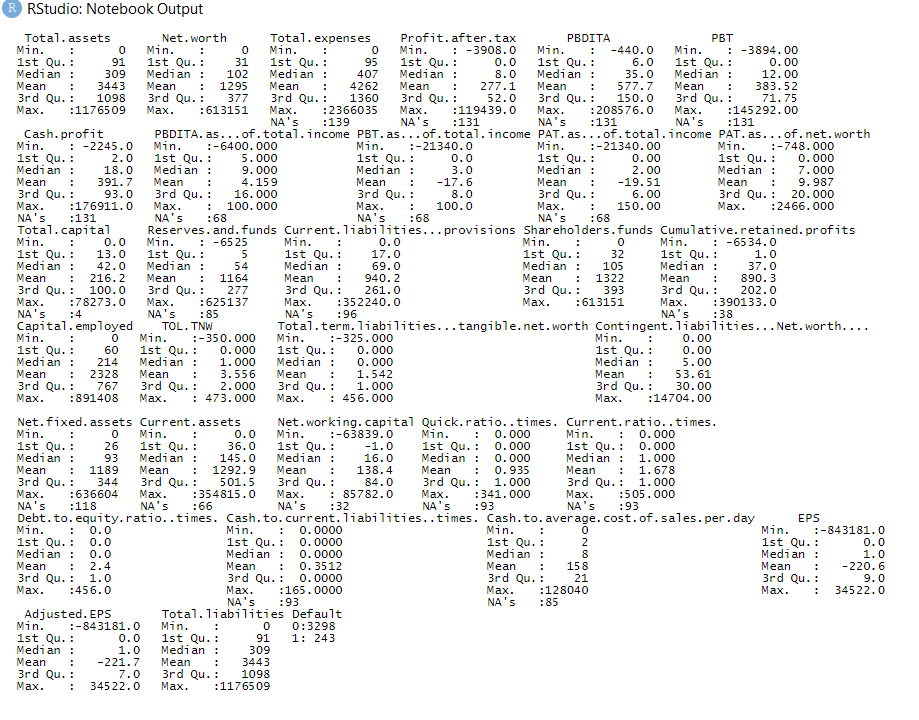
FRA Project

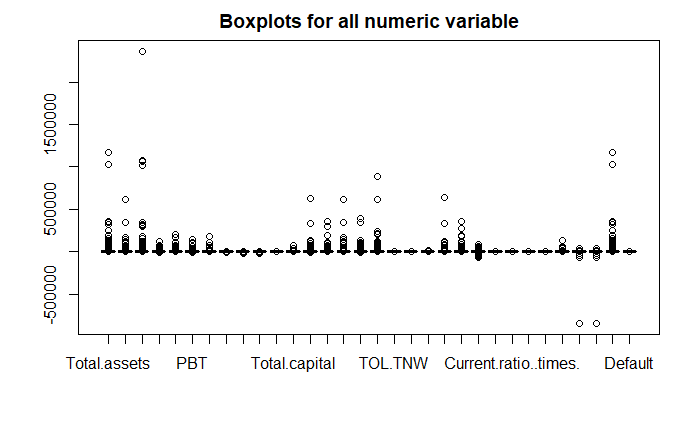
Objective: The data in hand is of a financial statement for various companies, the agenda of the project is to predict if a particular company is going to default given the various attributes in the data for which a logistic regression model will be used.

EDA: The data has 3541 observations and 51 variables and the total default rate in the data is nearly 7%, will be using the variable Net worth next year as dependent variable by converting into a factor named “Default”. The data also a lot of missing values. As a goof practise, have gone ahead and dropped variables that have missing values at least 5% to that of the total observations as imputing them may skew the model.

*5-point summary for remaining 32 variables:*

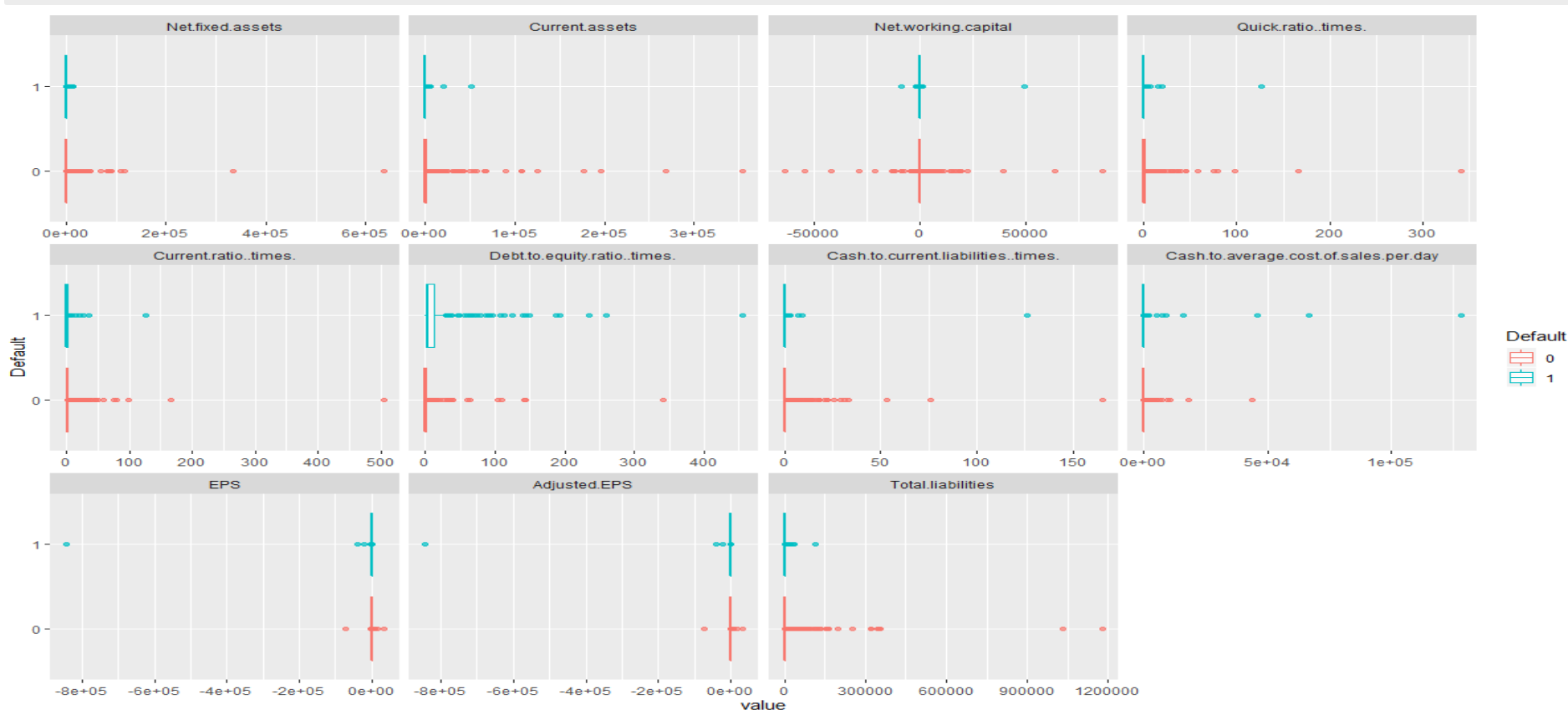
**

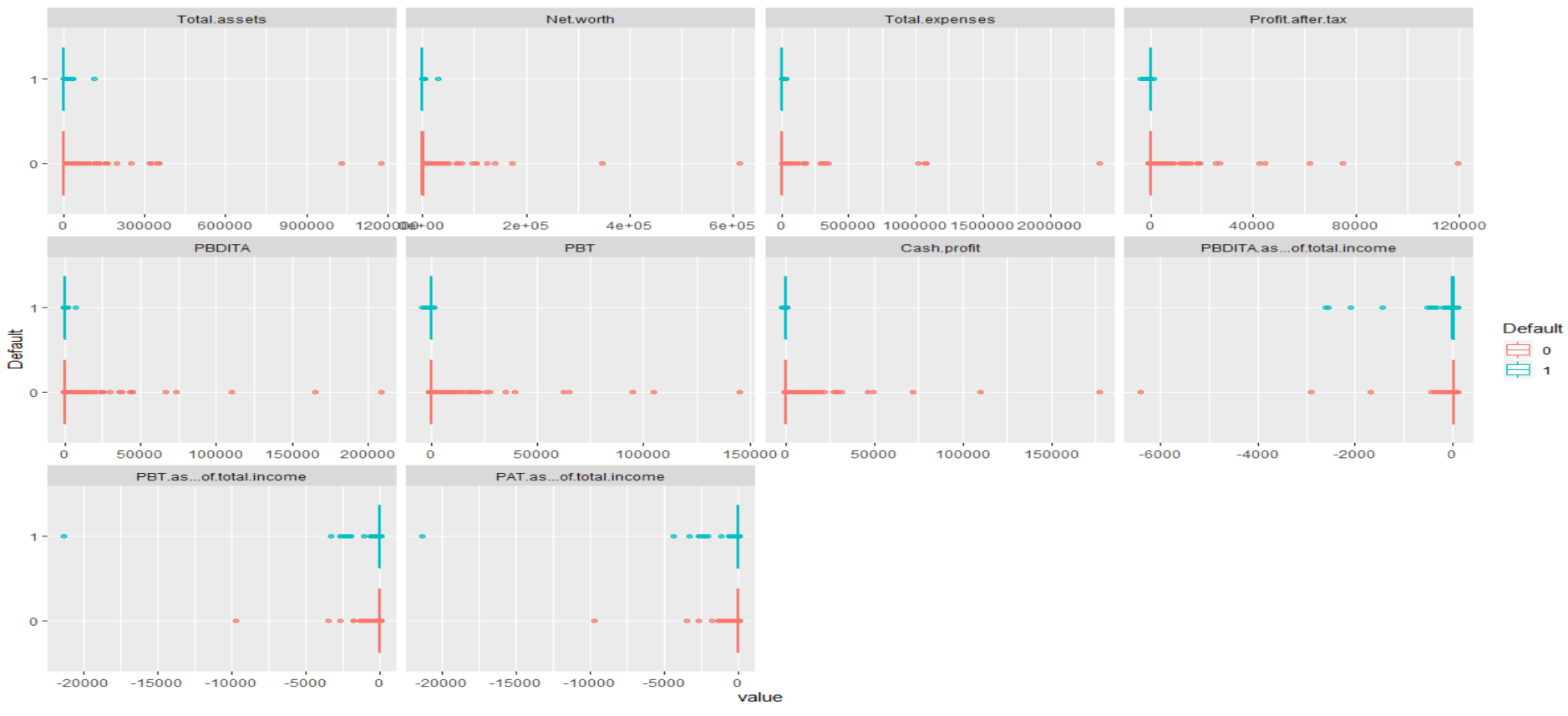
*Univariate analysis (Box plot):*

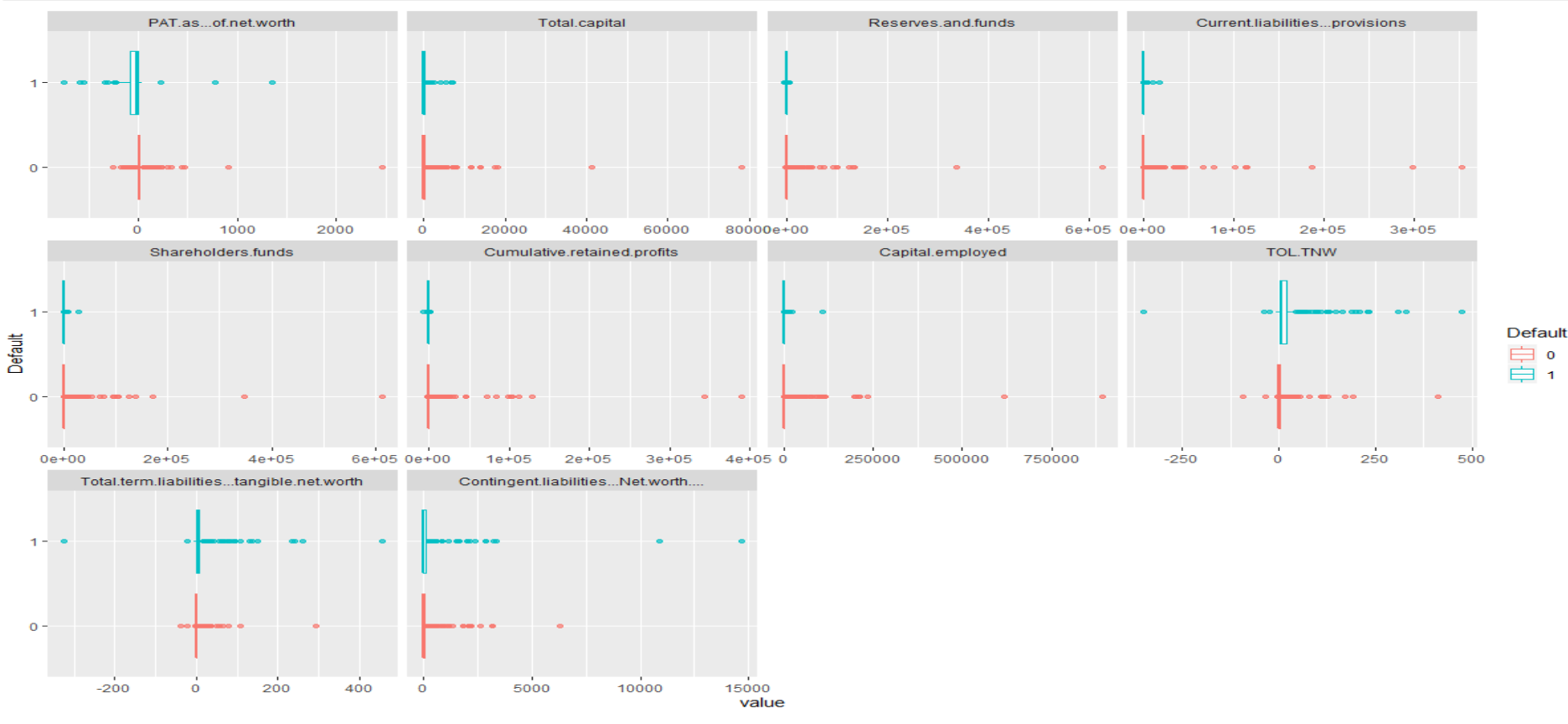
**

**Observation:** There are outliers in most of the variable, upon taking a closer look at the observations these appear to be valid data points hence will not be treating outlier and will be using the data as it is and the IQR for all variables are also very narrow.

*Bivariate Analysis with default as a factor (Box plot):*

**

**



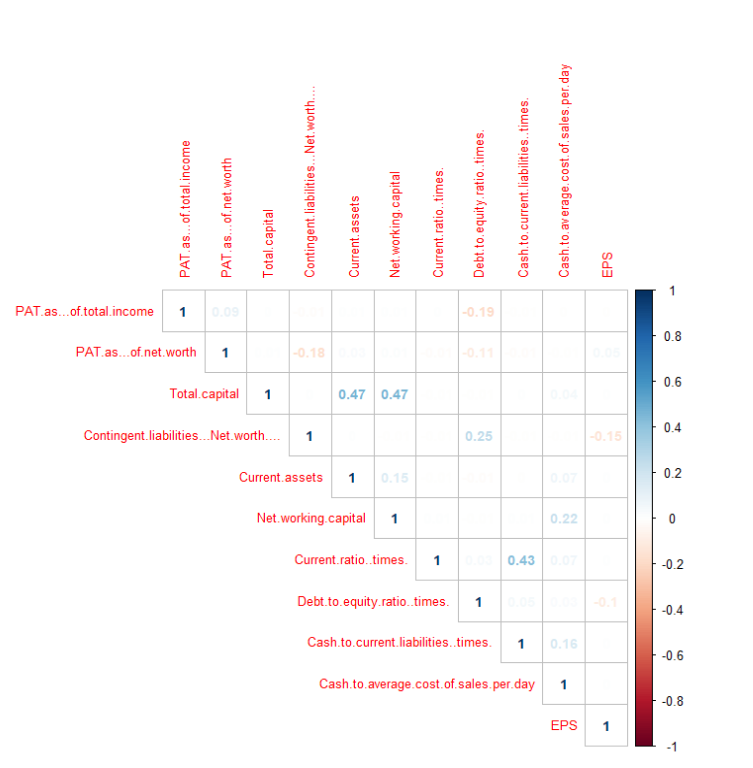
**Observation:** As seen earlier the IQR is narrowed around zero.

Data Preparation: Will be Creating 3 new variable in terms of ratios, plus we have current ratio times and Debt to equity ratio already in the data

* Asset Turnover: Sales/Total Assets (Efficiency)
* Return on asset: Total Income/Total Asset (Profitability)
* Return on Equity: Total Income/share holder funds (Profitability)
* Current ratio times: Current Asset/Current Liability (Liquidity)
* Debt to equity: Total Liability/Total Asset (Leverage)

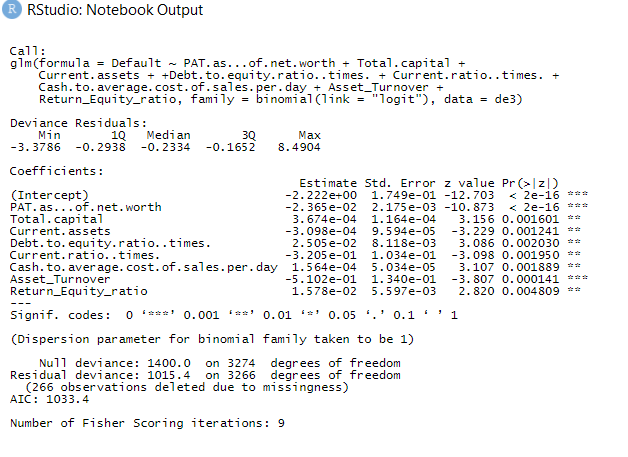
Missing Value Treatment: Have gone ahead and imputed missing values by using Predictive mean matching algorithm.

Multicollinearity Treatment:



**Observation:** Multicollinearity has been eliminated successfully by dropping variable that are highly correlated, we are left with 12 variables excluding the additional 3 variable that have been created.

Logistic Regression Model: After running the model on all 15 variables, based on p-value we get 8 significant variables.



**Interpretation:**

* The intercept value is -2.222
* A unit increase in PAT as % of net worth decrease the log odds ratio of default decrease by -0.0236.
* A unit increase in the Total Capital increases the log odds ratio of default by

0.000367.

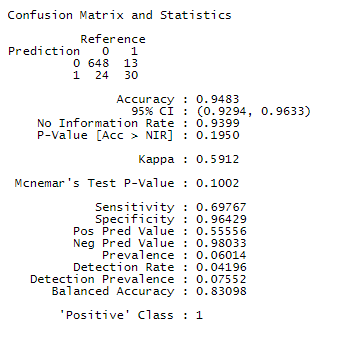
* A unit increase in Current assets decreases the log odds ratio of default by

-0.0003098.

* A unit increase in Debt to equity increase the log odds ratio of default by 0.02505.
* A unit increase in Current ratio times decreases the log odds ratio of default by -0.3205.
* A unit increase in cash to average sales per day increases the log odds ratio of default by 0.000564.
* A unit increase in asset turnover ratio decreases the log odds ratio of default by -0.5102.
* A unit increase return equity ratio increases the log odds ratio of default by 0.01578.

Validation on test data:

After having prepared the test data by imputing missing values and adding the 3 derived variables into the test data selecting the relevant variables for predicting and by finding the right threshold, we get a confusion matrix as below.



*Interpretation:* The model is able to capture the defaults and the non-defaults accurately with nearly 95% accuracy, the sensitivity or true positive rate is nearly 70% and the true negative rate is nearly 96%. Opportunity loss is nearly 4% and loss rate is nearly 2%.

* Data Validation by sorting predicted probability of score in descending order and dividing into deciles of 10 to check how much is default percentage in each decile shows 80% default in each decile.

