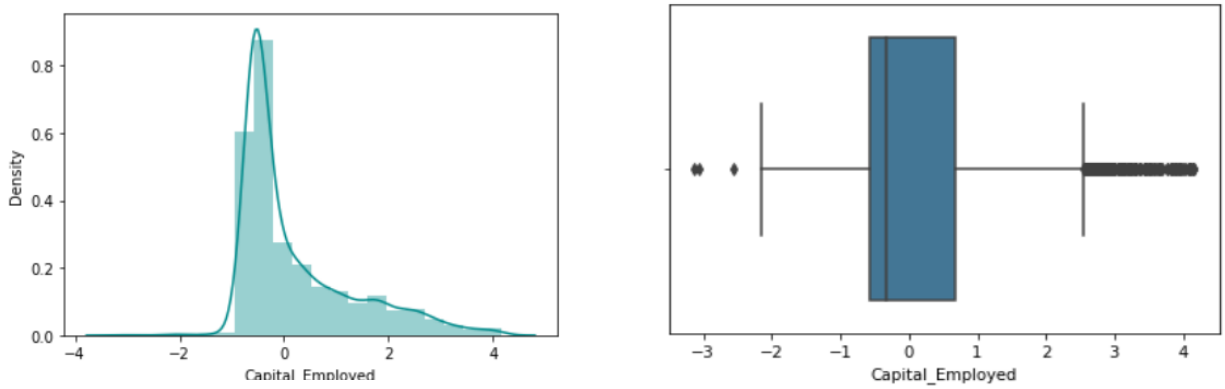
NETWORTH:



13.Univariate Analysis: Networth

The variable networth shows a normal distribution. The box plot for the same shows the presence of outliers.

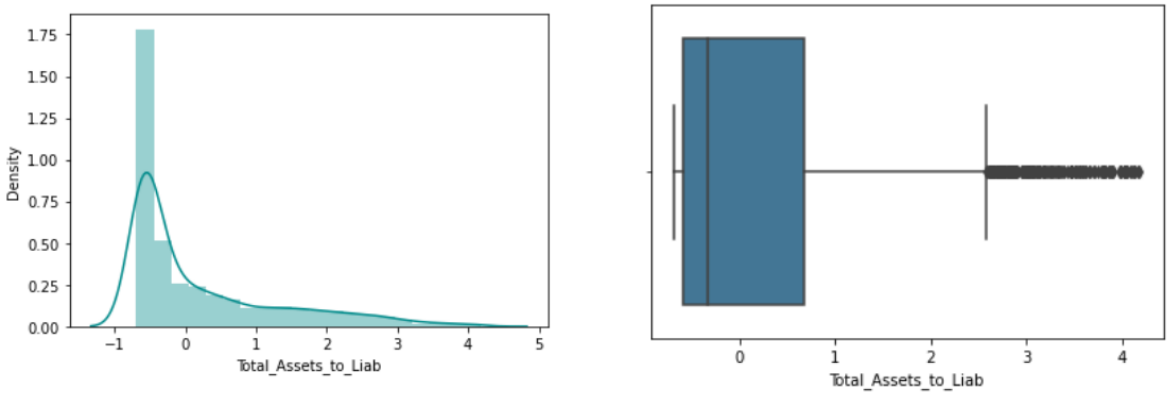
CAPITAL EMPLOYED:



14.Univariate Analysis:Capital Employed

The variable Captial_Employed is slightly right skewed and also shows the presence of outliers in the dataset.

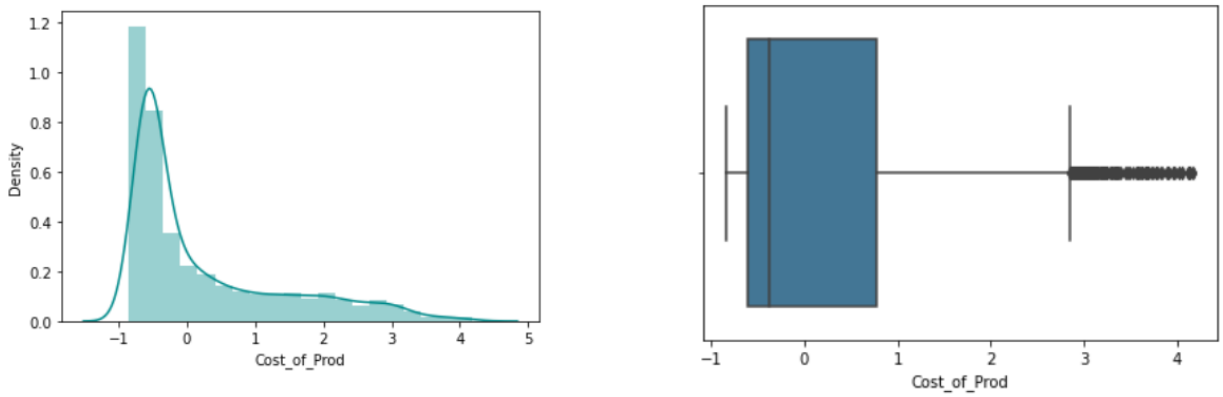
TOTAL ASSETS TO LIABILITY



15.Univariate Analysis: Total_Assets_to_Liab

The variable Total_Assets_to_Liab is highly right skewed and also shows the presence of outliers in the dataset.

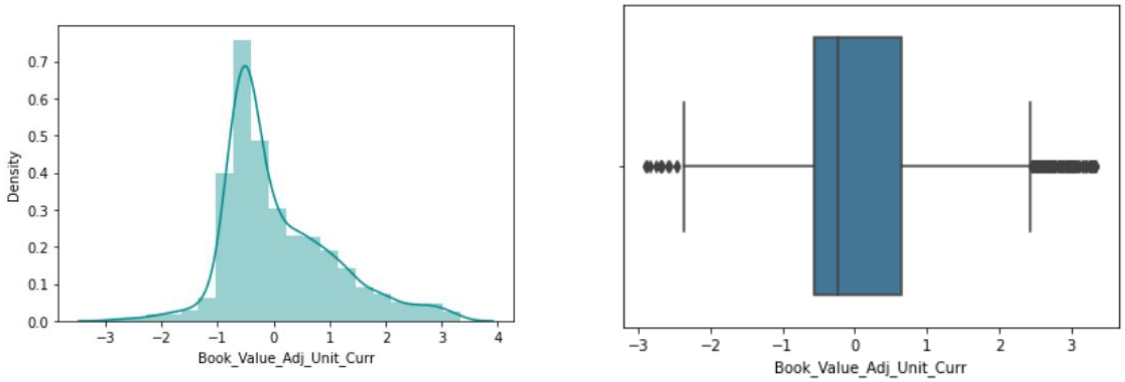
COST OF PRODUCTION



16.Univariate Analysis: Cost_of_Prod

The variable Cost_of_Prod is highly right skewed and also shows the presence of outliers in the dataset.

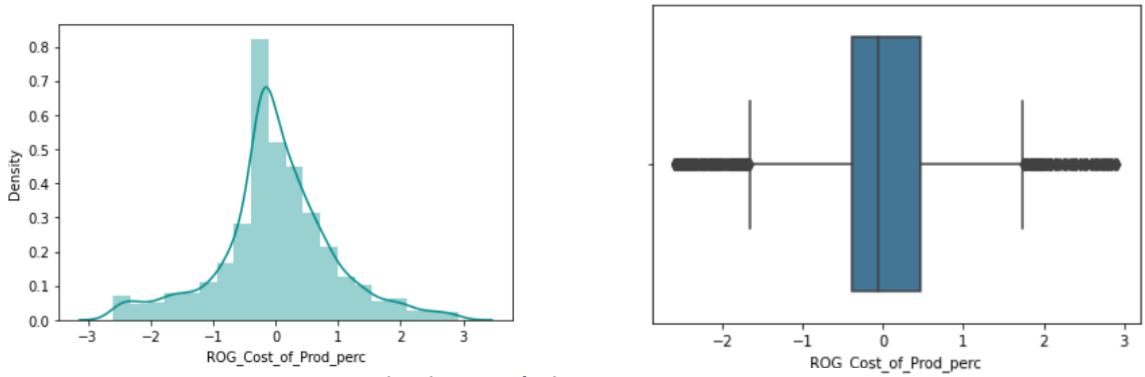
BOOK VALUE ADJ UNIT CURR



16.Univariate Analysis: Book_Value_Adj_Unit_Curr

The variable Book_Value_Adj_Unit_Curr is right skewed and also shows the presence of outliers in the dataset.

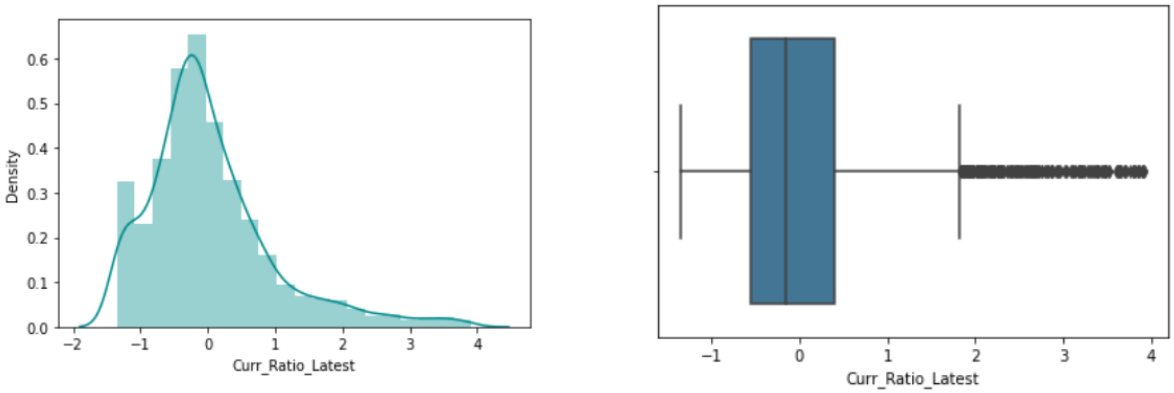
ROG_COST_OF_PROD_PERC



17.Univariate Analysis: ROG_Cost_of_Prod_perc

The variable ROG_Cost_of_Prod_perc is normally distributed. It has the presence of outliers in the dataset.

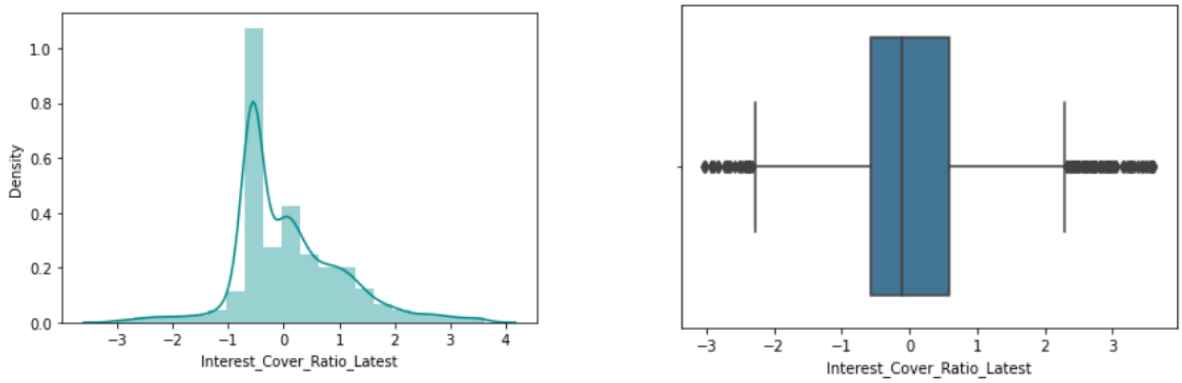
CURR_RATIO_LATEST



18.Univariate Analysis: ROG_Cost_of_Prod_perc

The variable is slightly right skewed and has the presence of outliers.

INTEREST COVER RATIO LATEST



19.Univariate Analysis: Interest_Cover_Ratio_Latest

SKEWNESS OF THE DATA

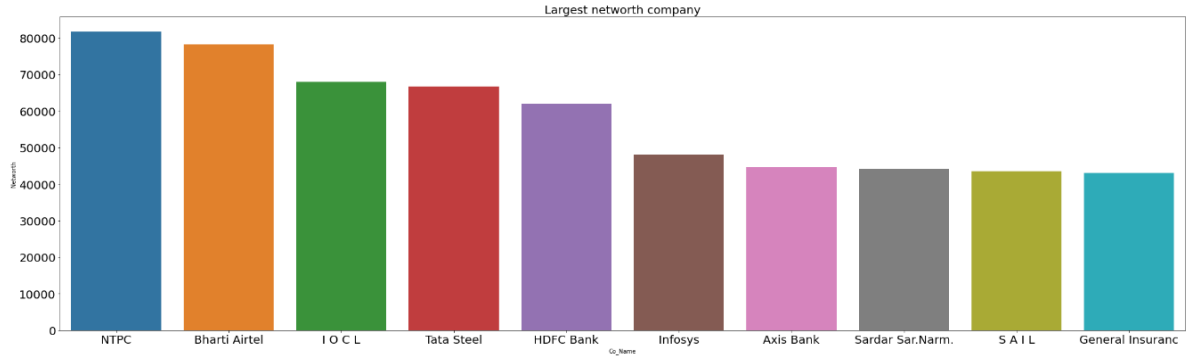
	Skewness
Book_Value_Adj_Unit_Curr	59.84
ROG_PBIT_perc	58.93
ROG_PBIDT_perc	58.88
ROG_PBDT_perc	58.41
ROG_PBT_perc	57.33
...	...
PBIDTM_perc_Latest	-30.93
PBITM_perc_Latest	-36.00
CPM_perc_Latest	-47.01
PBDTM_perc_Latest	-47.75
APATM_perc_Latest	-49.28

The data seems to be both highly negatively and positively skewed. They are skewed on either sides.

20.Skewness of the Data

BI-VARIATE ANALYSIS

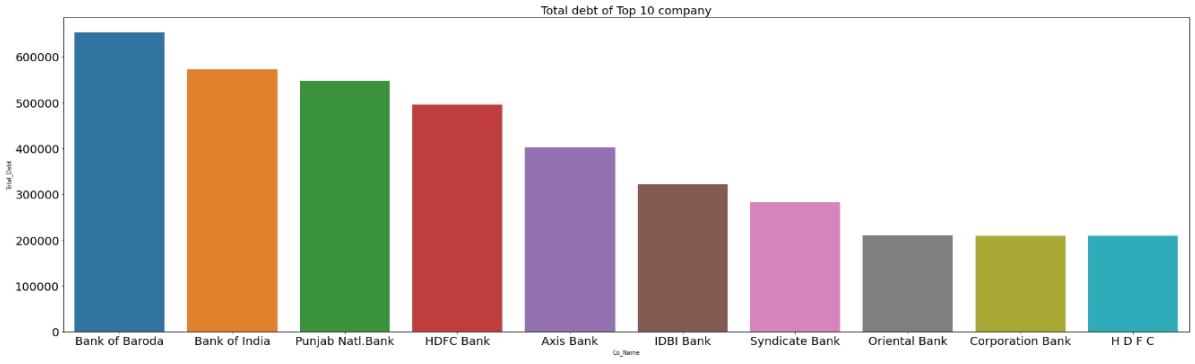
Networth & Co_name



21.Largest net worth of company

From the above bar graph, we can see that the highest net worth of company is for NTPC followed by Bharti Airtel and IOCL. The net worth for NTPC exceeds 80000.

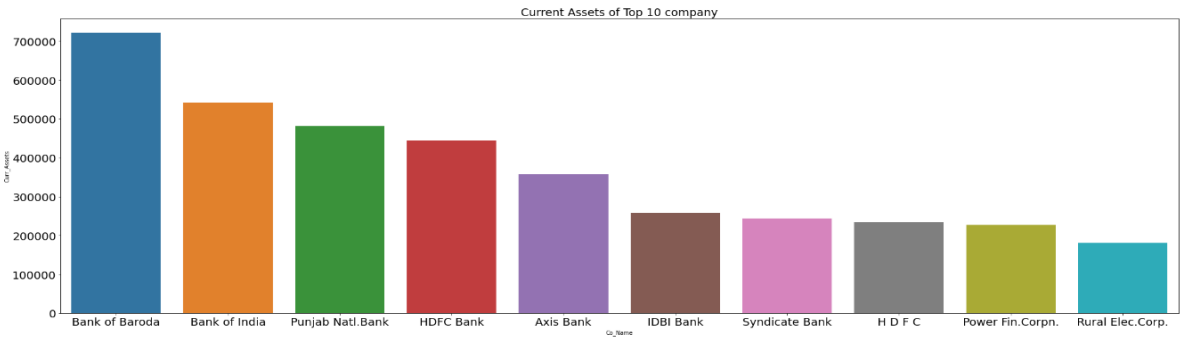
Total Debt of Company



22.Total Debt of company

Bank of Baroda has the highest Total debt among all the companies followed by Bank of India and Punjab National Bank. The total debt for Bank of Baroda exceeds 600000.

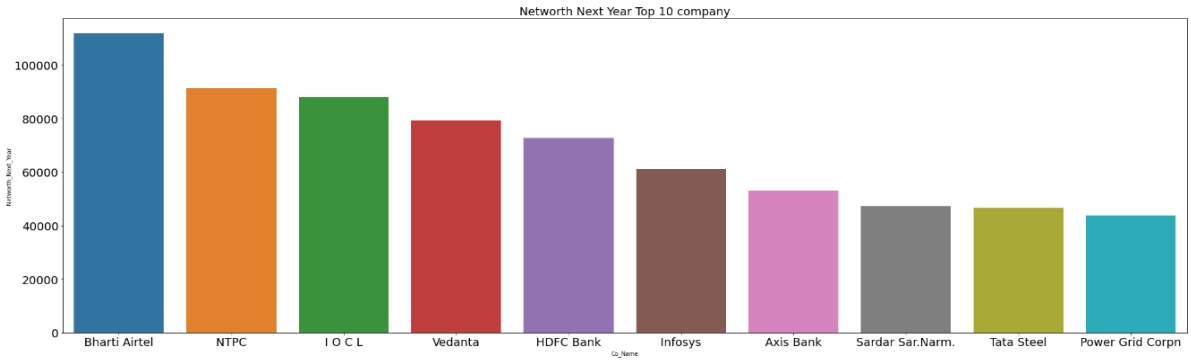
Current Assets of Company



23.Current Assets of company

The highest current assets belong to Bank of Baroda followed by Bank of India and Punjab National Bank just like total debt these companies also have highest Current Assets.

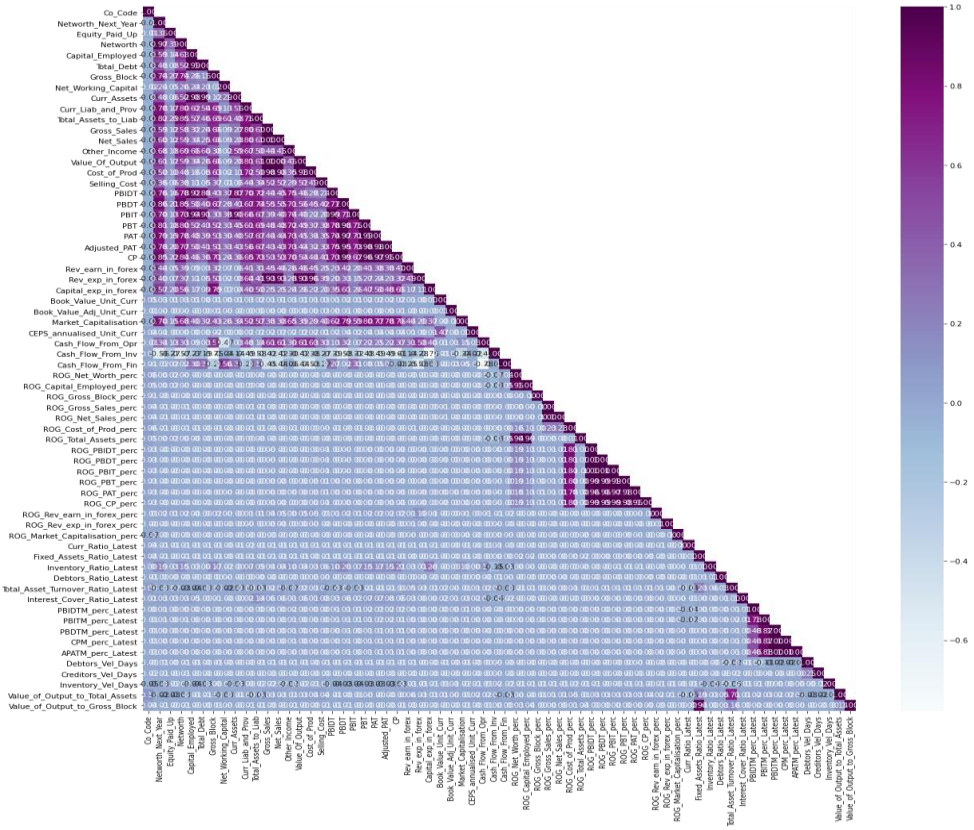
Networth Next Year



23.Networth Next Year

The net worth next year has been predicted the highest for Bharti Airtel, followed by NTPC and IOCL. We can see that through predictions of net worth next year Bharti Airtel may take over NTPC for net worth.

CORRELATION PLOT



24.Correlation Plot

1.5 Train Test Split

```
(2402, 62)          (2402,)
(1184, 62)          (1184,)
```

25.Train Test Split: X and y variables

On splitting the Train and test variables using sklearn.model_selection we see that the data has been split to 2402, 62 for X_train variables and 1184,62 for X_test variables.

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

Ain After splitting the variables into X_train, X_test and y_train and y_test, we have to use selected variables that will be needed in the model building.

Using VIF is a lenthy process as the dataset contains more than 60 columns and hence here we have used RFE(recursive feature elimination) from sklearn.feature_selection. The goal of RFE is to select features by recursively considering smaller and smaller sets of features. The following are the features that will be used for building the model as per RFE.

	Feature	Rank
1	Networth	1
2	Capital_Employed	1
6	Curr_Assets	1
8	Total_Assets_to_Liab	1
9	Gross_Sales	1
12	Value_Of_Output	1
13	Cost_of_Prod	1
19	PAT	1
20	Adjusted_PAT	1
25	Book_Value_Unit_Curr	1
26	Book_Value_Adj_Unit_Curr	1
37	ROG_Cost_of_Prod_perc	1
46	Curr_Ratio_Latest	1
47	Fixed_Assets_Ratio_Latest	1
51	Interest_Cover_Ratio_Latest	1

26.RFE feature selection

On using RFE we get the features that will be used for model building.

Model is built using Stats Model library, using all the features as per RFE, on building the model we see that many variables show a p value that is way above the alpha. We need a p-value that is 0.00 and hence we keep on reducing the features in the model until we get the aforementioned features with a p-value of 0.00.

Logit Regression Results

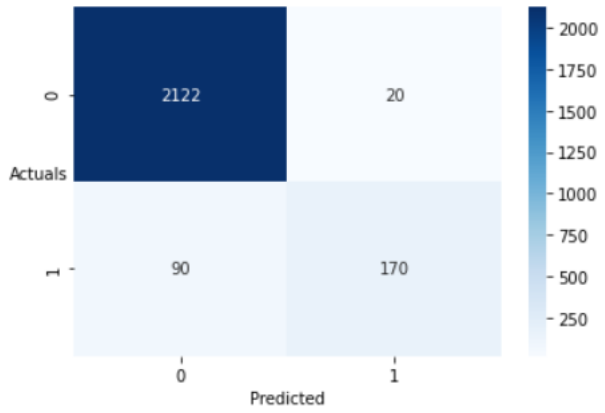
Dep. Variable:	default	**No. Observations:**	3586
Model:	Logit	**Df Residuals:**	3577
Method:	MLE	**Df Model:**	8
Date:	Sun, 12 Jun 2022	**Pseudo R-squ.:**	0.5624
Time:	18:08:07	**Log-Likelihood:**	-537.88
converged:	True	**LL-Null:**	-1229.0
Covariance Type:	nonrobust	**LLR p-value:**	3.771e-293

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.9126	0.209	-23.464	0.000	-5.323	-4.502
Networth	-1.4661	0.215	-6.812	0.000	-1.888	-1.044
Capital_Employed	-1.3988	0.244	-5.726	0.000	-1.878	-0.920
Total_Assets_to_Liab	1.7627	0.226	7.797	0.000	1.320	2.206
Cost_of_Prod	0.5482	0.130	4.213	0.000	0.293	0.803
Book_Value_Adj_Unit_Curr	-3.0566	0.236	-12.960	0.000	-3.519	-2.594
ROG_Cost_of_Prod_perc	-0.3905	0.094	-4.159	0.000	-0.575	-0.206
Curr_Ratio_Latest	-1.3513	0.131	-10.335	0.000	-1.608	-1.095
Interest_Cover_Ratio_Latest	-0.7637	0.124	-6.151	0.000	-1.007	-0.520

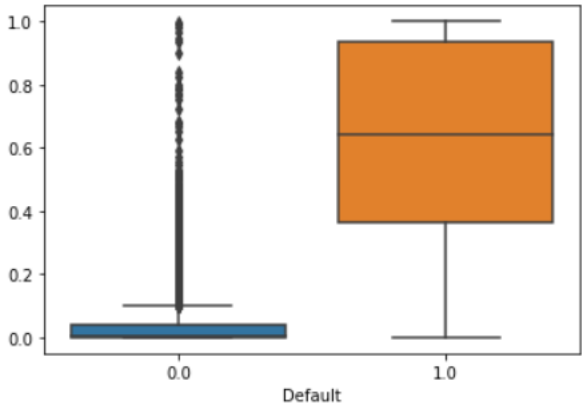
Possibly complete quasi-separation: A fraction 0.19 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

27. Model 8 summary using RFE and stats model

Model 8 shows the best features for model building as all the features gives us a p-value of 0.00 when other features with high p-value is removed.



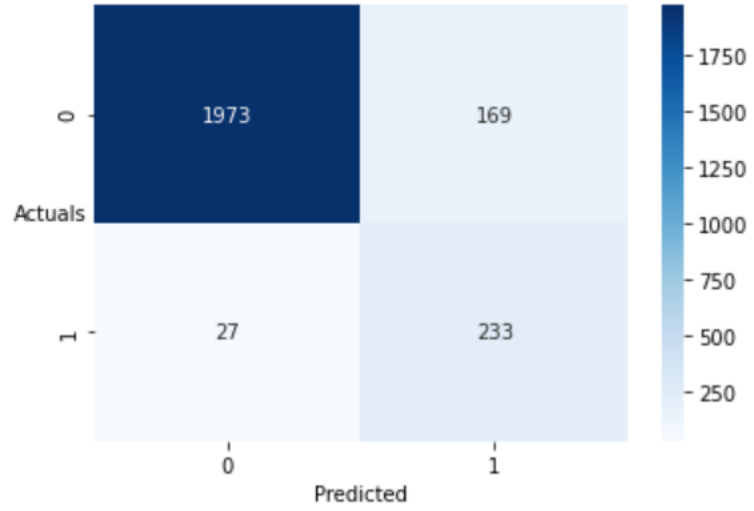
28. Confusion Matrix: y_predict



29.Default predicted using model 8

```
0.17405371823629606
```

30.Optimum cut-off



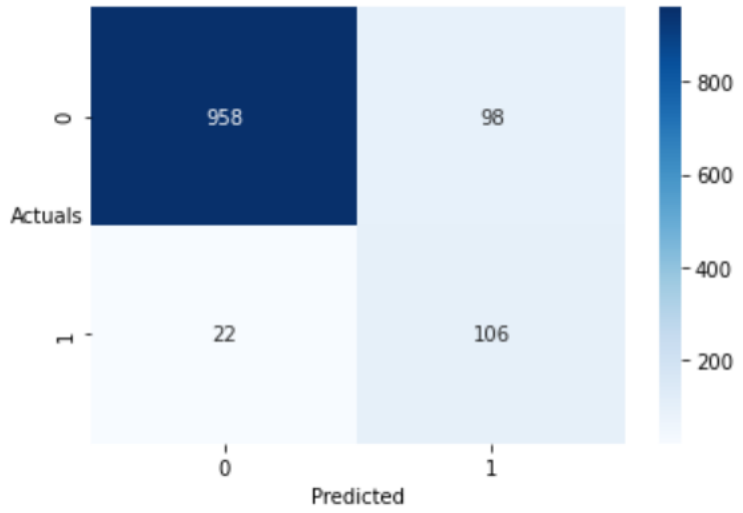
31.Confusion matrix: Revised matrix after using optimum cut off

	precision	recall	f1-score	support
0.0	0.987	0.921	0.953	2142
1.0	0.580	0.896	0.704	260
accuracy			0.918	2402
macro avg	0.783	0.909	0.828	2402
weighted avg	0.942	0.918	0.926	2402

32.Classification report: Train data data after using optimal cut off

On building the model without optimum cut off the data gives values that are wrongly predicted for 0 and 1. Using the optimum cut off at 0.17 the matrix gives us better values with lower wrong predicted variables.

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



33.Confusion matrix: Test Data

	precision	recall	f1-score	support
0.0	0.978	0.907	0.941	1056
1.0	0.520	0.828	0.639	128
accuracy			0.899	1184
macro avg	0.749	0.868	0.790	1184
weighted avg	0.928	0.899	0.908	1184

34.Classification report: Test data

From the above Classification report we can see that the recall has given us a good value at 0.828 there is no case of over fitting and under fitting. Using the optimum threshold value has helped with getting a good recall but the value for precision has gone lower than in train set. To ensure there is a trade off between recall and precision we can work on adjusting the threshold.