Financial Risk analysis project

Table of Contents

Problem Statement ............................................................................................................................ 4

Questions:

1. Outlier Treatment ......................................................................................................................... 5

1.2 Missing Value Treatment .......................................................................................................... 11

1.3 Transform Target variable into 0 and 1...................................................................................... 12

1.4 Univariate & Bivariate analysis with proper interpretation. (You may choose to include only

those variables which were significant in the model building). ..................................................... ….13

1.5 Train-Test Split. .......................................................................................................................... 19

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 ............ ……19

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

interpretation from the model. ........................................................................................................... 33

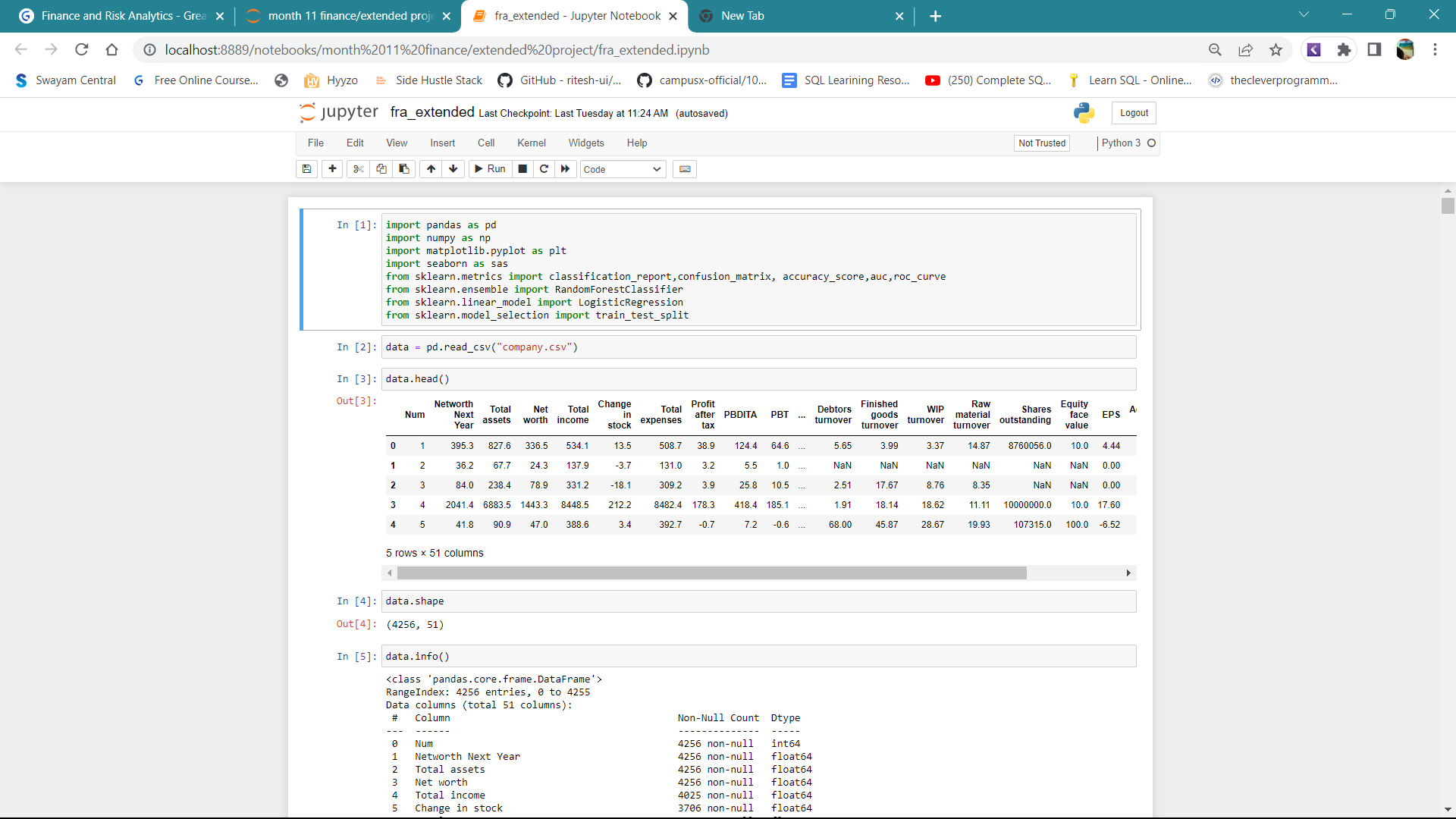
**Business Problem**

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.

* The first 5 row of the data set. Are as follow. It has 67 columns.



* No duplicate records
* Data type are as follows

|  |  |
| --- | --- |
| Data type | No of variable |
| object | 1 |
| float | 63 |
| int | 3 |
|  |  |

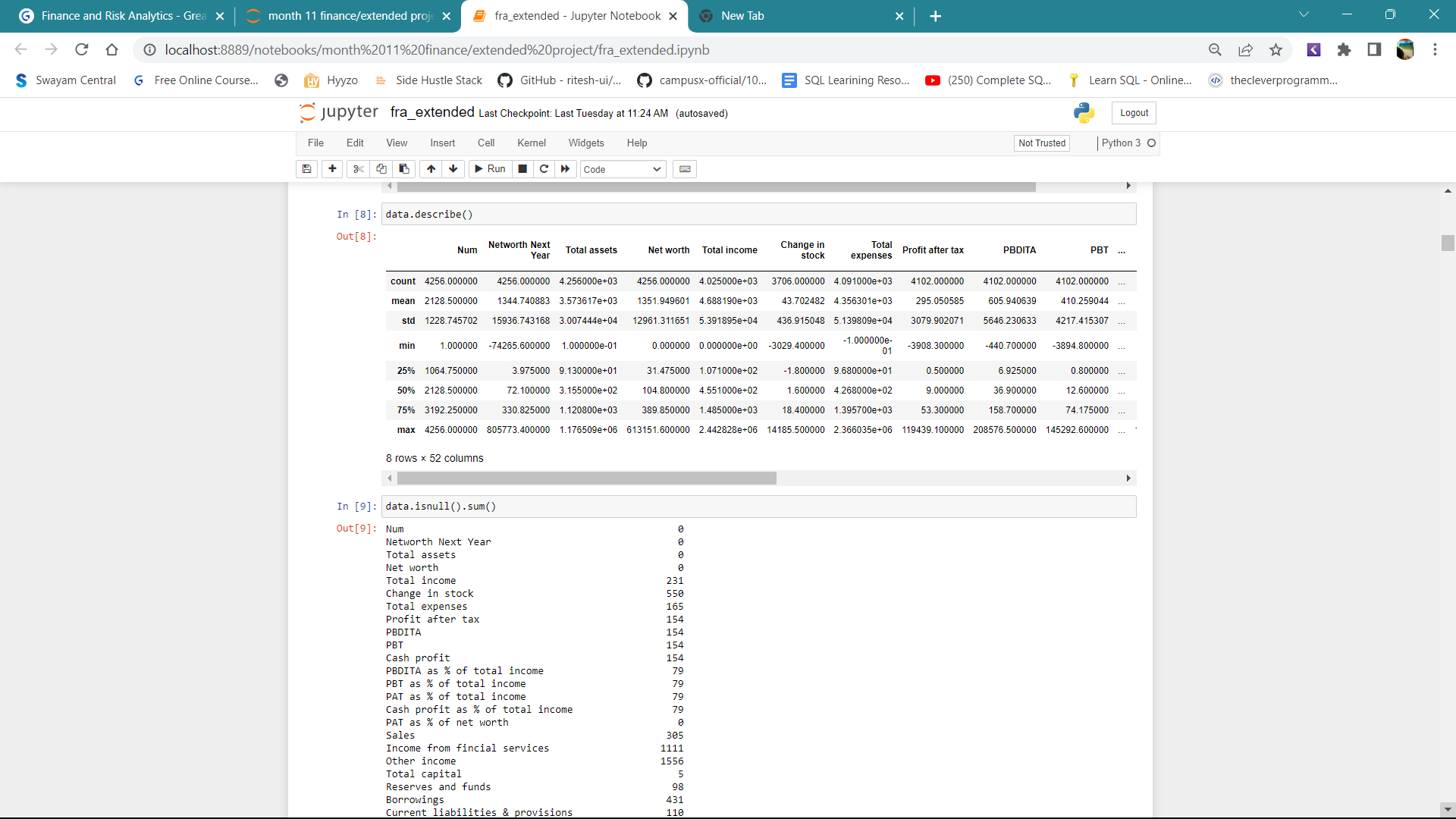
Co\_Code is treated as an integer because it has numeric values. It is a Categorical

variable and is therefore, converted to type ‘object’. However, this is not used in any

model building or EDA, and is therefore dropped.

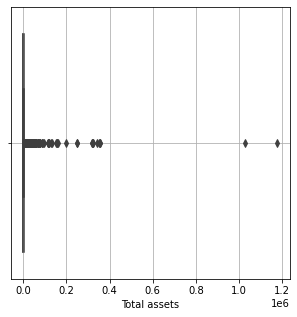
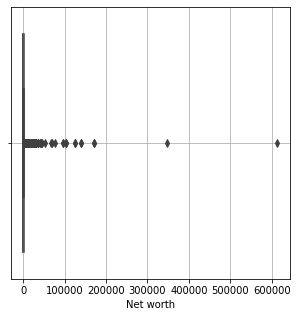
Co\_Name is the object variable and is dropped. We are then left with 65 variables.

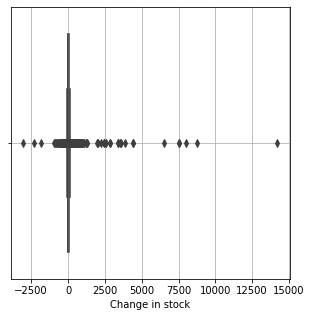
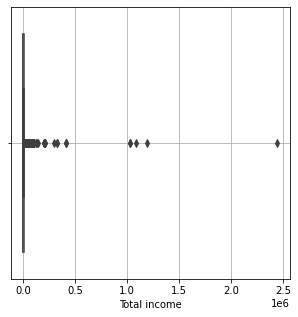
* A description of the first few columns of the dataset is given below:

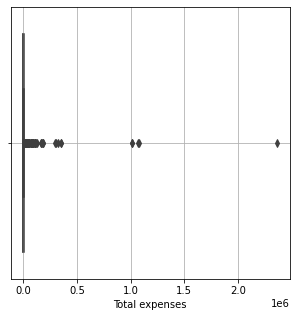
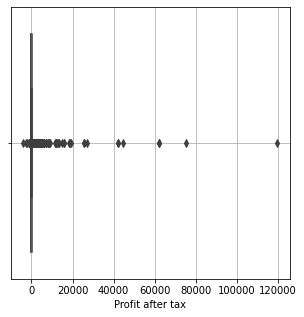


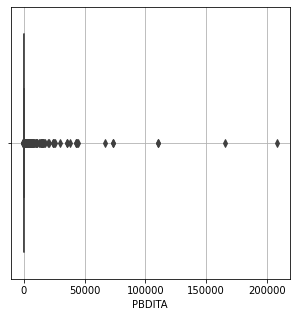
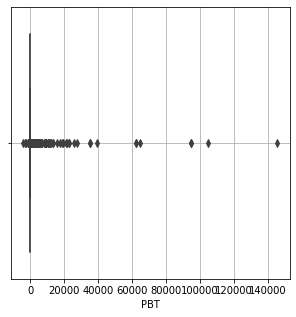
1.1 Outlier Treatment

Detecting Outliers

From the above boxplots, we can see there are outliers present in several variables

Another method of detecting Outliers:

Conventionally, outliers are identified based on the inter-quartile distance as follows:

Q1 – 25th Percentile

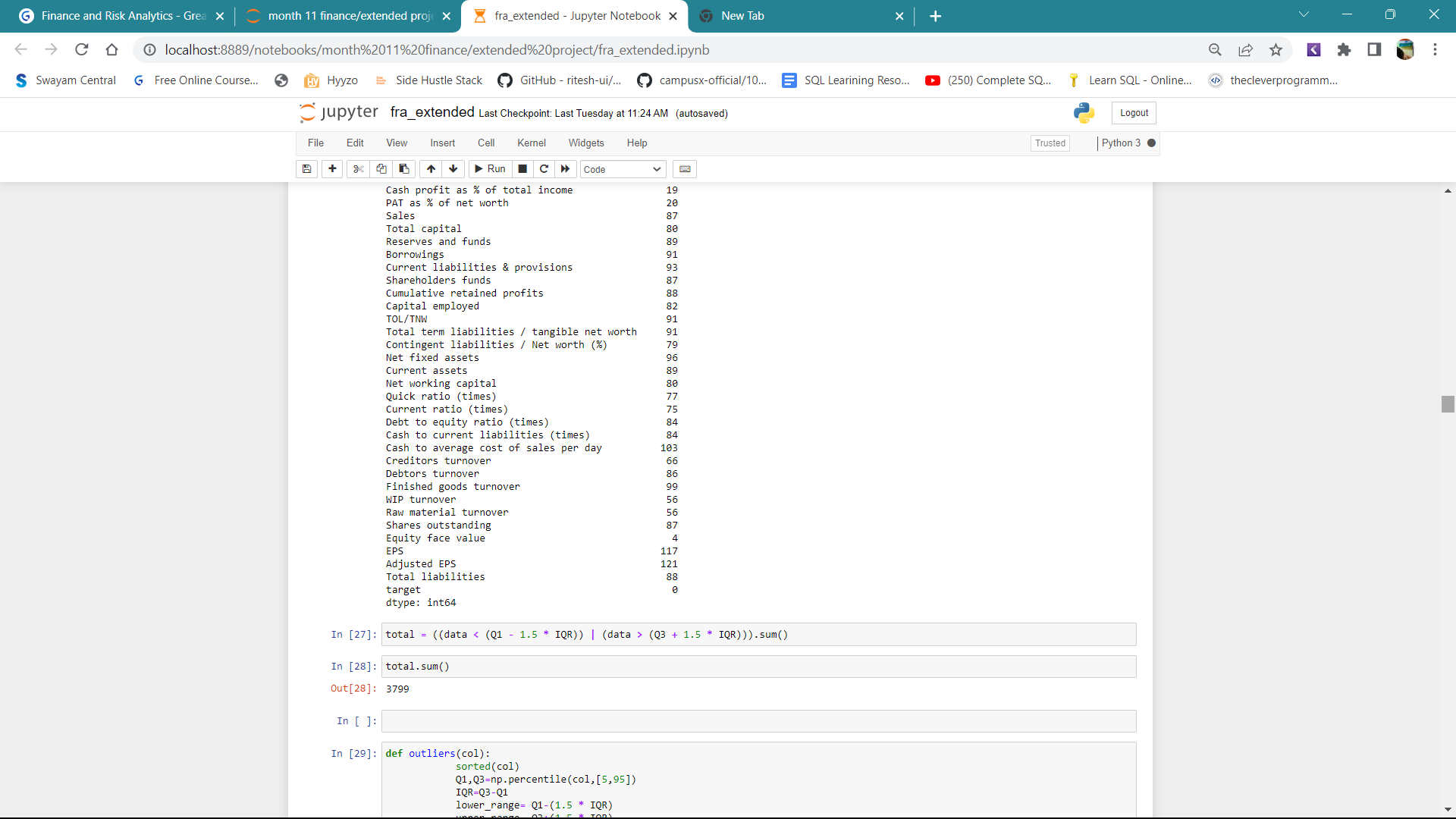
Q3 – 75th Percentile

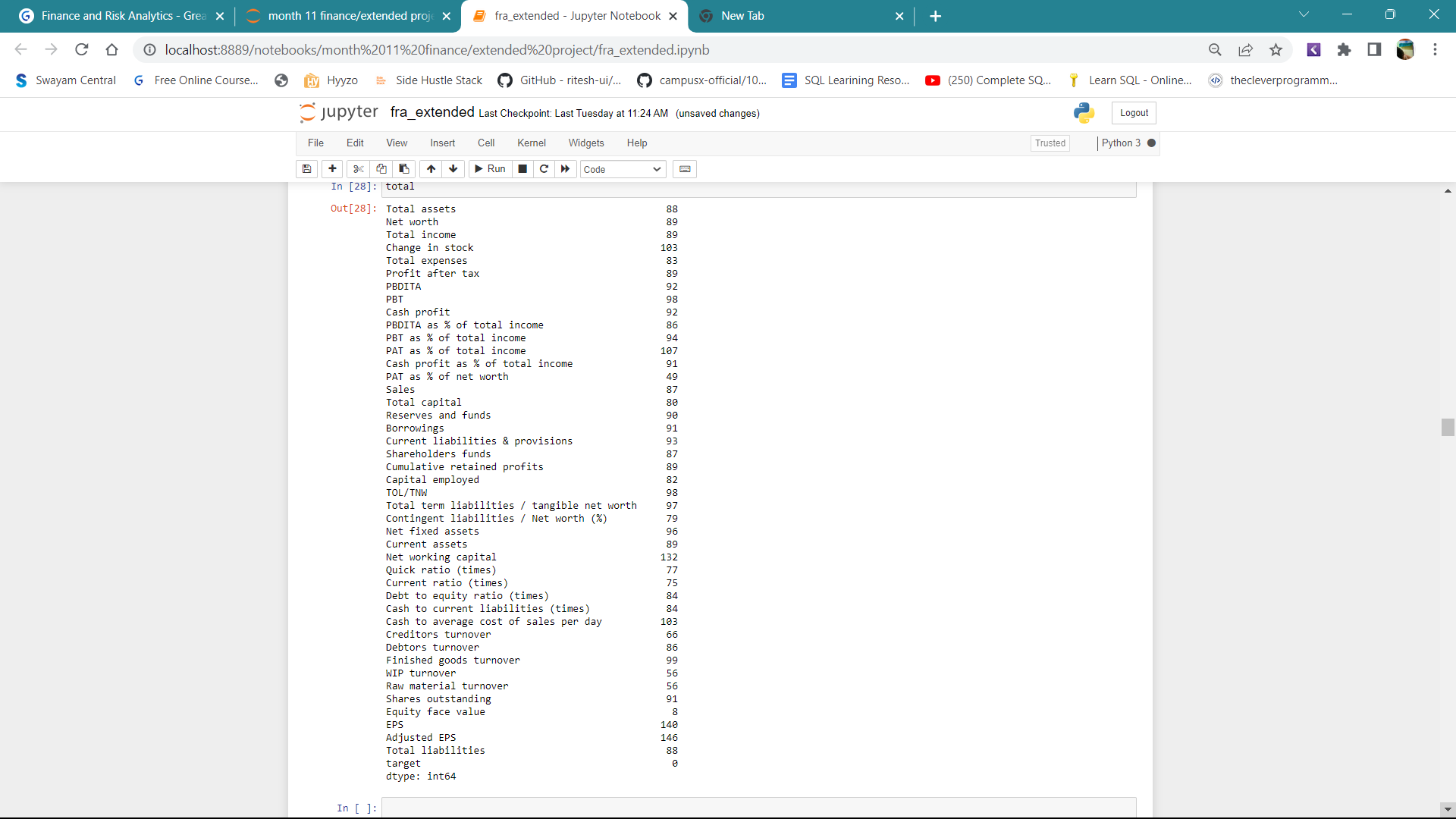
IQR = Q3 – Q1

Lower outlier = Value < 1.5 \* IQR

Upper Outlier = Value > 1.5 \* IQR

Based on this definition, the number of outliers in the dataset is as follows:





By further increasing the boundaries to 5th and 95th Percentile, the number of outliers

is as follows:

This seems to be a reasonable option, so, outliers have been treated as follows:

* Q1 – 25th Percentile
* Q3 – 75th Percentile
* IQR = Q3 – Q1
* Q05 – 5th Percentile
* Q95 – 95th Percentile

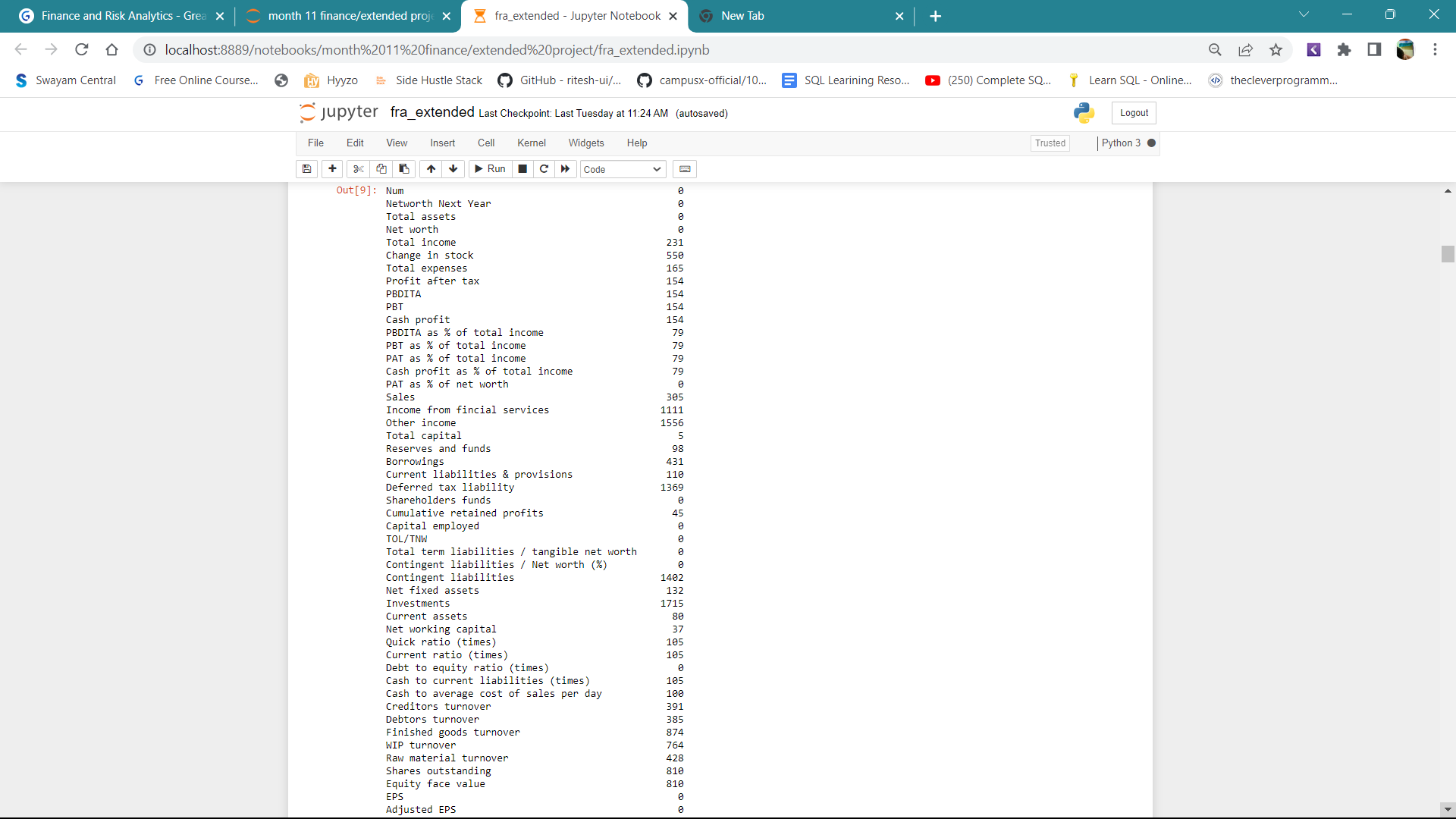
Rules:

* If Value < 1.5 \* IQR, then replace it with Q05 (5th percentile cap on lower outliers)
* If Value > 1.5 \* IQR, then replace it with Q95 (95th percentile cap on upper outliers

**1.2 Missing Value Treatment**

The dataset with the treated outliers is used for further analysis. There is a

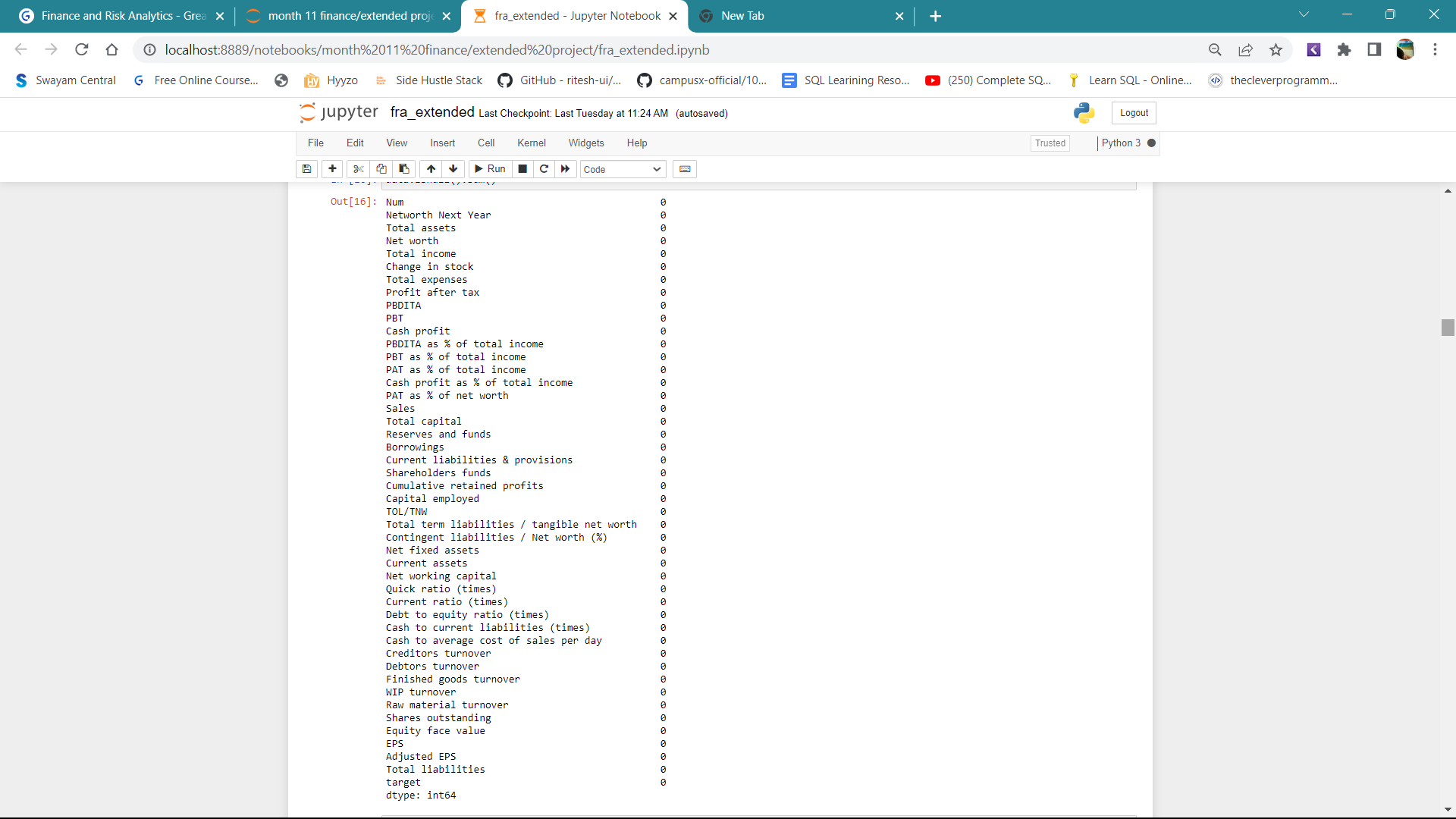
* total of 118 missing values as seen below:



**Approach**

* Since the variables have outliers, median is the best measure of central tendency to fill in missing values.
* Accordingly, the missing values have been treated using the mean.
* After imputing the missing values, it is confirmed that there are no

more missing values:

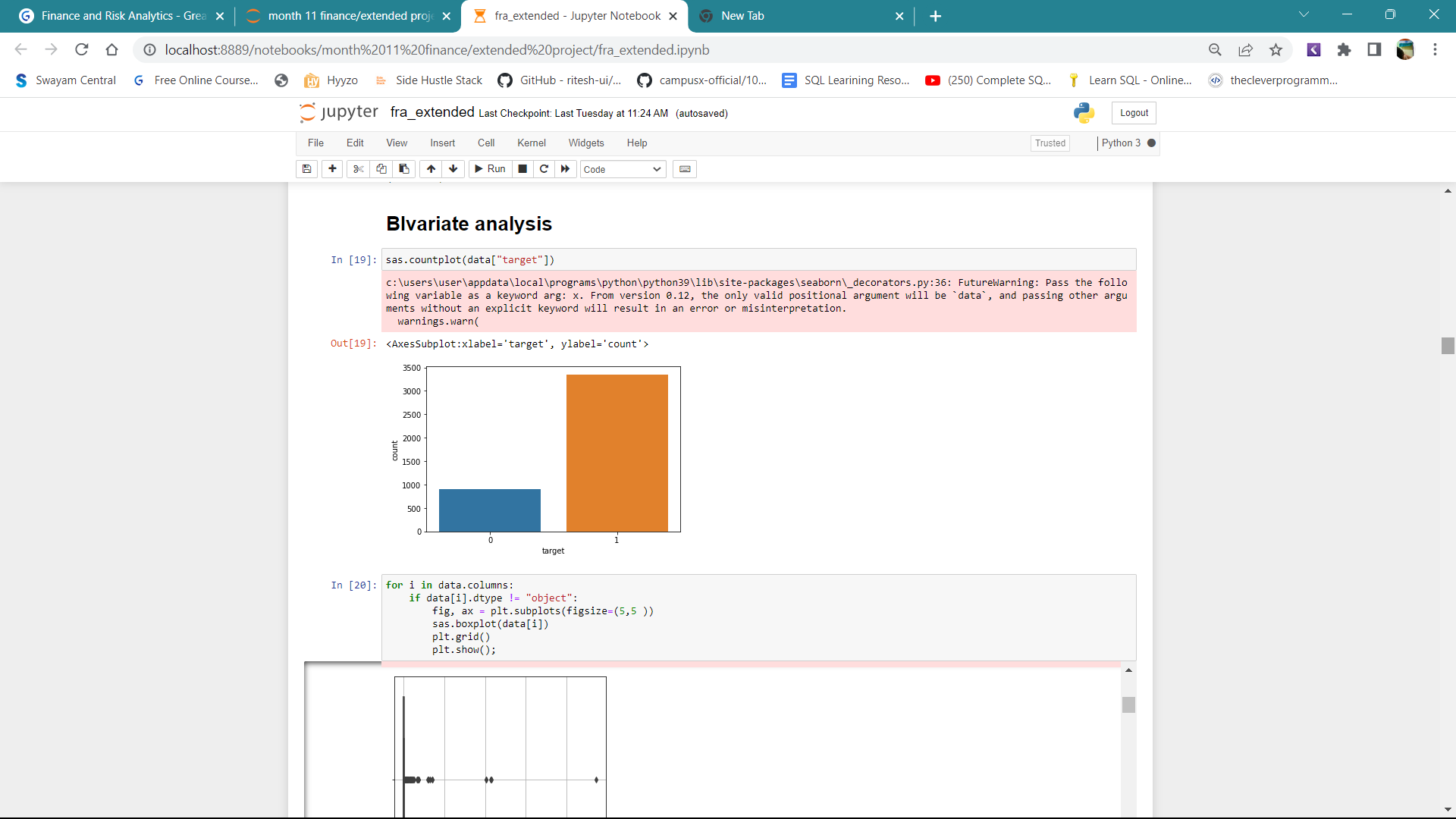


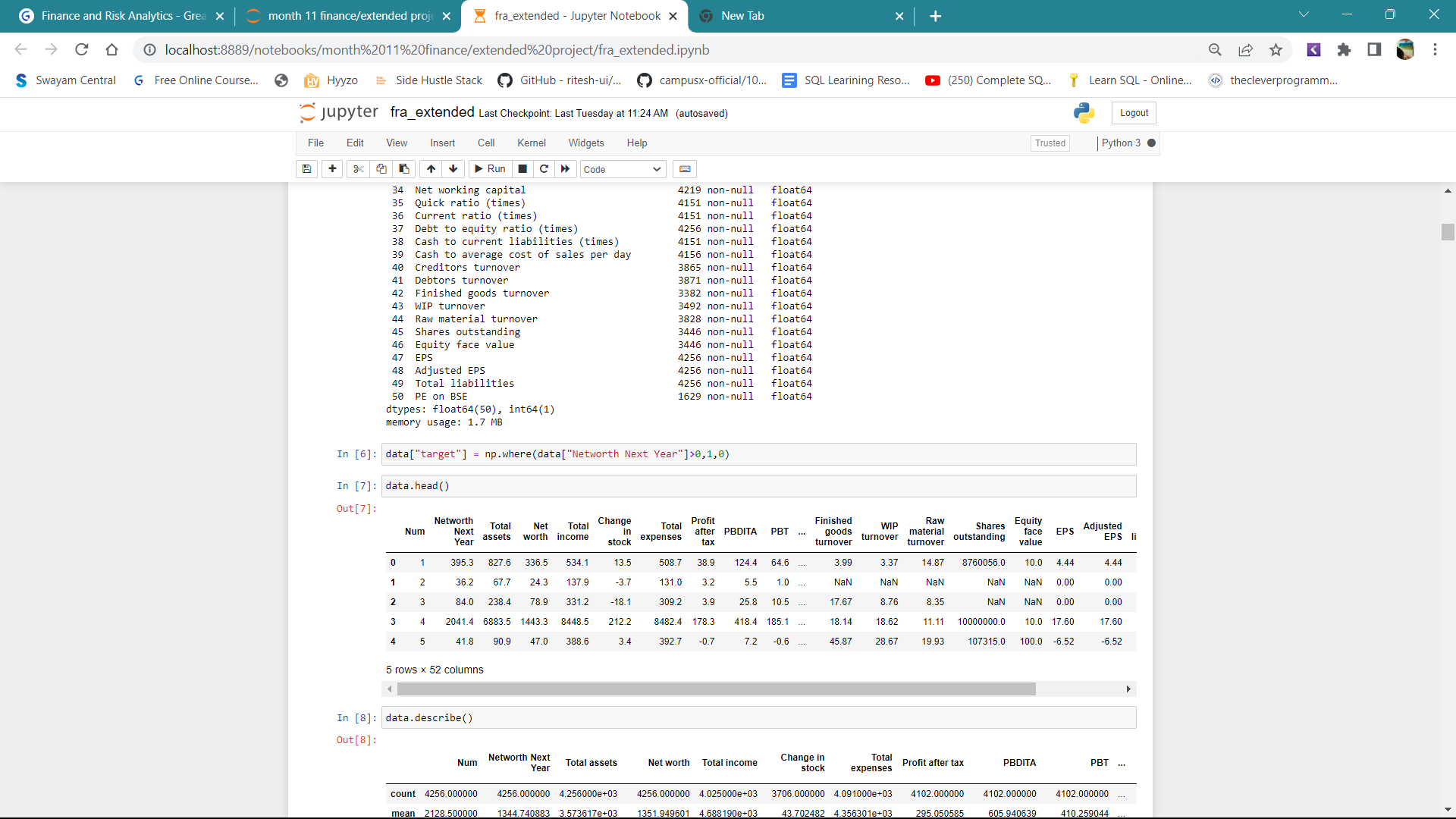
**1.3 Transform Target Variable into 0 and 1**

The definition of ‘default’ is as follows:

1.3.1 If ‘Networth\_Next\_Year’ < 0, default = 1

1.3.2 If ‘Networth\_Next\_Year’ > 0, default = 0



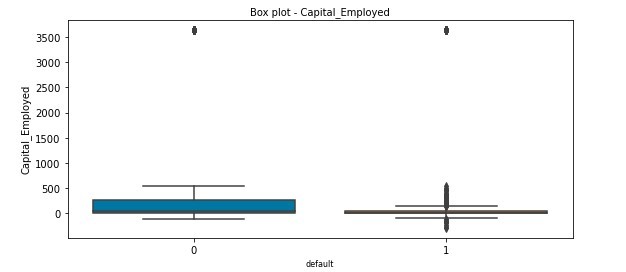
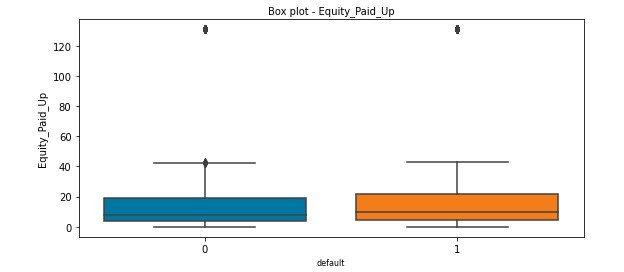
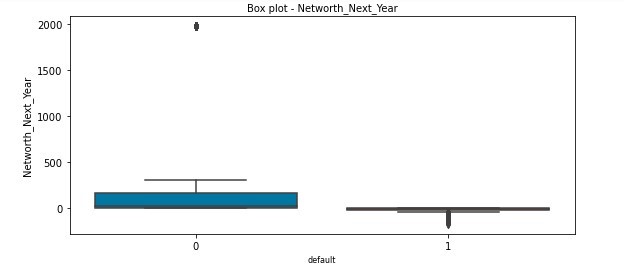


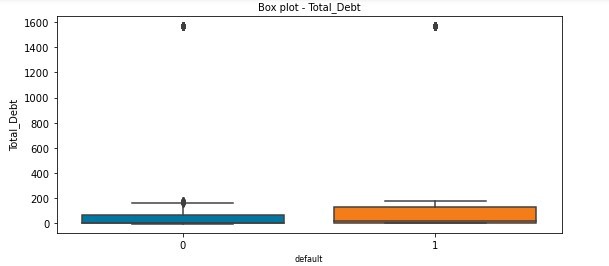
**1.4 Univariate & Bivariate analysis with proper interpretation**

**Univariate Analysis**

* The purpose of Univariate Analysis is to find out which variables have clear separation for the target variable ( default = 0 and 1 ).
* The boxplot is a good visual technique to identify such variables as

seen below for some of the independent variables:



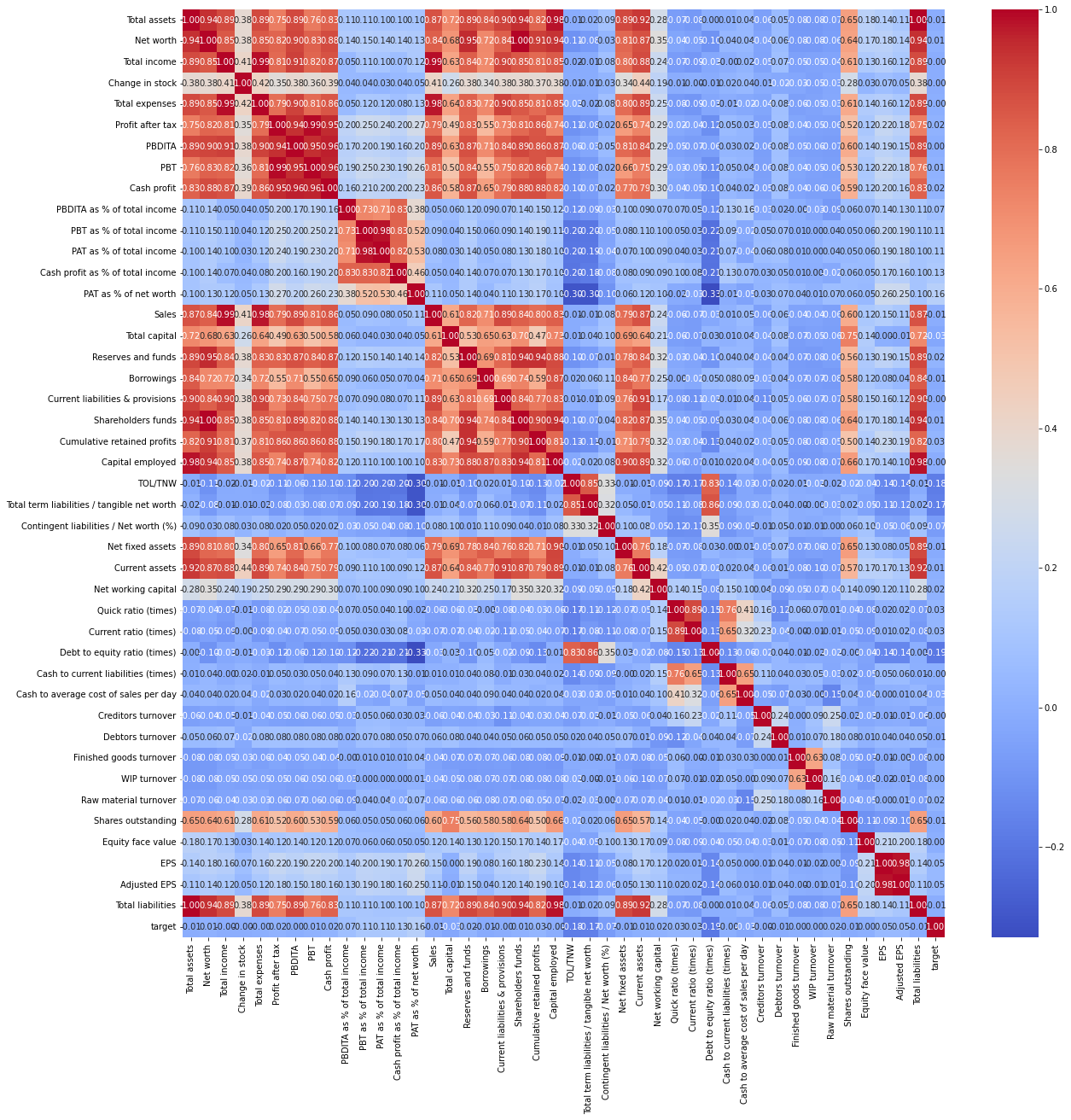


From the above , It is seen that there is no clear separation in the median values of the variables ‘Equity\_Paid\_Up’ and ‘Total\_Debt’ between companies who defaulted (1) and those who did not. Such variables are not good predictors.

On the other hand, variables like ‘Networth’ and ‘ Capital\_Employed’ have clearer separation. Such variables are better predictors

**Bivariate analysis**

Since all the variables are continuous variables, the heatmap of the correlation matrix can give a very good idea of the correlations between the independent variables and also with the dependent variable.

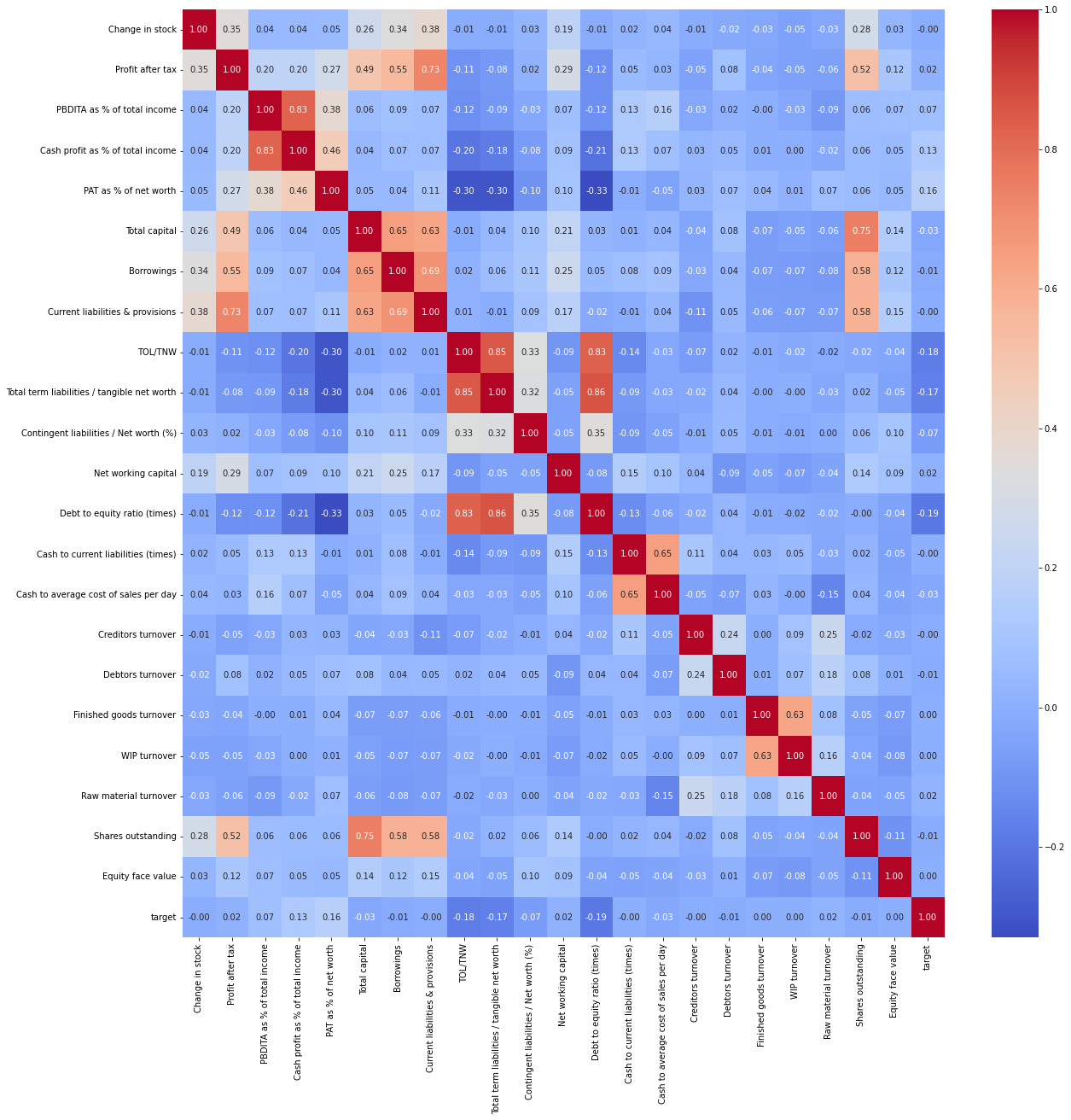
****

It is seen that there is a high correlation between several independent variables.

Thus, for eg, Networth is highly correlated to Total\_Debt and Net\_Working\_Capital; Gross\_Sales to Other\_Income, PBIDT is correlated to PBIT, PAT etc. This clearly establishes that the problem of Multi-collinearity exists.

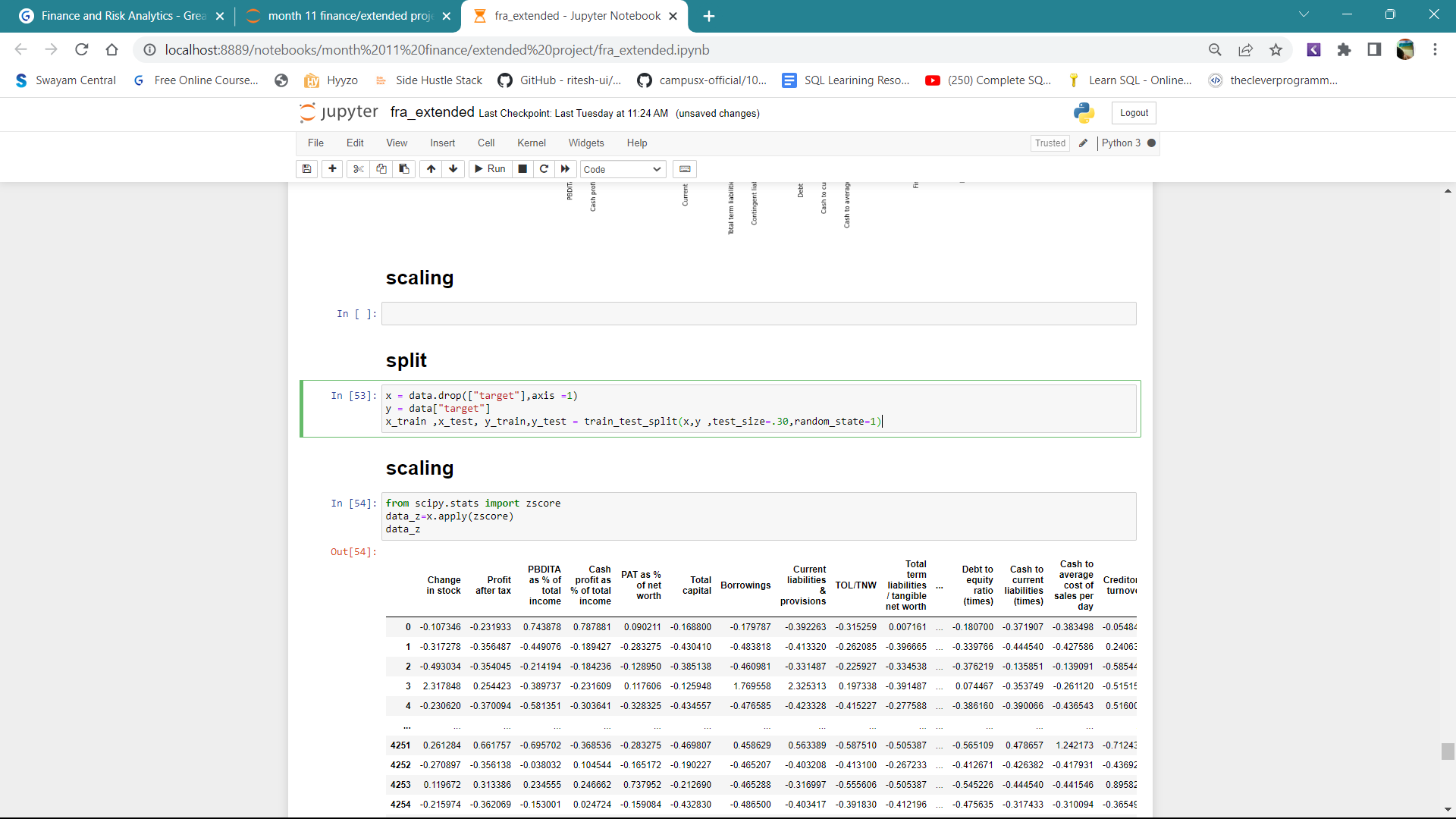
Therefore, we see that variables like Networth, Capital\_Employed, PBIDT, Curr\_Assets, PBDT are highly correlated with Networth\_Next\_Year

**To remove multi colinearity we use vif function to eleminate columns.**

* **There are 41 values which are above vif > 5 that we don’t use them in model building**
* **There are 23 values which are less than vif<=5, we use them in model building.**
* ****

**1.5 Train-Test Split**

As stated in the problem, the dataset has to be split in the ratio of 70:30 using random\_state 2. Accordingly, the datasets have been split as follows:

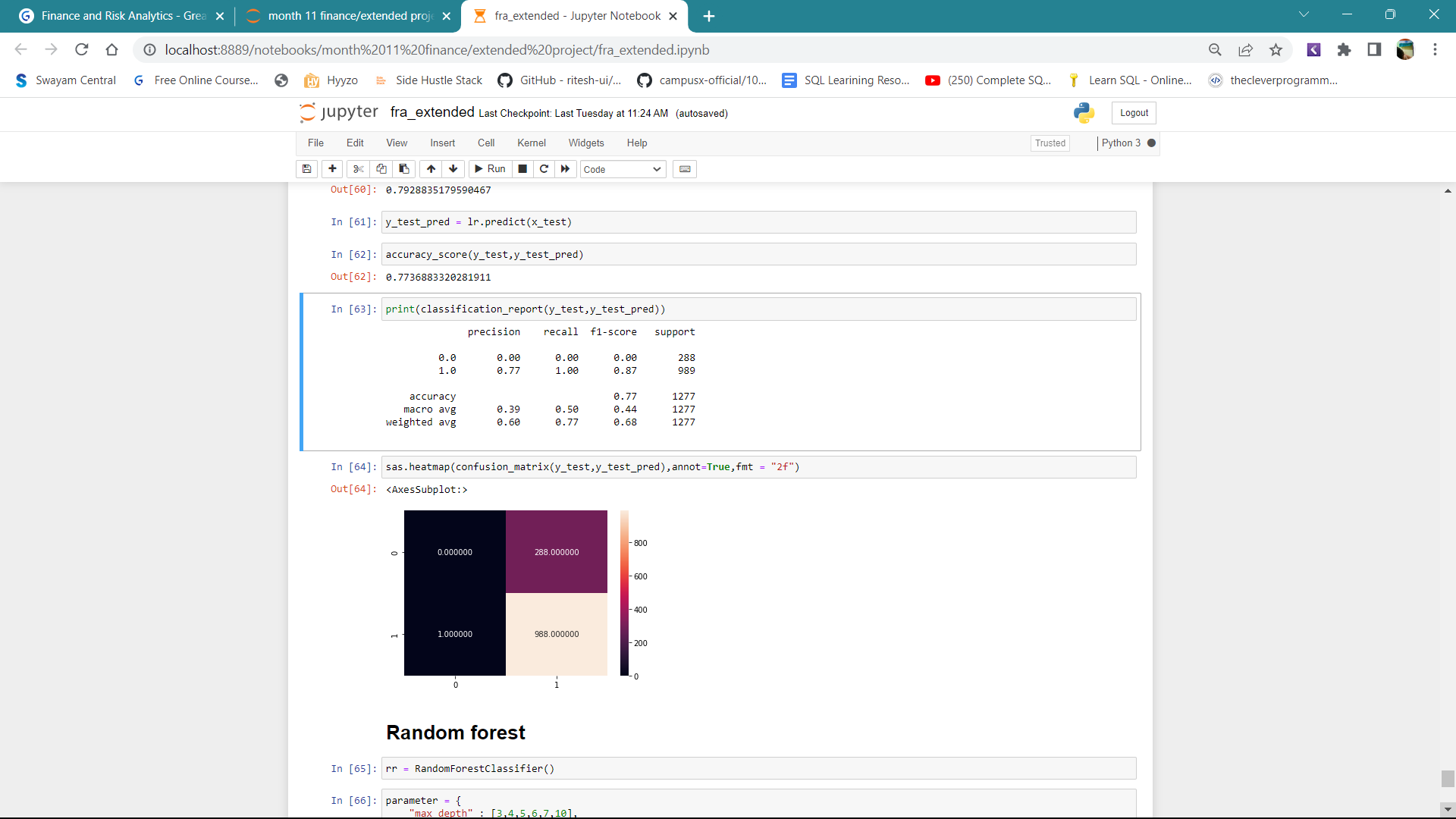


**1.6 This criterion is linked to a Learning OutcomeBuild Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach**

**and**

**1.7This criterion is linked to a Learning OutcomeValidate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model**

* We use logistic regression on the most important feature in the dataset
* We got the accuracy of .79288



**Approach for cutoff**

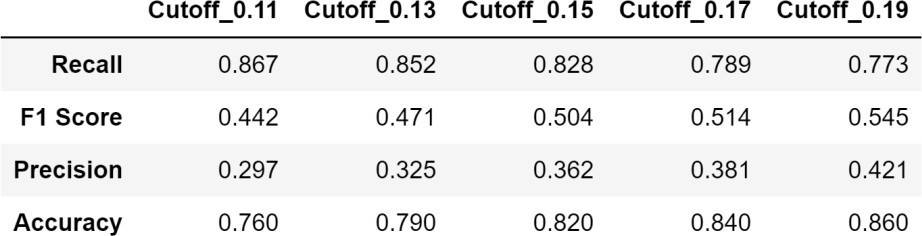
We are interested in predicting whether a company is likely to default, so that a bank or a lending company may either charge a higher premium in case of a high risk, or not give a loan at all. In such a case, Recall is the most important metric. A model which maximizes Recall (and therefore minimizes False Negatives) is preferred. This ensures

that loans are not given to companies which the model had predicted would not default, but later on, actually defaulted.

On the other hand, being over-cautious and not lending to any companies who have even a slight risk of default will also not be practical and will lead to loss of business and market share. Therefore, there needs to be a balance between the two.

The following table gives a comparison between the different performance measures at different cutoff levels (i.e. at what value of y would be differentiate between defaulters and non-defaulters on the train dataset). A minimum cutoff value of 0.11 is considered, since the proportion of defaulters in the original dataset is ~ 0.11. The

* models have been evaluated with 5 different cutoff values as shown below:

4

**Optimum Cutoff**

There needs to be a balance between the following 2 opposing factors:

**False Positives**

This means the model predicts the company is a defaulter, but in reality, is not a defaulter. This scenario represents a Lost Opportunity for the investor, since he would not have invested in the company, thinking it was a defaulter.

**False negatives**

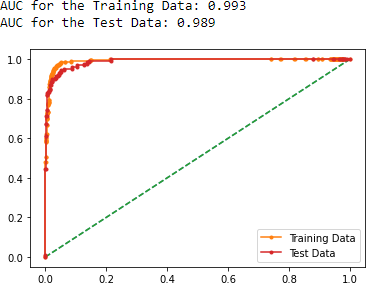
This is when the model predicts the company is not a defaulter, but in reality, actually defaults. This is a big loss to the investor, and therefore, needs to be minimized.

**1.8This criterion is linked to a Learning OutcomeBuild Random Forest Model on Train Data and validate on the Test Data**

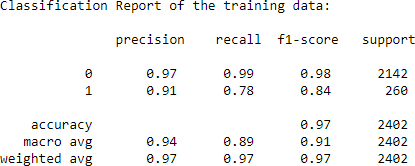
Random Forest Regression is **a supervised learning algorithm that uses ensemble learning method for regression**. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model

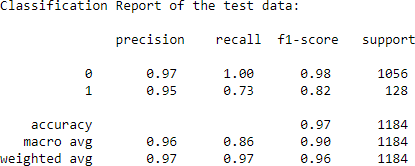


### AUC – ROC curve for train and test dataset

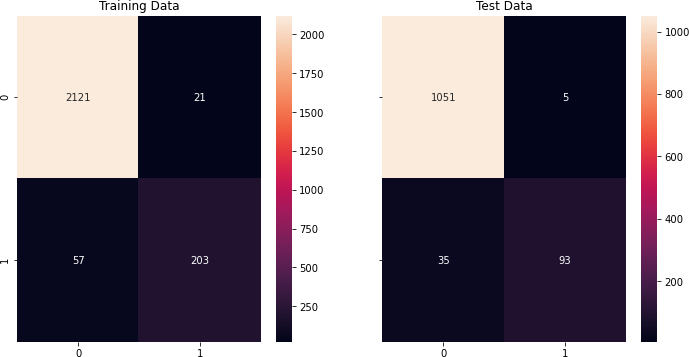


**Classification Report**





### Confusion matrix



Training and testing set results are almost similar, with overall measures on a decent side.

Net worth, Book\_Value\_Unit\_Curr and Book\_Value\_Adj\_Unit\_Curr are the most important variable for predicting claimed.

Recall of defaulters has come down with respect to the Logistic Regression model. Accuracy andprecision of defaulter class have good values compared to Logistic Regression.

**parameter used :-**

**parameter = {**

**"max\_depth" : [3,4,5,6,7,10],**

**"max\_features" : [ 3,4,5,6],**

**"min\_samples\_leaf" : [25,50,100,75 ],**

**"min\_samples\_split" : [ 50,100,150,300],**

**"n\_estimators" : [101,201,301,501],**

**"oob\_score" : [True , False]**

**}**

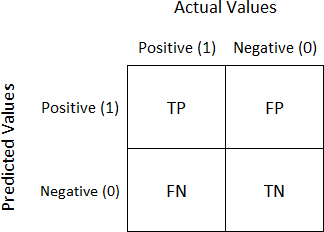
* There is no over-fitting.
* The error in the test data is slightly higher than the train data, which is absolutely fine because the error margin is low and the error in both train and test data is not too high. Thus, the model is not over-fitted or under-fitted

**1.8This criterion is linked to a Learning Outcome Compare the performance of Logistics Regression and Random Forest**

Models are evaluated on the basis of the below techniques to see how good it will perform for future records.

Some of the model evaluation techniques are:

* Accuracy – how precisely the model classifies the data points.



* Confusion Matrix – 2 \* 2 tabular structure reflecting the model performance in four blocks
* Receiver operating characteristics (ROC) curve – A technique to visualize classifier performance
* ROC\_AUC score – Area under curve, which is by calculating the percentage area below the curve.

### Best model for the given case study:

From the above results, we can say that Random Forest seem to be the optimized model for the given dataset. Since in this case study, classes are not well separated and there are lots of independent variables,.

Moreover, the training and test set results are in line for Random Forest than other models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Logistic**  **Regression** | | **Random Forest** | |
| **Metrics** | **Training set** | **Testing set** | **Training set** | **Testing set** |  |
| **Accuracy** | 79.89 | 77 | 96.75 | 96.62 |  |
| **Precision** | 0.55 | 0.49 | 0.91 | 0.95 |  |