Business Report

INDEX **Page No.**

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 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.

- Let's have a look at data
- Head of data looks like this. These are just few columns

	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A	_Cash_Flow_Per_Share
0	16974	Hind.Cables	8820000000.00	0.00	0.46	0.00	0.00	0.32
1	21214	Tata Tele. Mah.	9380000000.00	4230000000.00	0.46	0.00	0.00	0.32
2	14852	ABG Shipyard	3800000000.00	815000000.00	0.45	0.00	0.00	0.30
3	2439	GTL	6440000000.00	0.00	0.46	0.00	0.01	0.32
4	23505	Bharati Defence	3680000000.00	0.00	0.46	0.00	0.40	0.33

- The number of rows (observations) is 2058
- The number of columns (variables) is 58

- **Let’s check the data types**

#	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

- We can easily notice the missing values in data and most of the columns have data type as float.

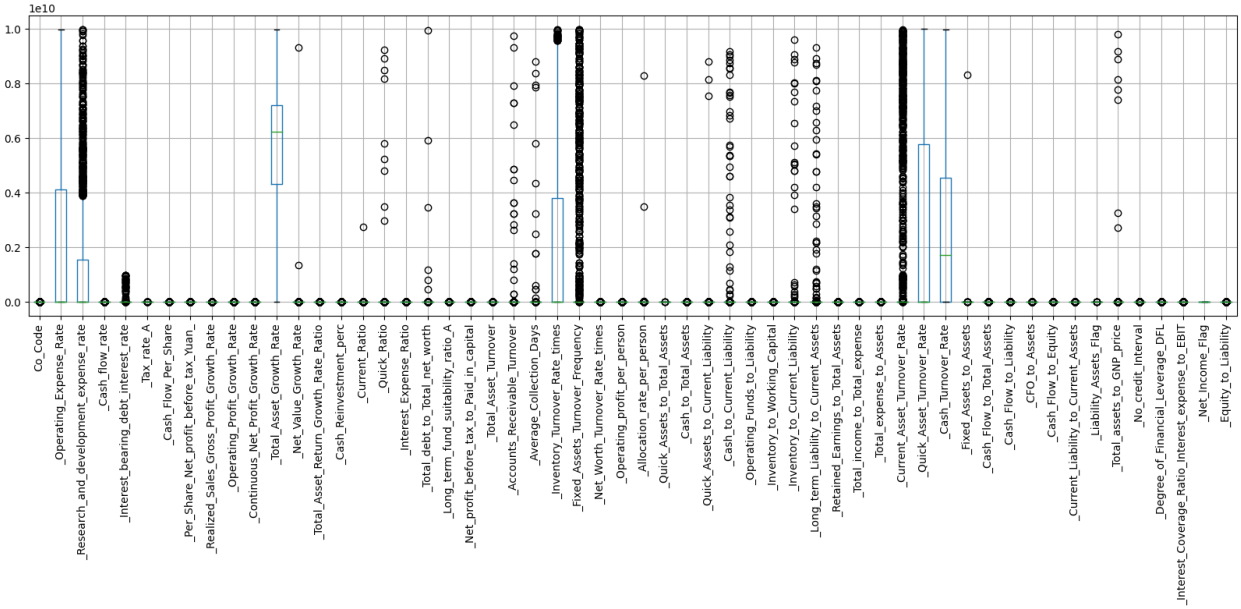
PART A: Outlier Treatment

- Let's check for the outliers

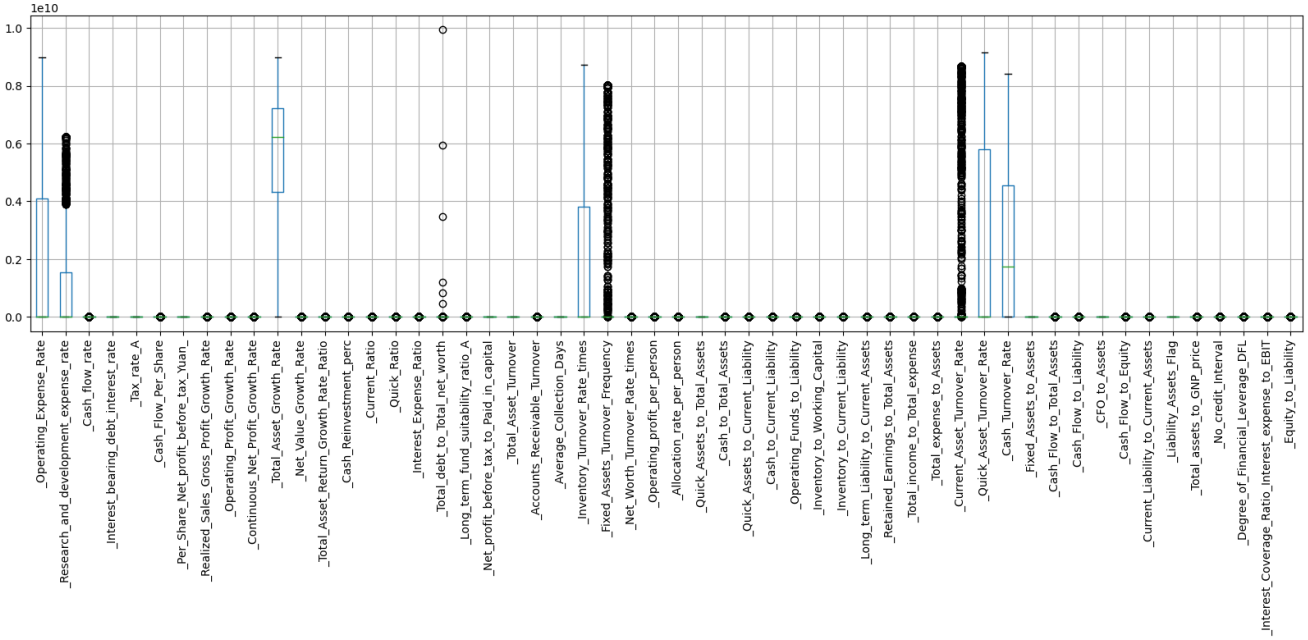
```
_Operating_Expense_Rate                                  0
_Research_and_development_expense_rate                 264
_Cash_flow_rate                                        139
_Interest_bearing_debt_interest_rate                     0
_Tax_rate_A                                              0
_Cash_Flow_Per_Share                                   146
_Per_Share_Net_profit_before_tax_Yuan_                   0
_Realized_Sales_Gross_Profit_Growth_Rate                207
_Operating_Profit_Growth_Rate                           317
_Continuous_Net_Profit_Growth_Rate                      340
_Total_Asset_Growth_Rate                                 0
_Net_Value_Growth_Rate                                 208
_Total_Asset_Return_Growth_Rate_Ratio                  123
_Cash_Reinvestment_perc                                141
_Current_Ratio                                         193
_Quick_Ratio                                           190
_Interest_Expense_Ratio                                328
_Total_debt_to_Total_net_worth                         105
_Long_term_fund_suitability_ratio_A                    234
_Net_profit_before_tax_to_Paid_in_capital                0
_Total_Asset_Turnover                                    0
_Accounts_Receivable_Turnover                          281
_Average_Collection_Days                                 0
_Inventory_Turnover_Rate_times                           0
_Fixed_Assets_Turnover_Frequency                       501
_Net_Worth_Turnover_Rate_times                         165
_Operating_profit_per_person                           357
_Allocation_rate_per_person                            200
_Quick_Assets_to_Total_Assets                            0
_Cash_to_Total_Assets                                  163
_Quick_Assets_to_Current_Liability                     185
_Cash_to_Current_Liability                             253
_Operating_Funds_to_Liability                          139
_Inventory_to_Working_Capital                          178
_Inventory_to_Current_Liability                        129
_Long_term_Liability_to_Current_Assets                 213
_Retained_Earnings_to_Total_Assets                     184
_Total_income_to_Total_expense                         116
_Total_expense_to_Assets                               168
_Current_Asset_Turnover_Rate                           464
_Quick_Asset_Turnover_Rate                               0
_Cash_Turnover_Rate                                      0
_Fixed_Assets_to_Assets                                  0
_Cash_Flow_to_Total_Assets                             317
_Cash_Flow_to_Liability                                407
_CFO_to_Assets                                           0
_Cash_Flow_to_Equity                                   306
_Current_Liability_to_Current_Assets                   121
_Liability_Assets_Flag                                   0
_Total_assets_to_GNP_price                             235
_No_credit_Interval                                    396
_Degree_of_Financial_Leverage_DFL                      438
_Interest_Coverage_Ratio_Interest_expense_to_EBIT      376
_Equity_to_Liability                                   190
```

- We have 9417 outliers in data and they make 8.47% of total data.
- We are capping them at 5% lower limit and 95% upper limit.

- **Before capping**



- **After capping**



PART A: Missing Value Treatment

- After outlier treatment let's have a look at missing values and their treatment.
- We have a total of 298 missing data points.

Column	Null Count
_Cash_Flow_Per_Share	167
_Total_debt_to_Total_net_worth	21
_Cash_to_Total_Assets	96
_Current_Liability_to_Current_Assets	14

- Only 4 columns have null values
- We are first scaling the data and then using KNN imputer with nearest neighbors = 10 and imputed the missing values.
- Missing values treated and outliers too.

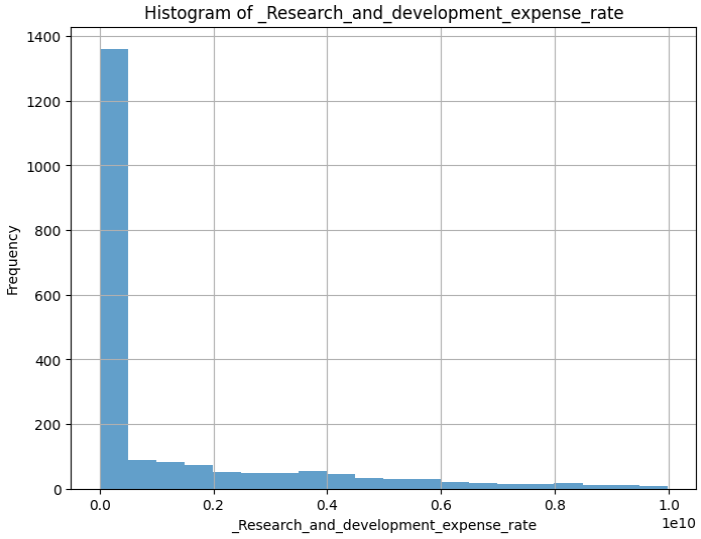
PART A: Univariate (4 marks) & Bivariate (6 marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)

We are just considering significant columns only-

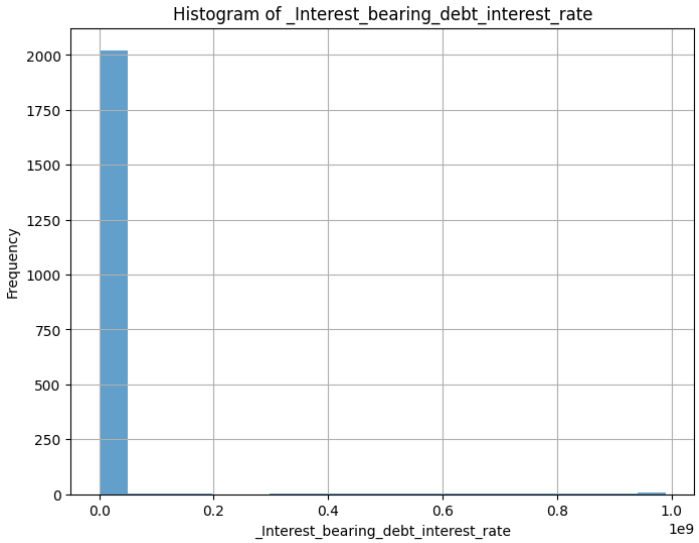
- '_Research_and_development_expense_rate'
- '_Interest_bearing_debt_interest_rate'
- '_Total_Asset_Growth_Rate'

- '_Net_Value_Growth_Rate'
- '_Interest_Expense_Ratio'
- '_Accounts_Receivable_Turnover'
- '_Fixed_Assets_Turnover_Frequency'
- '_Cash_to_Total_Assets'
- '_Cash_to_Current_Liability'
- '_Retained_Earnings_to_Total_Assets'
- '_Total_expense_to_Assets',
- '_Equity_to_Liability'
- Default

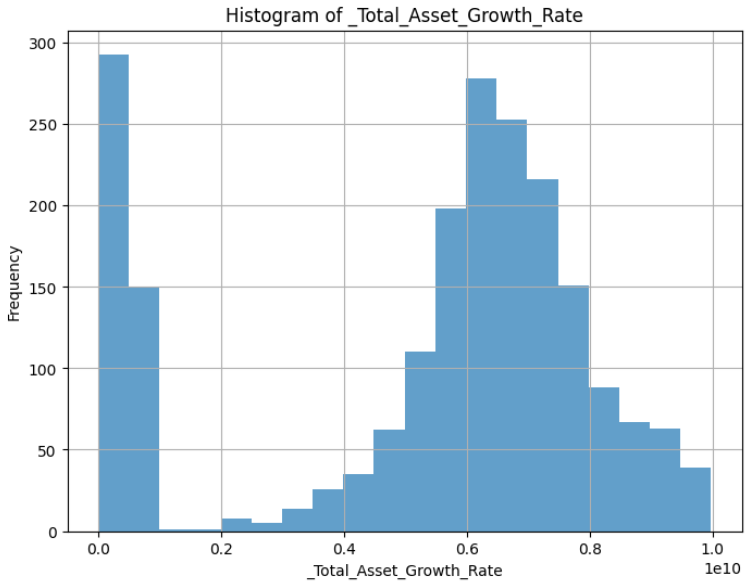
- **We will do EDA for these columns only-**



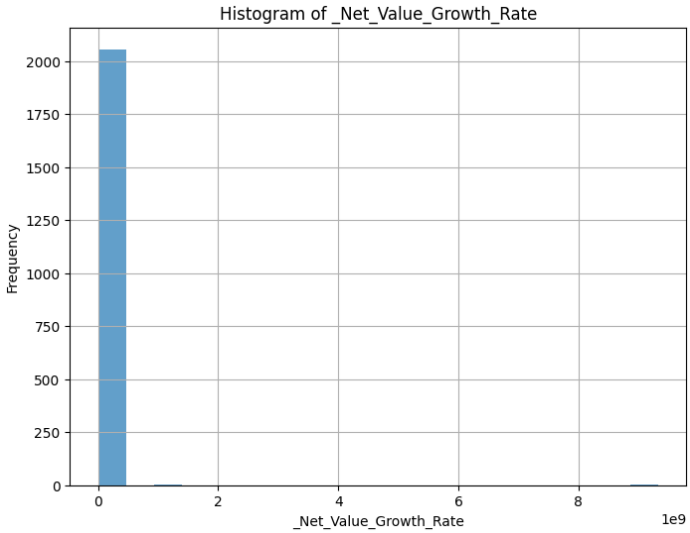
- Highly left skewed data. Few companies are spending a lot on R & D. But most companies are not spending that much.



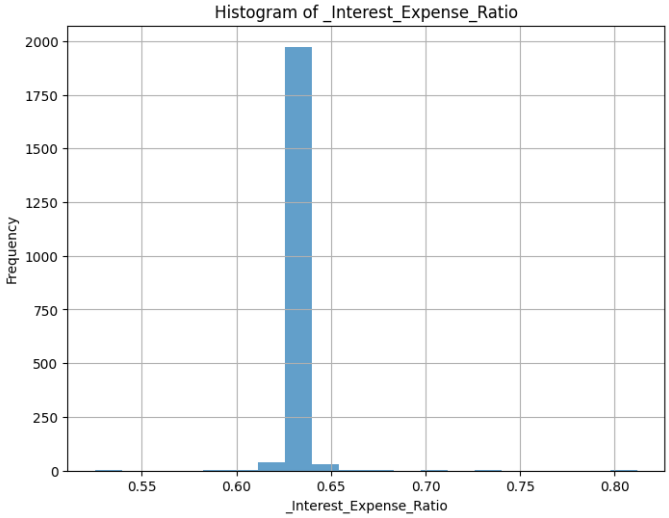
- Interest-bearing debt interest rate: Interest-bearing Debt/Equity is low, for most of the companies.



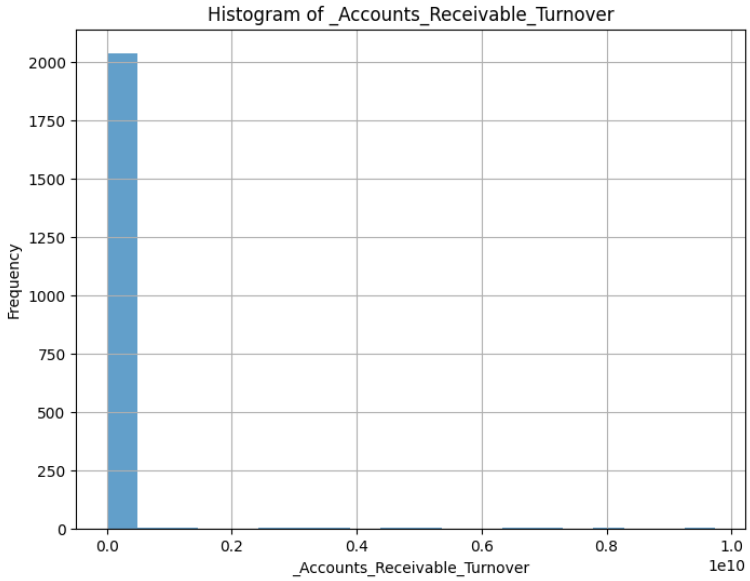
- Total Asset Growth Rate: Total Asset Growth. It is the rate at which how quickly the company has been growing its Assets. It can be seen that most companies are doing well to invest in growing their assets.



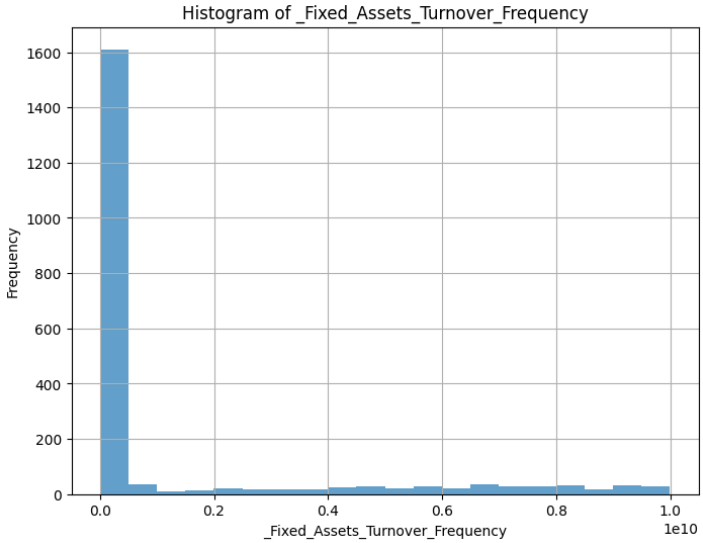
- Net value growth rate is low for most of the companies. Data is right skewed.



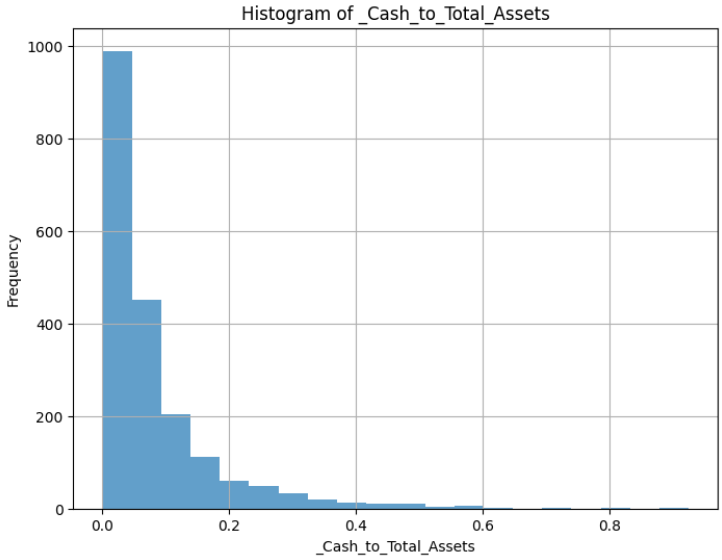
- Interest expense ratio for most of the companies is around 0.63 and 0.64. Most of companies are not that able to cover their interest expenses effectively.



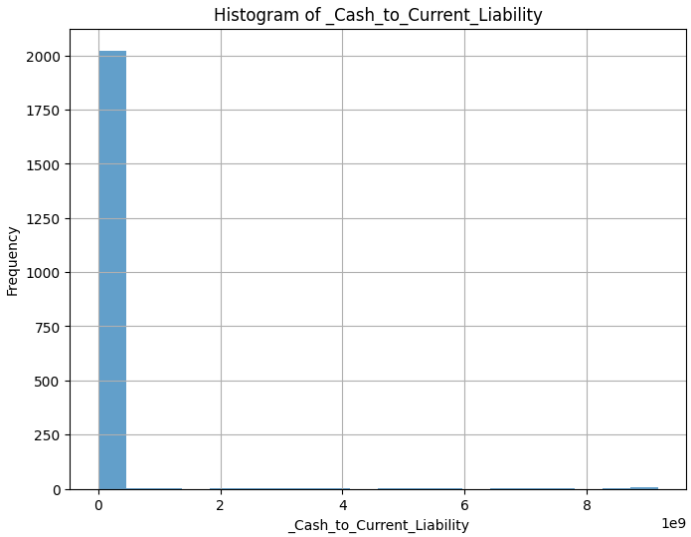
- Receivables turnover is used in business accounting to quantify how well companies are managing the credit that they extend to their customers by evaluating how long it takes to collect the outstanding debt throughout the accounting period. This is highly right skewed.



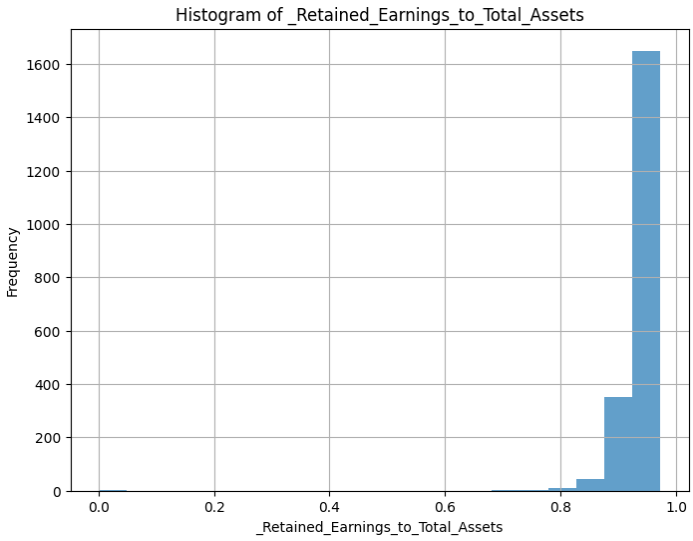
- Fixed Asset Turnover (FAT) is an efficiency ratio that indicates how well or efficiently a business uses fixed assets to generate sales. Lot of companies are doing pretty good to generate more sales with there fixed assets.



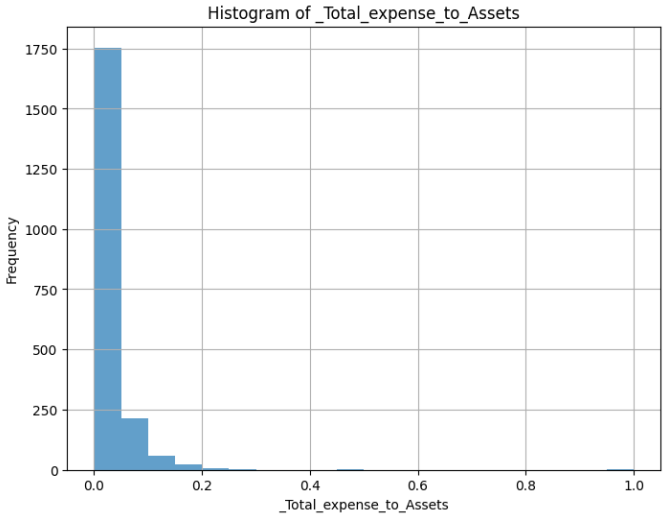
- Cash to total assets is highly right skewed. Cash for rotation is low for a lot of companies compared to there total asset values.



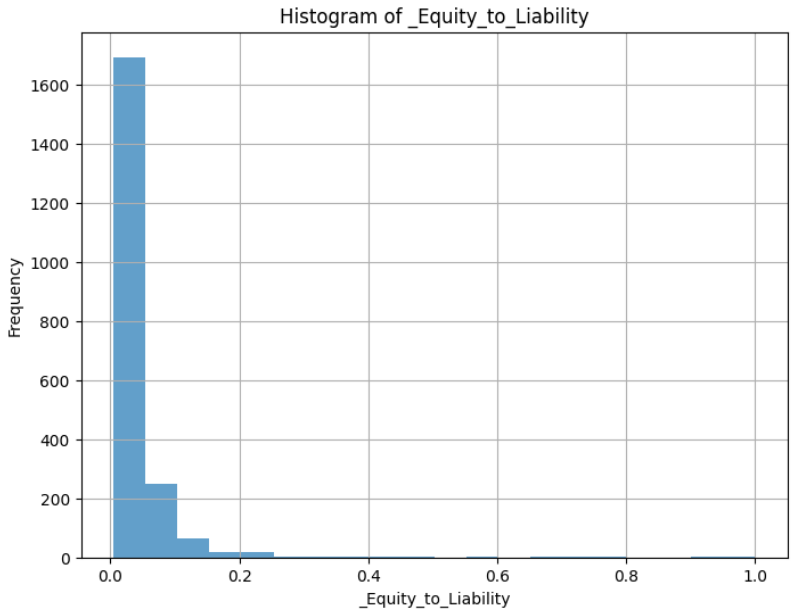
- Highly right skewed as cash to current liability is low and of the same range for most of the companies.



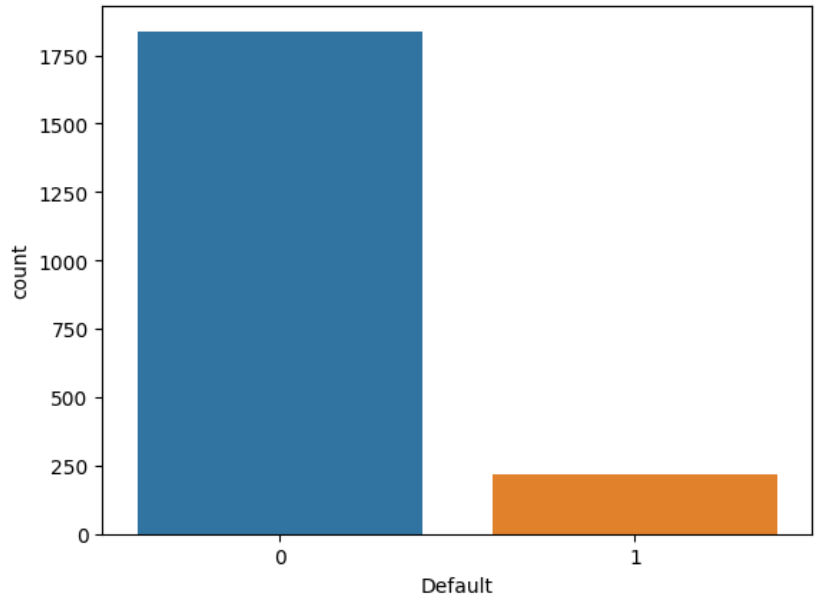
- Highly left skewed and most of values lie between 0.8 to 1.0 for retained earning to total assets.



- Total expense to assets ratio lies between 0 to 0.2 for most of the companies.

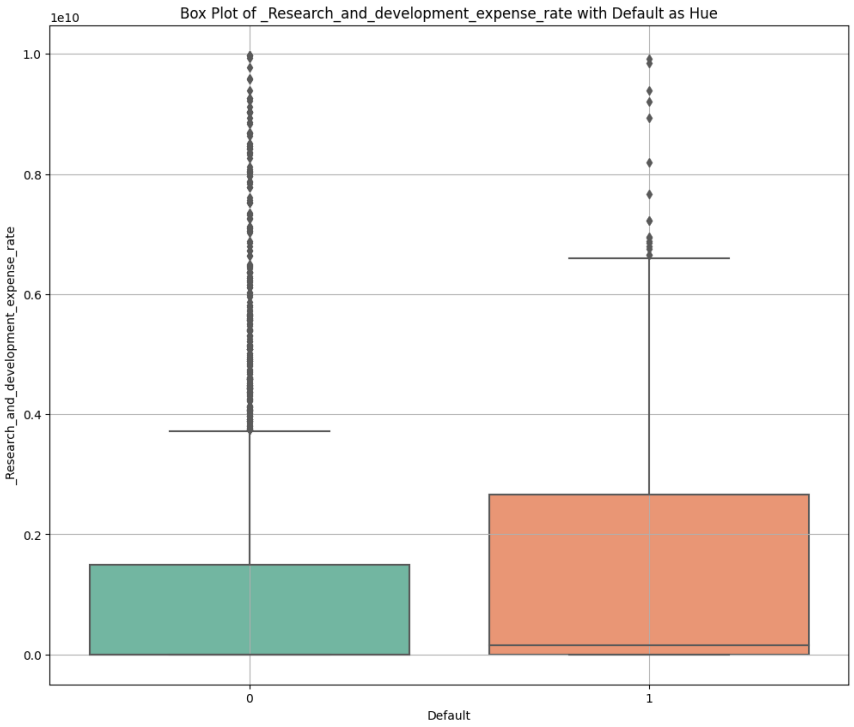


- Equity to liability ratio lies between 0 to 0.2 for most of the companies.

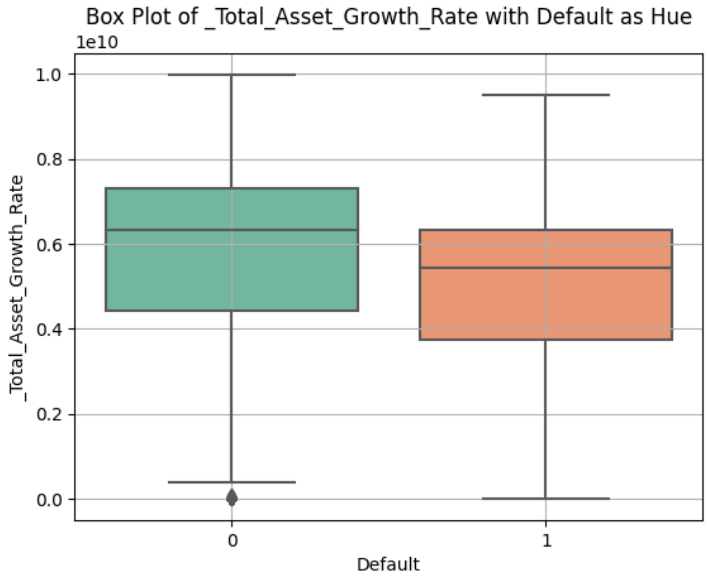


- Status 0 (Not Defaulted): Around 89.38% of companies have not defaulted.
- Status 1 (Defaulted): Approximately 10.62% of companies have defaulted.

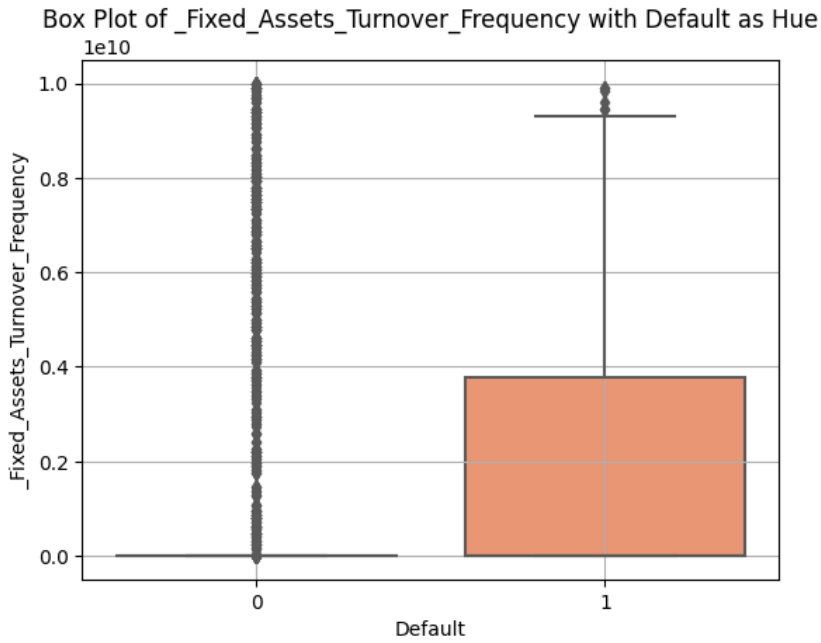
Bivariate analysis-



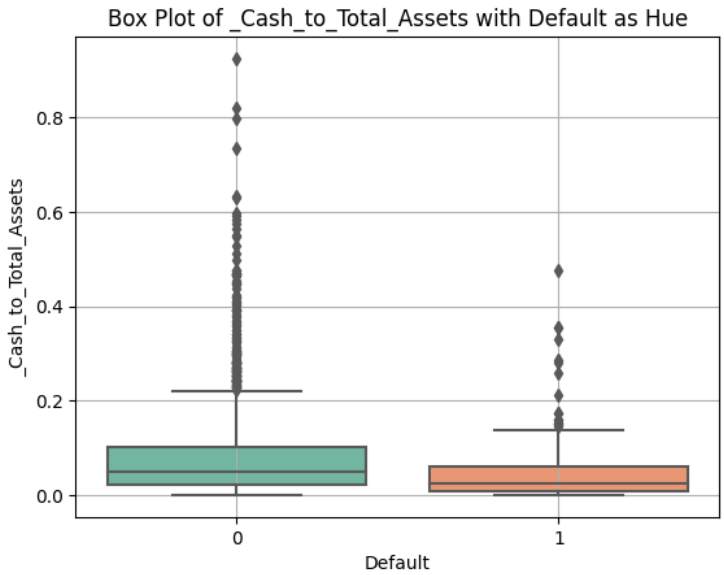
- Clear differences can be seen in distribution for R & D expense, considering separation on the default column. Upper whisker is quite high for defaulters compared to non-defaulters.



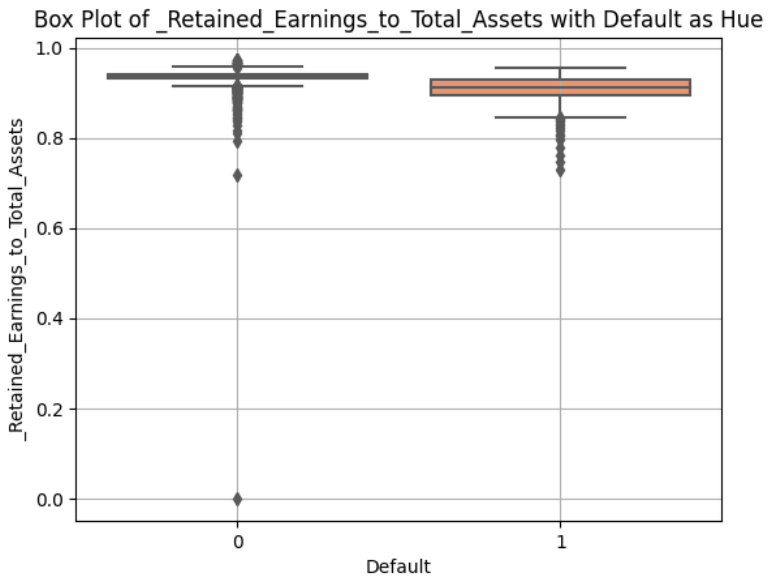
- Companies that have defaulted tend to have lower growth rates in their total assets compared to companies that have not defaulted.



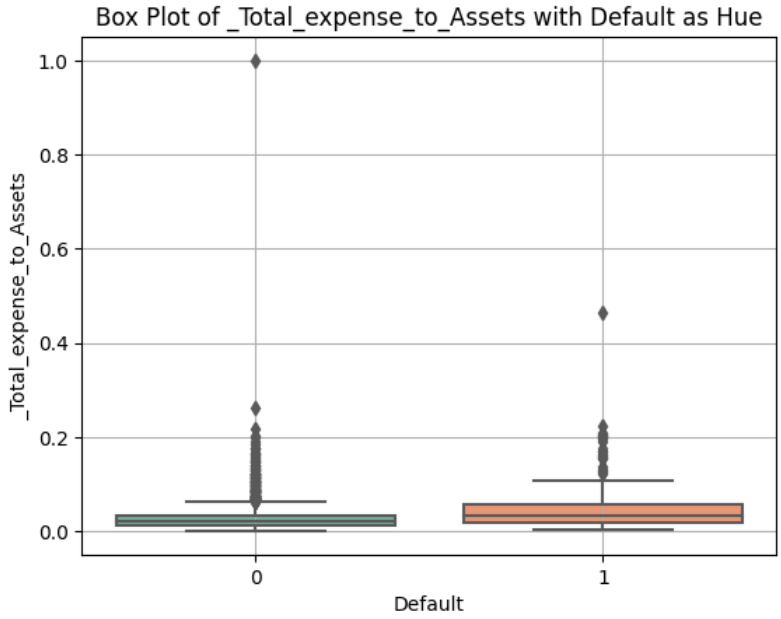
- Fixed assets turnover frequency is higher for defaulters compared to non defaulters.



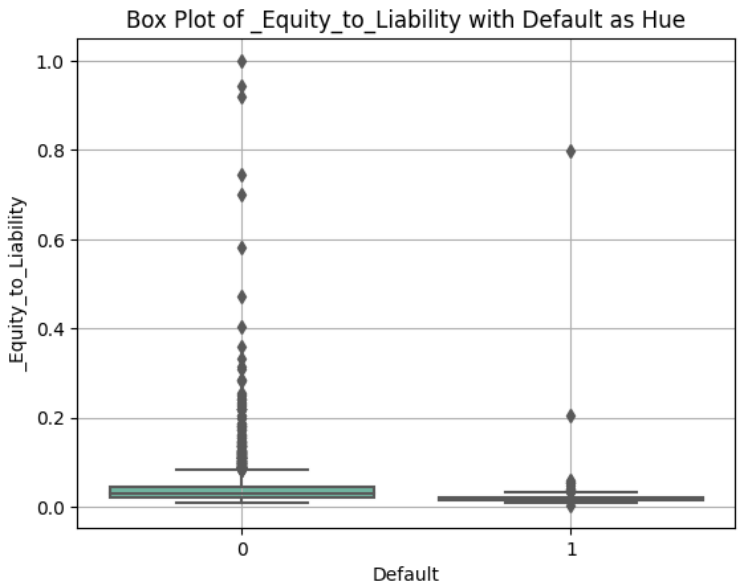
- Cash to total asset ratio is low for defaulters.



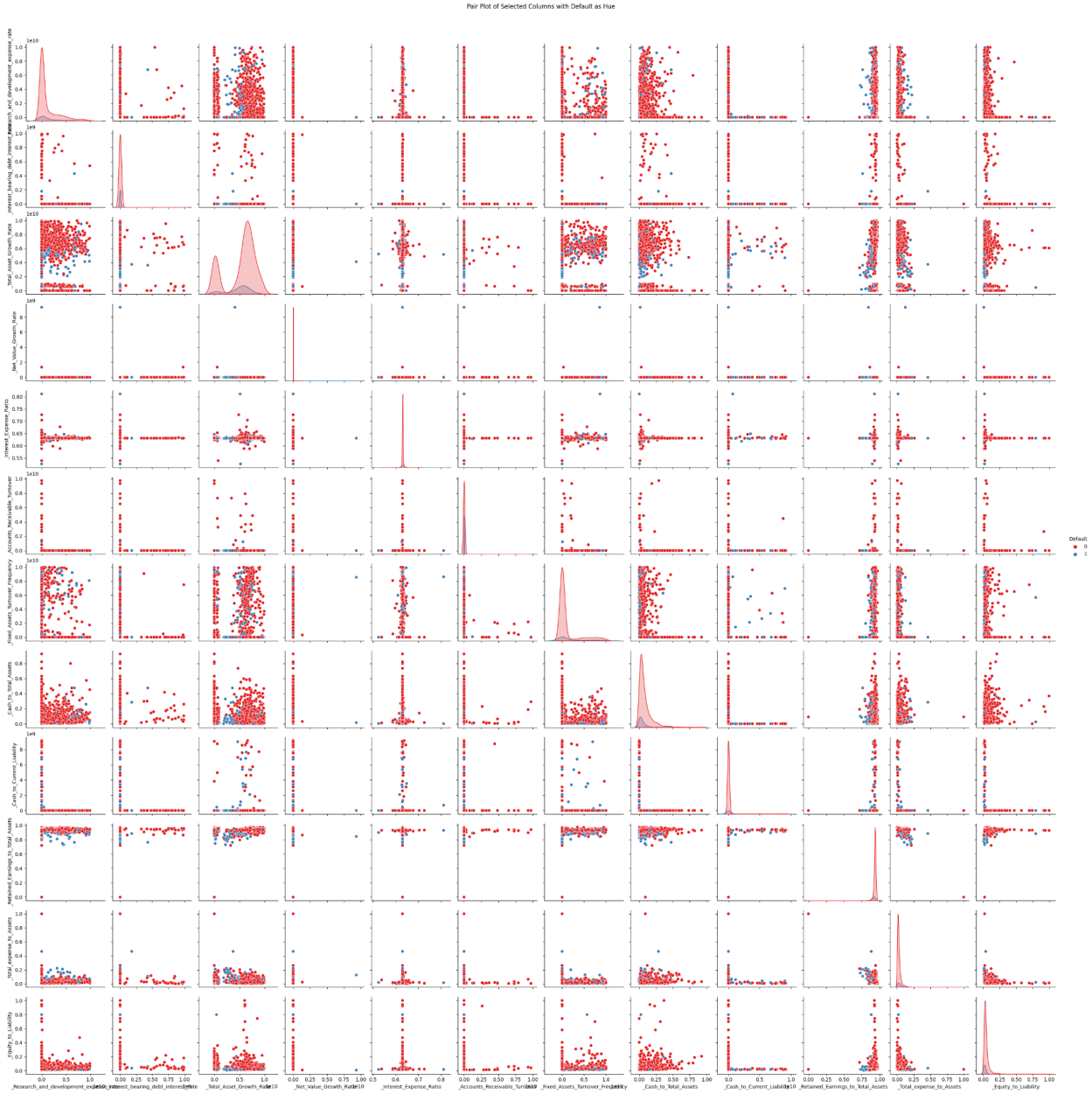
- Retained earning to total assets is low for defaulters as compared to non defaulters.



- Total expense to assets is high for defaulters compared to non defaulters.



- Equity to liability is low for defaulters compared to non defaulters.



- We can have a look at pair plot too to have a better understanding of differentiation in distribution for various columns considering companies that defaulted and that have not defaulted.

Summary:

- **Research and Development (R&D) Expense:**
    - Clear differences in distribution can be observed when analyzing R&D expenses while considering the separation based on the default column.

    - Defaulting companies tend to have higher variability and potentially higher R&D expenses, as indicated by the upper whisker extending further compared to non-defaulting companies.

- **Total Asset Growth Rate:**
    - Companies that have defaulted demonstrate a tendency to exhibit lower growth rates in their total assets compared to their non-defaulting counterparts.
    - This divergence might reflect financial challenges faced by defaulting companies, resulting in subdued expansion and possibly indicating financial distress.
- **Fixed Assets Turnover Frequency:**
    - The fixed assets turnover frequency appears to be higher for companies that have defaulted when compared to non-defaulting companies.
    - This could imply that defaulting companies are utilizing their fixed assets more frequently in generating revenue, possibly indicating operational efficiency or risk-taking behavior.
- **Cash to Total Asset Ratio:**
    - Defaulting companies exhibit a lower cash-to-total-asset ratio compared to non-defaulting companies.
    - This lower ratio may highlight liquidity challenges among defaulting companies, implying a potential inability to cover short-term obligations using available cash reserves.
- **Retained Earnings to Total Assets:**
    - Defaulting companies tend to have lower retained earnings in relation to their total assets compared to non-defaulting companies.
    - This disparity might signify that defaulting companies have struggled to accumulate earnings over time, affecting their financial stability and ability to cover obligations.
- **Total Expense to Assets:**

    - Defaulting companies display higher values in the total expense-to-assets ratio compared to non-defaulting companies.
    - This could suggest that defaulting companies allocate a larger portion of their assets towards expenses, indicating potential inefficiencies in cost management.

- **Equity to Liability:**
    - The equity-to-liability ratio is lower for defaulting companies in comparison to non-defaulting companies.
    - This lower ratio might imply that defaulting companies have a higher proportion of liabilities in relation to their equity, indicating potential leverage-related financial stress.

PART A: Train Test Split

- Prior moving forward we dropped the column '_Net_Income_Flag'as it is consistent, and has one unique value in all rows.
- We also dropped the columns 'Co_Code', 'Co_Name'as they are also like ID columns and have no use in analysis.

We do Train test split:

- X_train has a shape of (1378, 54), which means it contains 1378 samples and each sample has 54 features.
- X_test has a shape of (680, 54), which means it contains 680 samples and each sample also has 54 features
- Percentage of X_train: (1378 / (1378 + 680)) * 100 = 66.97%
- Percentage of X_test: (680 / (1378 + 680)) * 100 = 33.03%
- As we used stratified sampling let's check values for default:

For y_train:

- The value count for "Default" being 0 (indicating non-default) is 1231.

- The value count for "Default" being 1 (indicating default) is 147.

For y_test:

- The value count for "Default" being 0 (indicating non-default) is 607.
- The value count for "Default" being 1 (indicating default) is 73.

PART A: 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

- Before we proceed with this part, we have initiated a few steps. As we have 54 independent columns and one dependent column.
- We want to remove the multicollinearity in between these columns. For this we have used VIF on X_train and it has dropped below named columns. We have used the threshold value as 3 for VIF.

```
Removing variable _Per_Share_Net_profit_before_tax_Yuan_ with VIF 82.95608830545275
Removing variable _Cash_Flow_to_Total_Assets with VIF 49.591715108401274
Removing variable _CFO_to_Assets with VIF 33.443562818018606
Removing variable _Quick_Assets_to_Current_Liability with VIF 21.99515757558882
Removing variable _Operating_Funds_to_Liability with VIF 17.68478006734544
Removing variable _Total_Asset_Turnover with VIF 9.756740129944738
Removing variable _Current_Ratio with VIF 9.322516604773304
Removing variable _Net_profit_before_tax_to_Paid_in_capital with VIF 6.937731737801428
Removing variable _Quick_Assets_to_Total_Assets with VIF 5.33521317106477
Removing variable _Interest_Coverage_Ratio_Interest_expense_to_EBIT with VIF 4.765570535427797
Removing variable _Cash_Reinvestment_perc with VIF 4.248439615269727
Removing variable _Cash_Flow_to_Liability with VIF 4.102921235494216
Removing variable _Quick_Ratio with VIF 3.604266066073344
Removing variable _Total_income_to_Total_expense with VIF 3.467362908778745
Removing variable _Fixed_Assets_to_Assets with VIF 3.0231182819176747
```

We are left with 39 columns now, once we remove these columns.

After this we run the stats model on remaining columns and below are the results

	coef	std err	t	P>|t|	[0.025	0.975]
_Operating_Expense_Rate	0.0100	0.009	1.143	0.253	-0.007	0.027
_Research_and_development_expense_rate	0.0313	0.008	3.860	0.000	0.015	0.047
_Cash_flow_rate	0.0133	0.012	1.094	0.274	-0.011	0.037
_Interest_bearing_debt_interest_rate	0.0202	0.008	2.483	0.013	0.004	0.036
_Tax_rate_A	-0.0124	0.009	-1.312	0.190	-0.031	0.006
_Cash_Flow_Per_Share	-0.0054	0.011	-0.495	0.621	-0.027	0.016
_Realized_Sales_Gross_Profit_Growth_Rate	-0.0092	0.011	-0.819	0.413	-0.031	0.013
_Operating_Profit_Growth_Rate	0.0137	0.013	1.081	0.280	-0.011	0.039
_Continuous_Net_Profit_Growth_Rate	-0.0222	0.013	-1.773	0.076	-0.047	0.002
_Total_Asset_Growth_Rate	-0.0170	0.008	-2.018	0.044	-0.034	-0.000
_Net_Value_Growth_Rate	-0.0301	0.011	-2.729	0.006	-0.052	-0.008
_Total_Asset_Return_Growth_Rate_Ratio	0.0012	0.013	0.097	0.922	-0.023	0.026
_Interest_Expense_Ratio	-0.0248	0.012	-2.087	0.037	-0.048	-0.001
_Total_debt_to_Total_net_worth	0.0074	0.007	1.126	0.260	-0.005	0.020
_Long_term_fund_suitability_ratio_A	0.0057	0.009	0.605	0.545	-0.013	0.024
_Accounts_Receivable_Turnover	-0.0227	0.010	-2.313	0.021	-0.042	-0.003
_Average_Collection_Days	0.0127	0.010	1.284	0.199	-0.007	0.032
_Inventory_Turnover_Rate_times	0.0060	0.008	0.721	0.471	-0.010	0.022
_Fixed_Assets_Turnover_Frequency	0.0217	0.009	2.480	0.013	0.005	0.039
_Net_Worth_Turnover_Rate_times	0.0151	0.011	1.325	0.185	-0.007	0.037
_Operating_profit_per_person	0.0179	0.011	1.583	0.114	-0.004	0.040
_Allocation_rate_per_person	0.0202	0.011	1.767	0.077	-0.002	0.043
_Cash_to_Total_Assets	-0.0263	0.012	-2.145	0.032	-0.050	-0.002
_Cash_to_Current_Liability	0.0293	0.012	2.525	0.012	0.007	0.052
_Inventory_to_Working_Capital	-0.0158	0.009	-1.764	0.078	-0.033	0.002
_Inventory_to_Current_Liability	-0.0120	0.010	-1.222	0.222	-0.031	0.007
_Long_term_Liability_to_Current_Assets	-0.0080	0.009	-0.901	0.368	-0.026	0.009
_Retained_Earnings_to_Total_Assets	-0.0885	0.012	-7.310	0.000	-0.112	-0.065
_Total_expense_to_Assets	0.0294	0.011	2.793	0.005	0.009	0.050
_Current_Asset_Turnover_Rate	-0.0026	0.009	-0.294	0.769	-0.020	0.015
_Quick_Asset_Turnover_Rate	-0.0006	0.009	-0.065	0.949	-0.018	0.017
_Cash_Turnover_Rate	-0.0144	0.008	-1.755	0.080	-0.030	0.002
_Cash_Flow_to_Equity	-0.0090	0.009	-0.963	0.336	-0.027	0.009
_Current_Liability_to_Current_Assets	0.0200	0.011	1.867	0.062	-0.001	0.041
_Liability_Assets_Flag	-6.06e-18	3.3e-18	-1.837	0.066	-1.25e-17	4.11e-19
_Total_assets_to_GNP_price	0.0133	0.009	1.417	0.157	-0.005	0.032
No credit Interval	-0.0136	0.009	-1.448	0.148	-0.032	0.005
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
_Degree_of_Financial_Leverage_DFL	-0.0004	0.011	-0.032	0.974	-0.022	0.021
_Equity_to_Liability	-0.0477	0.012	-3.835	0.000	-0.072	-0.023

We are going to drop columns with p-values greater than 0.05 because they are considered statistically insignificant and don't contribute meaningfully to explaining the variation in the dependent variable.

After dropping other columns , significant columns are:

- '_Research_and_development_expense_rate'
- '_Interest_bearing_debt_interest_rate'
- '_Total_Asset_Growth_Rate'
- '_Net_Value_Growth_Rate'
- '_Interest_Expense_Ratio'
- '_Accounts_Receivable_Turnover'
- '_Fixed_Assets_Turnover_Frequency'
- '_Cash_to_Total_Assets'
- '_Cash_to_Current_Liability'
- '_Retained_Earnings_to_Total_Assets'
- '_Total_expense_to_Assets',
- '_Equity_to_Liability’

- We are left with these 12 columns only.
- We run stats model again and out_put is:

OLS Regression Results

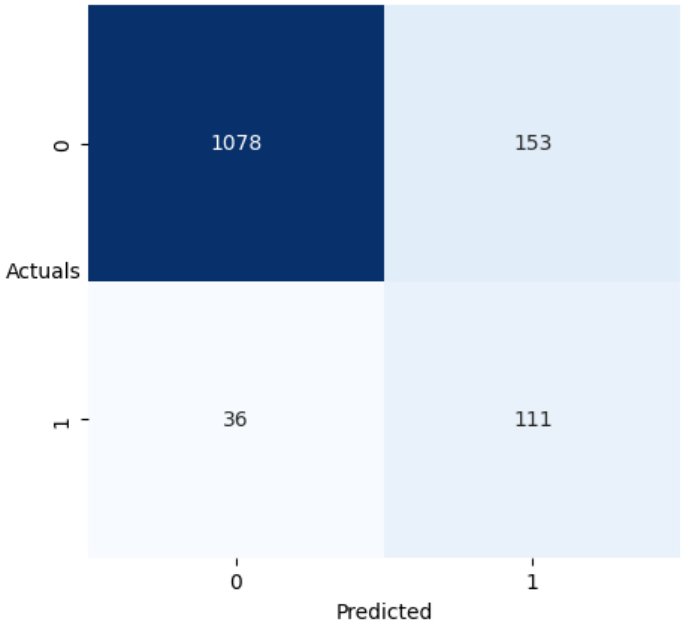
Dep. Variable:	Default	R-squared (uncentered):	0.240
Model:	OLS	Adj. R-squared (uncentered):	0.233
Method:	Least Squares	F-statistic:	35.92
Date:	Sun, 13 Aug 2023	Prob (F-statistic):	4.62e-73
Time:	05:25:11	Log-Likelihood:	-224.39
No. Observations:	1378	AIC:	472.8
Df Residuals:	1366	BIC:	535.5
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
_Research_and_development_expense_rate	0.0284	0.008	3.599	0.000	0.013	0.044
_Interest_bearing_debt_interest_rate	0.0175	0.008	2.210	0.027	0.002	0.033
_Total_Asset_Growth_Rate	-0.0184	0.008	-2.210	0.027	-0.035	-0.002
_Net_Value_Growth_Rate	-0.0393	0.009	-4.263	0.000	-0.057	-0.021
_Interest_Expense_Ratio	-0.0292	0.008	-3.499	0.000	-0.046	-0.013
_Accounts_Receivable_Turnover	-0.0132	0.008	-1.657	0.098	-0.029	0.002
_Fixed_Assets_Turnover_Frequency	0.0249	0.008	3.142	0.002	0.009	0.040
_Cash_to_Total_Assets	-0.0346	0.011	-3.208	0.001	-0.056	-0.013
_Cash_to_Current_Liability	0.0384	0.011	3.479	0.001	0.017	0.060
_Retained_Earnings_to_Total_Assets	-0.0917	0.010	-9.534	0.000	-0.111	-0.073
_Total_expense_to_Assets	0.0204	0.009	2.307	0.021	0.003	0.038
_Equity_to_Liability	-0.0587	0.010	-6.131	0.000	-0.077	-0.040

Omnibus:	420.266	Durbin-Watson:	1.714
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1107.818
Skew:	1.608	Prob(JB):	2.76e-241
Kurtosis:	5.992	Cond. No.	2.79

- Once this model has given us the predictions for the train dataset. Then we convert these predictions to 0 and 1 based on the optimum threshold.
- By identifying the threshold that maximizes the difference between TPR and FPR, we're essentially finding the point on the ROC curve where the trade-off between sensitivity and specificity is optimal.
- **Threshold we got is 0.137**
- Recall (also known as sensitivity or true positive rate) plays a crucial role, especially when dealing with imbalanced datasets or situations where one class is more important to correctly identify than the other. Here's how recall is important in this case:
- Recall measures the proportion of actual positive cases (default cases) that the model correctly identifies as positive. Mathematically, it's defined as:
- Recall = True Positives / (True Positives + False Negatives)

- True Positives (TP): The number of actual positive cases correctly predicted as positive by the model.
- False Negatives (FN): The number of actual positive cases incorrectly predicted as negative by the model.
- A high recall helps ensure that the model captures as many actual default cases as possible, reducing the risk of financial losses due to missed default predictions.
- To optimize recall, we would generally want to identify the threshold that provides the highest true positive rate (TPR) while accepting some increase in false positive rate (FPR).
- This threshold would ensure that the model captures as many true positive cases (actual defaults) as possible, even if it means tolerating a higher number of false positives (non-defaults misclassified as defaults).
- Let's check how this model has performed on train data-



- We have 36 actual defaulters classified as non defaulters.

	precision	recall	f1-score	support
0.0	0.968	0.876	0.919	1231
1.0	0.420	0.755	0.540	147
accuracy			0.863	1378
macro avg	0.694	0.815	0.730	1378
weighted avg	0.909	0.863	0.879	1378

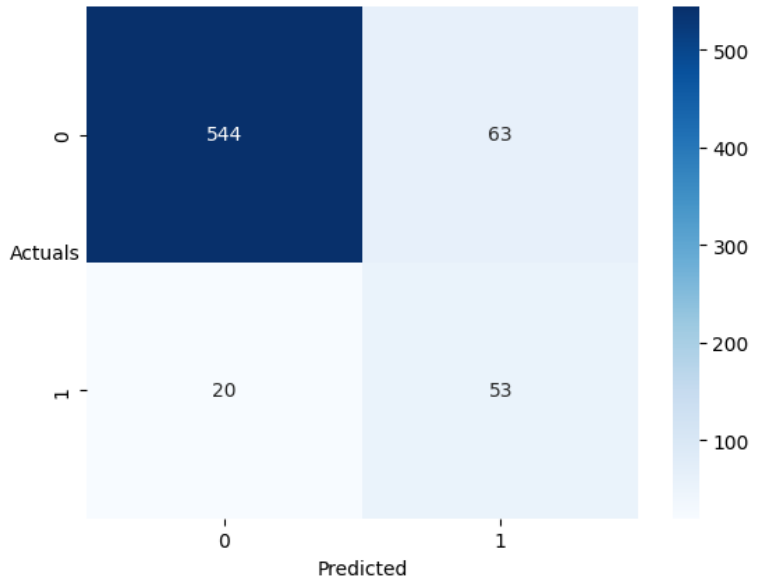
- Recall is 75.5% in predicting defaulters for train data who are actually defaulters.

PART A: Validate the Model on Test Dataset and state the performance metrics. Also state interpretation from the model

- Let's validate on test set:

	precision	recall	f1-score	support
0.0	0.965	0.896	0.929	607
1.0	0.457	0.726	0.561	73
accuracy			0.878	680
macro avg	0.711	0.811	0.745	680
weighted avg	0.910	0.878	0.890	680

- Recall is 72.6% in predicting defaulters for test data who are actually defaulters.



- Out of a total 73 actual defaulters, our model has rightly captured 53 defaulters in the test data set.
- Let's check the equation that is behind this model.

y = 0.03(_Research_and_development_expense_rate) +*

0.02(_Interest_bearing_debt_interest_rate) - 0.02*(_Total_Asset_Growth_Rate) -*

0.04(_Net_Value_Growth_Rate) - 0.03*(_Interest_Expense_Ratio) -*

0.01(_Accounts_Receivable_Turnover) + 0.02*(_Fixed_Assets_Turnover_Frequency) -*

0.03(_Cash_to_Total_Assets) + 0.04*(_Cash_to_Current_Liability) -*

0.09(_Retained_Earnings_to_Total_Assets) + 0.02*(_Total_expense_to_Assets) -*

0.06(_Equity_to_Liability)*

Explanation:

Variable	Impact on Predicted Probability of Default
_Research_and_development_expense_rate	Increase in R&D expenses leads to an increase in predicted probability of default
_Interest_bearing_debt_interest_rate	Increase in interest rates leads to an increase in predicted probability of default
_Total_Asset_Growth_Rate	Decrease in asset growth rate leads to an increase in predicted probability of default
_Net_Value_Growth_Rate	Decrease in net value growth rate leads to an increase in predicted probability of default
_Interest_Expense_Ratio	Decrease in interest expense ratio leads to an increase in predicted probability of default
_Accounts_Receivable_Turnover	Decrease in turnover leads to a slight increase in predicted probability of default
_Fixed_Assets_Turnover_Frequency	Increase in turnover frequency leads to an increase in predicted probability of default
_Cash_to_Total_Assets	Decrease in cash ratio leads to an increase in predicted probability of default

_Cash_to_Current_Liability	Increase in cash ratio leads to an increase in predicted probability of default
_Retained_Earnings_to_Total_Assets	Decrease in retained earnings ratio leads to an increase in predicted probability of default
_Total_expense_to_Assets	Increase in expense ratio leads to an increase in predicted probability of default
_Equity_to_Liability	Decrease in equity ratio leads to an increase in predicted probability of default

PART A: Build a Random Forest Model on a Train Dataset. Also showcase your model building approach

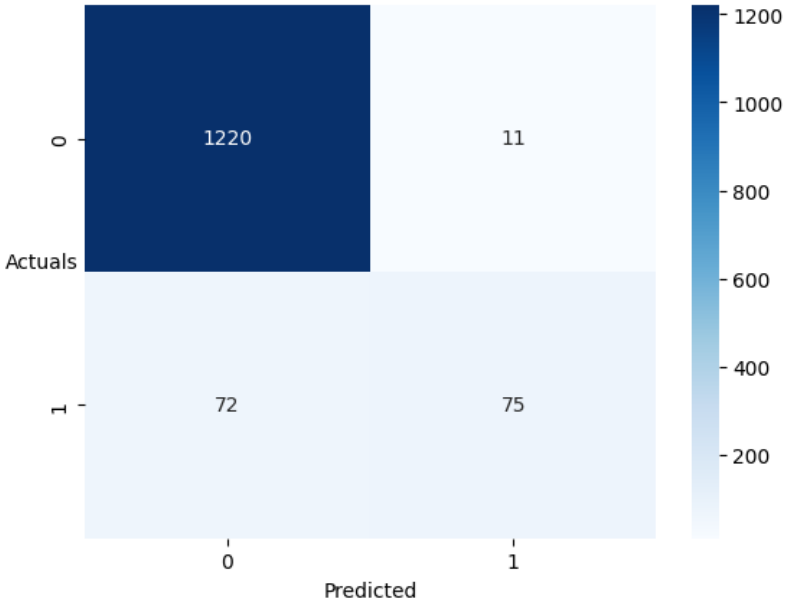
- We have taken the same columns that we have used for linear model, we are using grid search CV to tune the parameters

```
{'max_depth': 5,
 'min_samples_leaf': 5,
 'min_samples_split': 15,
 'n_estimators': 50}
```

- These are the parameters selected by grid_search cv as best parameters. Once we build model on train data using these tuned hyper-parameters we get:

	precision	recall	f1-score	support
0.0	0.94	0.99	0.97	1231
1.0	0.87	0.51	0.64	147
accuracy			0.94	1378
macro avg	0.91	0.75	0.81	1378
weighted avg	0.94	0.94	0.93	1378

- We are only getting 51% of recall only. These are not so great results.



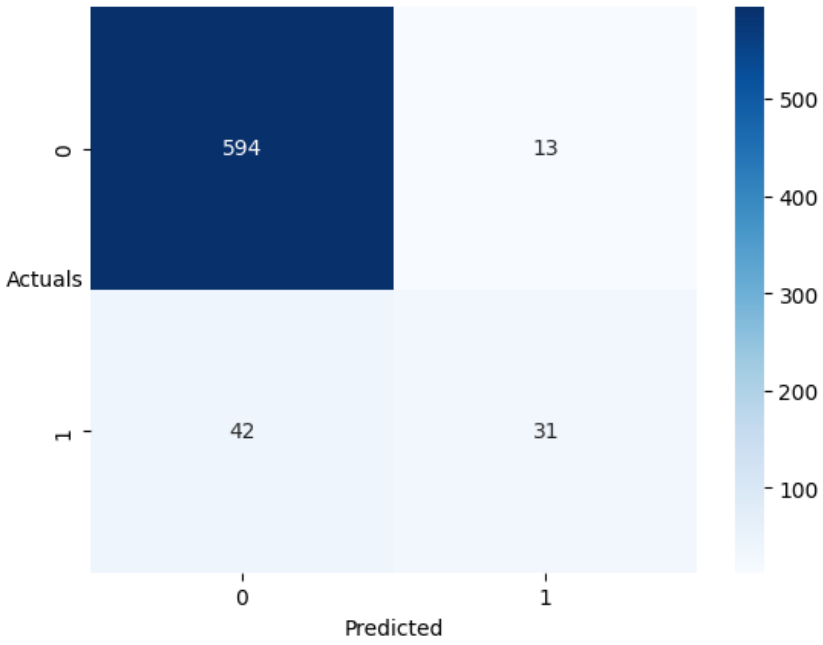
- We made 72 misclassifications, where we predicted actual defaulters as non-defaulters.

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

Let's check on test data set-

	precision	recall	f1-score	support
0.0	0.93	0.98	0.96	607
1.0	0.70	0.42	0.53	73
accuracy			0.92	680
macro avg	0.82	0.70	0.74	680
weighted avg	0.91	0.92	0.91	680

- We are only getting 42% of recall only. This is not a good model to use.



- We made 42 misclassifications, where we predicted actual defaulters as non-defaulters in test data.

Interpretation from model-

- The hyperparameters we've listed were selected by GridSearchCV to optimize the performance of a Random Forest model:
- **max_depth:** This limits how deep each decision tree can grow. A value of 5 means each tree will have a maximum of 5 levels, preventing overly complex trees.
- **min_samples_leaf**: This specifies the minimum number of samples required for a leaf node. With a value of 5, it ensures that leaf nodes contain a minimum of 5 data points, promoting generalization.
- **min_samples_split:** This sets the minimum number of samples required to split an internal node. A threshold of 15 ensures that a node will only be split if it contains at least 15 samples.
- **n_estimators:** This determines the number of decision trees in the Random Forest. Having 50 trees contributes to model stability and improved prediction quality.

- Overall, these settings were chosen to strike a balance between model complexity, avoiding overfitting, and achieving good predictive performance.
- With this we are not getting the desired results.

PART A: Build a LDA Model on Train Dataset. Also showcase your model building approach

- We have taken the same columns that we have used for the linear model.
- Equation we got is:

*y = 0.38043858 * _Research_and_development_expense_rate*

*+ 0.27726807 * _Interest_bearing_debt_interest_rate*

*- 0.20104734 * _Total_Asset_Growth_Rate*

*- 0.55002414 * _Net_Value_Growth_Rate*

*- 0.40064468 * _Interest_Expense_Ratio*

*- 0.18327681 * _Accounts_Receivable_Turnover*

*+ 0.33187783 * _Fixed_Assets_Turnover_Frequency*

*- 0.47777908 * _Cash_to_Total_Assets*

*+ 0.53143399 * _Cash_to_Current_Liability*

*- 1.32075294 * _Retained_Earnings_to_Total_Assets*

*+ 0.25247948 * _Total_expense_to_Assets*

*- 0.78894195 * _Equity_to_Liability*

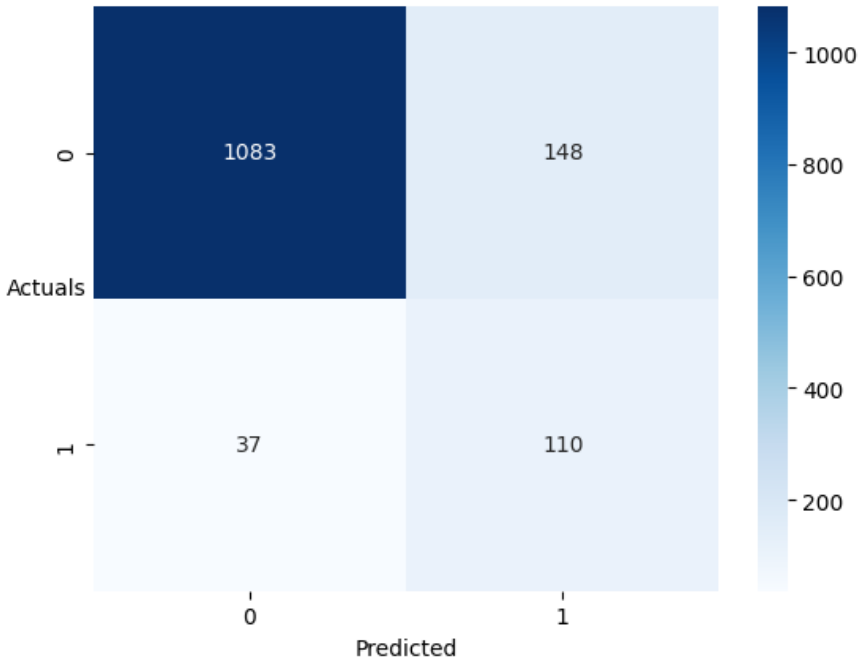
- 3.62639994

- Once this model has given us the predictions for the train dataset. Then we convert these predictions to 0 and 1 based on the optimum threshold.
- By identifying the threshold that maximizes the difference between TPR and FPR, we're essentially finding the point on the ROC curve where the trade-off between sensitivity and specificity is optimal.

- **Threshold we got is 0.165**
- Recall (also known as sensitivity or true positive rate) plays a crucial role, especially when dealing with imbalanced datasets or situations where one class is more important to correctly identify than the other. Here's how recall is important in this case:
- Recall measures the proportion of actual positive cases (default cases) that the model correctly identifies as positive.
- Let's check performance on train data:

	precision	recall	f1-score	support
0.0	0.967	0.880	0.921	1231
1.0	0.426	0.748	0.543	147
accuracy			0.866	1378
macro avg	0.697	0.814	0.732	1378
weighted avg	0.909	0.866	0.881	1378

- Recall is 74.8% in predicting defaulters for train data who are actually defaulters.



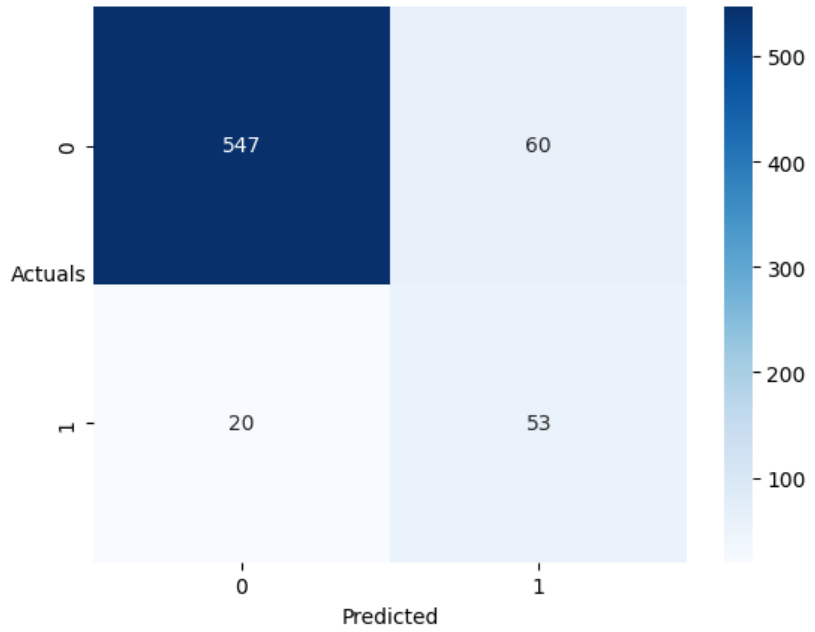
- We have 37 actual defaulters classified as non defaulters.

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

- Let's validate on test set:

	precision	recall	f1-score	support
0.0	0.965	0.901	0.932	607
1.0	0.469	0.726	0.570	73
accuracy			0.882	680
macro avg	0.717	0.814	0.751	680
weighted avg	0.912	0.882	0.893	680

- Recall is 72.6% in predicting defaulters for test data who are actually defaulters.



- Out of a total 73 actual defaulters, our model has rightly captured 53 defaulters in the test data set.
- Let's check the equation that is behind this model.

$$y = 0.38043858 * (\_Research\_and\_development\_expense\_rate)$$

$$+ 0.27726807 * (\_Interest\_bearing\_debt\_interest\_rate)$$

*- 0.20104734 * (_Total_Asset_Growth_Rate)*

*- 0.55002414 *(_Net_Value_Growth_Rate)*

*- 0.40064468 * (_Interest_Expense_Ratio)*

*- 0.18327681 * (_Accounts_Receivable_Turnover)*

*+ 0.33187783 * (_Fixed_Assets_Turnover_Frequency)*

*- 0.47777908 * (_Cash_to_Total_Assets)*

*+ 0.53143399 * (_Cash_to_Current_Liability)*

*- 1.32075294 * (_Retained_Earnings_to_Total_Assets)*

*+ 0.25247948 * (_Total_expense_to_Assets)*

*- 0.78894195 * (_Equity_to_Liability)*

- 3.62639994

Explanation:

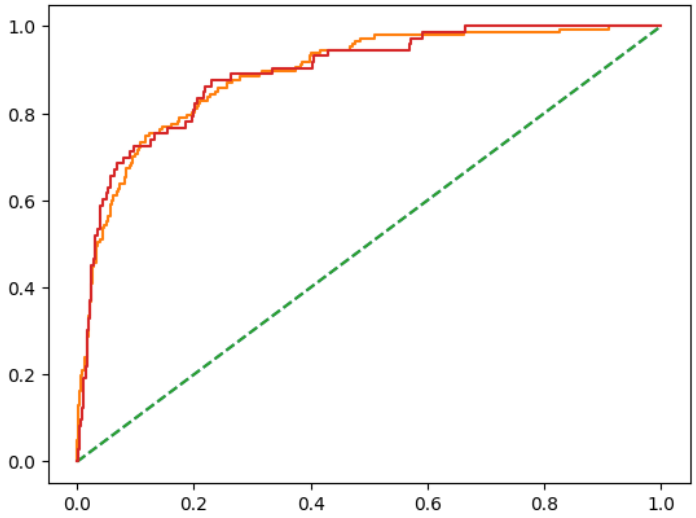
Variable	Impact on Predicted Probability of Default
_Research_and_development_expense_rate	Increase in R&D expenses leads to an increase in predicted probability of default
_Interest_bearing_debt_interest_rate	Increase in interest rates leads to an increase in predicted probability of default
_Total_Asset_Growth_Rate	Decrease in asset growth rate leads to an increase in predicted probability of default
_Net_Value_Growth_Rate	Decrease in net value growth rate leads to an increase in predicted probability of default

_Interest_Expense_Ratio	Decrease in interest expense ratio leads to an increase in predicted probability of default
_Accounts_Receivable_Turnover	Decrease in turnover leads to a slight increase in predicted probability of default
_Fixed_Assets_Turnover_Frequency	Increase in turnover frequency leads to an increase in predicted probability of default
_Cash_to_Total_Assets	Decrease in cash ratio leads to an increase in predicted probability of default
_Cash_to_Current_Liability	Increase in cash ratio leads to an increase in predicted probability of default
_Retained_Earnings_to_Total_Assets	Decrease in retained earnings ratio leads to an increase in predicted probability of default
_Total_expense_to_Assets	Increase in expense ratio leads to an increase in predicted probability of default
_Equity_to_Liability	Decrease in equity ratio leads to an increase in predicted probability of default

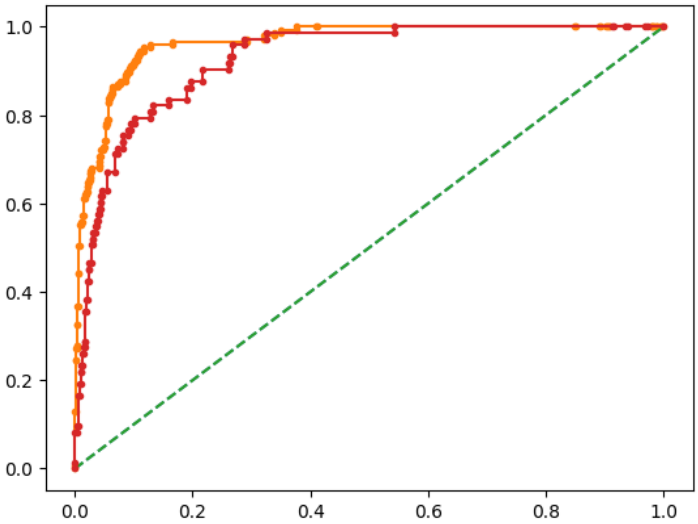
PART A: Compare the performances of Logistic Regression, Random Forest, and LDA models (include ROC curve)

- We will look at recalls that these models have given and the AUC scores to get to the best model.

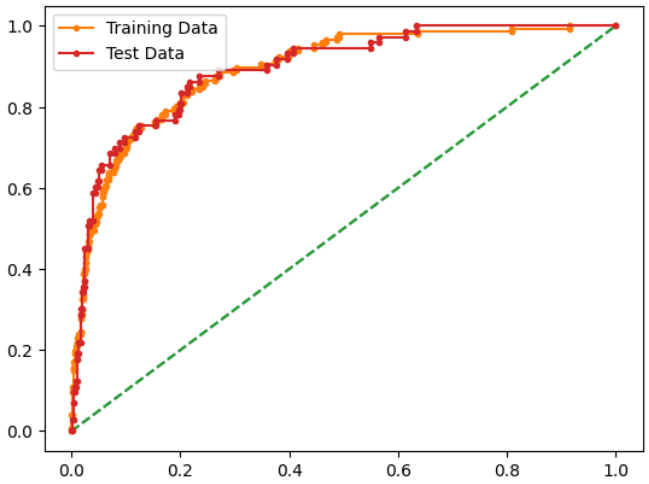
- Let's check AUC scores-
- **Logistic regression-Train_AUC: 0.891 , Test_AUC: 0.893**



- **Random Forest- Train_AUC: 0.964, Test_AUC: 0.926**



- **LDA-Train_AUC: 0.892 , Test_AUC: 0.895**



Let's check different parameters (in %)

Model	Train_recall	Test_recall	AUC_train	AUC_test
Logistic Regression	75.5	72.6	89.1	89.3
Random Forest	51	42	96.4	92.6
LDA	74.8	72.6	89.2	89.5

- While Random Forest has high AUC scores, its recall values are comparatively lower, suggesting that it might not perform as well in identifying actual default cases. LDA performs similarly to Logistic Regression in terms of recall and AUC scores.
- Overall, considering both recall and AUC scores, Logistic Regression seems to be the best model among the three presented. It strikes a balance between correctly identifying default cases and maintaining a good performance on the overall data, making it a suitable choice for predicting default probabilities.

PART A: Conclusions and Recommendations

- When making investment decisions, we can factor in the probability of default as inferred from the model's predictions. Here's how we can approach it based on the impact of these parameters on the predicted probability of default:
    - **Research and Development Expense Rate:** Look for companies that manage their R&D expenses well, as higher R&D expenses might contribute to an increased probability of default.
    - **Interest Bearing Debt Interest Rate:** Companies with higher interest rates on their debt might have an elevated probability of default. Seek those with manageable interest rates for a more stable investment.
    - **Total Asset Growth Rate:** Companies with decreasing asset growth rates could have an increased probability of default. Consider those with consistent or positive growth in assets.
    - **Net Value Growth Rate:** Companies with decreasing net value growth rates might have a higher likelihood of default. Prioritize those with positive growth in net value for a more promising investment.
    - **Interest Expense Ratio:** Lowering the interest expense ratio can lead to a decreased probability of default. Companies with effective management of interest expenses could be more resilient investments.
    - **Accounts Receivable Turnover:** Companies with decreasing turnover in accounts receivable might have an increased probability of default. Look for those with efficient turnover for better cash flow management.
    - **Fixed Assets Turnover Frequency:** Companies with decreasing turnover in fixed assets could face a higher probability of default. Consider those with a higher frequency of turnover for more effective asset utilization.
    - **Cash to Total Assets:** A lower cash-to-assets ratio can contribute to an elevated probability of default. Opt for companies with healthier liquidity positions for a safer investment.

- **Cash to Current Liability:** Companies with lower cash-to-liabilities ratios might face an increased probability of default. Choose those with stronger cash reserves relative to their liabilities.
- **Retained Earnings to Total Assets:** Companies with a decreasing ratio of retained earnings to total assets might have an elevated probability of default. Seek those with higher retained earnings for financial stability.