LeTicia Cancel

DATA622 Homework #4

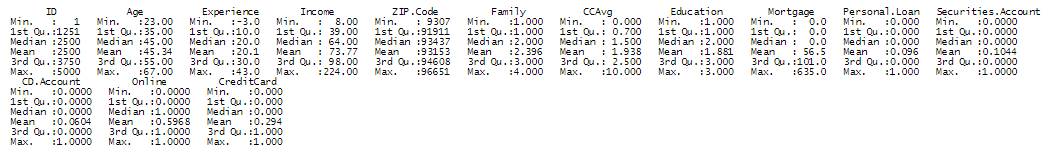
December 22, 2022

The dataset for this assignment is the [Bank Loan Modelling](https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling/code?datasetId=48024&language=R) data from Kaggle. The data consists of customer information for a campaign to get customers to open bank loans. The goal is to create a model to identify customers who have a higher probability of opening a loan.

Data column descriptions

|  |  |
| --- | --- |
| Age | Customer's age in completed years |
| Experience | #years of professional experience |
| Income | Annual income of the customer ($000) |
| Zip | Home Address ZIP code. |
| Family | Family size of the customer |
| CCAvg | 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? |
| CeditCard | Does the customer use a credit card issued by UniversalBank? |

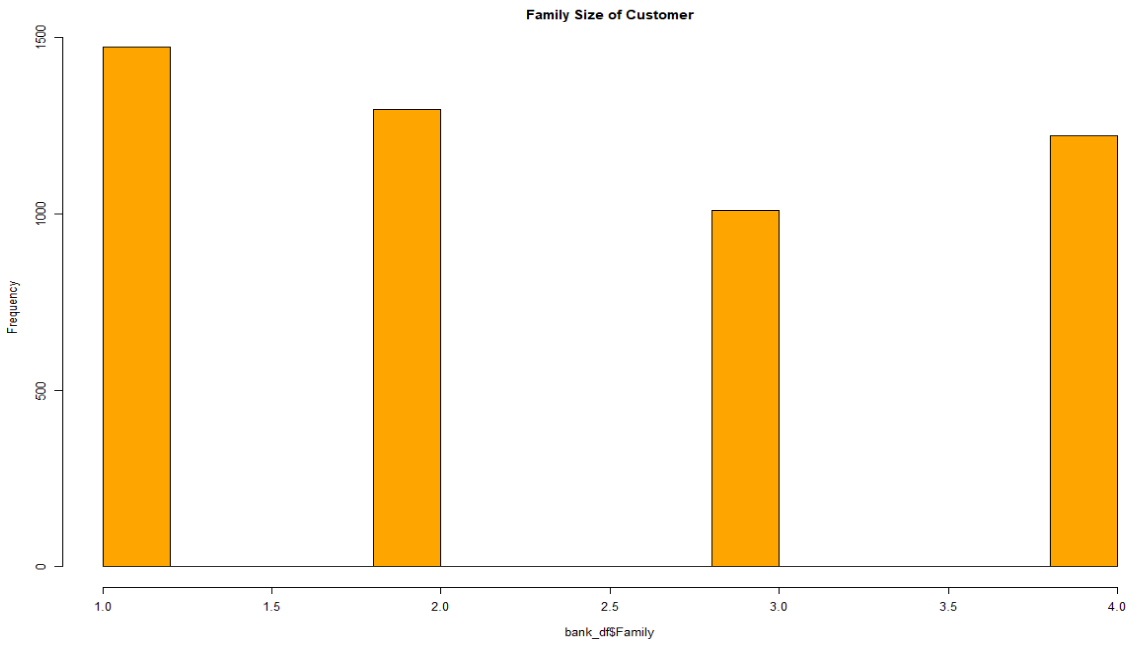
Summary of the dataframe. There are no N/A values

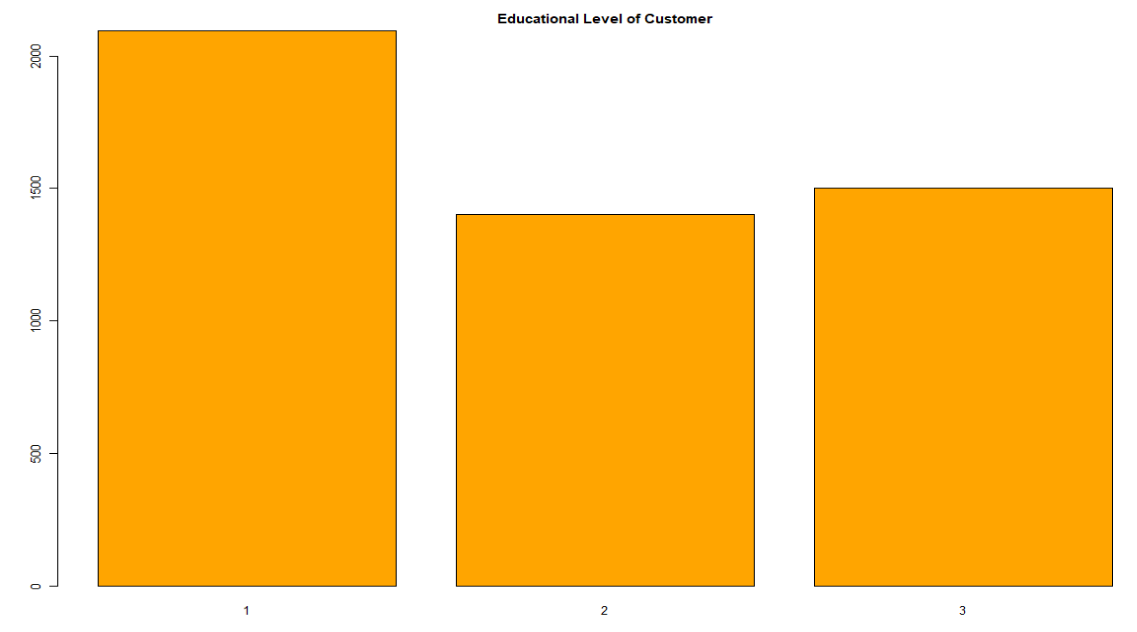


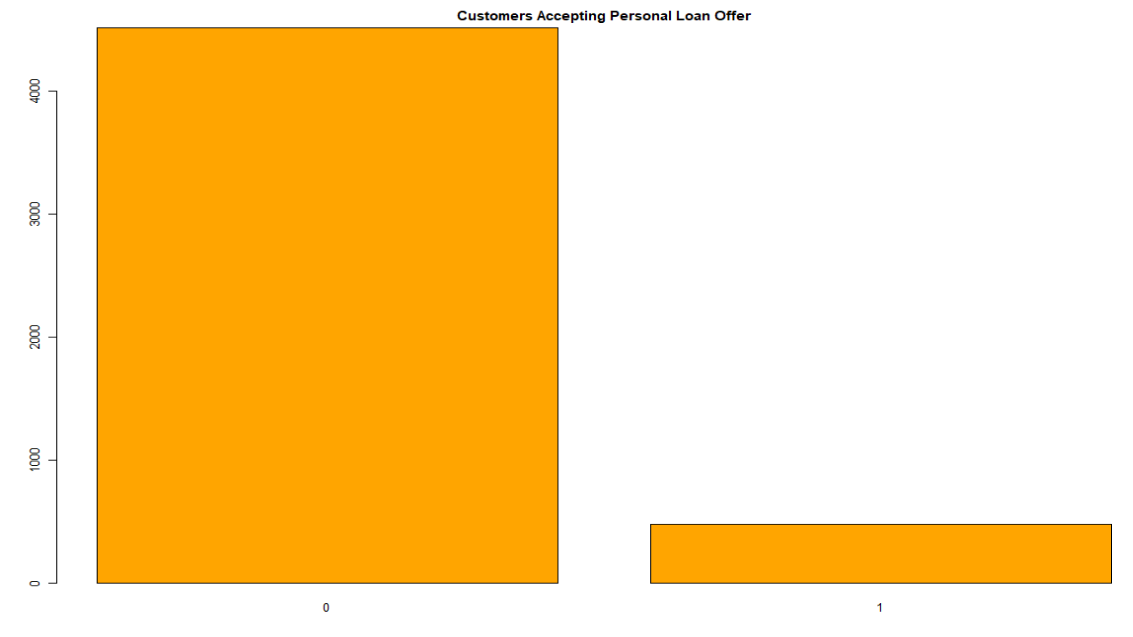
Convert categorical and bit columns to Factor:

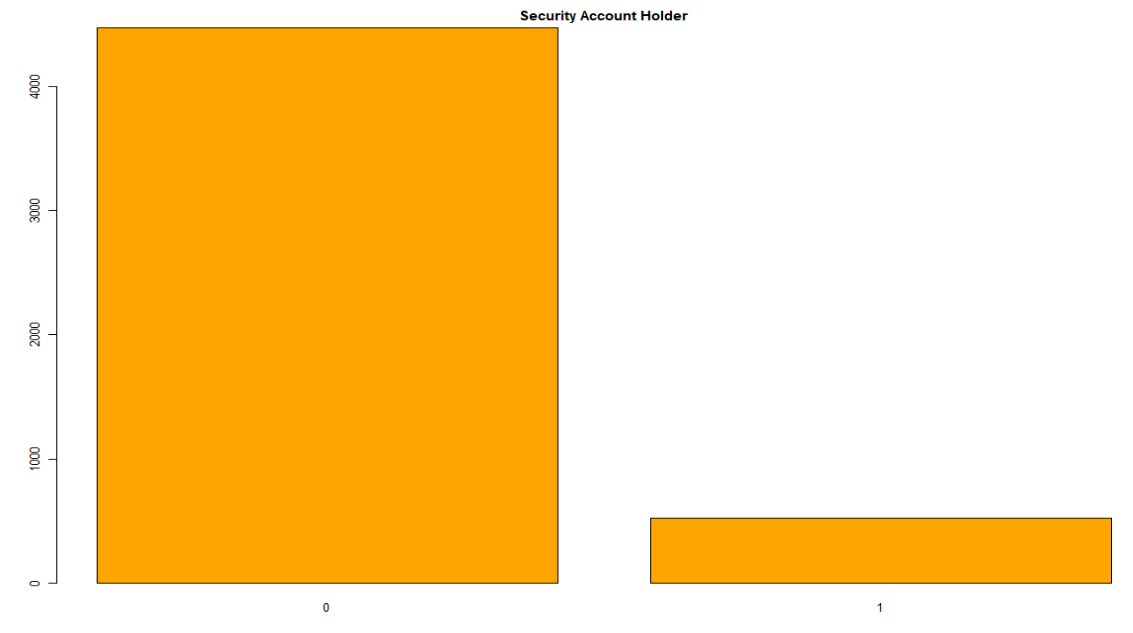
* Education
* Personal Loan
* Securities Account
* CD Account
* Online
* CreditCard

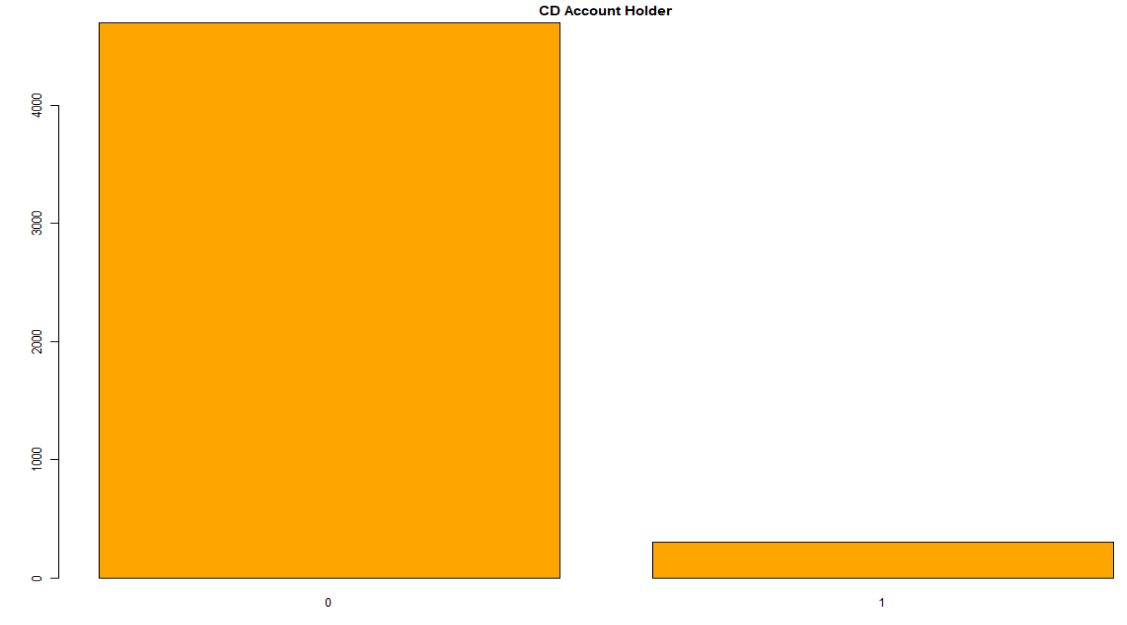
Exploring the columns

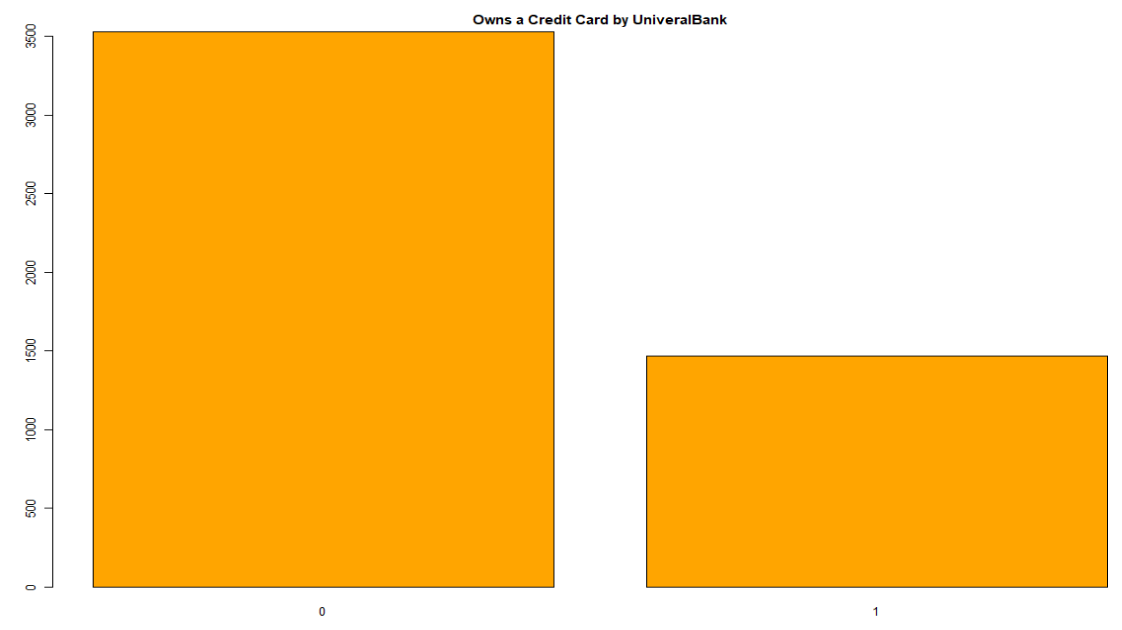






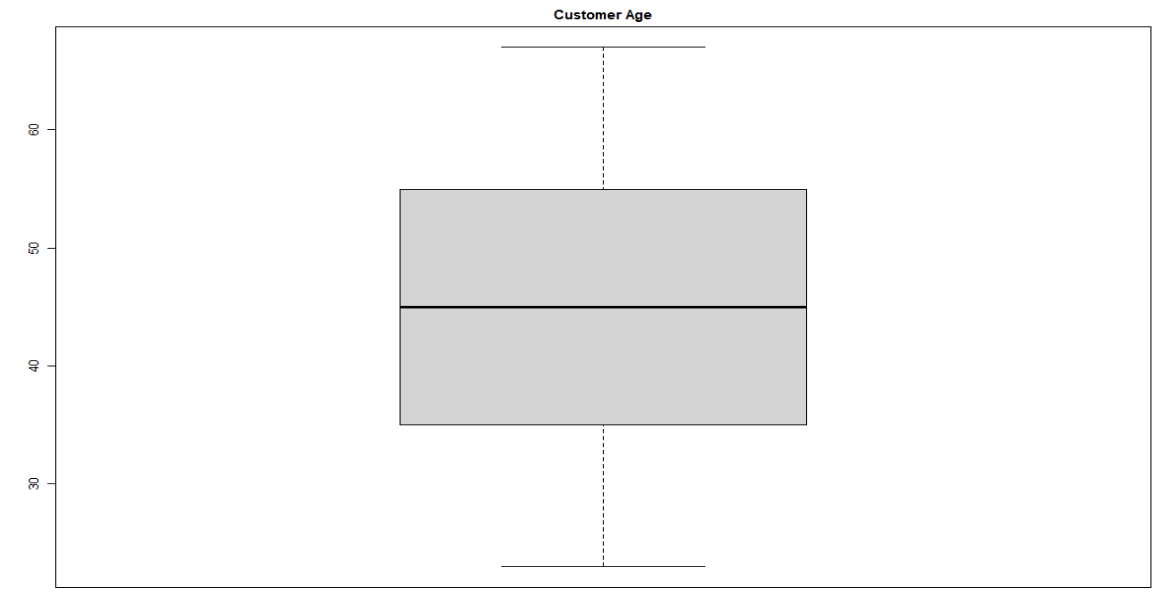


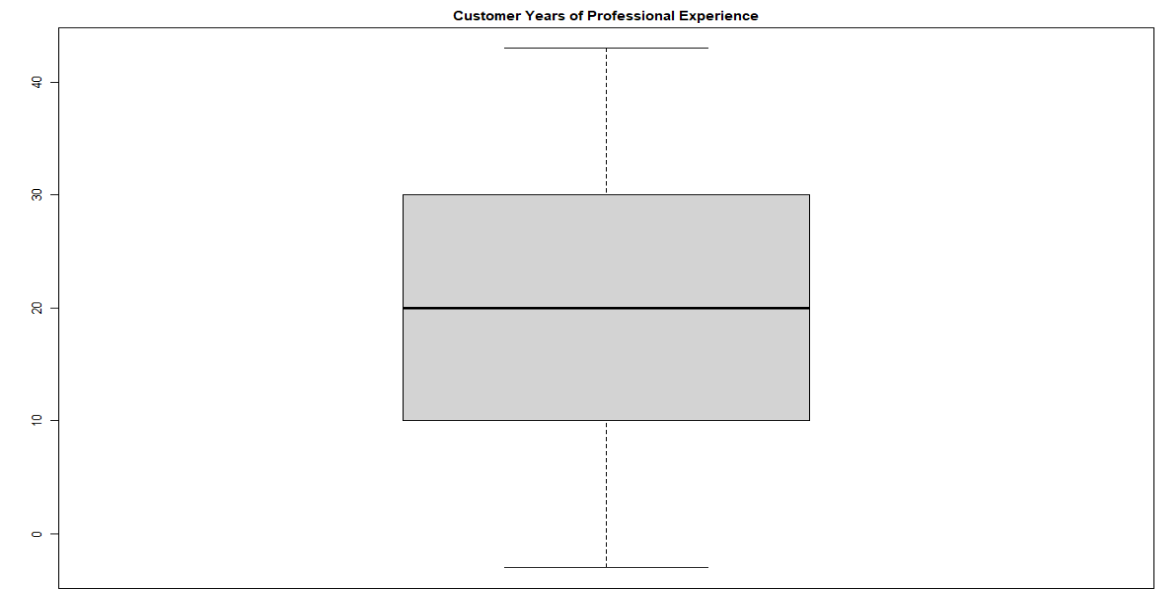


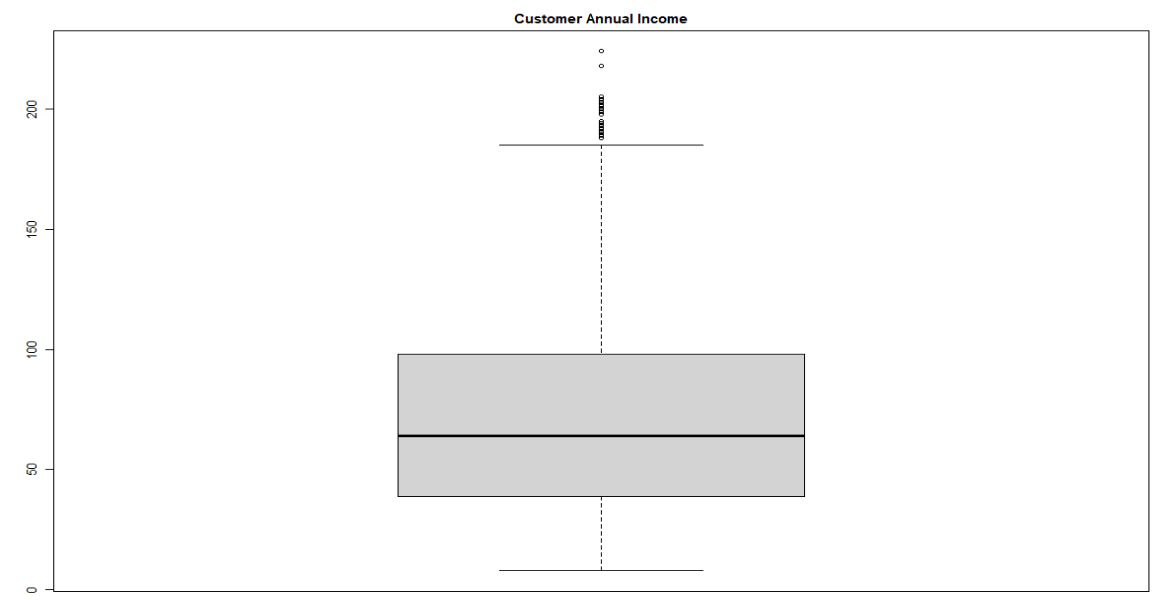


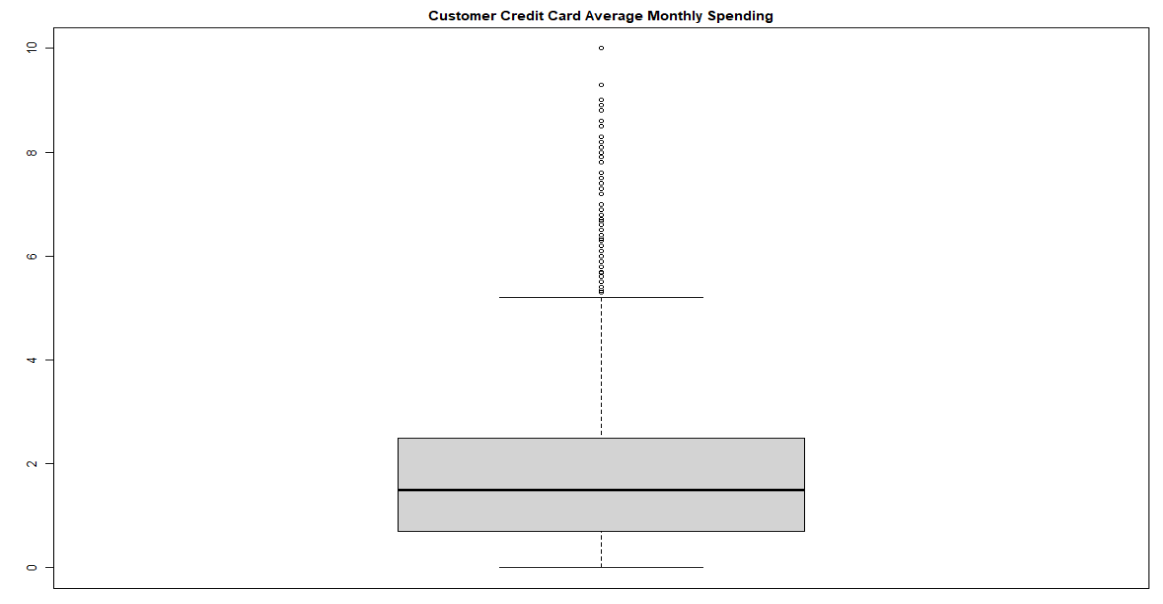
Boxplot of continuous variables

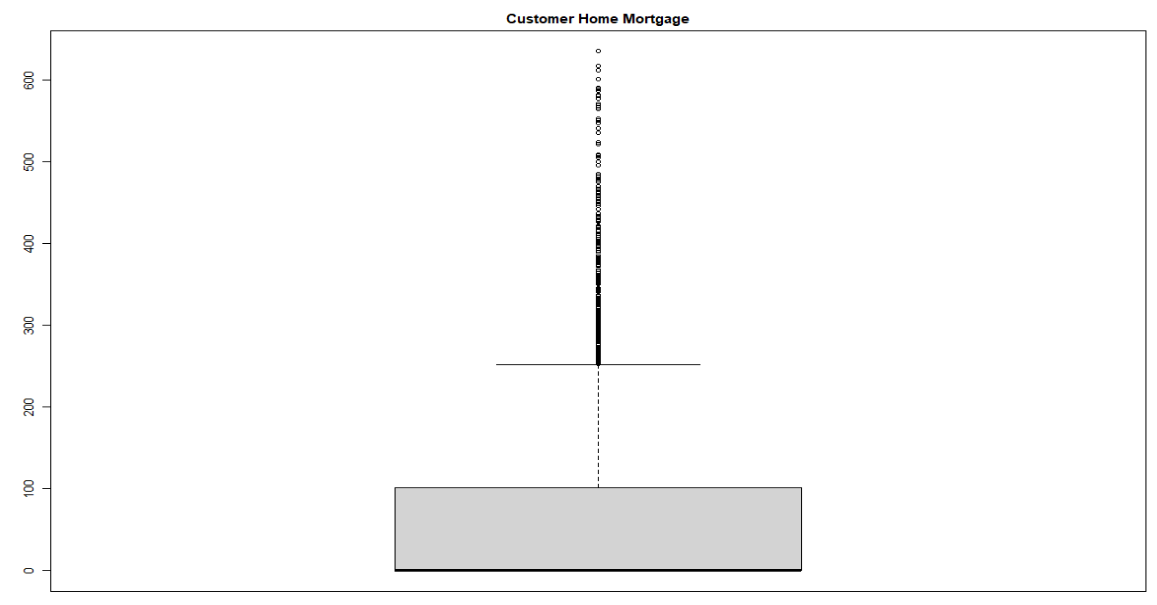
Customer age and years of professional experience have a normal, symmetrical distribution. The median customer age is in the 40's and the median years of professional experience is 20 years.



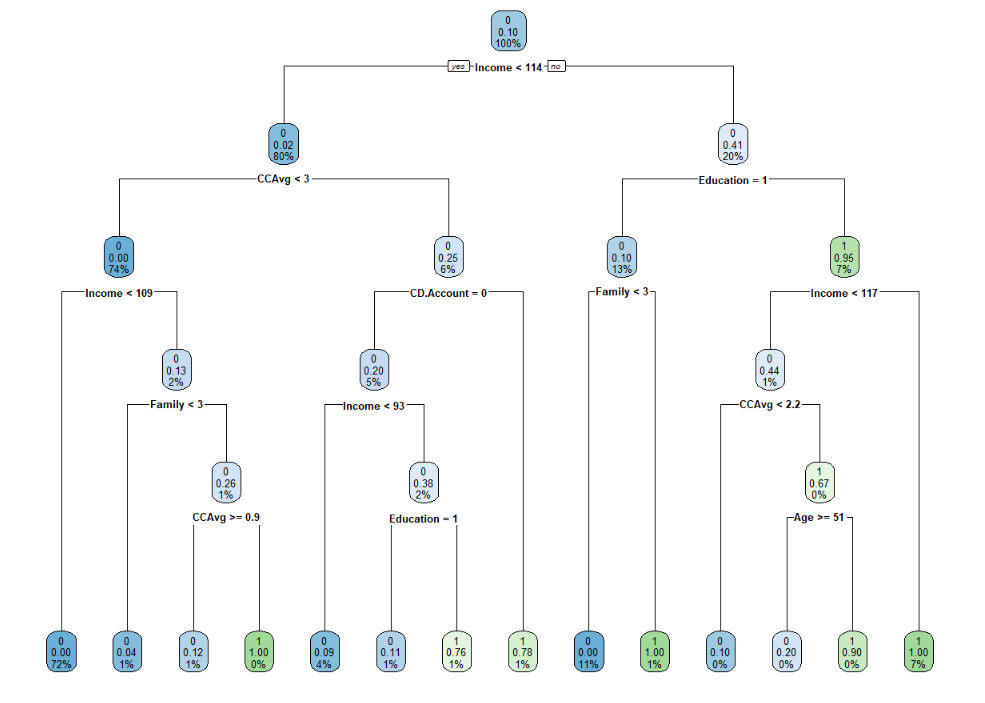


The last three boxplots contain some outliers and home mortgage has the most outliers. 

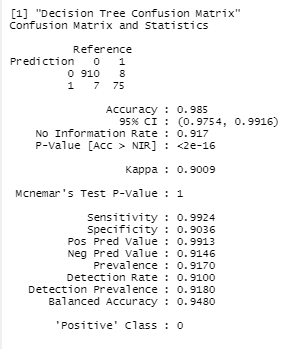




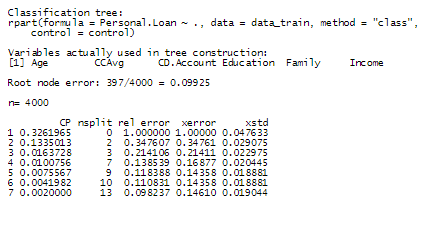
Modeling

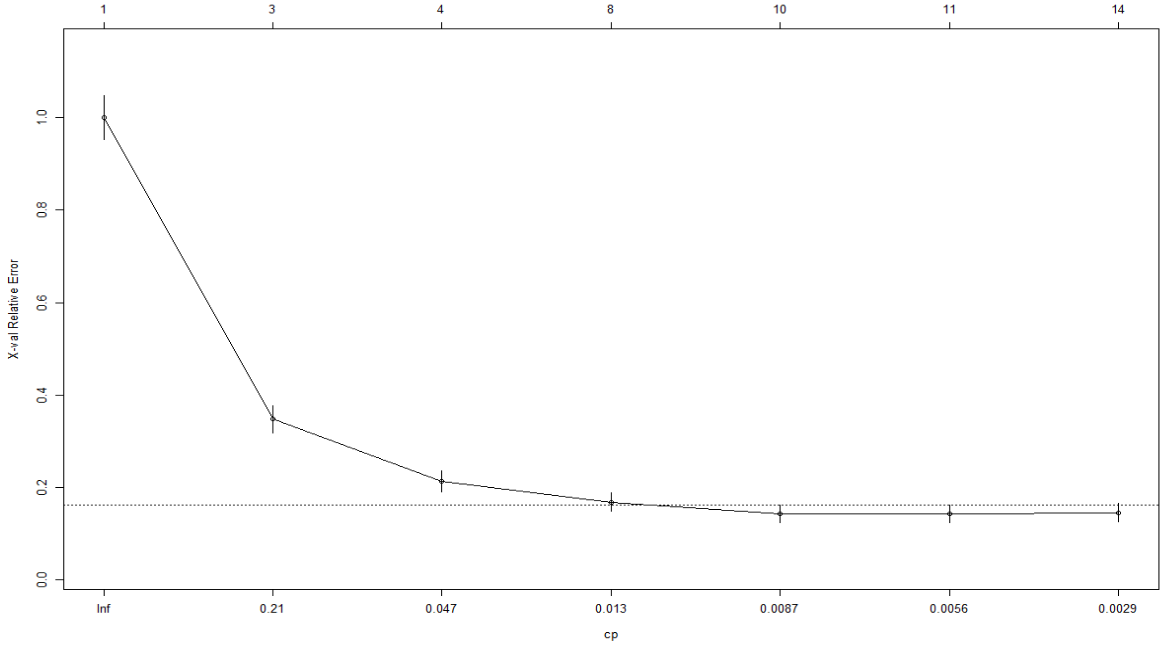


Confusion Matrix of Decision Tree has a 98.5% accuracy.



CP table. CP 5 & 6 have equal xerror and are the lowest error rates.





Accuracy comparison, base accuracy versus postpruning accuracy. It is exactly the same.