|  |
| --- |
|  |
| CP G1 Group Assignment |
| Thera Bank – Loan Purchase Modelling |

Amit Sharma

Amitesh Bajpai

Sahil Sachdeva

Kushal Maheshwari

Neeraj Singh

Contents

[1. Project Objective 3](#_Toc22587522)

[2. Exploratory Data Analysis – Step by step approach 3](#_Toc22587523)

[Assumptions made before starting our analysis: 3](#_Toc22587524)

[2.1 Environment Set up and Data Import 4](#_Toc22587525)

[2.1.1 Installation of necessary packages and calling libraries 4](#_Toc22587526)

[2.1.2 Set up working directory 5](#_Toc22587527)

[2.1.3 Import and read the dataset 5](#_Toc22587528)

[2.2 Variable Identification 6](#_Toc22587529)

[2.3 Missing Value Identification 7](#_Toc22587530)

[2.4 Negative Value Treatment 8](#_Toc22587531)

[2.5 Outlier Detection 8](#_Toc22587532)

[2.6 Removing unnecessary variables 9](#_Toc22587533)

[2.7 Data Visualization 9](#_Toc22587534)

[2.7.1 Univariate Analysis 9](#_Toc22587535)

[2.7.2 Multivariate Analysis 11](#_Toc22587536)

[2.7.3 Correlation between variables 13](#_Toc22587537)

[3. K-Means Clustering 14](#_Toc22587538)

[4. CART and Evaluation of Model Performance 17](#_Toc22587539)

[Model Performance 20](#_Toc22587540)

[Pruning 20](#_Toc22587541)

[5. RANDOM FOREST 21](#_Toc22587542)

[Model Performance 24](#_Toc22587543)

[6. CONCLUSION 25](#_Toc22587544)

# 1. Project Objective

The objective of the report is to explore the dataset named “[Thera Bank-Data Set.xlsx](https://olympus.greatlearning.in/courses/5171/files/470597/download?wrap=1)” in R and generate insights about the same. This exploration report will consist of the following:

* Exploratory Data Analysis
* Data Clustering
* Building decision tree - CART
* Building decision tree – Random Forest
* Checking model performance
* Insights from the dataset

# 2. Exploratory Data Analysis – Step by step approach

We shall follow these steps to go through our data exploration activity:

1. Environment Set up and Data Import
2. Variable Identification
3. Variable Transformation / Feature Creation
4. Missing Value Treatment
5. Negative Value Treatment
6. Outlier Treatment
7. Removing unnecessary variables
8. Data Visualization
   1. Univariate Analysis
   2. Multivariate Analysis
   3. Correlation Matrix

## Assumptions made before starting our analysis:

* There is at least some correlation amongst the variables.
* Our sample size is large enough to yield reliable estimates of correlations among the variables.
* Our sample size is large enough to build, train and test models.

Here Personal Loan is considered the Dependent variable and all other attributes as Independent variables.

The data includes the demographic information of customer like (Age, Income, Experience, Family, zip code, family members, Education) which represent the customer behavior, so we need to take these columns under consideration.

The columns like (Mortgage, Securities, CD Acc, online, credit card) helps us to understand facilities avail by the customer with bank and represents the customer satisfaction with bank to encourage its customer to go for personal loan, so we need to consider these too.

Here we should not consider the ID as its completely unique for each customer and does not help in model building.

## 2.1 Environment Set up and Data Import

We need to set up our working directory, install all the necessary packages and import the Bank dataset in question, to proceed with our study.

### 2.1.1 Installation of necessary packages and calling libraries

The below mentioned libraries are necessary for our analyses in R.

*#Importing and calling libraries*

*install.packages("readxl")*

*install.packages("caTools")*

*install.packages("rpart")*

*install.packages("rpart.plot")*

*install.packages("rattle")*

*install.packages("RColorBrewer")*

*install.packages("data.table")*

*install.packages("ROCR")*

*install.packages("randomForest")*

*install.packages("corrplot")*

*install.packages("ineq")*

*install.packages("ggplot2")*

*install.packages("factoextra")*

*install.packages("psych")*

*#Loading all the required libraries*

*library(readxl)*

*library(caTools)*

*library(rpart)*

*library(rpart.plot)*

*library(rattle)*

*library(RColorBrewer)*

*library(data.table)*

*library(ROCR)*

*library(randomForest)*

*library(corrplot)*

*library(ineq)*

*library(ggplot2)*

*library(factoextra)*

*library(psych)*

### 2.1.2 Set up working directory

Setting a working directory at the beginning of an R session makes importing and exporting data files and code files easier. Basically, working directory is the location/folder on the PC where we have the data, codes, etc. related to our project.

*> setwd("C:/Users/neera/OneDrive/Desktop/R Programming")*

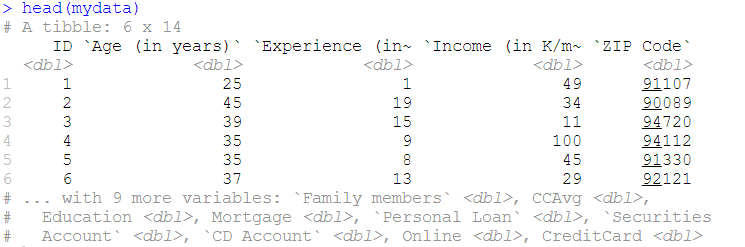
### 2.1.3 Import and read the dataset

The given dataset is in .xlsx format. Hence, the command ‘read.excel’ is used for importing the file. There are two sheets in the file, and we are interested in Sheet #2 and hence we import only the 2nd sheet.

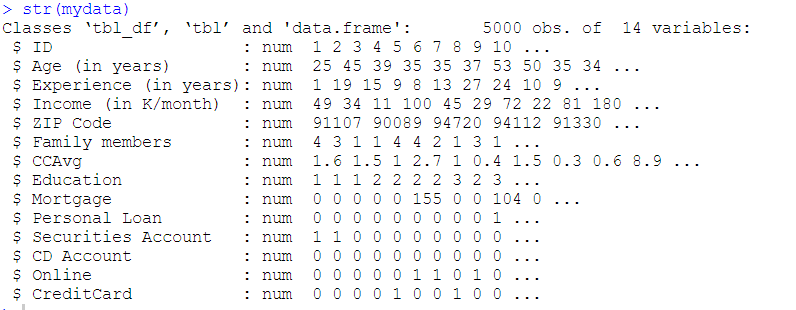
*> mydata=read\_excel("Thera Bank-Data Set.xlsx',sheet = 2)*

## 2.2 Variable Identification

The “head” function is used to view the top few observations of our dataset.

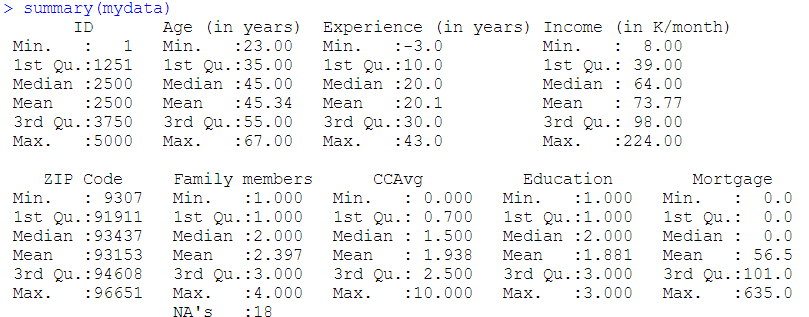


The “str” function shows us the structure of the dataset.



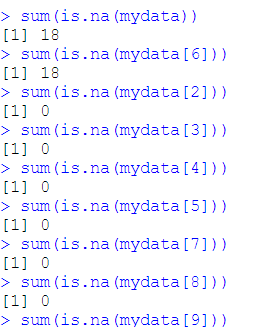
The “summary” function performs a Univariate analysis on our dataset. The summary shows us that the mean and median are very close for most of the variables. Also, we can conclude below points.

1. Mortgage, PersonalLoan, SecuritiesAccount, CD.Account, Online, Credit card columns all are having only 0 or 1 as a value.
2. Family Members column is having 18 Null value for which we must do null value analysis.
3. Personal loan is having mean of 0.096 which infers having 9.6 % success rate in last year campaign.

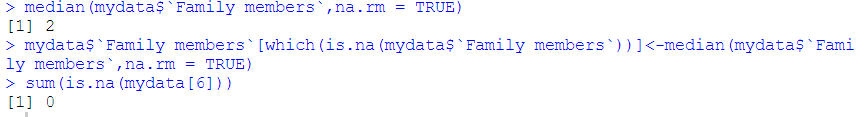


## 2.3 Missing Value Identification

There are 18 missing datapoints in the file. All missing datapoints are from column 6 i.e. Family members.



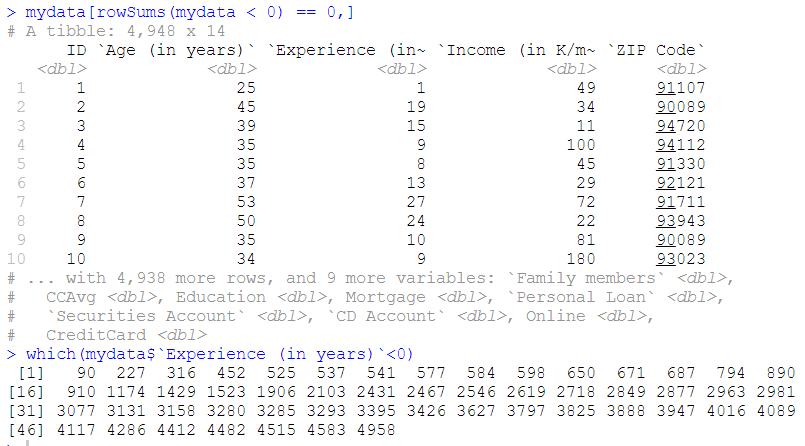
We transform the missing values in the following manner.



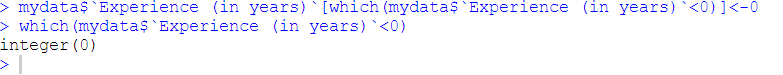
The 18 missing values in Family members column has been replaced by the median value of the Family members variable.

## 2.4 Negative Value Treatment

We identify the negative values present in the dataset.

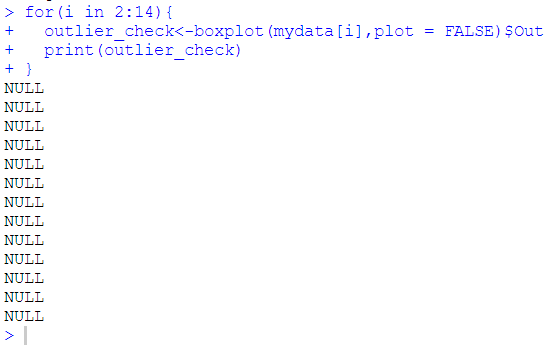


There are 52 missing values in the Experience column, and we replace them again with the value 0 which means there is 0 professional experience.



## 2.5 Outlier Detection

We check for outliers using the below code.



## 2.6 Removing unnecessary variables

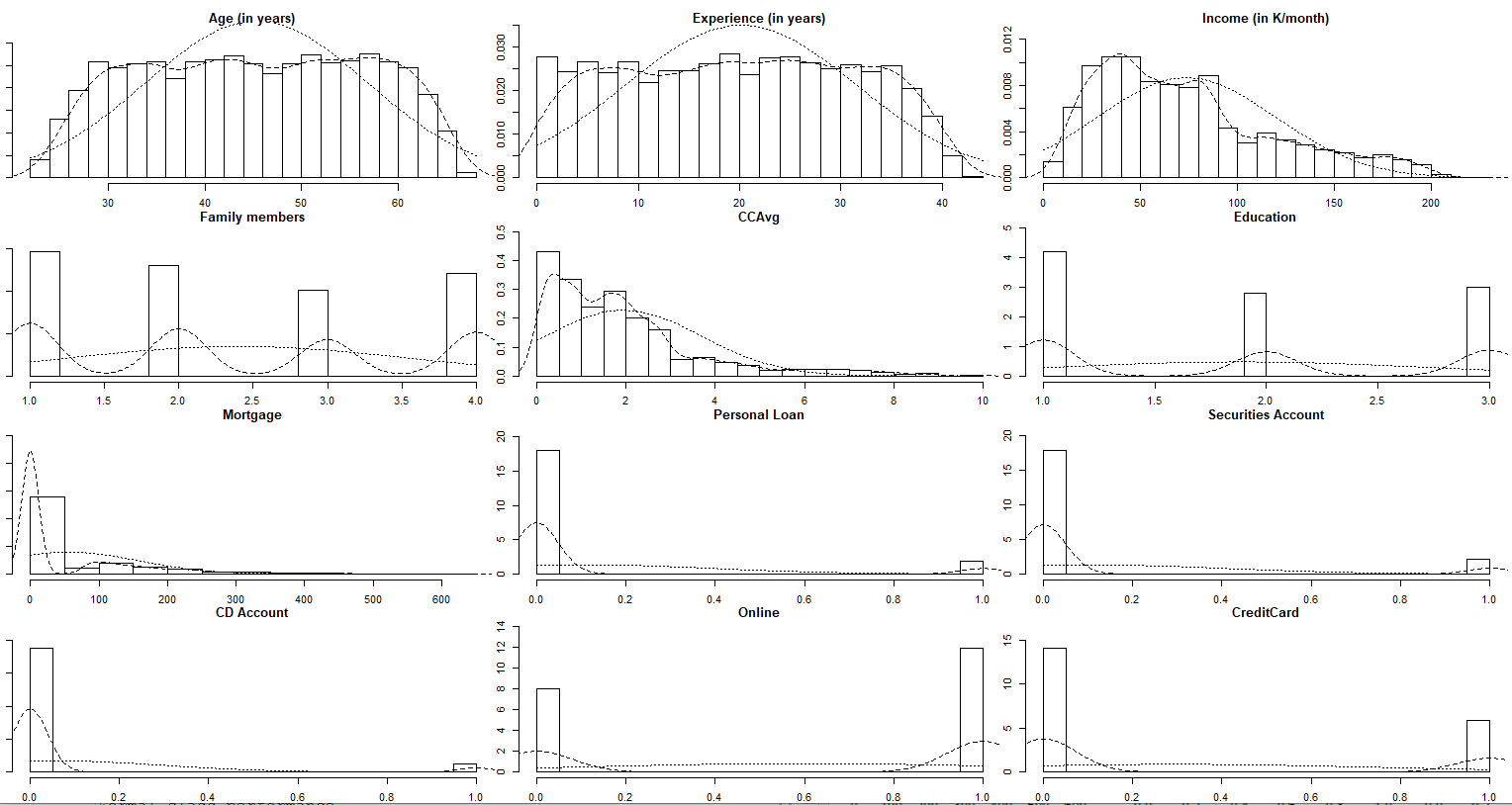
It is visible that the variables ID and ZIPcode have no relationship with whether a customer will avail a Personal Loan. Hence, we remove these columns from the dataset for the ease of our analysis.



## 2.7 Data Visualization

### 2.7.1 Univariate Analysis

We plot histograms for all the 12 variables as seen below.



Most customers do not have a Personal Loan, Securities account, or CD account as seen above.

Here we are dealing with imbalanced data, only 10% of the population accepted the offer.

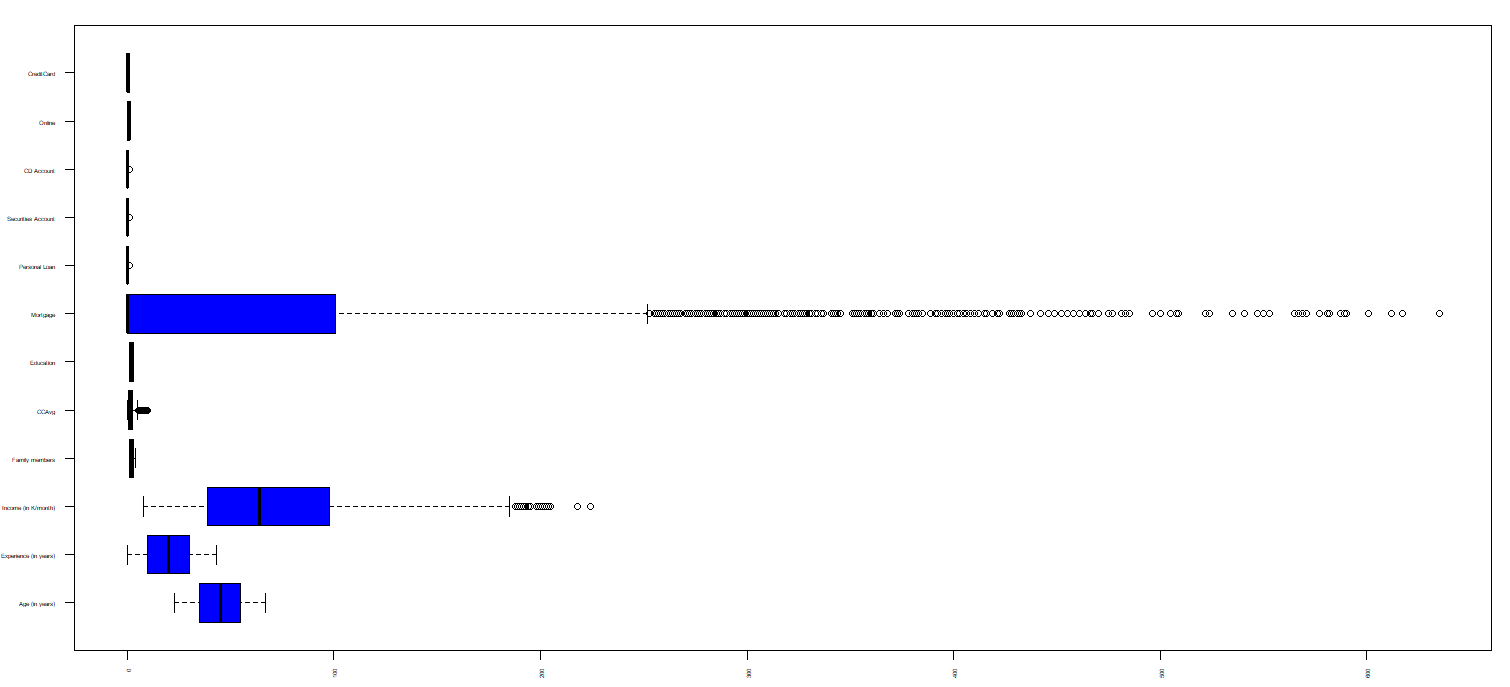
Majority of the customers do not have a securities account - 90%

Majority of the customers do not have a deposits account - 94%

Mortgage data is highly skewed, means only few customers have house mortgage

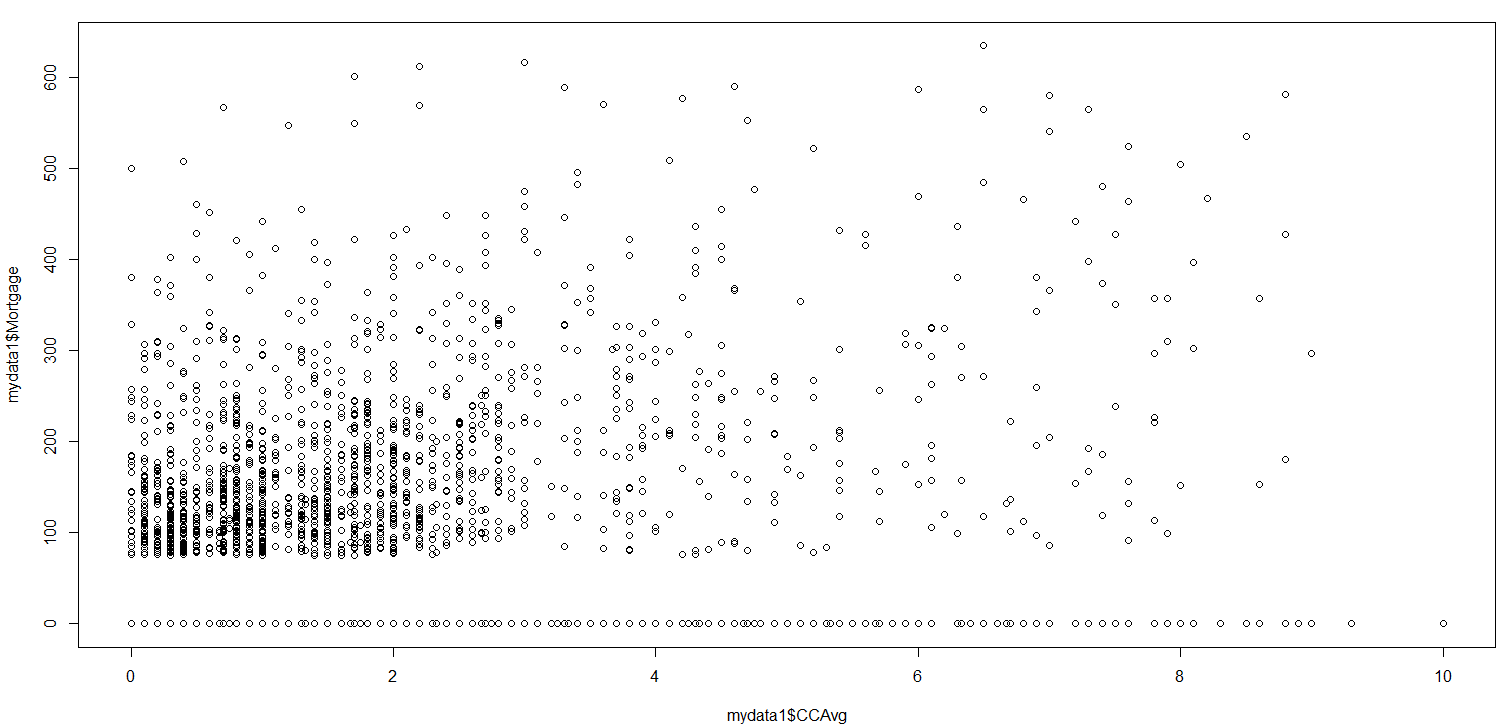
Age, professional experience has normal distribution

We plot the boxplots of all the individual variables as well.



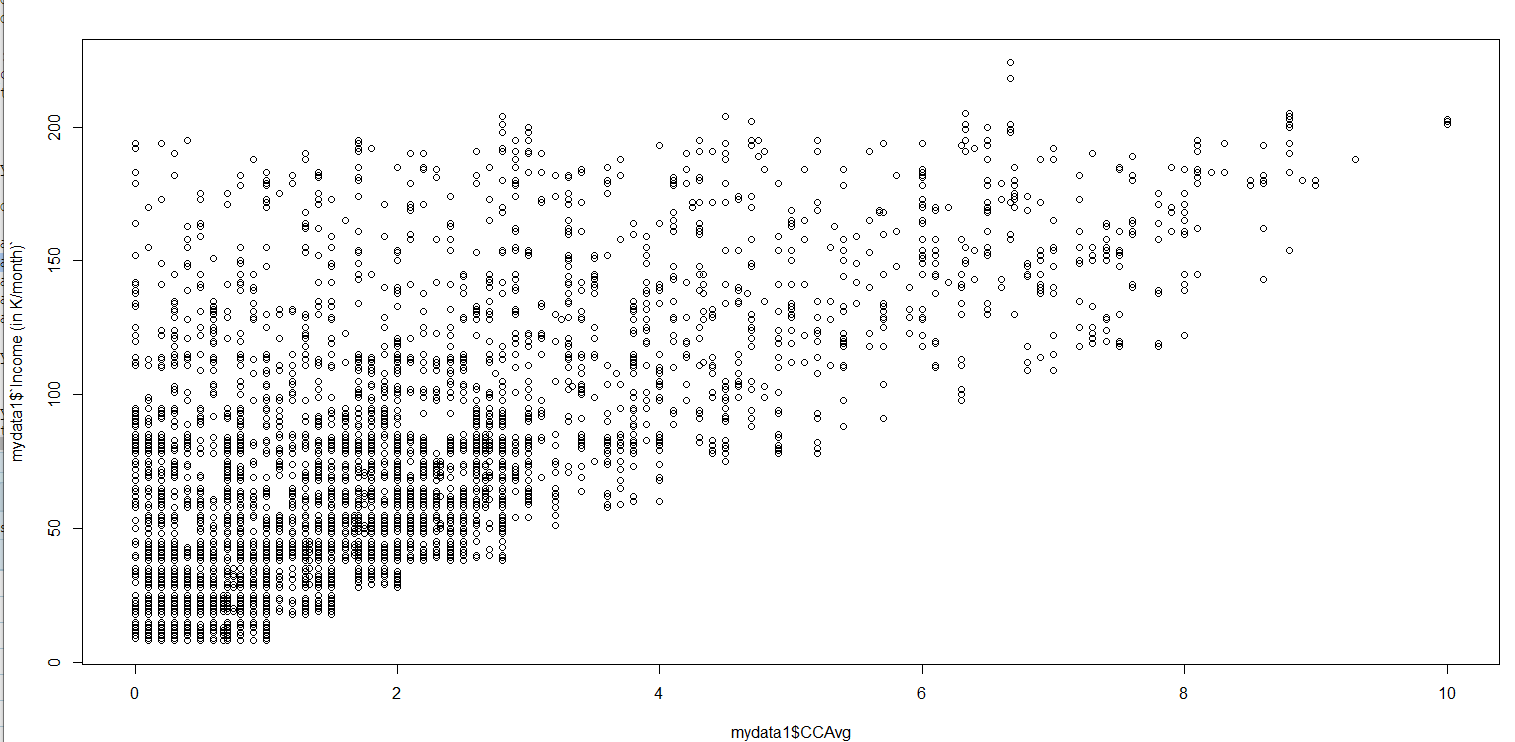
### 2.7.2 Multivariate Analysis

Average Credit card spending vs. Mortgage value:



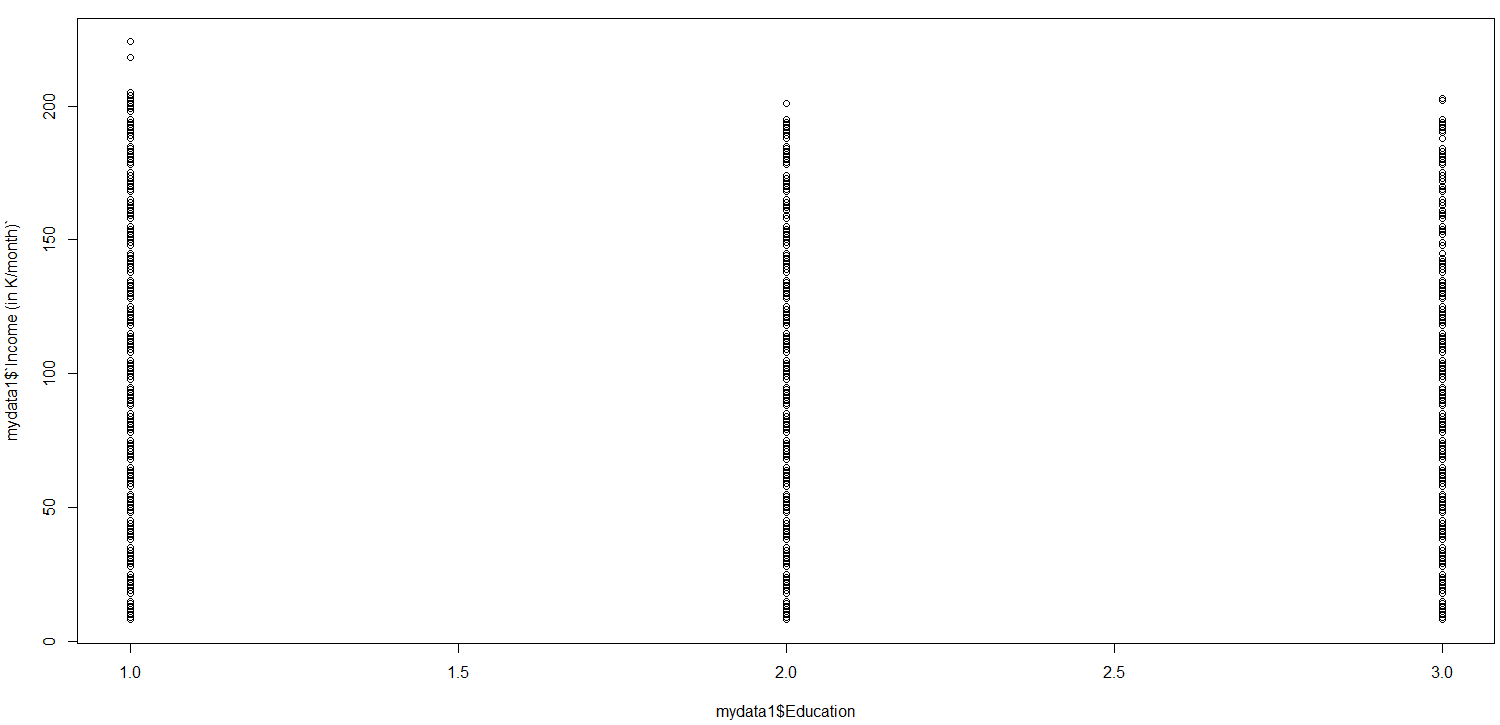
The relationship between Mortgage value and Credit card spending is scattered. With low value of mortgage, the spending can be huge. With high value of mortgage, the spending is less. But again, there are cases where the spending is quite high even with high value of mortgage.

Average Credit card spending vs. Income :



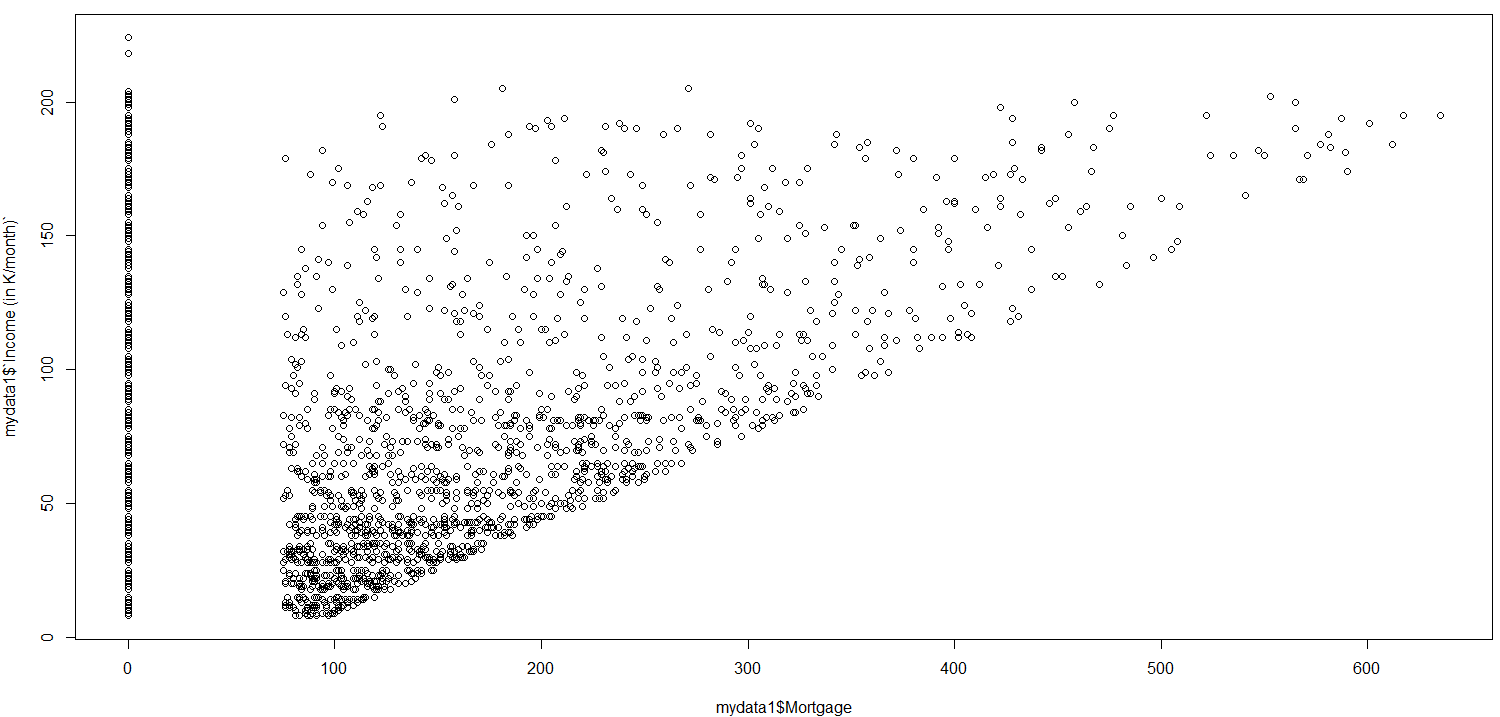
The average credit card spending increases with Income in many scenarios, but it is also true that the spending can be more even with low income and vice versa.

Education vs. Income:



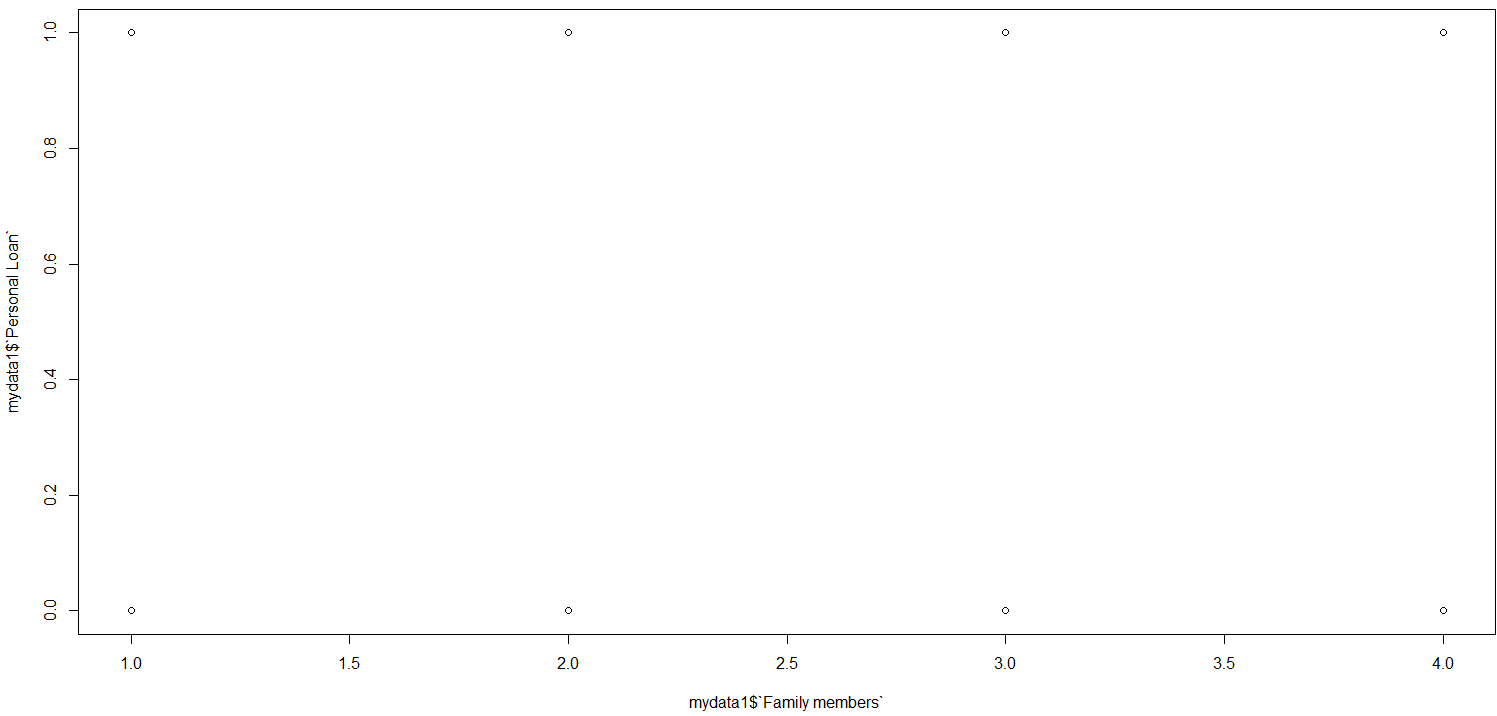
Customers with Education Level 1 has more income than Education levels 2 and 3.

Mortgage vs. Income :



The mortgage value is usually increasing with increasing income.

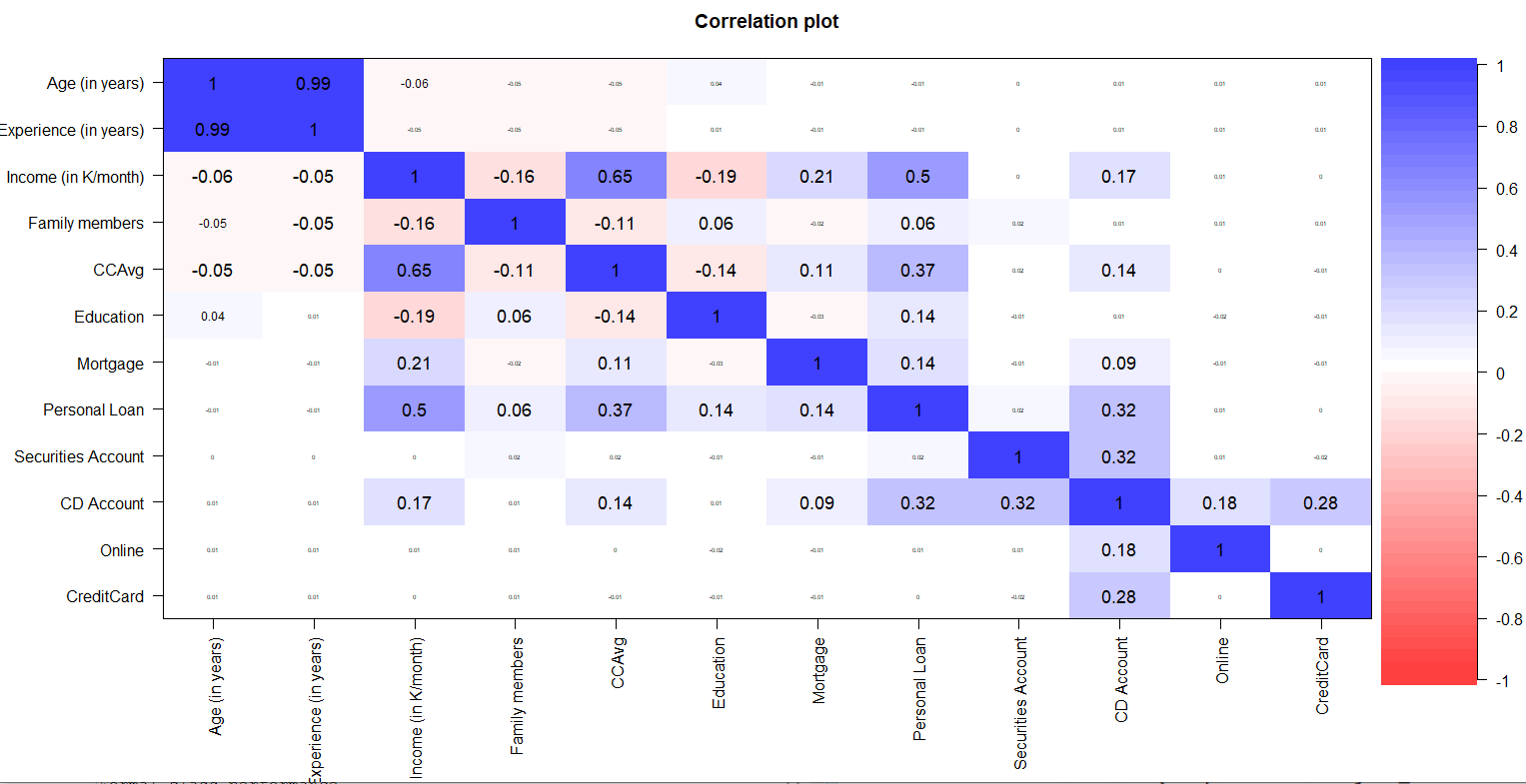
Family size vs. Personal Loan :



Family size has no clear impact on Personal Loan.

### 2.7.3 Correlation between variables

We plot the correlation matrix.



The variables with the highest correlation have been clustered together.

We can see there is high degree of correlation between Age of the person and professional experience.

We can see there is some degree of correlation between income and spending on credit card.

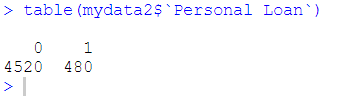
We can see there is some degree of correlation between income and mortgage amount.

There is no significant correlation between years of experience and income per month.

# 3. K-Means Clustering

In K-means clustering, observations are divided into K groups and reshuffled to form the most cohesive clusters possible according to a given criterion. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Our target variable is Personal Loan and before we start our analysis, we count the actual counts of customers with and without Personal Loans.

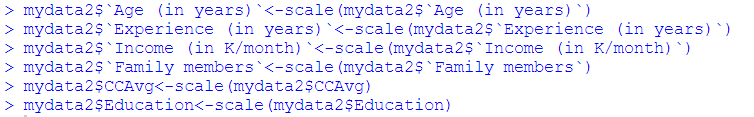


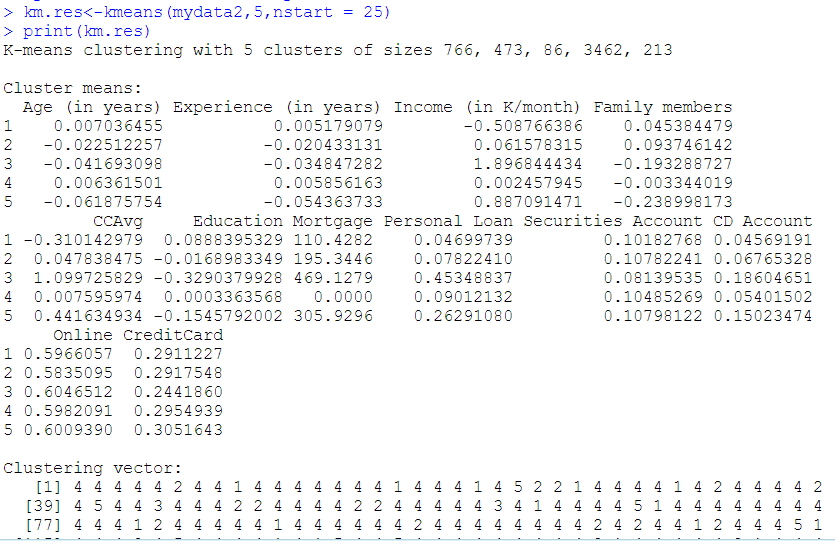
There are 480 customers with personal loans and the rest are not yet converted.

We first scale out dataset to match all the variables with each other.

After scaling, we apply K-means clustering which is a centroid based method and will help us understand our data better.

We choose k = 5.







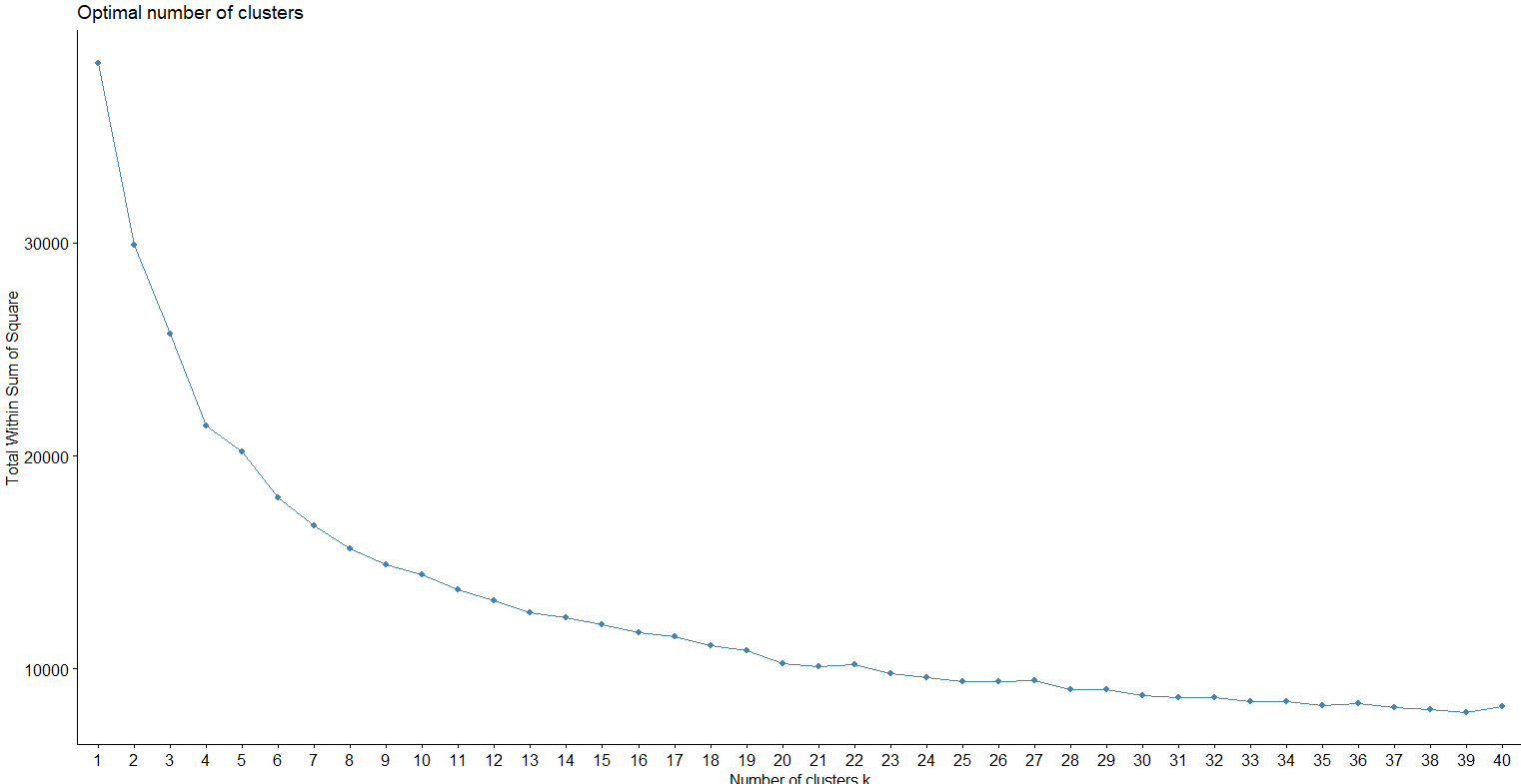
To choose the optimal value of k, we need to apply the Elbow rule. This method uses within-group homogeneity or within-group heterogeneity to evaluate the variability. We want to understand the percentage of the variance explained by each cluster. We want to reach that value of k that is beyond the diminishing returns. Adding a new cluster will not improve the variability in the data because very few information will be left to explain.



This above function will return the total within clusters sum of squares after running k times.

We run this over a range of k (we selected a max of 40).

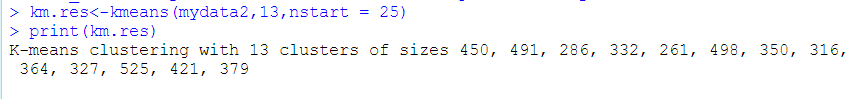
We then plot the graph to check the optimal value of k.



We observe that the optimal value of k is 13 and below that the curve has diminishing returns.

We therefore rerun the kmeans clustering, using the optimal value of k, i.e., 13.

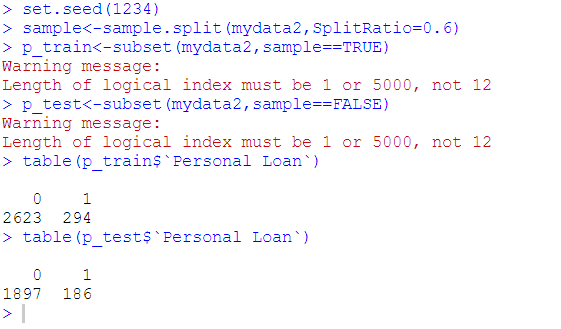
The sizes of our clusters are:



# 4. CART and Evaluation of Model Performance

Our first step is to split our dataset into training and test sets. We use a split ratio of 60% and 40%, respectively.

Checking customers with Personal Loan availed in each of the sets:



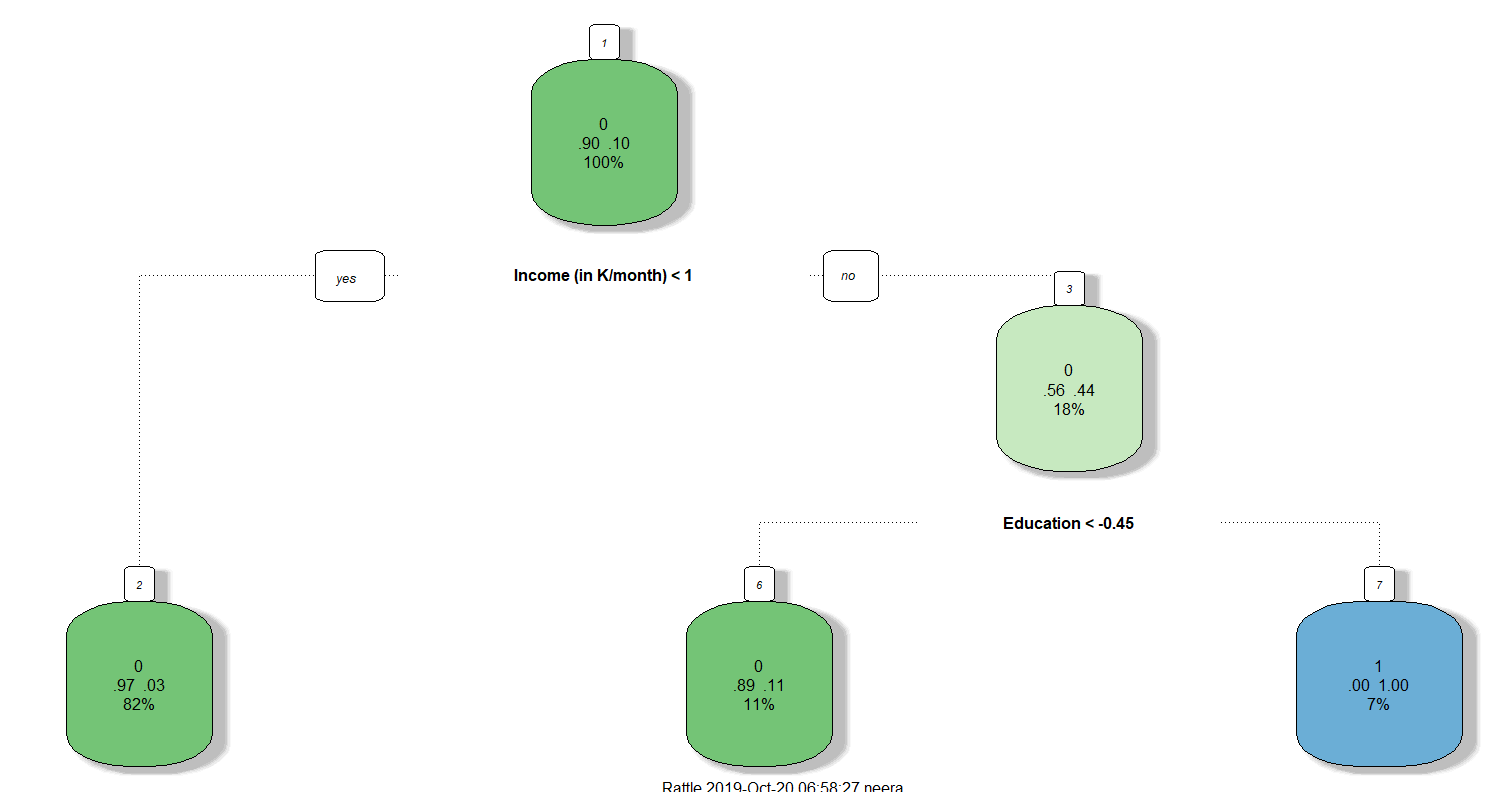
The customers with Personal Loans have been uniformly divided into the training and testing datasets.

We next set out control parameters and build the CART model and below are the results.





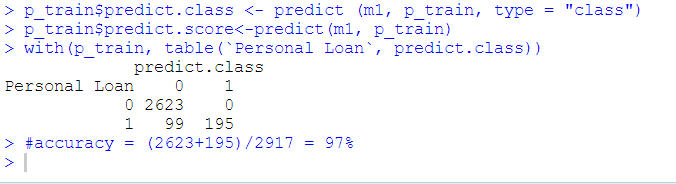
We display the decision tree.



Root node is Income.

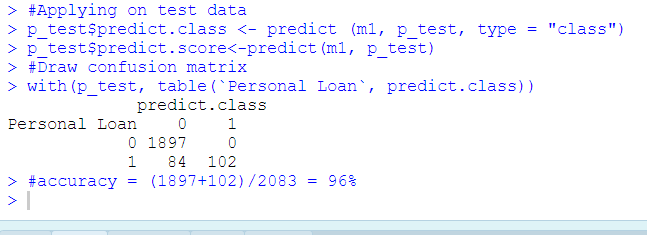
We next test our model on the testing dataset.

The confusion matrix of the training set is below:



The accuracy of the model when used on training set is ~ 97%.

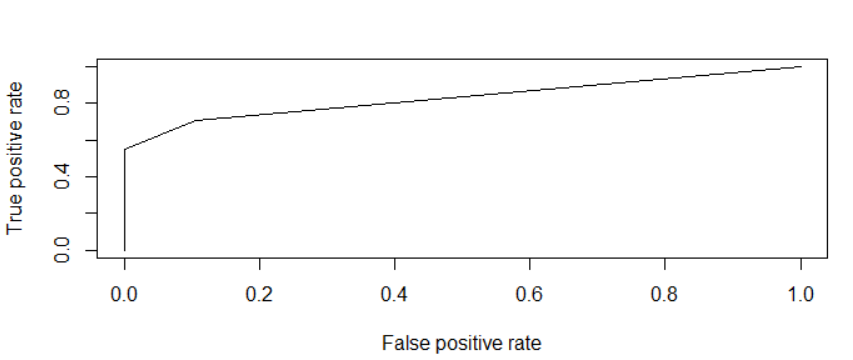
The confusion matrix of testing data set is below:



The accuracy of the model when used on testing set is ~ 96%.

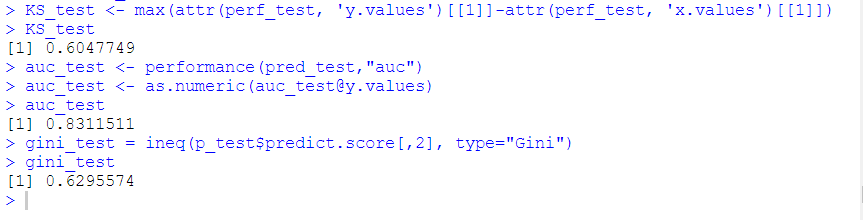
Next, we plot the ROC curve.





## Model Performance

We validate our model performance.

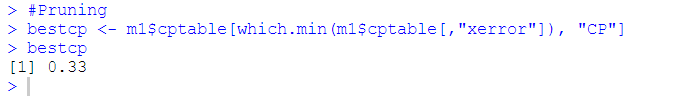


We get decent AUC, KS and GINI values which is showing model is built good.

AUC = 0.83: The model has 83% chance that model will be able to distinguish between customers availing personal loans and customers not availing personal loans.

K-S = 0.60: The degree of separation between the customers availing and not availing personal loans is 60%.

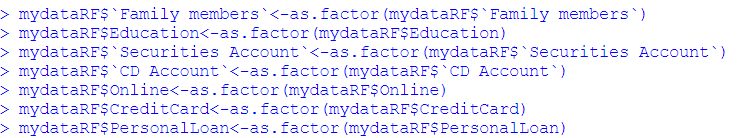
## Pruning



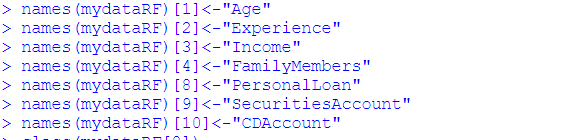
# 5. RANDOM FOREST

We first check the numerical variables and the categorical variables. Numerical variables are Age, Experience, Income, CCAvg, Mortgage. Categorical variables are Family, Education, SecuritiesAccount, CDAccount, Online, Creditcard and PersonalLoan.

We convert all the categorical variables into factors to better create random forest model.

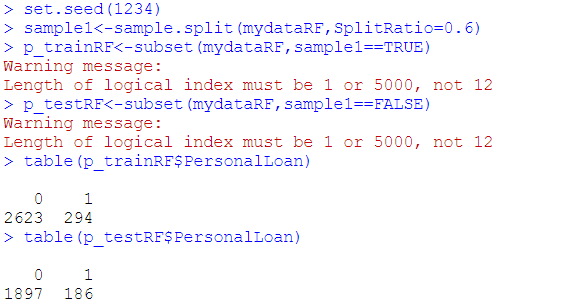


We rename some of the columns like below so as to avoid errors due to special characters and space while running Random Forest.

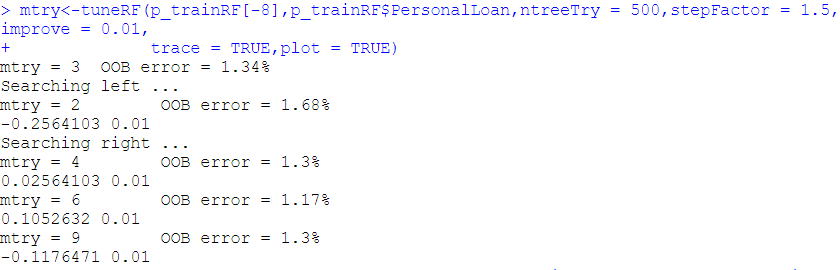


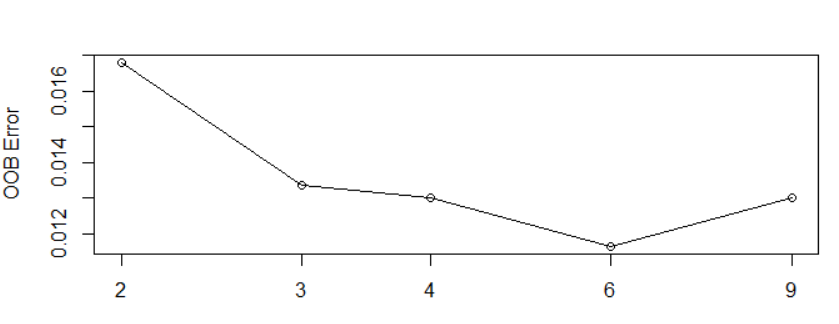
We again split our dataset into training and testing datasets as 60% and 40% respectively.

Number of customers availing Personal Loan in each of the sets are:



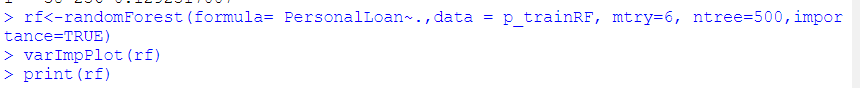
We start the Random Forest by getting the appropriate **mtry** value for creating the model.

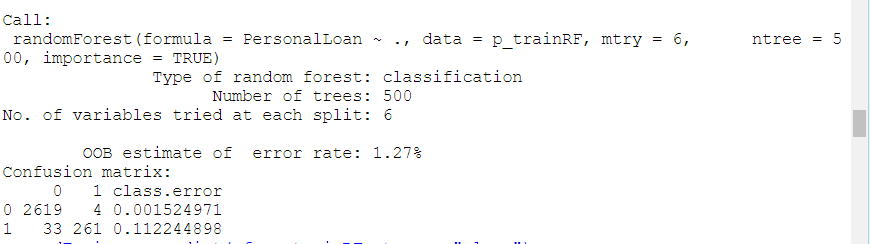




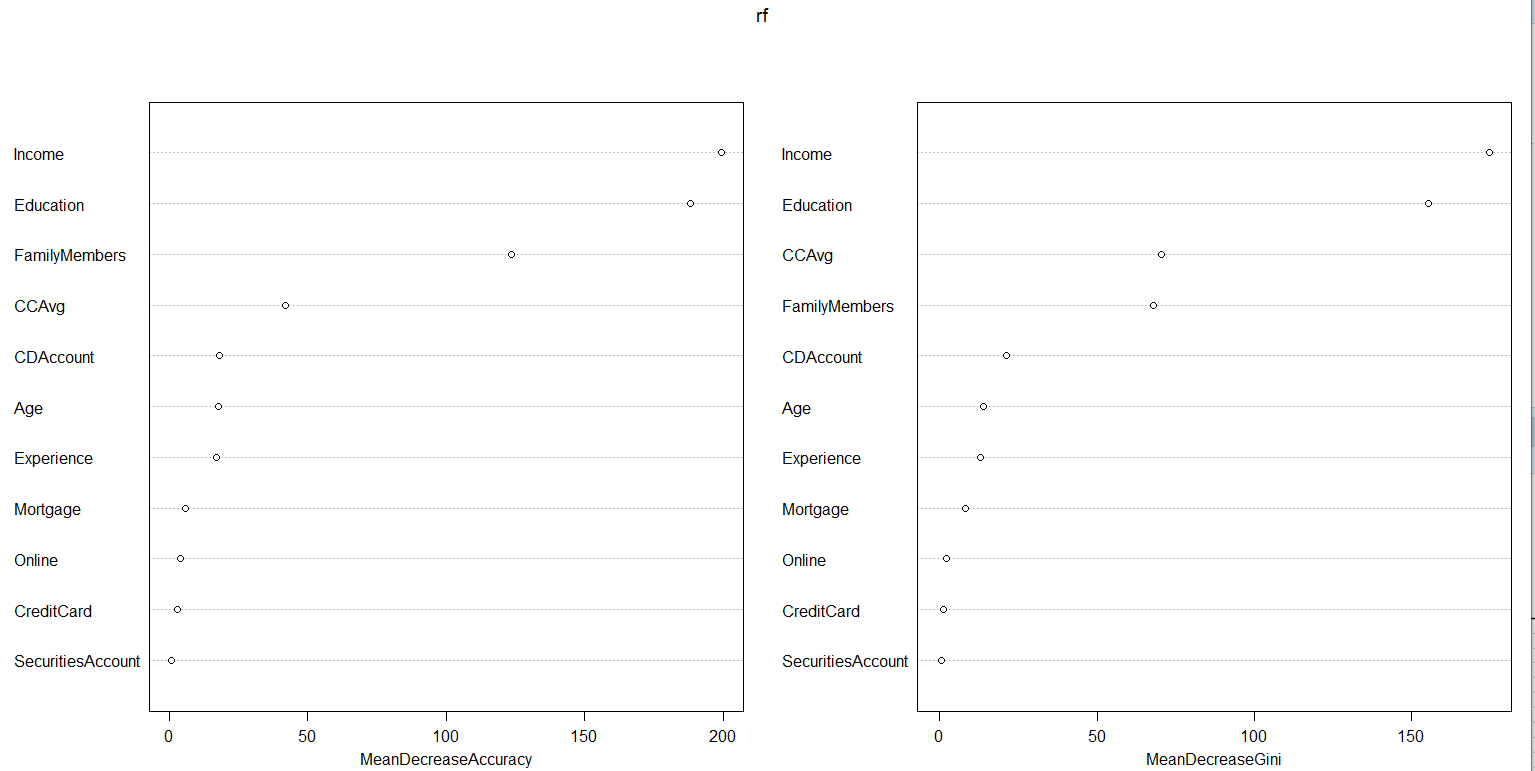
From the console result and resulting select, it is clear for 6 the OOB error is least and so we select **mtry** value as 6 for creating the model.

We create the model and plot the same.



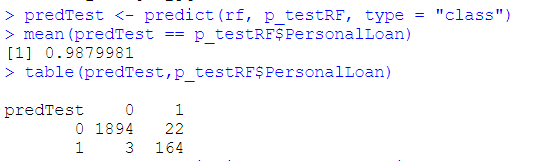


From the above confusion matrix, we can get the accuracy of training set which is around 98.7%.



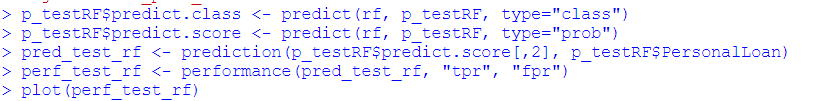
As seen above, Income is the most important factor in availing of Personal Loan, followed by Education and Credit Card spending.

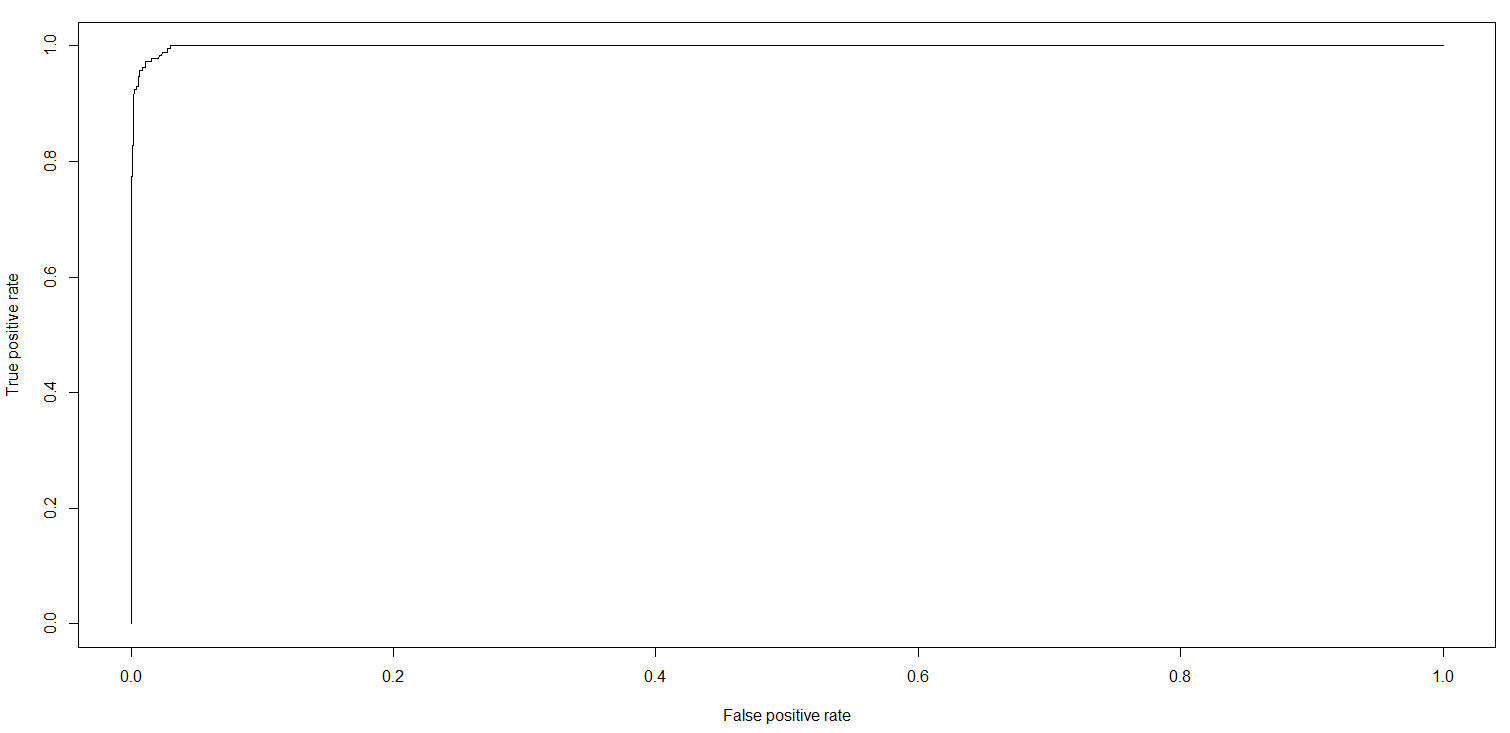
We next check the accuracy of testing set. Confusion matrix is given below.



This accuracy also comes out to be around 98.7%.

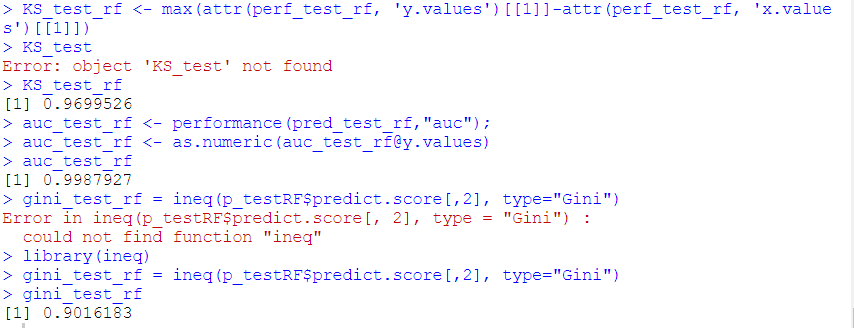
Next, we plot the ROC curve.





## Model Performance

We validate our model performance.



We get fantastic AUC, KS and GINI values which is showing model is exceptionally good.

AUC = 0.99: The model has 99% chance that model will be able to distinguish between customers availing personal loans and customers not availing personal loans.

K-S = 0.96: The degree of separation between the customers availing and not availing personal loans is 96%.

# 6. CONCLUSION

The goal of Thera bank is to convert the liability customers into loan customers.

We built two models – CART and Random Forest. The random forest model works better as the accuracy of the model is 98.7% as compared to CART model. The optimal number of trees has been determined and that also shows this model is better suited to work on our dataset at hand. Thera Bank can use the Random Forest model to predict the customers who are likely to accept the personal loans.

