

Thera Bank loan

DEBORAH OSILADE



Table of Contents

[1. Project Story 1](#_Toc41024100)

[2. Project Objective 1](#_Toc41024101)

[3. Data Description 2](#_Toc41024102)

[4. Exploratory Data Analysis 2](#_Toc41024103)

[4.1. Descriptive Statistics 3](#_Toc41024104)

[4.2. Visual EDA 4](#_Toc41024105)

[4.3. Bivariate analysis 13](#_Toc41024106)

[5. Clustering 14](#_Toc41024107)

# Project Story

This case is about a bank (Thera Bank) which has a growing customer base.

Majority of these customers are depositors with varying size of deposits.

The number of customers who are also borrowers is quite small, and the bank is interested in expanding its loan base rapidly to bring in more loan business and in the process, earn more through the interest on loans.

In particular, the management wants to explore ways of converting its depositors into personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for its depositors showed a healthy conversion rate of over 9% success.

This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with a minimal budget.

What Thera Bank wants

* The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

The dataset has data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

You are brought in as a consultant and your job is to build the best model which can classify the right customers who have a higher probability of purchasing the loan.

# Project Objective

* EDA of the data available. Showcase the results using appropriate graphs
* Apply appropriate clustering on the data and interpret the output (Thera Bank wants to understand what kind of customers exist in their database and hence we need to do customer segmentation).
* Build appropriate models on both the test and train data (CART & Random Forest). Interpret all the model outputs and do the necessary modifications wherever eligible (such as pruning).
* Check the performance of all the models that you have built (test and train). Use all the model Performa.

# Data Description

|  |  |
| --- | --- |
| **Data Description:** |  |
|  |  |
| ID | Customer ID |
| Age | Customer's age in years |
| Experience | Years of professional experience |
| Income | Annual income of the customer ($000) |
| ZIP Code | Home Address ZIP code. |
| Family | Family size of the customer |
| CC Avg | Avg. spending on credit cards per month ($000) |
| Education | Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional |
| Mortgage | Value of house mortgage if any. ($000) |
| Personal Loan | Did this customer accept the personal loan offered in the last campaign? |
| Securities Account | Does the customer have a securities account with the bank? |
| CD Account | Does the customer have a certificate of deposit (CD) account with the bank? |
| Online | Does the customer use internet banking facilities? |
| Credit Card | Does the customer use a credit card issued by the bank? |
|  |  |
|  |  |
| Note: Data is hypothetical |  |

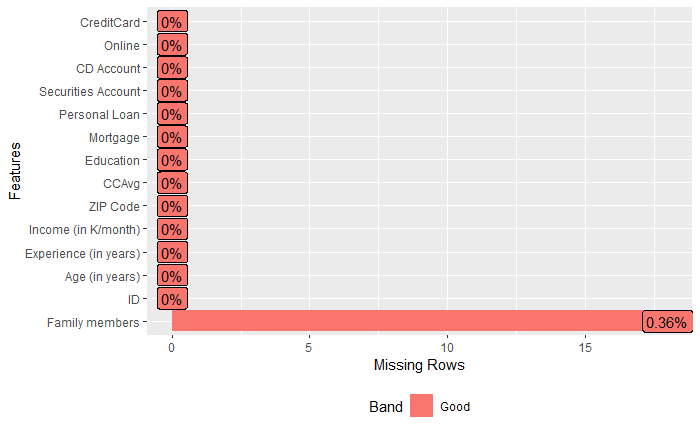
Education has a spilt of 1-3,

|  |  |
| --- | --- |
| 1 | Undergraduate |
| 2 | Graduate |
| 3 | Advanced Professional |

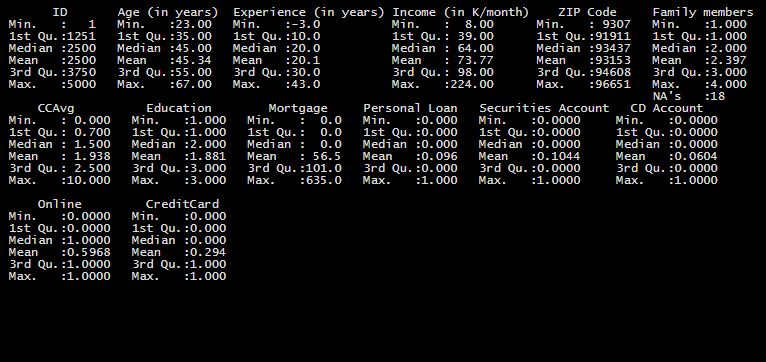
# Exploratory Data Analysis

* Number of Rows = 5000
* Number of Columns = 14
* Grouped into customer demographic - age, income, id, zip code, experience, family, education
* Customer relationship - mortgage, securities account, CD account, CC Avg
* Customer response - personal loan, online, credit card
* The columns Experience in years, Zip Code, Family Members, column does not hold any statistical significance on the data set.
* 480 Customers took personal loan while 4520 didn’t collect loan (find the percentage of customers that did not collect loan since only 9.6% collected loan.
* 1470 Customers have Credit cards while 3530 are without credit cards.
* 522 Customers own a securities account, 4478 have no security account.
* 2984 have online accounts, 2016 of the customers are not online.
* 302 have Certificate of deposit account, 4698 no Certificate of deposits.
* 3462 are without a mortgage plan.

## Descriptive Statistics



* Shows family members has NA (missing variables)



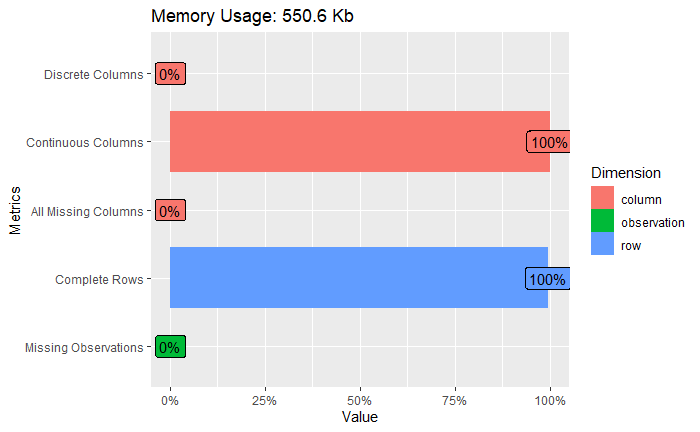
Summary () shows that 5 out of 14 variables are in the same range of 0-1 and this falls under customer relationship and response, and the remaining variables have a different maximum and minimum value. Hence scaling could be required to make all levels the same using its standard deviations.

* There are missing values in the dataset. However, missing values comes from an insignificant variable family member, and is not significant to the dataset.
* ID variable shows 1-5000, signifying that up to 5000 customers have been recorded in the Thera Bank dataset.
* Age indicates that the bank customers are between the age of 23-67 years, 75% of customers are aged and 25% of customers are middle-aged customers.
* Experience in years shows a -3 minimum which could be an error as it is impossible to have a negative experience in years.
* 25% of Thera customers are on a mortgage plan, with the value of the house above $101,000.
* The bank records customers deposits for income per month between $8,000 to $224k and the average income earned is $73,770.
* Thera bank customers have a maximum average spending on credit cards of $10,000 per month.

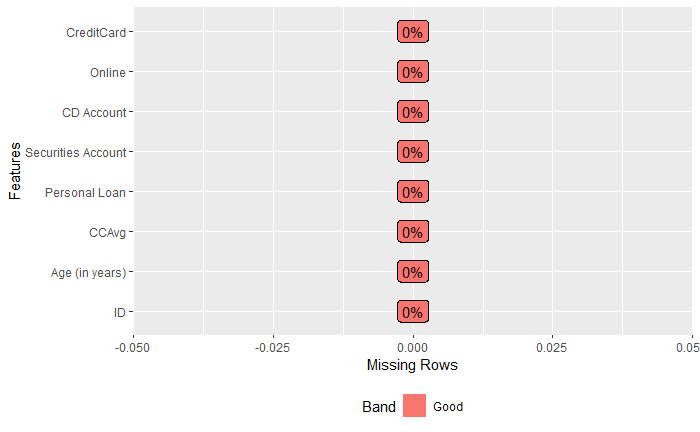
## Visual EDA



* Removed Column names were insignificant to the dataset.
* Experience in years, Zip Code, Family Members, column does not hold any statistical significance on the data set

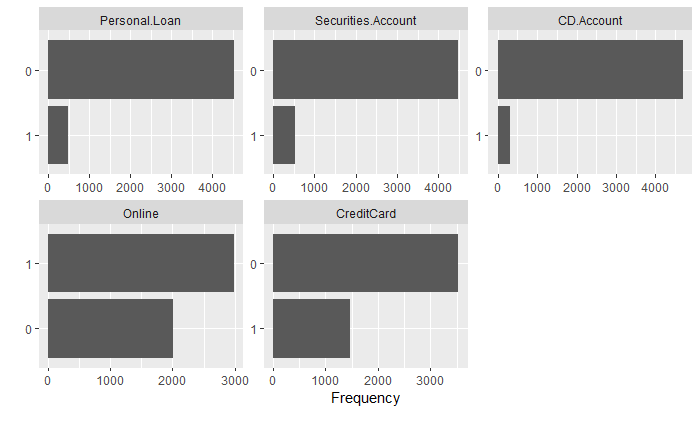


* All rows and columns are complete
* Shows the absence of missing observations, columns and no discrete columns.



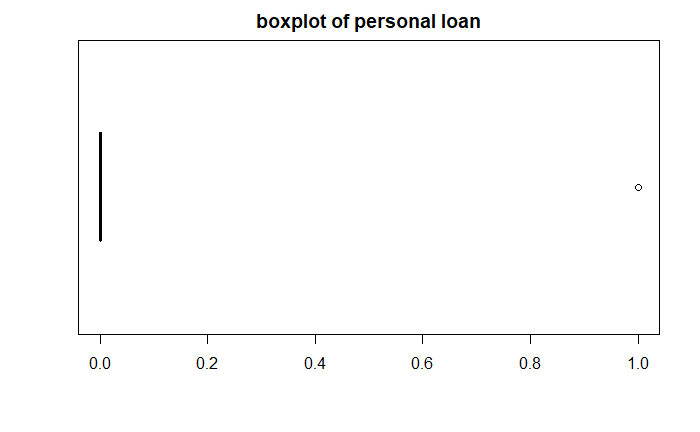
* Shows no missing values in the dataset after removing insignificant columns

### Bar plot

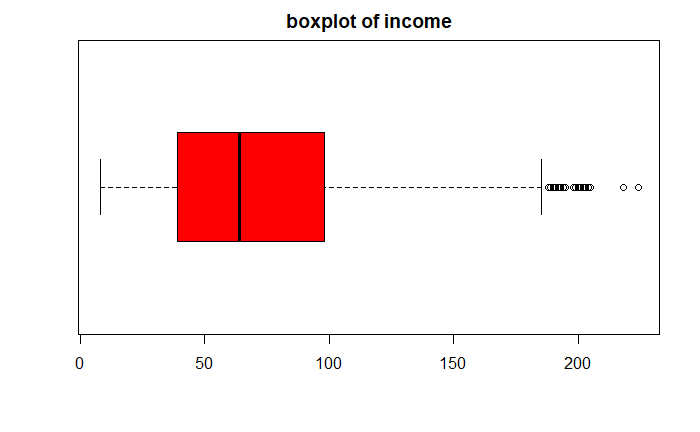


* Personal Loan, Online, Credit card, securities Account, CD Account, could be used as categorical to group those with and without a loan, a card, or an account. This is records as 0 and 1. 0 for No and 1 for yes.

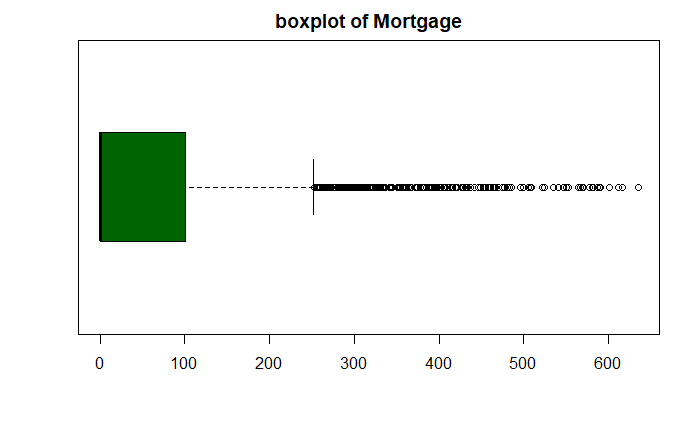
### Boxplot

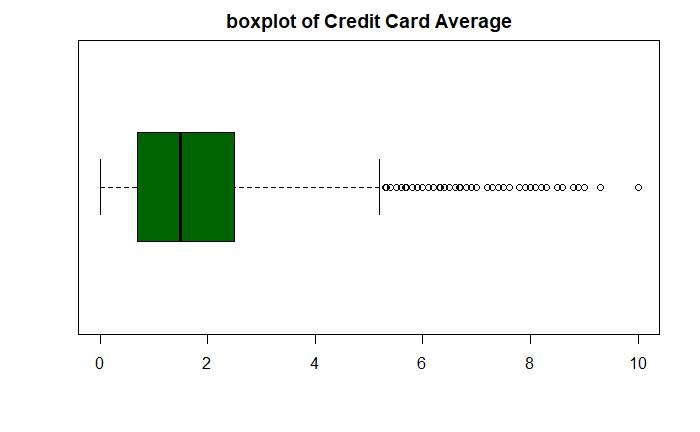


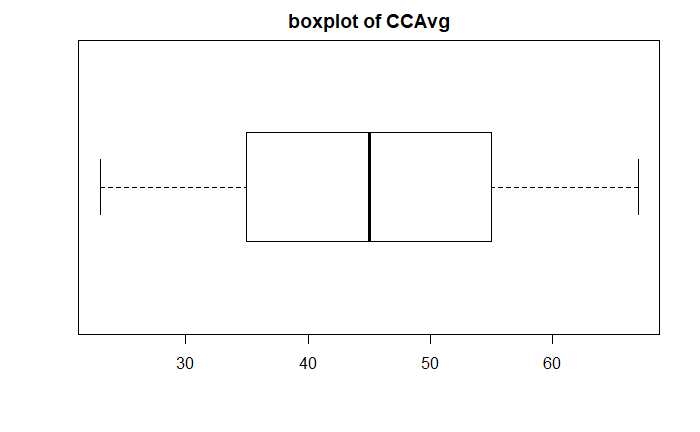
* Shows and outlier of 1, which represent customers that responded to the campaign and took personal loans. Which is more categorical than numeric.
* Target Variable in the dataset.



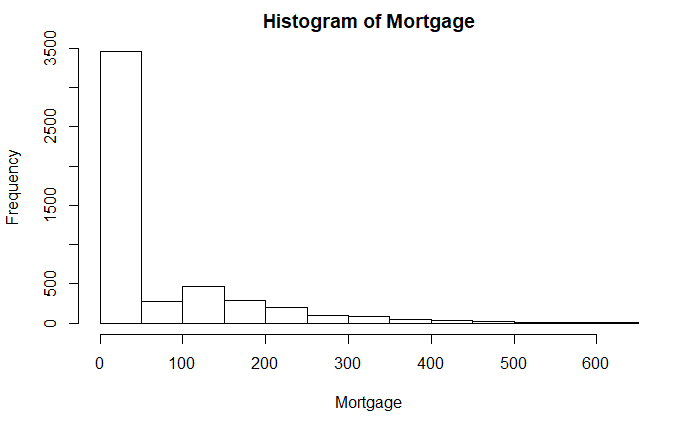
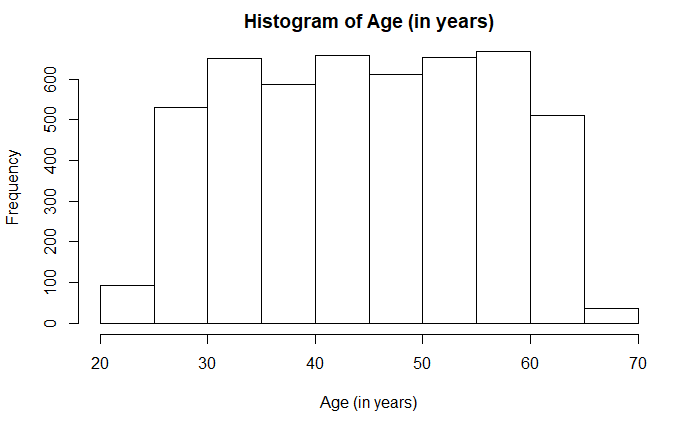
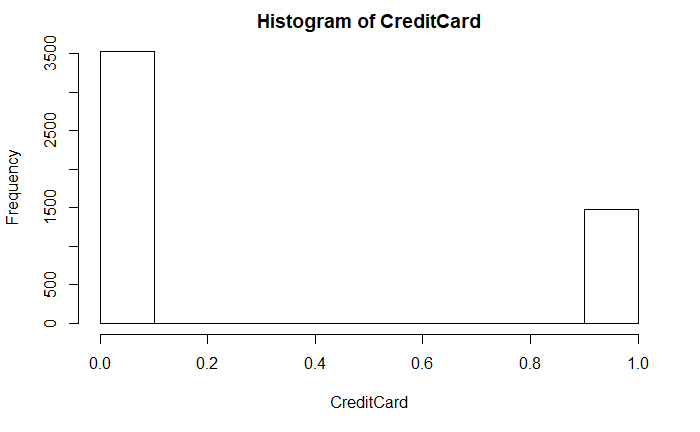
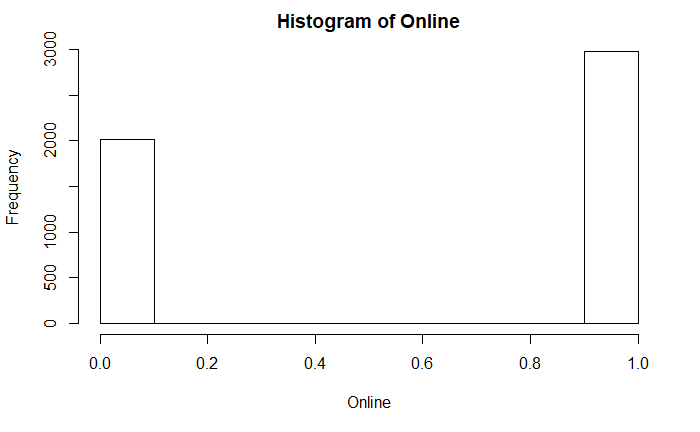
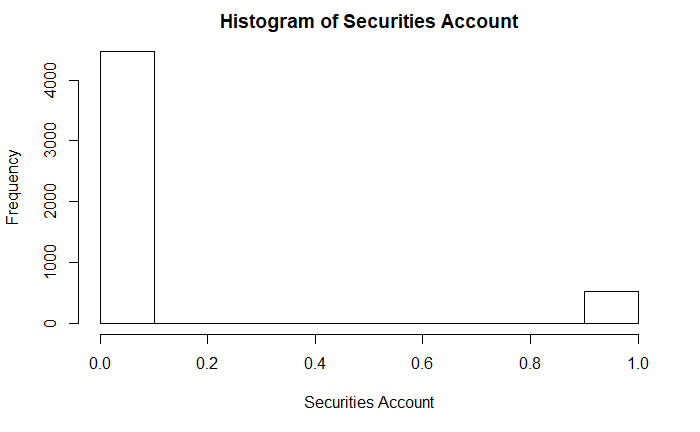
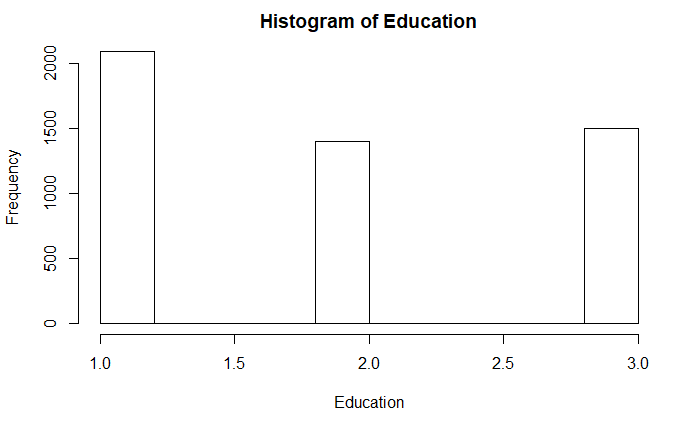
* Maximum value excluding outlier fall under $180k, with outliers the maximum income is $224k.
* 25% of income is less than $39k.
* 50% of customers earn above $64k.

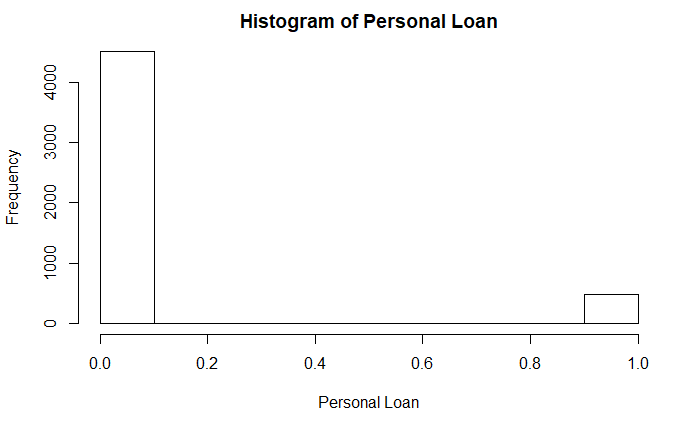
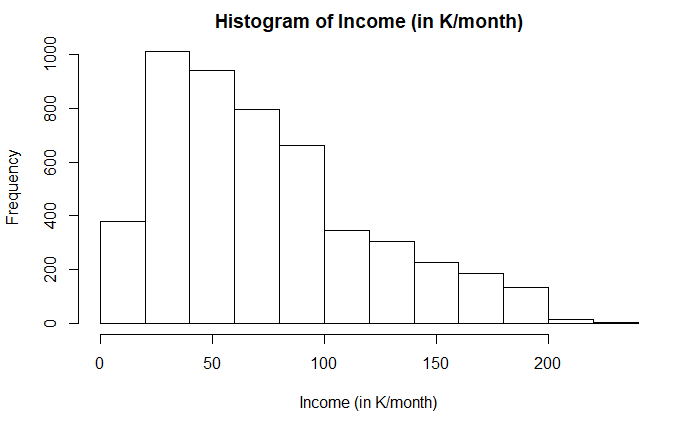
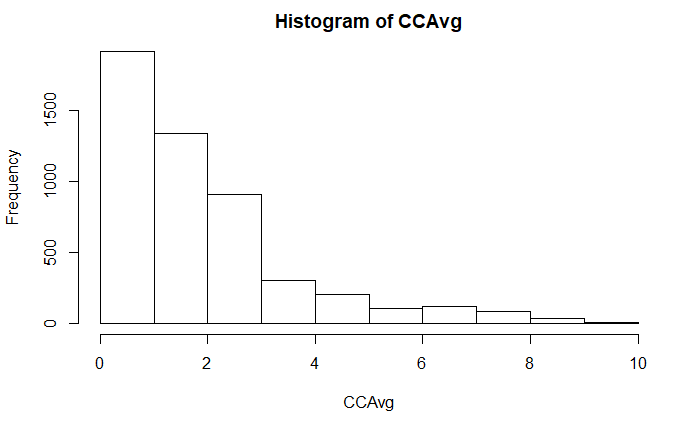






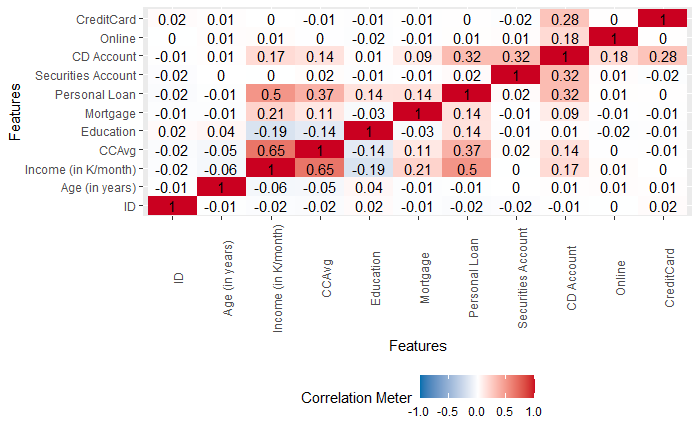
Histogram





## Bivariate analysis

* Correlation



* Income in K month and CCAvg = 0
* Personal Loan and Income in K month = 3.560296e-318
* Securities Account and CD Account = 3.859242e-117
* Personal Loan and CD Account = 1.278403e-116

Income, cc avg, have p.valuess << 0.001 indicating these correlations are significant

# Clustering

K- Clustering – Deals efficiently with large data sets.

* Set the seed – Seed was set to 1000, to ensure k-means clustering results are reproducible.
* The function NbClust() to determine optimal number of clusters

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Among all indices:

\* 7 proposed 2 as the best number of clusters

\* 2 proposed 3 as the best number of clusters

\* 13 proposed 4 as the best number of clusters

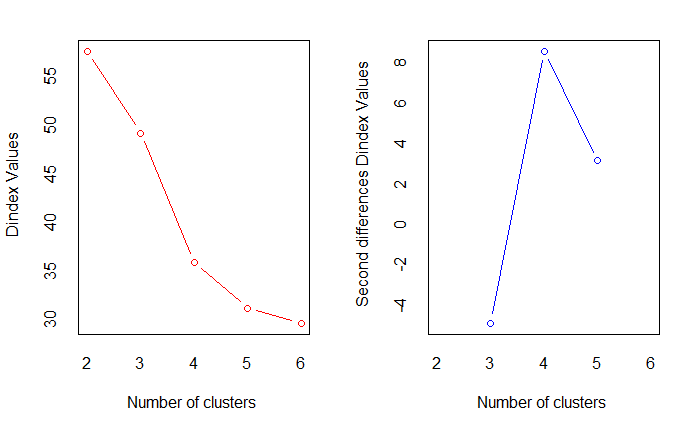
\* 2 proposed 6 as the best number of clusters

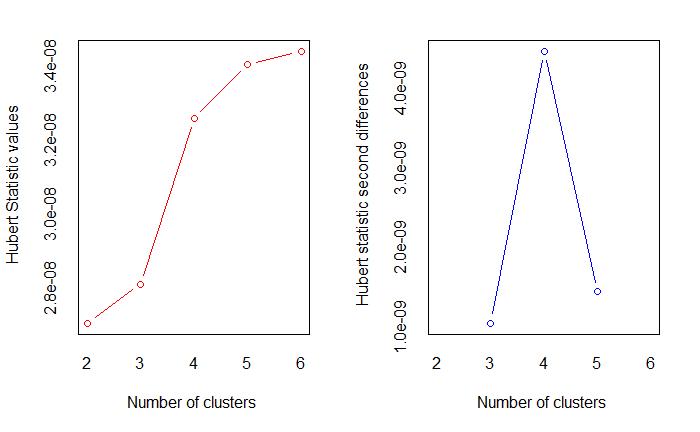
\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 4

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

* 13 indices indicate that optimal number of clusters is 4 and 2 indicate that the optimal number of clusters is 3.
* The value of K getting the highest number of votes is 4

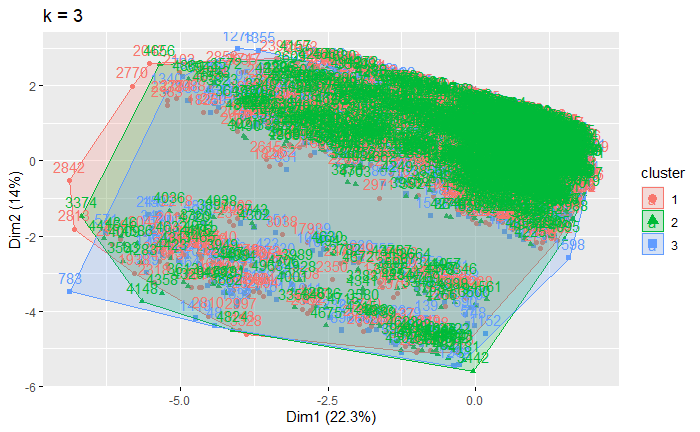


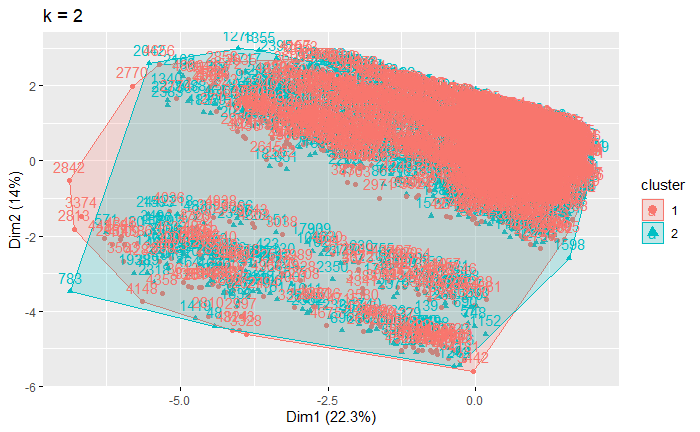


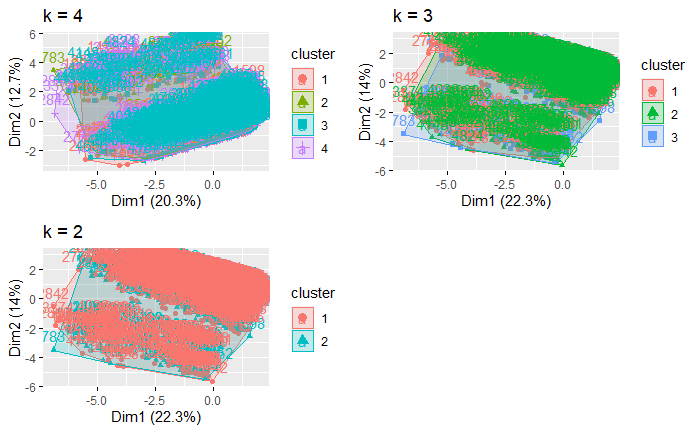
Options for optimum number of clusters

* We will use k = 4, 3, 2 and analyse them to find optimal k for Thera bank loan









Clust 4



Clust 3



Clust 2



K = 4 gives 4 clusters, out of which cluster 2 is a cluster of a slightly high no of customers compared to other clusters and the 4th cluster the lowest group of customers. While the remaining two others clusters have similar customer. Cluster 2 has customers with a higher income parameter and the highest group for almost all parameters.

K=3 shows cluster 1 and 3 as clusters with the same customer no, however the 3-cluster group shows a higher income than the remaining clusters and slightly high values in a few parameters.

K=2 Shows a group of 2 clusters with the second cluster showing the high values in all parameters

Cluster 1 = shows customers with high mortgage, good income and a satisfactory level of customer relationships and response.

Cluster 2 = Have a high credit card spending average, an outstanding income earner, they show and impressive customer response to purchasing a loan

# Decision Trees

n= 3500

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 3500 336 0 (0.90400000 0.09600000)

2) Income..in.K.month.< 113.5 2804 57 0 (0.97967190 0.02032810)

4) CCAvg< 2.95 2599 9 0 (0.99653713 0.00346287) \*

5) CCAvg>=2.95 205 48 0 (0.76585366 0.23414634)

10) CD.Account< 0.5 185 33 0 (0.82162162 0.17837838)

20) Income..in.K.month.< 92.5 119 10 0 (0.91596639 0.08403361) \*

21) Income..in.K.month.>=92.5 66 23 0 (0.65151515 0.34848485)

42) Education< 1.5 38 4 0 (0.89473684 0.10526316) \*

43) Education>=1.5 28 9 1 (0.32142857 0.67857143) \*

11) CD.Account>=0.5 20 5 1 (0.25000000 0.75000000) \*

3) Income..in.K.month.>=113.5 696 279 0 (0.59913793 0.40086207)

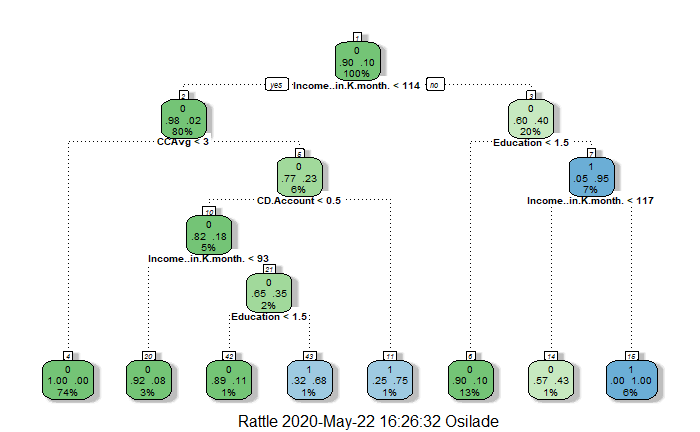
6) Education< 1.5 451 47 0 (0.89578714 0.10421286) \*

7) Education>=1.5 245 13 1 (0.05306122 0.94693878)

14) Income..in.K.month.< 116.5 23 10 0 (0.56521739 0.43478261) \*

15) Income..in.K.month.>=116.5 222 0 1 (0.00000000 1.00000000) \*

* Output 0 denotes customers without loans and 1 for customers with a personal loan



Interpreting CART.

* First Node shows that there are 10% of chance of being a personal loan potential customer and 90% of customers yet to stand a chance.
* Income per month is the most important variable while building this strategy as the first variable from the split.
* Customers with Income greater than $114k, who are graduates have better chances of being potential customers.
* Customers who are Graduates and earn above $117k per month have the highest probability of purchasing the loan customers.
* If income is less than 114 then there is a 2% probability of being a potential customer, hence if the customer has an average spending limit of less than $3k then the customers loses all the chance of purchasing a loan.
* A customer with a certificate of deposit has a 75% chance of being a potential customer.
* If a customer is without a deposit certificate, then is the income greater than $93k, hence chances increases to 35%.
* 74% stands no chance of purchasing a loan.

Strategy that can be implemented to increase purchasing rate

* Accept customers who earn above $114k.
* Validate the customer is a graduate, who probably earns above $117k as the best candidate and less than $117k -$114k as potential customer.
* Accept customers application whose credit average spending is above $3k and has a certificate of deposit.
* Consider customers who has no certificate of deposit but has an average spending above $3k, earns above $93K and is a graduate.

# Random Forest

OOB estimate of error rate: 2.97%

### Importance

Optimum number of random trees

0 1 MeanDecreaseAccuracy MeanDecreaseGini

ID -2.2868421 -3.266536 -3.723085 26.016102

Age..in.years. 15.7716589 -1.841356 13.097479 25.878383

Income..in.K.month. 107.5447465 87.075101 128.629599 200.560399

CCAvg 26.7016701 27.807912 31.880226 94.193727

Education 121.2455437 65.489332 124.742757 137.786876

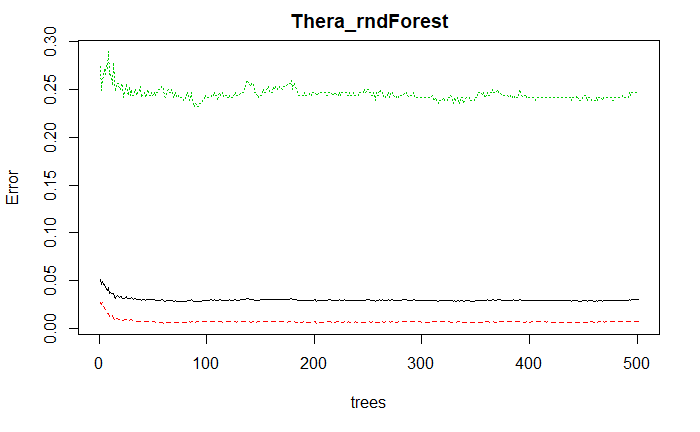
Mortgage 8.8878935 -1.244292 7.879358 20.226742

Securities.Account 2.1950136 1.156991 2.675688 2.646631

CD.Account 16.8944348 20.043575 24.461208 36.289189

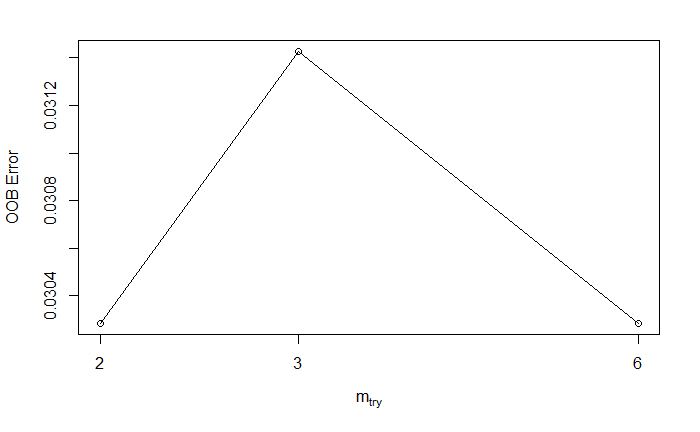
Online -0.2723488 3.187130 1.423903 3.766752

CreditCard 9.1448344 2.146020 9.033291 4.921453



### Tune random

* mtry = 3 OOB error = 3.03%
* mtry = 2 OOB error = 3.06% -0.009433962 TRUE
* mtry = 6 OOB error = 3.03%



### Refined random forest

Out of bag error rate is 3%

OOB estimate of error rate: 3%

Confusion matrix:

0 1 class.error

0 3139 25 0.007901391

1 80 256 0.238095238

### Confusion matrix

Thera\_random\_predict

0 1

0 3139 25

1 80 256

Thera\_test\_prediction

0 1

0 1351 5

1 37 107

From the train data

* 3139 was predicted correctly and 25 was predicted correctly as 1 and its false
* 80 was predicted incorrectly as 0 and 256 was predicted incorrectly and its true
* Accuracy = 3139+256 = 3,395/3500 = 0.97
* Classification error rate = 25+80 =105/ 3500 =0.03
* True positive = 3139/3139 +80 =81
* True Negative = 256/256+80 = 81

From the test data

* Out of 1388, 1351 was predicted correctly and its true and 5 was predicted correctly and its false
* 37 was predicted negative and its false while 107 was predicted incorrectly and its true

### Error Rate

* Train data - 0.1645714
* Test data - 0.02933333

### Accuracy

* Train - 0.8354286
* Test - 0.9706667