



# Company Default Prediction Modelling

FRA Project -  
Milestone 1

---

BY ANOOP RAUT (PGDBA)

# Content

---

- 1. Problem Statement**
- 2. Data Overview**
- 3. Data Cleaning & Processing**
- 4. Data Analysis**
- 5. Model Building**
- 6. Prediction & Validation**



# 1. 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 labelled field.

Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

# 2. Data Overview

## 2.1 Glimpse of Dataset

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
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	1116.85	109.60	0.00	0.00
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	1585.74	6043.94	2892.73	2892.73
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	4601.39	12316.07	392.13	392.13
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	3646.54	6000.42	1354.39	1354.39
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	2849.58	7524.91	38.72	38.72

Value Of Output	Cost of Production	Selling Cost	PBDT	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
-0.07	137.67	0.00	-179.06	-926.52	-185.53	-932.99	-932.99	-937.85	-926.52	0.00	0.00	0.00	-167.58	-167.58	0.00
2900.71	2572.46	40.51	646.46	-4.32	35.53	-615.25	-615.25	-617.14	-4.32	6.35	143.42	141.17	-15.18	-15.18	1544.39
301.16	408.51	54.83	-281.92	-1086.71	-381.10	-1185.89	-897.70	-873.39	-798.52	0.00	86.36	2.27	94.14	94.14	1220.81
1350.14	1326.99	3.34	-213.01	-677.57	-336.73	-801.29	-801.29	-770.18	-677.57	0.89	28.88	0.00	-39.64	-39.64	194.27
38.72	186.29	1.97	-647.86	-944.42	-710.13	-1006.69	-864.58	-327.77	-802.31	0.00	15.62	0.00	-212.89	-212.89	113.68

CEPS (annualised) (Unit Curr)	Cash Flow From Operating Activities	Cash Flow From Investing Activities	Cash Flow From Financing Activities	ROG- Net Worth (%)	ROG- Capital Employed (%)	ROG- Gross Block (%)	ROG- Gross Sales (%)	ROG- Net Sales (%)	ROG- Cost of Production (%)	ROG- Total Assets (%)	ROG- PBDT (%)	ROG- PBDT (%)	ROG- PBIT (%)	ROG- PBIT (%)	ROG- PBT (%)	ROG- PBT (%)	ROG- PAT (%)	ROG- PAT (%)	ROG- CP (%)	ROG- CP (%)
-22.09	-102.47	1.46	92.58	-15.31	-20.76	-9.75	0.00	0.00	1.38	-25.75	5.23	-19.29	4.43	-19.33	-19.33	-19.29	-19.29	-19.29	-19.29	
-0.02	635.91	-785.00	176.93	-26.15	6.30	3.17	5.92	5.92	13.16	6.66	5.24	-108.77	614.89	-9.85	-9.85	-108.77	-108.77	-108.77	-108.77	
-148.31	-873.40	-458.27	1187.51	-61.86	15.66	-2.07	-75.87	-75.87	-69.93	-0.78	-169.69	-431.68	-221.85	-300.41	-350.43	-642.67	-642.67	-642.67	-642.67	
-43.08	324.47	17.31	-412.55	-450.67	-40.84	-1.80	8.33	8.33	16.59	-10.12	-203.84	-102.72	-455.28	-80.23	-70.57	-88.52	-88.52	-88.52	-88.52	
-159.50	-191.54	61.34	143.65	-559.83	-11.76	0.34	-80.61	-80.61	-48.80	-5.81	-40.07	3.28	-38.72	1.87	-2.59	-1.13	-1.13	-1.13	-1.13	

## 2. Data Overview

### 2.1 Glimpse of Dataset

A dark background image showing a person's hand holding a smartphone. The screen of the phone is blurred, appearing as a grid of numbers and text, suggesting a financial dataset or spreadsheet.

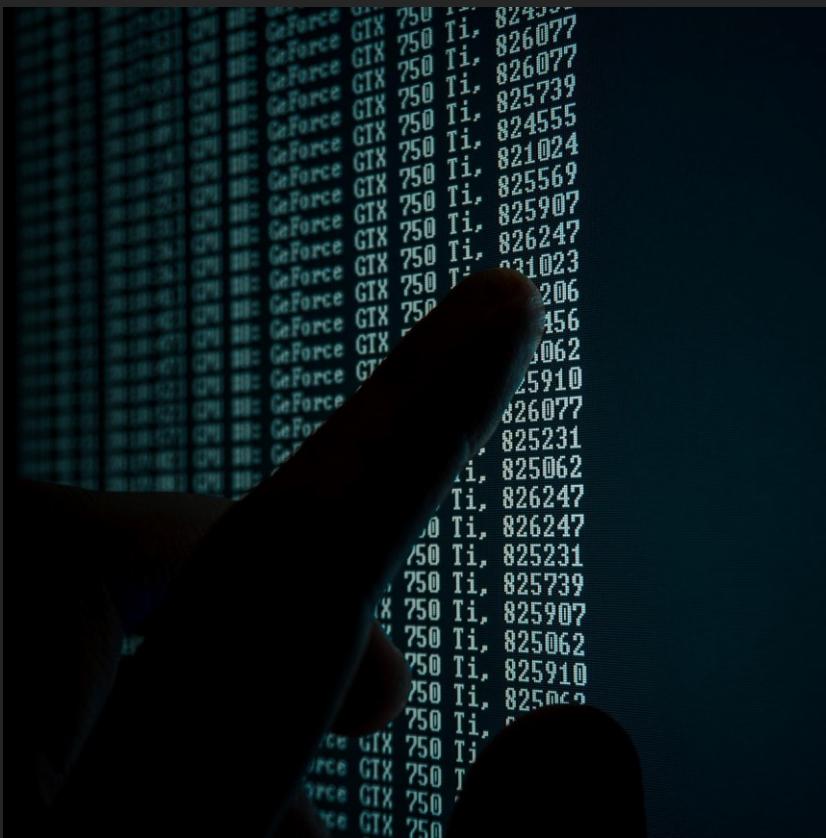
Product ID	Category	Item Name	Unit Price	Stock Level	Last Update
1234567890	Electronics	Smartphone X	\$599.99	250	2023-10-01
1234567891	Electronics	Smartphone Y	\$499.99	300	2023-10-02
1234567892	Electronics	Smartphone Z	\$699.99	200	2023-10-03
1234567893	Electronics	Smartphone A	\$549.99	280	2023-10-04
1234567894	Electronics	Smartphone B	\$449.99	320	2023-10-05
1234567895	Electronics	Smartphone C	\$599.99	220	2023-10-06
1234567896	Electronics	Smartphone D	\$499.99	340	2023-10-07
1234567897	Electronics	Smartphone E	\$699.99	200	2023-10-08
1234567898	Electronics	Smartphone F	\$549.99	260	2023-10-09
1234567899	Electronics	Smartphone G	\$449.99	310	2023-10-10
1234567890	Electronics	Smartphone H	\$599.99	240	2023-10-11
1234567891	Electronics	Smartphone I	\$499.99	330	2023-10-12
1234567892	Electronics	Smartphone J	\$699.99	210	2023-10-13
1234567893	Electronics	Smartphone K	\$549.99	270	2023-10-14
1234567894	Electronics	Smartphone L	\$449.99	350	2023-10-15
1234567895	Electronics	Smartphone M	\$599.99	230	2023-10-16
1234567896	Electronics	Smartphone N	\$499.99	320	2023-10-17
1234567897	Electronics	Smartphone O	\$699.99	200	2023-10-18
1234567898	Electronics	Smartphone P	\$549.99	280	2023-10-19
1234567899	Electronics	Smartphone Q	\$449.99	300	2023-10-20
1234567890	Electronics	Smartphone R	\$599.99	250	2023-10-21
1234567891	Electronics	Smartphone S	\$499.99	310	2023-10-22
1234567892	Electronics	Smartphone T	\$699.99	220	2023-10-23
1234567893	Electronics	Smartphone U	\$549.99	260	2023-10-24
1234567894	Electronics	Smartphone V	\$449.99	330	2023-10-25
1234567895	Electronics	Smartphone W	\$599.99	200	2023-10-26
1234567896	Electronics	Smartphone X	\$499.99	280	2023-10-27
1234567897	Electronics	Smartphone Y	\$699.99	240	2023-10-28
1234567898	Electronics	Smartphone Z	\$549.99	300	2023-10-29
1234567899	Electronics	Smartphone A	\$449.99	210	2023-10-30
1234567890	Electronics	Smartphone B	\$599.99	270	2023-10-31
1234567891	Electronics	Smartphone C	\$499.99	340	2023-11-01
1234567892	Electronics	Smartphone D	\$699.99	200	2023-11-02
1234567893	Electronics	Smartphone E	\$549.99	260	2023-11-03
1234567894	Electronics	Smartphone F	\$449.99	310	2023-11-04
1234567895	Electronics	Smartphone G	\$599.99	230	2023-11-05
1234567896	Electronics	Smartphone H	\$499.99	320	2023-11-06
1234567897	Electronics	Smartphone I	\$699.99	200	2023-11-07
1234567898	Electronics	Smartphone J	\$549.99	280	2023-11-08
1234567899	Electronics	Smartphone K	\$449.99	350	2023-11-09
1234567890	Electronics	Smartphone L	\$599.99	240	2023-11-10
1234567891	Electronics	Smartphone M	\$499.99	330	2023-11-11
1234567892	Electronics	Smartphone N	\$699.99	210	2023-11-12
1234567893	Electronics	Smartphone O	\$549.99	270	2023-11-13
1234567894	Electronics	Smartphone P	\$449.99	300	2023-11-14
1234567895	Electronics	Smartphone Q	\$599.99	220	2023-11-15
1234567896	Electronics	Smartphone R	\$499.99	310	2023-11-16
1234567897	Electronics	Smartphone S	\$699.99	200	2023-11-17
1234567898	Electronics	Smartphone T	\$549.99	280	2023-11-18
1234567899	Electronics	Smartphone U	\$449.99	340	2023-11-19
1234567890	Electronics	Smartphone V	\$599.99	250	2023-11-20
1234567891	Electronics	Smartphone W	\$499.99	320	2023-11-21
1234567892	Electronics	Smartphone X	\$699.99	210	2023-11-22
1234567893	Electronics	Smartphone Y	\$549.99	260	2023-11-23
1234567894	Electronics	Smartphone Z	\$449.99	330	2023-11-24
1234567895	Electronics	Smartphone A	\$599.99	230	2023-11-25
1234567896	Electronics	Smartphone B	\$499.99	300	2023-11-26
1234567897	Electronics	Smartphone C	\$699.99	200	2023-11-27
1234567898	Electronics	Smartphone D	\$549.99	280	2023-11-28
1234567899	Electronics	Smartphone E	\$449.99	350	2023-11-29
1234567890	Electronics	Smartphone F	\$599.99	240	2023-11-30
1234567891	Electronics	Smartphone G	\$499.99	330	2023-12-01
1234567892	Electronics	Smartphone H	\$699.99	210	2023-12-02
1234567893	Electronics	Smartphone I	\$549.99	270	2023-12-03
1234567894	Electronics	Smartphone J	\$449.99	300	2023-12-04
1234567895	Electronics	Smartphone K	\$599.99	220	2023-12-05
1234567896	Electronics	Smartphone L	\$499.99	310	2023-12-06
1234567897	Electronics	Smartphone M	\$699.99	200	2023-12-07
1234567898	Electronics	Smartphone N	\$549.99	280	2023-12-08
1234567899	Electronics	Smartphone O	\$449.99	340	2023-12-09
1234567890	Electronics	Smartphone P	\$599.99	250	2023-12-10
1234567891	Electronics	Smartphone Q	\$499.99	320	2023-12-11
1234567892	Electronics	Smartphone R	\$699.99	200	2023-12-12
1234567893	Electronics	Smartphone S	\$549.99	260	2023-12-13
1234567894	Electronics	Smartphone T	\$449.99	330	2023-12-14
1234567895	Electronics	Smartphone U	\$599.99	230	2023-12-15
1234567896	Electronics	Smartphone V	\$499.99	300	2023-12-16
1234567897	Electronics	Smartphone W	\$699.99	210	2023-12-17
1234567898	Electronics	Smartphone X	\$549.99	270	2023-12-18
1234567899	Electronics	Smartphone Y	\$449.99	350	2023-12-19
1234567890	Electronics	Smartphone Z	\$599.99	240	2023-12-20
1234567891	Electronics	Smartphone A	\$499.99	330	2023-12-21
1234567892	Electronics	Smartphone B	\$699.99	210	2023-12-22
1234567893	Electronics	Smartphone C	\$549.99	260	2023-12-23
1234567894	Electronics	Smartphone D	\$449.99	300	2023-12-24
1234567895	Electronics	Smartphone E	\$599.99	220	2023-12-25
1234567896	Electronics	Smartphone F	\$499.99	310	2023-12-26
1234567897	Electronics	Smartphone G	\$699.99	200	2023-12-27
1234567898	Electronics	Smartphone H	\$549.99	280	2023-12-28
1234567899	Electronics	Smartphone I	\$449.99	340	2023-12-29
1234567890	Electronics	Smartphone J	\$599.99	250	2023-12-30
1234567891	Electronics	Smartphone K	\$499.99	320	2023-12-31

ROG-Revenue earnings in forex (%)	ROG-Revenue expenses in forex (%)	ROG-Market Capitalisation (%)	Current Ratio[Latest]	Fixed Assets Ratio[Latest]	Inventory Ratio[Latest]	Debtors Ratio[Latest]	Total Asset Turnover Ratio[Latest]	Interest Cover Ratio[Latest]	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]
0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	-0.20	0.00	0.00	0.00	0.00
1.93	34.00	6.61	0.08	0.25	804.44	10.35	0.30	-0.84	-10.30	-39.74	-57.74	-57.74
-100.00	-81.21	-6.31	1.06	0.03	0.01	0.42	0.00	-2.21	-5279.14	-5516.98	-7780.25	-7723.67
-52.91	-33.85	-13.94	0.09	5.08	411.15	9.26	0.00	-0.16	-3.33	-7.21	-48.13	-47.70
-100.00	-91.17	-12.91	0.50	0.05	0.02	1.28	0.01	-0.73	-295.55	-400.55	-845.88	379.79

APATM (%) [Latest]	Debtors Velocity (Days)	Creditors Velocity (Days)	Inventory Velocity (Days)	Value of Output/Total Assets	Value of Output/Gross Block
0.00	0	0	45.0	0.00	0.00
-87.18	29	101	2.0	0.31	0.24
-7961.51	97	558	0.0	-0.03	-0.26
-51.58	93	63	2.0	0.24	1.90
274.79	3887	346	0.0	0.01	0.05

## 2. Data Overview

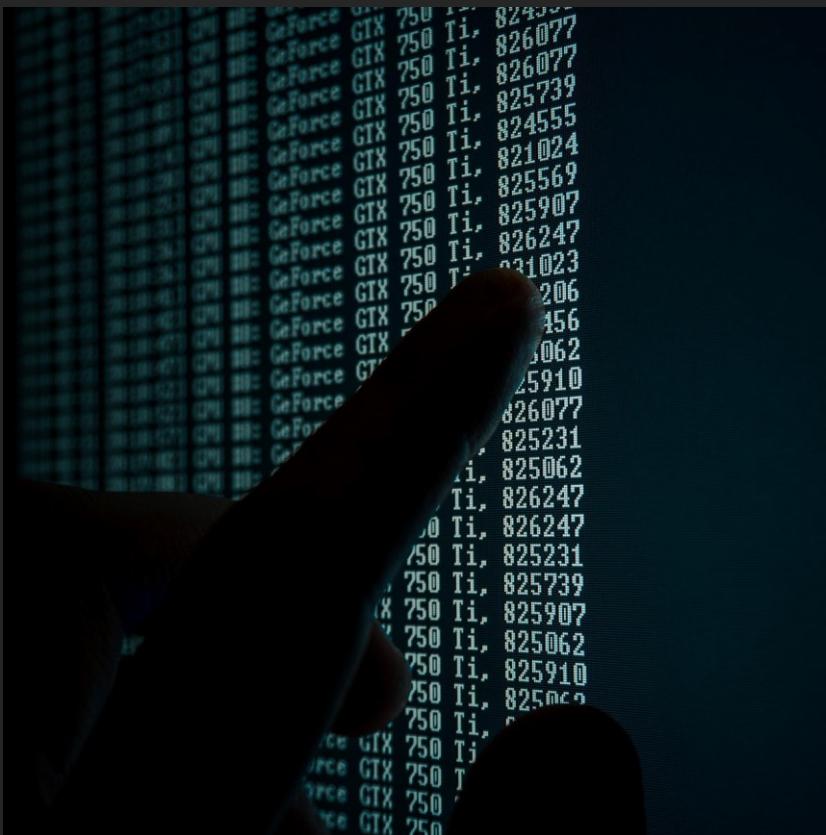
### 2.2 Data Dictionary



Col.No	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 liabilities 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

## 2. Data Overview

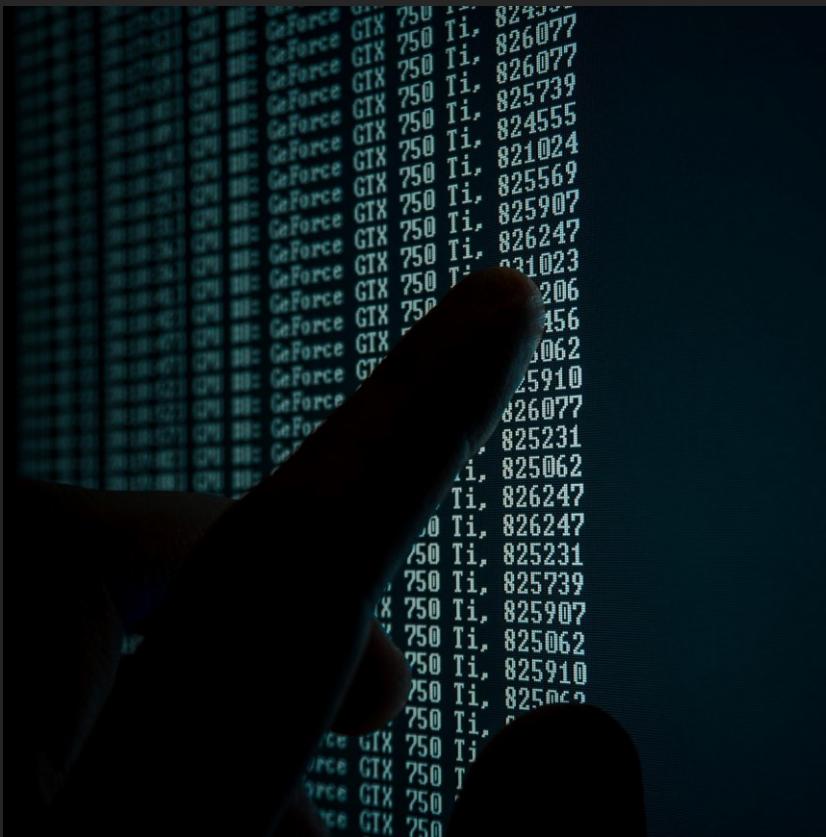
### 2.2 Data Dictionary



Col.No	Field Name	Description
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

## 2. Data Overview

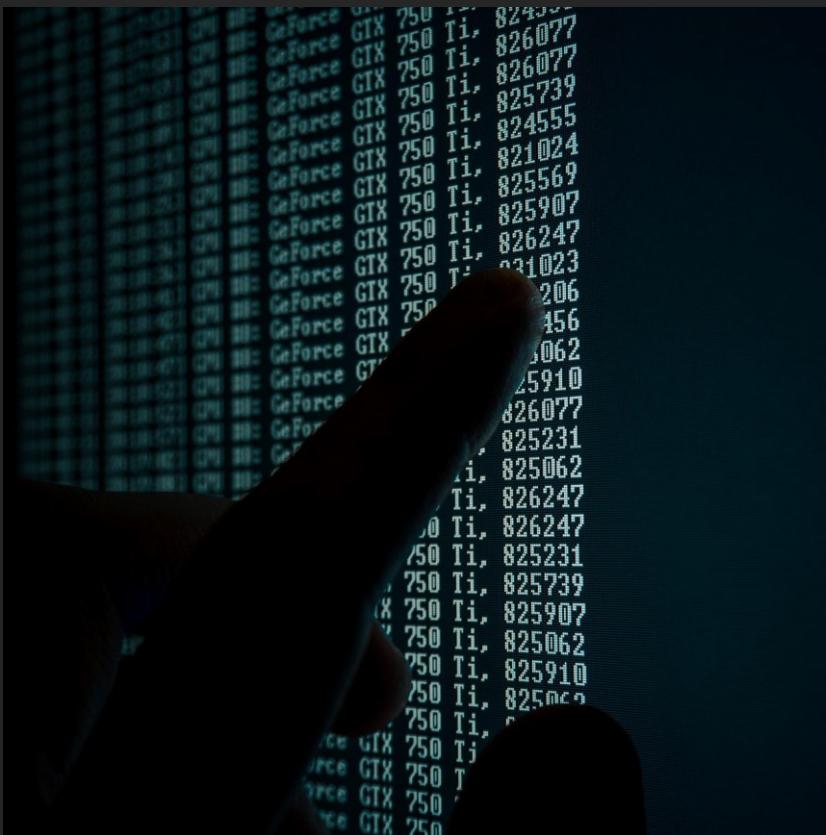
### 2.2 Data Dictionary



Col.No	Field Name	Description
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

## 2. Data Overview

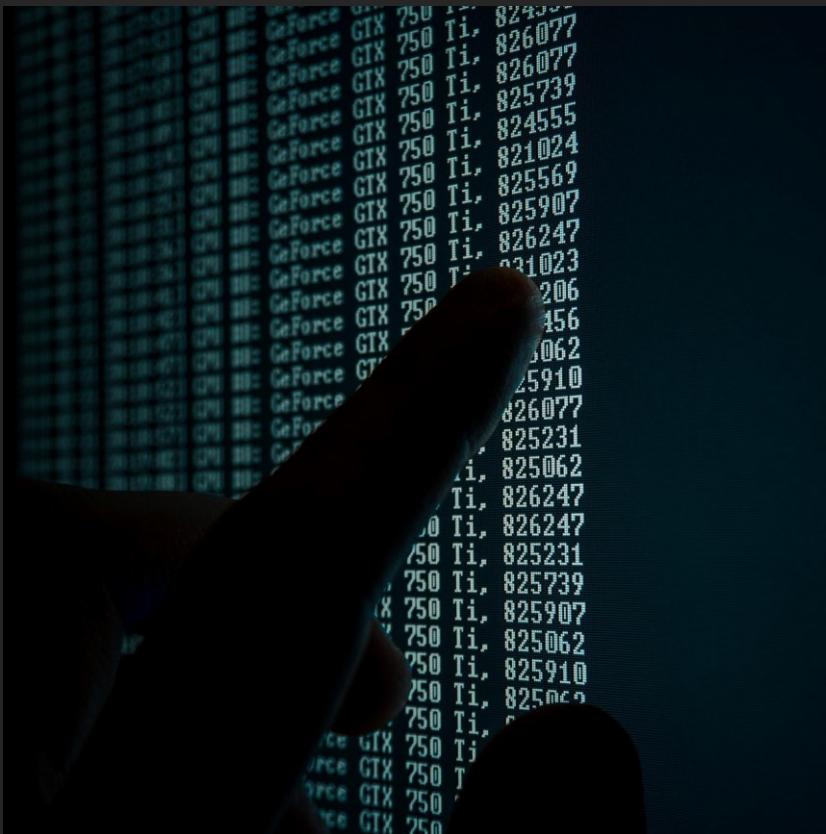
### 2.2 Data Dictionary



Col.No	Field Name	Description
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

## 2. Data Overview

### 2.2 Data Dictionary



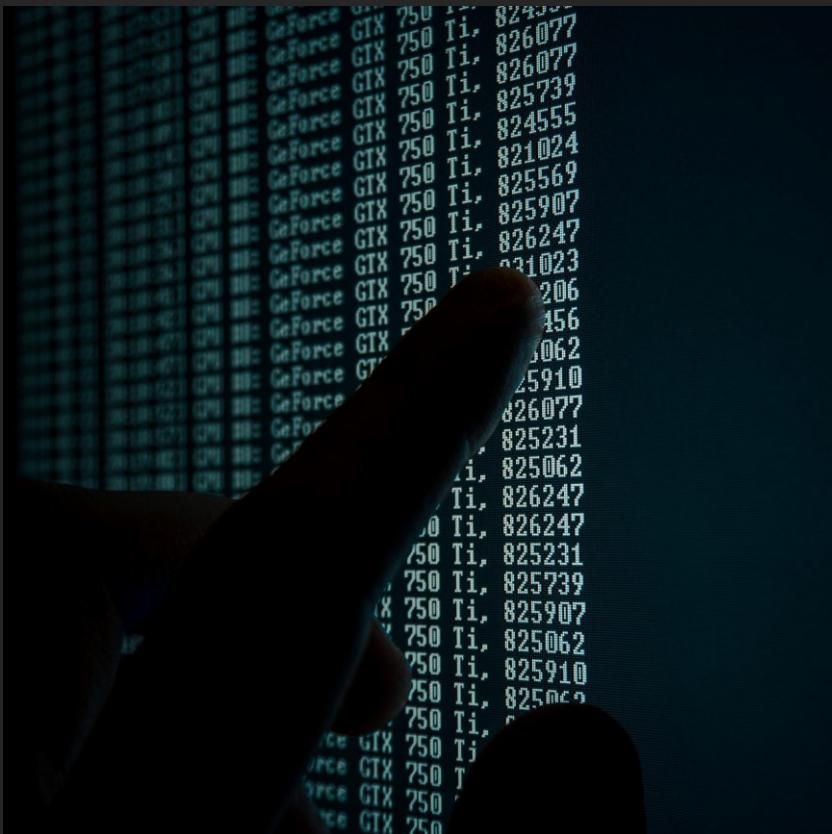
Col.No	Field Name	Description
64	Debtors Velocity (Days)	Average days required for receiving the payments
65	Creditors Velocity (Days)	Average number of days company takes to pay suppliers
66	Inventory Velocity (Days)	Average number of days the company needs to turn its inventory into sales
67	Value 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

There are total 68 variables in this dataset.

It contains various measures related to company business.

## 2. Data Overview

### 2.2 Data info



The number of rows (observations) is 3586.  
The number of columns (variables) is 67.

There are 63 float variables, 3 integer variables and 1 object variable.

There are no duplicates present in the dataset.

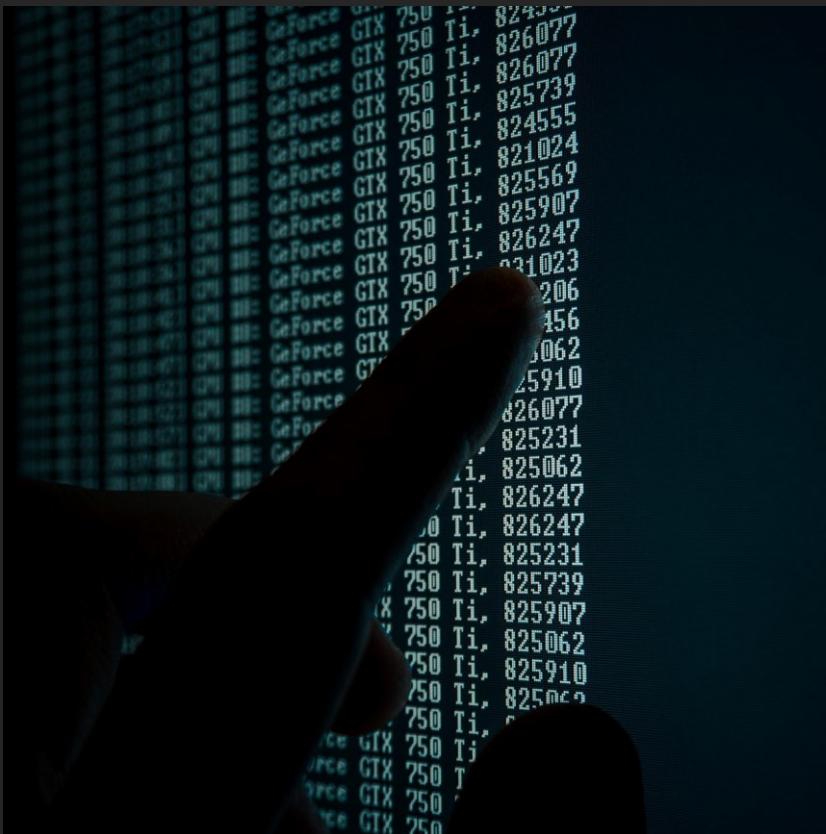
All the data has been renamed for ease as suggested in original dictionary dataset.

Networth\_next\_Year will be taken as 'default' column or output variable.

We are dropping the company code and company name for modelling purpose.

## 2. Data Overview

### 2.3 Statistical Summary



	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	Total	Gross Sales	Net Sales
count	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	3586.00	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	1778.45	1123.74	1079.70	
std	19776.82	4769.68	778.76	4091.99	26975.14	23652.84	4871.55	6301.22	22577.57	2675.00	11437.57	10603.70	9996.57	
min	4.00	-8021.60	0.00	-7027.48	-1824.75	-0.72	-41.19	-13162.42	-0.91	-0.23	-4.51	-62.59	-62.59	
25%	3029.25	3.98	3.75	3.89	7.60	0.03	0.57	0.94	4.00	0.73	10.56	1.44	1.44	
50%	6077.50	19.02	8.29	18.58	39.09	7.49	15.87	10.14	24.54	9.23	52.01	31.21	30.44	
75%	24269.50	123.80	19.52	117.30	226.60	72.35	131.90	61.18	135.28	65.65	310.54	242.25	234.44	
max	72493.00	111729.10	42263.46	81657.35	714001.25	652823.81	128477.59	223257.56	721166.00	83232.98	254737.22	474182.94	443775.16	

The statistical summary shows that there are many outliers present in the dataset in almost all the variables.

The median Networth is 18.58 units where as minimum is -7027 units and maximum is 81657 units. This shows that how much deviation is there which is also proved by the difference in mean and std. deviation.

ROG-Revenue earnings in forex and expense variable shows that the 25<sup>th</sup> & 75<sup>th</sup> percentile are 0 which means they have a large chunk of data having zero which might not be contributing enough towards the output and can be eliminated.

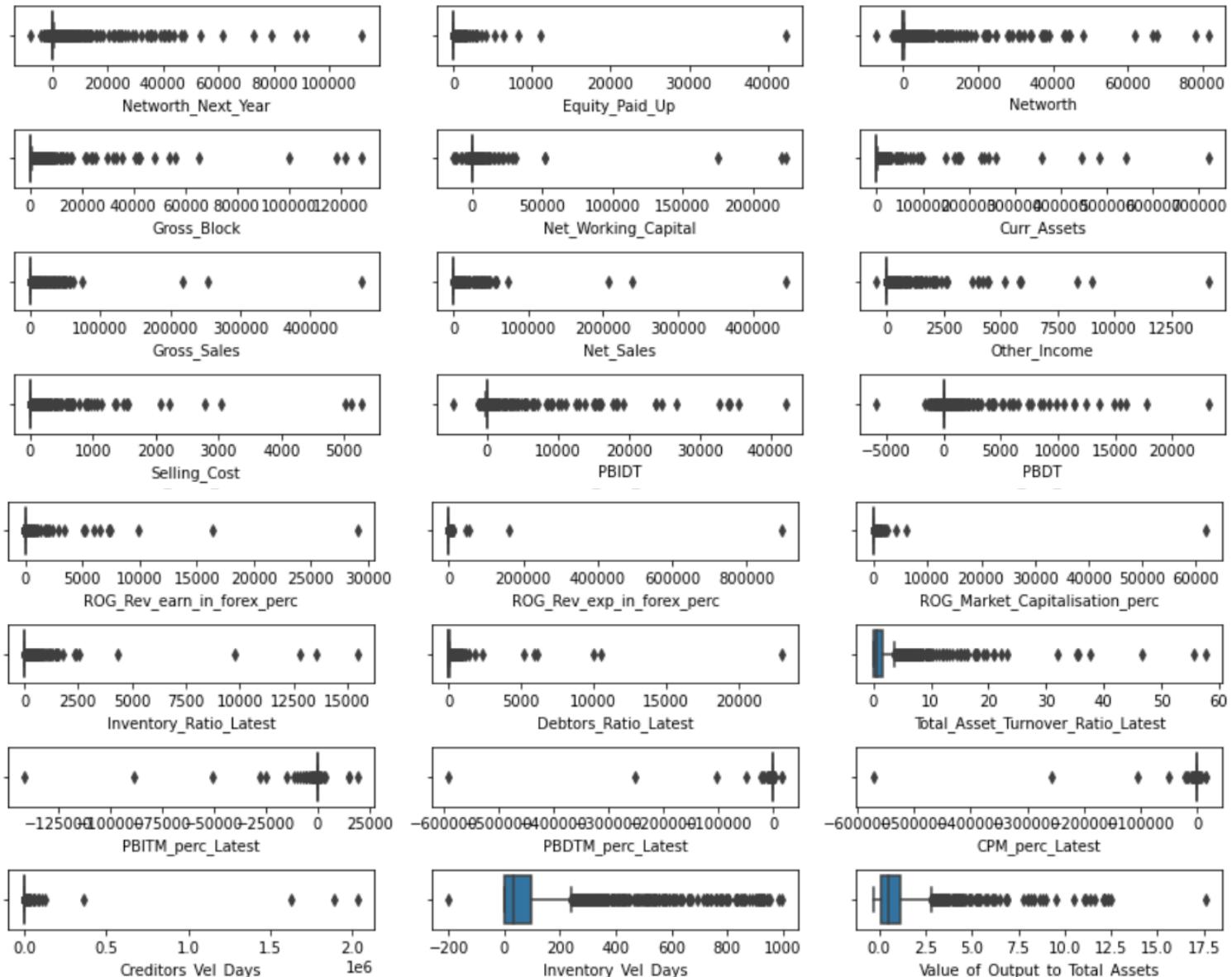
We will look further into the outliers to see how bad the situation is, whether any treatment is required or not.

# 3. Data Cleaning & Processing

## 3.1 Outlier Treatment



# Box Plots showing Outliers

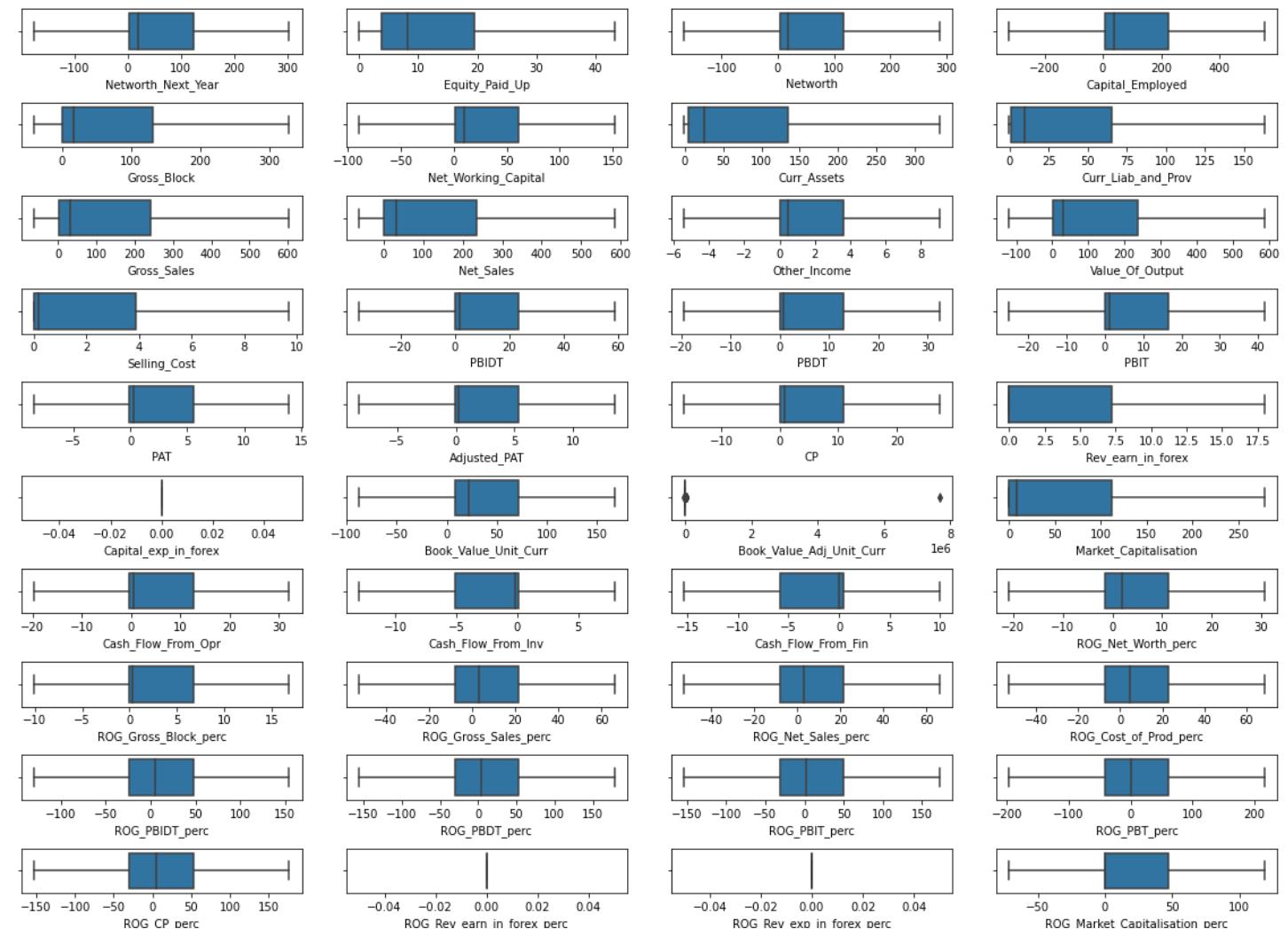


# 3. Data Cleaning & Processing

## 3.1 Outlier Treatment



# After Outlier Treatment



### 3. Data Cleaning & Processing

#### 3.2 Transform Target Variable



Target value 'Netwoth\_Next\_Year' is transform into 0 and 1 s.

Netwoth\_Next\_Year <0 (negative) then target variable = 1

Netwoth\_Next\_Year >=0 (positive) then target variable = 0

1 means 'Company might Default'

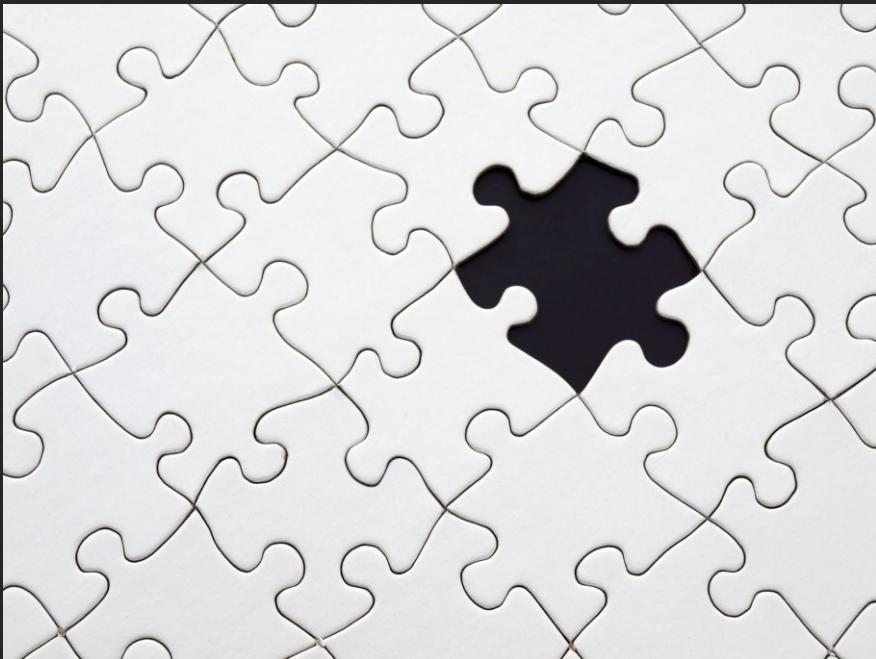
0 means 'Company might not Default'

Debtors_Vel_Days	Creditors_Vel_Days	Inventory_Vel_Days	Value_of_Output_to_Total_Assets	Value_of_Output_to_Gross_Block	default
0.0	0.0	45.0	0.00	0.00	1
29.0	101.0	2.0	0.31	0.24	1
97.0	210.5	0.0	-0.03	-0.26	1
93.0	63.0	2.0	0.24	1.90	1
253.0	210.5	0.0	0.01	0.05	1

Totals number of 0's are 3198 and total number of 1's are 388.

# 3. Data Cleaning & Processing

## 3.3 Missing Value Treatment



There are 118 missing values present in the dataset.

		missing values in percentage	
Inventory_Vel_Days	103	Inventory_Vel_Days	2.872281
Book_Value_Adj_Unit_Curr	4	Book_Value_Adj_Unit_Curr	0.111545
PBITM_perc_Latest	1	PBITM_perc_Latest	0.027886
Curr_Ratio_Latest	1	Curr_Ratio_Latest	0.027886
Fixed_Assets_Ratio_Latest	1	Fixed_Assets_Ratio_Latest	0.027886
Inventory_Ratio_Latest	1	Inventory_Ratio_Latest	0.027886
Debtors_Ratio_Latest	1	Debtors_Ratio_Latest	0.027886
Interest_Cover_Ratio_Latest	1	Interest_Cover_Ratio_Latest	0.027886
PBIDTM_perc_Latest	1	PBIDTM_perc_Latest	0.027886
Total_Asset_Turnover_Ratio_Latest	1	Total_Asset_Turnover_Ratio_Latest	0.027886
PBDTM_perc_Latest	1	PBDTM_perc_Latest	0.027886
CPM_perc_Latest	1	CPM_perc_Latest	0.027886
APATM_perc_Latest	1	APATM_perc_Latest	0.027886

The missing value is quite less in percentage, hence shouldn't be removed, rather imputed.

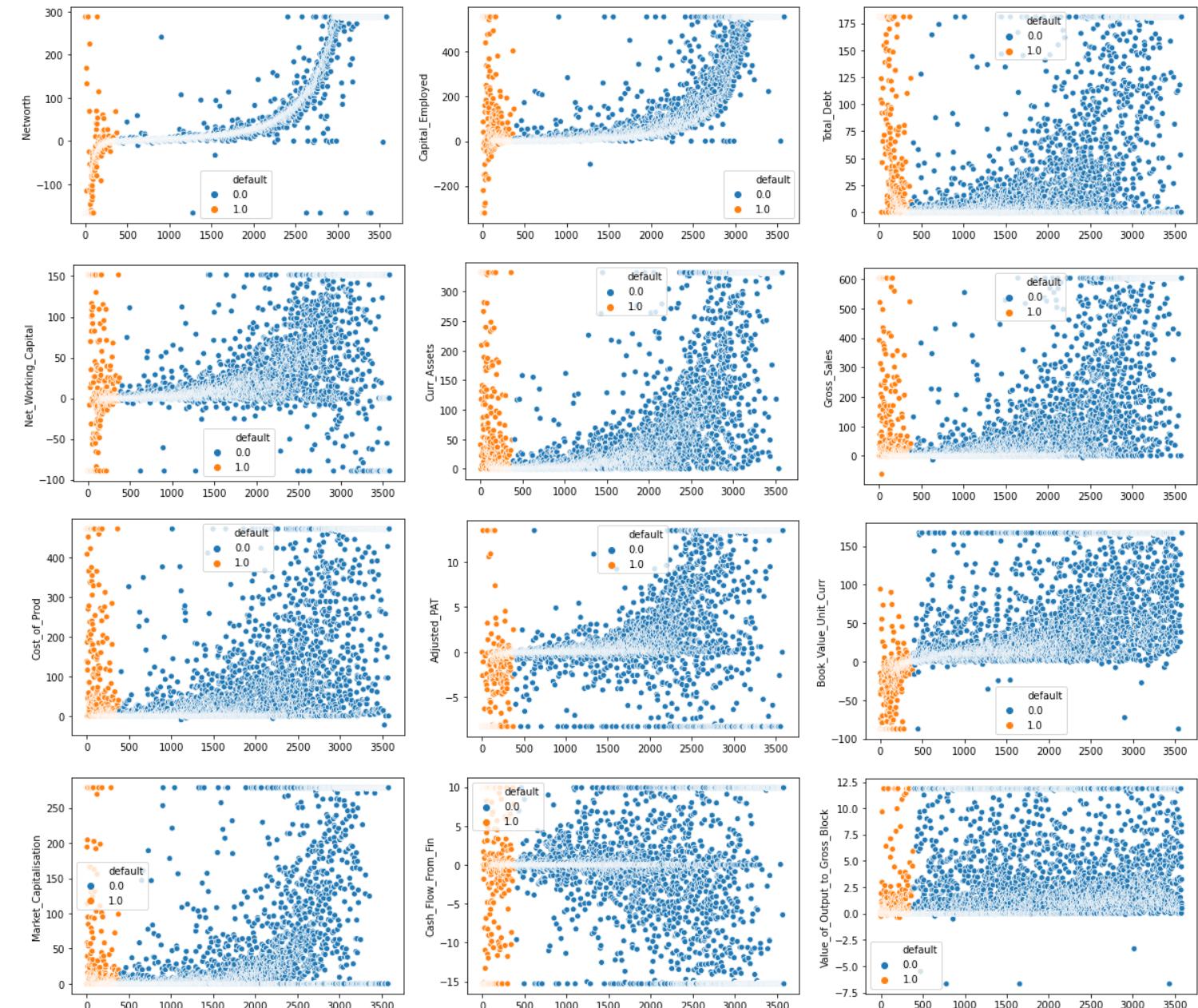
Inventory_Vel_Days	0
Book_Value_Adj_Unit_Curr	0
PBITM_perc_Latest	0
Curr_Ratio_Latest	0
Fixed_Assets_Ratio_Latest	0
Inventory_Ratio_Latest	0
Debtors_Ratio_Latest	0
Interest_Cover_Ratio_Latest	0
PBIDTM_perc_Latest	0
Total_Asset_Turnover_Ratio_Latest	0
PBDTM_perc_Latest	0
CPM_perc_Latest	0
APATM_perc_Latest	0

# 4. Data Analysis

## 4.1 Univariate Analysis



# Scatter Plot of few Variables

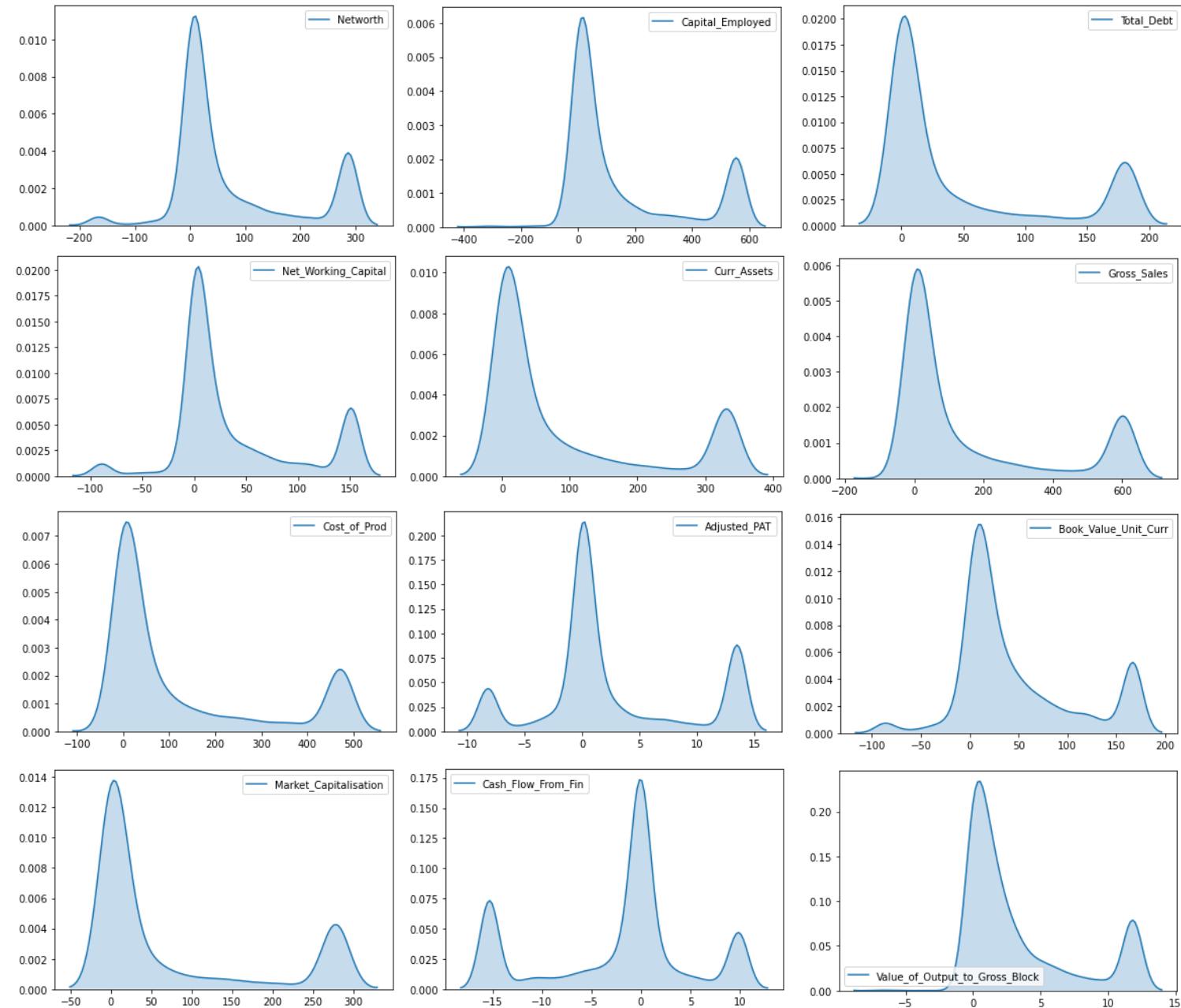


# 4. Data Analysis

## 4.1 Univariate Analysis



# Density Plot of few Variables



# 4. Data Analysis

## 4.1 Univariate Analysis

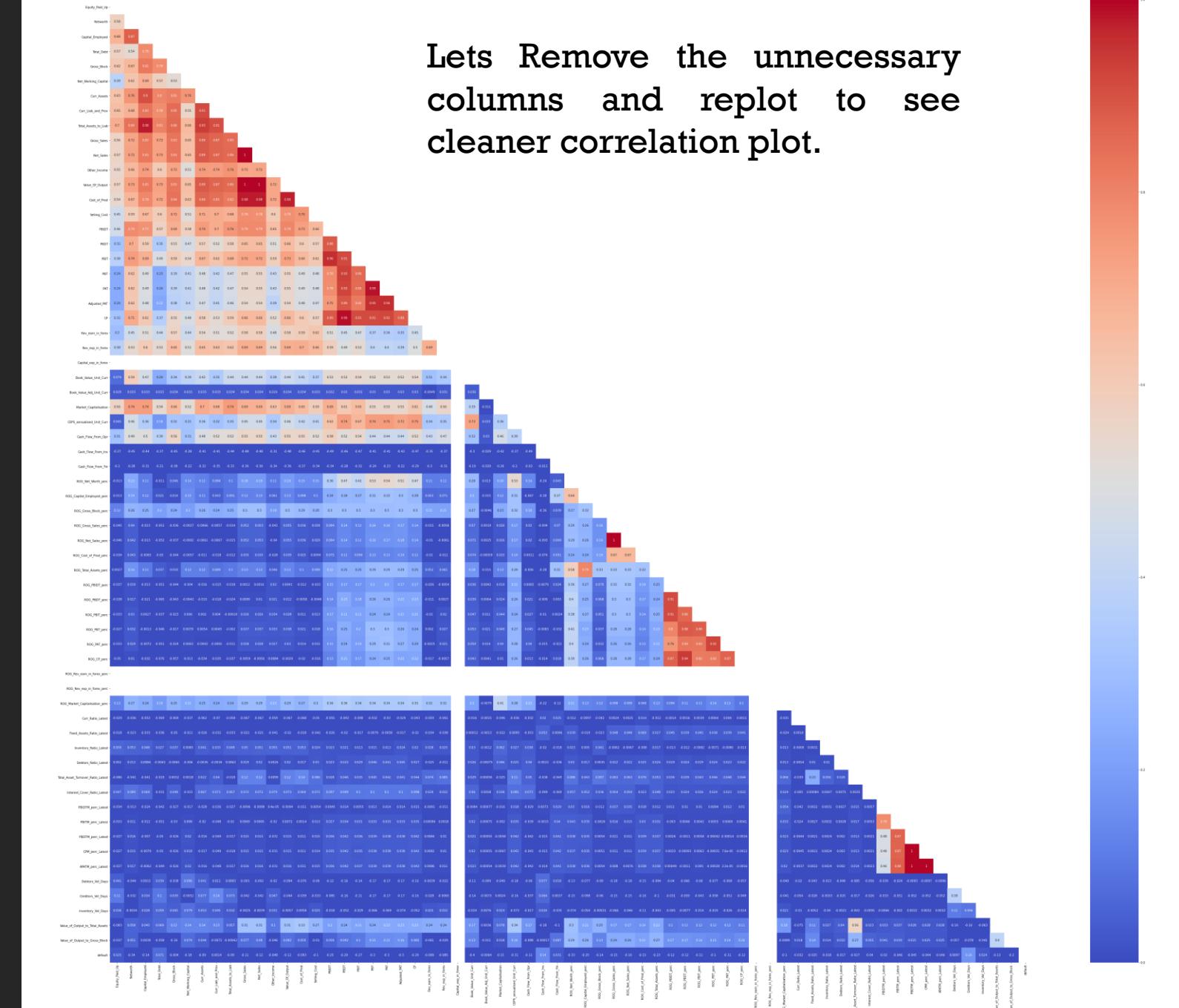


### # Histogram



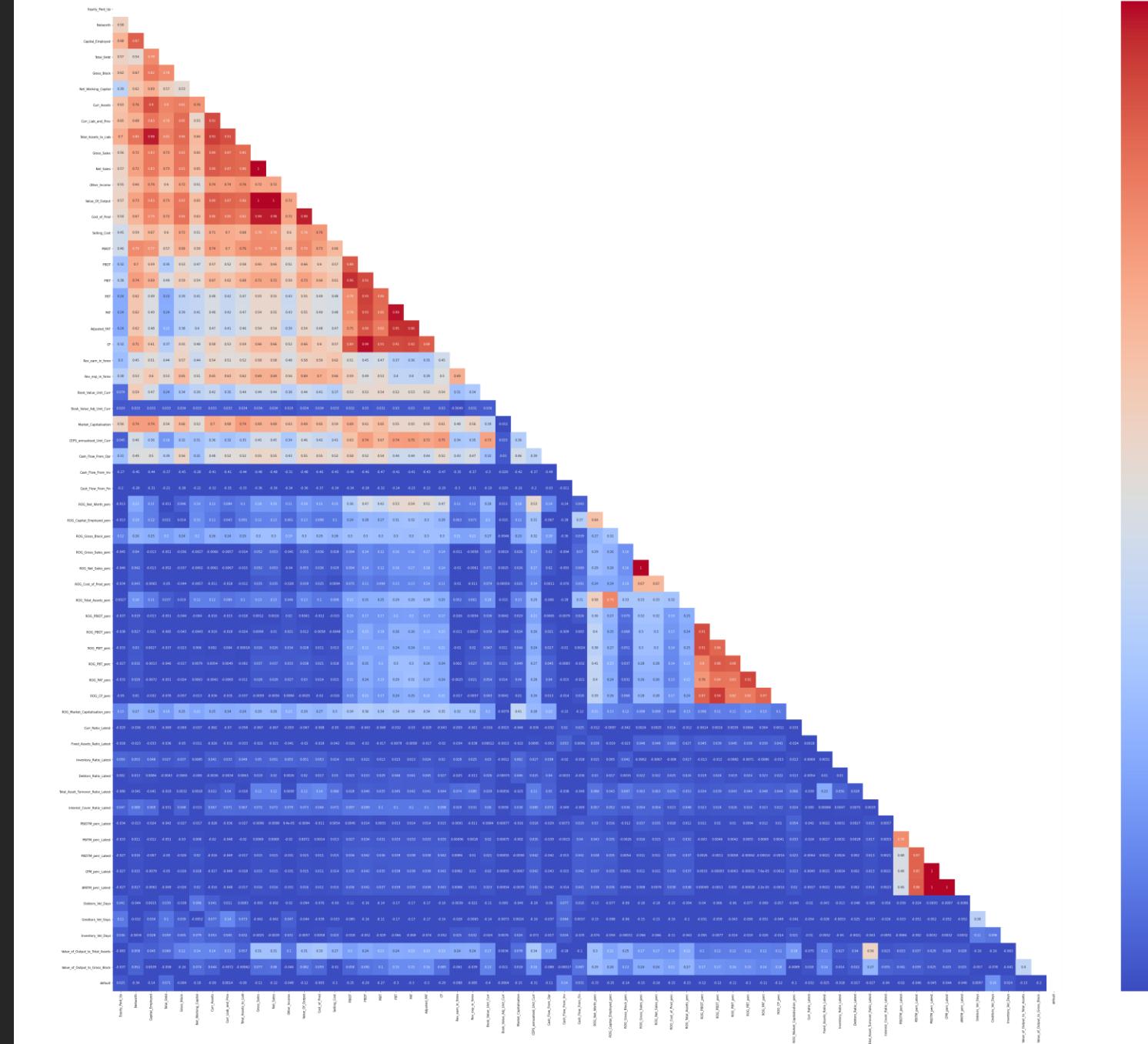
# 4. Data Analysis

## 4.2 Correlation Plot



# 4. Data Analysis

## 4.2 Correlation Plot



# 4. Data Analysis

## 4.2 Correlation Plot



Equity_Paid_Up -	Networth -	Capital_Employed -	Total_Debt -	Gross_Block -	Net_Working_Capital -	Curr_Assets -	Curr_Liab_and_Prov -	Total_Assets_to_Liab -	Gross_Sales -	Net_Sales -	Other_Income -	Value_Of_Output -	Cost_of_Prod -	Selling_Cost -	PBDT -	PBDT -	PBIT -	PAT -	Adjusted_PAT -	CP -	Rev_earn_in_forex -	Rev_exp_in_forex -	Book_Value_Unit_Curr -	Book_Value_Adj_Unit_Curr -	Market_Capitalisation -	CEPS_annualised_Unit_Curr -																				
	0.58																																													
Networth -		0.68	0.68																																											
Capital_Employed -				0.54	0.78																																									
Total_Debt -						0.62	0.67	0.82	0.78																																					
Gross_Block -							0.62	0.67	0.82	0.78																																				
Net_Working_Capital -								0.39	0.62	0.69	0.57	0.53																																		
Curr_Assets -									0.63	0.76	0.9	0.8	0.81	0.76																																
Curr_Liab_and_Prov -										0.65	0.68	0.83	0.78	0.85	0.55	0.91																														
Total_Assets_to_Liab -										0.7	0.84	0.98	0.81	0.86	0.66	0.93	0.91																													
Gross_Sales -											0.56	0.72	0.83	0.73	0.83	0.65	0.89	0.87	0.85																											
Net_Sales -											0.57	0.72	0.83	0.73	0.83	0.65	0.89	0.87	0.86	1																										
Other_Income -											0.55	0.66	0.74	0.6	0.72	0.51	0.74	0.74	0.76	0.72	0.72																									
Value_Of_Output -												0.57	0.73	0.83	0.73	0.83	0.65	0.89	0.87	0.86	1	1	0.72																							
Cost_of_Prod -													0.54	0.67	0.79	0.72	0.84	0.63	0.86	0.85	0.82	0.98	0.98	0.72	0.98																					
Selling_Cost -														0.45	0.59	0.67	0.6	0.72	0.51	0.71	0.7	0.68	0.78	0.78	0.6	0.78	0.76																			
PBDT -															0.46	0.79	0.77	0.57	0.69	0.58	0.74	0.7	0.76	0.79	0.79	0.65	0.79	0.73	0.66																	
PBDT -															0.32	0.7	0.59	0.35	0.53	0.47	0.57	0.52	0.58	0.65	0.65	0.51	0.66	0.6	0.57	0.88																
PBIT -																0.38	0.74	0.69	0.49	0.59	0.54	0.67	0.62	0.68	0.72	0.72	0.59	0.73	0.66	0.61	0.95	0.91	0.86													
PAT -																	0.24	0.62	0.49	0.23	0.39	0.41	0.48	0.42	0.47	0.55	0.55	0.43	0.55	0.49	0.48	0.78	0.93	0.85	0.99											
Adjusted_PAT -																		0.24	0.62	0.48	0.22	0.38	0.4	0.47	0.41	0.46	0.54	0.54	0.39	0.54	0.48	0.47	0.75	0.89	0.82	0.95	0.96									
CP -																			0.32	0.71	0.61	0.37	0.55	0.48	0.58	0.53	0.59	0.66	0.66	0.52	0.66	0.6	0.57	0.89	0.99	0.91	0.92	0.88								
Rev_earn_in_forex -																				0.3	0.45	0.51	0.44	0.57	0.44	0.51	0.52	0.58	0.58	0.58	0.62	0.51	0.45	0.47	0.37	0.36	0.35	0.45								
Rev_exp_in_forex -																				0.38	0.53	0.6	0.53	0.65	0.51	0.63	0.62	0.69	0.69	0.69	0.56	0.69	0.7	0.66	0.59	0.49	0.53	0.4	0.4	0.39	0.5	0.69				
Book_Value_Unit_Curr -																				0.074	0.59	0.47	0.24	0.34	0.39	0.42	0.35	0.44	0.44	0.44	0.38	0.44	0.41	0.37	0.53	0.53	0.54	0.52	0.54	0.31	0.34					
Book_Value_Adj_Unit_Curr -																				0.029	0.033	0.033	0.033	0.034	0.033	0.033	0.034	0.034	0.034	0.034	0.019	0.034	0.034	0.033	0.032	0.03	0.03	0.03	-0.0049	0.031	0.036					
Market_Capitalisation -																				0.56	0.74	0.74	0.54	0.66	0.52	0.7	0.68	0.74	0.69	0.69	0.63	0.69	0.65	0.59	0.69	0.61	0.65	0.55	0.55	0.61	0.48	0.56	0.39	0.033	0.36	
CEPS_annualised_Unit_Curr -																				0.045	0.46	0.36	0.18	0.32	0.31	0.36	0.32	0.35	0.45	0.45	0.34	0.46	0.42	0.41	0.63	0.74	0.67	0.74	0.75	0.72	0.75	0.34	0.35	0.73	0.033	0.36

A closer look to the variables which are highly positively correlated.

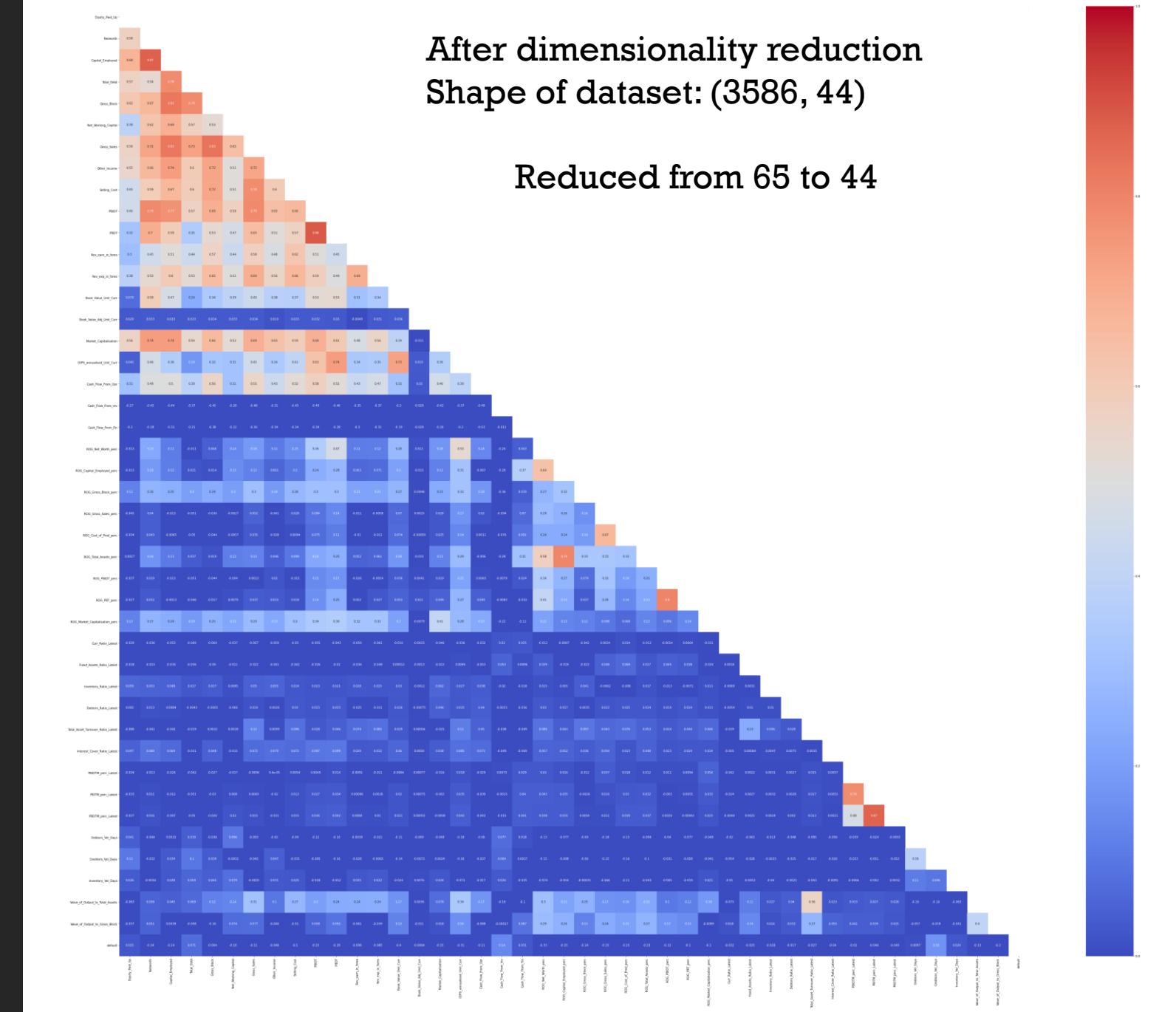
- Net\_sales is highly related with gross\_sales.
- Value\_of\_Ouput is related to gross\_sales and net\_sales
- Cost\_of\_prod is highly corelated with value\_of\_output.

The list of highly correlated variables with more than 90% of correlation

'Curr\_Assets', 'Curr\_Liab\_and\_Prov', 'Total\_Assets\_to\_Liab', 'Net\_Sales', 'Value\_Of\_Output', 'Cost\_of\_Prod', 'PBIT', 'PBT', 'PAT', 'Adjusted\_PAT', 'CP', 'ROG\_Net\_Sales\_perc', 'ROG\_PBDT\_perc', 'ROG\_PBIT\_perc', 'ROG\_PAT\_perc', 'ROG\_CP\_perc', 'CPM\_perc\_Latest', 'APATM\_perc\_Latest'

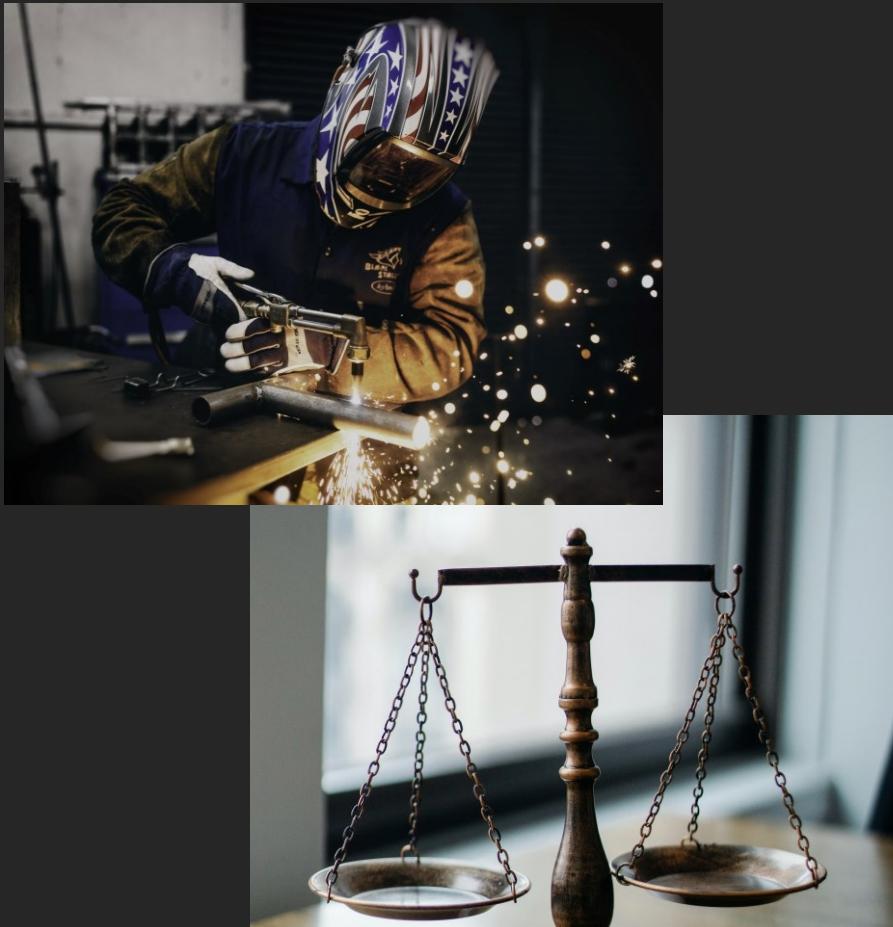
# 4. Data Analysis

## 4.2 Correlation Plot



# 5. Model Building

## 5.1 Target Variable & Scaling



- Output variable [y= default] is separated from the dataset which is renamed as X after separation.
- Now, scaling is required before proceeding with any model building as different variables are at different scale.
- Scaling was done using standard scaler. Glimpse of scaled dataset is as shown below.

	Equity_Paid_Up	Networth	Capital_Employed	Total_Debt	Gross_Block	Net_Working_Capital	Gross_Sales	Other_Income	Selling_Cost	PBDT	PBDT
0	2.08394	-2.124536	-2.277679	1.955653	2.011189	-2.126873	-0.711425	1.501295	-0.666143	-2.077966	-1.859258
1	2.08394	-2.124536	1.937167	1.955653	2.011189	-2.126873	1.980032	1.924923	1.942957	1.877612	-0.797237
2	2.08394	1.892586	1.937167	1.955653	2.011189	1.933358	1.037488	1.924923	1.942957	-2.077966	-1.859258
3	2.08394	-2.124536	1.937167	1.955653	2.011189	-2.126873	1.980032	1.924923	0.231669	-2.077966	-1.859258
4	2.08394	-2.124536	1.937167	1.955653	2.011189	1.933358	-0.538733	1.924923	-0.136595	-2.077966	-1.859258

- Dataset is cleaned from any NaN data.
- Train & Test dataset is created with ratio of 67:33 & random-state=42.
- Train has 43 columns and 2402 rows.
- Test has 43 columns and 1184 rows.
- 43 variables are a huge and can mess up our prediction. Hence we will do initial feature selection using recursive feature elimination and pick the best 15 features among rest.

# 5. Model Building

## 5.2 Recursive Feature Elimination using LR



	Feature	Rank
0	Equity_Paid_Up	28
1	Networth	1
2	Capital_Employed	1
3	Total_Debt	1
4	Gross_Block	20
5	Net_Working_Capital	4
6	Gross_Sales	10
7	Other_Income	8
8	Selling_Cost	1
9	PBIDT	1
10	PBDT	18
11	Rev_earn_in_forex	13
12	Rev_exp_in_forex	1
13	Book_Value_Unit_Curr	1
14	Book_Value_Adj_Unit_Curr	29
15	Market_Capitalisation	1
16	CEPS_annualised_Unit_Curr	1
17	Cash_Flow_From_Opr	26
18	Cash_Flow_From_Inv	14
19	Cash_Flow_From_Fin	27
20	ROG_Net_Worth_perc	1
21	ROG_Capital_Employed_perc	6
22	ROG_Gross_Block_perc	12
23	ROG_Gross_Sales_perc	19
24	ROG_Cost_of_Prod_perc	2
25	ROG_Total_Assets_perc	5
26	ROG_PBIDT_perc	24
27	ROG_PBT_perc	15
28	ROG_Market_Capitalisation_perc	21
29	Curr_Ratio_Latest	1
30	Fixed_Assets_Ratio_Latest	3
31	Inventory_Ratio_Latest	1
32	Debtors_Ratio_Latest	1
33	Total_Asset_Turnover_Ratio_Latest	17
34	Interest_Cover_Ratio_Latest	1
35	PBIDTM_perc_Latest	23
36	PBITM_perc_Latest	22
37	PBDTM_perc_Latest	9
38	Debtors_Vel_Days	7
39	Creditors_Vel_Days	16
40	Inventory_Vel_Days	25
41	Value_of_Output_to_Total_Assets	11
42	Value_of_Output_to_Gross_Block	1

Selected top 15 features by recursive elimination using Logistic Regression

	Feature	Rank
1	Networth	1
2	Capital_Employed	1
3	Total_Debt	1
8	Selling_Cost	1
9	PBIDT	1
12	Rev_exp_in_forex	1
13	Book_Value_Unit_Curr	1
15	Market_Capitalisation	1
16	CEPS_annualised_Unit_Curr	1
20	ROG_Net_Worth_perc	1
29	Curr_Ratio_Latest	1
31	Inventory_Ratio_Latest	1
32	Debtors_Ratio_Latest	1
34	Interest_Cover_Ratio_Latest	1
42	Value_of_Output_to_Gross_Block	1

# 5. Model Building

## 5.3 Modelling using Logistic Regression & Feature Selection



# Modelling Results from Model\_1 started with 15 features

Logit Regression Results

Dep. Variable:	default	No. Observations:	2402 <th data-cs="4" data-kind="parent"></th> <th data-kind="ghost"></th> <th data-kind="ghost"></th> <th data-kind="ghost"></th>				
Model:	Logit	Df Residuals:	2386				
Method:	MLE	Df Model:	15				
Date:	Mon, 06 Dec 2021	Pseudo R-squ.:	0.5980				
Time:	11:40:46	Log-Likelihood:	-331.01				
converged:	True	LL-Null:	-823.47				
Covariance Type:	nonrobust	LLR p-value:	2.301e-200				
	coef	std err	z	P> z	[0.025	0.975]	
Intercept	-7.5241	0.505	-14.887	0.000	-8.515	-6.533	
Networth	-0.6488	0.417	-1.557	0.119	-1.465	0.168	
Capital_Employed	-0.5778	0.566	-1.020	0.308	-1.688	0.532	
Total_Debt	1.3585	0.375	3.622	0.000	0.623	2.094	
Selling_Cost	-0.3243	0.278	-1.168	0.243	-0.868	0.220	
PBIDT	-0.5765	0.329	-1.752	0.080	-1.221	0.068	
Rev_exp_in_forex	0.3099	0.226	1.372	0.170	-0.133	0.753	
Book_Value_Unit_Curr	-6.0852	0.643	-9.460	0.000	-7.346	-4.824	
Market_Capitalisation	-0.5763	0.307	-1.880	0.060	-1.177	0.024	
CEPS_annualised_Unit_Curr	-0.4984	0.354	-1.406	0.160	-1.193	0.196	
ROG_Net_Worth_perc	-0.4011	0.132	-3.045	0.002	-0.659	-0.143	
Curr_Ratio_Latest	-0.6999	0.651	-1.074	0.283	-1.977	0.577	
Inventory_Ratio_Latest	-1.5284	1.127	-1.356	0.175	-3.738	0.681	
Debtors_Ratio_Latest	-1.1137	1.821	-0.612	0.541	-4.683	2.455	
Interest_Cover_Ratio_Latest	-0.4314	0.329	-1.312	0.190	-1.076	0.213	
Value_of_Output_to_Gross_Block	-0.4849	0.160	-3.027	0.002	-0.799	-0.171	

There are few insignificant variables that need to be eliminated further.

# 5. Model Building

## 5.3.1 VIF for Multicollinearity check



# using VIF for multi-collinearity check

	variables	VIF
1	Capital_Employed	10.551030
0	Networth	7.834030
4	PBIDT	4.958160
8	CEPS_annualised_Unit_Curr	3.867861
2	Total_Debt	3.726908
6	Book_Value_Unit_Curr	2.883421
7	Market_Capitalisation	2.637396
3	Selling_Cost	2.561793
5	Rev_exp_in_forex	2.035738
9	ROG_Net_Worth_perc	1.669261
14	Value_of_Output_to_Gross_Block	1.119559
13	Interest_Cover_Ratio_Latest	1.059009
12	Debtors_Ratio_Latest	1.013748
11	Inventory_Ratio_Latest	1.013310
10	Curr_Ratio_Latest	1.007535



Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

VIF should be below 5.  
'Capital\_Employed' has the highest VIF value above 7, hence should be eliminated and then rechecked.

	variables	VIF
3	PBIDT	4.937241
0	Networth	4.186332
7	CEPS_annualised_Unit_Curr	3.863714
5	Book_Value_Unit_Curr	2.882601
6	Market_Capitalisation	2.620175
2	Selling_Cost	2.558161
1	Total_Debt	2.056565
4	Rev_exp_in_forex	2.026963
8	ROG_Net_Worth_perc	1.652877
13	Value_of_Output_to_Gross_Block	1.118411
12	Interest_Cover_Ratio_Latest	1.049056
11	Debtors_Ratio_Latest	1.013583
10	Inventory_Ratio_Latest	1.013265
9	Curr_Ratio_Latest	1.007420

Now all the VIF values are under 5 and can be proceeded.

# 5. Model Building

## 5.3.2 Modelling using Logistic Regression & Feature Selection



### # Modelling Results from Model\_8

Logit Regression Results

Dep. Variable:	default	No. Observations:	2402
Model:	Logit	Df Residuals:	2393
Method:	MLE	Df Model:	8
Date:	Mon, 06 Dec 2021	Pseudo R-squ.:	0.5926
Time:	12:04:09	Log-Likelihood:	-335.46
converged:	True	LL-Null:	-823.47
Covariance Type:	nonrobust	LLR p-value:	2.242e-205

	coef	std err	z	P> z	[0.025	0.975]
Intercept	-7.3385	0.483	-15.186	0.000	-8.286	-6.391
Networth	-0.8121	0.354	-2.297	0.022	-1.505	-0.119
Total_Debt	1.1298	0.199	5.677	0.000	0.740	1.520
PBDT	-0.7743	0.286	-2.712	0.007	-1.334	-0.215
Book_Value_Unit_Curr	-6.2276	0.640	-9.732	0.000	-7.482	-4.973
Market_Capitalisation	-0.7470	0.286	-2.608	0.009	-1.308	-0.186
ROG_Net_Worth_perc	-0.4691	0.122	-3.852	0.000	-0.708	-0.230
Inventory_Ratio_Latest	-1.9615	1.283	-1.529	0.126	-4.475	0.552
Value_of_Output_to_Gross_Block	-0.4720	0.158	-2.990	0.003	-0.781	-0.163

**Features eliminated:**

Capital\_Employed >  
Debtors\_Ratio\_Latest >  
Curr\_Ratio\_Latest >  
Selling\_Cost >  
Rev\_exp\_in\_forex >  
CEPS\_annualised\_Unit\_Curr >  
Interest\_Cover\_Ratio\_Latest

There are still one insignificant variable (Inventory\_Ratio\_Latest) that needs to be eliminated.

# 5. Model Building

## 5.3.4 Final Model after Feature Selection



### # Modelling Results from Model\_9

Logit Regression Results

Dep. Variable:	default	No. Observations:	2402			
Model:	Logit	Df Residuals:	2394			
Method:	MLE	Df Model:	7			
Date:	Mon, 06 Dec 2021	Pseudo R-squ.:	0.5914			
Time:	12:05:31	Log-Likelihood:	-336.44			
converged:	True	LL-Null:	-823.47			
Covariance Type:	nonrobust	LLR p-value:	4.858e-206			
	coef	std err	z	P> z	[0.025	0.975]
Intercept	-7.2051	0.466	-15.468	0.000	-8.118	-6.292
Networth	-0.7906	0.351	-2.250	0.024	-1.479	-0.102
Total_Debt	1.1148	0.198	5.620	0.000	0.726	1.504
PBIDT	-0.7480	0.285	-2.625	0.009	-1.306	-0.190
Book_Value_Unit_Curr	-6.2362	0.639	-9.761	0.000	-7.488	-4.984
Market_Capitalisation	-0.7612	0.284	-2.680	0.007	-1.318	-0.205
ROG_Net_Worth_perc	-0.4728	0.121	-3.894	0.000	-0.711	-0.235
Value_of_Output_to_Gross_Block	-0.4841	0.157	-3.080	0.002	-0.792	-0.176

All the p values are now in acceptable range. Hence, we can finalise here with model 9 and proceed further with the prediction and validation.

# 6. Prediction & Validation

## 6.1 Prediction



### # Prediction on Train & Test dataset

Predicted Train dataset	842	0.0	251	1.0	Predicted Test dataset
	1057	0.0	3493	0.0	
	1595	0.0	3063	0.0	
	100	1.0	2384	0.0	
	1191	0.0	1679	0.0	
	2163	0.0	604	0.0	
	2763	0.0	3434	0.0	
	2701	0.0	2244	0.0	
	2072	0.0	2523	0.0	
	2349	0.0	2162	0.0	
	1392	0.0	3102	0.0	
	1621	0.0	1638	0.0	
	1960	0.0	2046	0.0	
	2148	0.0	1241	0.0	
	571	0.0	133	0.0	
	1984	0.0	2294	0.0	
	1592	0.0	2139	0.0	
	3110	0.0	2844	0.0	
	1564	0.0	1360	0.0	

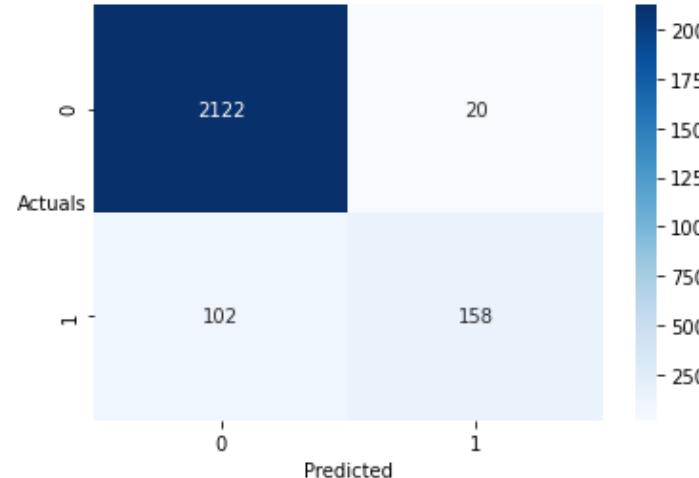
# 6. Prediction & Validation

## 6.2.1 Validation (Thres=0.5)

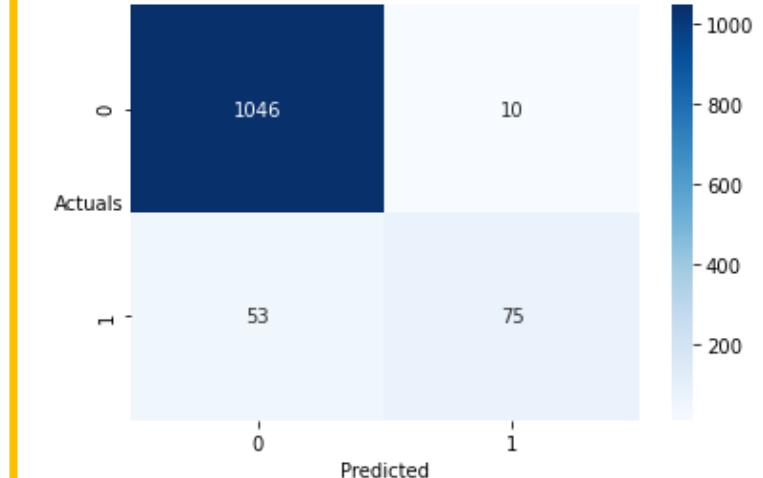


# Confusion Matrix & Classification Report  
for Train & Test dataset (with threshold=0.5)

Predicted Train dataset



Predicted Test dataset



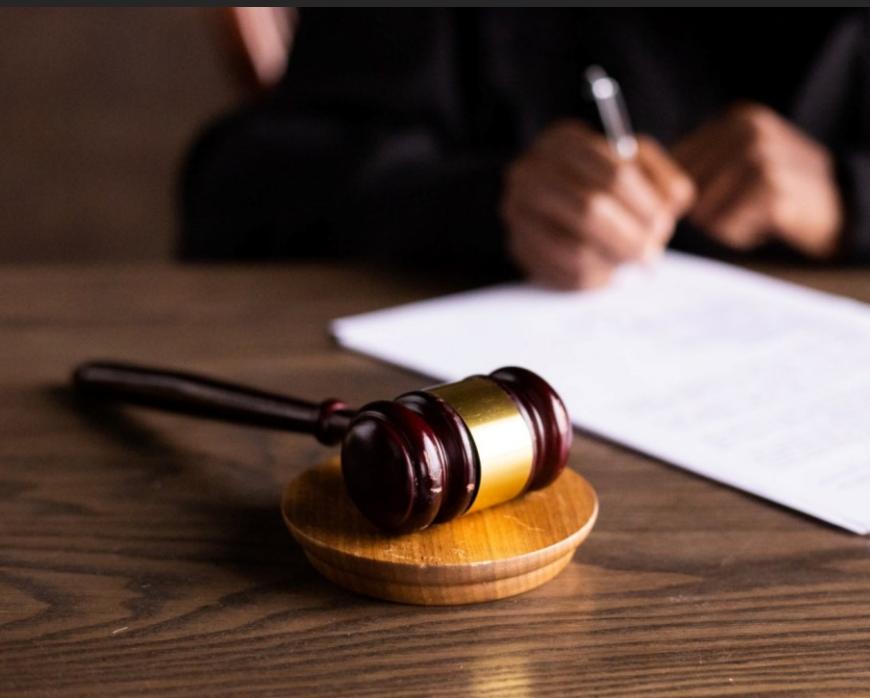
	precision	recall	f1-score	support
0.0	0.954	0.991	0.972	2142
1.0	0.888	0.608	0.721	260
accuracy			0.949	2402
macro avg	0.921	0.799	0.847	2402
weighted avg	0.947	0.949	0.945	2402

	precision	recall	f1-score	support
0.0	0.952	0.991	0.971	1056
1.0	0.882	0.586	0.704	128
accuracy			0.947	1184
macro avg	0.917	0.788	0.837	1184
weighted avg	0.944	0.947	0.942	1184

Precision to measure default on test dataset is 88% with recall of 58% and accuracy of 94%.

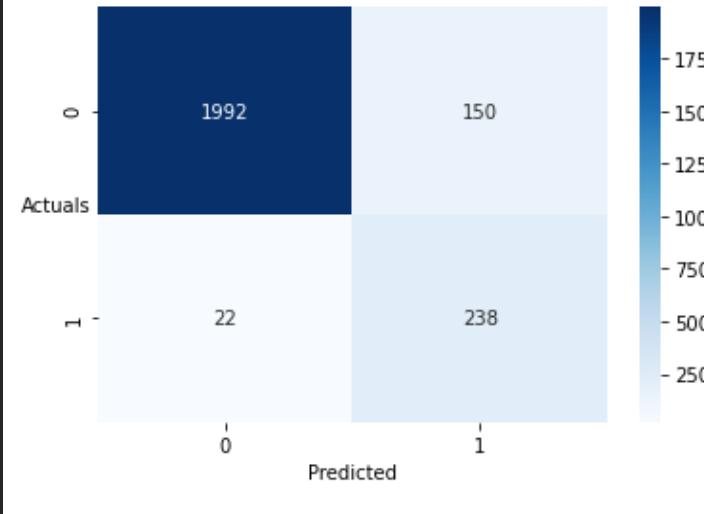
## 6. Prediction & Validation

### 6.2.2 Validation (Thres= 0.168)

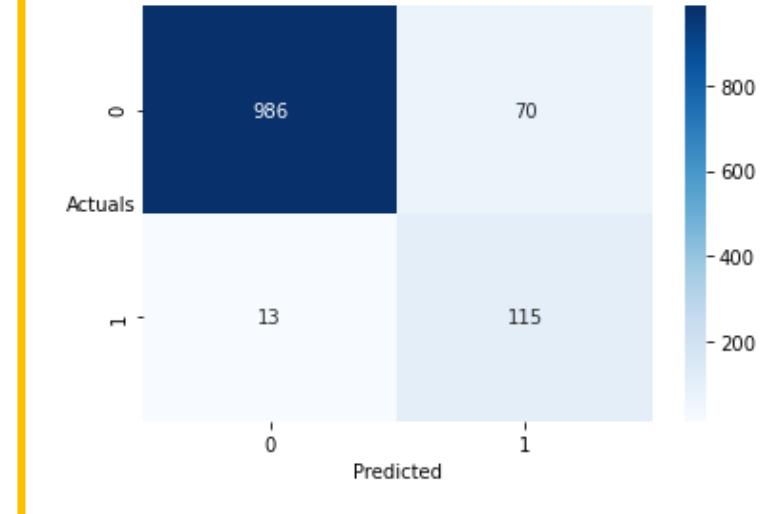


# Confusion Matrix & Classification Report  
for Train & Test dataset (with threshold=0.1689)

Predicted Train dataset



Predicted Test dataset



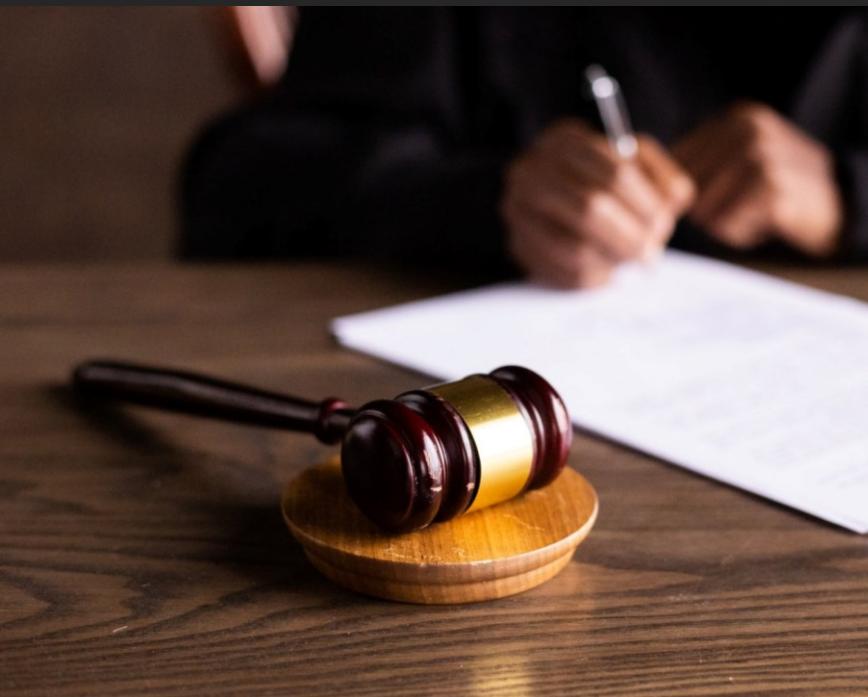
	precision	recall	f1-score	support
0.0	0.989	0.930	0.959	2142
1.0	0.613	0.915	0.735	260
accuracy			0.928	2402
macro avg	0.801	0.923	0.847	2402
weighted avg	0.948	0.928	0.934	2402

	precision	recall	f1-score	support
0.0	0.987	0.934	0.960	1056
1.0	0.622	0.898	0.735	128
accuracy			0.930	1184
macro avg	0.804	0.916	0.847	1184
weighted avg	0.947	0.930	0.935	1184

Precision to measure default on test dataset is 62% with recall of 90% and accuracy of 93%.

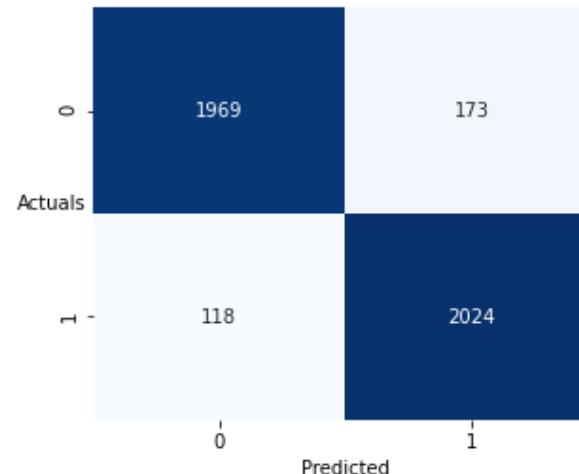
# 6. Prediction & Validation

## 6.2.3 Validation (SMOTE)

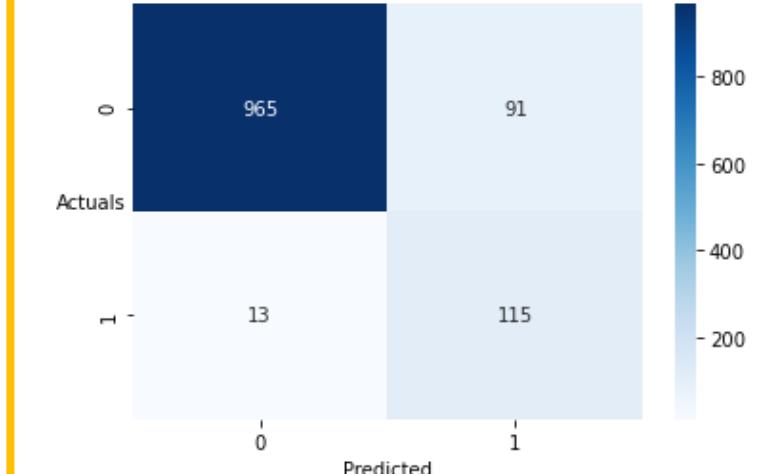


# Confusion Matrix & Classification Report  
for Train & Test dataset (using SMOTE)

Predicted Train dataset



Predicted Test dataset

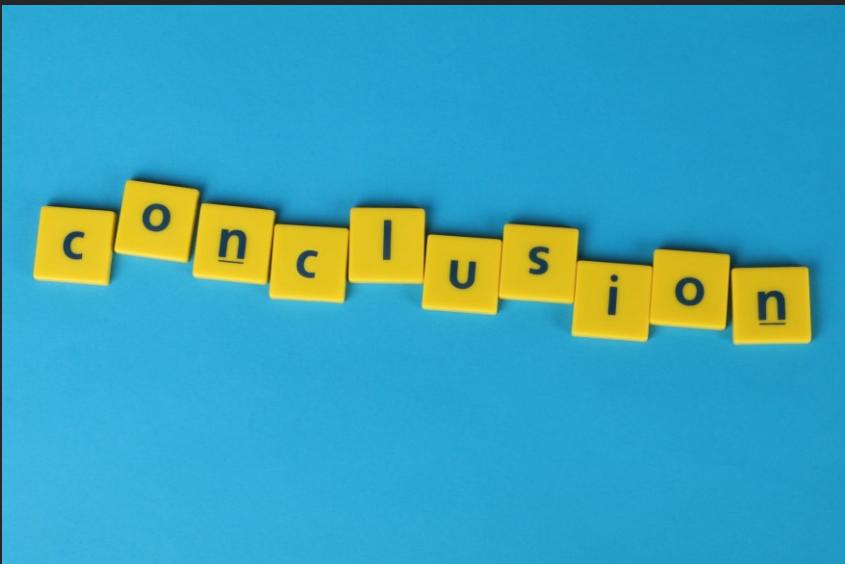


	precision	recall	f1-score	support
0.0	0.94	0.92	0.93	2142
1.0	0.92	0.94	0.93	2142
accuracy			0.93	4284
macro avg	0.93	0.93	0.93	4284
weighted avg	0.93	0.93	0.93	4284

	precision	recall	f1-score	support
0.0	0.0	0.99	0.91	1056
1.0	0.56	0.90	0.69	128
accuracy			0.91	1184
macro avg	0.77	0.91	0.82	1184
weighted avg	0.94	0.91	0.92	1184

Precision to measure default on test dataset is 56% with recall of 90% and accuracy of 91%.

## 7. Conclusion



### # Conclusion

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