Financial Risk Analytics Project(credit-risk)

by Anisha Sharma

PGPDSBA.O.Mar23.A

Great Learning

Problem 1: 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.

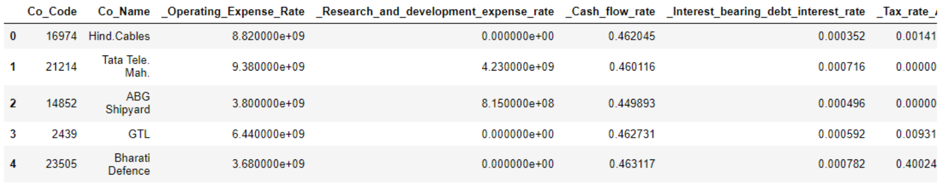
Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

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

Exploratory data analysis

- Dataset has 58 variables of which 53 are of float data type, 4 are integer type and 1 is object type.

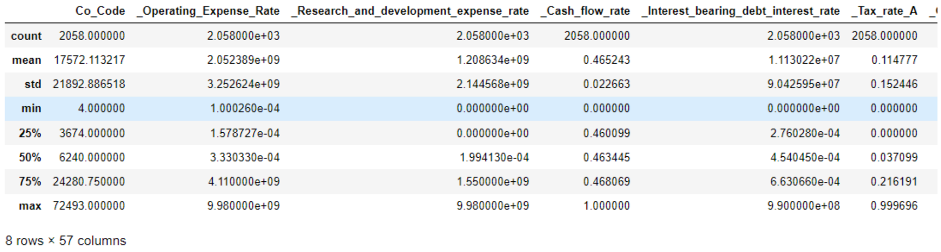
The head of the dataset is as below:



	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_
0	16974	Hind.Cables	8.820000e+09	0.000000e+00	0.462045	0.000352	0.00141
1	21214	Tata Tele. Mah.	9.380000e+09	4.230000e+09	0.460116	0.000716	0.00000
2	14852	ABG Shipyard	3.800000e+09	8.150000e+08	0.449893	0.000496	0.00000
3	2439	GTL	6.440000e+09	0.000000e+00	0.462731	0.000592	0.00931
4	23505	Bharati Defence	3.680000e+09	0.000000e+00	0.463117	0.000782	0.40024

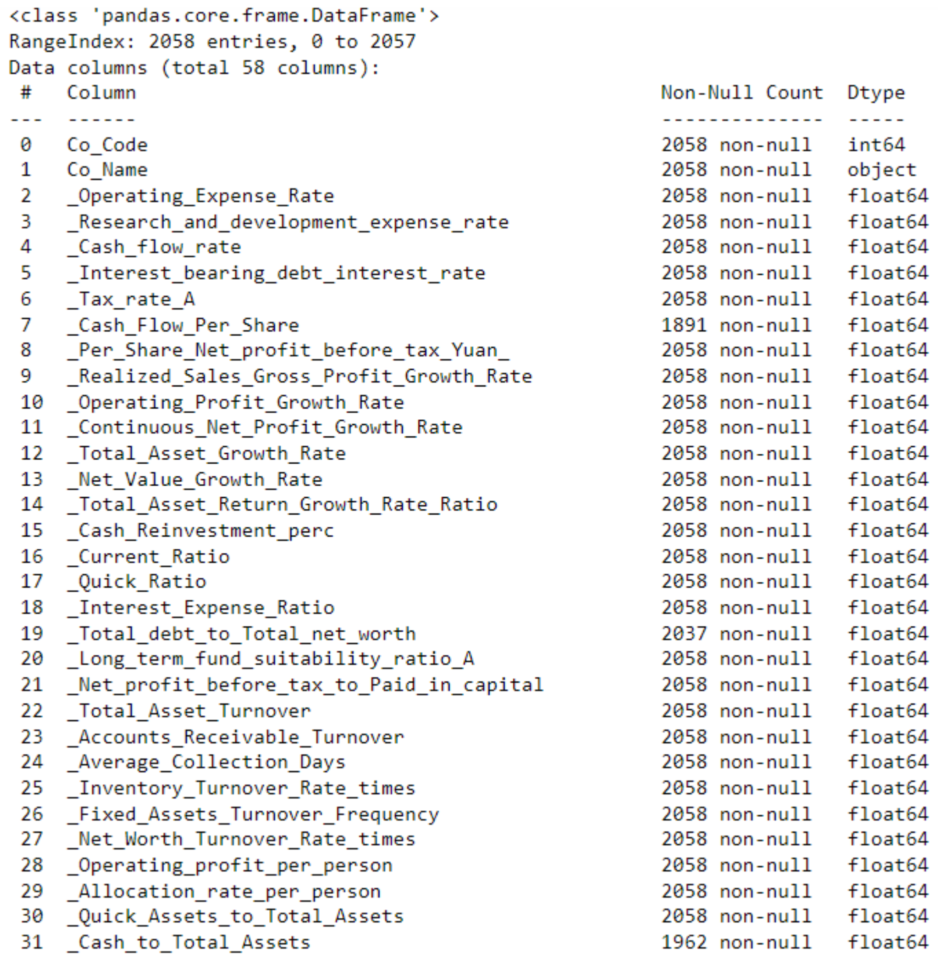
- The data has 2058 rows and 58 columns.
- No duplicate data is present in the dataset.
- There are 298 null values present in the dataset.
- We remove unwanted variables ‘Co_Code’ and ‘Co_Name’ since it does not add value to analysis.

Discriptive statistics

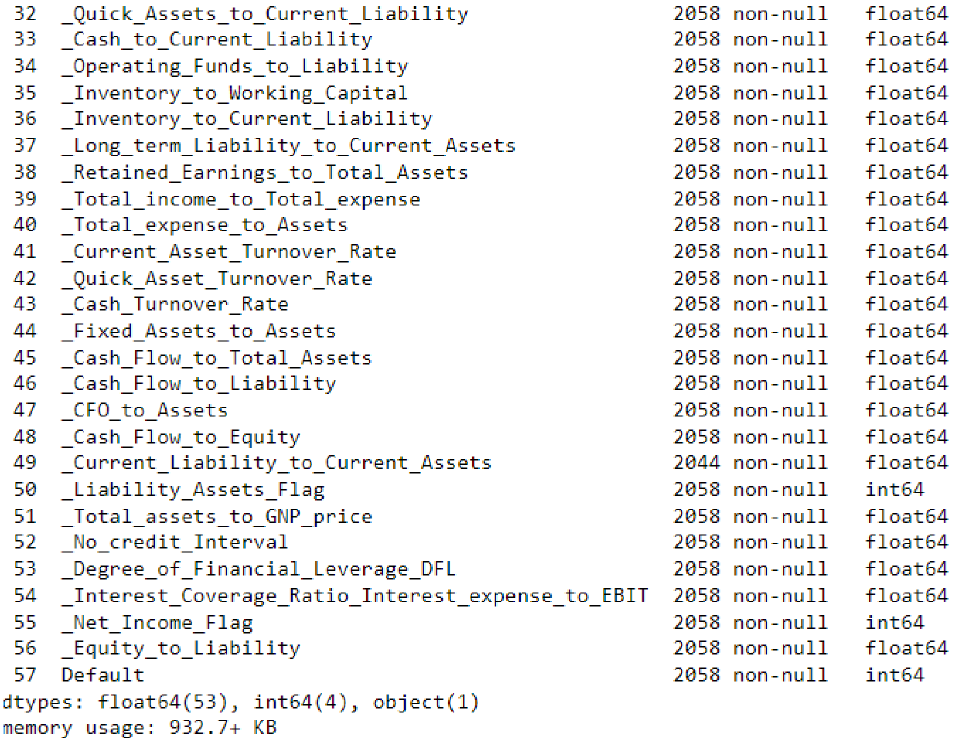


	Co_Code	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A	_
count	2058.000000	2.058000e+03	2.058000e+03	2058.000000	2.058000e+03	2058.000000	
mean	17572.113217	2.052389e+09	1.208634e+09	0.465243	1.113022e+07	0.114777	
std	21892.886518	3.252624e+09	2.144568e+09	0.022663	9.042595e+07	0.152446	
min	4.000000	1.000260e-04	0.000000e+00	0.000000	0.000000e+00	0.000000	
25%	3674.000000	1.578727e-04	0.000000e+00	0.460099	2.760280e-04	0.000000	
50%	6240.000000	3.330330e-04	1.994130e-04	0.463445	4.540450e-04	0.037099	
75%	24280.750000	4.110000e+09	1.550000e+09	0.468069	6.630660e-04	0.216191	
max	72493.000000	9.980000e+09	9.980000e+09	1.000000	9.900000e+08	0.999696	

8 rows × 57 columns



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2058 entries, 0 to 2057
Data columns (total 58 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Co_Code                                         2058 non-null   int64  
 1   Co_Name                                         2058 non-null   object 
 2   _Operating_Expense_Rate                         2058 non-null   float64
 3   _Research_and_development_expense_rate          2058 non-null   float64
 4   _Cash_flow_rate                                 2058 non-null   float64
 5   _Interest_bearing_debt_interest_rate            2058 non-null   float64
 6   _Tax_rate_A                                     2058 non-null   float64
 7   _Cash_Flow_Per_Share                            1891 non-null   float64
 8   _Per_Share_Net_profit_before_tax_Yuan_          2058 non-null   float64
 9   _Realized_Sales_Gross_Profit_Growth_Rate        2058 non-null   float64
 10  _Operating_Profit_Growth_Rate                   2058 non-null   float64
 11  _Continuous_Net_Profit_Growth_Rate              2058 non-null   float64
 12  _Total_Asset_Growth_Rate                        2058 non-null   float64
 13  _Net_Value_Growth_Rate                          2058 non-null   float64
 14  _Total_Asset_Return_Growth_Rate_Ratio           2058 non-null   float64
 15  _Cash_Reinvestment_perc                         2058 non-null   float64
 16  _Current_Ratio                                  2058 non-null   float64
 17  _Quick_Ratio                                    2058 non-null   float64
 18  _Interest_Expense_Ratio                         2058 non-null   float64
 19  _Total_debt_to_Total_net_worth                  2037 non-null   float64
 20  _Long_term_fund_suitability_ratio_A             2058 non-null   float64
 21  _Net_profit_before_tax_to_Paid_in_capital       2058 non-null   float64
 22  _Total_Asset_Turnover                           2058 non-null   float64
 23  _Accounts_Receivable_Turnover                   2058 non-null   float64
 24  _Average_Collection_Days                        2058 non-null   float64
 25  _Inventory_Turnover_Rate_times                  2058 non-null   float64
 26  _Fixed_Assets_Turnover_Frequency                2058 non-null   float64
 27  _Net_Worth_Turnover_Rate_times                  2058 non-null   float64
 28  _Operating_profit_per_person                    2058 non-null   float64
 29  _Allocation_rate_per_person                     2058 non-null   float64
 30  _Quick_Assets_to_Total_Assets                   2058 non-null   float64
 31  _Cash_to_Total_Assets                           1962 non-null   float64
```

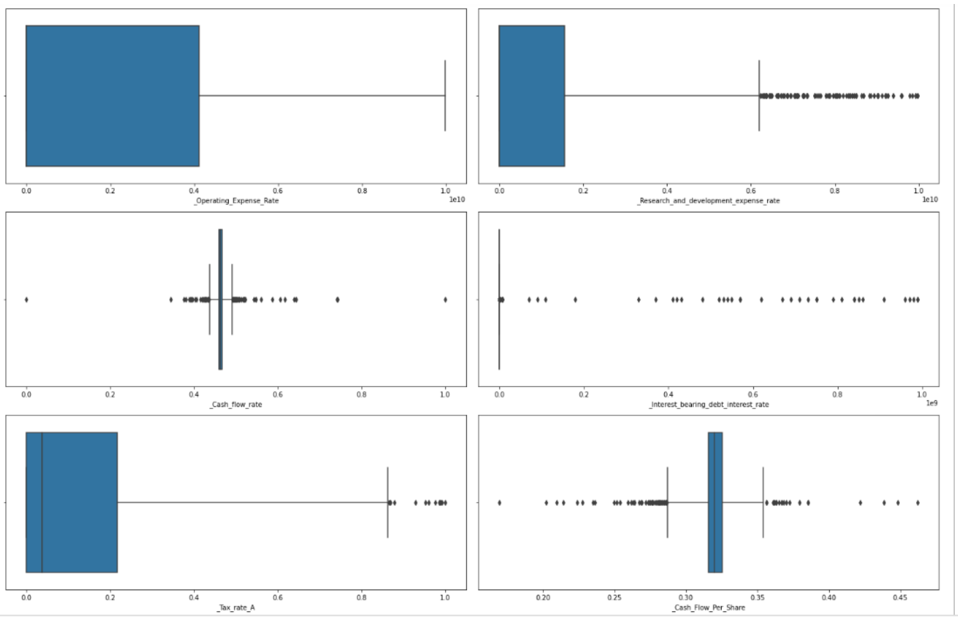


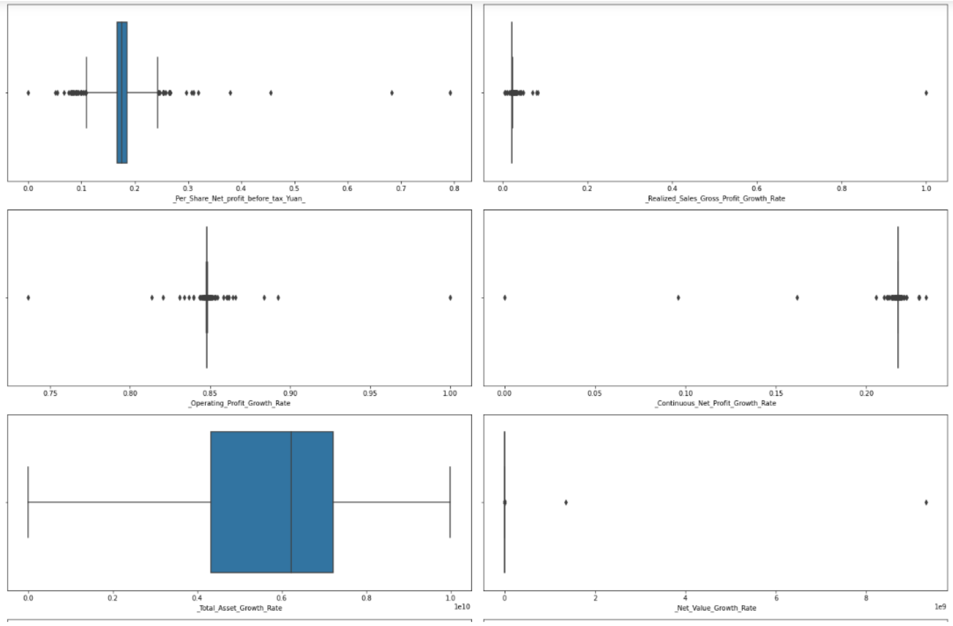
```
32  _Quick_Assets_to_Current_Liability               2058 non-null   float64
33  _Cash_to_Current_Liability                       2058 non-null   float64
34  _Operating_Funds_to_Liability                    2058 non-null   float64
35  _Inventory_to_Working_Capital                    2058 non-null   float64
36  _Inventory_to_Current_Liability                  2058 non-null   float64
37  _Long_term_Liability_to_Current_Assets           2058 non-null   float64
38  _Retained_Earnings_to_Total_Assets               2058 non-null   float64
39  _Total_income_to_Total_expense                   2058 non-null   float64
40  _Total_expense_to_Assets                         2058 non-null   float64
41  _Current_Asset_Turnover_Rate                     2058 non-null   float64
42  _Quick_Asset_Turnover_Rate                       2058 non-null   float64
43  _Cash_Turnover_Rate                              2058 non-null   float64
44  _Fixed_Assets_to_Assets                          2058 non-null   float64
45  _Cash_Flow_to_Total_Assets                       2058 non-null   float64
46  _Cash_Flow_to_Liability                          2058 non-null   float64
47  _CFO_to_Assets                                   2058 non-null   float64
48  _Cash_Flow_to_Equity                             2058 non-null   float64
49  _Current_Liability_to_Current_Assets             2044 non-null   float64
50  _Liability_Assets_Flag                           2058 non-null   int64  
51  _Total_assets_to_GNP_price                       2058 non-null   float64
52  _No_credit_Interval                              2058 non-null   float64
53  _Degree_of_Financial_Leverage_DFL                2058 non-null   float64
54  _Interest_Coverage_Ratio_Interest_expense_to_EBIT 2058 non-null  float64
55  _Net_Income_Flag                                 2058 non-null   int64  
56  _Equity_to_Liability                             2058 non-null   float64
57  Default                                          2058 non-null   int64  
dtypes: float64(53), int64(4), object(1)
memory usage: 932.7+ KB
```

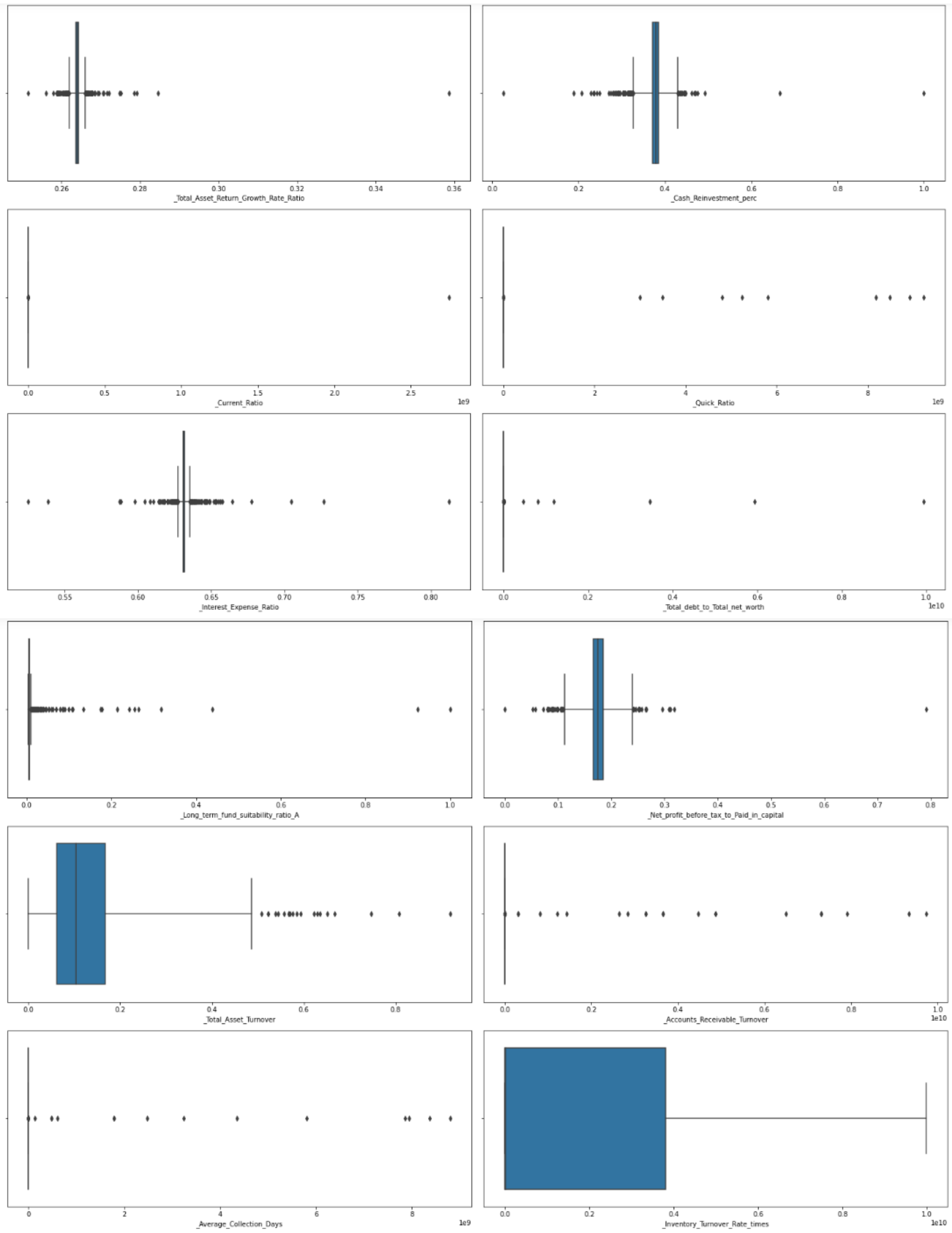
Problem 1: Outlier treatment

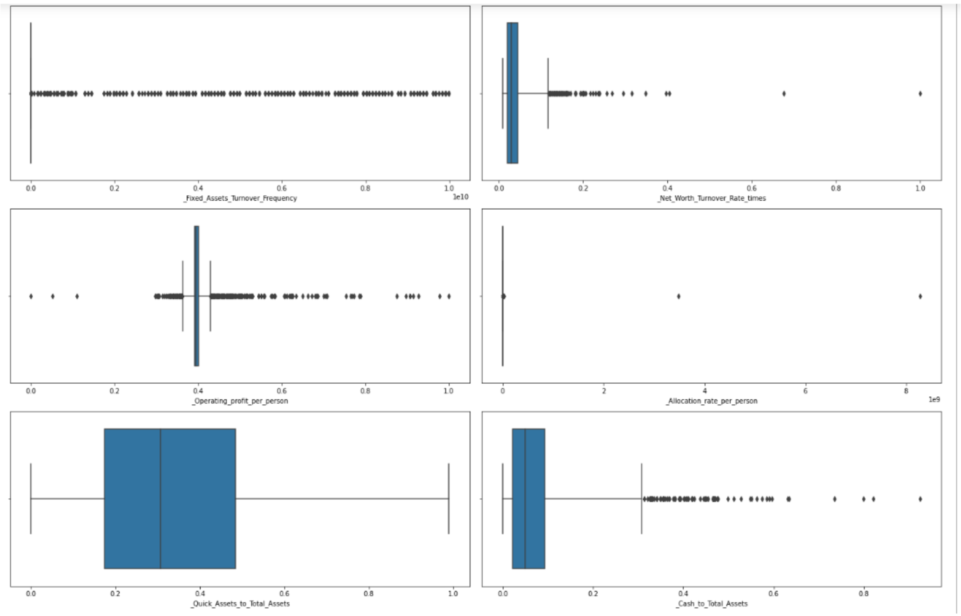
Describing the data:

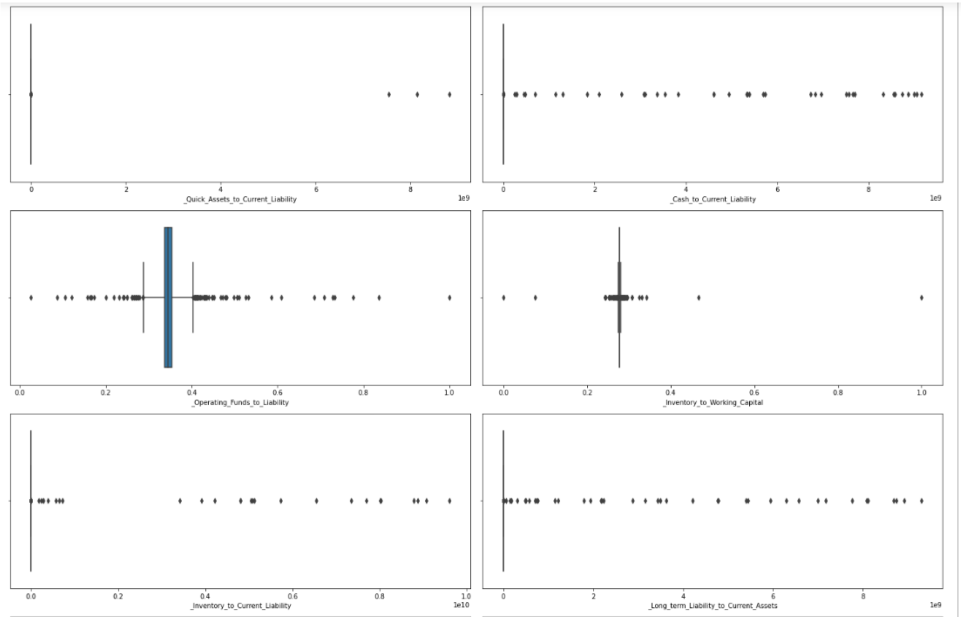
- First we import all the necessary libraries in Python, and then import the data file which is ‘Company_Data2015-1’. Once we import the file we confirm whether the data has been uploaded correctly or not using ‘head’ function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
- Then using the ‘shape’ function we can understand how many row and columns are there in our data set.
- To check the data type of all the columns and also to check the null values, ‘info’ function. Has been used.
- To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc.
- Using the ‘isnull’ function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
- Using the ‘dups’ function we check for the duplicates and there were no duplicate values.
- We also identified the unique values in categorical data. We used 3 times the IQR range as the criteria to determine the outliers. Our analysis gave significant chunk of outliers in the data. Below are boxplots which were plotted to analyze this data.

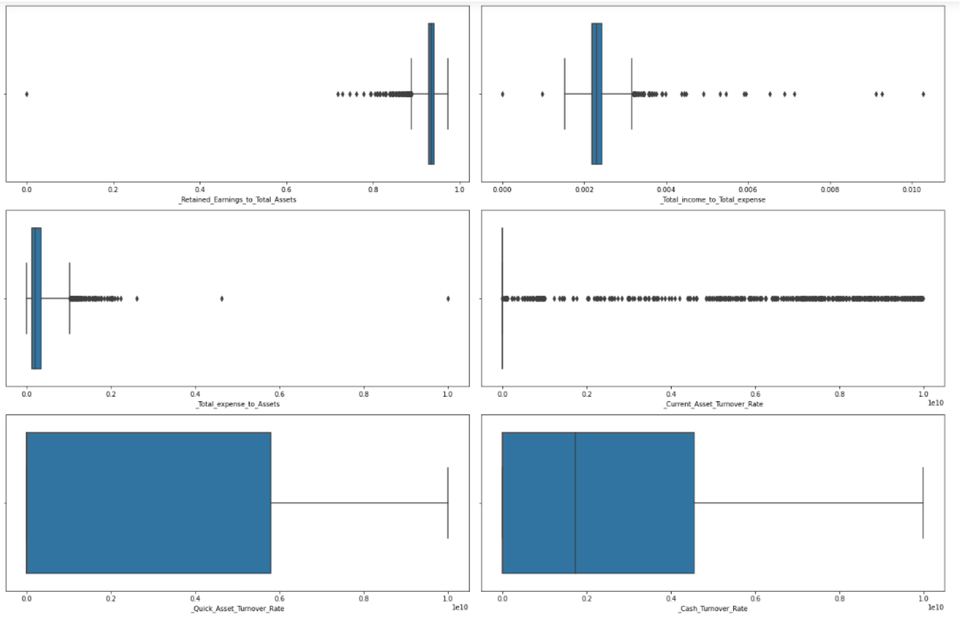

0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Operating_Expense_Rate
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Research_and_development_expense_rate
0.0
0.2
0.4
0.6
0.8
1.0
_Cash_flow_rate
0.0
0.2
0.4
0.6
0.8
1.0
1e9
_Interest_bearing_debt_interest_rate
0.0
0.2
0.4
0.6
0.8
1.0
_Tax_rate_A
0.20
0.25
0.30
0.35
0.40
0.45
_Cash_Flow_Per_Share

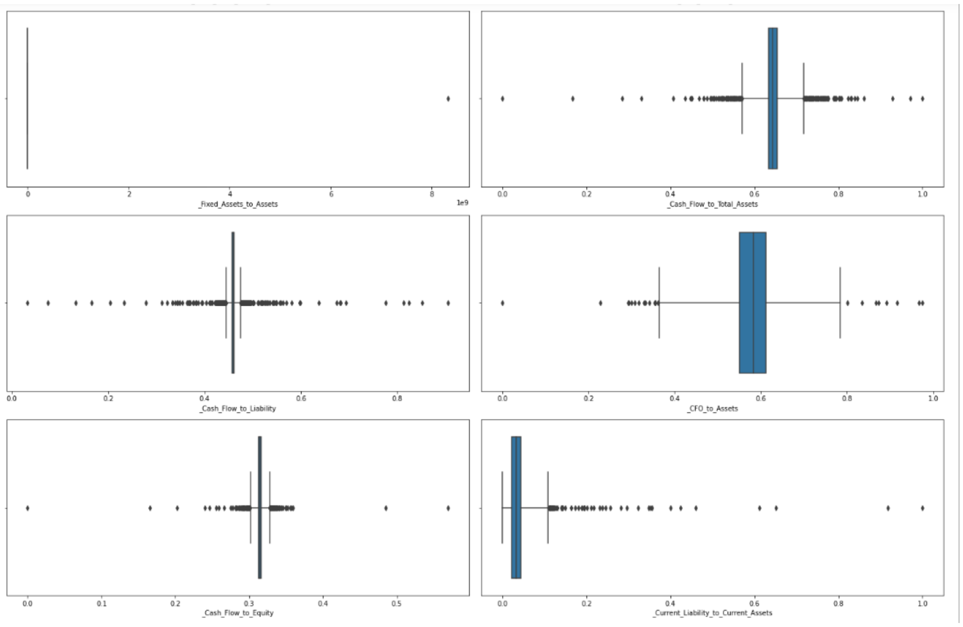

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
_Per_Share_Net_profit_before_tax_Yuan_
0.0
0.2
0.4
0.6
0.8
1.0
_Realized_Sales_Gross_Profit_Growth_Rate
0.75
0.80
0.85
0.90
0.95
1.00
_Operating_Profit_Growth_Rate
0.00
0.05
0.10
0.15
0.20
_Continuous_Net_Profit_Growth_Rate
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Total_Asset_Growth_Rate
0
2
4
6
8
1e9
_Net_Value_Growth_Rate

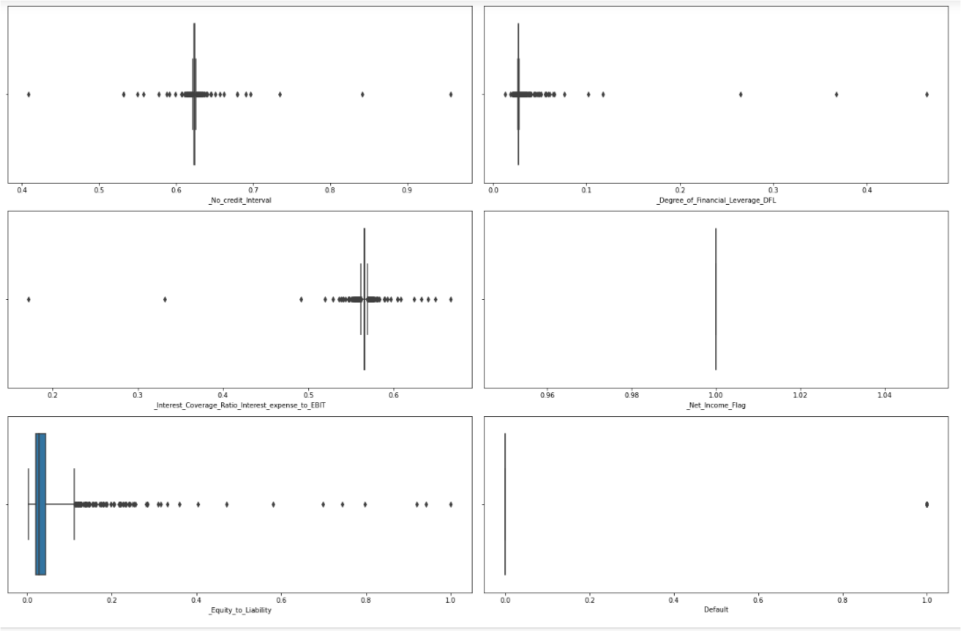

0.26
0.28
0.30
0.32
0.34
0.36
_Total_Asset_Return_Growth_Rate_Ratio
0.0
0.2
0.4
0.6
0.8
1.0
_Cash_Reinvestment_perc
0.0
0.5
1.0
1.5
2.0
2.5
1e9
_Current_Ratio
0
2
4
6
8
1e9
_Quick_Ratio
0.55
0.60
0.65
0.70
0.75
0.80
_Interest_Expense_Ratio
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Total_debt_to_Total_net_worth
0.0
0.2
0.4
0.6
0.8
1.0
_Long_term_fund_suitability_ratio_A
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
_Net_profit_before_tax_to_Paid_in_capital
0.0
0.2
0.4
0.6
0.8
_Total_Asset_Turnover
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Accounts_Receivable_Turnover
0
2
4
6
8
1e9
_Average_Collection_Days
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Inventory_Turnover_Rate_times


0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Fixed_Assets_Turnover_Frequency
0.0
0.2
0.4
0.6
0.8
1.0
_Net_Worth_Turnover_Rate_times
0.0
0.2
0.4
0.6
0.8
1.0
_Operating_profit_per_person
0
2
4
6
8
1e9
_Allocation_rate_per_person
0.0
0.2
0.4
0.6
0.8
1.0
_Quick_Assets_to_Total_Assets
0.0
0.2
0.4
0.6
0.8
_Cash_to_Total_Assets


0
2
4
6
8
1e9
_Quick_Assets_to_Current_Liability
0
2
4
6
8
1e9
_Cash_to_Current_Liability
0.0
0.2
0.4
0.6
0.8
1.0
_Operating_Funds_to_Liability
0.0
0.2
0.4
0.6
0.8
1.0
_Inventory_to_Working_Capital
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Inventory_to_Current_Liability
0
2
4
6
8
1e9
_Long_term_Liability_to_Current_Assets


0.0
0.2
0.4
0.6
0.8
1.0
_Retained_Earnings_to_Total_Assets
0.000
0.002
0.004
0.006
0.008
0.010
_Total_income_to_Total_expense
0.0
0.2
0.4
0.6
0.8
1.0
_Total_expense_to_Assets
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Current_Asset_Turnover_Rate
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Quick_Asset_Turnover_Rate
0.0
0.2
0.4
0.6
0.8
1.0
1e10
_Cash_Turnover_Rate

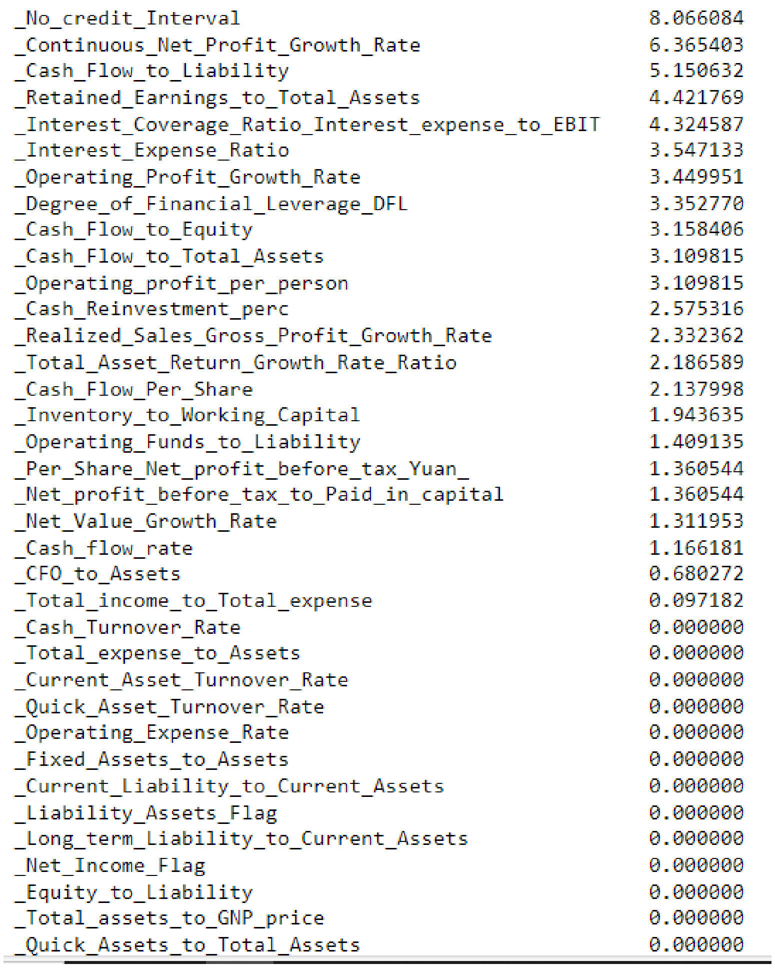

0
2
4
6
8
1e9
_Fixed_Assets_to_Assets
0.0
0.2
0.4
0.6
0.8
1.0
_Cash_Flow_to_Total_Assets
0.0
0.2
0.4
0.6
0.8
_Cash_Flow_to_Liability
0.0
0.2
0.4
0.6
0.8
1.0
_CFO_to_Assets
0.0
0.1
0.2
0.3
0.4
0.5
_Cash_Flow_to_Equity
0.0
0.2
0.4
0.6
0.8
1.0
_Current_Liability_to_Current_Assets



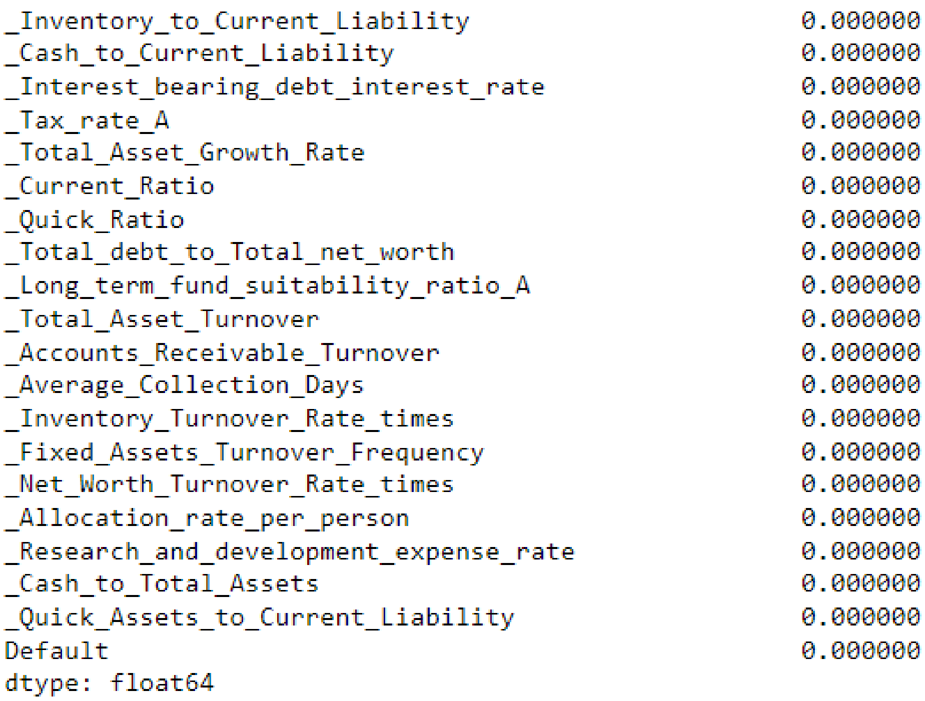
OUTLIER TREATMENT

Significant number of outliers were present for almost all the variables. We captured the actual percentage of data which was above and below the third and first quintiles respectively.

Data above third quintile

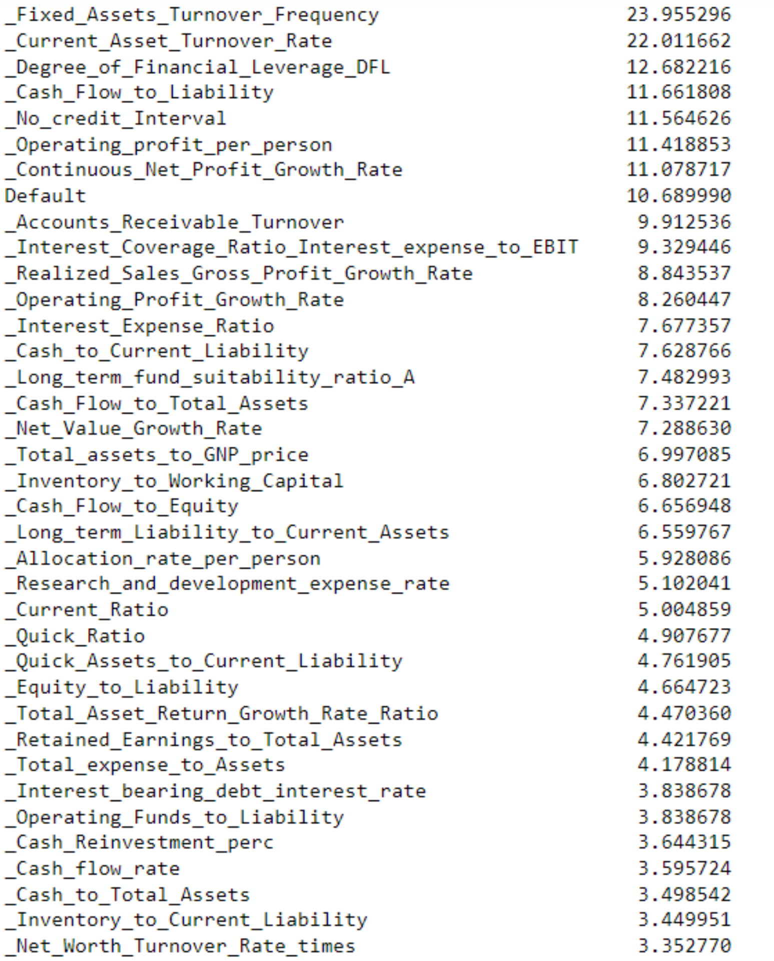


```
_No_credit_Interval                                8.066084
_Continuous_Net_Profit_Growth_Rate                 6.365403
_Cash_Flow_to_Liability                            5.150632
_Retained_Earnings_to_Total_Assets                 4.421769
_Interest_Coverage_Ratio_Interest_expense_to_EBIT  4.324587
_Interest_Expense_Ratio                            3.547133
_Operating_Profit_Growth_Rate                      3.449951
_Degree_of_Financial_Leverage_DFL                  3.352770
_Cash_Flow_to_Equity                               3.158406
_Cash_Flow_to_Total_Assets                         3.109815
_Operating_profit_per_person                       3.109815
_Cash_Reinvestment_perc                            2.575316
_Realized_Sales_Gross_Profit_Growth_Rate           2.332362
_Total_Asset_Return_Growth_Rate_Ratio              2.186589
_Cash_Flow_Per_Share                               2.137998
_Inventory_to_Working_Capital                      1.943635
_Operating_Funds_to_Liability                      1.409135
_Per_Share_Net_profit_before_tax_Yuan_             1.360544
_Net_profit_before_tax_to_Paid_in_capital          1.360544
_Net_Value_Growth_Rate                             1.311953
_Cash_flow_rate                                    1.166181
_CFO_to_Assets                                     0.680272
_Total_income_to_Total_expense                     0.097182
_Cash_Turnover_Rate                                0.000000
_Total_expense_to_Assets                           0.000000
_Current_Asset_Turnover_Rate                       0.000000
_Quick_Asset_Turnover_Rate                         0.000000
_Operating_Expense_Rate                            0.000000
_Fixed_Assets_to_Assets                            0.000000
_Current_Liability_to_Current_Assets               0.000000
_Liability_Assets_Flag                             0.000000
_Long_term_Liability_to_Current_Assets             0.000000
_Net_Income_Flag                                   0.000000
_Equity_to_Liability                               0.000000
_Total_assets_to_GNP_price                         0.000000
_Quick_Assets_to_Total_Assets                      0.000000
```

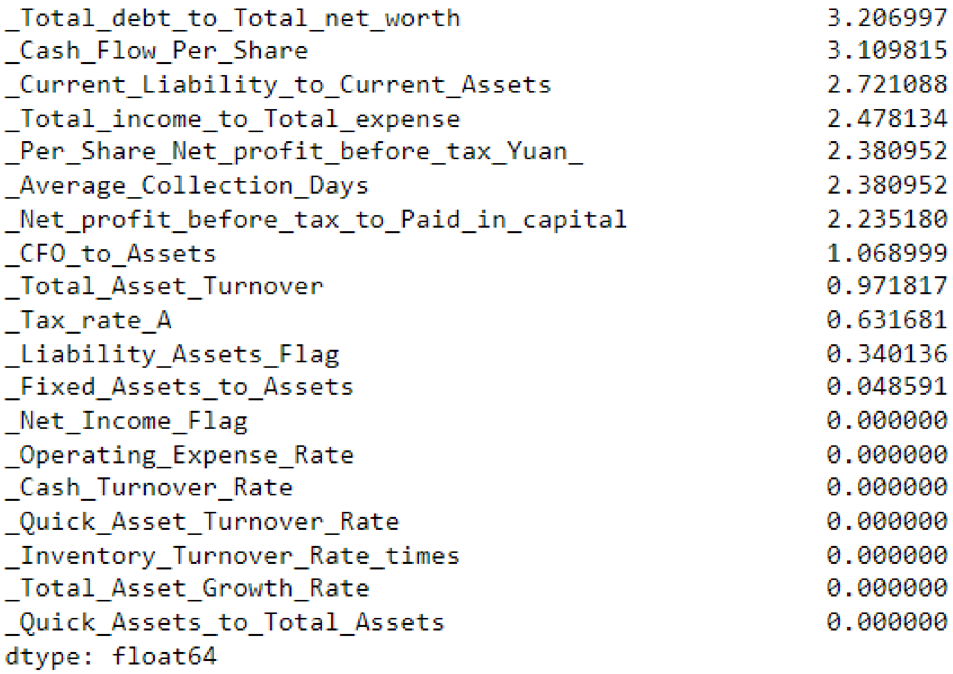


```
_Inventory_to_Current_Liability                    0.000000
_Cash_to_Current_Liability                         0.000000
_Interest_bearing_debt_interest_rate               0.000000
_Tax_rate_A                                        0.000000
_Total_Asset_Growth_Rate                           0.000000
_Current_Ratio                                     0.000000
_Quick_Ratio                                       0.000000
_Total_debt_to_Total_net_worth                     0.000000
_Long_term_fund_suitability_ratio_A                0.000000
_Total_Asset_Turnover                              0.000000
_Accounts_Receivable_Turnover                      0.000000
_Average_Collection_Days                           0.000000
_Inventory_Turnover_Rate_times                     0.000000
_Fixed_Assets_Turnover_Frequency                   0.000000
_Net_Worth_Turnover_Rate_times                     0.000000
_Allocation_rate_per_person                        0.000000
_Research_and_development_expense_rate             0.000000
_Cash_to_Total_Assets                              0.000000
_Quick_Assets_to_Current_Liability                 0.000000
Default                                            0.000000
dtype: float64
```

Data above first quartile



```
_Fixed_Assets_Turnover_Frequency                   23.955296
_Current_Asset_Turnover_Rate                       22.011662
_Degree_of_Financial_Leverage_DFL                  12.682216
_Cash_Flow_to_Liability                            11.661808
_No_credit_Interval                                11.564626
_Operating_profit_per_person                       11.418853
_Continuous_Net_Profit_Growth_Rate                 11.078717
Default                                            10.689990
_Accounts_Receivable_Turnover                       9.912536
_Interest_Coverage_Ratio_Interest_expense_to_EBIT   9.329446
_Realized_Sales_Gross_Profit_Growth_Rate            8.843537
_Operating_Profit_Growth_Rate                       8.260447
_Interest_Expense_Ratio                             7.677357
_Cash_to_Current_Liability                          7.628766
_Long_term_fund_suitability_ratio_A                 7.482993
_Cash_Flow_to_Total_Assets                          7.337221
_Net_Value_Growth_Rate                              7.288630
_Total_assets_to_GNP_price                          6.997085
_Inventory_to_Working_Capital                       6.802721
_Cash_Flow_to_Equity                                6.656948
_Long_term_Liability_to_Current_Assets              6.559767
_Allocation_rate_per_person                         5.928086
_Research_and_development_expense_rate              5.102041
_Current_Ratio                                      5.004859
_Quick_Ratio                                        4.907677
_Quick_Assets_to_Current_Liability                  4.761905
_Equity_to_Liability                                4.664723
_Total_Asset_Return_Growth_Rate_Ratio               4.470360
_Retained_Earnings_to_Total_Assets                  4.421769
_Total_expense_to_Assets                            4.178814
_Interest_bearing_debt_interest_rate                3.838678
_Operating_Funds_to_Liability                       3.838678
_Cash_Reinvestment_perc                             3.644315
_Cash_flow_rate                                     3.595724
_Cash_to_Total_Assets                               3.498542
_Inventory_to_Current_Liability                     3.449951
_Net_Worth_Turnover_Rate_times                      3.352770
```



```
_Total_debt_to_Total_net_worth                   3.206997
_Cash_Flow_Per_Share                             3.109815
_Current_Liability_to_Current_Assets             2.721088
_Total_income_to_Total_expense                   2.478134
_Per_Share_Net_profit_before_tax_Yuan_           2.380952
_Average_Collection_Days                         2.380952
_Net_profit_before_tax_to_Paid_in_capital        2.235180
_CFO_to_Assets                                   1.068999
_Total_Asset_Turnover                            0.971817
_Tax_rate_A                                      0.631681
_Liability_Assets_Flag                           0.340136
_Fixed_Assets_to_Assets                          0.048591
_Net_Income_Flag                                 0.000000
_Operating_Expense_Rate                          0.000000
_Cash_Turnover_Rate                              0.000000
_Quick_Asset_Turnover_Rate                       0.000000
_Inventory_Turnover_Rate_times                   0.000000
_Total_Asset_Growth_Rate                         0.000000
_Quick_Assets_to_Total_Assets                    0.000000
dtype: float64
```

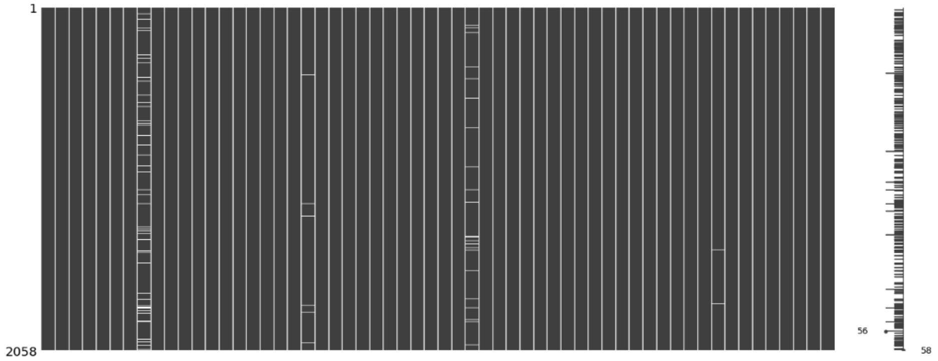
Since the number of outliers are too large in number to be treated, as treated such large number of records would mean changing the essence of the data. Also given the fact that this is a financial data and the outliers might very well reflect the information which is genuine in nature. Since there is data captured for small, medium as well as large companies. Hence we decided against treating the outliers in this data set.

PROBLEM 1.2

Missing Value Treatment Resolution:

Given the size of the data set i.e. 2058 rows, there were not many missing values to start with. There were a total of 298 missing records observed in the entire data.

Snapshot from missingno library has been published below for reference



There are 4 variables containing null values which are '_cash_flow_per_share' , '_cash_to_total_assets' , '_total_debt to_total_net_worth' and '_current_liability_to_current_assets'.

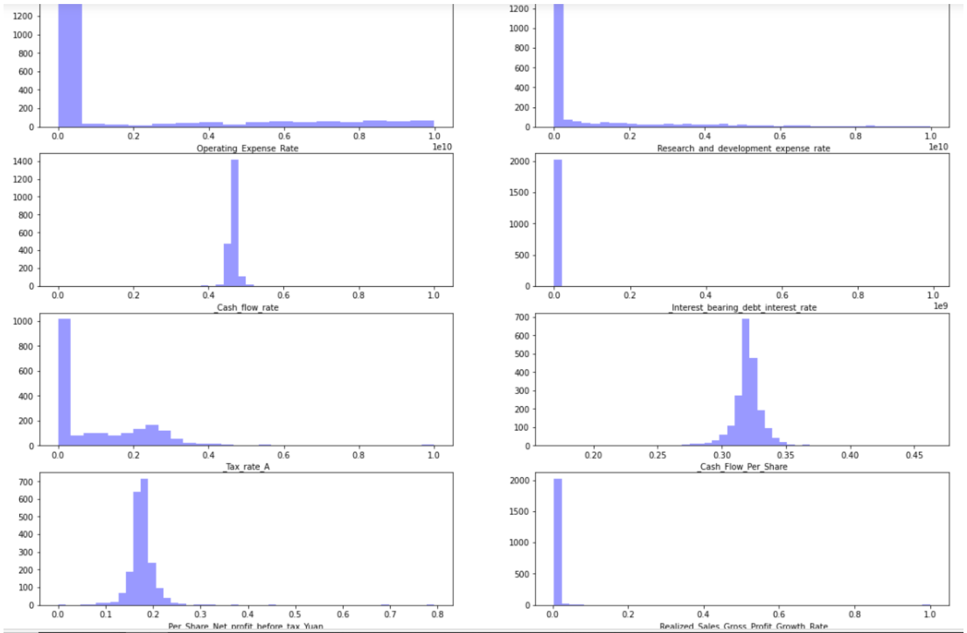
Records with missing value in this 4 column were imputed with the average value. No more missing values were present after treatment.

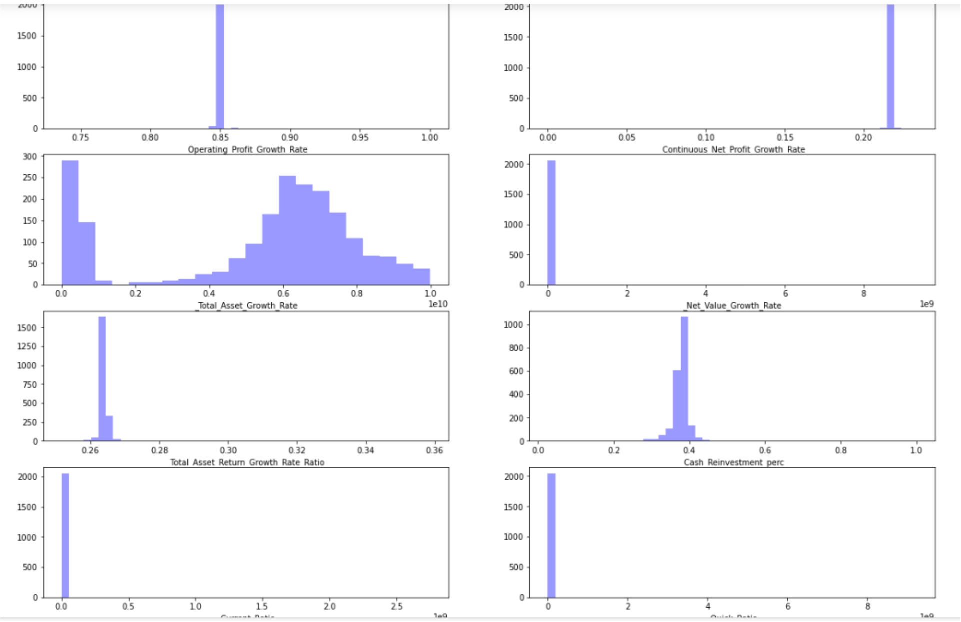
PROBLEM 1.4

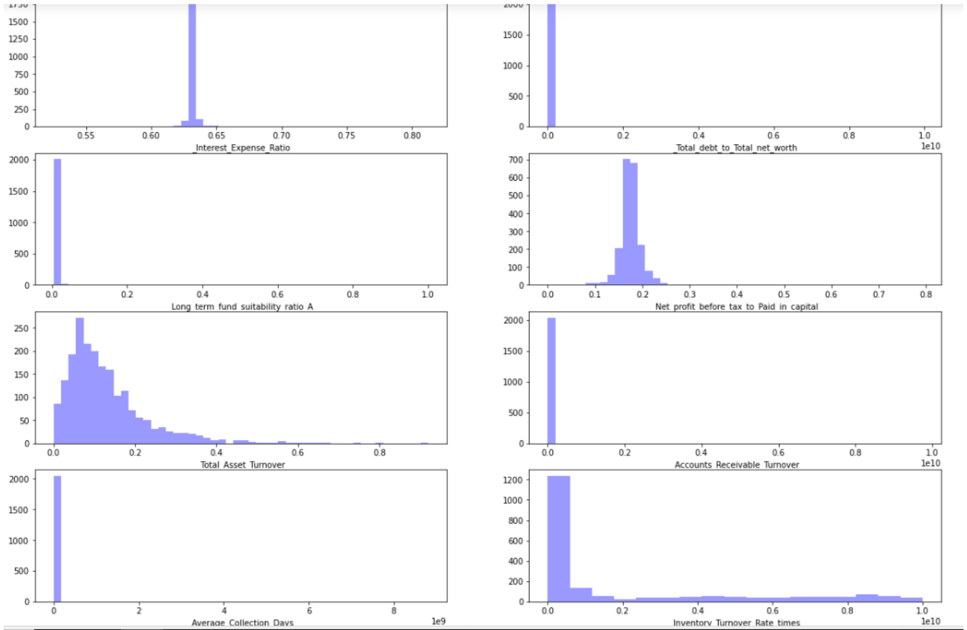
Univariate & Bivariate analysis with proper interpretation.

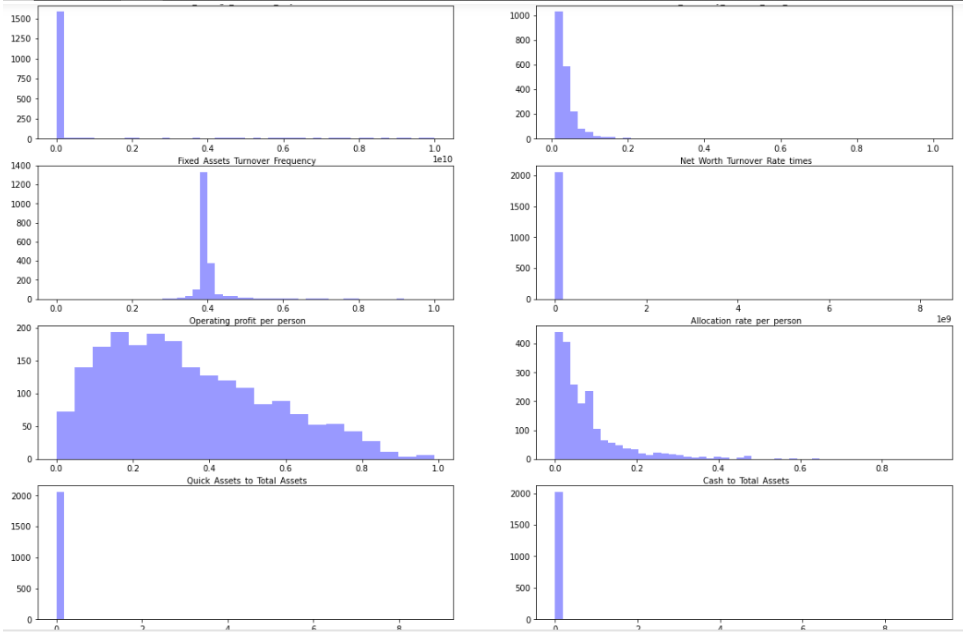
Resolution:

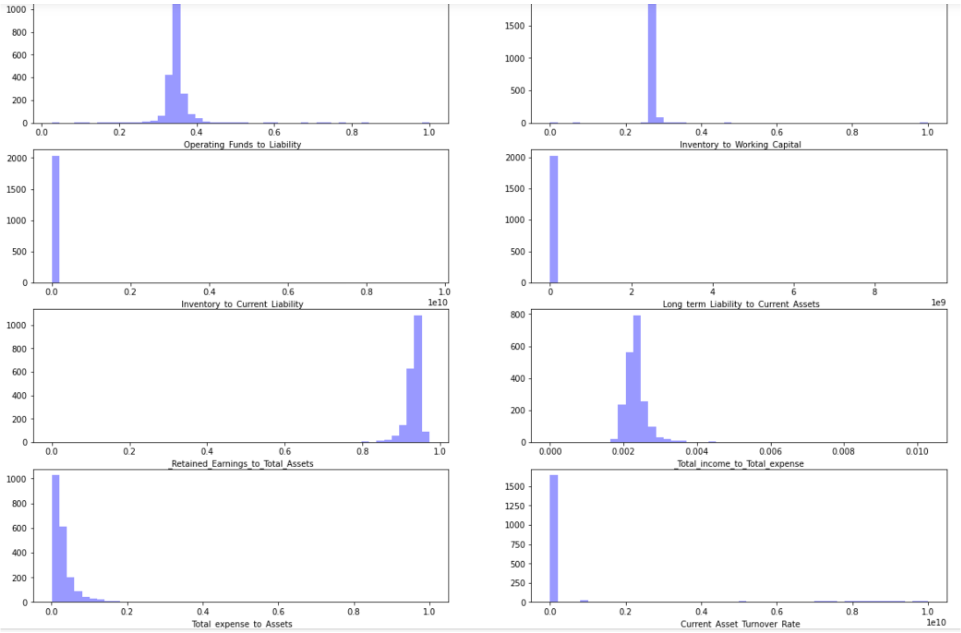
Distplot were plotted for all the variables to analyze the distribution of all the variables.

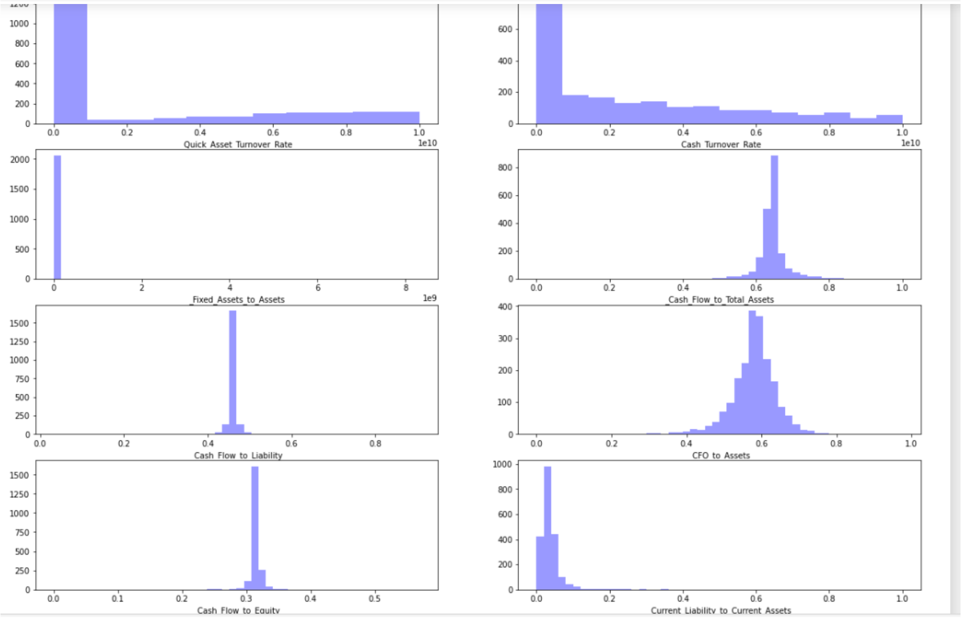


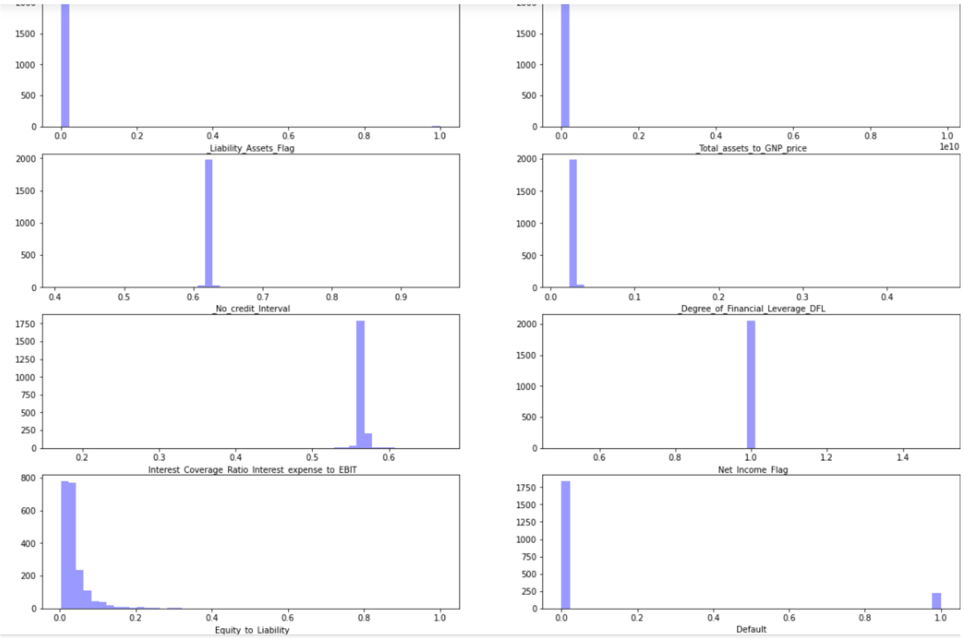



Interest_Expense_Ratio
Total_debt_to_Total_net_worth
1e10
Long_term_fund_suitability_ratio_A
Net_profit_before_tax_to_Paid_in_capital
Total_Asset_Turnover
Accounts_Receivable_Turnover
1e10
Average_Collection_Days
1e9
Inventory_Turnover_Rate_times
1e10


Fixed_Assets_Turnover_Frequency
1e10
Net_Worth_Turnover_Rate_times
Operating_profit_per_person
Allocation_rate_per_person
1e9
Quick_Assets_to_Total_Assets
Cash_to_Total_Assets

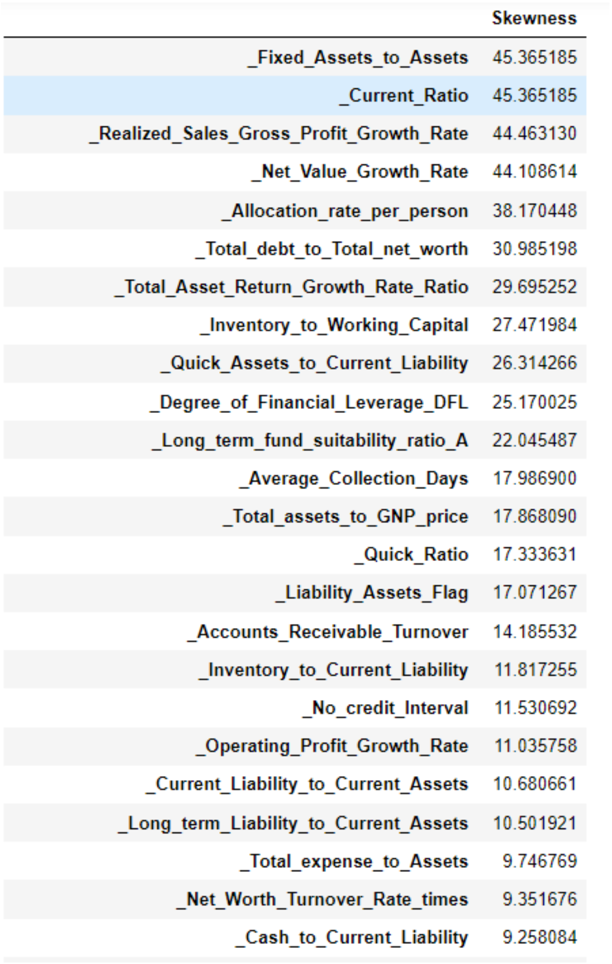

Operating_Funds_to_Liability
Inventory_to_Working_Capital
Inventory_to_Current_Liability
1e10
Long_term_Liability_to_Current_Assets
1e9
_Retained_Earnings_to_Total_Assets
_Total_income_to_Total_expense
Total_expense_to_Assets
Current_Asset_Turnover_Rate
1e10


Quick_Asset_Turnover_Rate
1e10
Cash_Turnover_Rate
1e10
_Fixed_Assets_to_Assets
1e9
_Cash_Flow_to_Total_Assets
Cash_Flow_to_Liability
CFO_to_Assets
Cash_Flow_to_Equity
Current_Liability_to_Current_Assets

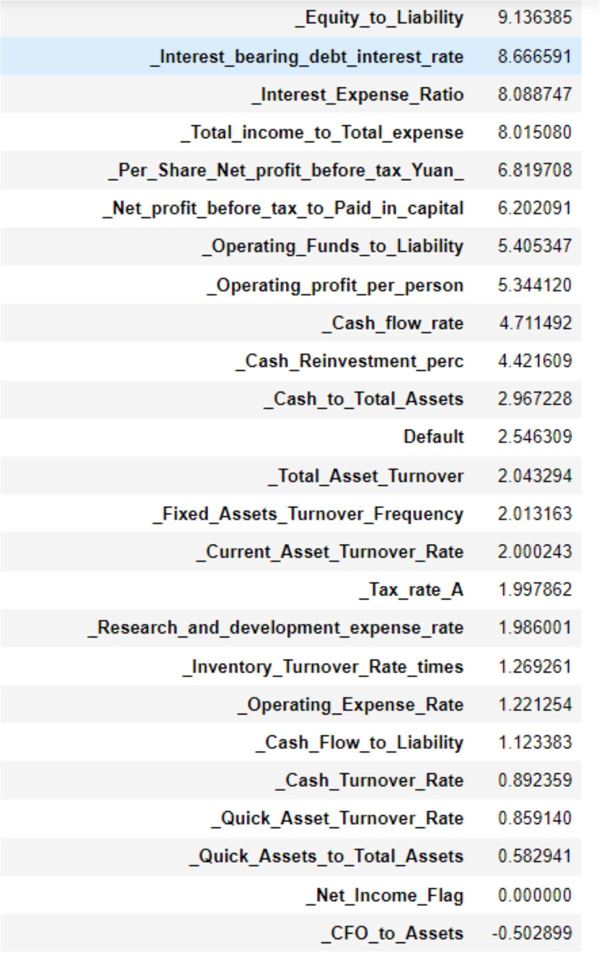


None of the variables show perfect normal distribution. Few of the variables have skewness in data. There are no duplicate values. Skewness was observed in almost all the variables.

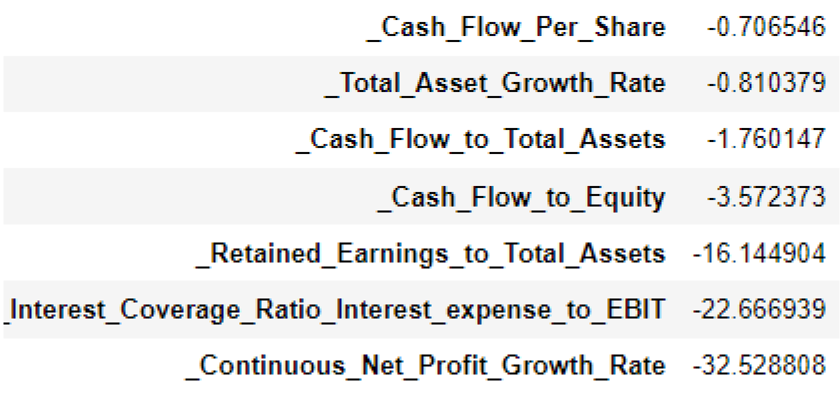
Skewness in the dataset.



	Skewness
_Fixed_Assets_to_Assets	45.365185
_Current_Ratio	45.365185
_Realized_Sales_Gross_Profit_Growth_Rate	44.463130
_Net_Value_Growth_Rate	44.108614
_Allocation_rate_per_person	38.170448
_Total_debt_to_Total_net_worth	30.985198
_Total_Asset_Return_Growth_Rate_Ratio	29.695252
_Inventory_to_Working_Capital	27.471984
_Quick_Assets_to_Current_Liability	26.314266
_Degree_of_Financial_Leverage_DFL	25.170025
_Long_term_fund_suitability_ratio_A	22.045487
_Average_Collection_Days	17.986900
_Total_assets_to_GNP_price	17.868090
_Quick_Ratio	17.333631
_Liability_Assets_Flag	17.071267
_Accounts_Receivable_Turnover	14.185532
_Inventory_to_Current_Liability	11.817255
_No_credit_Interval	11.530692
_Operating_Profit_Growth_Rate	11.035758
_Current_Liability_to_Current_Assets	10.680661
_Long_term_Liability_to_Current_Assets	10.501921
_Total_expense_to_Assets	9.746769
_Net_Worth_Turnover_Rate_times	9.351676
_Cash_to_Current_Liability	9.258084



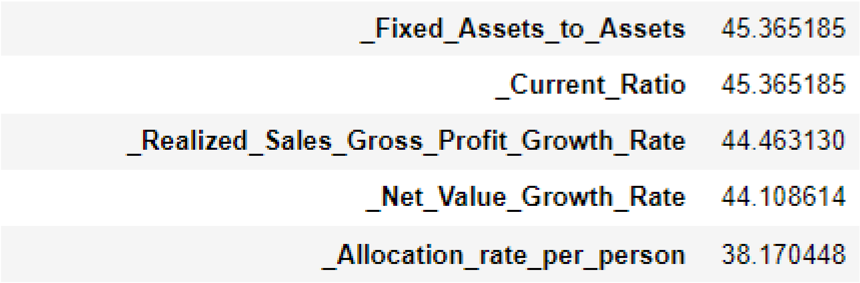
_Equity_to_Liability	9.136385
_Interest_bearing_debt_interest_rate	8.666591
_Interest_Expense_Ratio	8.088747
_Total_income_to_Total_expense	8.015080
_Per_Share_Net_profit_before_tax_Yuan_	6.819708
_Net_profit_before_tax_to_Paid_in_capital	6.202091
_Operating_Funds_to_Liability	5.405347
_Operating_profit_per_person	5.344120
_Cash_flow_rate	4.711492
_Cash_Reinvestment_perc	4.421609
_Cash_to_Total_Assets	2.967228
Default	2.546309
_Total_Asset_Turnover	2.043294
_Fixed_Assets_Turnover_Frequency	2.013163
_Current_Asset_Turnover_Rate	2.000243
_Tax_rate_A	1.997862
_Research_and_development_expense_rate	1.986001
_Inventory_Turnover_Rate_times	1.269261
_Operating_Expense_Rate	1.221254
_Cash_Flow_to_Liability	1.123383
_Cash_Turnover_Rate	0.892359
_Quick_Asset_Turnover_Rate	0.859140
_Quick_Assets_to_Total_Assets	0.582941
_Net_Income_Flag	0.000000
_CFO_to_Assets	-0.502899



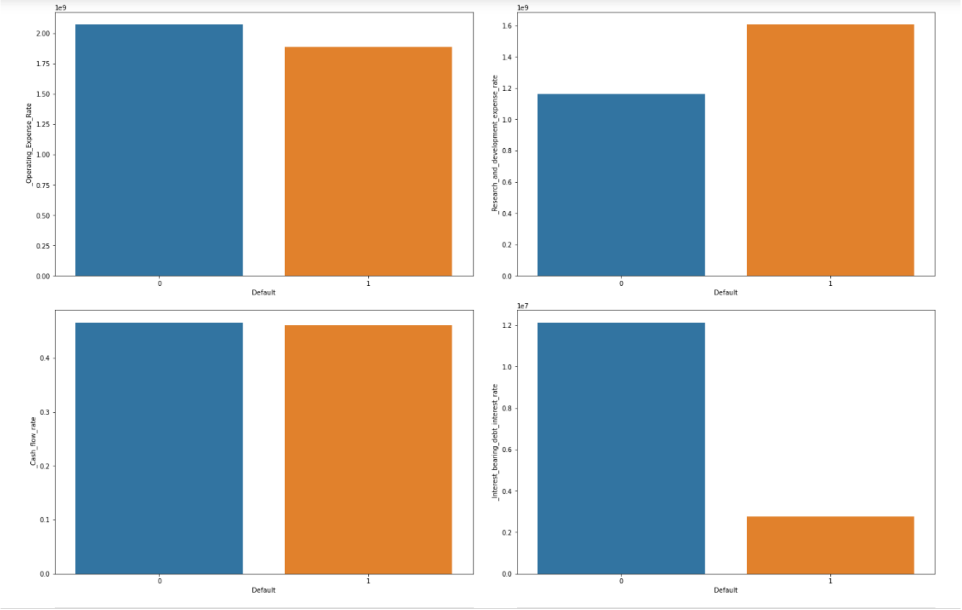
_Cash_Flow_Per_Share	-0.706546
_Total_Asset_Growth_Rate	-0.810379
_Cash_Flow_to_Total_Assets	-1.760147
_Cash_Flow_to_Equity	-3.572373
_Retained_Earnings_to_Total_Assets	-16.144904
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	-22.666939
_Continuous_Net_Profit_Growth_Rate	-32.528808

Univariate Analysis

Data is highly skewed. Most variables were found having tails to the right and hence were right skewed. The top 5 variables that have the highest skew are:

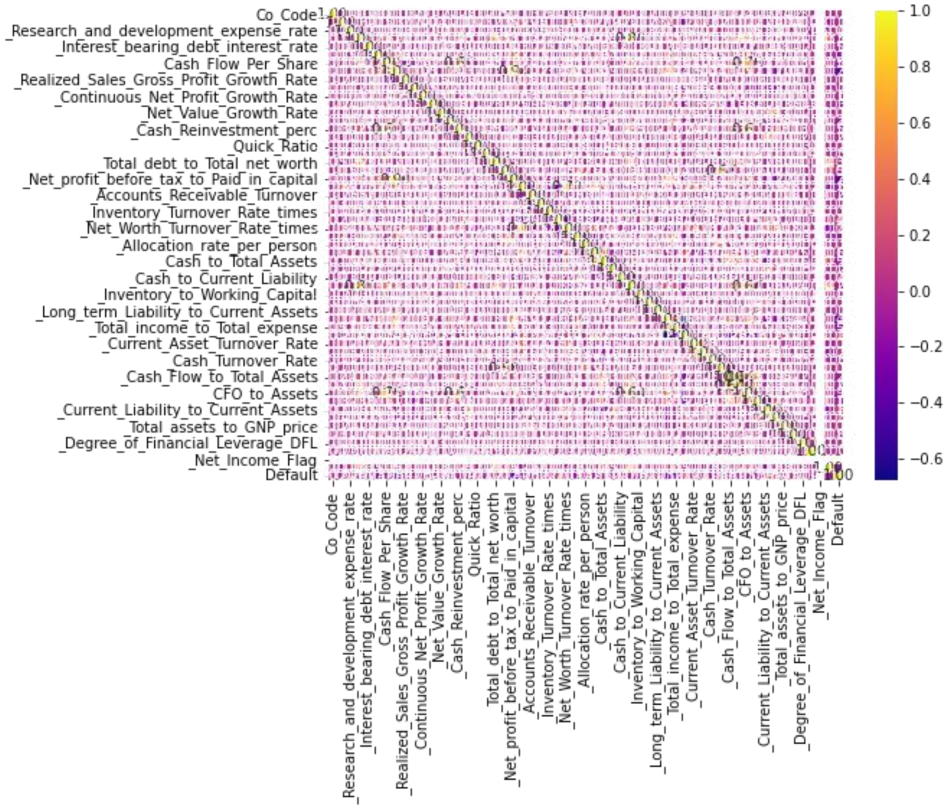


_Fixed_Assets_to_Assets	45.365185
_Current_Ratio	45.365185
_Realized_Sales_Gross_Profit_Growth_Rate	44.463130
_Net_Value_Growth_Rate	44.108614
_Allocation_rate_per_person	38.170448



Multivariate Analysis

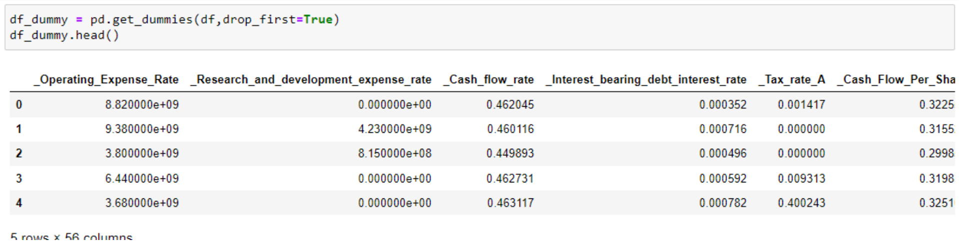
We also performed multi Variate analysis on the data to see if there are any correlation that are observed within the data. Correlations function was used and seaborn cluster map was used to plot the correlations and to make better sense of the data.



PROBLEM 1.5

Train Test Split Resolution:

Since there was a great imbalance in the data set, we also created a parallel data set with SMOTE and evaluated the performance on smote as well as non smote data.



```
df_dummy = pd.get_dummies(df,drop_first=True)
df_dummy.head()
```

	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A	_Cash_Flow_Per_Sha
0	8.820000e+09	0.000000e+00	0.462045	0.000352	0.001417	0.3225
1	9.380000e+09	4.230000e+09	0.460116	0.000716	0.000000	0.3155
2	3.800000e+09	8.150000e+08	0.449893	0.000496	0.000000	0.2998
3	6.440000e+09	0.000000e+00	0.462731	0.000592	0.009313	0.3198
4	3.680000e+09	0.000000e+00	0.463117	0.000782	0.400243	0.3251

5 rows × 56 columns

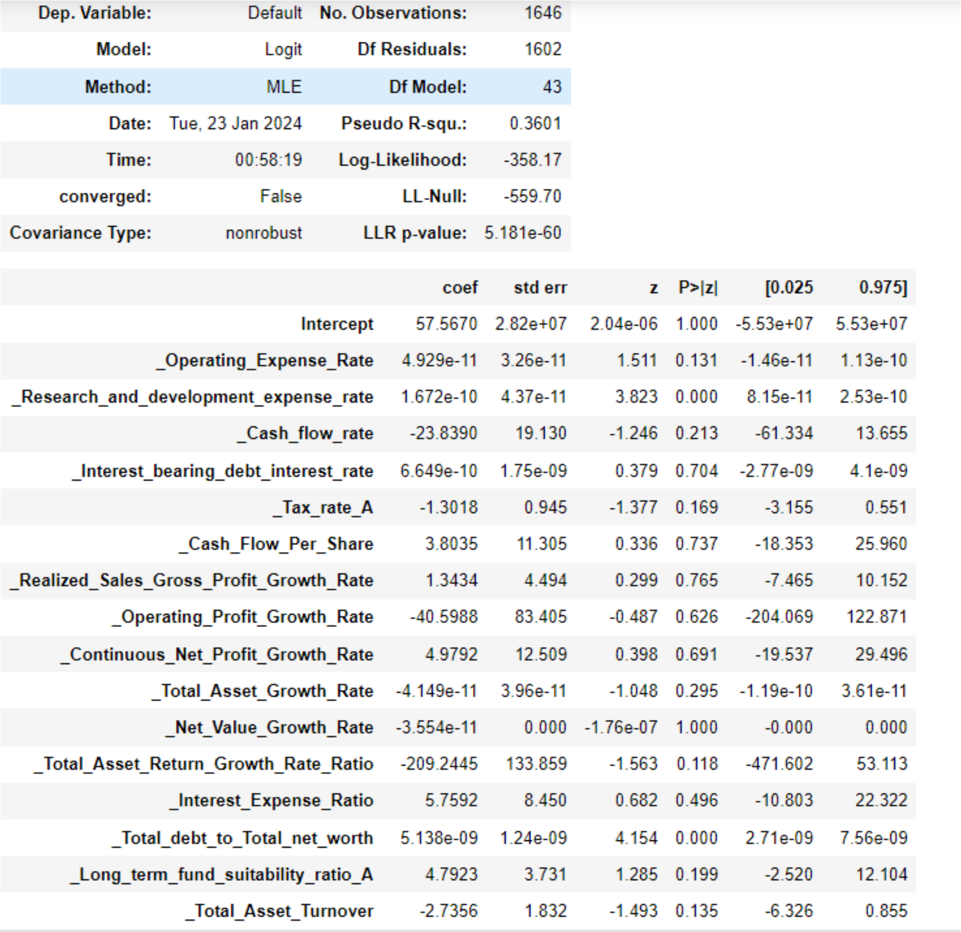
Data was split in the 67:33 ratio as per project notes using sklearn's train_test_split function. Also seed value of 42 was used.

PROBLEM 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

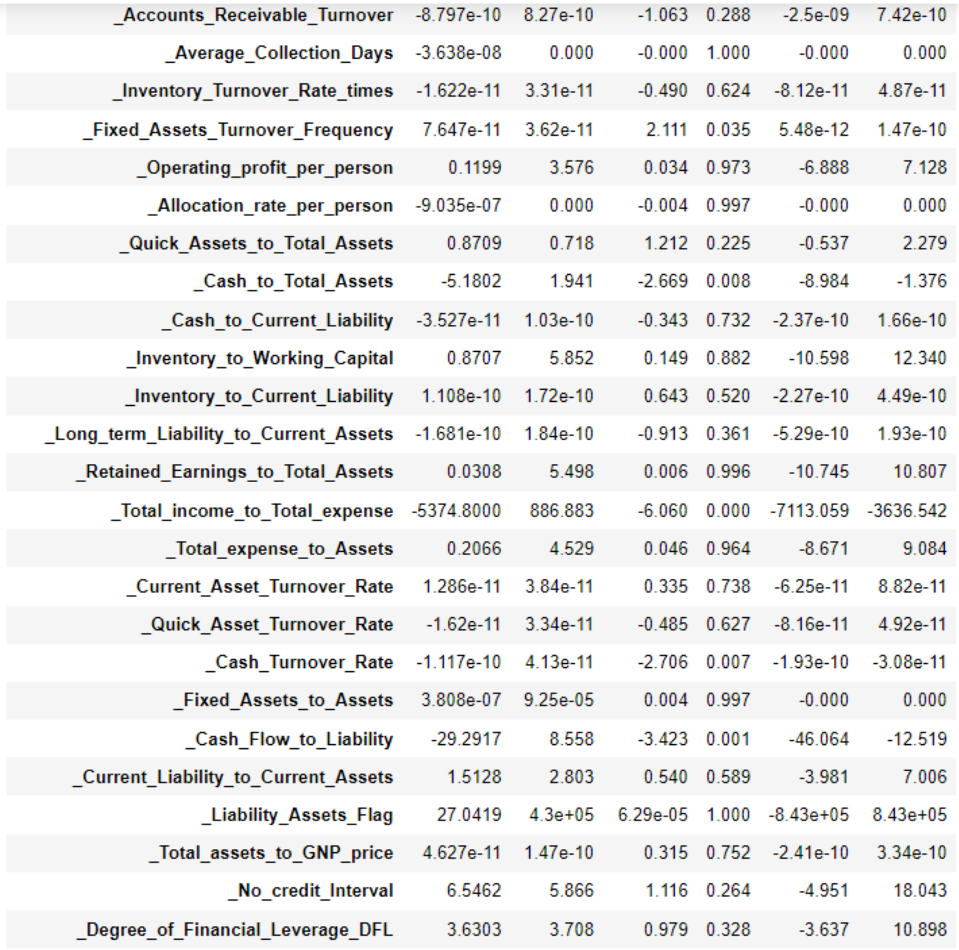
Resolution:

First model-



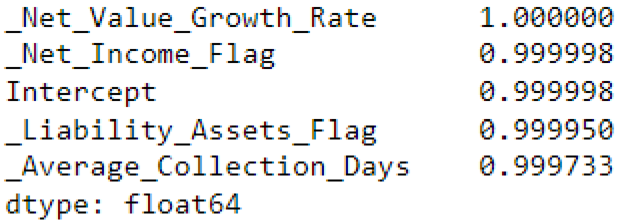
Dep. Variable:	Default	**No. Observations:**	1646
Model:	Logit	**Df Residuals:**	1602
Method:	MLE	**Df Model:**	43
Date:	Tue, 23 Jan 2024	**Pseudo R-squ.:**	0.3601
Time:	00:58:19	**Log-Likelihood:**	-358.17
converged:	False	**LL-Null:**	-559.70
Covariance Type:	nonrobust	**LLR p-value:**	5.181e-60

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	57.5670	2.82e+07	2.04e-06	1.000	-5.53e+07	5.53e+07
_Operating_Expense_Rate	4.929e-11	3.26e-11	1.511	0.131	-1.46e-11	1.13e-10
_Research_and_development_expense_rate	1.672e-10	4.37e-11	3.823	0.000	8.15e-11	2.53e-10
_Cash_flow_rate	-23.8390	19.130	-1.246	0.213	-61.334	13.655
_Interest_bearing_debt_interest_rate	6.649e-10	1.75e-09	0.379	0.704	-2.77e-09	4.1e-09
_Tax_rate_A	-1.3018	0.945	-1.377	0.169	-3.155	0.551
_Cash_Flow_Per_Share	3.8035	11.305	0.336	0.737	-18.353	25.960
_Realized_Sales_Gross_Profit_Growth_Rate	1.3434	4.494	0.299	0.765	-7.465	10.152
_Operating_Profit_Growth_Rate	-40.5988	83.405	-0.487	0.626	-204.069	122.871
_Continuous_Net_Profit_Growth_Rate	4.9792	12.509	0.398	0.691	-19.537	29.496
_Total_Asset_Growth_Rate	-4.149e-11	3.96e-11	-1.048	0.295	-1.19e-10	3.61e-11
_Net_Value_Growth_Rate	-3.554e-11	0.000	-1.76e-07	1.000	-0.000	0.000
_Total_Asset_Return_Growth_Rate_Ratio	-209.2445	133.859	-1.563	0.118	-471.602	53.113
_Interest_Expense_Ratio	5.7592	8.450	0.682	0.496	-10.803	22.322
_Total_debt_to_Total_net_worth	5.138e-09	1.24e-09	4.154	0.000	2.71e-09	7.56e-09
_Long_term_fund_suitability_ratio_A	4.7923	3.731	1.285	0.199	-2.520	12.104
_Total_Asset_Turnover	-2.7356	1.832	-1.493	0.135	-6.326	0.855



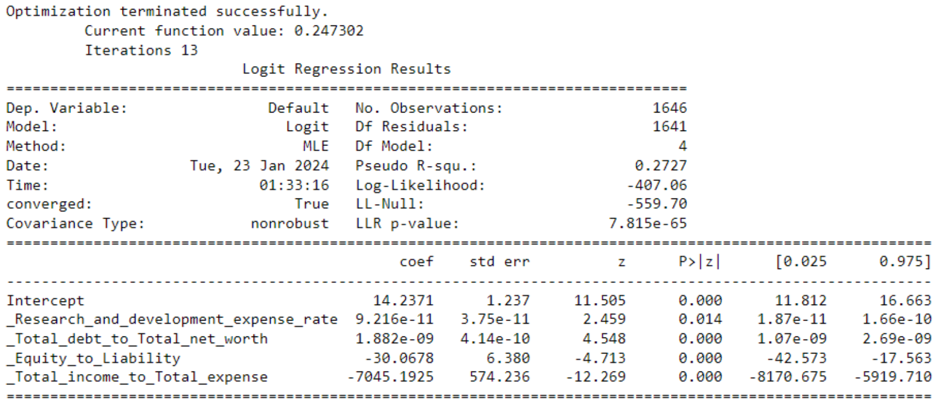
_Accounts_Receivable_Turnover	-8.797e-10	8.27e-10	-1.063	0.288	-2.5e-09	7.42e-10
_Average_Collection_Days	-3.638e-08	0.000	-0.000	1.000	-0.000	0.000
_Inventory_Turnover_Rate_times	-1.622e-11	3.31e-11	-0.490	0.624	-8.12e-11	4.87e-11
_Fixed_Assets_Turnover_Frequency	7.647e-11	3.62e-11	2.111	0.035	5.48e-12	1.47e-10
_Operating_profit_per_person	0.1199	3.576	0.034	0.973	-6.888	7.128
_Allocation_rate_per_person	-9.035e-07	0.000	-0.004	0.997	-0.000	0.000
_Quick_Assets_to_Total_Assets	0.8709	0.718	1.212	0.225	-0.537	2.279
_Cash_to_Total_Assets	-5.1802	1.941	-2.669	0.008	-8.984	-1.376
_Cash_to_Current_Liability	-3.527e-11	1.03e-10	-0.343	0.732	-2.37e-10	1.66e-10
_Inventory_to_Working_Capital	0.8707	5.852	0.149	0.882	-10.598	12.340
_Inventory_to_Current_Liability	1.108e-10	1.72e-10	0.643	0.520	-2.27e-10	4.49e-10
_Long_term_Liability_to_Current_Assets	-1.681e-10	1.84e-10	-0.913	0.361	-5.29e-10	1.93e-10
_Retained_Earnings_to_Total_Assets	0.0308	5.498	0.006	0.996	-10.745	10.807
_Total_income_to_Total_expense	-5374.8000	886.883	-6.060	0.000	-7113.059	-3636.542
_Total_expense_to_Assets	0.2066	4.529	0.046	0.964	-8.671	9.084
_Current_Asset_Turnover_Rate	1.286e-11	3.84e-11	0.335	0.738	-6.25e-11	8.82e-11
_Quick_Asset_Turnover_Rate	-1.62e-11	3.34e-11	-0.485	0.627	-8.16e-11	4.92e-11
_Cash_Turnover_Rate	-1.117e-10	4.13e-11	-2.706	0.007	-1.93e-10	-3.08e-11
_Fixed_Assets_to_Assets	3.808e-07	9.25e-05	0.004	0.997	-0.000	0.000
_Cash_Flow_to_Liability	-29.2917	8.558	-3.423	0.001	-46.064	-12.519
_Current_Liability_to_Current_Assets	1.5128	2.803	0.540	0.589	-3.981	7.006
_Liability_Assets_Flag	27.0419	4.3e+05	6.29e-05	1.000	-8.43e+05	8.43e+05
_Total_assets_to_GNP_price	4.627e-11	1.47e-10	0.315	0.752	-2.41e-10	3.34e-10
_No_credit_Interval	6.5462	5.866	1.116	0.264	-4.951	18.043
_Degree_of_Financial_Leverage_DFL	3.6303	3.708	0.979	0.328	-3.637	10.898

P-value in descending order



```
_Net_Value_Growth_Rate       1.000000
_Net_Income_Flag             0.999998
Intercept                    0.999998
_Liability_Assets_Flag       0.999950
_Average_Collection_Days     0.999733
dtype: float64
```

It is evident from the above image that the variable _Net_Value_Growth_Rate has a p-value of 1.00000. Since this is higher than 0.05 and the highest of all the variables, we will drop this variable in subsequent models. This process of dropping variables based on p-values and modeling continued until a model where all the p-values were relevant was achieved. The iterative process got stopped at Model11 which has 4 independent variables and each of them were relevant.



```
Optimization terminated successfully.
         Current function value: 0.247302
         Iterations 13
                           Logit Regression Results                           
==============================================================================
Dep. Variable:                Default   No. Observations:                 1646
Model:                          Logit   Df Residuals:                     1641
Method:                           MLE   Df Model:                            4
Date:                Tue, 23 Jan 2024   Pseudo R-squ.:                  0.2727
Time:                        01:33:16   Log-Likelihood:                -407.06
converged:                       True   LL-Null:                       -559.70
Covariance Type:            nonrobust   LLR p-value:                 7.815e-65
==========================================================================================================
                                             coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
Intercept                                 14.2371      1.237     11.505      0.000      11.812      16.663
_Research_and_development_expense_rate 9.216e-11   3.75e-11      2.459      0.014    1.87e-11    1.66e-10
_Total_debt_to_Total_net_worth          1.882e-09   4.14e-10      4.548      0.000    1.07e-09    2.69e-09
_Equity_to_Liability                     -30.0678      6.380     -4.713      0.000     -42.573     -17.563
_Total_income_to_Total_expense         -7045.1925    574.236    -12.269      0.000   -8170.675   -5919.710
==========================================================================================================
```

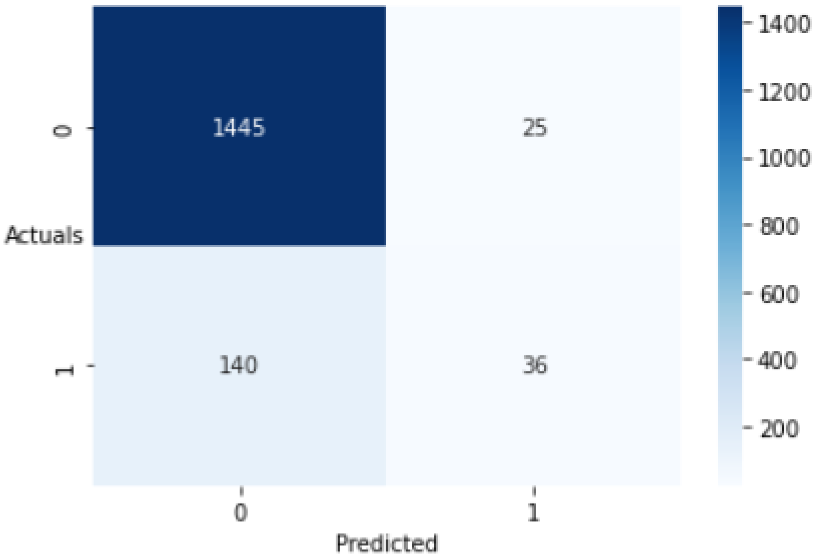
P-values of all the variables are less than 0.05 and thus all the coefficients are relevant. _Total_income_to_Total_expense has the highest coefficient and _Research_and_development_expense_rate the least of all. This model will be used to validate the test dataset.

PROBLEM 1.7

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

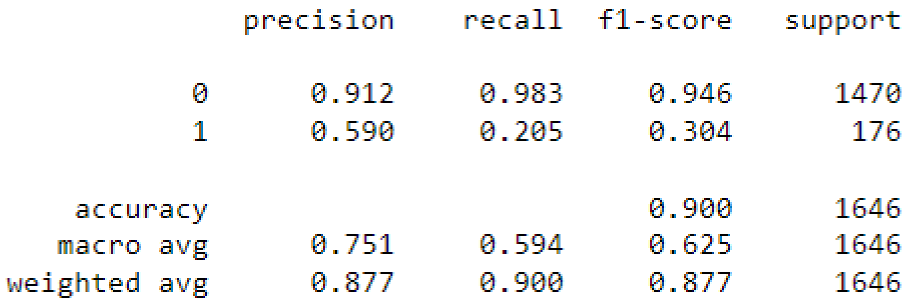
Resolution:

With default probability threshold of 0.5, the confusion matrix for the train set is as follows



Correctly predicted = 1445 incorrectly predicted records = 36

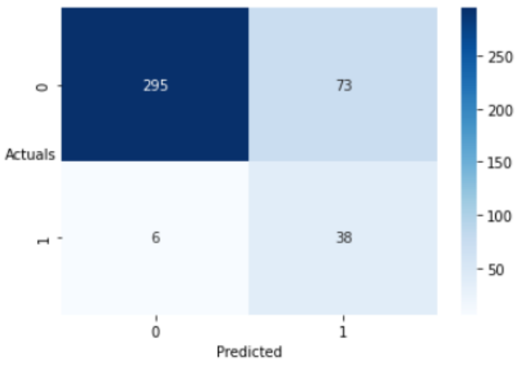
This was pretty good result on its own, however to further improve the on the results. We decided to look for the optimum threshold. After evaluating using the optimal threshold. Below was the new classification matrix.

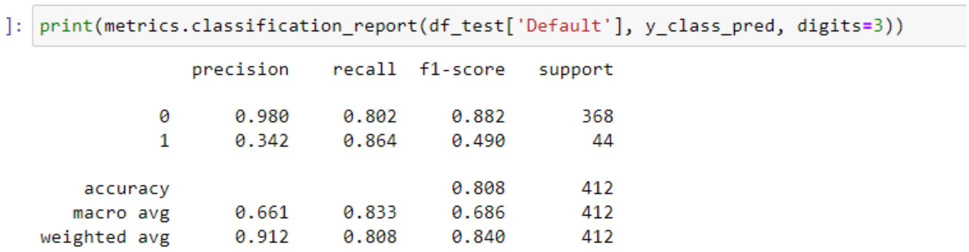


```
              precision    recall  f1-score   support

           0      0.912     0.983     0.946      1470
           1      0.590     0.205     0.304       176

    accuracy                          0.900      1646
   macro avg      0.751     0.594     0.625      1646
weighted avg      0.877     0.900     0.877      1646
```

Accuracy about 80% was achieved while recall, precision and f1 score were also very high at 80%,90% and 83% respectively.

We also evaluated the test data set for the same model which was built after the above mentioned re-iterative process. Below are statistics for the test model.





```
]: print(metrics.classification_report(df_test['Default'], y_class_pred, digits=3))
```

```
              precision    recall  f1-score   support

           0      0.980     0.802     0.882       368
           1      0.342     0.864     0.490        44

    accuracy                          0.808       412
   macro avg      0.661     0.833     0.686       412
weighted avg      0.912     0.808     0.840       412
```

Correctly predicted = 295 incorrectly predicted records = 38

Accuracy of 80% and very high recall, precision and f1 score of 80% ,91% and 84% respectively were also observed on the test set.