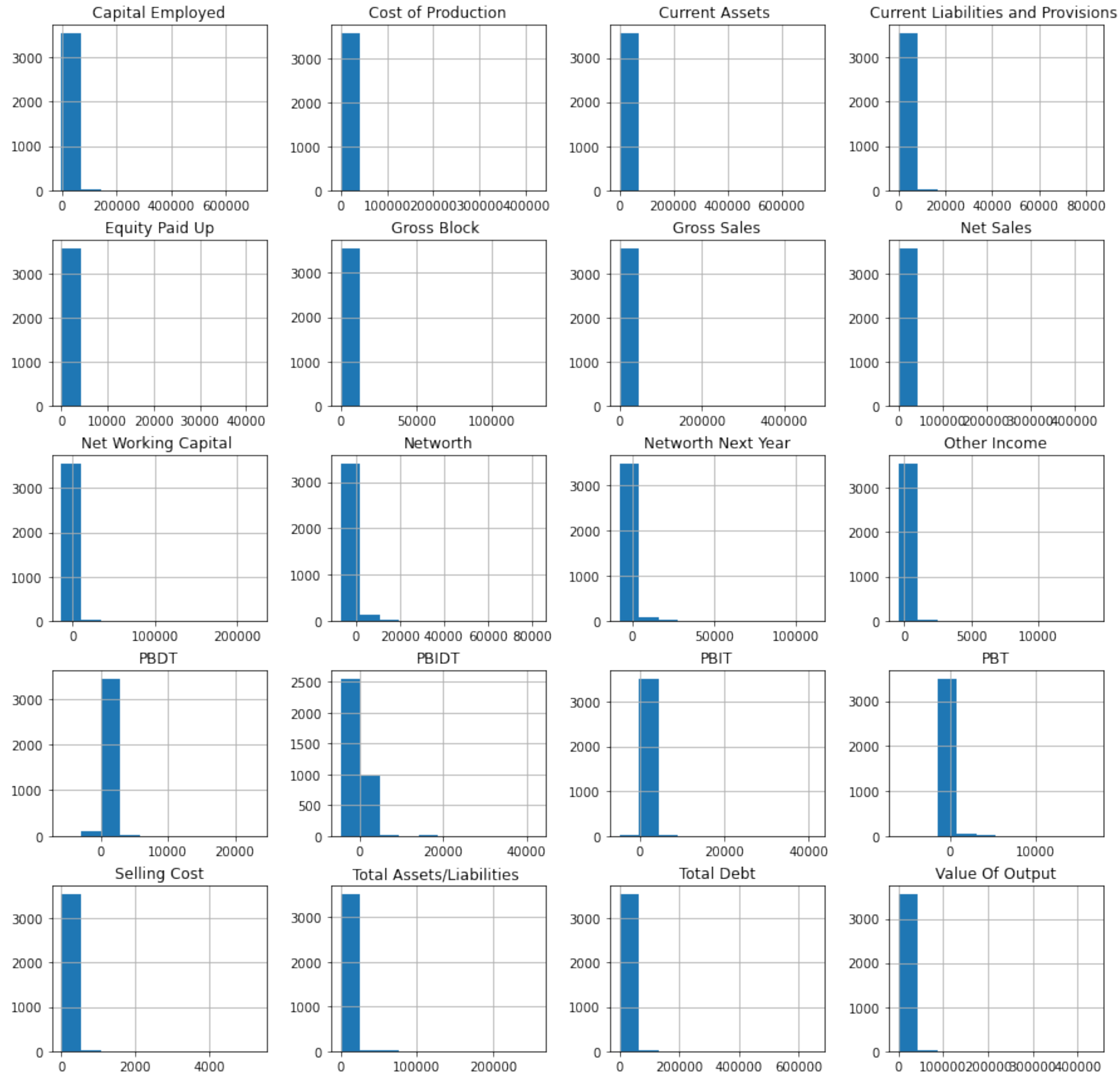
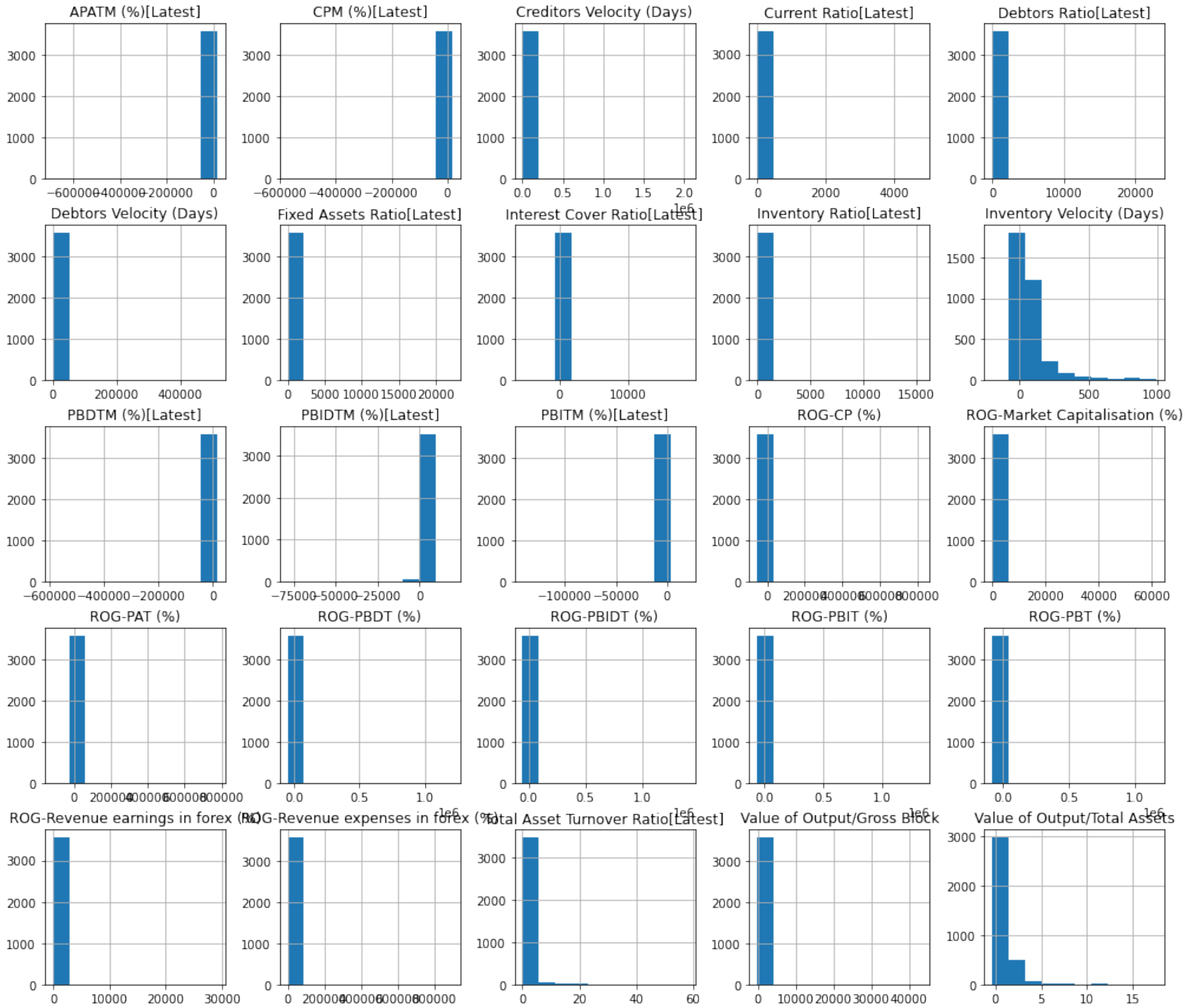
Table of Contents

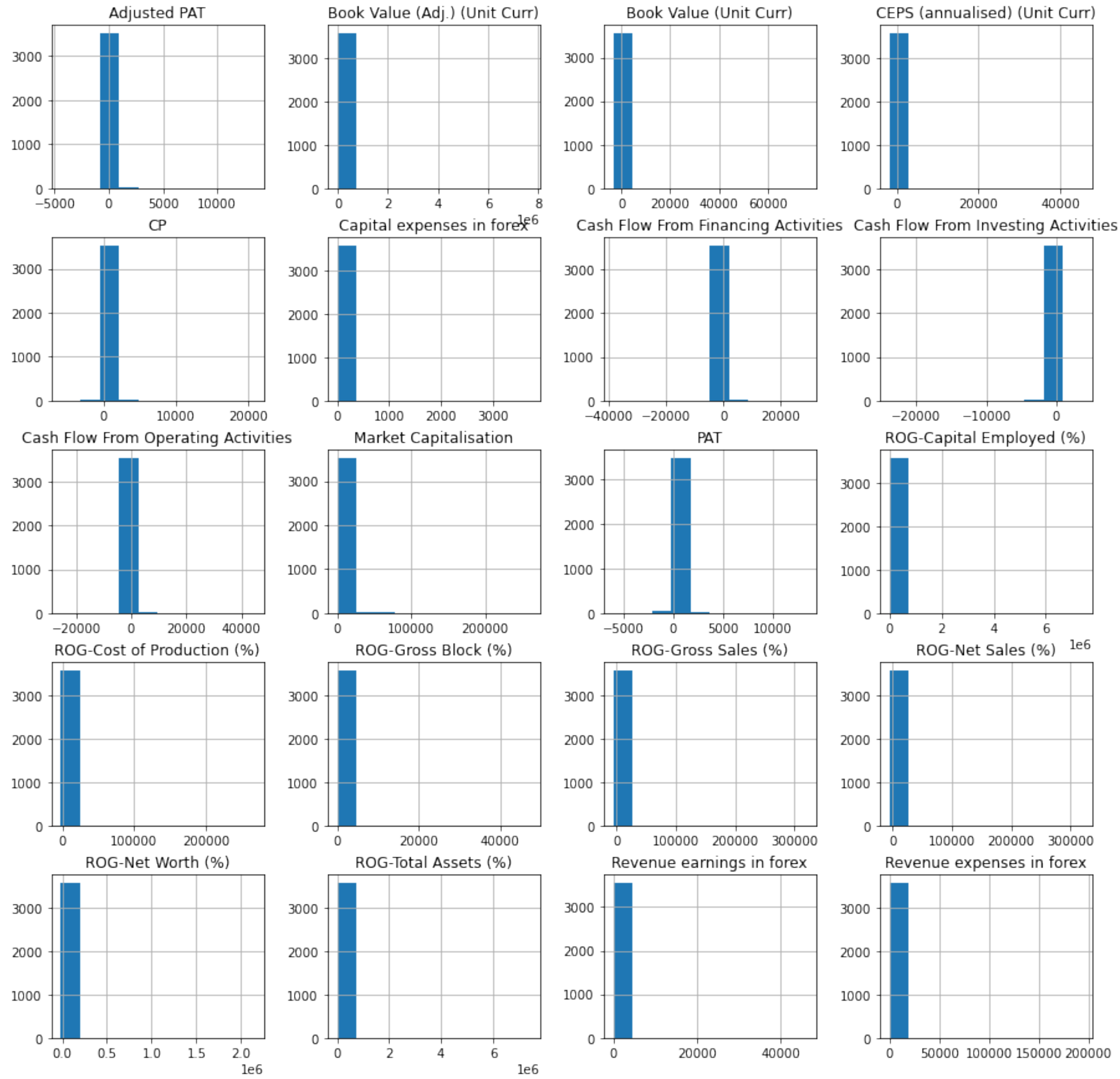
Dataset that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Net worth of the company in the following year (2016) is provided which can be used to drive the labelled field.

Dataset has 3587 entries of different companies across 67 columns.

Let’s check out the distribution of every variable. Below are the histograms:



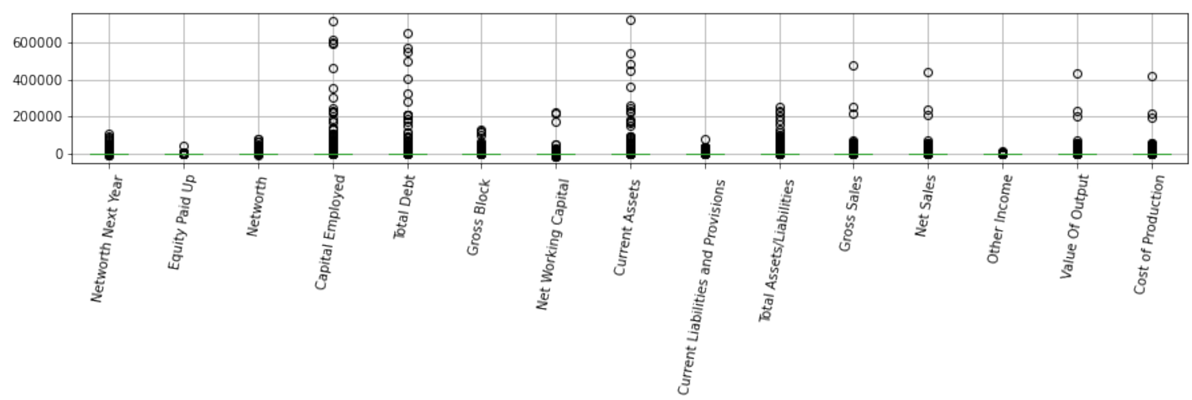

APATM (%)[Latest]
CPM (%)[Latest]
Creditors Velocity (Days)
Current Ratio[Latest]
Debtors Ratio[Latest]
Debtors Velocity (Days)
Fixed Assets Ratio[Latest]
Interest Cover Ratio[Latest]
Inventory Ratio[Latest]
Inventory Velocity (Days)
PBDTM (%)[Latest]
PBIDTM (%)[Latest]
PBITM (%)[Latest]
ROG-CP (%)
ROG-Market Capitalisation (%)
ROG-PAT (%)
ROG-PBDT (%)
ROG-PBIDT (%)
ROG-PBIT (%)
ROG-PBT (%)
ROG-Revenue earnings in forex (%)
ROG-Revenue expenses in forex (%)
Total Asset Turnover Ratio[Latest]
Value of Output/Gross Block
Value of Output/Total Assets

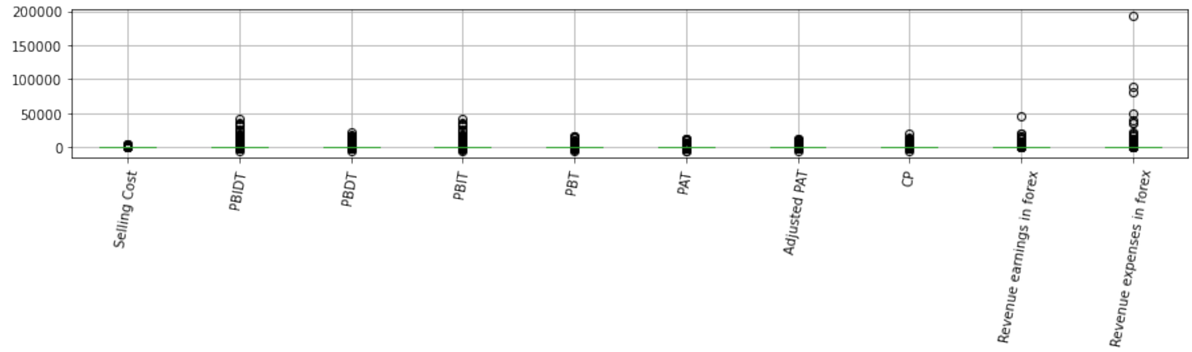


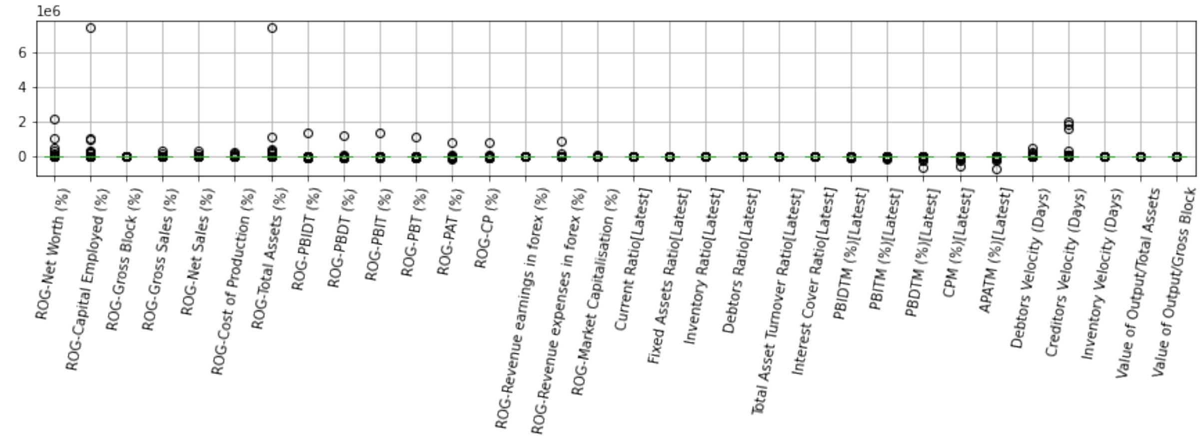
The distribution for almost all the variables looks extremely skewed to the left (i.e. in the first 25% in the quantile). Few Variables are right skewed.

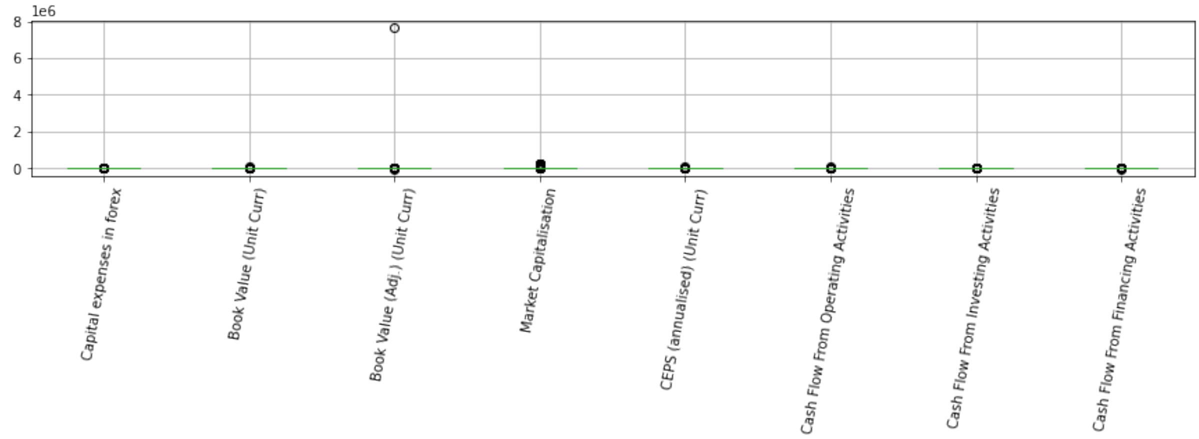
1.1 Outlier Treatment

All the 67 variables have outliers, please see the below boxplots to observe outliers in each.

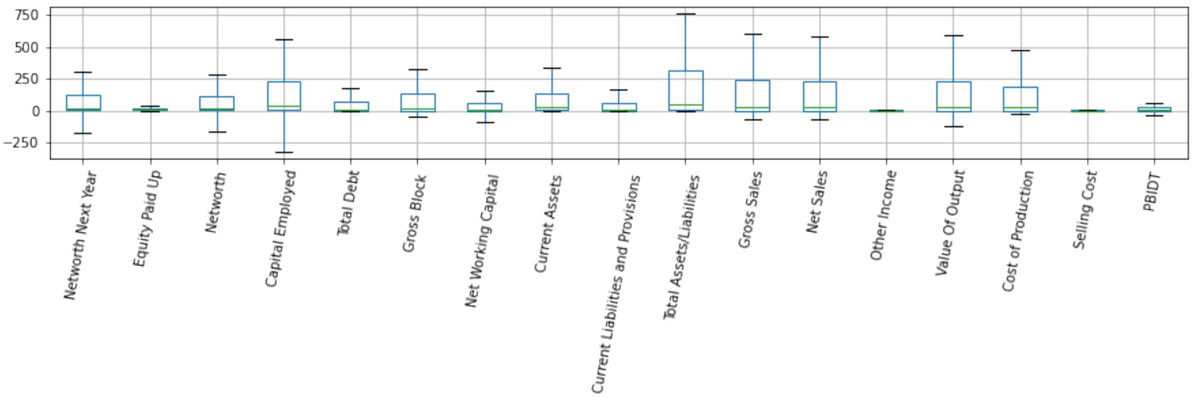


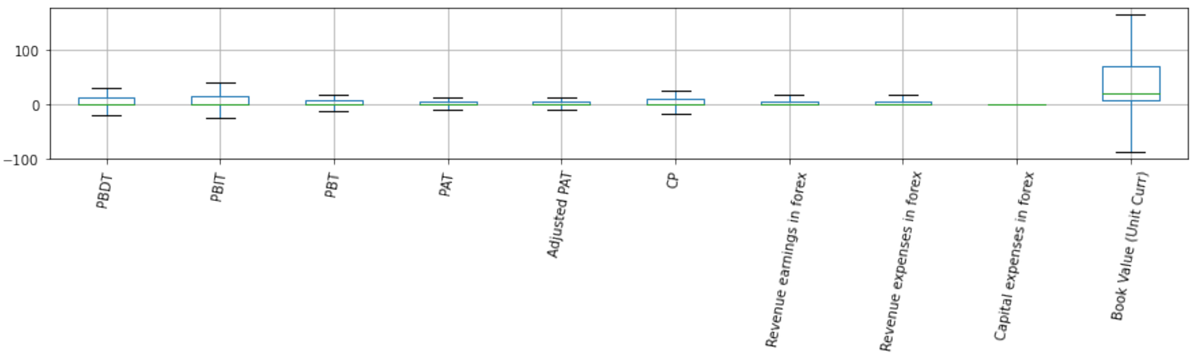


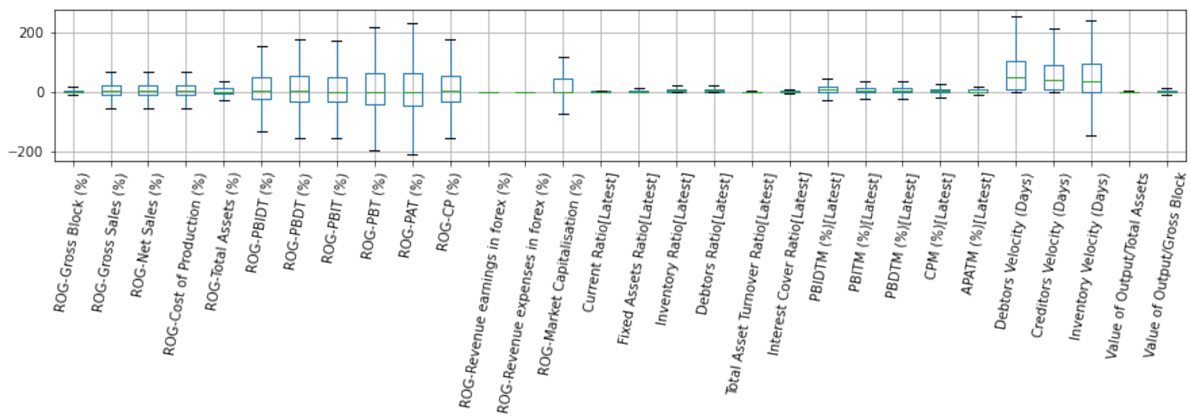


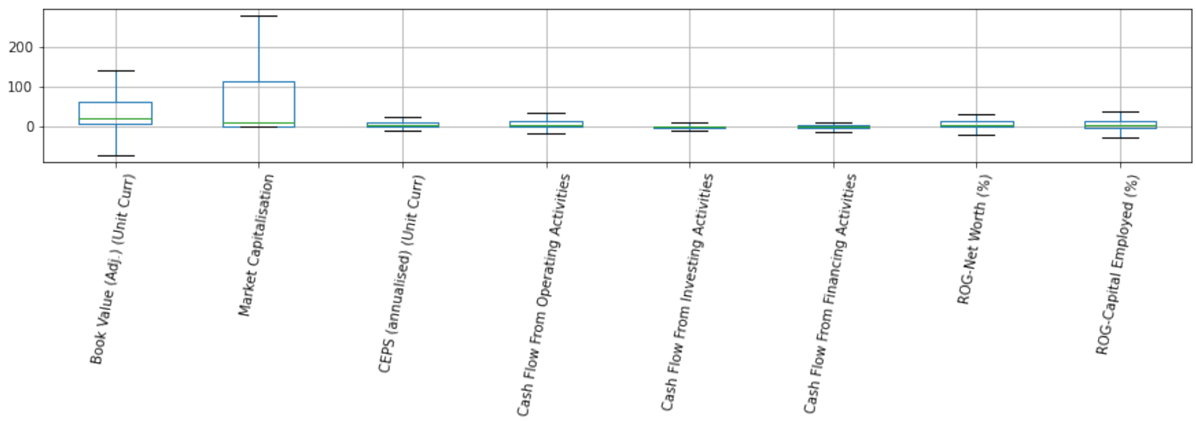


The outliers are treated using IQR (Inter Quantile Range). Below are the boxplots after the treatment.


750
500
250
0
−250
Networth Next Year
Equity Paid Up
Networth
Capital Employed
Total Debt
Gross Block
Net Working Capital
Current Assets
Current Liabilities and Provisions
Total Assets/Liabilities
Gross Sales
Net Sales
Other Income
Value Of Output
Cost of Production
Selling Cost
PBIDT


100
0
−100
PBDT
PBIT
PBT
PAT
Adjusted PAT
CP
Revenue earnings in forex
Revenue expenses in forex
Capital expenses in forex
Book Value (Unit Curr)


200
0
−200
ROG-Gross Block (%)
ROG-Gross Sales (%)
ROG-Net Sales (%)
ROG-Cost of Production (%)
ROG-Total Assets (%)
ROG-PBIDT (%)
ROG-PBDT (%)
ROG-PBIT (%)
ROG-PBT (%)
ROG-PAT (%)
ROG-CP (%)
ROG-Revenue earnings in forex (%)
ROG-Revenue expenses in forex (%)
ROG-Market Capitalisation (%)
Current Ratio[Latest]
Fixed Assets Ratio[Latest]
Inventory Ratio[Latest]
Debtors Ratio[Latest]
Total Asset Turnover Ratio[Latest]
Interest Cover Ratio[Latest]
PBIDTM (%)[Latest]
PBITM (%)[Latest]
PBDTM (%)[Latest]
CPM (%)[Latest]
APATM (%)[Latest]
Debtors Velocity (Days)
Creditors Velocity (Days)
Inventory Velocity (Days)
Value of Output/Total Assets
Value of Output/Gross Block


200
100
0
Book Value (Adj.) (Unit Curr)
Market Capitalisation
CEPS (annualised) (Unit Curr)
Cash Flow From Operating Activities
Cash Flow From Investing Activities
Cash Flow From Financing Activities
ROG-Net Worth (%)
ROG-Capital Employed (%)

1.2 Missing Value Treatment

Of the 67 columns, only few have missing values, below are those shows:

```
Book Value (Adj.) (Unit Curr)          4
Current Ratio[Latest]                  1
Fixed Assets Ratio[Latest]             1
Inventory Ratio[Latest]                1
Debtors Ratio[Latest]                  1
Total Asset Turnover Ratio[Latest]     1
Interest Cover Ratio[Latest]           1
PBIDTM (%)[Latest]                     1
PBITM (%)[Latest]                      1
PBDTM (%)[Latest]                      1
CPM (%)[Latest]                        1
APATM (%)[Latest]                      1
Debtors Velocity (Days)                0
Creditors Velocity (Days)              0
Inventory Velocity (Days)            103
```

The variable 'Inventory Velocity (Days)' has 103 missing value, the variable was just treated for outliers hence mean was used to substitute for missing values.
Similarly, the variable Book Value (Adj.) (Unit Curr), Current Ratio[Latest], Fixed Assets Ratio[Latest], Inventory Ratio[Latest], Debtors Ratio[Latest], Total Asset Turnover Ratio[Latest], Interest Cover Ratio[Latest], PBIDTM (%)[Latest], PBITM (%)[Latest], PBDTM (%)[Latest], CPM (%)[Latest], APATM (%)[Latest], Debtors Velocity (Days) were treated for missing values using mean.

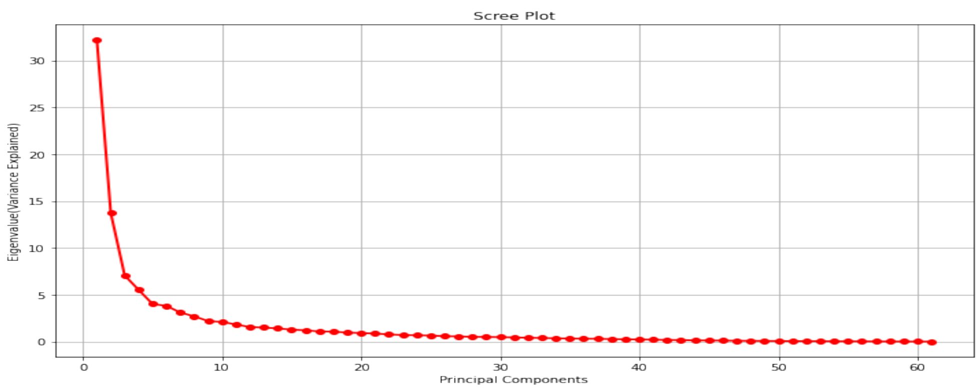
Scaling:

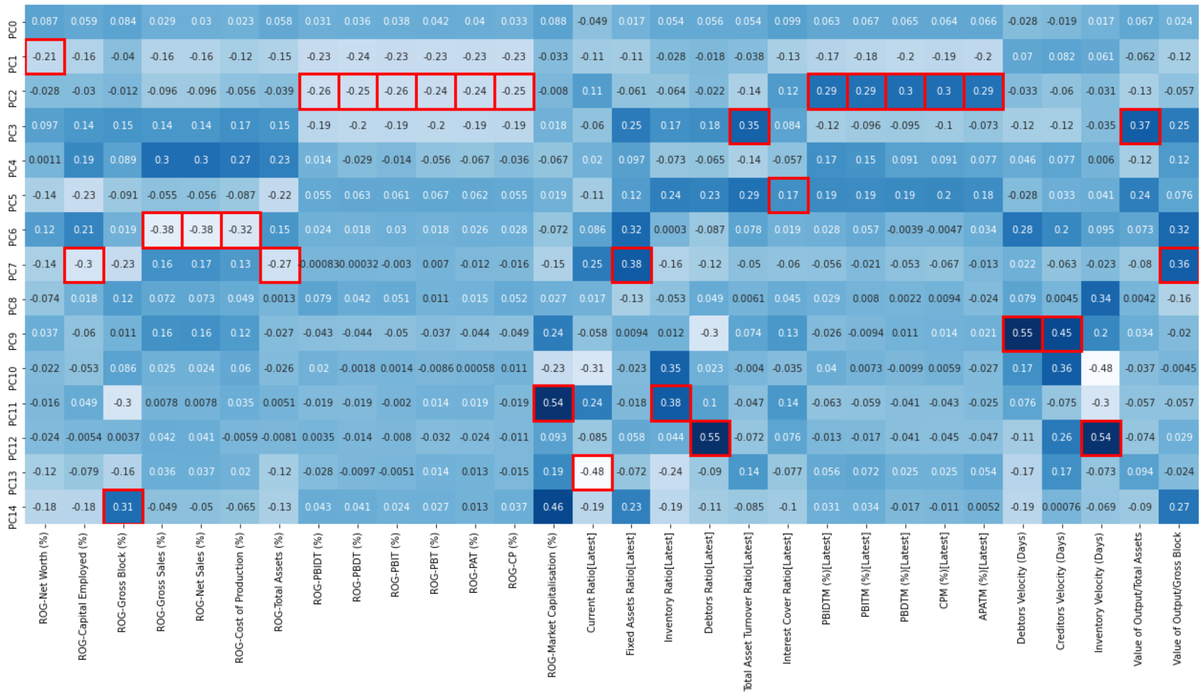
The dataset was scaled using z-score to standardize the physical measurements of the variables. E.g. Gross Sales was in 1000s whereas Profits were in %, they were standardize using zscore.

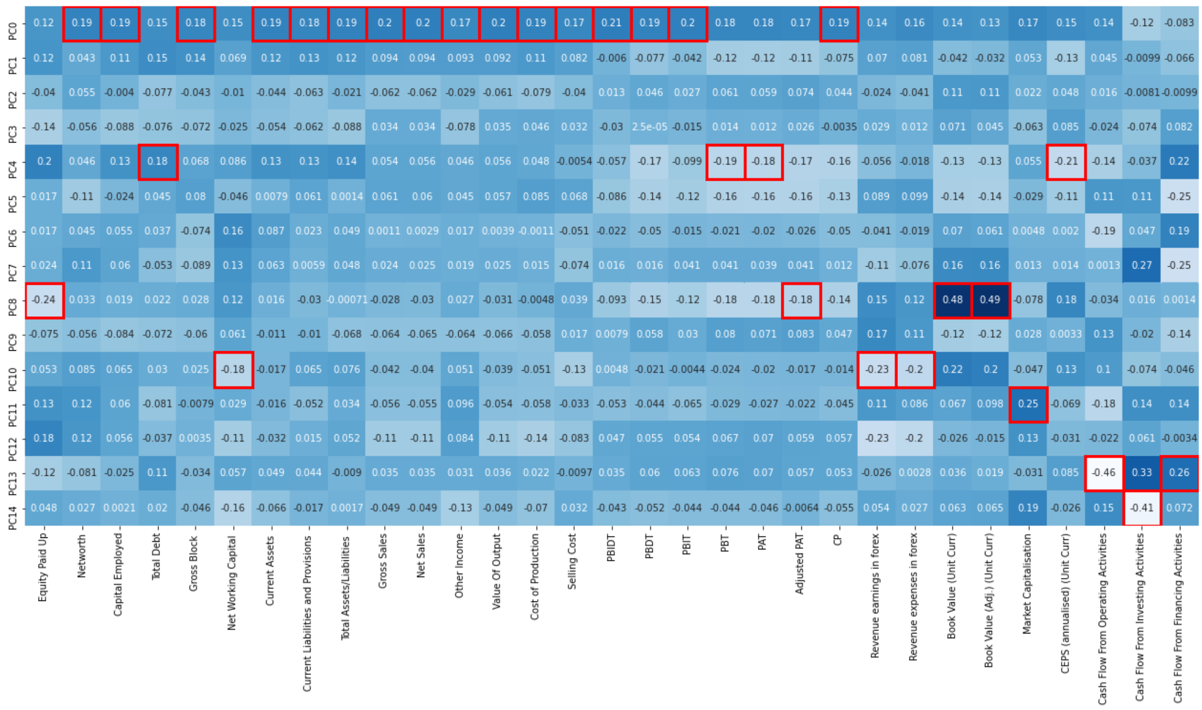
PCA:

Since we had 67 variables, dimensionality reduction technique was use to further extract fewer variable. PCA was use to capture most variance of data in top 13 components.

See below the Scree plot showing the principal components capturing variance, also see the table showing correlation between PCs and the actual variables:







The top variable showing strong correlation with every PC:

- PC0 - Assets/Costs/Profits
- PC1 - Equity Paid Up
- PC2 - ROG on Profits - Tax Margin
- PC3 - Asset Turnover/Output
- PC4 - CEPS
- PC5 - Debtors/ Total Assets/ Interest Cover
- PC6- ROG Sales
- PC7- ROG Capital employed/Total Asset

- PC8 - Book Value (NAV)
- PC9- Debtors/Creditors Velocity(Days)
- PC10 - Revenue earning/expenses in Forex
- PC11 - ROG on gross/capitalisation
- PC12- Debtors Ration/ Inventory velocity
- PC13- Cash Flow Operating/Investing/Financing

1.3 Transform Target variable into 0 and 1

Default variable is created to take the value of 1 when net worth next year is negative & 0 when net worth is positive.

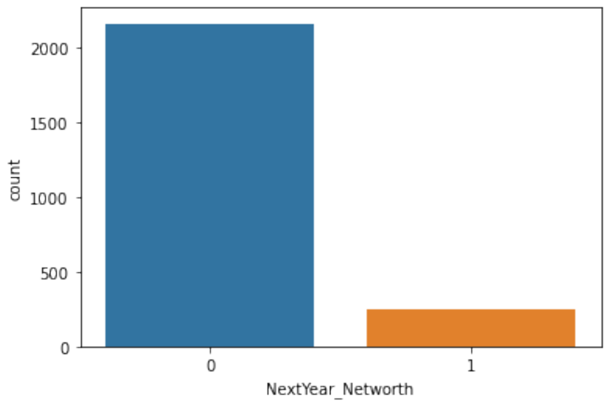
Showing the unique count in the 'Networth_NextYear' Variable:

```
array ([1., 0.])
```

1.4 Univariate & Bivariate analysis with proper interpretation

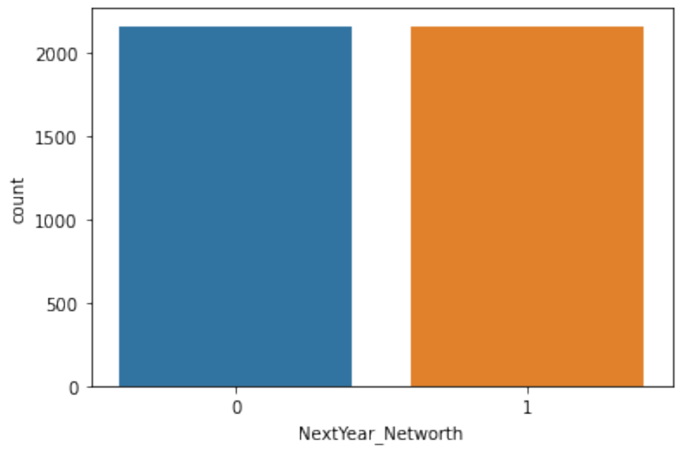
Univariate:

We have already seen Histograms and Boxplots as part of the univariate Analysis in previous topics. Let's have a look at the Count Plot of the dependant variable to check if the data is unbalanced or not.




```
0    0.898002
1    0.101998
```

since the data is heavily unbalanced, SMOTE was performed on Train set to get it balanced. See the count plot after SMOTE:

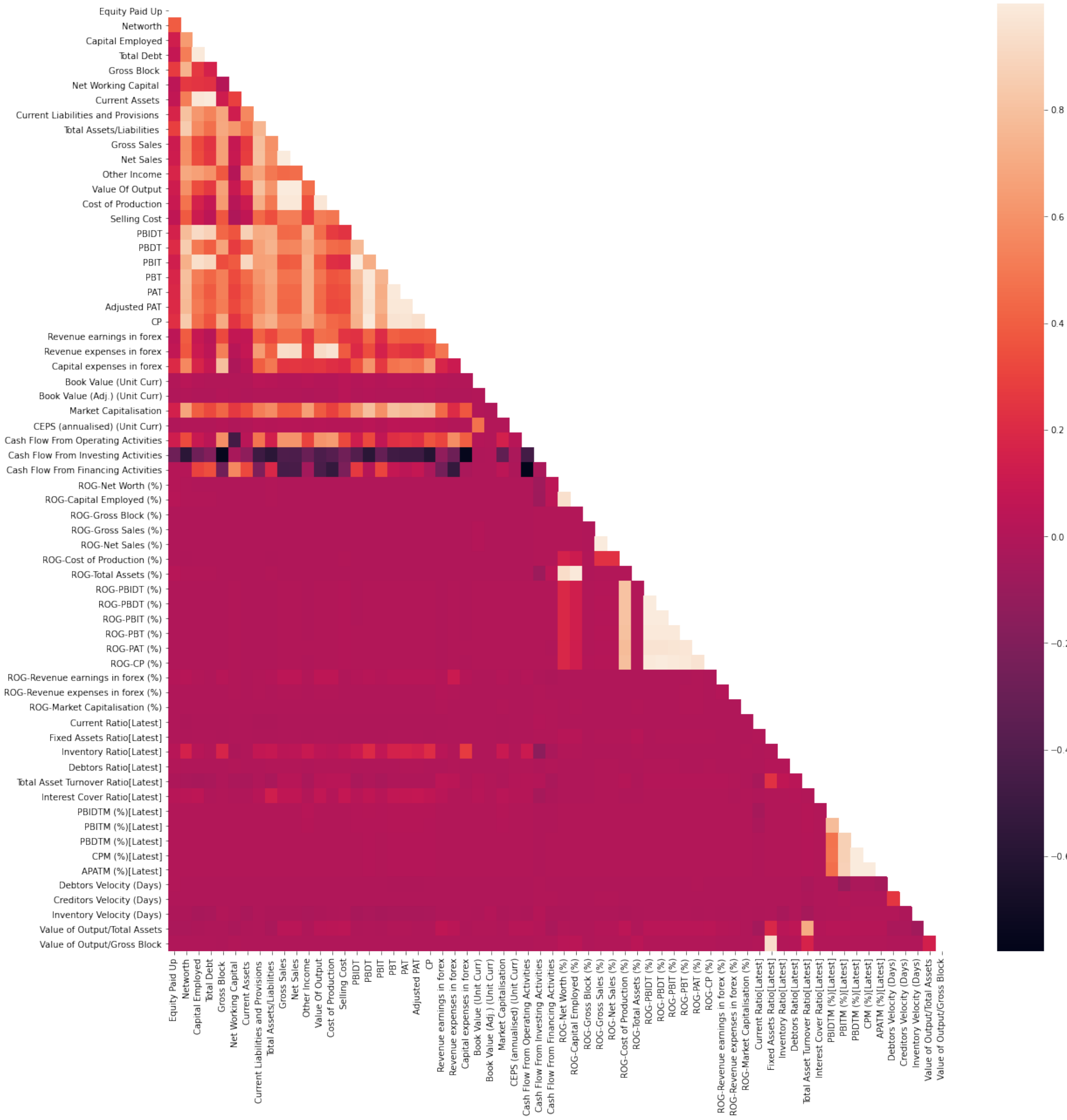



```
1    0.5
0    0.5
```

Bivariate:

Heatmaps:

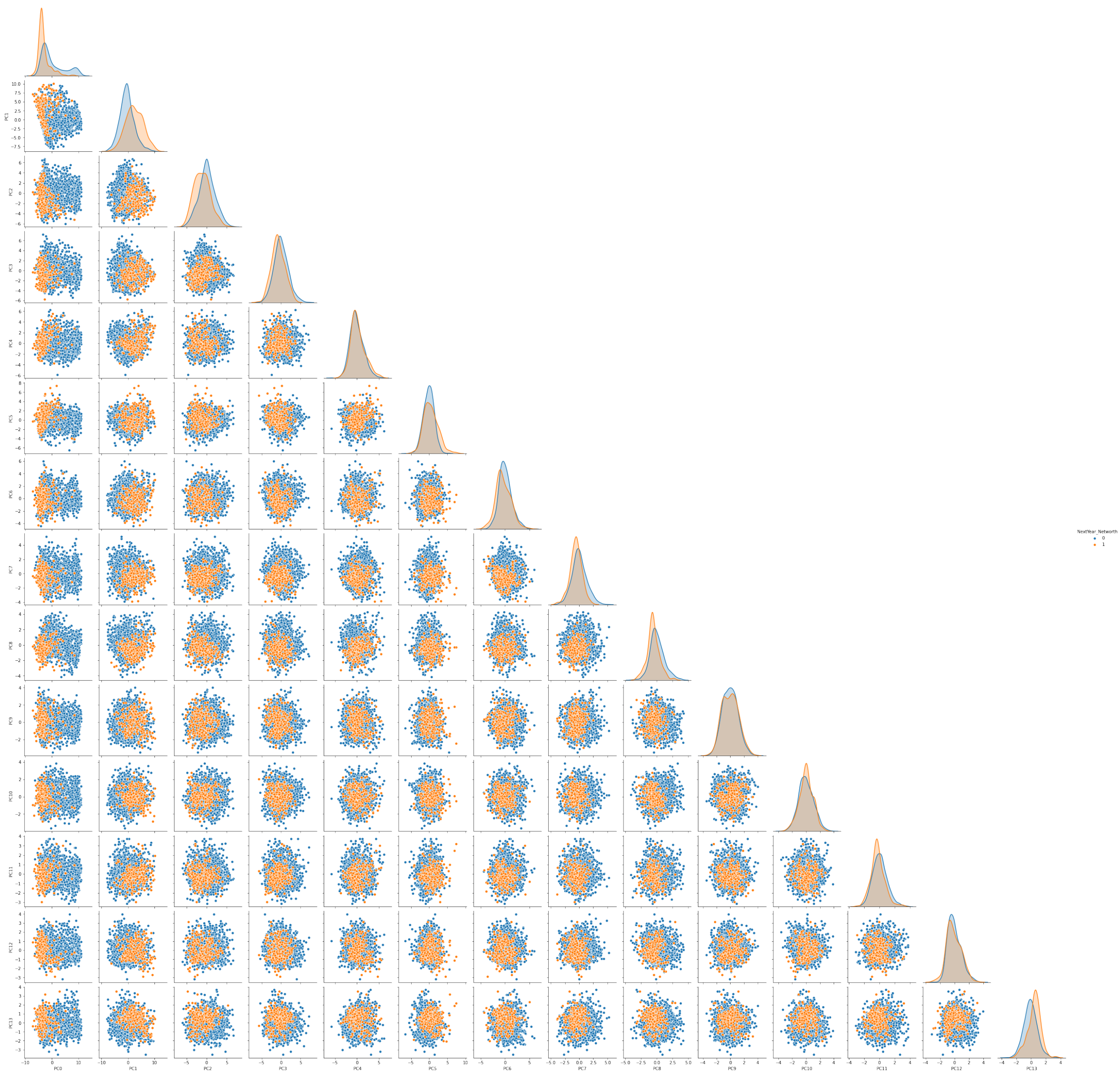
Correlation between the actual variables (Before PCA) is checked using Heatmaps.



A lot of variables show strong co-relations amongst each other. For e.g. variable representing “Profits” have strong relation with the variables representing Cash Flow Activities. The Total asset, liability and cost/sales variables also show strong relationship with one another.

Pairplot Plots:

Pairplot shows the relationship between the principal components in the form of scatter plots as also the the distribution of every component for Networth Next Year ‘0’ and ‘1’



The scatter plots appear cloudy for almost all the PCs, so we can see that they do not depict strong linear relation with one another.

1.5 Train Test Split

The dataset is split into Train and Test dataset in a ratio of 67:33 and Model is Built on Train Dataset and Validated on Test Dataset.

```
'Train Data (x):' (2402, 14)
'Train Data (y):' (2402,)
'Test Data (x):' (1184, 14)
'Test Data (y):' (1184,)
'Train Data after SMOTE (x):'(4314, 14)
'Train Data after SMOTE (y):'(4314,)
```

1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff

Logit Regression Results

Dep. Variable:	NextYear_Networth	**No. Observations:**	4314
Model:	Logit	**Df Residuals:**	4299
Method:	MLE	**Df Model:**	14
Date:	Mon, 08 Feb 2021	**Pseudo R-squ.:**	0.7282
Time:	13:44:38	**Log-Likelihood:**	-812.69
converged:	True	**LL-Null:**	-2990.2
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-5.1025	0.218	-23.369	0.000	-5.530	-4.675
PC0	-0.8534	0.045	-18.882	0.000	-0.942	-0.765
PC1	0.7522	0.034	21.916	0.000	0.685	0.819
PC2	-1.0687	0.053	-20.284	0.000	-1.172	-0.965
PC3	-0.4305	0.045	-9.500	0.000	-0.519	-0.342
PC4	0.7836	0.074	10.650	0.000	0.639	0.928
PC5	0.7256	0.054	13.368	0.000	0.619	0.832
PC6	-0.5352	0.053	-10.067	0.000	-0.639	-0.431
PC7	-1.3354	0.077	-17.452	0.000	-1.485	-1.185
PC8	-2.4797	0.140	-17.680	0.000	-2.755	-2.205
PC9	0.4911	0.071	6.933	0.000	0.352	0.630
PC10	-0.9999	0.098	-10.228	0.000	-1.191	-0.808
PC11	-1.0604	0.089	-11.928	0.000	-1.235	-0.886
PC12	-0.4190	0.088	-4.788	0.000	-0.591	-0.247
PC13	1.3110	0.100	13.074	0.000	1.114	1.508

From the above summary it is observed that all the 14 independent variables have p- values less than 0.05 so we can say they have significant relationship with networth next Year

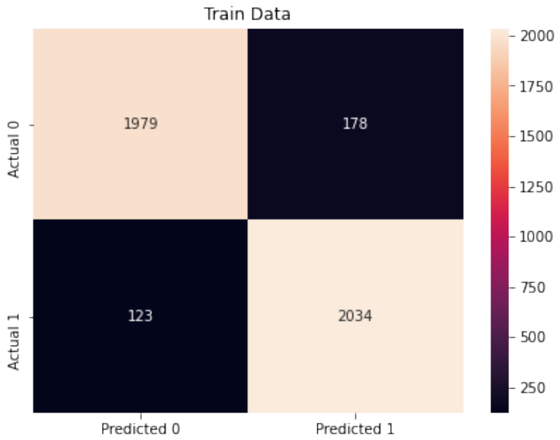
PC8, PC7, PC10, PC2, PC0 seem to have more impact on the networth next year variable than others.

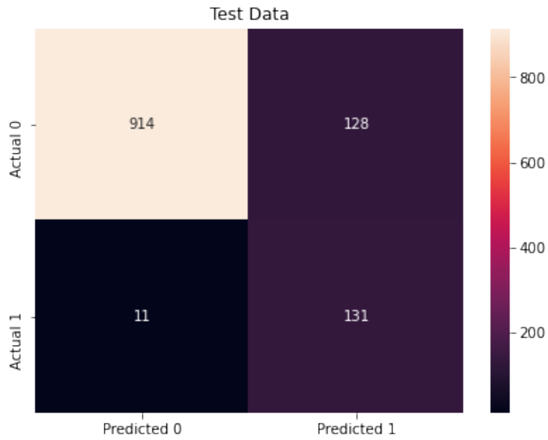
Having close look at these components we can say,

Book Value(NAV), Revenue in Forex due to expenses/earnings, Rate of growth due to Capital employed/Total Asset and Profits and Equity Paid up are influencing the networth of following year for the companies.

1.7 Validate the Model on Test Dataset and state the performance matrices

Please see below the confusion matrix and Classification report:





Train Data:

	precision	recall	f1-score	support
0	0.94	0.92	0.93	2157
1	0.92	0.94	0.93	2157
accuracy			0.93	4314
macro avg	0.93	0.93	0.93	4314
weighted avg	0.93	0.93	0.93	4314

Test Data:

	precision	recall	f1-score	support
0	0.99	0.88	0.93	1042
1	0.51	0.92	0.65	142
accuracy			0.88	1184
macro avg	0.75	0.90	0.79	1184
weighted avg	0.93	0.88	0.90	1184

The accuracy us 88% on test data as opposed to 93% on train data. The model slightly over fits the train data.

The precision for negative networth net year i.e. ‘1’ in dependant variable shows poor precision on test data. The overall recall seems okay at 92%.