FINANCIAL RISK ANALYTICS PROJECT BUSINESS REPORT

PG-DSBA

Written by

Priyamvada Singh

Dated: **20-09-2023**

(Format: dd-mm-yyyy)

Table of Contents

Table of Figures

PART A

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 interest 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.

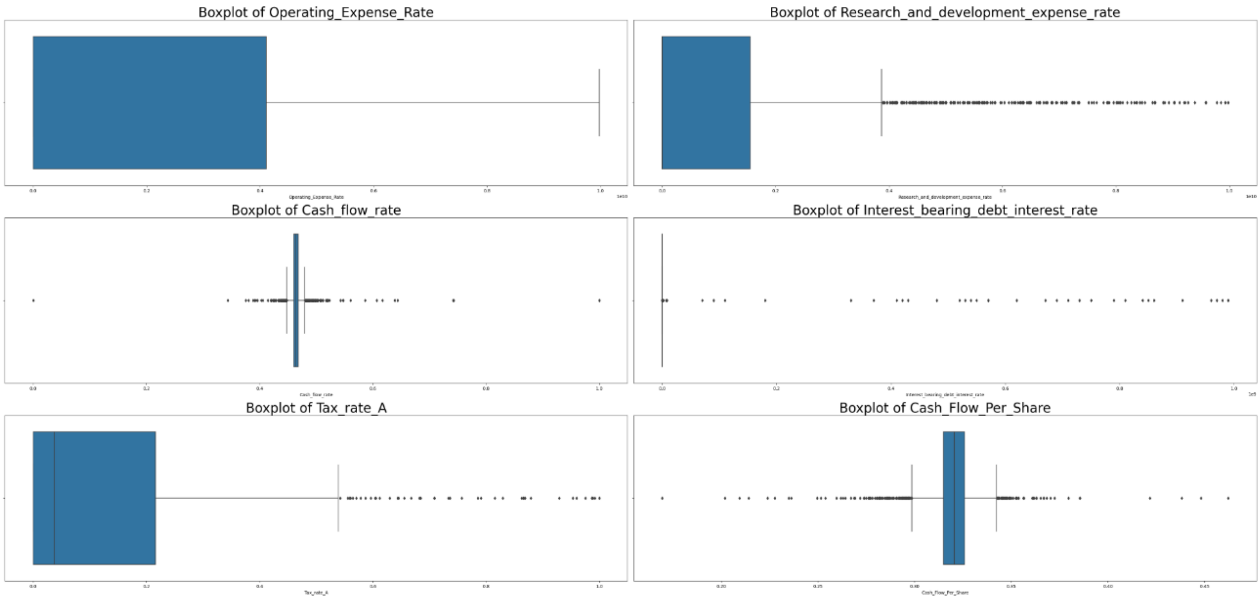
Data that is available includes information from the financial statement of the companies for the previous year.

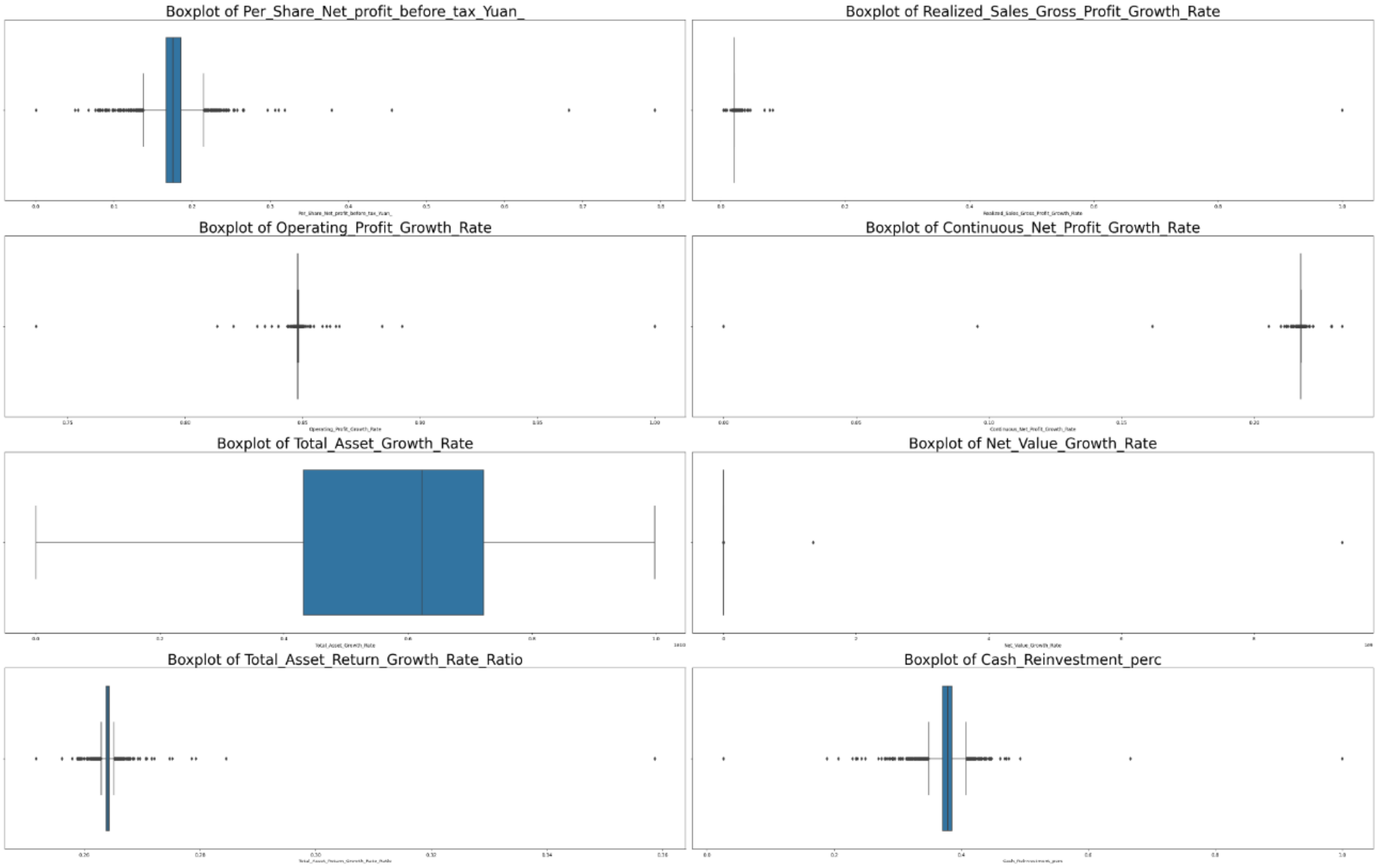
Dependent variable - No need to create any new variable, as the 'Default' variable is already provided in the dataset, which can be considered as the dependent variable.

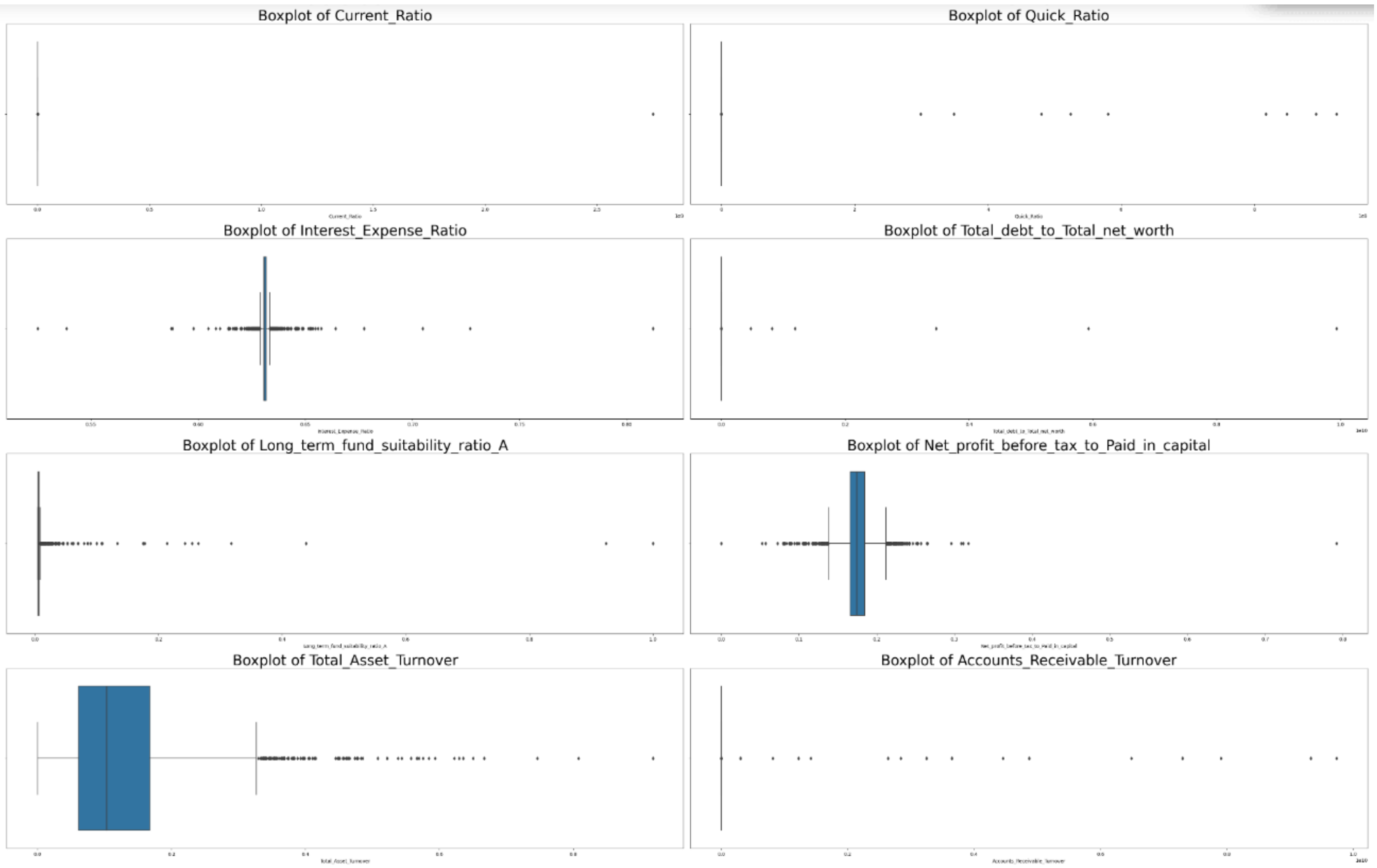
Test Train Split - Split the data into train and test datasets in the ratio of 67:33 and use a random state of 42 (*random_state=42*). Model building is to be done on the train dataset and model validation is to be done on the test dataset.

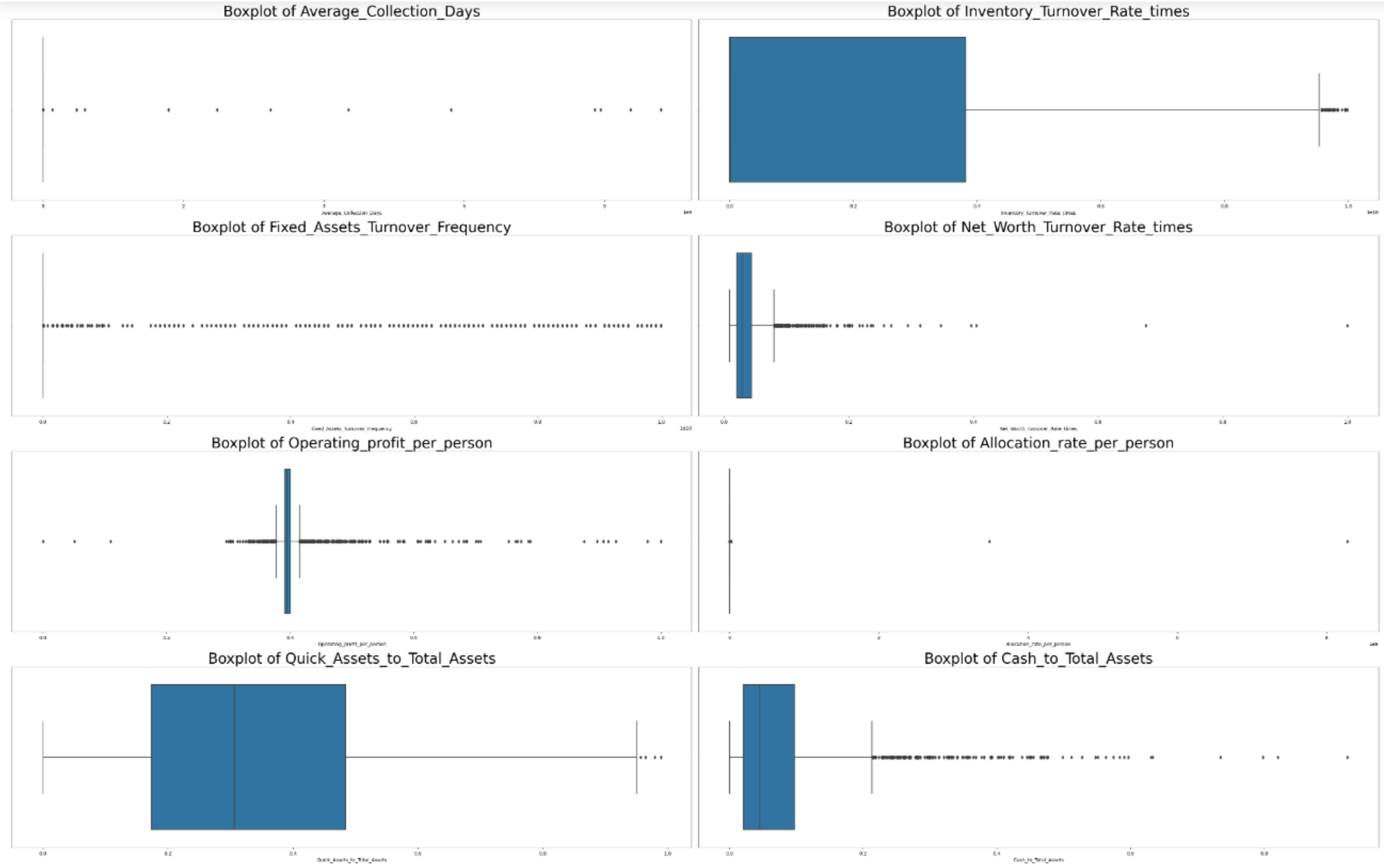
1. Outlier Treatment

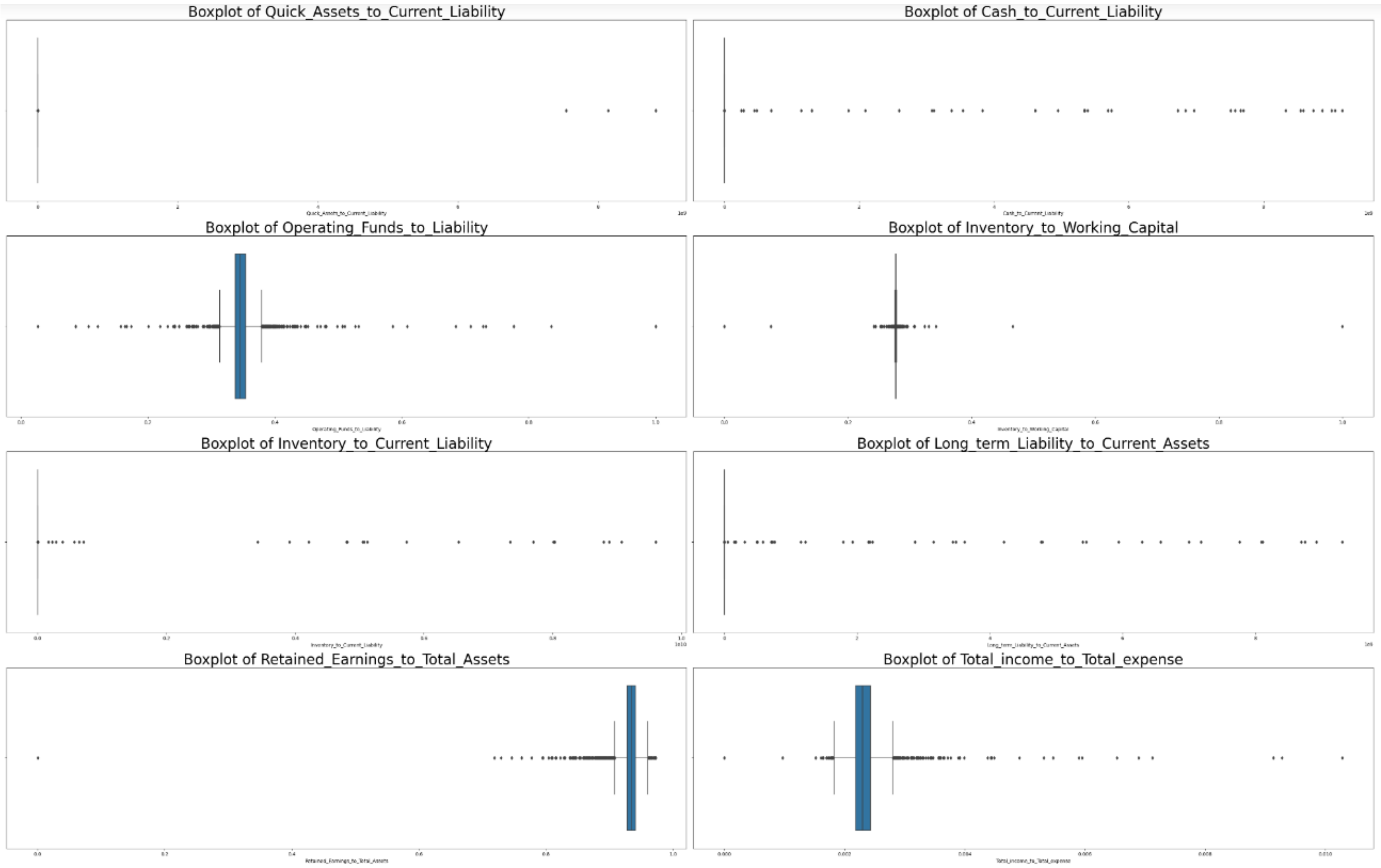
Figure 1 – Outliers - I

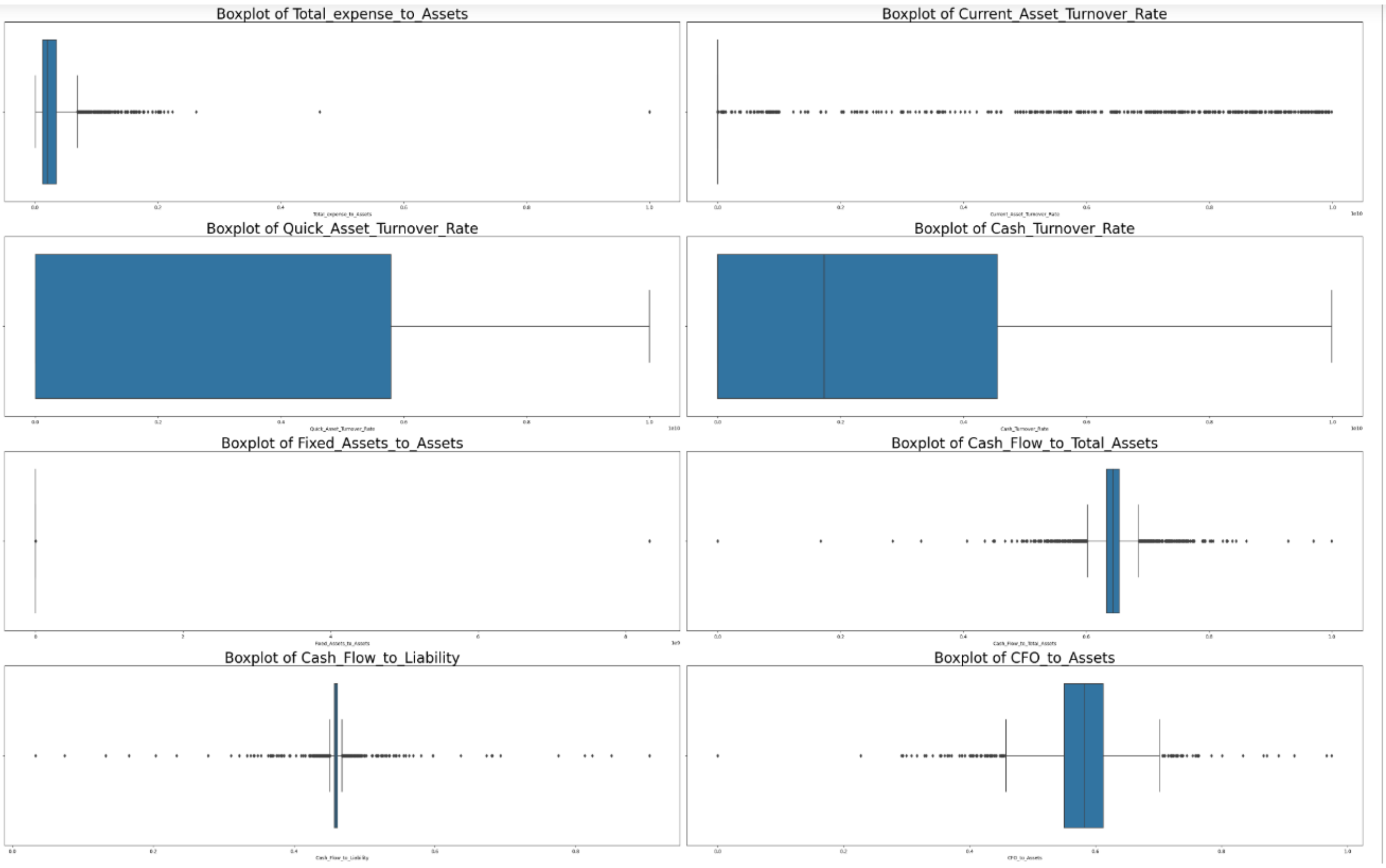


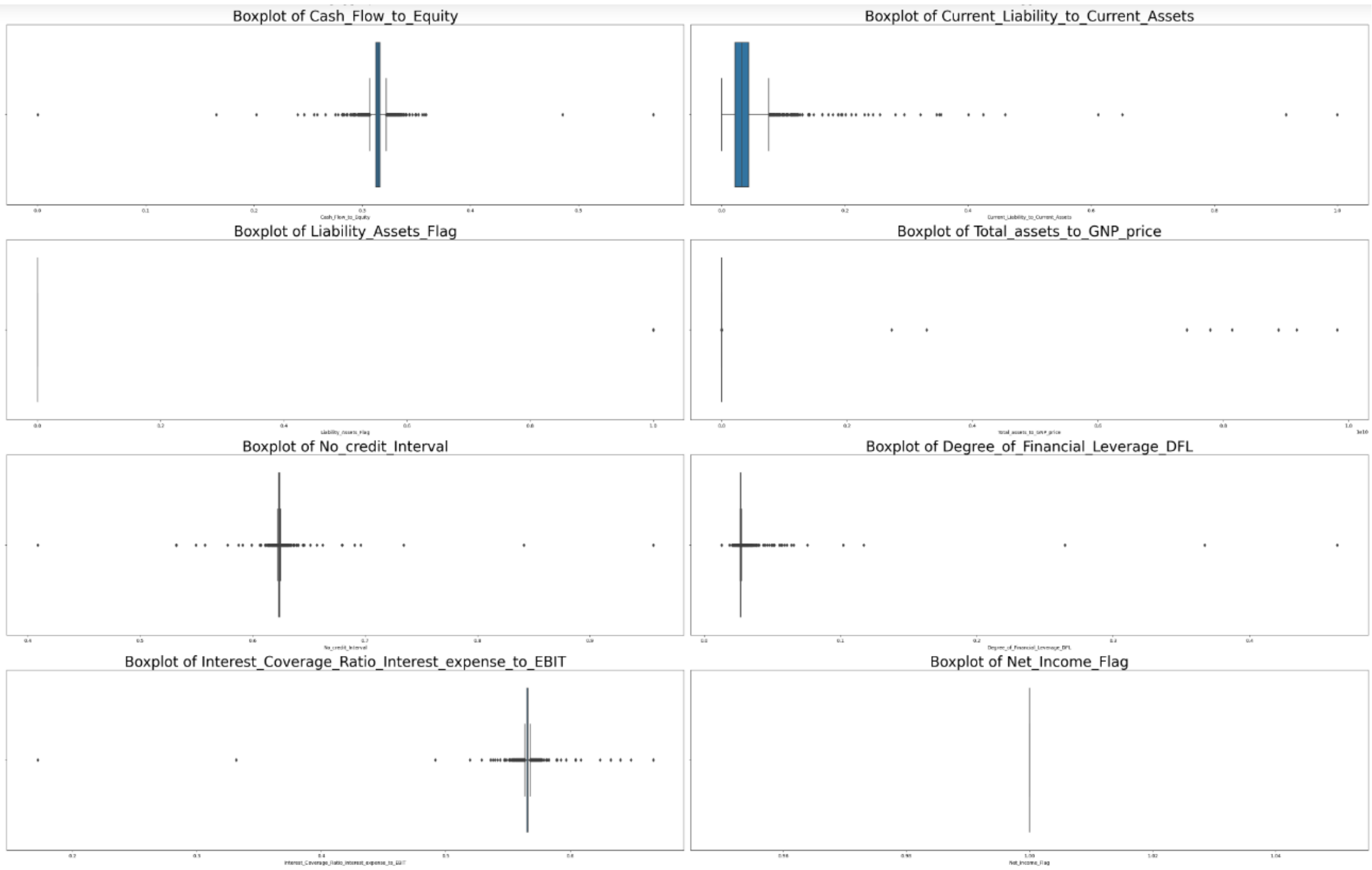

Boxplot of Per_Share_Net_profit_before_tax_Yuan_
Boxplot of Realized_Sales_Gross_Profit_Growth_Rate
Boxplot of Operating_Profit_Growth_Rate
Boxplot of Continuous_Net_Profit_Growth_Rate
Boxplot of Total_Asset_Growth_Rate
Boxplot of Net_Value_Growth_Rate
Boxplot of Total_Asset_Return_Growth_Rate_Ratio
Boxplot of Cash_Reinvestment_perc

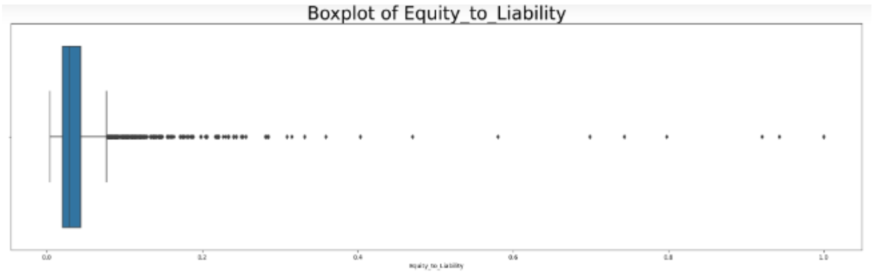

Boxplot of Current_Ratio
Boxplot of Quick_Ratio
Boxplot of Interest_Expense_Ratio
Boxplot of Total_debt_to_Total_net_worth
Boxplot of Long_term_fund_suitability_ratio_A
Boxplot of Net_profit_before_tax_to_Paid_in_capital
Boxplot of Total_Asset_Turnover
Boxplot of Accounts_Receivable_Turnover


Boxplot of Average_Collection_Days
Boxplot of Inventory_Turnover_Rate_times
Boxplot of Fixed_Assets_Turnover_Frequency
Boxplot of Net_Worth_Turnover_Rate_times
Boxplot of Operating_profit_per_person
Boxplot of Allocation_rate_per_person
Boxplot of Quick_Assets_to_Total_Assets
Boxplot of Cash_to_Total_Assets


Boxplot of Quick_Assets_to_Current_Liability
Boxplot of Cash_to_Current_Liability
Boxplot of Operating_Funds_to_Liability
Boxplot of Inventory_to_Working_Capital
Boxplot of Inventory_to_Current_Liability
Boxplot of Long_term_Liability_to_Current_Assets
Boxplot of Retained_Earnings_to_Total_Assets
Boxplot of Total_income_to_Total_expense


Boxplot of Total_expense_to_Assets
Boxplot of Current_Asset_Turnover_Rate
Boxplot of Quick_Asset_Turnover_Rate
Boxplot of Cash_Turnover_Rate
Boxplot of Fixed_Assets_to_Assets
Boxplot of Cash_Flow_to_Total_Assets
Boxplot of Cash_Flow_to_Liability
Boxplot of CFO_to_Assets

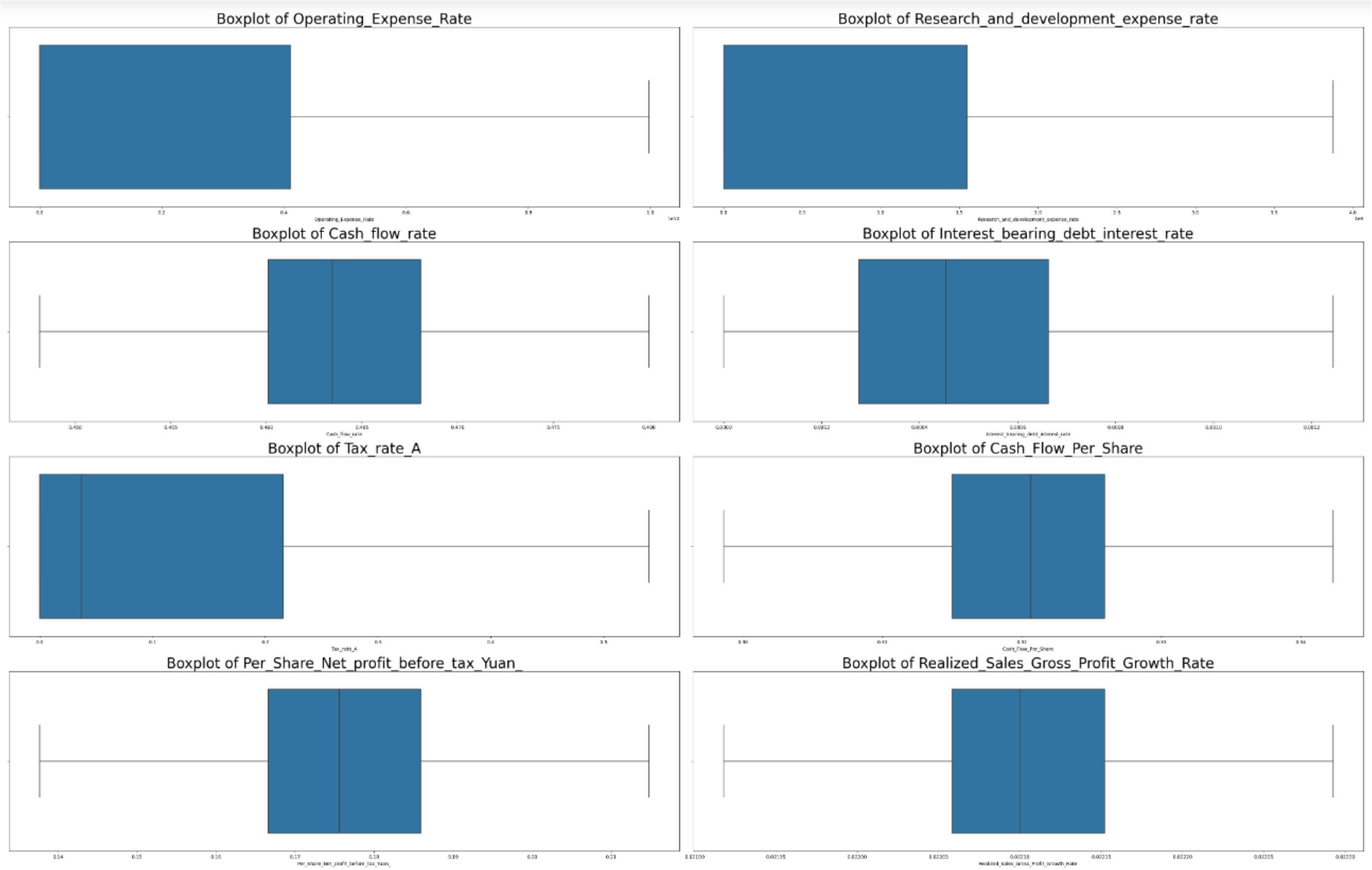

Boxplot of Cash_Flow_to_Equity
Boxplot of Current_Liability_to_Current_Assets
Boxplot of Liability_Assets_Flag
Boxplot of Total_assets_to_GNP_price
Boxplot of No_credit_Interval
Boxplot of Degree_of_Financial_Leverage_DFL
Boxplot of Interest_Coverage_Ratio_Interest_expense_to_EBIT
Boxplot of Net_Income_Flag

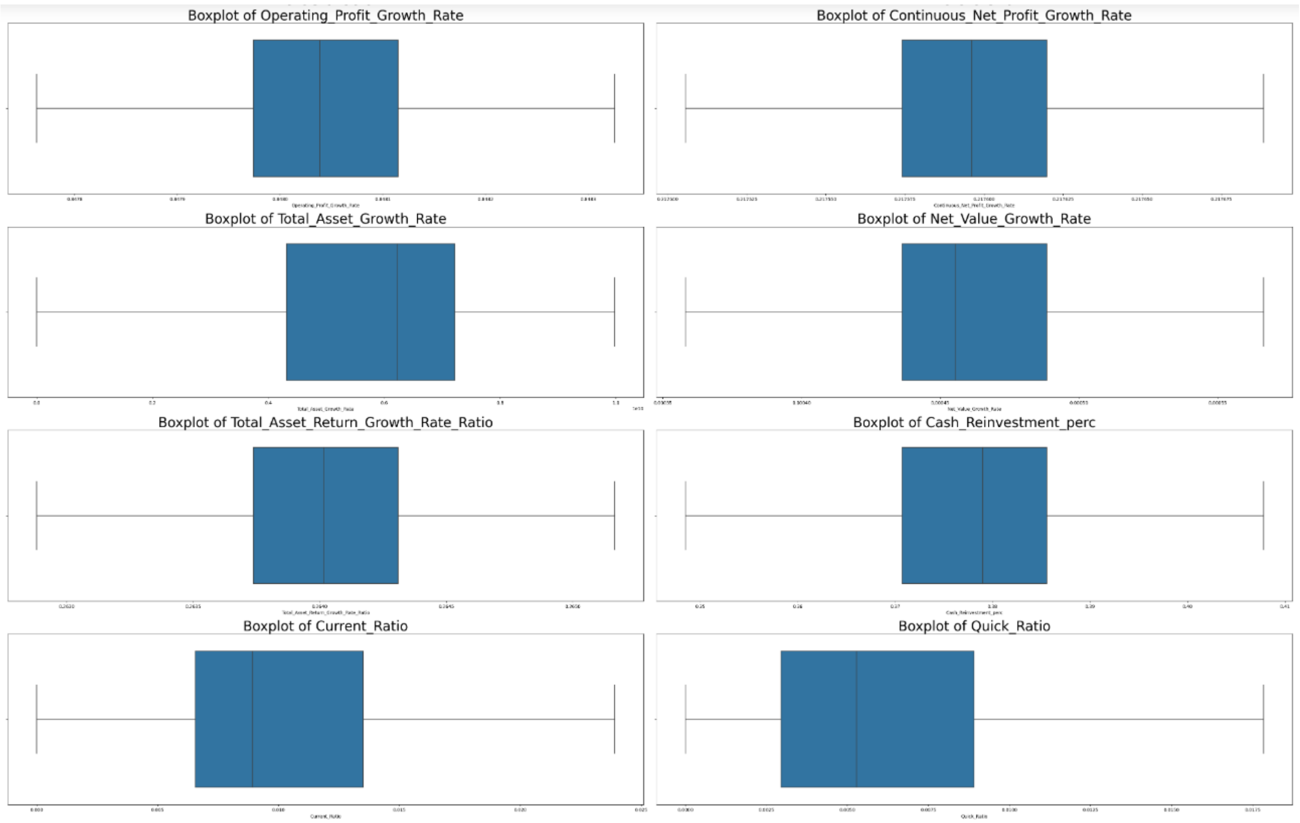


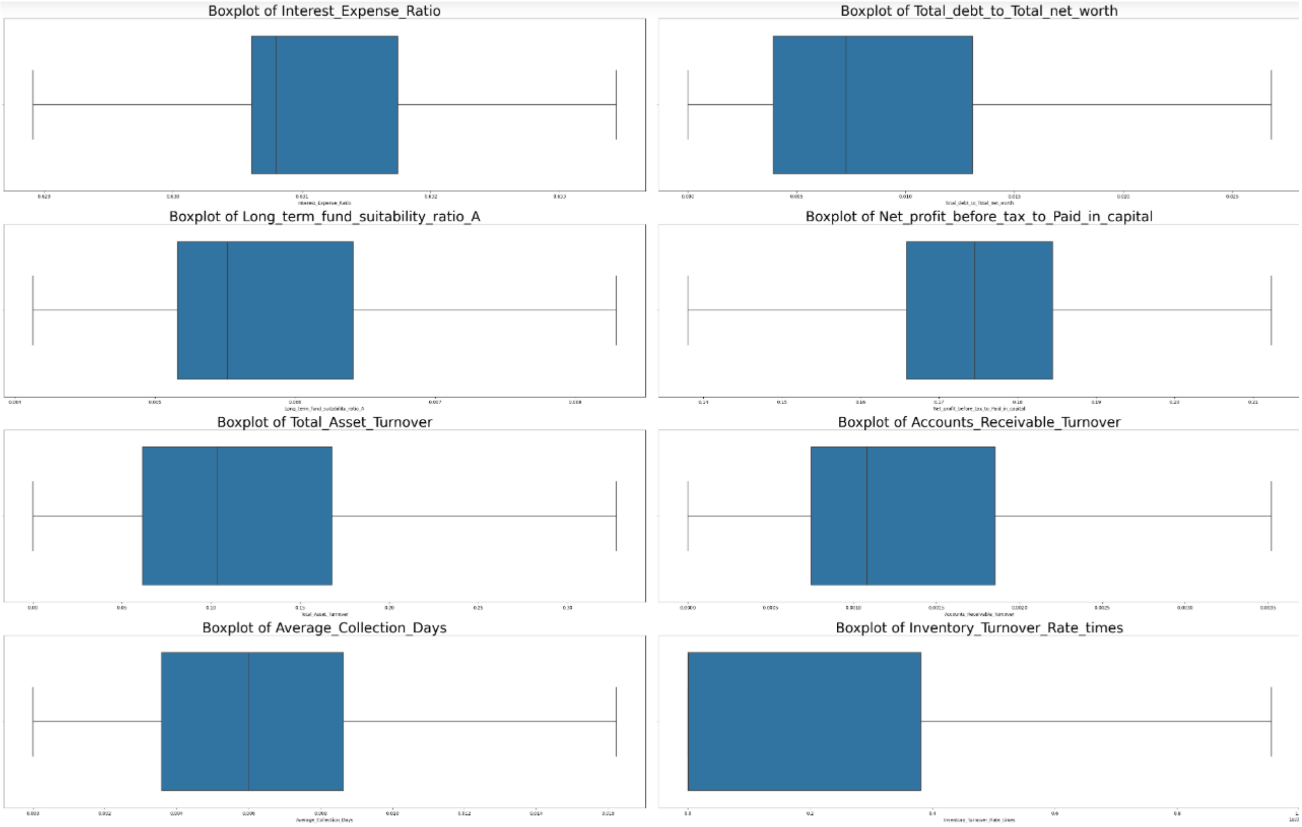
i. <u>Inferences summary:</u>

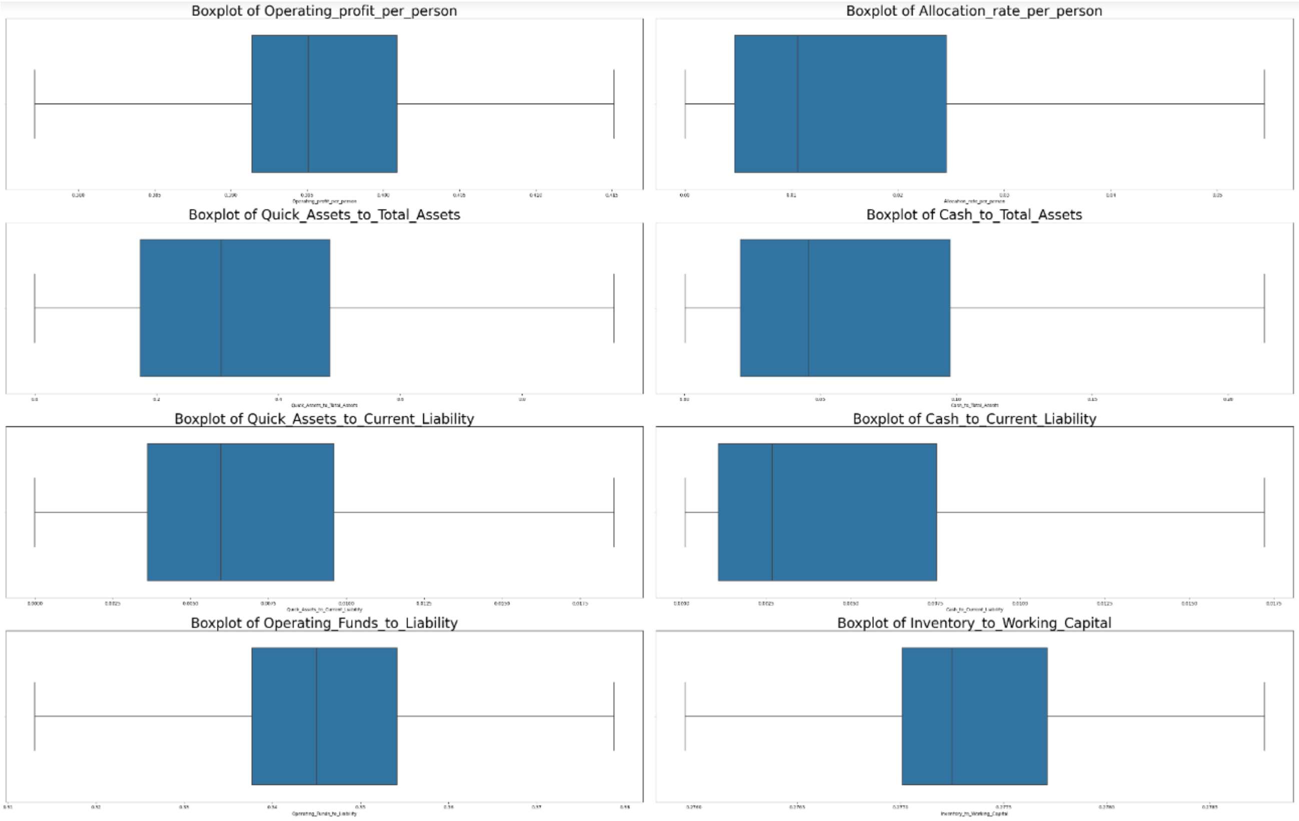
- There are 9.598% outliers out of the total values in the dataset.
- Outlier percentage calculated as (10864/113190)*100
- All the variables except Quick_Asset_Turnover_Rate, Cash_Turnover_Rate, Net_Income_Flag, and Liability_Assets_Flags have outliers.
- The response variable has been excluded from this exercise. The two values that it takes is 0 and 1.
- Outliers have now been treated. The below screenshot shows the revised boxplots:

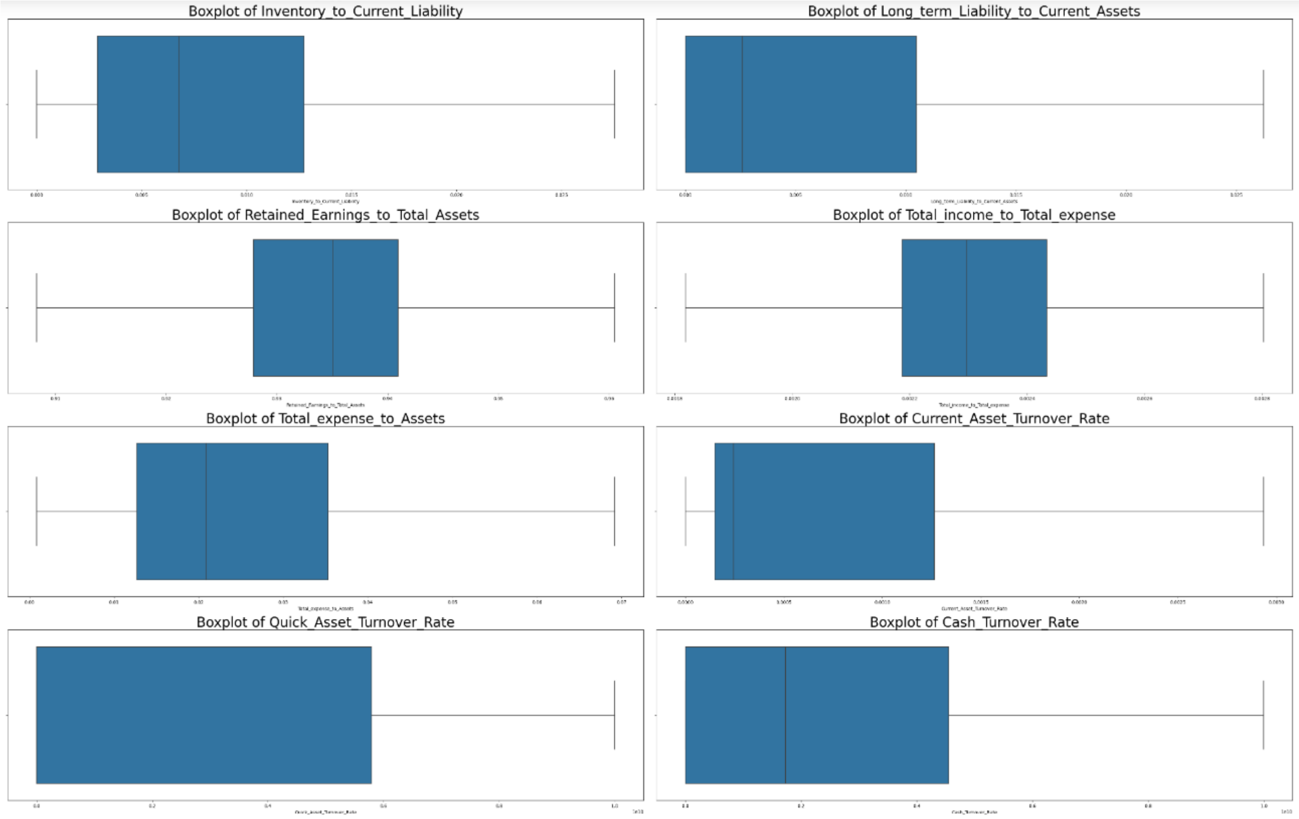
Figure 2 – Outlier Treatment With IQR

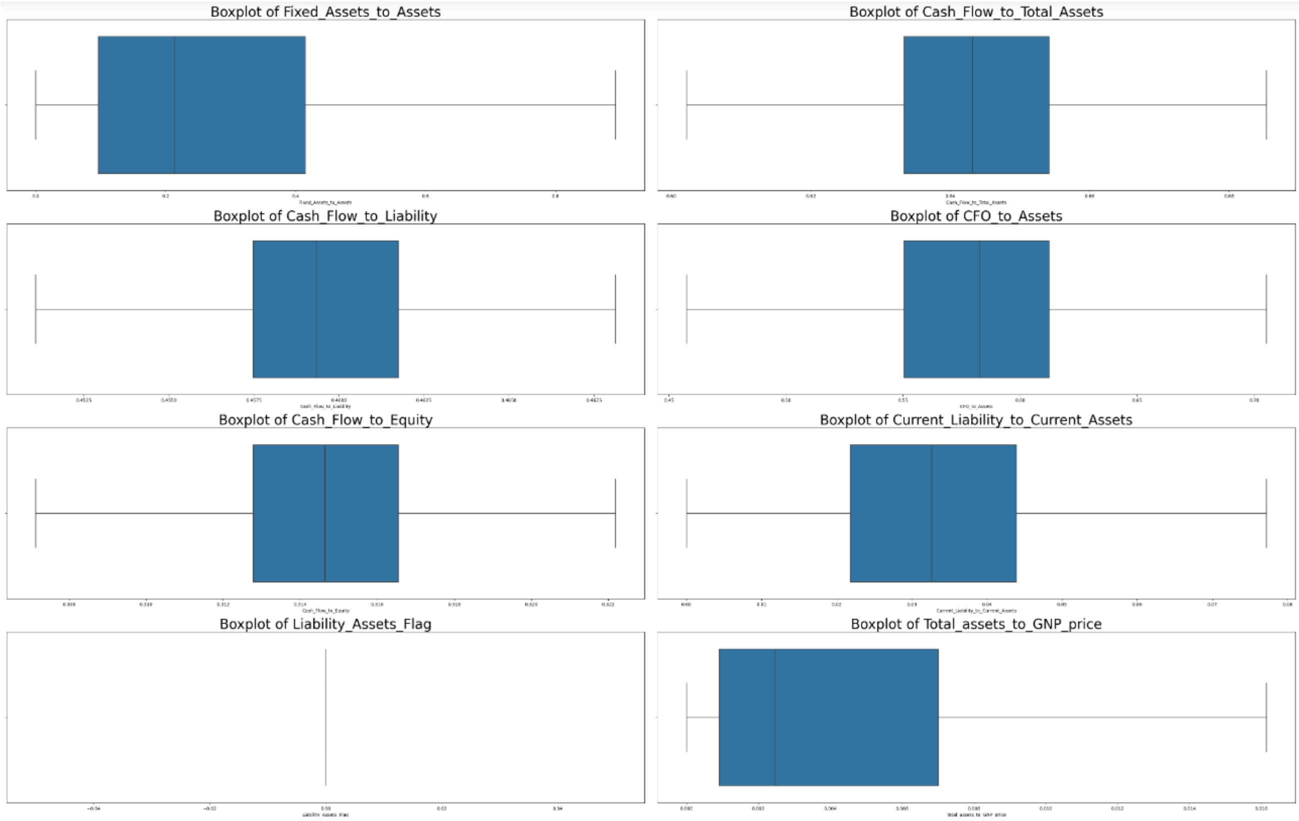


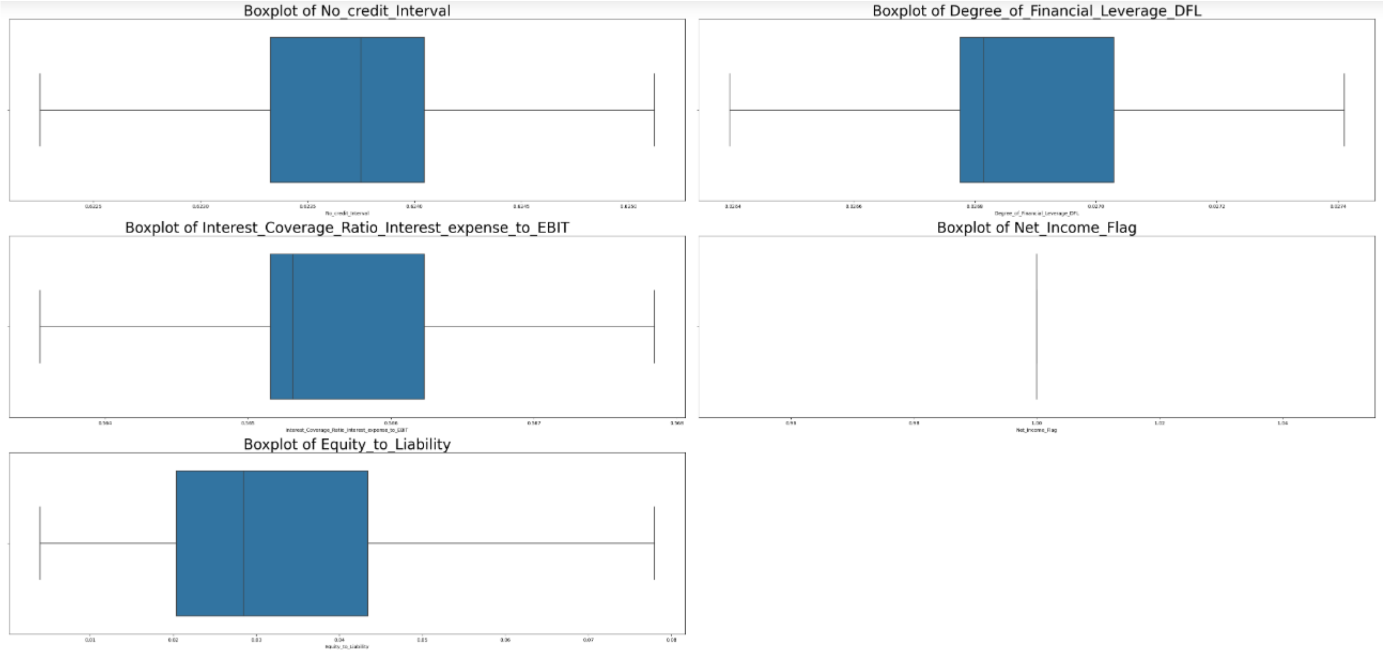

Boxplot of Operating_Profit_Growth_Rate
Boxplot of Continuous_Net_Profit_Growth_Rate
Boxplot of Total_Asset_Growth_Rate
Boxplot of Net_Value_Growth_Rate
Boxplot of Total_Asset_Return_Growth_Rate_Ratio
Boxplot of Cash_Reinvestment_perc
Boxplot of Current_Ratio
Boxplot of Quick_Ratio


Boxplot of Interest_Expense_Ratio
Boxplot of Total_debt_to_Total_net_worth
Boxplot of Long_term_fund_suitability_ratio_A
Boxplot of Net_profit_before_tax_to_Paid_in_capital
Boxplot of Total_Asset_Turnover
Boxplot of Accounts_Receivable_Turnover
Boxplot of Average_Collection_Days
Boxplot of Inventory_Turnover_Rate_times


Boxplot of Operating_profit_per_person
Boxplot of Allocation_rate_per_person
Boxplot of Quick_Assets_to_Total_Assets
Boxplot of Cash_to_Total_Assets
Boxplot of Quick_Assets_to_Current_Liability
Boxplot of Cash_to_Current_Liability
Boxplot of Operating_Funds_to_Liability
Boxplot of Inventory_to_Working_Capital


Boxplot of Inventory_to_Current_Liability
Boxplot of Long_term_Liability_to_Current_Assets
Boxplot of Retained_Earnings_to_Total_Assets
Boxplot of Total_income_to_Total_expense
Boxplot of Total_expense_to_Assets
Boxplot of Current_Asset_Turnover_Rate
Boxplot of Quick_Asset_Turnover_Rate
Boxplot of Cash_Turnover_Rate





2. Missing Value Treatment

i. There are 0.25% missing values in the data.

ii. Following are the variables with missing values:

- Cash_Flow_Per_Share ➔ 167
- Total_debt_to_Total_net_worth ➔ 21

- Cash_to_Total_Assets ➔ 96
- Current_Liability_to_Current_Assets ➔ 14

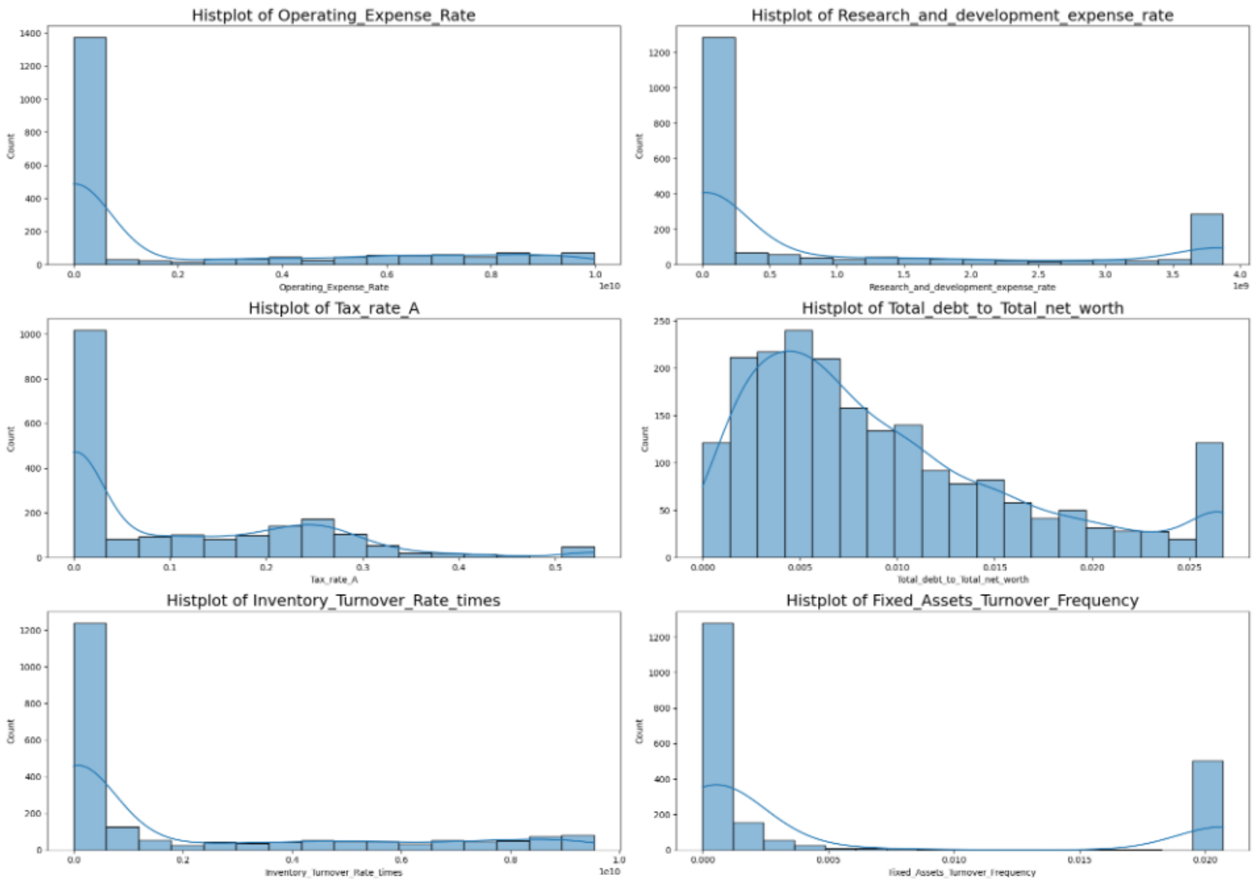
iii. These values have been treated through the KNN imputer where the 'n' neighbours value is set to 5. This imputer only uses the 5 nearest neighbours to impute the missing values instead of the whole column, making it a better way to impute missing values.

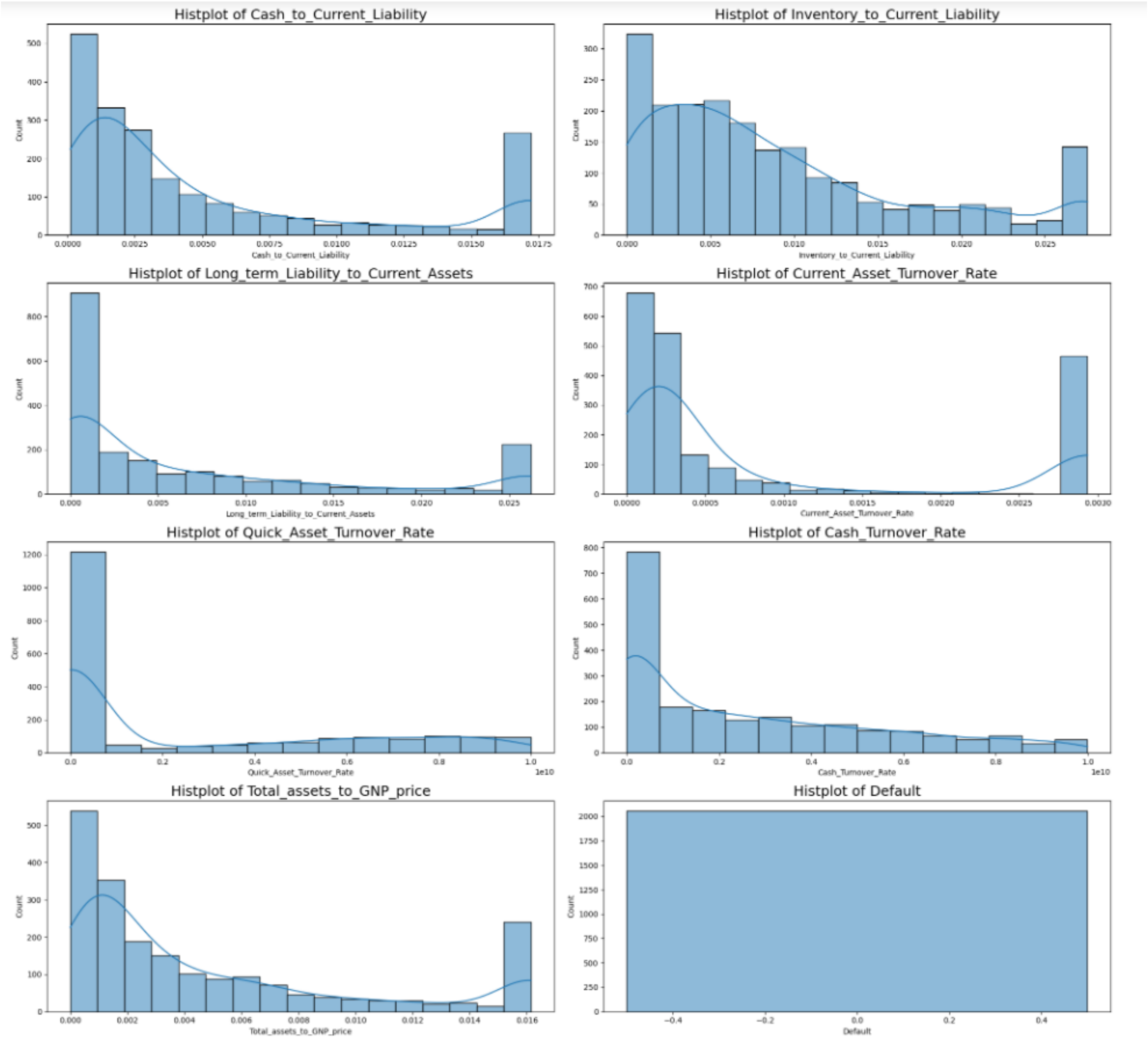
3. Univariate (4 marks) & Bivariate (6 marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

i. Univariate Analysis

- Distribution for each significant variable used in the model building:

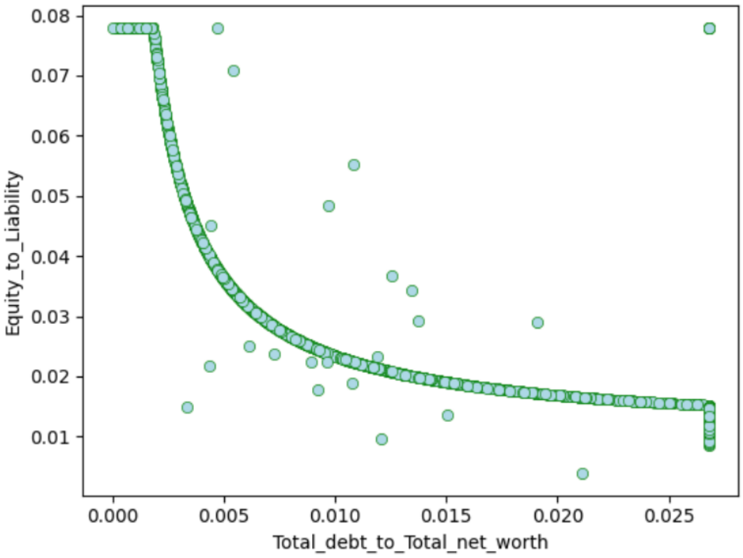
Figure 3 - Histogram for all significant variables





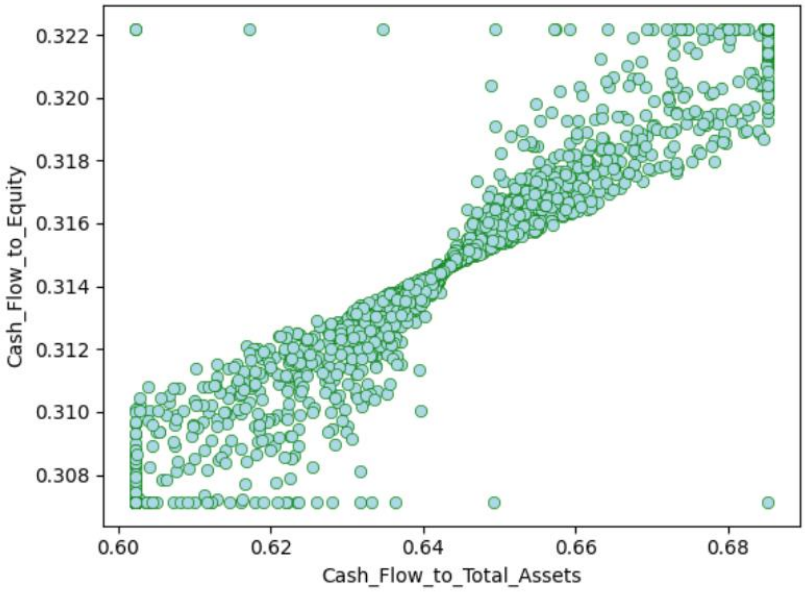
ii. Bivariate Analysis

Figure 4 - Scatterplot I



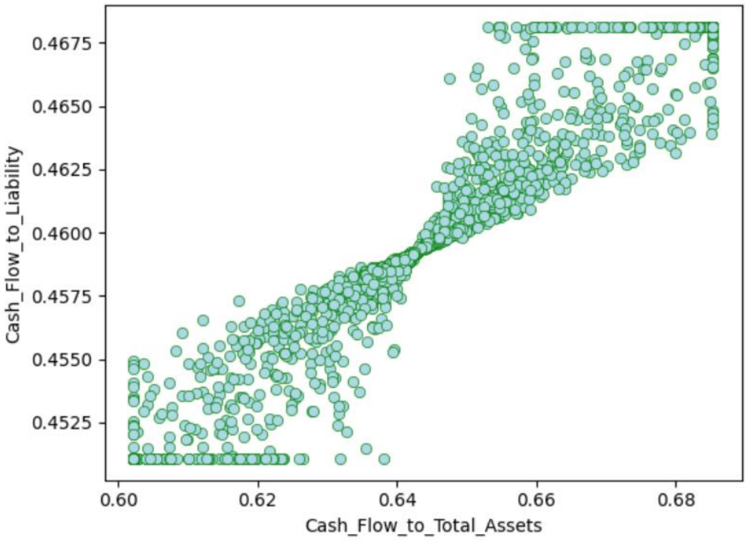
- **Total debt to total net worth** and **equity to liability** are showing a negative correlation of -0.75 as shown in the above graph.

Figure 5 - Scatterplot II



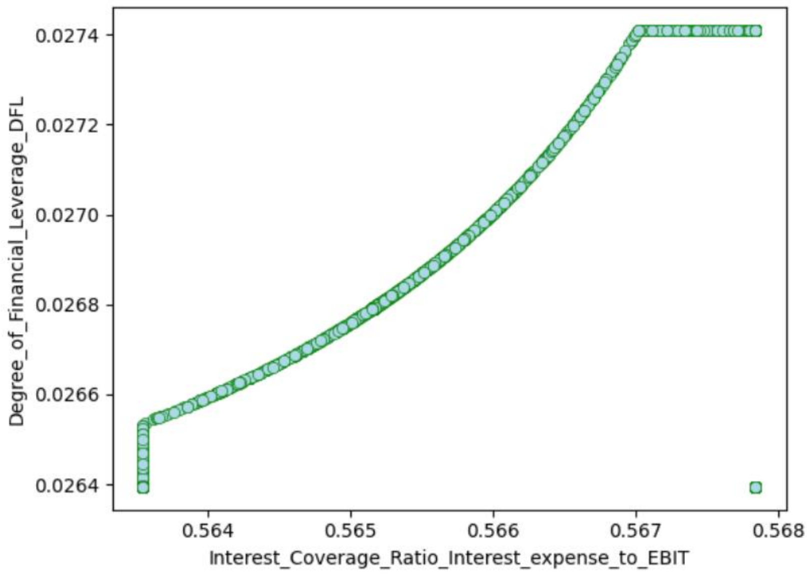
- **Cash flow to total assets** and **cash flow to equity** are showing a strong positive correlation of 0.95 as shown in the above graph.

Figure 6 - Scatterplot III



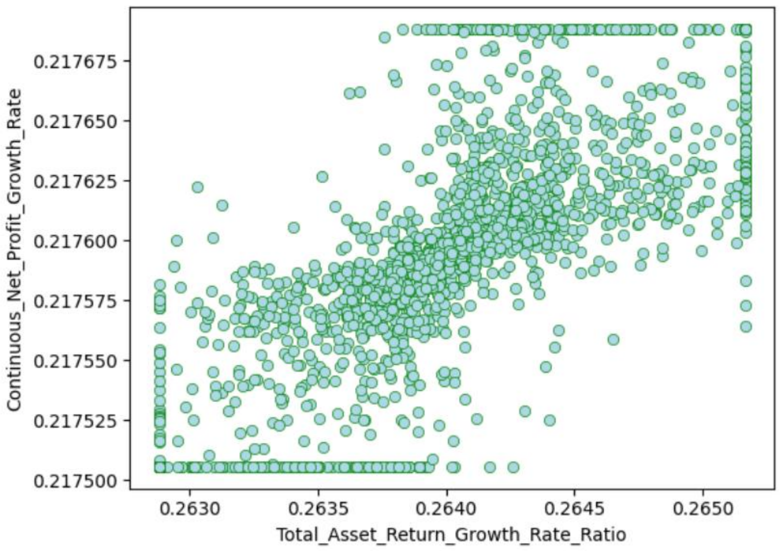
- **Cash flow to total assets** and **cash flow to liability** are showing a strong positive correlation of 0.96 as shown in the above graph.

Figure 7 - Scatterplot IV



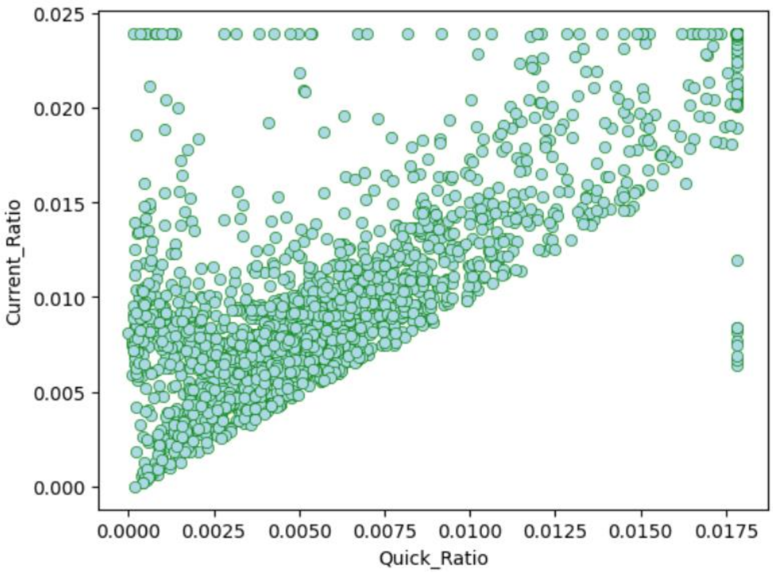
- **Interest coverage ratio interest expense to EBIT** and **degree of financial leverage DFL** are showing a positive correlation of 0.81 as shown in the above graph.

Figure 8 - Scatterplot V



- **Total asset return growth rate ratio** and **continuous net profit growth rate** are showing a positive correlation of 0.75 as shown in the above graph.

Figure 9 - Scatterplot VI

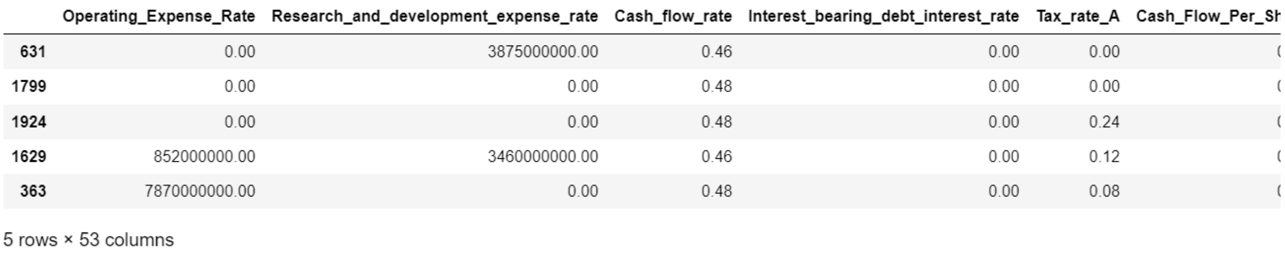


- **Quick ratio** and **current ratio** are showing a positive correlation of 0.81 as shown in the above graph.

4. Train Test Split

i. X_train sample:

Figure 10 - X_train dataset sample

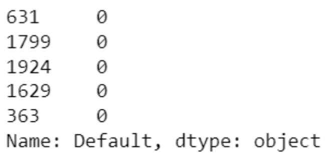


	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Sh
631	0.00	3875000000.00	0.46	0.00	0.00	(
1799	0.00	0.00	0.48	0.00	0.00	(
1924	0.00	0.00	0.48	0.00	0.24	(
1629	852000000.00	3460000000.00	0.46	0.00	0.12	(
363	7870000000.00	0.00	0.48	0.00	0.08	(

5 rows × 53 columns

iii. y_train sample:

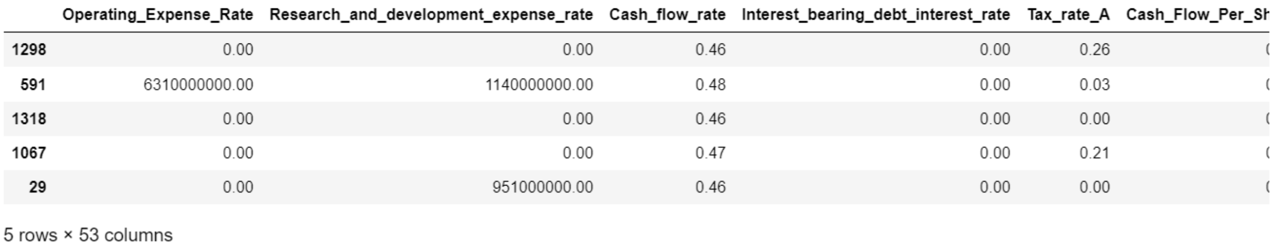
Figure 11 - y_train dataset sample



```
631     0
1799    0
1924    0
1629    0
363     0
Name: Default, dtype: object
```

iv. X_test sample:

Figure 12 - X_test dataset sample

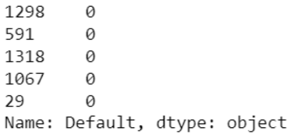


	Operating_Expense_Rate	Research_and_development_expense_rate	Cash_flow_rate	Interest_bearing_debt_interest_rate	Tax_rate_A	Cash_Flow_Per_Sh
1298	0.00	0.00	0.46	0.00	0.26	(
591	6310000000.00	1140000000.00	0.48	0.00	0.03	(
1318	0.00	0.00	0.46	0.00	0.00	(
1067	0.00	0.00	0.47	0.00	0.21	(
29	0.00	951000000.00	0.46	0.00	0.00	(

5 rows × 53 columns

v. y_test sample:

Figure 13 - y_test dataset sample

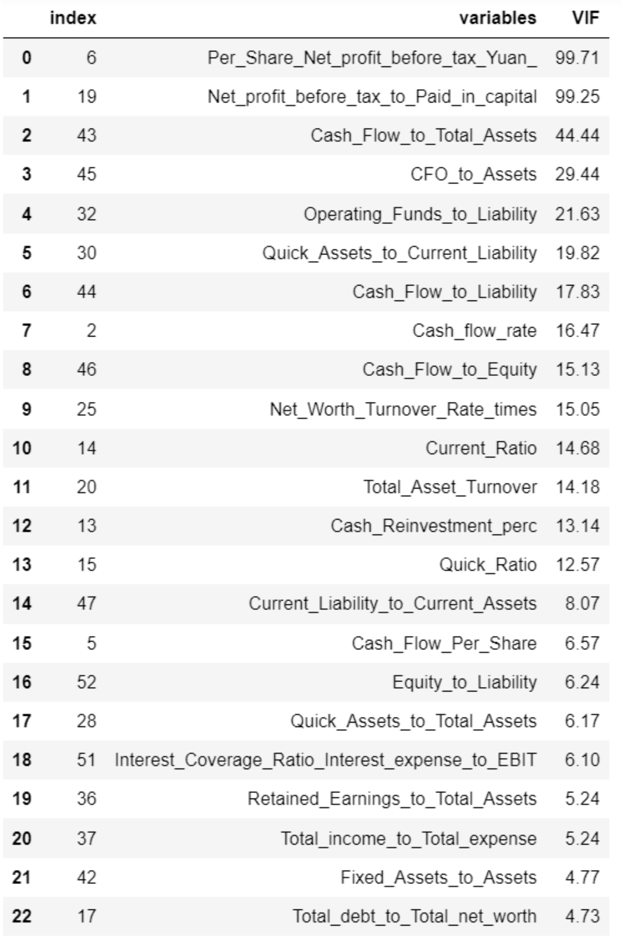


```
1298    0
591     0
1318    0
1067    0
29      0
Name: Default, dtype: object
```

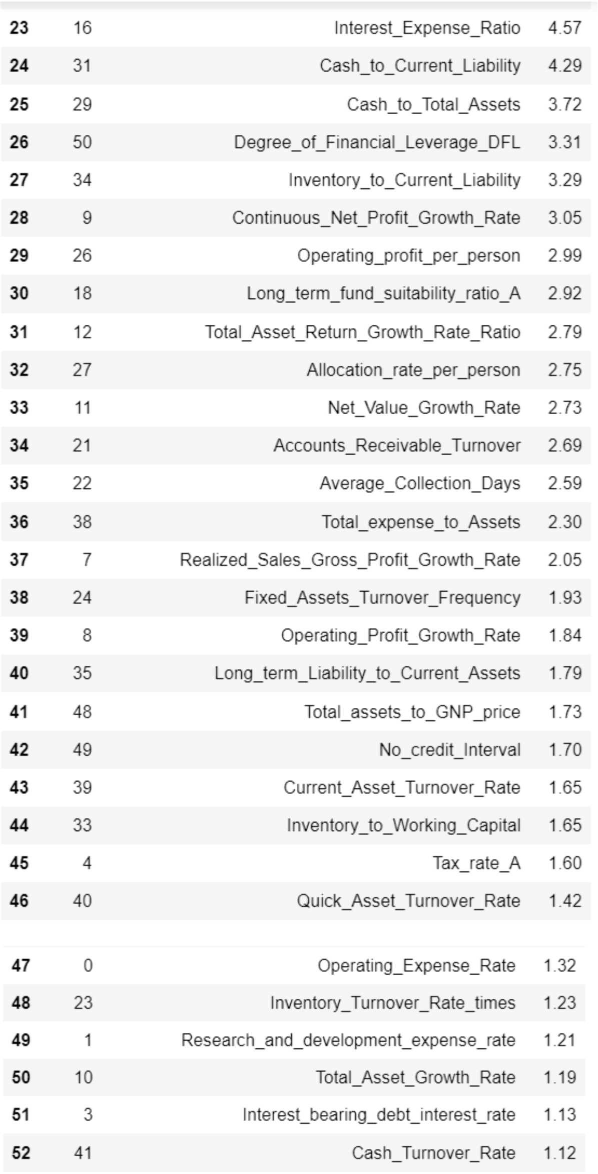
5. Build Logistic Regression Model (using statsmodels library) on most important variables on train dataset and choose the optimum cut-off. Also showcase your model building approach

i. Variance Inflation Factor for each variable:

Figure 14 - Variance Inflation Factor for all variables



	index	variables	VIF
0	6	Per_Share_Net_profit_before_tax_Yuan_	99.71
1	19	Net_profit_before_tax_to_Paid_in_capital	99.25
2	43	Cash_Flow_to_Total_Assets	44.44
3	45	CFO_to_Assets	29.44
4	32	Operating_Funds_to_Liability	21.63
5	30	Quick_Assets_to_Current_Liability	19.82
6	44	Cash_Flow_to_Liability	17.83
7	2	Cash_flow_rate	16.47
8	46	Cash_Flow_to_Equity	15.13
9	25	Net_Worth_Turnover_Rate_times	15.05
10	14	Current_Ratio	14.68
11	20	Total_Asset_Turnover	14.18
12	13	Cash_Reinvestment_perc	13.14
13	15	Quick_Ratio	12.57
14	47	Current_Liability_to_Current_Assets	8.07
15	5	Cash_Flow_Per_Share	6.57
16	52	Equity_to_Liability	6.24
17	28	Quick_Assets_to_Total_Assets	6.17
18	51	Interest_Coverage_Ratio_Interest_expense_to_EBIT	6.10
19	36	Retained_Earnings_to_Total_Assets	5.24
20	37	Total_income_to_Total_expense	5.24
21	42	Fixed_Assets_to_Assets	4.77
22	17	Total_debt_to_Total_net_worth	4.73

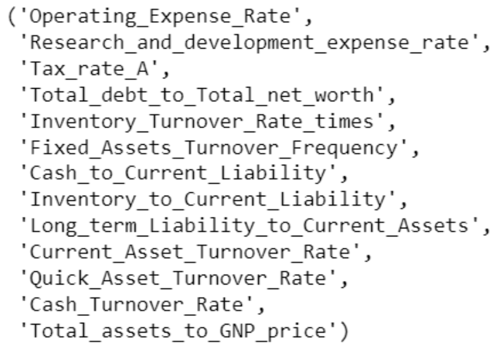


23	16	Interest_Expense_Ratio	4.57
24	31	Cash_to_Current_Liability	4.29
25	29	Cash_to_Total_Assets	3.72
26	50	Degree_of_Financial_Leverage_DFL	3.31
27	34	Inventory_to_Current_Liability	3.29
28	9	Continuous_Net_Profit_Growth_Rate	3.05
29	26	Operating_profit_per_person	2.99
30	18	Long_term_fund_suitability_ratio_A	2.92
31	12	Total_Asset_Return_Growth_Rate_Ratio	2.79
32	27	Allocation_rate_per_person	2.75
33	11	Net_Value_Growth_Rate	2.73
34	21	Accounts_Receivable_Turnover	2.69
35	22	Average_Collection_Days	2.59
36	38	Total_expense_to_Assets	2.30
37	7	Realized_Sales_Gross_Profit_Growth_Rate	2.05
38	24	Fixed_Assets_Turnover_Frequency	1.93
39	8	Operating_Profit_Growth_Rate	1.84
40	35	Long_term_Liability_to_Current_Assets	1.79
41	48	Total_assets_to_GNP_price	1.73
42	49	No_credit_Interval	1.70
43	39	Current_Asset_Turnover_Rate	1.65
44	33	Inventory_to_Working_Capital	1.65
45	4	Tax_rate_A	1.60
46	40	Quick_Asset_Turnover_Rate	1.42
47	0	Operating_Expense_Rate	1.32
48	23	Inventory_Turnover_Rate_times	1.23
49	1	Research_and_development_expense_rate	1.21
50	10	Total_Asset_Growth_Rate	1.19
51	3	Interest_bearing_debt_interest_rate	1.13
52	41	Cash_Turnover_Rate	1.12

- Generally, if the VIF is below 5, the variable is considered to not contribute to the multicollinearity in the data.
- However, multiple iterations of VIF were run on the data to shortlist the 13 top variables to be used in the Logit model.

vi. The first Logit model used the following variables:

Figure 15 - Final variables used in the Logit model



```
('Operating_Expense_Rate',
 'Research_and_development_expense_rate',
 'Tax_rate_A',
 'Total_debt_to_Total_net_worth',
 'Inventory_Turnover_Rate_times',
 'Fixed_Assets_Turnover_Frequency',
 'Cash_to_Current_Liability',
 'Inventory_to_Current_Liability',
 'Long_term_Liability_to_Current_Assets',
 'Current_Asset_Turnover_Rate',
 'Quick_Asset_Turnover_Rate',
 'Cash_Turnover_Rate',
 'Total_assets_to_GNP_price')
```

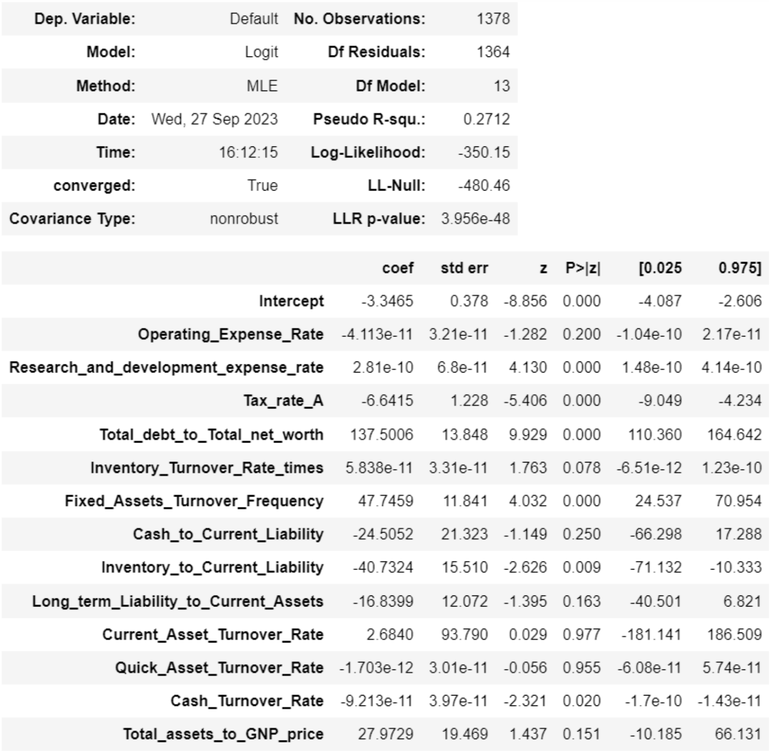
vii. The Logit formula to build the model is as follows:

- *'Default ~ Operating_Expense_Rate + Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Cash_to_Current_Liability + Inventory_to_Current_Liability + Long_term_Liability_to_Current_Assets + Current_Asset_Turnover_Rate + Quick_Asset_Turnover_Rate + Cash_Turnover_Rate + Total_assets_to_GNP_price'*

viii. Unlike the scikit-learn library, statsmodel uses manual insertion of the Logit formula. It also does not require response and predictor variables to be split into two separate data frames. Only two datasets were required to build this model – train and test datasets.

- Logit Model 1:

Figure 16 - Logit model I



Dep. Variable:	Default	**No. Observations:**	1378
Model:	Logit	**Df Residuals:**	1364
Method:	MLE	**Df Model:**	13
Date:	Wed, 27 Sep 2023	**Pseudo R-squ.:**	0.2712
Time:	16:12:15	**Log-Likelihood:**	-350.15
converged:	True	**LL-Null:**	-480.46
Covariance Type:	nonrobust	**LLR p-value:**	3.956e-48

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-3.3465	0.378	-8.856	0.000	-4.087	-2.606
Operating_Expense_Rate	-4.113e-11	3.21e-11	-1.282	0.200	-1.04e-10	2.17e-11
Research_and_development_expense_rate	2.81e-10	6.8e-11	4.130	0.000	1.48e-10	4.14e-10
Tax_rate_A	-6.6415	1.228	-5.406	0.000	-9.049	-4.234
Total_debt_to_Total_net_worth	137.5006	13.848	9.929	0.000	110.360	164.642
Inventory_Turnover_Rate_times	5.838e-11	3.31e-11	1.763	0.078	-6.51e-12	1.23e-10
Fixed_Assets_Turnover_Frequency	47.7459	11.841	4.032	0.000	24.537	70.954
Cash_to_Current_Liability	-24.5052	21.323	-1.149	0.250	-66.298	17.288
Inventory_to_Current_Liability	-40.7324	15.510	-2.626	0.009	-71.132	-10.333
Long_term_Liability_to_Current_Assets	-16.8399	12.072	-1.395	0.163	-40.501	6.821
Current_Asset_Turnover_Rate	2.6840	93.790	0.029	0.977	-181.141	186.509
Quick_Asset_Turnover_Rate	-1.703e-12	3.01e-11	-0.056	0.955	-6.08e-11	5.74e-11
Cash_Turnover_Rate	-9.213e-11	3.97e-11	-2.321	0.020	-1.7e-10	-1.43e-11
Total_assets_to_GNP_price	27.9729	19.469	1.437	0.151	-10.185	66.131

- While the VIF helps treat the multicollinearity of the data, the p-values of the variables shown in the snapshot above talk about the importance of a predictor feature with respect to the response variable.
- We will keep running iterations of the Logit model each time we eliminate a variable with the p-value of more than 0.05.

- Logit Model 2 with the updated formula (Quick_Asset_Turnover_Rate removed):

  'Default ~ Operating_Expense_Rate + Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Cash_to_Current_Liability + Inventory_to_Current_Liability + Long_term_Liability_to_Current_Assets + Current_Asset_Turnover_Rate + Cash_Turnover_Rate + Total_assets_to_GNP_price'
- Logit Model 3 with the updated formula (Current_Asset_Turnover_Rate removed):

  'Default ~ Operating_Expense_Rate + Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Cash_to_Current_Liability + Inventory_to_Current_Liability + Long_term_Liability_to_Current_Assets + Cash_Turnover_Rate + Total_assets_to_GNP_price'

- Logit Model 4 with the updated formula (Cash_to_Current_Liability removed):

  'Default ~ Operating_Expense_Rate + Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Inventory_to_Current_Liability +

Long_term_Liability_to_Current_Assets + Cash_Turnover_Rate + Total_assets_to_GNP_price'

- Logit Model 5 with the updated formula (Operating_Expense_Rate removed):

  'Default ~ Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Inventory_to_Current_Liability + Long_term_Liability_to_Current_Assets + Cash_Turnover_Rate + Total_assets_to_GNP_price'

- Logit Model 6 with the updated formula (Total_assets_to_GNP_price removed):

  'Default ~ Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Inventory_to_Current_Liability + Long_term_Liability_to_Current_Assets + Cash_Turnover_Rate'

- Logit Model 7 with the updated formula (Long_term_Liability_to_Current_Assets removed):

  'Default ~ Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Inventory_Turnover_Rate_times + Fixed_Assets_Turnover_Frequency + Inventory_to_Current_Liability + Cash_Turnover_Rate'

- Logit Model 8 with the updated formula (Inventory_Turnover_Rate_times removed):

  'Default ~ Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Fixed_Assets_Turnover_Frequency + Inventory_to_Current_Liability + Cash_Turnover_Rate'

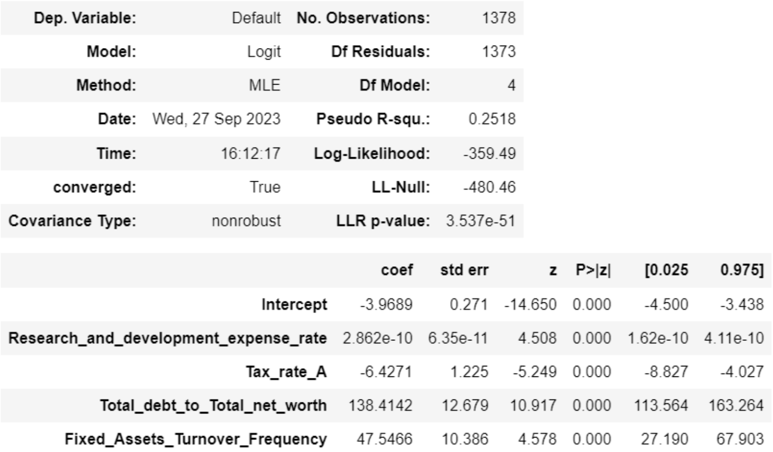
- Logit Model 9 with the updated formula (Cash_Turnover_Rate removed):

  'Default ~ Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Fixed_Assets_Turnover_Frequency + Inventory_to_Current_Liability'
- Logit Model 10 with the updated formula (Cash_Turnover_Rate removed):

  'Default ~ Research_and_development_expense_rate + Tax_rate_A + Total_debt_to_Total_net_worth + Fixed_Assets_Turnover_Frequency'

ix. Final Logit model after 9 iterations (Model 10):

Figure 17 - Final Logit model



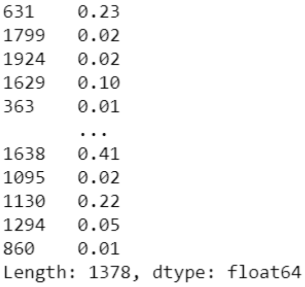
Dep. Variable:	Default	No. Observations:	1378
Model:	Logit	Df Residuals:	1373
Method:	MLE	Df Model:	4
Date:	Wed, 27 Sep 2023	Pseudo R-squ.:	0.2518
Time:	16:12:17	Log-Likelihood:	-359.49
converged:	True	LL-Null:	-480.46
Covariance Type:	nonrobust	LLR p-value:	3.537e-51

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.9689	0.271	-14.650	0.000	-4.500	-3.438
Research_and_development_expense_rate	2.862e-10	6.35e-11	4.508	0.000	1.62e-10	4.11e-10
Tax_rate_A	-6.4271	1.225	-5.249	0.000	-8.827	-4.027
Total_debt_to_Total_net_worth	138.4142	12.679	10.917	0.000	113.564	163.264
Fixed_Assets_Turnover_Frequency	47.5466	10.386	4.578	0.000	27.190	67.903

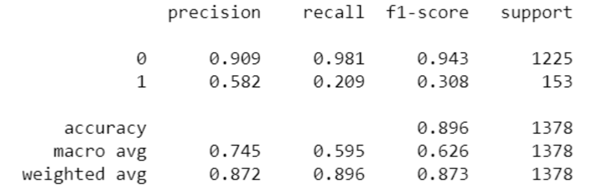
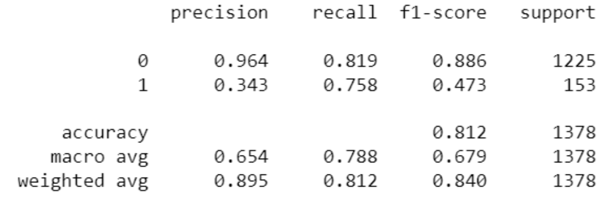
- Only four variables were left after eliminating variables with p-value of greater than 0.05 with each iteration. The final variables used in the model are:

    - Research_and_development_expense_rate
    - Tax_rate_A
    - Total_debt_to_Total_net_worth
    - Fixed_Assets_Turnover_Frequency'

- Predicted probabilities for the train set:

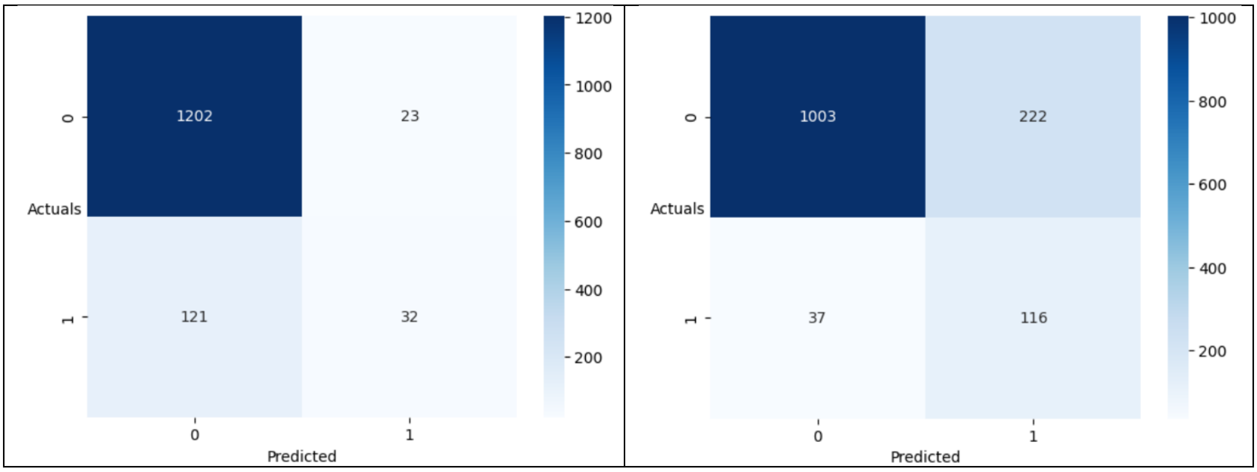


```
631     0.23
1799    0.02
1924    0.02
1629    0.10
363     0.01
        ...
1638    0.41
1095    0.02
1130    0.22
1294    0.05
860     0.01
Length: 1378, dtype: float64
```

Model summary with the threshold of 0.5	Model summary with the threshold of 0.13
	
Confusion matrix for 0.5 threshold	Confusion matrix for 0.13 threshold

```
              precision    recall  f1-score   support

           0      0.909     0.981     0.943      1225
           1      0.582     0.209     0.308       153

    accuracy                          0.896      1378
   macro avg      0.745     0.595     0.626      1378
weighted avg      0.872     0.896     0.873      1378
```

```
              precision    recall  f1-score   support

           0      0.964     0.819     0.886      1225
           1      0.343     0.758     0.473       153

    accuracy                          0.812      1378
   macro avg      0.654     0.788     0.679      1378
weighted avg      0.895     0.812     0.840      1378
```



- The recall for class 1 has improved drastically from 21% to 76%.

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

i. Results and performance metrics for the Test dataset

Figure 18 - Logit Confusion Matrix

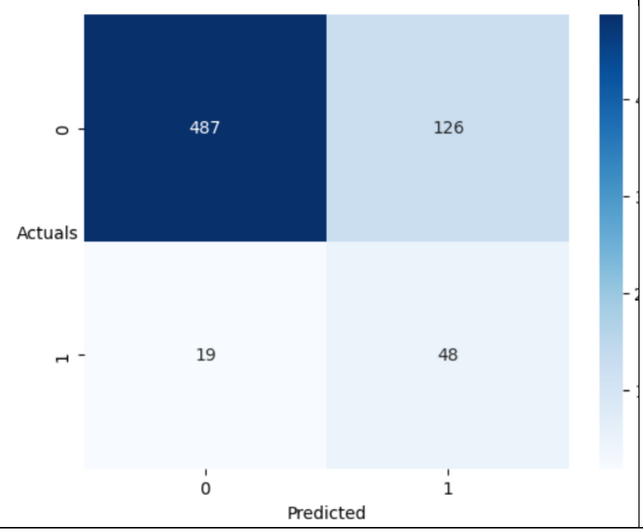
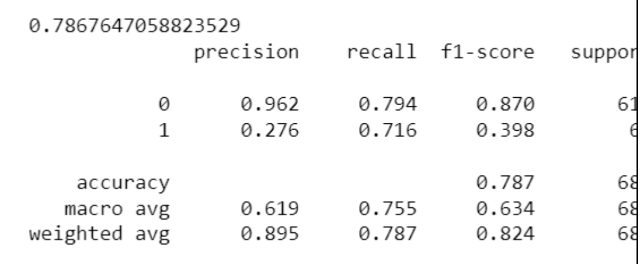


Figure 19 - Logit Model Summary



```
0.7867647058823529
              precision    recall  f1-score   suppor

           0      0.962     0.794     0.870      61
           1      0.276     0.716     0.398       6

    accuracy                          0.787      68
   macro avg      0.619     0.755     0.634      68
weighted avg      0.895     0.787     0.824      68
```

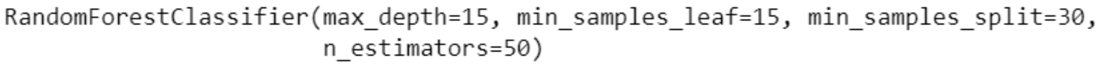
- Recall for test data for class 1 is 72%.
- The model was able to make 48 correct predictions for class 1.
- The overall accuracy for the model is 79%.

7. Build a Random Forest Model on Train Dataset. Also showcase your model building approach

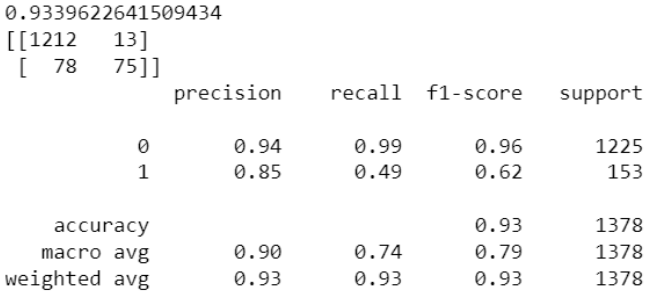
i. After running a Grid Search, the best parameters for the Random Forest Model were:

'max_depth': 15, 'min_samples_leaf': 15, 'min_samples_split': 30, 'n_estimators': 50

ii. The model was built using the above parameters.



```
RandomForestClassifier(max_depth=15, min_samples_leaf=15, min_samples_split=30,
                       n_estimators=50)
```



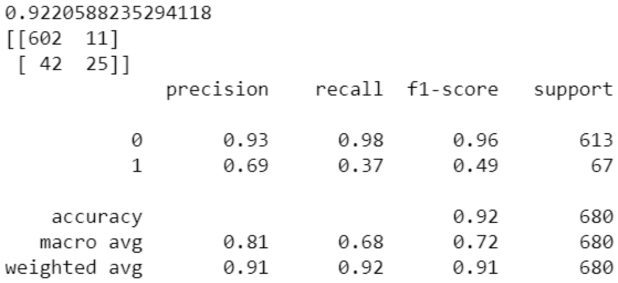
```
0.9339622641509434
[[1212   13]
 [  78   75]]
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      1225
           1       0.85      0.49      0.62       153

    accuracy                           0.93      1378
   macro avg       0.90      0.74      0.79      1378
weighted avg       0.93      0.93      0.93      1378
```

- The model made 75 correct predictions for class 1 on the training data.
- Its recall is precision is 85% and recall is 49% for class 1.
- The overall model accuracy is 93% but most of it comes from class 0 predictions, since the data is highly imbalanced and the values of 0 are much higher in number than class 1.

8. Validate the Random Forest Model on test Dataset and state the performance metrics. Also state interpretation from the model

i. Random Forest Model test data summary and confusion matrix:

Figure 20 - Random Forest Model

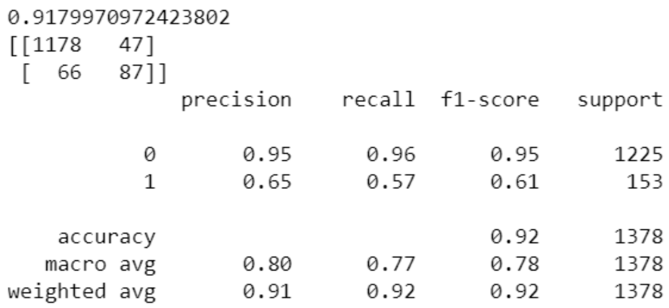


```
0.9220588235294118
[[602  11]
 [ 42  25]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.96       613
           1       0.69      0.37      0.49        67

    accuracy                           0.92       680
   macro avg       0.81      0.68      0.72       680
weighted avg       0.91      0.92      0.91       680
```

- The model delivered 25 correct predictions for the class 1.
- It has a 69% precision and 37% recall for class 1 and an overall accuracy of 92% which majorly comes from the class 0, since the data is imbalanced.

9. Build a LDA Model on Train Dataset. Also showcase your model building approach

i. LDA model on train dataset summary and confusion matrix:



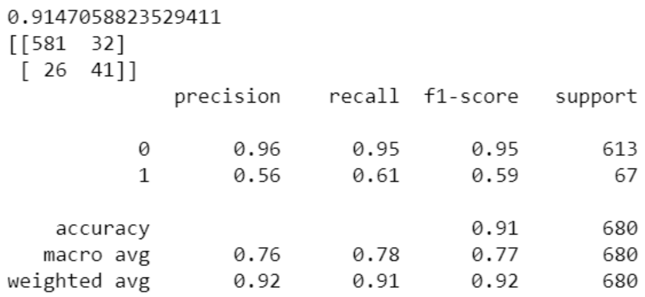
```
0.9179970972423802
[[1178   47]
 [  66   87]]
              precision    recall  f1-score   support

           0       0.95      0.96      0.95      1225
           1       0.65      0.57      0.61       153

    accuracy                           0.92      1378
   macro avg       0.80      0.77      0.78      1378
weighted avg       0.91      0.92      0.92      1378
```

- The model shows 65% precision and 57% recall for class 1 with an overall accuracy of 92%.
- It made 87 correct predictions for class 1.

10. Validate the LDA Model on test Dataset and state the performance metrics. Also state interpretation from the model

i. LDA Model test data summary and confusion matrix:

Figure 21 - LDA Model Summary



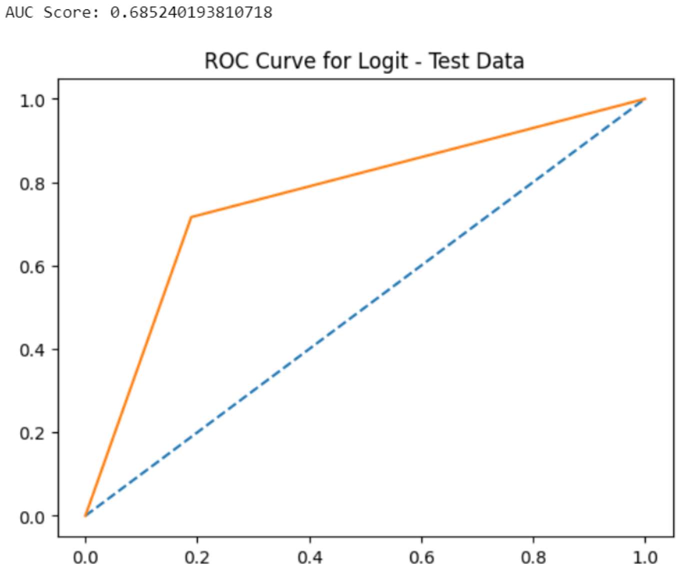
```
0.9147058823529411
[[581  32]
 [ 26  41]]
              precision    recall  f1-score   support

           0       0.96      0.95      0.95       613
           1       0.56      0.61      0.59        67

    accuracy                           0.91       680
   macro avg       0.76      0.78      0.77       680
weighted avg       0.92      0.91      0.92       680
```

- The model is showing 61% recall and 56% precision for class 1.
- It has an overall model accuracy of 91% due to the data being highly imbalanced with 0 class as the mode.
- It made 41 correct predictions for class 1.

11. Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve)

i. ROC Curve for the Logit model:

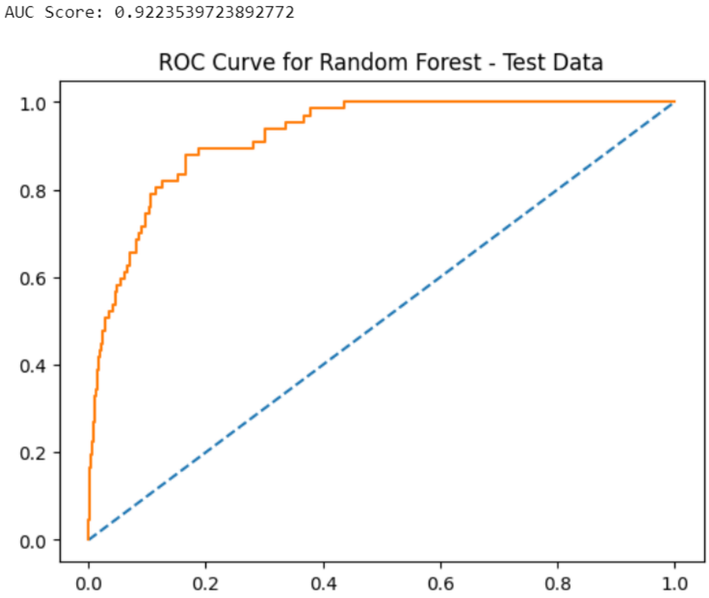
Figure 22 - Logit ROC Curve



```
AUC Score: 0.685240193810718
```

ii. ROC Curve for the Random Forest Model:

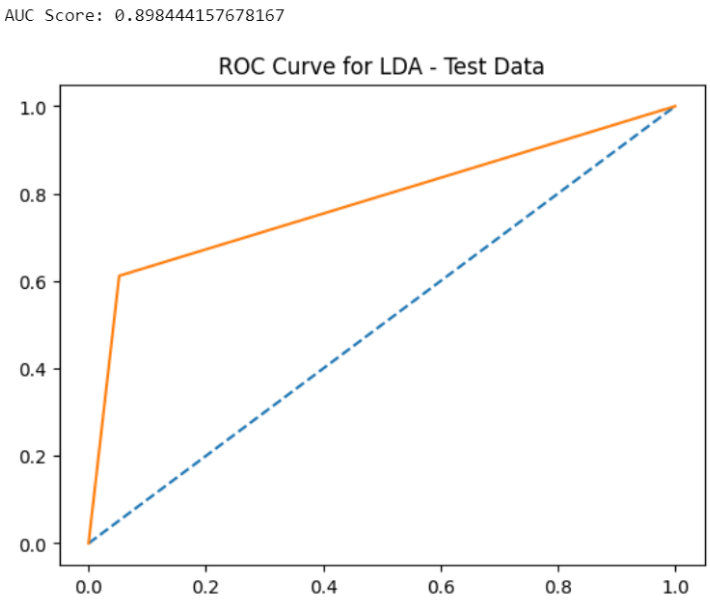
Figure 23 - Random Forest ROC Curve



```
AUC Score: 0.9223539723892772
```

iii. ROC Curve for the LDA model:

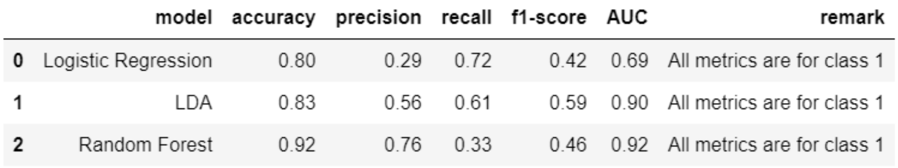
Figure 24 - LDA ROC Curve



```
AUC Score: 0.898444157678167
```

iv. Following is the comparison of all the models in a tabular form:

Figure 25 - Model comparison with performance metrics



	model	accuracy	precision	recall	f1-score	AUC	remark
0	Logistic Regression	0.80	0.29	0.72	0.42	0.69	All metrics are for class 1
1	LDA	0.83	0.56	0.61	0.59	0.90	All metrics are for class 1
2	Random Forest	0.92	0.76	0.33	0.46	0.92	All metrics are for class 1

12. PART A: Conclusions and Recommendations

i. Random Forest has performed the best on the test data among the three models.

ii. It has an accuracy score of 92% with a precision of 76% against Logit that has given a recall of 72%.

iii. LDA has given the highest f1-score but its recall or precision is not high enough for class 1 on the test data.

iv. Random Forest has scored the highest AUC Score of 0.92 against Logit that has only scored 0.69 and LDA which has scored 0.90.

v. However, when it comes to using one model for further improvement and optimisation, Logit would do the best job since it is giving the highest recall.

vi. In the case of credit risk where it is more important to catch false negatives of defaulters, recall should be prioritized.

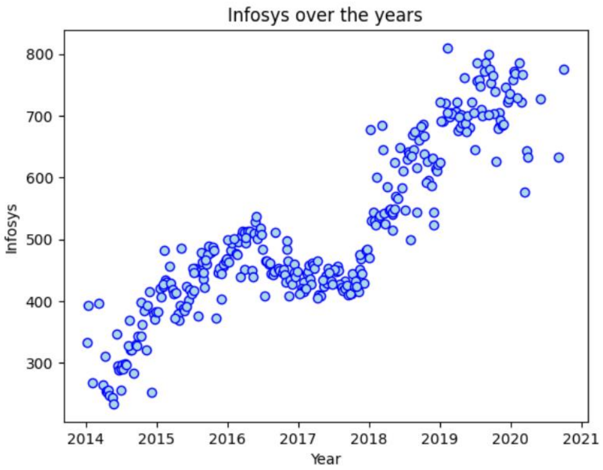
PART-B

The dataset contains 6 years of information (weekly stock information) on the stock prices of 10 different Indian Stocks. Calculate the mean and standard deviation on the stock returns and share insights. You are expected to do the Market Risk Analysis using Python.

1. Draw Stock Price Graph (Stock Price vs Time) for any 2 given stocks with inference

i. Stock price graph for Infosys:

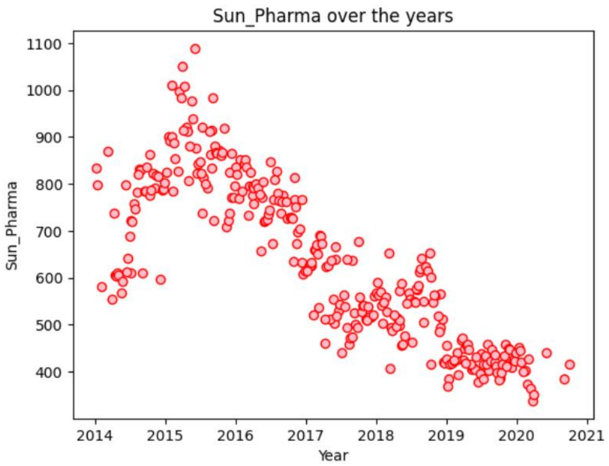
Figure 26 - Stock price graph for Infosys



- Infosys shows a positive trend from 2014 to 2021 as the stock price keeps climbing up as the timeline proceeds.

ii. Stock price graph for Sun Pharma:

Figure 27 - Stock price graph for Sun Pharma

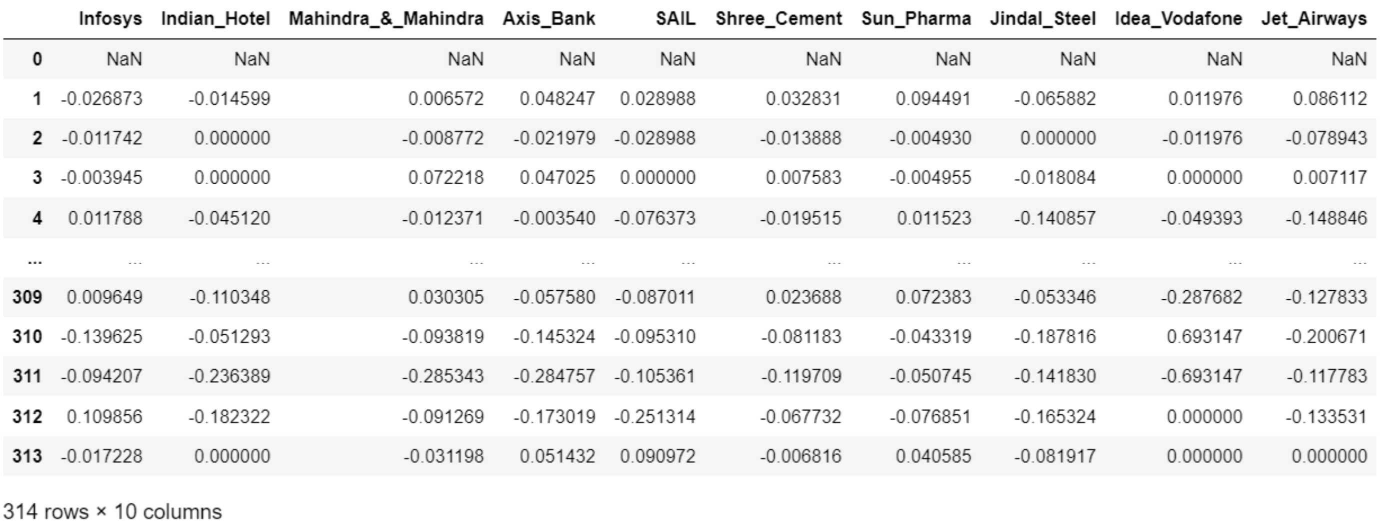


- The overall stock price trend for Sun Pharma is downward; the stock has not performed well for the 2014 to 2021 timeline.
- From 2014 to 2015, an uptrend can be soon, which soon plummeted.

2. Calculate Returns for all stocks with inference

i. Following is the sample data of all the stock returns calculated for all the companies:

Figure 28 - Stock returns for all stocks



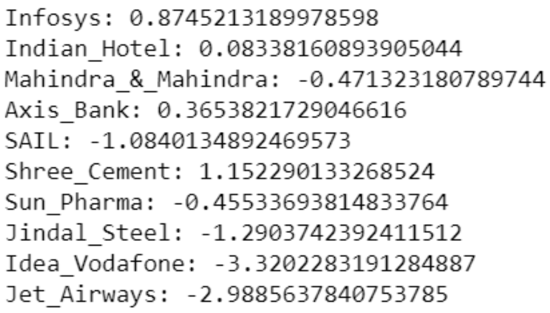
	Infosys	Indian_Hotel	Mahindra_&_Mahindra	Axis_Bank	SAIL	Shree_Cement	Sun_Pharma	Jindal_Steel	Idea_Vodafone	Jet_Airways
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	-0.026873	-0.014599	0.006572	0.048247	0.028988	0.032831	0.094491	-0.065882	0.011976	0.086112
2	-0.011742	0.000000	-0.008772	-0.021979	-0.028988	-0.013888	-0.004930	0.000000	-0.011976	-0.078943
3	-0.003945	0.000000	0.072218	0.047025	0.000000	0.007583	-0.004955	-0.018084	0.000000	0.007117
4	0.011788	-0.045120	-0.012371	-0.003540	-0.076373	-0.019515	0.011523	-0.140857	-0.049393	-0.148846
...	...	...	...	...	...	...	...	...	...	...
309	0.009649	-0.110348	0.030305	-0.057580	-0.087011	0.023688	0.072383	-0.053346	-0.287682	-0.127833
310	-0.139625	-0.051293	-0.093819	-0.145324	-0.095310	-0.081183	-0.043319	-0.187816	0.693147	-0.200671
311	-0.094207	-0.236389	-0.285343	-0.284757	-0.105361	-0.119709	-0.050745	-0.141830	-0.693147	-0.117783
312	0.109856	-0.182322	-0.091269	-0.173019	-0.251314	-0.067732	-0.076851	-0.165324	0.000000	-0.133531
313	-0.017228	0.000000	-0.031198	0.051432	0.090972	-0.006816	0.040585	-0.081917	0.000000	0.000000

314 rows × 10 columns

ii. These stock returns are weekly.

iii. Shree Cement has the highest stock returns over the years (=1.15).

iv. Following is the chart of all the stock returns in the descending order:

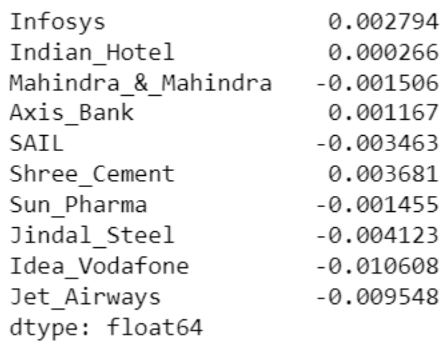


```
Infosys: 0.8745213189978598
Indian_Hotel: 0.08338160893905044
Mahindra_&_Mahindra: -0.471323180789744
Axis_Bank: 0.3653821729046616
SAIL: -1.0840134892469573
Shree_Cement: 1.152290133268524
Sun_Pharma: -0.45533693814833764
Jindal_Steel: -1.2903742392411512
Idea_Vodafone: -3.3202283191284887
Jet_Airways: -2.9885637840753785
```

v. Idea_Vodafone has the lowest stock returns (in negative) from 2014 to 2021 (=-3.32).

3. Calculate Stock Means and Standard Deviation for all stocks with inference

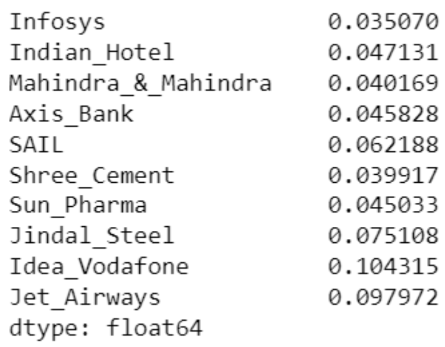
i. Stock Means:



```
Infosys                 0.002794
Indian_Hotel            0.000266
Mahindra_&_Mahindra    -0.001506
Axis_Bank               0.001167
SAIL                   -0.003463
Shree_Cement            0.003681
Sun_Pharma             -0.001455
Jindal_Steel           -0.004123
Idea_Vodafone          -0.010608
Jet_Airways            -0.009548
dtype: float64
```

- Shree Cement has the highest average stock returns over the years (=0.0035).
- Idea Vodafone has a negative and the lowest of all mean stock returns (-0.01).
- Long-term investments made in Shree Cement would have gained a good profit as opposed to other stocks.
- The stocks that made losses are Mahindra & Mahindra, SAIL, Sun_Pharma, Jindal_Steel, Idea_Vodafone, Jet_Airways.

ii. Standard Deviation for all the stocks:



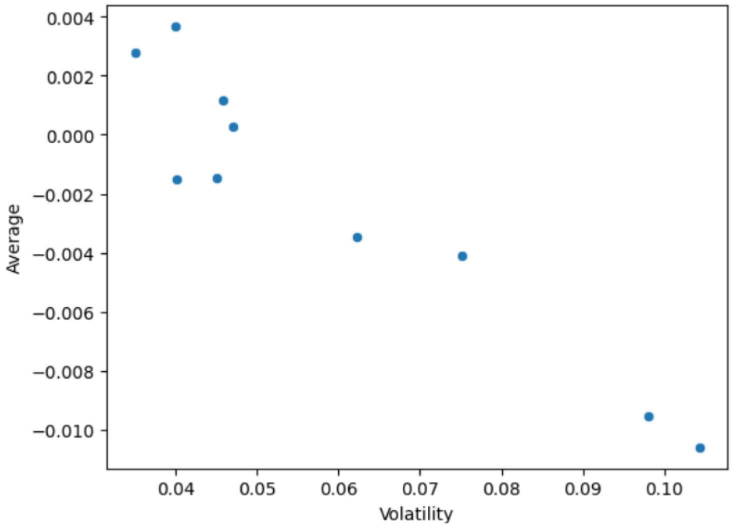
```
Infosys                 0.035070
Indian_Hotel            0.047131
Mahindra_&_Mahindra     0.040169
Axis_Bank               0.045828
SAIL                    0.062188
Shree_Cement            0.039917
Sun_Pharma              0.045033
Jindal_Steel            0.075108
Idea_Vodafone           0.104315
Jet_Airways             0.097972
dtype: float64
```

- The standard deviation is the volatility and therefore risk posed by a stock.
- Idea Vodafone is the highest in volatility. This seems to be correlated with the losses made while calculating the average stock return.
- The least volatile stocks are Infosys and Shree Cement. They have also given the highest mean stock returns.
- A good investment is done using portfolio diversification where the risk and reward is optimized to suit the needs of an investor.

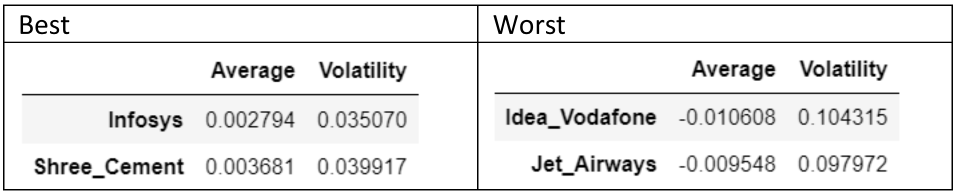
4. Draw a plot of Stock Means vs Standard Deviation and state your inference.

i. Below is the graph for Stock Means on the y-axis and Standard Deviation (Volatility) on the x-axis:

Figure 29 - Stock Means vs. Stock Standard Deviations graph



- The best stocks are those that are low on volatility but high on the average stock return.
- Infosys and Shree Cement are such stocks that have performed extremely well; they have the highest returns and the lowest volatility values.
- Following are the tables for the two best and two worst stocks:



Best			Worst		
	Average	**Volatility**		**Average**	**Volatility**
Infosys	0.002794	0.035070	**Idea_Vodafone**	-0.010608	0.104315
Shree_Cement	0.003681	0.039917	**Jet_Airways**	-0.009548	0.097972

5. Conclusions and Recommendations

i. Shree Cement is the best stock available in the dataset.

ii. The highest stock returns are shown by Shree Cement and Infosys. The lowest volatility is also shown by Infosys and Shree Cement.

iii. The worst stocks in the dataset are Idea Vodafone and Jet Airways. They are highly volatile and therefore risky, and have the lowest average returns through the years.

iv. If an investor wants to make a healthy investment, Infosys and Shree Cement would be the best options.

-----THE DOCUMENT ENDS HERE-----