

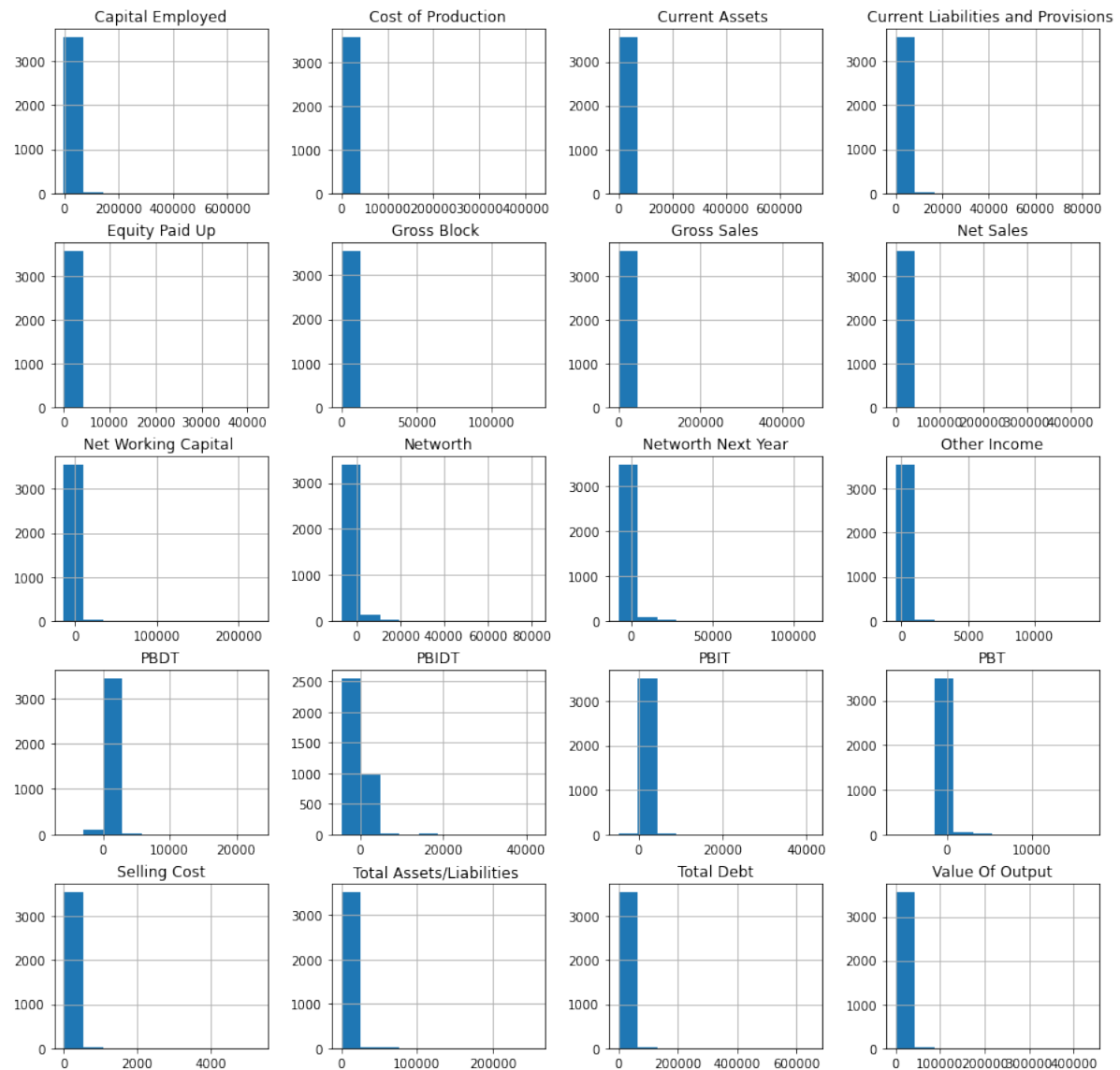
Table of Contents

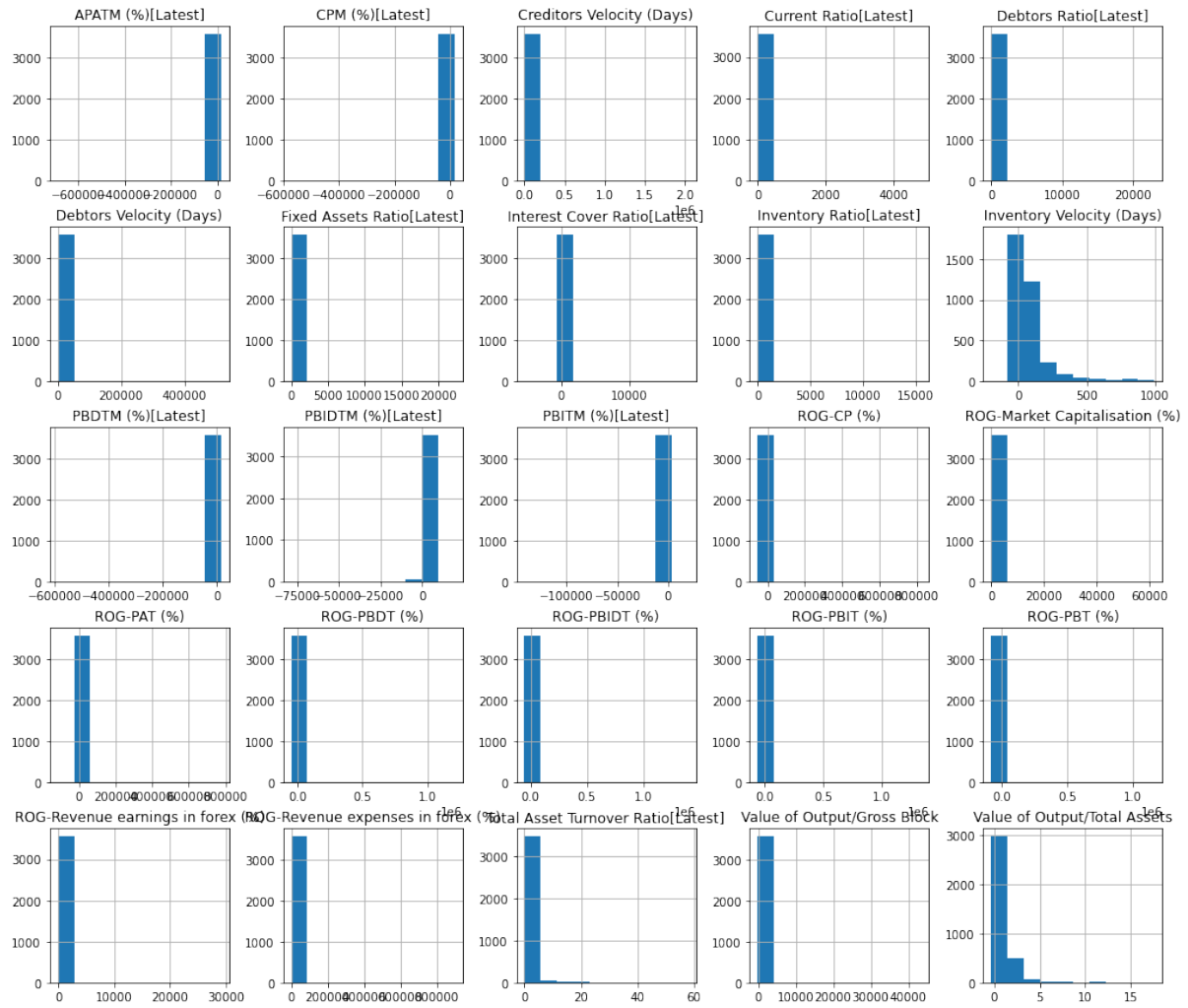
1.1 Outlier Treatment.....	4
1.2 Missing Value Treatment.....	7
1.3 Transform Target variable into 0 and 1	9
1.4 Univariate & Bivariate analysis with proper interpretation	9
1.5 Train Test Split.....	12
1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff.....	13
1.7 Validate the Model on Test Dataset and state the performance matrices	13

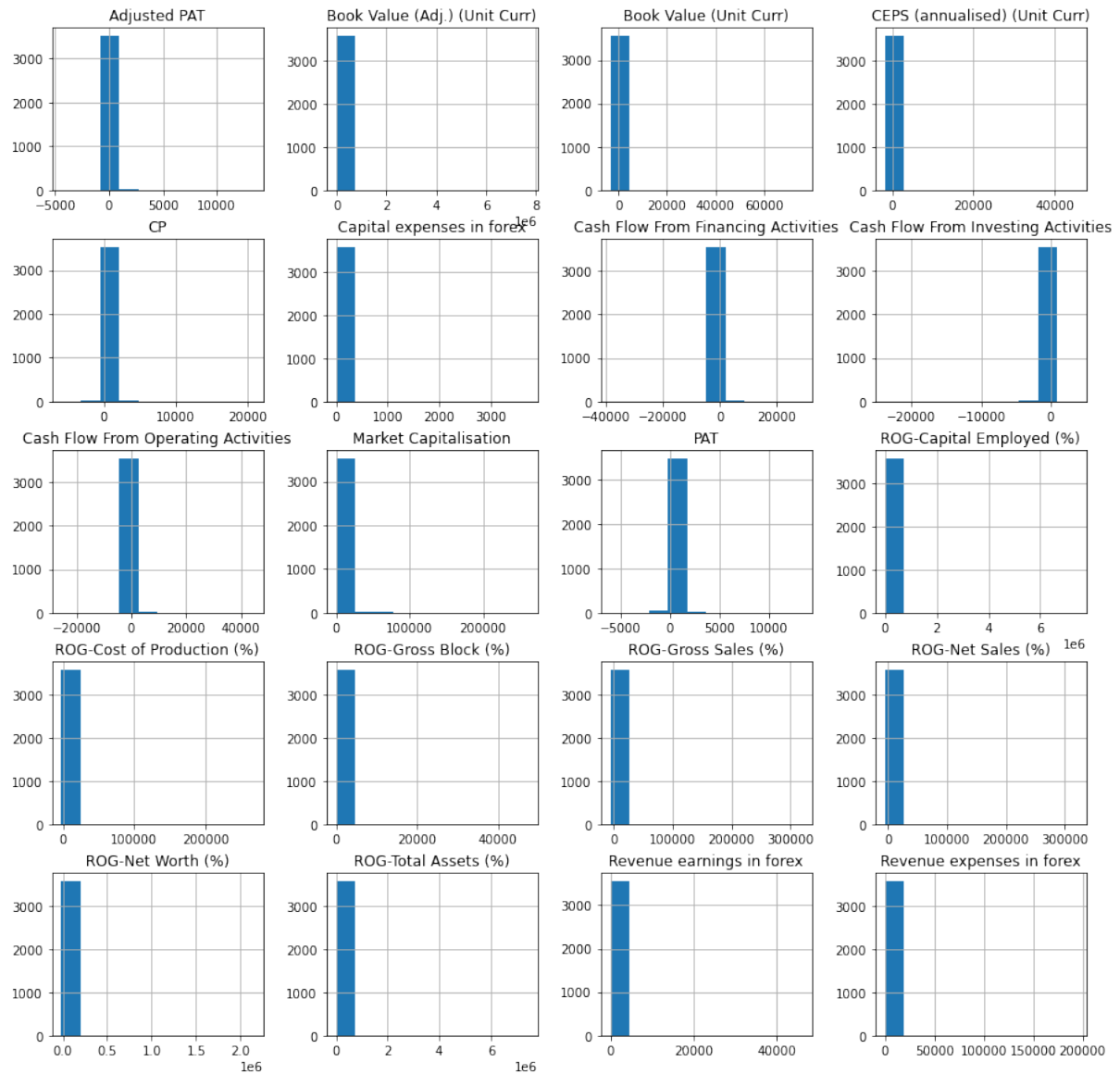
Dataset that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Net worth of the company in the following year (2016) is provided which can be used to drive the labelled field.

Dataset has 3587 entries of different companies across 67 columns.

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



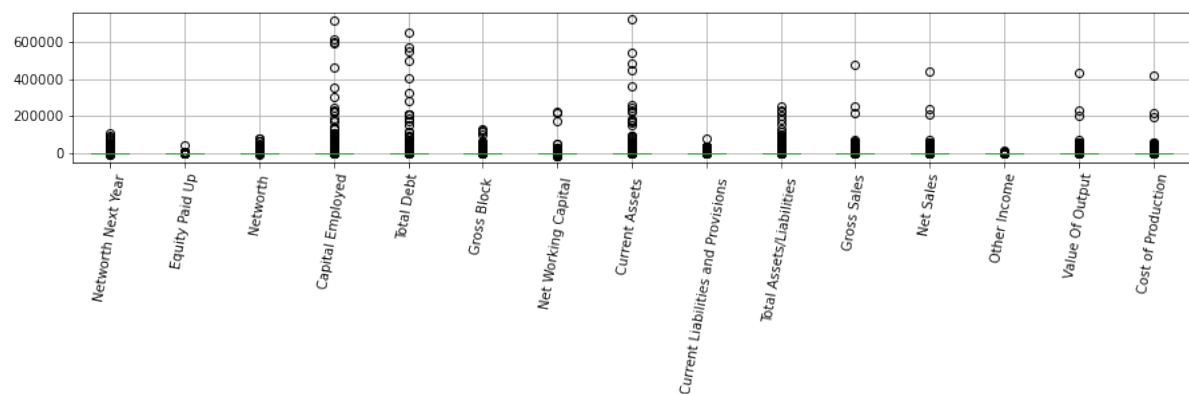


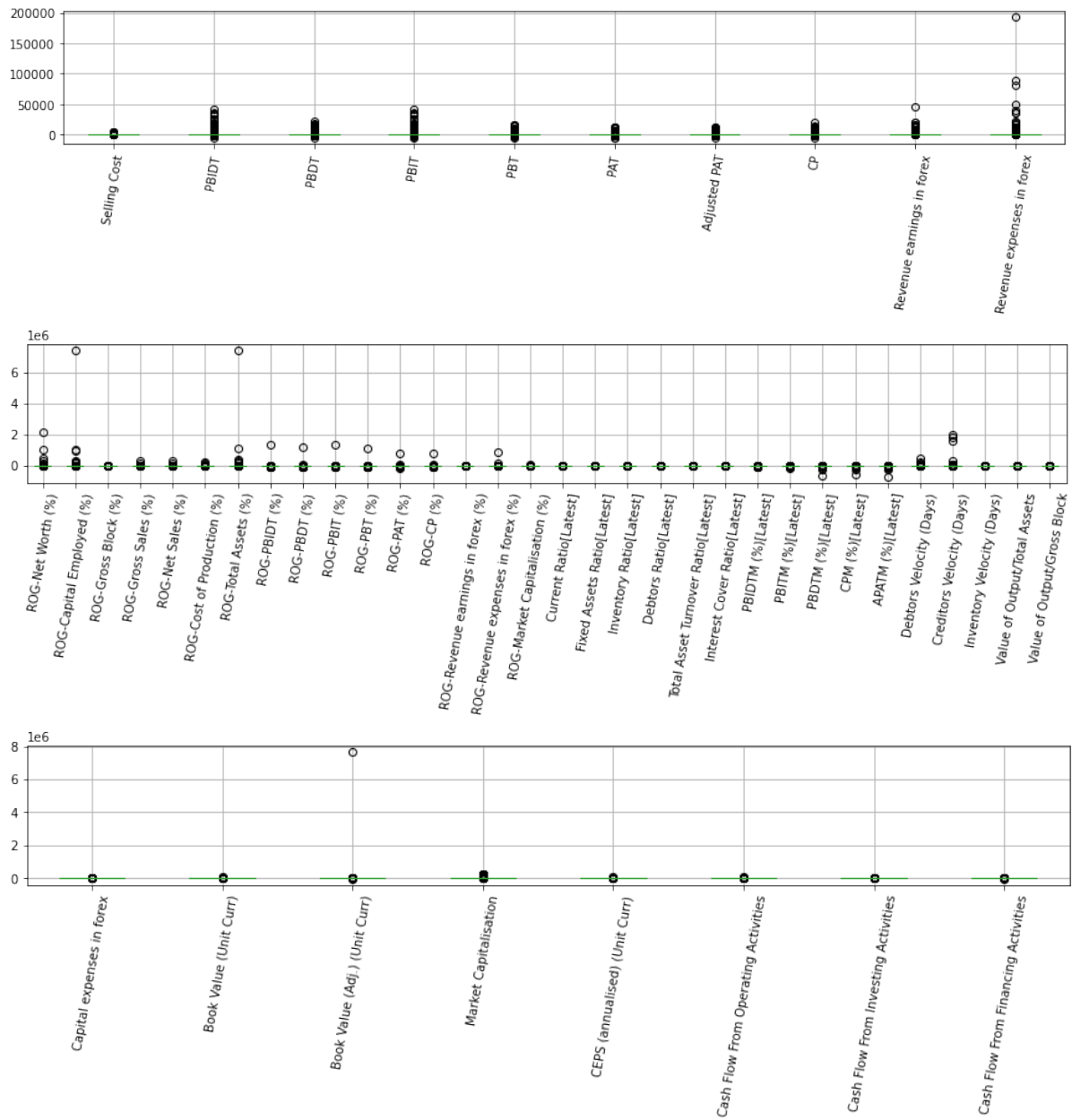


The distribution for almost all the variables looks extremely skewed to the left (i.e. in the first 25% in the quantile). Few Variables are right skewed.

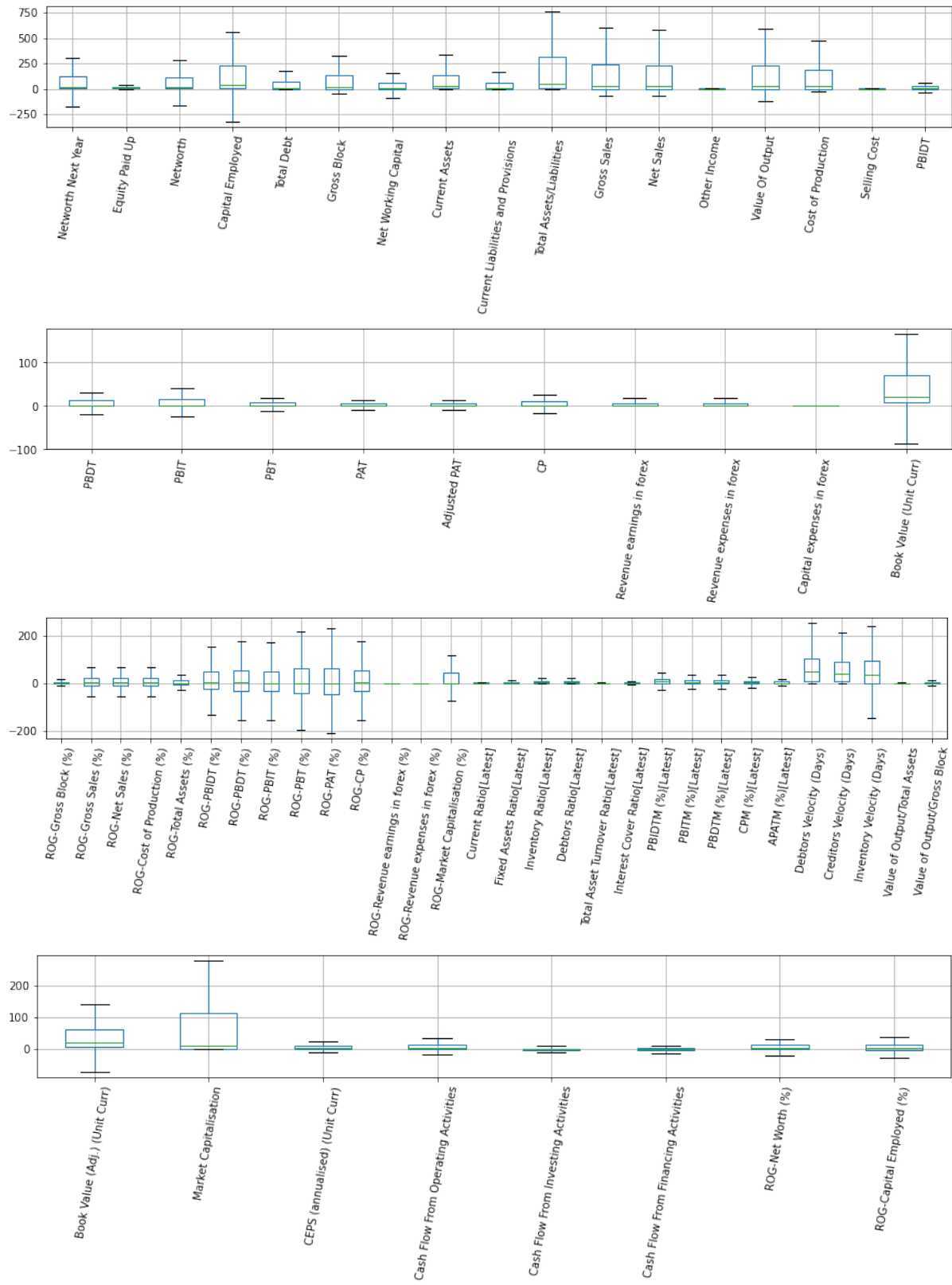
1.1 Outlier Treatment

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.



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.

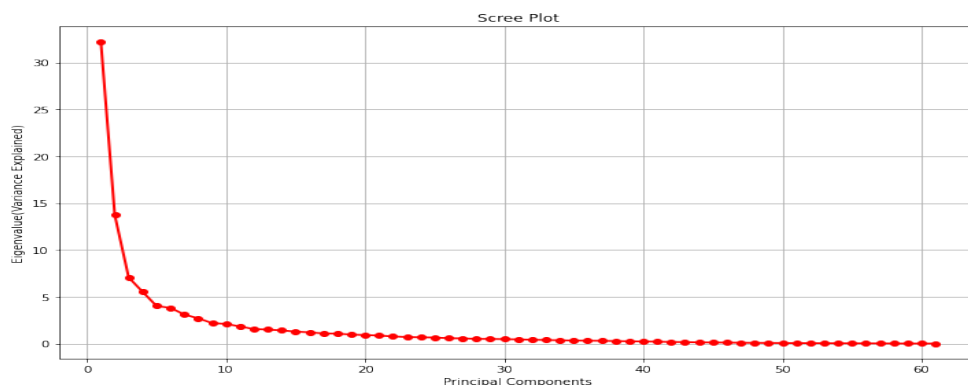
Scaling:

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:



PC0	0.087	0.059	0.084	0.029	0.03	0.023	0.058	0.031	0.036	0.038	0.042	0.04	0.033	0.088	-0.049	0.017	0.054	0.056	0.054	0.099	0.063	0.067	0.065	0.064	0.066	-0.028	-0.019	0.017	0.067	0.024
PC1	-0.21	-0.16	-0.04	-0.16	-0.16	-0.12	-0.15	-0.23	-0.24	-0.23	-0.23	-0.23	-0.23	-0.033	-0.11	-0.11	-0.028	-0.018	-0.038	-0.13	-0.17	-0.18	-0.2	-0.19	-0.2	0.07	0.082	0.061	-0.062	-0.12
PC2	-0.028	-0.03	-0.012	-0.096	-0.096	-0.056	-0.039	-0.26	-0.25	-0.26	-0.24	-0.24	-0.25	-0.008	0.11	-0.061	-0.064	-0.022	-0.14	0.12	0.29	0.29	0.3	0.3	0.29	-0.033	-0.06	-0.031	-0.13	-0.057
PC3	0.097	0.14	0.15	0.14	0.14	0.17	0.15	-0.19	-0.2	-0.19	-0.2	-0.19	-0.19	0.018	-0.06	0.25	0.17	0.18	0.35	0.084	-0.12	-0.096	-0.095	-0.1	-0.073	-0.12	-0.12	-0.035	0.37	0.25
PC4	0.0011	0.19	0.089	0.3	0.3	0.27	0.23	0.014	-0.029	-0.014	-0.056	-0.067	-0.036	-0.067	0.02	0.097	-0.073	-0.065	-0.14	-0.057	0.17	0.15	0.091	0.091	0.077	0.046	0.077	0.006	-0.12	0.12
PC5	-0.14	-0.23	-0.091	-0.055	-0.056	-0.087	-0.22	0.055	0.063	0.061	0.067	0.062	0.055	0.019	-0.11	0.12	0.24	0.23	0.29	0.17	0.19	0.19	0.2	0.18	-0.028	0.033	0.041	0.24	0.076	
PC6	0.12	0.21	0.019	-0.38	-0.38	-0.32	0.15	0.024	0.018	0.03	0.018	0.026	0.028	-0.072	0.086	0.32	0.0003	-0.087	0.078	0.019	0.028	0.057	-0.0039	-0.0047	0.034	0.28	0.2	0.095	0.073	0.32
PC7	-0.14	-0.3	-0.23	0.16	0.17	0.13	-0.27	0.000830	0.00032	-0.003	0.007	-0.012	-0.016	-0.15	0.25	0.38	-0.16	-0.12	-0.05	-0.06	-0.056	-0.021	-0.053	-0.067	-0.013	0.022	-0.063	-0.023	-0.08	0.36
PC8	-0.074	0.018	0.12	0.072	0.073	0.049	0.0013	0.079	0.042	0.051	0.011	0.015	0.052	0.027	0.017	-0.13	-0.053	0.049	0.0061	0.045	0.029	0.008	0.0022	0.0094	-0.024	0.079	0.0045	0.34	0.0042	-0.16
PC9	0.037	-0.06	0.011	0.16	0.16	0.12	-0.027	-0.043	-0.044	-0.05	-0.037	-0.044	-0.049	0.24	-0.058	0.0094	0.012	-0.3	0.074	0.13	-0.026	-0.0094	0.011	0.014	0.021	0.55	0.45	0.2	0.034	-0.02
PC10	-0.022	-0.053	0.086	0.025	0.024	0.06	-0.026	0.02	-0.0018	0.0014	-0.0086	0.00058	0.011	-0.23	-0.31	-0.023	0.35	0.023	-0.004	-0.035	0.04	0.0073	-0.0099	0.0059	-0.027	0.17	0.36	-0.48	-0.037	-0.0045
PC11	-0.016	0.049	-0.3	0.0078	0.0078	0.035	0.0051	-0.019	-0.019	-0.002	0.014	0.019	-0.019	0.54	0.24	-0.018	0.38	0.1	-0.047	0.14	-0.063	-0.059	-0.041	-0.043	-0.025	0.076	-0.075	-0.3	-0.057	-0.057
PC12	-0.024	-0.0054	0.0037	0.042	0.041	-0.0059	-0.0081	0.0035	-0.014	-0.008	-0.032	-0.024	-0.011	0.093	-0.085	0.058	0.044	0.55	-0.072	0.076	-0.013	-0.017	-0.041	-0.045	-0.047	-0.11	0.26	0.54	-0.074	0.029
PC13	-0.12	-0.079	-0.16	0.036	0.037	0.02	-0.12	-0.028	-0.0097	-0.0051	0.014	0.013	-0.015	0.19	-0.48	-0.072	-0.24	-0.09	0.14	-0.077	0.056	0.072	0.025	0.025	0.054	-0.17	0.17	-0.073	0.094	-0.024
PC14	-0.18	-0.18	0.31	-0.049	-0.05	-0.065	-0.13	0.043	0.041	0.024	0.027	0.013	0.037	0.46	-0.19	0.23	-0.19	-0.11	-0.085	-0.1	0.031	0.034	-0.017	-0.011	0.0052	-0.19	0.00076	-0.069	-0.09	0.27
	ROG-Net Worth (%)	ROG-Capital Employed (%)	ROG-Gross Block (%)	ROG-Gross Sales (%)	ROG-Net Sales (%)	ROG-Cost of Production (%)	ROG-Total Assets (%)	ROG-PBDT (%)	ROG-PBDT (%)	ROG-PBIT (%)	ROG-PBT (%)	ROG-PAT (%)	ROG-CP (%)	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)	PBDTM (%) (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

PC0	0.12	0.19	0.19	0.15	0.18	0.15	0.19	0.18	0.19	0.2	0.2	0.17	0.2	0.19	0.17	0.21	0.19	0.2	0.18	0.18	0.17	0.19	0.14	0.16	0.14	0.13	0.17	0.15	0.14	-0.12	-0.083
PC1	0.12	0.043	0.11	0.15	0.14	0.069	0.12	0.13	0.12	0.094	0.094	0.093	0.092	0.11	0.082	-0.006	-0.077	-0.042	-0.12	-0.12	-0.11	-0.075	0.07	0.081	-0.042	-0.032	0.053	-0.13	0.045	-0.0099	-0.066
PC2	-0.04	0.055	-0.004	-0.077	-0.043	-0.01	-0.044	-0.063	-0.021	-0.062	-0.062	-0.029	-0.061	-0.079	-0.04	0.013	0.046	0.027	0.061	0.059	0.074	0.044	-0.024	-0.041	0.11	0.11	0.022	0.048	0.016	-0.0081	-0.0099
PC3	-0.14	-0.056	-0.088	-0.076	-0.072	-0.025	-0.054	-0.062	-0.088	0.034	0.034	-0.078	0.035	0.046	0.032	-0.03	2.5e-05	-0.015	0.014	0.012	0.026	-0.0035	0.029	0.012	0.071	0.045	-0.063	0.085	-0.024	-0.074	0.082
PC4	0.2	0.046	0.13	0.18	0.068	0.086	0.13	0.13	0.14	0.054	0.056	0.046	0.056	0.048	-0.0054	-0.057	-0.17	-0.099	-0.19	-0.18	-0.17	-0.16	-0.056	-0.018	-0.13	-0.13	0.055	-0.21	-0.14	-0.037	0.22
PC5	0.017	-0.11	-0.024	0.045	0.08	-0.046	0.0079	0.061	0.0014	0.061	0.06	0.045	0.057	0.085	0.068	-0.086	-0.14	-0.12	-0.16	-0.16	-0.16	-0.13	0.089	0.099	-0.14	-0.14	-0.029	-0.11	0.11	0.11	-0.25
PC6	0.017	0.045	0.055	0.037	-0.074	0.16	0.087	0.023	0.049	0.0011	0.0029	0.017	0.0039	-0.0011	-0.051	-0.022	-0.05	-0.015	-0.021	-0.02	-0.026	-0.05	-0.041	-0.019	0.07	0.061	0.0048	0.002	-0.19	0.047	0.19
PC7	0.024	0.11	0.06	-0.053	-0.089	0.13	0.063	0.0059	0.048	0.024	0.025	0.019	0.025	0.015	-0.074	0.016	0.016	0.041	0.041	0.039	0.041	0.012	-0.11	-0.076	0.16	0.16	0.013	0.014	0.0013	0.27	-0.25
PC8	-0.24	0.033	0.019	0.022	0.028	0.12	0.016	-0.03	-0.00071	-0.028	-0.03	0.027	-0.031	-0.0048	0.039	-0.093	-0.15	-0.12	-0.18	-0.18	-0.18	-0.14	0.15	0.12	0.48	0.49	-0.078	0.18	-0.034	0.016	0.0014
PC9	-0.075	-0.056	-0.084	-0.072	-0.06	0.061	-0.011	-0.01	-0.068	-0.064	-0.065	-0.064	-0.066	-0.058	0.017	0.0079	0.058	0.03	0.08	0.071	0.083	0.047	0.17	0.11	-0.12	-0.12	0.028	0.0033	0.13	-0.02	-0.14
PC10	0.053	0.085	0.065	0.03	0.025	-0.18	-0.017	0.065	0.076	-0.042	-0.04	0.051	-0.039	-0.051	-0.13	0.0048	-0.021	-0.0044	-0.024	-0.02	-0.017	-0.014	-0.23	-0.2	0.22	0.2	-0.047	0.13	0.1	-0.074	-0.046
PC11	0.13	0.12	0.06	-0.081	-0.0079	0.029	-0.016	-0.052	0.034	-0.056	-0.055	0.096	-0.054	-0.058	-0.033	-0.053	-0.044	-0.065	-0.029	-0.027	-0.022	-0.045	0.11	0.086	0.067	0.098	-0.25	-0.069	-0.18	0.14	0.14
PC12	0.18	0.12	0.056	-0.037	0.0035	-0.11	-0.032	0.015	0.052	-0.11	-0.11	0.084	-0.11	-0.14	-0.083	0.047	0.055	0.054	0.067	0.07	0.059	0.057	-0.23	-0.2	-0.026	-0.015	0.13	-0.031	-0.022	0.061	-0.0034
PC13	-0.12	-0.081	-0.025	0.11	-0.034	0.057	0.049	0.044	-0.009	0.035	0.035	0.031	0.036	0.022	-0.0097	0.035	0.06	0.063	0.076	0.07	0.057	0.053	-0.026	0.0028	0.036	0.019	-0.031	0.085	-0.46	0.33	0.26
PC14	0.048	0.027	0.0021	0.02	-0.046	-0.16	-0.066	-0.017	0.0017	-0.049	-0.049	-0.13	-0.049	-0.07	0.032	-0.043	-0.052	-0.044	-0.044	-0.046	-0.064	-0.055	0.054	0.027	0.063	0.065	0.19	-0.026	0.15	-0.41	0.072
	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	PBDT	PBDT	PBT	PBT	PAT	Adjusted PAT	CP	Revenue earnings in forex	Revenue expenses in forex	Book Value (Unit Curr)	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

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.

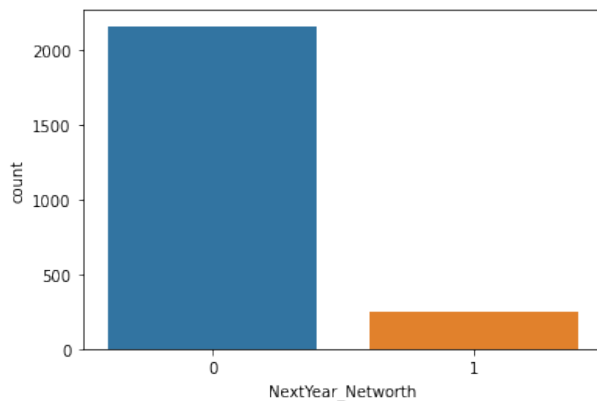
Showing the unique count in the 'Networth_NextYear' Variable:

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

1.4 Univariate & Bivariate analysis with proper interpretation

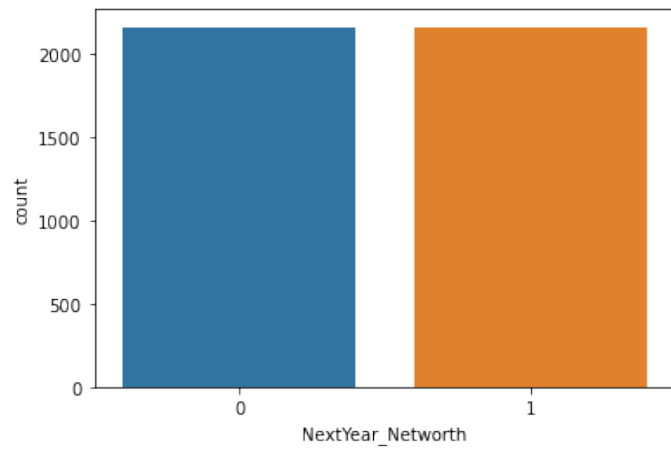
Univariate:

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:

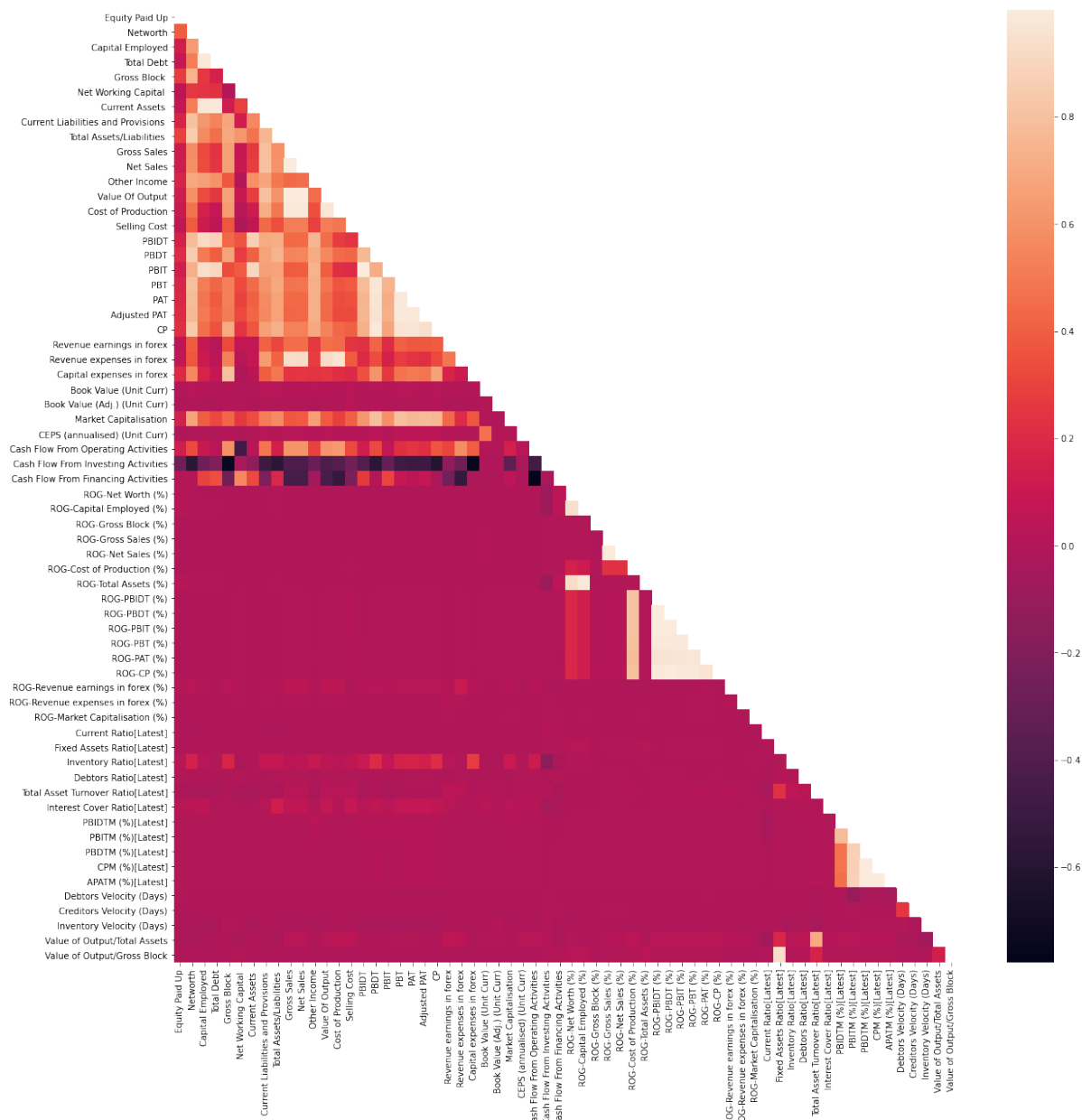


1	0.5
0	0.5

Bivariate:

Heatmaps:

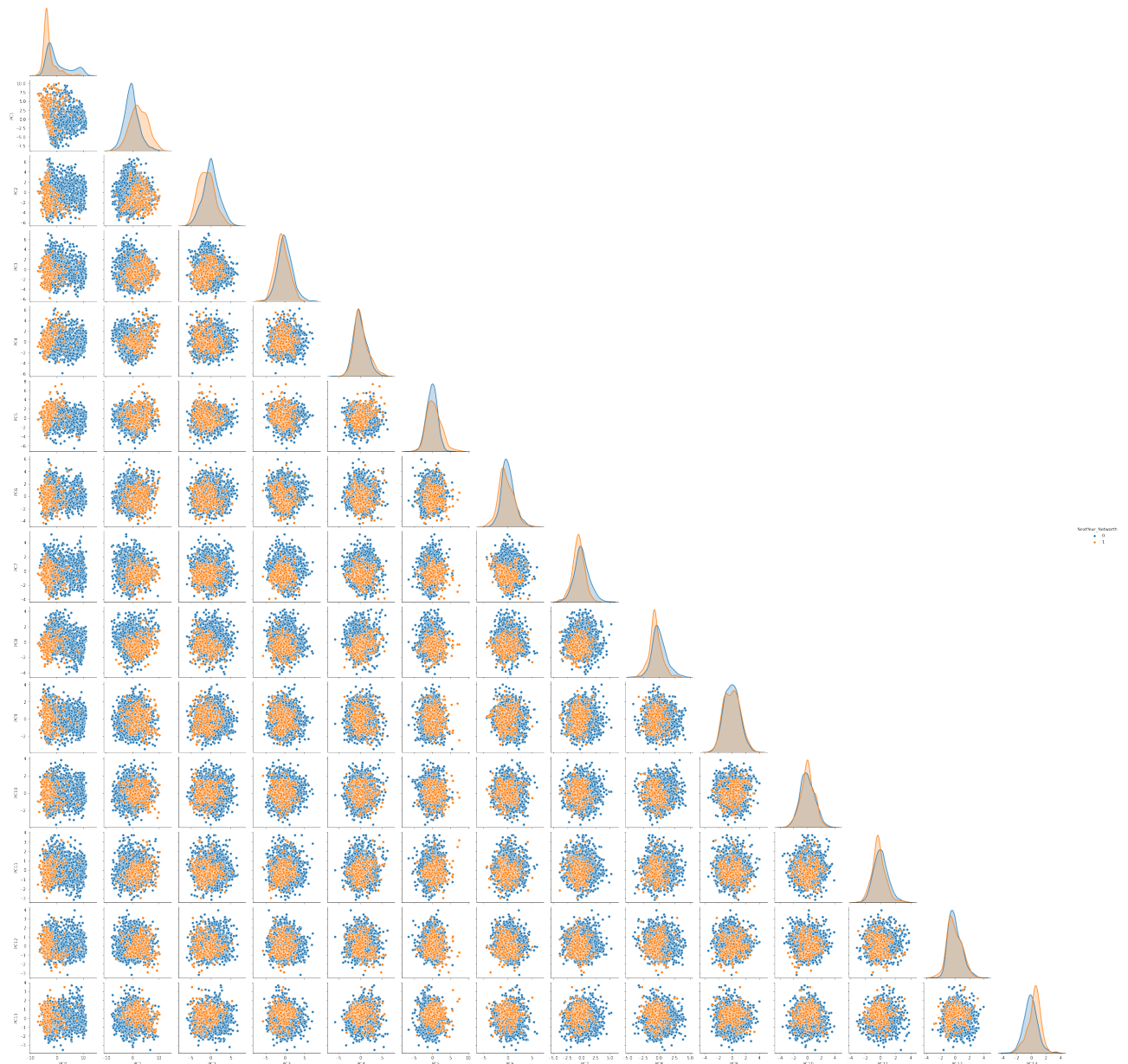
Correlation between the actual variables (Before PCA) is checked using Heatmaps.



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 Network 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 network next Year

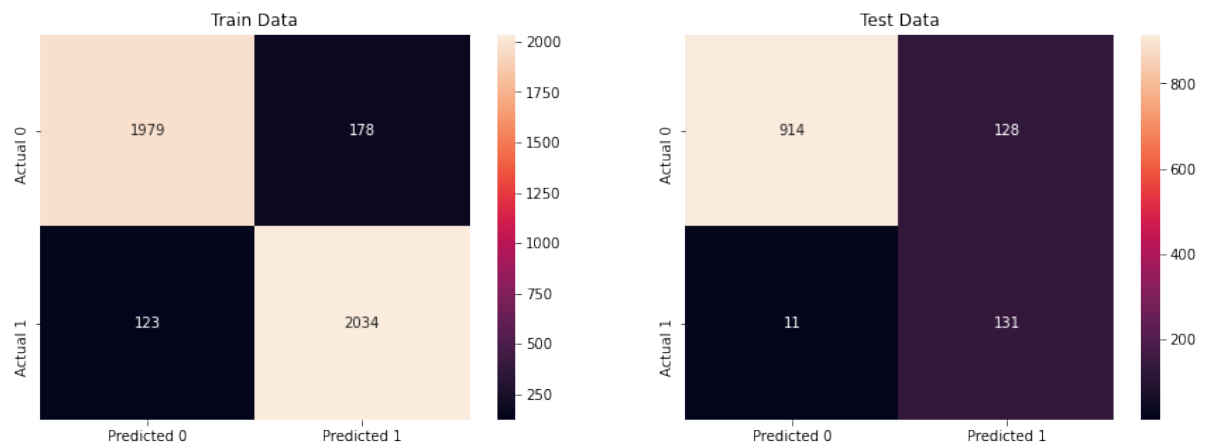
PC8, PC7, PC10, PC2, PC0 seem to have more impact on the network next year variable than others.

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 network 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 is 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%.