LeTicia Cancel

DATA622 Homework #4

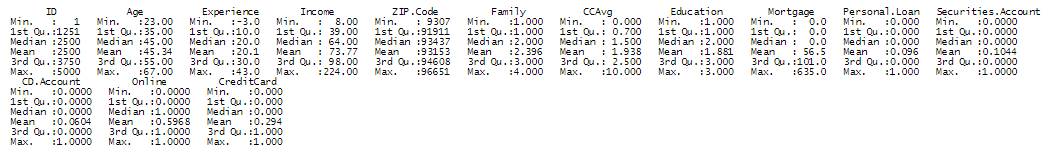
December 22, 2022

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.

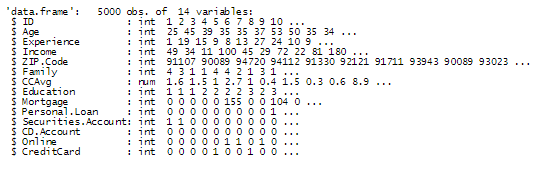
The bank collected customer demographic information and the types of accounts the customer has with the bank. Below is a list of all columns in the dataset and a description for each column.

|  |  |
| --- | --- |
| **Column** | **Description** |
| 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? |

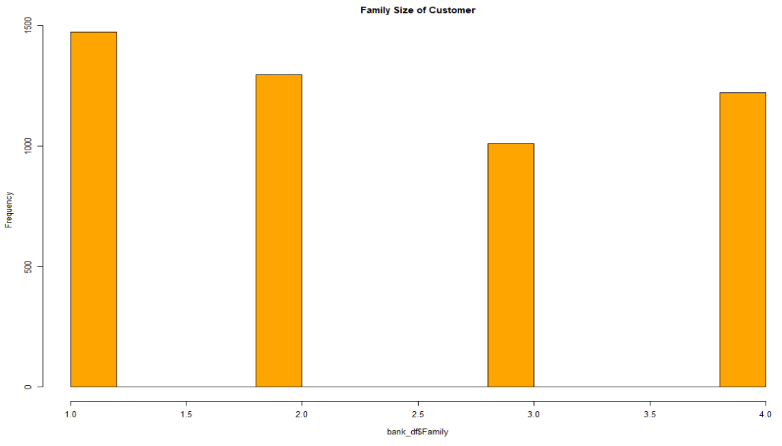
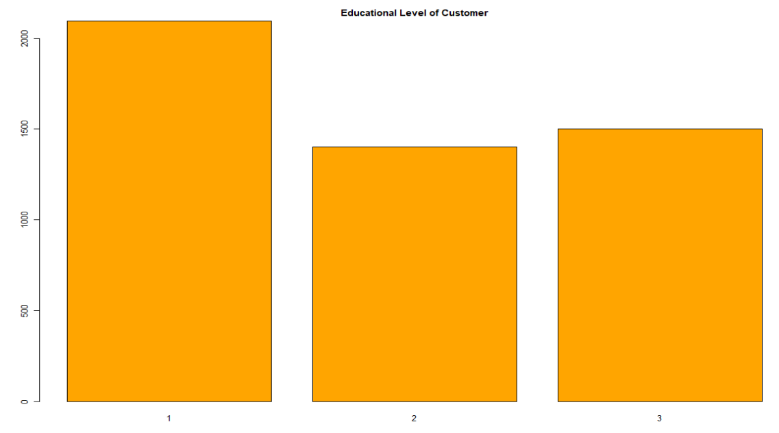
If we look at the summary of the dataframe we see that there are no N/A values. Some columns have a minimum value of zero but the zero has a meaning so we do not have to do not have to clean or remove any of the values.

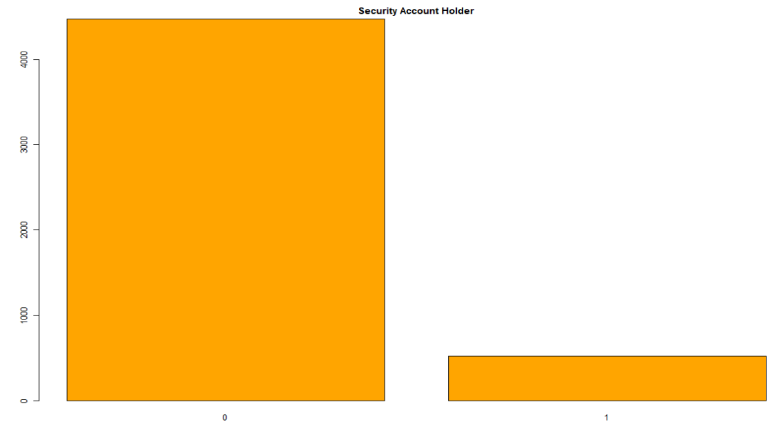


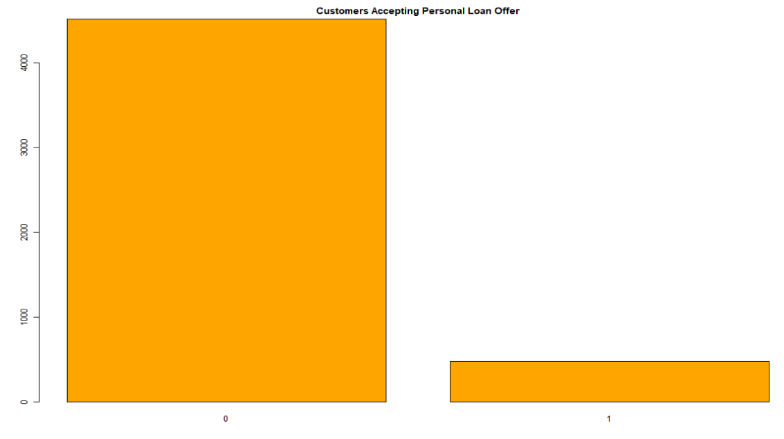
The columns are a mix of continuous and categorical so we will convert the categorical columns to factors.

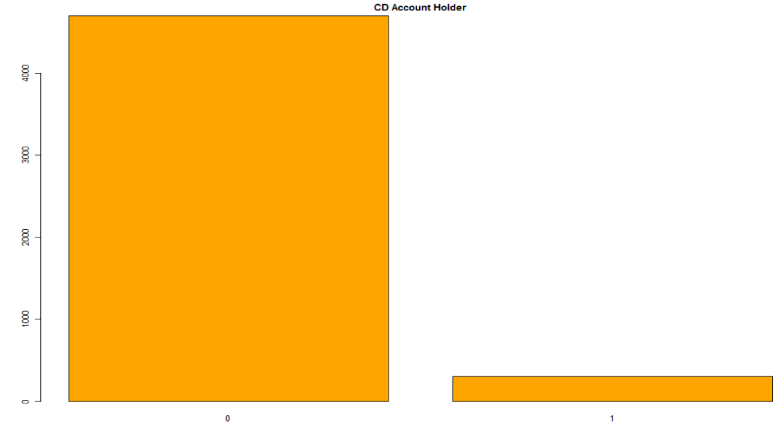
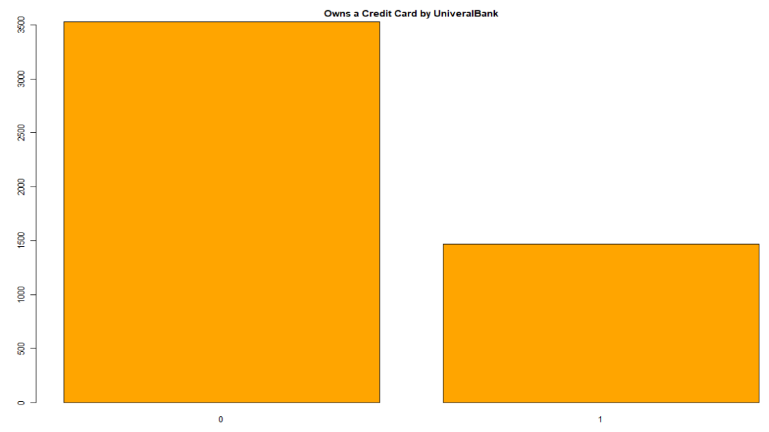


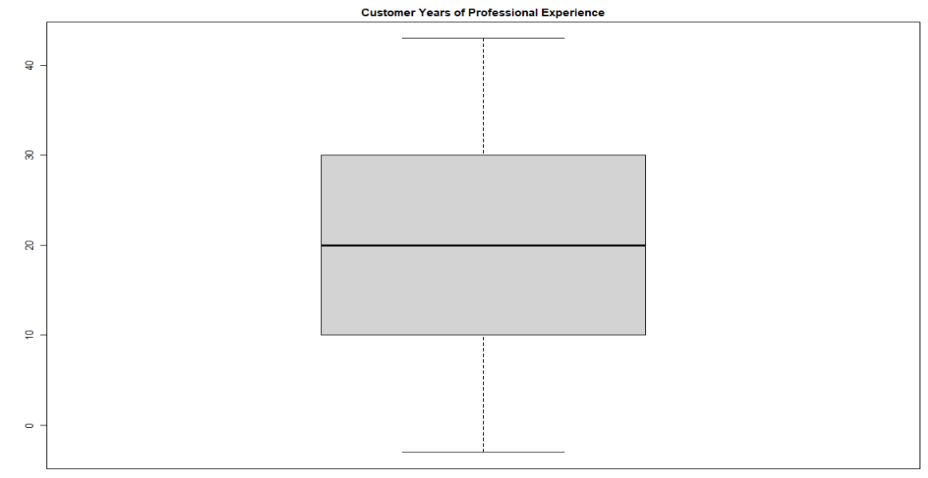
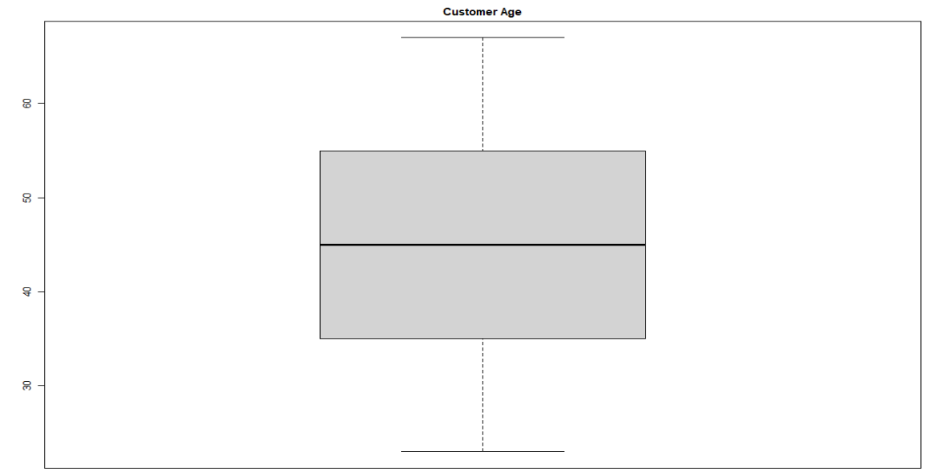
After converting Education, Personal Loan, Securities Account, CD Account, Online, and CreditCard to factors, we can explore the counts for each column. Most of the customers are a household of 1 or 2. The highest education level for most customers is Undergrad (1), then Advanced/Professional (3), and last is Graduate (2).



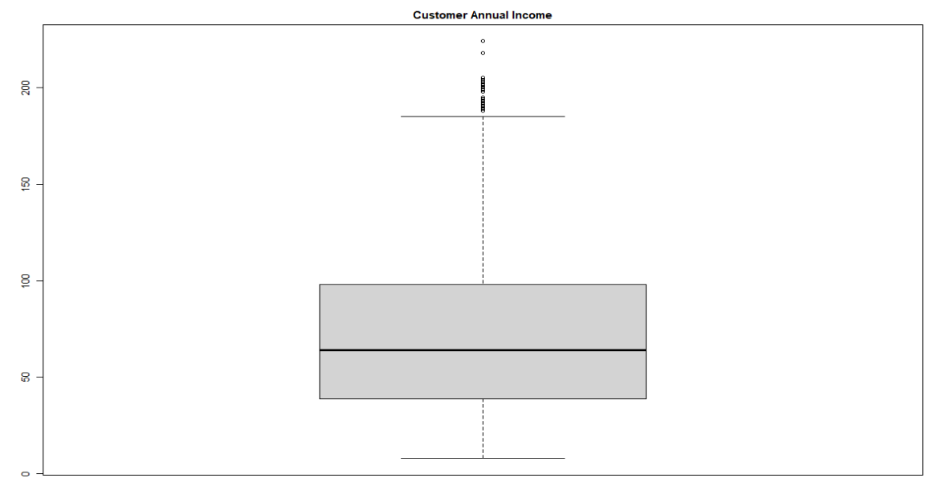
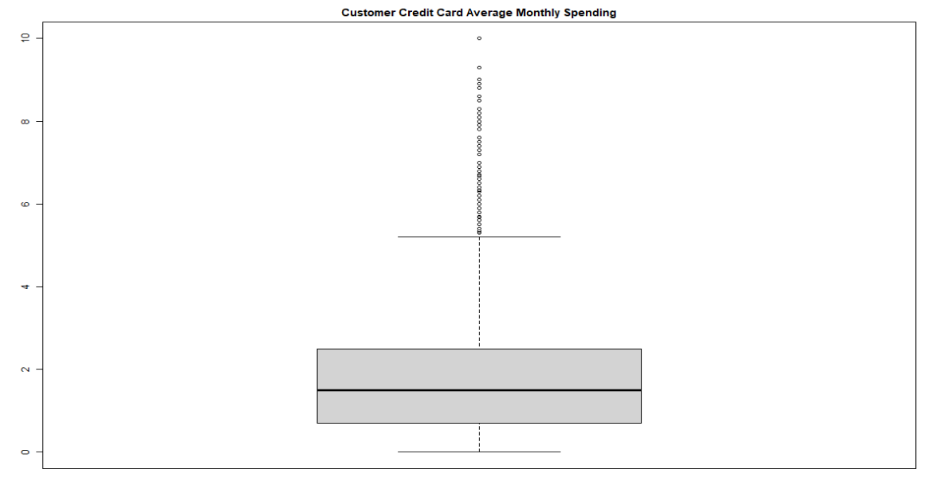
The following charts are columns that only contain Boolean values, a one or a zero (Yes/No). Most of the customers did not accept the Personal Loan offer, they do not have a Security Account or a CD account, and most do not own a credit card from UniversalBank.

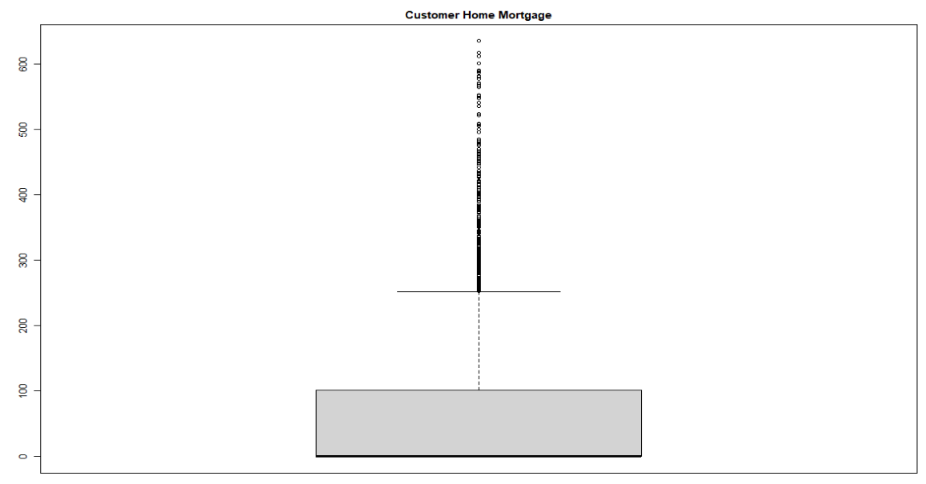




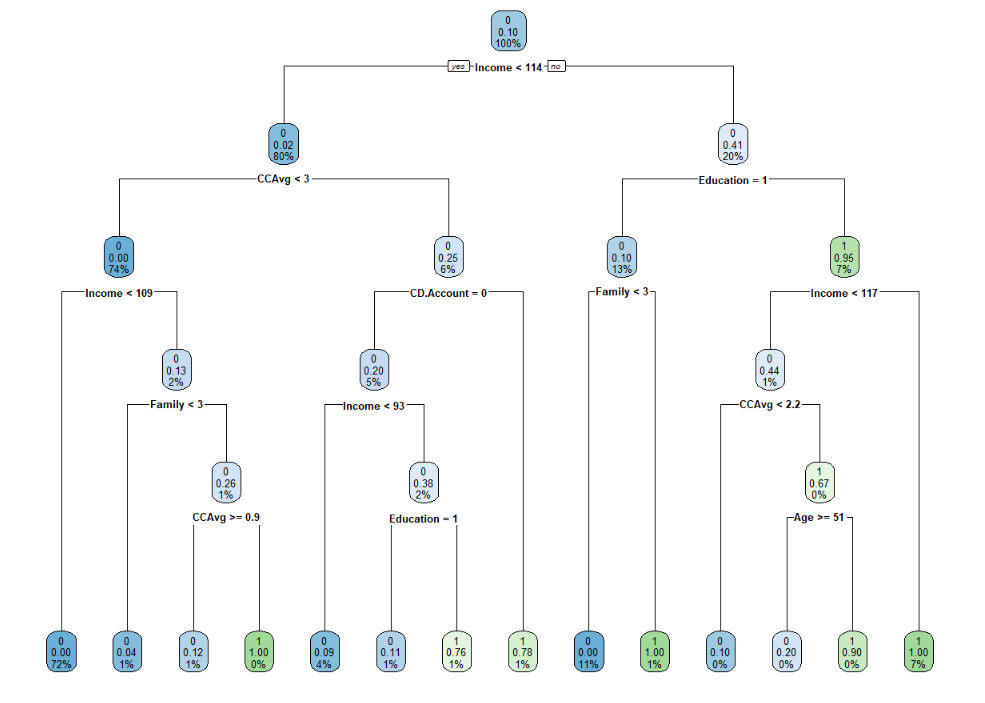
We will use a Boxplot to explore the columns with continuous variables. Customer age and years of professional experience have a normal, symmetrical distribution. The median customer age is in the 40's and the median years of professional experience is 20 years.

The last three boxplots contain some outliers with Home Mortgage having the most outliers.

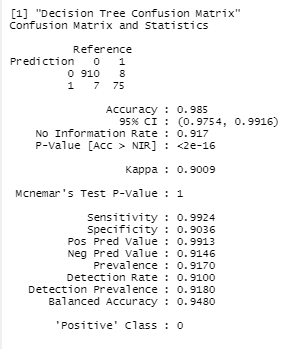




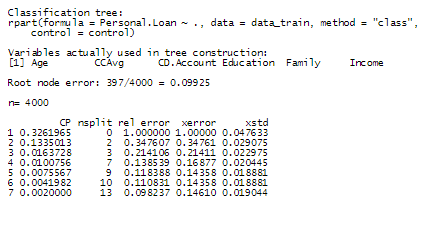
The model used for this project is a decision tree. We used an 80/20 split for the train and test dataframes and used the Personal Loan column as the predictor. A zero means the customer did not accept the Personal Loan offer and a one means the customer did accept the personal loan offer. We know from looking at the bar chart from earlier that most customers did not accept the Personal Loan offer. If we follow one of the branches that leads to a 1 (accepted loan offer), customers with an income over $114,00, Education Level above Undergraduate, and income over $117,000 (there are two Income nodes on the right side of the decision tree). A simplified interpretation of that is customers with a higher level of education and a higher income were more likely to accept the Personal Loan offer.

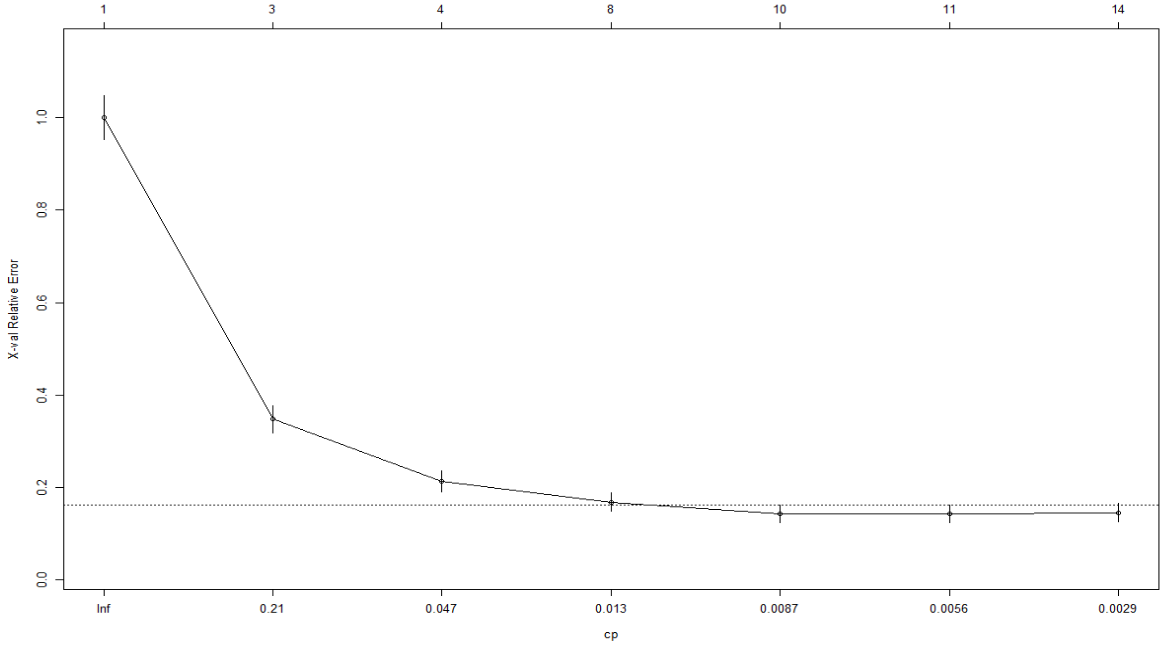


A Confusion Matrix was used to test the accuracy of the Decision Tree, so we can see that is has a 98.5% accuracy.

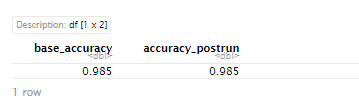


Although the confusion matrix shows a high accuracy, we will experiment with pruning the tree to see if it makes a difference. A Complexity Parameter (CP) table is used to select the optimal size of the decision tree. CP 5 & 6 have equal and lowest xerror.





Using the cp value for row 5 (0.0041982) we run the tree and prediction one more time. If we compare the accuracy from the original prediction to the postruning, we see that the values are exactly the same. So, pruning the tree did not make a difference.



If I were to do this analysis again, I would filter the data by customers who accepted the loan only. The current decision tree mostly displays information for customers who did not accept the loan which is useful information. If the bank wanted to run this campaign again, they can see the characteristics of the customers who said no. However, I think it would also be useful to have a tree that displays the characteristics of the customers who said yes.