FINANCE AND RISK ANALYSIS

KIRAN.N - GREAT LEARNING

Milestone-1

Table of Contents

Table Of Figures

List Of Tables

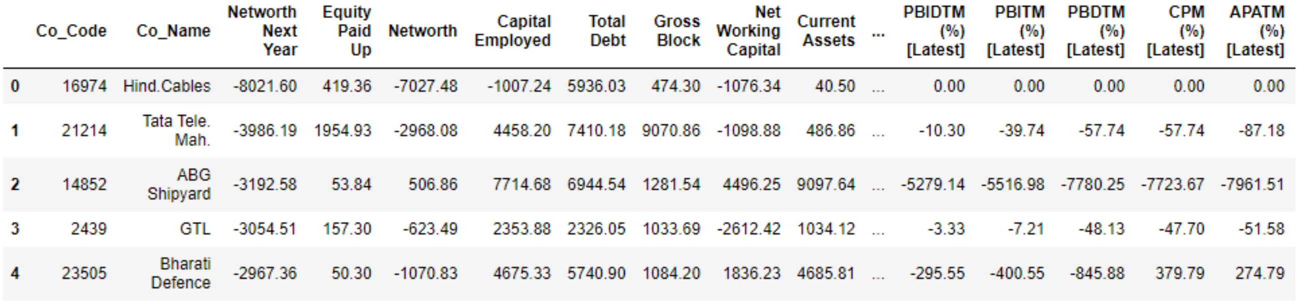
Problem Statement

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

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

Sample Of Dataset



	Co_Code	Co_Name	Networth Next Year	Equity Paid Up	Networth	Capital Employed	Total Debt	Gross Block	Net Working Capital	Current Assets	...	PBIDTM (%) [Latest]	PBITM (%) [Latest]	PBDTM (%) [Latest]	CPM (%) [Latest]	APATM (%) [Latest]
0	16974	Hind.Cables	-8021.60	419.36	-7027.48	-1007.24	5936.03	474.30	-1076.34	40.50	...	0.00	0.00	0.00	0.00	0.00
1	21214	Tata Tele. Mah.	-3986.19	1954.93	-2968.08	4458.20	7410.18	9070.86	-1098.88	486.86	...	-10.30	-39.74	-57.74	-57.74	-87.18
2	14852	ABG Shipyard	-3192.58	53.84	506.86	7714.68	6944.54	1281.54	4496.25	9097.64	...	-5279.14	-5516.98	-7780.25	-7723.67	-7961.51
3	2439	GTL	-3054.51	157.30	-623.49	2353.88	2326.05	1033.69	-2612.42	1034.12	...	-3.33	-7.21	-48.13	-47.70	-51.58
4	23505	Bharati Defence	-2967.36	50.30	-1070.83	4675.33	5740.90	1084.20	1836.23	4685.81	...	-295.55	-400.55	-845.88	379.79	274.79

Table 1: Sample Dataset

Dataset has 67 columns with 3586 rows. Each row in the dataset corresponds to one individual with their financial statement details.

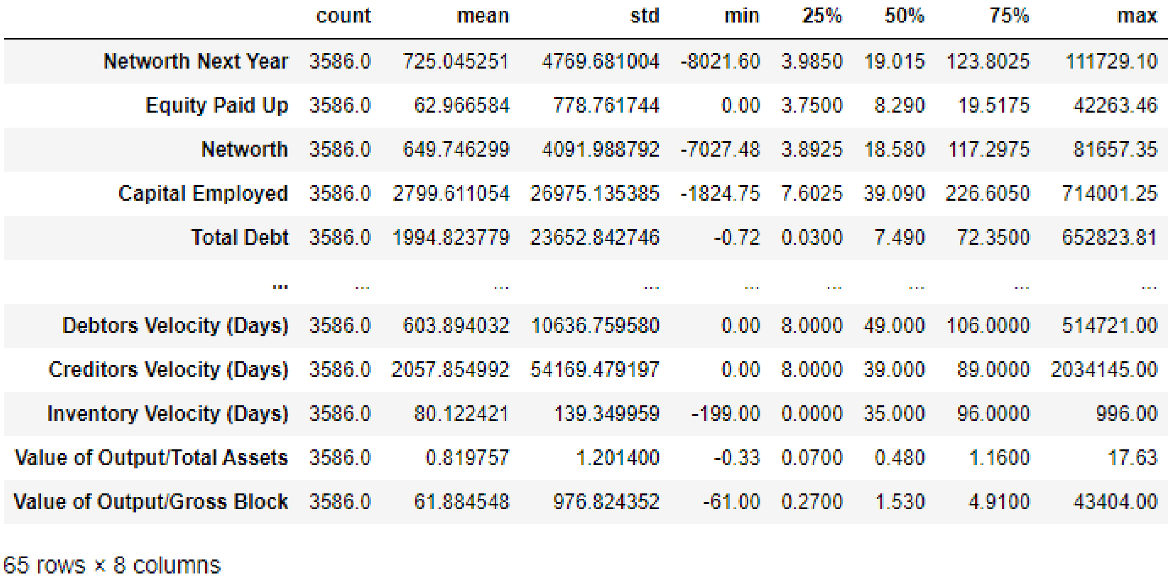
Exploratory Analysis

There are total 3586 rows and 67 columns in the dataset. Out of 67, 1 column is of object type, 3 columns are of integer type and rest 63 is of float data type.

There are total 118 null values are missing values resent in given dataset.

We are using 'ffill' method to impute null values in dataframe. This method replaces the NULL values with the value from the previous row.

Descriptive Statistics



	count	mean	std	min	25%	50%	75%	max
Networth Next Year	3586.0	725.045251	4769.681004	-8021.60	3.9850	19.015	123.8025	111729.10
Equity Paid Up	3586.0	62.966584	778.761744	0.00	3.7500	8.290	19.5175	42263.46
Networth	3586.0	649.746299	4091.988792	-7027.48	3.8925	18.580	117.2975	81657.35
Capital Employed	3586.0	2799.611054	26975.135385	-1824.75	7.6025	39.090	226.6050	714001.25
Total Debt	3586.0	1994.823779	23652.842746	-0.72	0.0300	7.490	72.3500	652823.81
...	...	...	...	...	...	...	...	...
Debtors Velocity (Days)	3586.0	603.894032	10636.759580	0.00	8.0000	49.000	106.0000	514721.00
Creditors Velocity (Days)	3586.0	2057.854992	54169.479197	0.00	8.0000	39.000	89.0000	2034145.00
Inventory Velocity (Days)	3586.0	80.122421	139.349959	-199.00	0.0000	35.000	96.0000	996.00
Value of Output/Total Assets	3586.0	0.819757	1.201400	-0.33	0.0700	0.480	1.1600	17.63
Value of Output/Gross Block	3586.0	61.884548	976.824352	-61.00	0.2700	1.530	4.9100	43404.00

65 rows × 8 columns

Table 2: Descriptive Statistics of Data

Q 1.1 Outlier Treatment

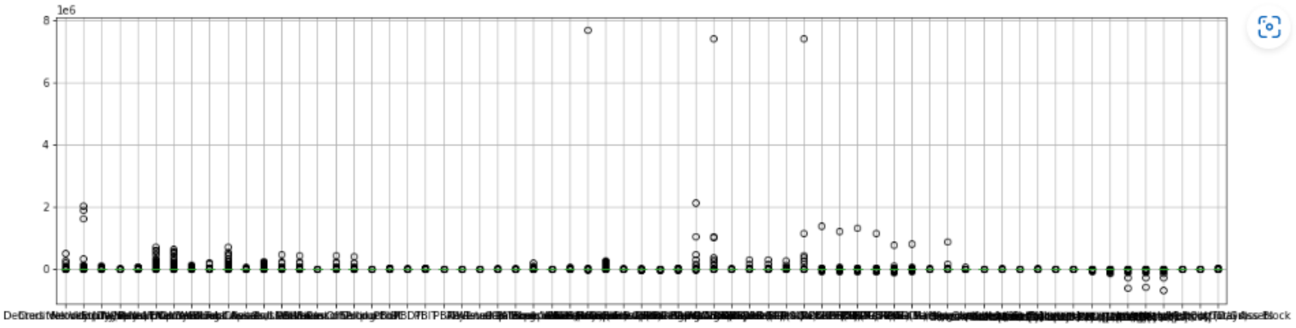


Figure 1: Box Plot of 65 Numeric Columns

Out of 66 numeric columns in dataframe 'Co_Code' column is a kind of Identifier; it has nothing to do with numeric calculations.

From the above figure it is evident that outliers are present all 65 numeric columns.

Outliers are data points in a dataset that are considered to be extreme, false, or not representative of what the data is describing. These outliers can be caused by either incorrect data collection or genuine outlying observations. Removing these outliers will often help your model to generalize better as these long tail observations could skew the learning.

Any data point below than Q1 – 1.5 times Inter Quartile Range and above Q3 + 1.5 times Inter Quartile Range is considered as outlier.

As part of Outlier treatment, values less than (Q1 – 1.5 * IQR) is set to lower boundary i.e (Q1 – 1.5 * IQR) and values more than (Q3 + 1.5 * IQR) is set to upper boundary i.e (Q3 + 1.5 * IQR).

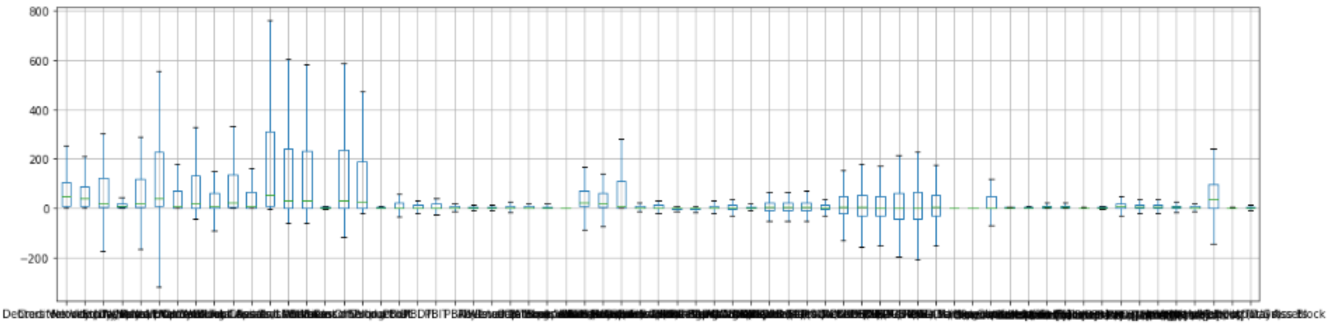


Figure 2: Box Plot of 65 Numeric Columns After Outlier Treatment

Q 1.2 Missing Value Treatment

There were total 118 null values are missing values resent in given dataset.

We are using ‘ffill’ method to impute null values in dataframe. This method replaces the NULL values with the value from the previous row.

Q 1.3 Transform Target variable into 0 and 1

We create a new column ‘default’ in dataframe. This column will be a dependent column. It takes the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

There is total 3199 rows with value as ‘0’ and 387 rows with value as ‘1’.

Nearly 11% of rows have value as ‘1’. The given dataset is not balanced dataset.

Q 1.5 Train Test Split

We use train_test_split functionality to split the given data into Test and Train split in required proportion randomly.

After splitting the data in 67:33 ratio:

- Test Split has 2402 rows of data.
- Train split has 1184 rows of data.

Q 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

Correlation heatmap is added in the IPYNB file. (Due to size constraint it is not added here.)

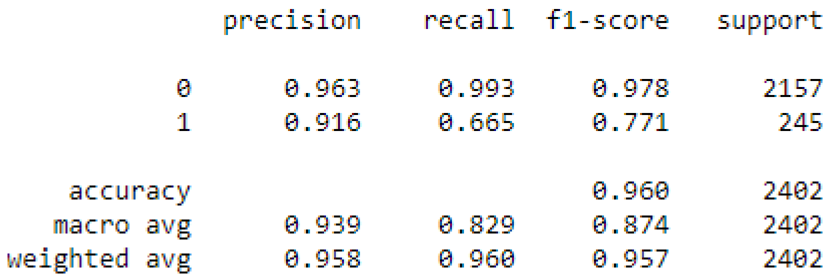
From the Correlation Heatmap we observe there is strong correlation among the variable(columns).
Using Variance Inflation Factor (VIF), we remove a column with highest VIF value and again calculate
VIF for remaining columns and remove the one with highest value. We continue this process until we have columns with VIF values less than '5'.

In this way we have eliminated 30+ columns from our given dataset.

Then we build a Logistic regression model. By going through the model summary we again remove all the columns whose 'p-value is greater than 0.05.

By following this procedure, we finally arrived at a model which was built using most important 8 independent columns which are Equity_Paid_Up, Total_Debt, Book_Value_Adj_Unit_Curr, CEPS, Current_Ratio, PBITM, Value_of_Output_Gross_Block, Debtors_Velocity.

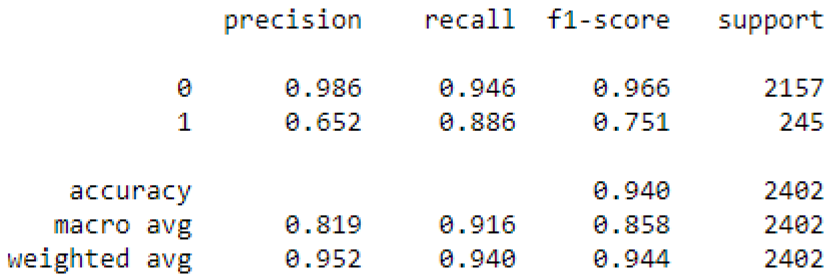
Following is the Classification Report with default threshold of 0.5



```
              precision    recall  f1-score   support

           0      0.963     0.993     0.978      2157
           1      0.916     0.665     0.771       245

    accuracy                          0.960      2402
   macro avg      0.939     0.829     0.874      2402
weighted avg      0.958     0.960     0.957      2402
```

Table 3: Train Data Classification Report For 0.5 Threshold

Using ROC we found optimal threshold as 0.203



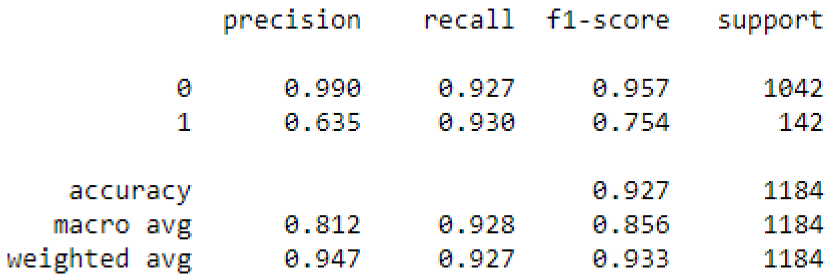
```
              precision    recall  f1-score   support

           0      0.986     0.946     0.966      2157
           1      0.652     0.886     0.751       245

    accuracy                          0.940      2402
   macro avg      0.819     0.916     0.858      2402
weighted avg      0.952     0.940     0.944      2402
```

Table 4: Train Data Classification Report For 0.203 Threshold

With the optimal Threshold value we observe there is a better Recall value compared to previous one.

Q1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model

To validate the model built, we will use the model to predict the ‘default’ values of test set with the same optimul threshold of 0.203.



```
              precision    recall  f1-score   support

           0      0.990     0.927     0.957      1042
           1      0.635     0.930     0.754       142

    accuracy                          0.927      1184
   macro avg      0.812     0.928     0.856      1184
weighted avg      0.947     0.927     0.933      1184
```

Table 5: Test Data Classification Report For 0.203 Threshold

By comparing Table 4 and 5 we infer the following:

- Accuracy is 90+ so the model built is a good model.
- Precision & Recall values of both tables are quite similar, so the model is neither over fit nor under fit.
- Precision value is quite good so model returns more relevant results than irrelevant ones.

Q 1.4 Univariate & Bivariate analysis with proper interpretation.

Univariate Analysis

1) Equity_Paid_Up

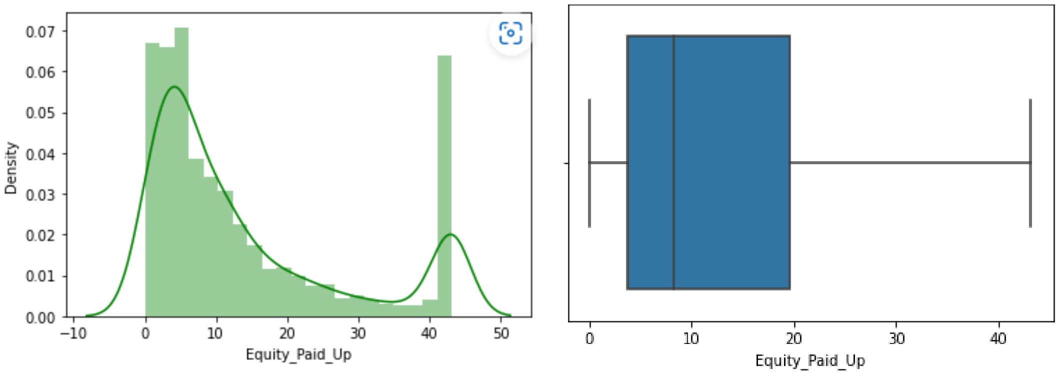


Figure 3: Histogram & Boxplot of Equity Paid Up

Here Data is Rightly skewed.

2) Total_Debt

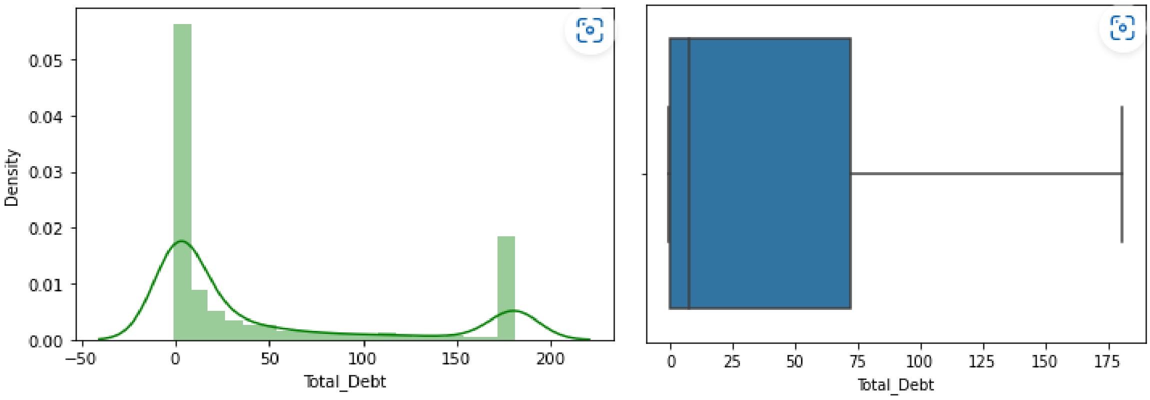


Figure 4: Histogram & Boxplot of Total Debt

Here Data is Rightly skewed.

3) Book_Value_Adj_Unit_Curr

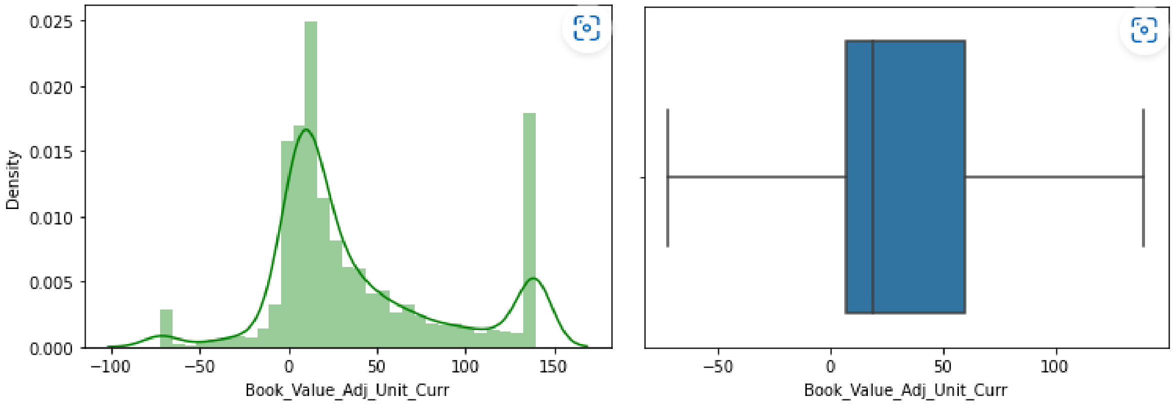


Figure 5: Histogram & Boxplot of Book_Value_Adj_Unit_Curr

Data is normally Distributed.

4) CEPS

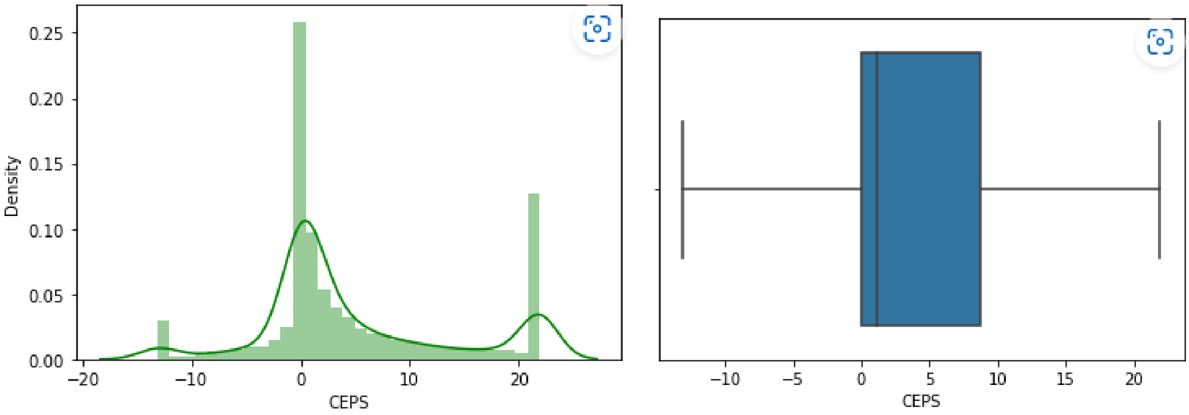


Figure 6: Histogram & Boxplot of CEPS

Data is Normally Distributed.

5) Current_Ratio

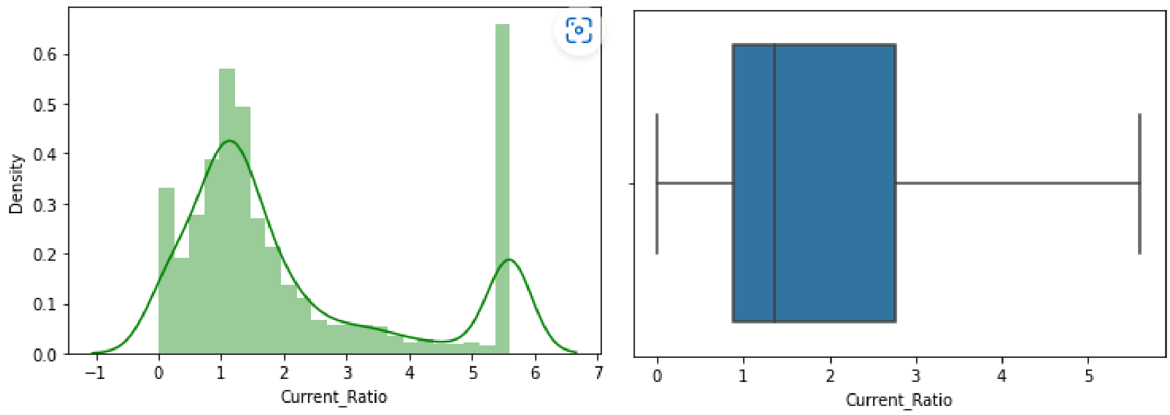


Figure 7: Histogram & Boxplot of Current Ratio

Here Data is Rightly skewed.

6) PBITM

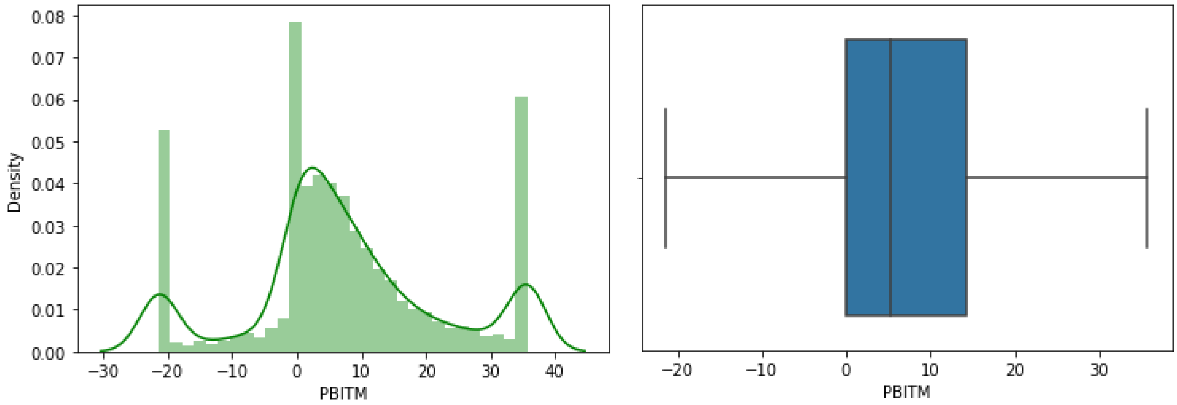


Figure 8: Histogram & Boxplot of PBITM

Data is Normally Distributed.

7) Value_of_Output_Gross_Block

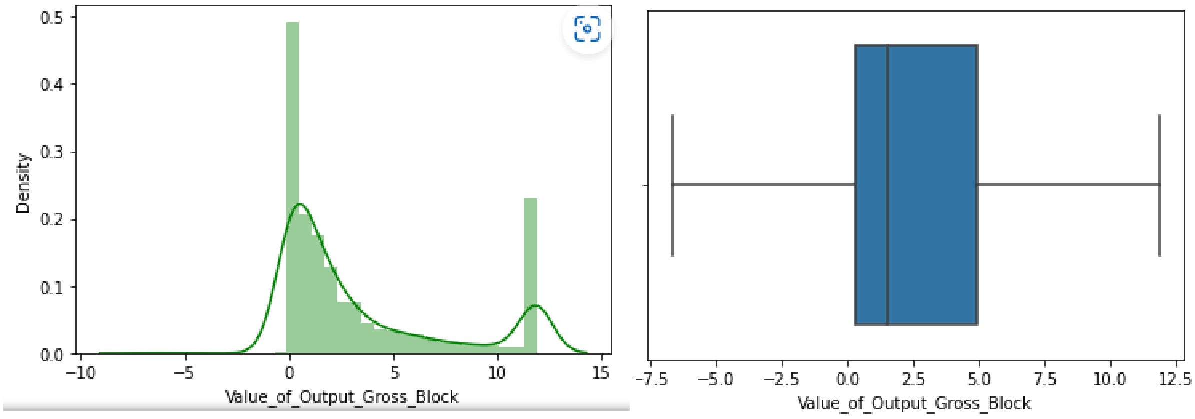


Figure 9: Histogram & Boxplot of Value_of_Output_Gross_Block

Here data is slightly Right Skewed.

8) Debtors_Velocity

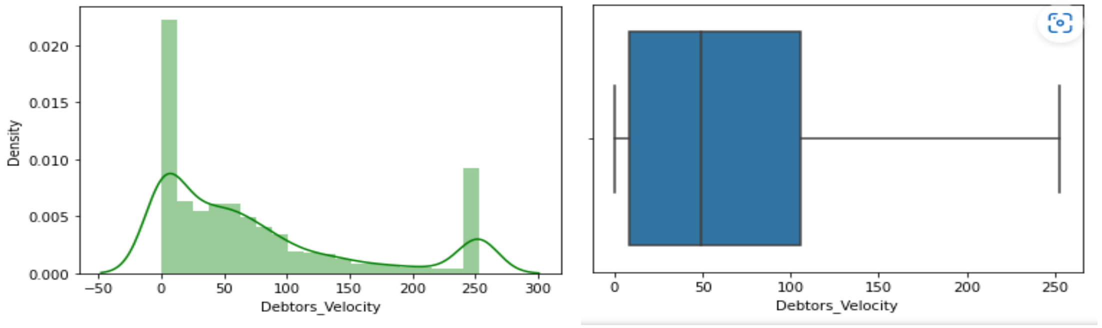


Figure 10: Histogram & Boxplot of Debtors Velocity

Here Data is Rightly skewed.

Bi-Variate Analysis

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

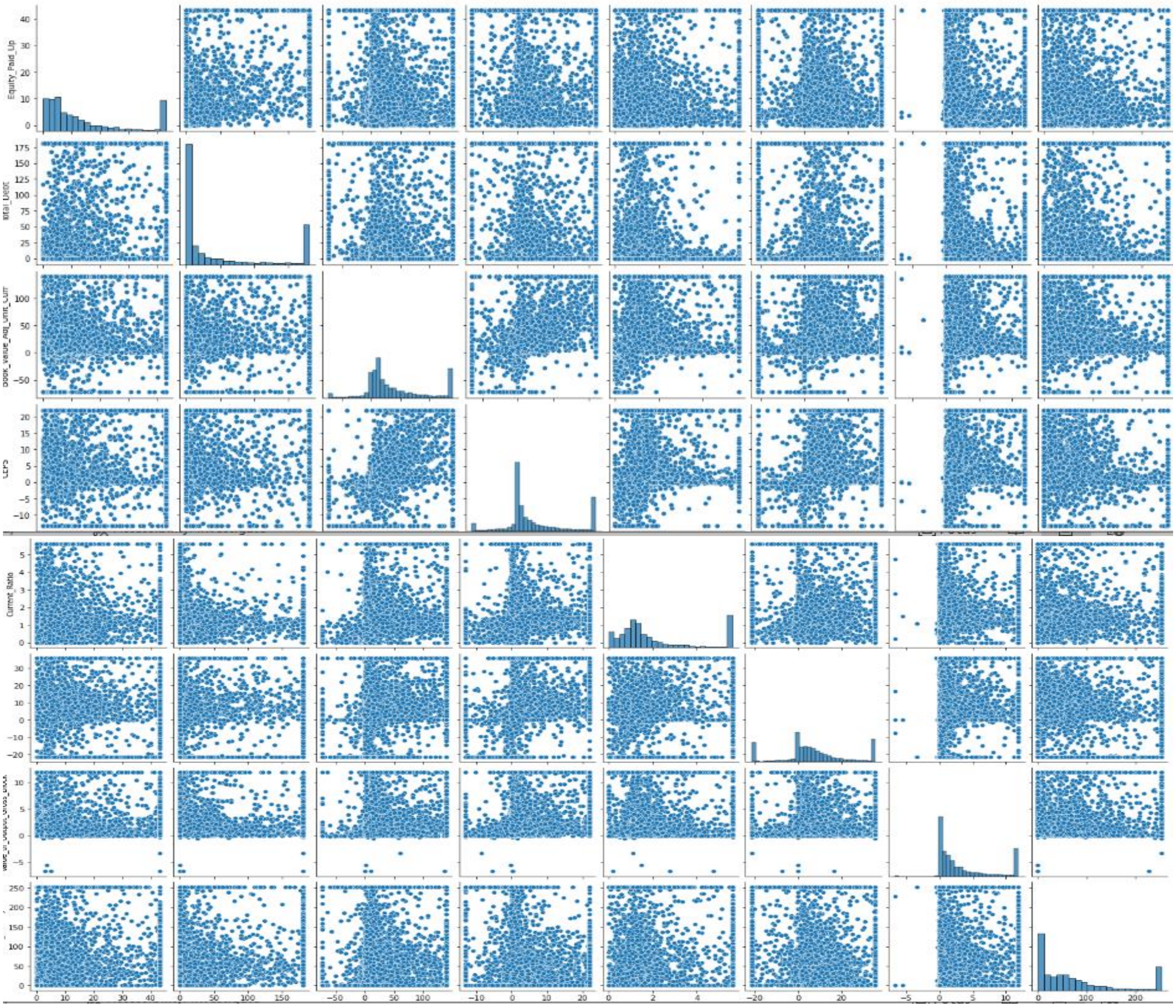


Figure 11: Pair Plot of Important Columns

In Pair plot we can analyse how 1 variable varies w.r.t other variable.

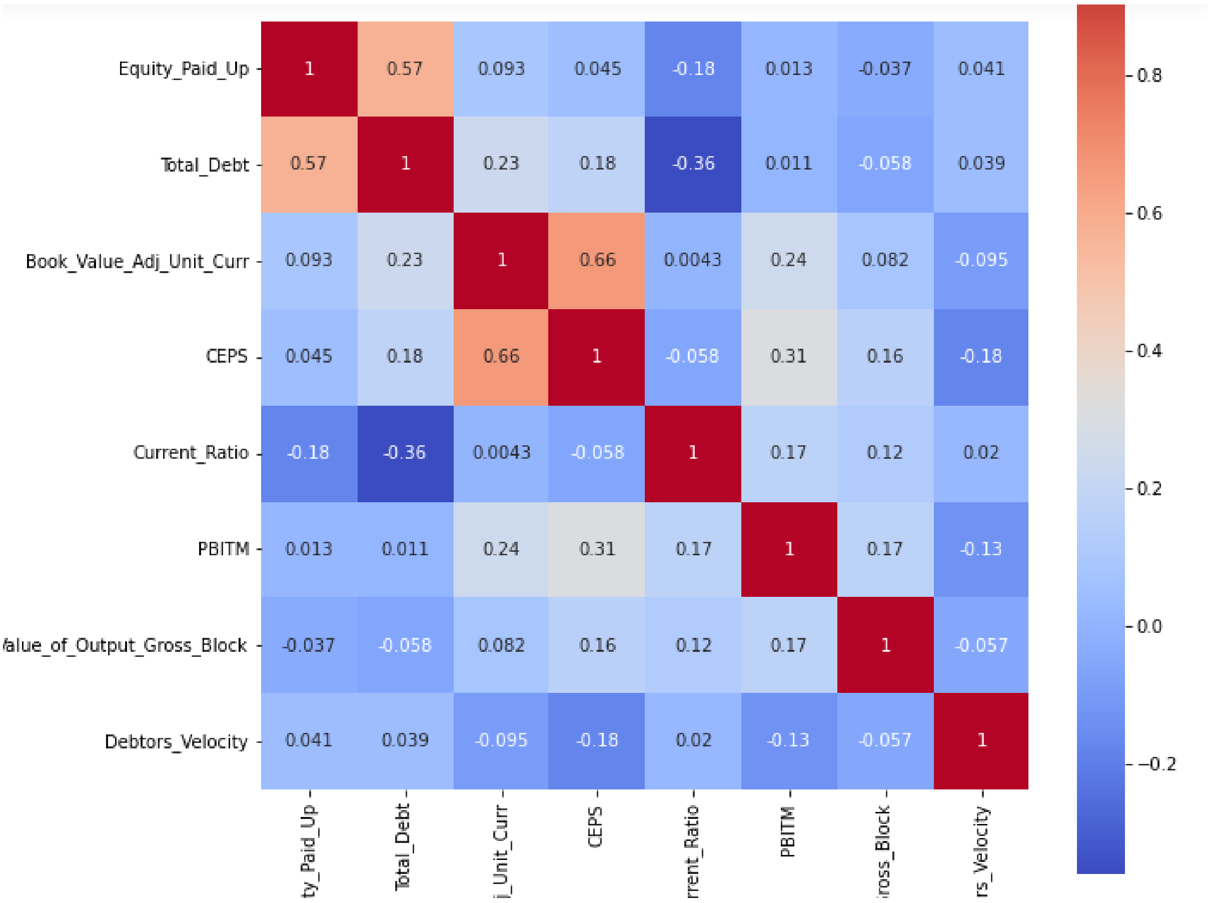


Figure 12: Correlation Heat Map of Important Columns