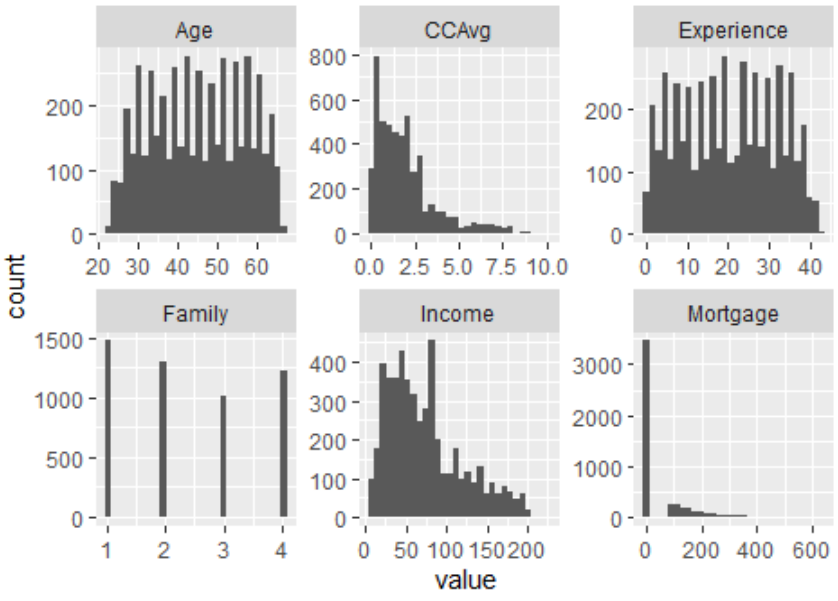
Data Description

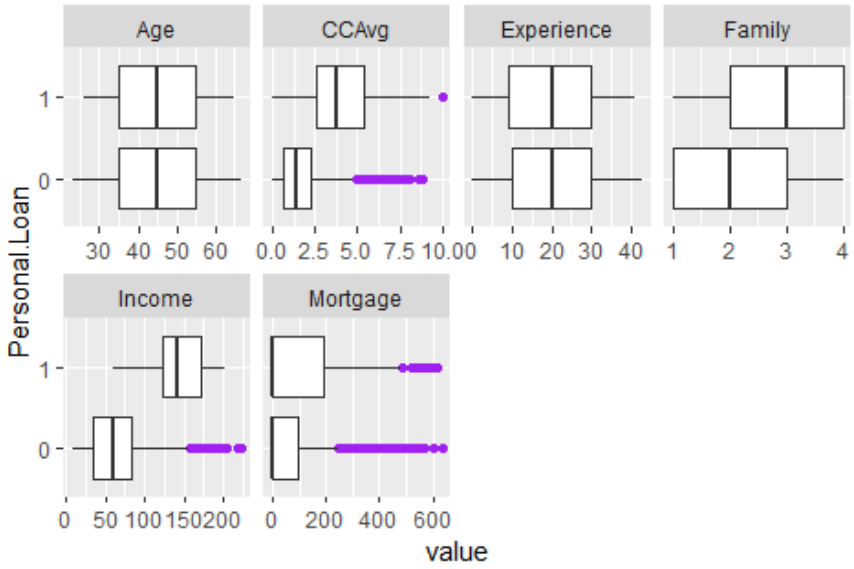
ID: Customer ID (unique)
Age: Customer's age
Experience: Number of years of professional experience
Income: Annual income of the customer in $1, 000
ZIP Code: Customer's home address ZIP code
Family: Number of customer's family member
CCAvg: Customer's average spending on credit cards per month in $1,000
Education: Customer's education Level (1: Undergrad; 2: Graduate; 3: Advanced/Professional)
Mortgage: The value of house mortgage in $1,000
Personal Loan: If this customer accept the personal loan offered in the last campaign? (1 - Yes, 0 - No)
Securities Account: Does the customer have a securities account with the bank? (1 - Yes, 0 - No)
CD Account: Does the customer have a certificate of deposit (CD) account with the bank? (1 - Yes, 0 - No)
Online: Does the customer use internet banking facilities? (1 - Yes, 0 - No)
CreditCard: Does the customer use a credit card issued by this Bank? (1 - Yes, 0 - No)

Exploratory data analysis

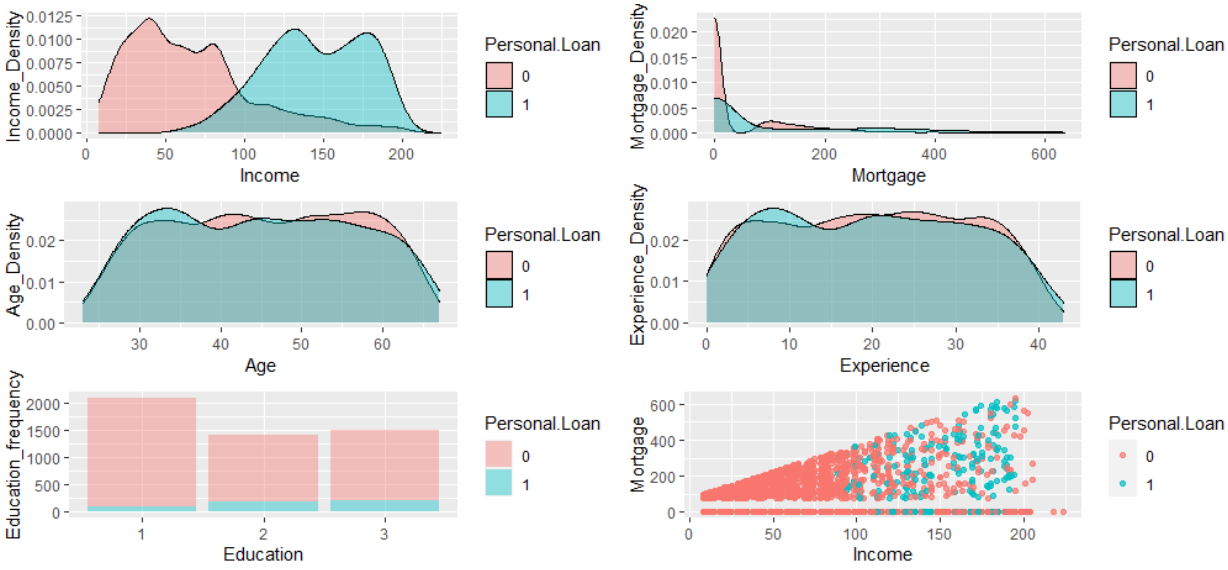
1. The following graph shows the count of customers in various categories as divided by variables like age, average spending on credit cards/month, no of years of professional experience, family size of the customer, annual income of customer and value of house mortgage if any.



2. **Explored box plots to explain the relationship between personal loan and other factors with the purple colored values indicating outliers.**



3. **Density plots across various attributes and by Personal Loan availed or not**



4. **Frequency plots across various attributes and by Personal Loan availed or not**

Employed Methods and Analysis Results

Two types of analysis were performed:

Consumer segmentation

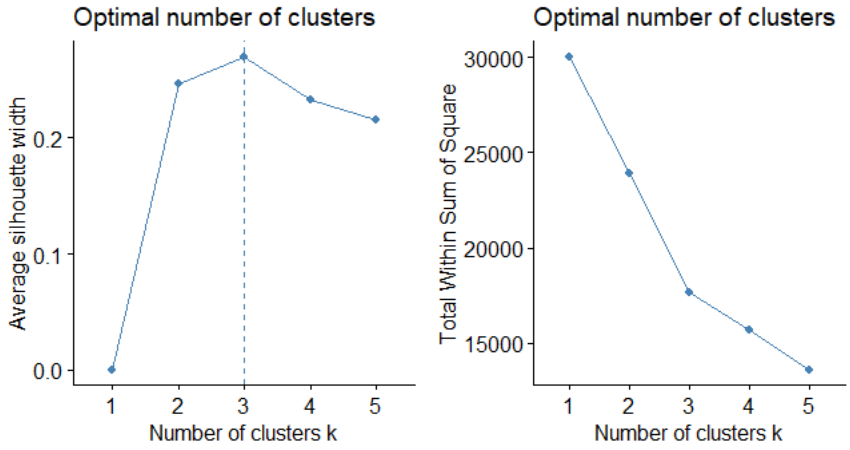
Method used: k-means clustering

1) **Classification**

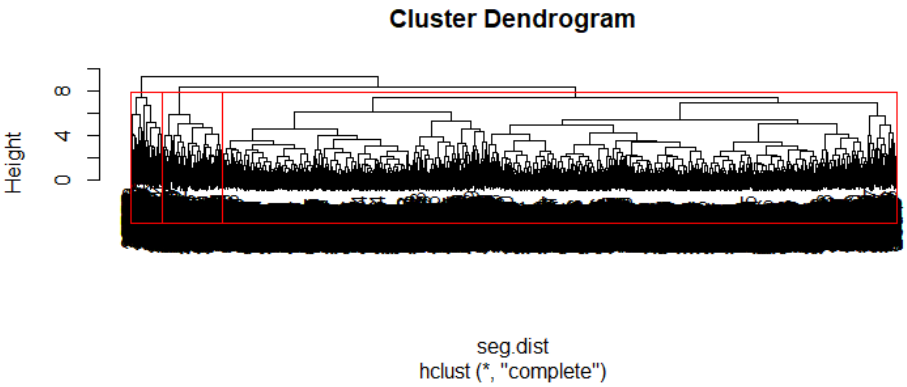
Methods used: a) Logistic Regression
b) Decision Tree
c) Random Forest

Consumer Segmentation:

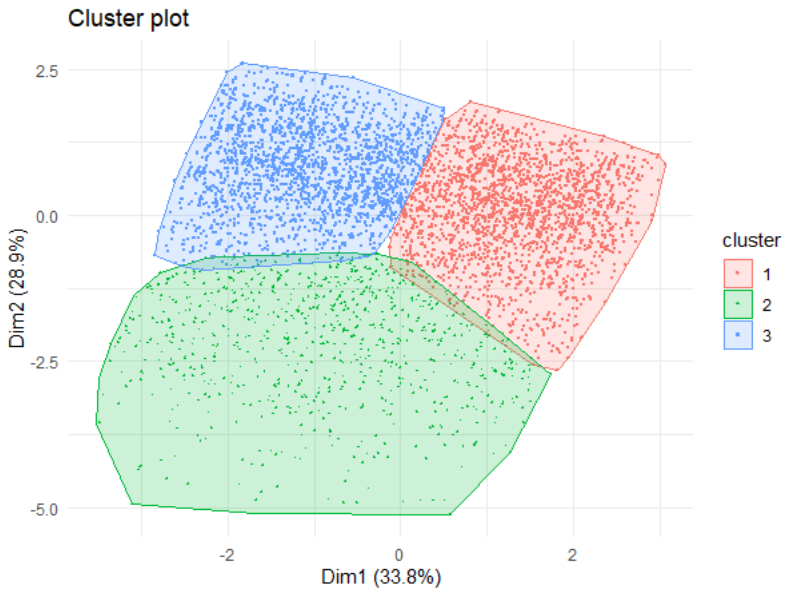
Firstly, numerical columns are scaled, then package “factoextra” is installed, it provides some easy-to-use functions to extract and visualize the output of multivariate data analyses". To determine and visualize the optimal number of clusters, function “fviz_nbclust()” is used. Optimal number of clusters were found using two different methods: within cluster sums of squares and average silhouette. Visualization can be seen below for both the methods:



Both methods point to 3 being the optimal number of clusters. As an additional check, hierarchical clustering is also performed and the resulting Dendrogram is checked for 3 clusters. Visualization can be seen below:



Next, k-means clustering is implemented by specifying to generate 3 clusters . Resulting clustering plot can be visualized as seen below:



After aggregating all the attributes with respect to each cluster, following results are seen

```
> seg.summ(seg.df.num, df.k$cluster)
  Group.1      Age Experience    Income   Family    CCAvg Education  Mortgage Personal.Loan Securities.Account CD.Account    Online CreditCard
1       1 55.52952  30.235704  58.24733 2.380753 1.354891  1.944677  44.73175    0.03765690          0.1032078 0.04416550 0.6071595  0.2993956
2       2 43.69212  18.682578 146.44869 1.905728 4.792709  1.594272 112.81623    0.38663484          0.1062053 0.15393795 0.6109785  0.3007160
3       3 35.12382   9.935356  60.09796 2.617603 1.371969  1.932372  45.61711    0.03729488          0.1049229 0.03878667 0.5798110  0.2854301
```

Cluster 2 seems to be the segment which is most likely to avail a personal loan. This segment belongs to middle-aged people, belonging to high income category, small family size, high credit card spending, mortgage value of house being high and high probability of having Deposit account(CD.Account)

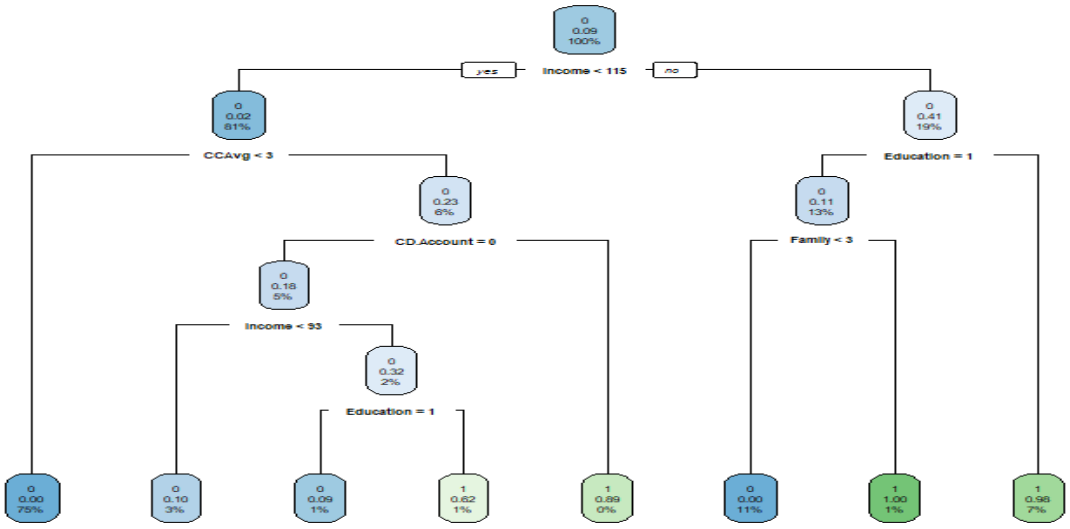
Classification:

Cart Model:

To start with, the dataset is split randomly into training dataset and test dataset with 80% for training and 20% left for testing. Next a decision tree model is fitted on the training dataset, using “rpart” package. Variable importance is checked for and as seen below, Education appears to be the most important, followed by income, family size and credit card spending

```
> dt$variable.importance
 Education     Income     Family      CCAvg CD.Account   Mortgage        Age     Online Experience
273.685516 192.583376 171.222498 101.225326  64.373342  23.355176   3.390302   3.130238   3.015400
```

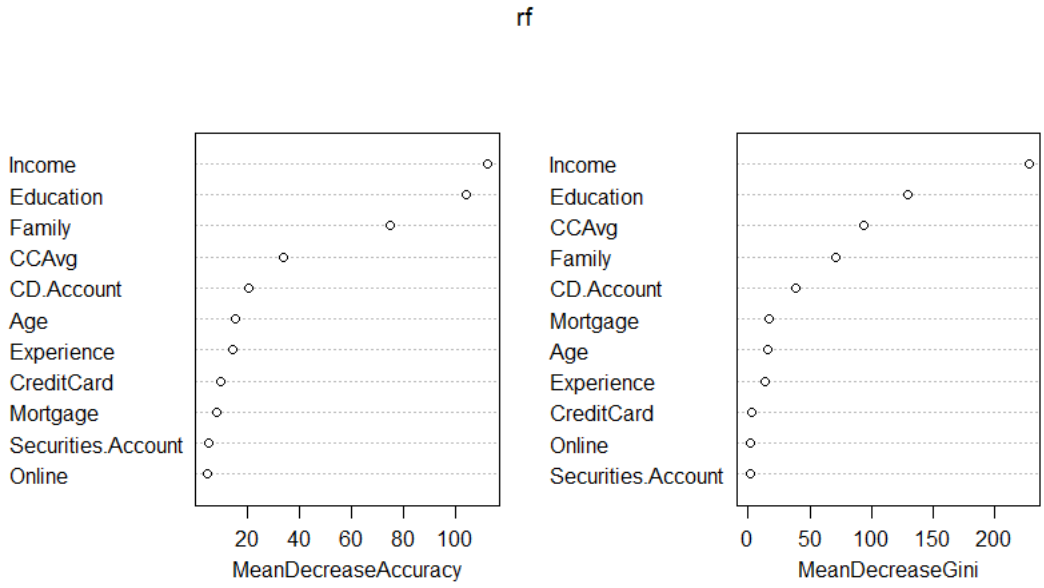
The tree plot can be seen below:



After using the fitted model on test data set, it can be seen that an accuracy of 98.4% is achieved. The confusion matrix is displayed below:

```
  pred.cart
     0   1
0  892   7
1    9  92
```

Random Forest Method: To see if better results, Random Forest method has been deployed where 100 trees are utilized for the bootstrapping purpose with minimum size of terminal nodes to be 10 and the importance of each variable has also been checked as seen below



An accuracy of 98.1% is achieved with Random forest model.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 896   3
         1  16  85
```

After comparison, it can be seen that Decision tree (Cart) model is better in terms of prediction.

Actionable Marketing Implications

The bank can effectively target the segment which is well educated, earning high income, with strong credit card spending history for selling more personal loans. The bank should think of more innovative schemes to attract elderly and young people with not so high incomes and spending history. These are the people who earn between 40000$ to $100,000 with credit card spending less than 3000$ per month. Bank may try to deliver loan schemes with low interest to attract these people, and in the case of elderly have dedicated staff to help senior citizens avail a loan account in a smooth manner. In the case of young people, bank should think of potential spending on attractive marketing campaigns online, specifically targeting online places where young people are likely to engage mostly. Banks can also engage HR teams of companies and promote the loan schemes to young people working in the company.