Celina McBrayer

C744 – Data Mining & Analytics

May 29, 2020

Performance Assessment

I. Tool Selection

A. Benefits of R

R is a prominent programming language for statistics. It is free and open-source allowing for new developments to be released often through packages. However, its capabilities are on the lower end as it loads all of the data into RAM causing a dramatic reduction in speed.

B. Objective of Analysis

The main objective of this analysis is a focus on customer attrition. By analyzing and identifying customer habits, the telecommunications company can mitigate customer loss. In this case, the focus will be on the variable “Churn” and how it is impacted by other factors such as gender, age, monthly charges, etc.

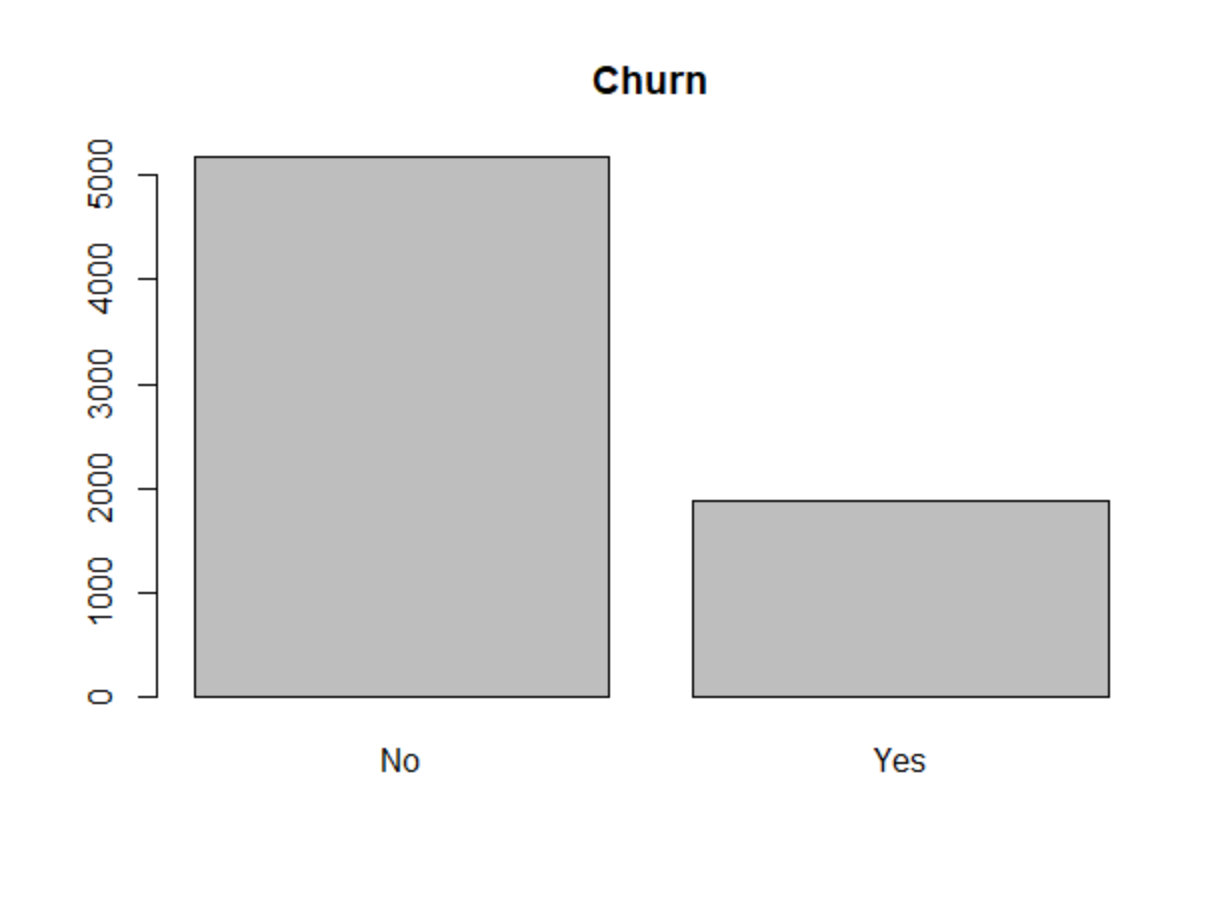
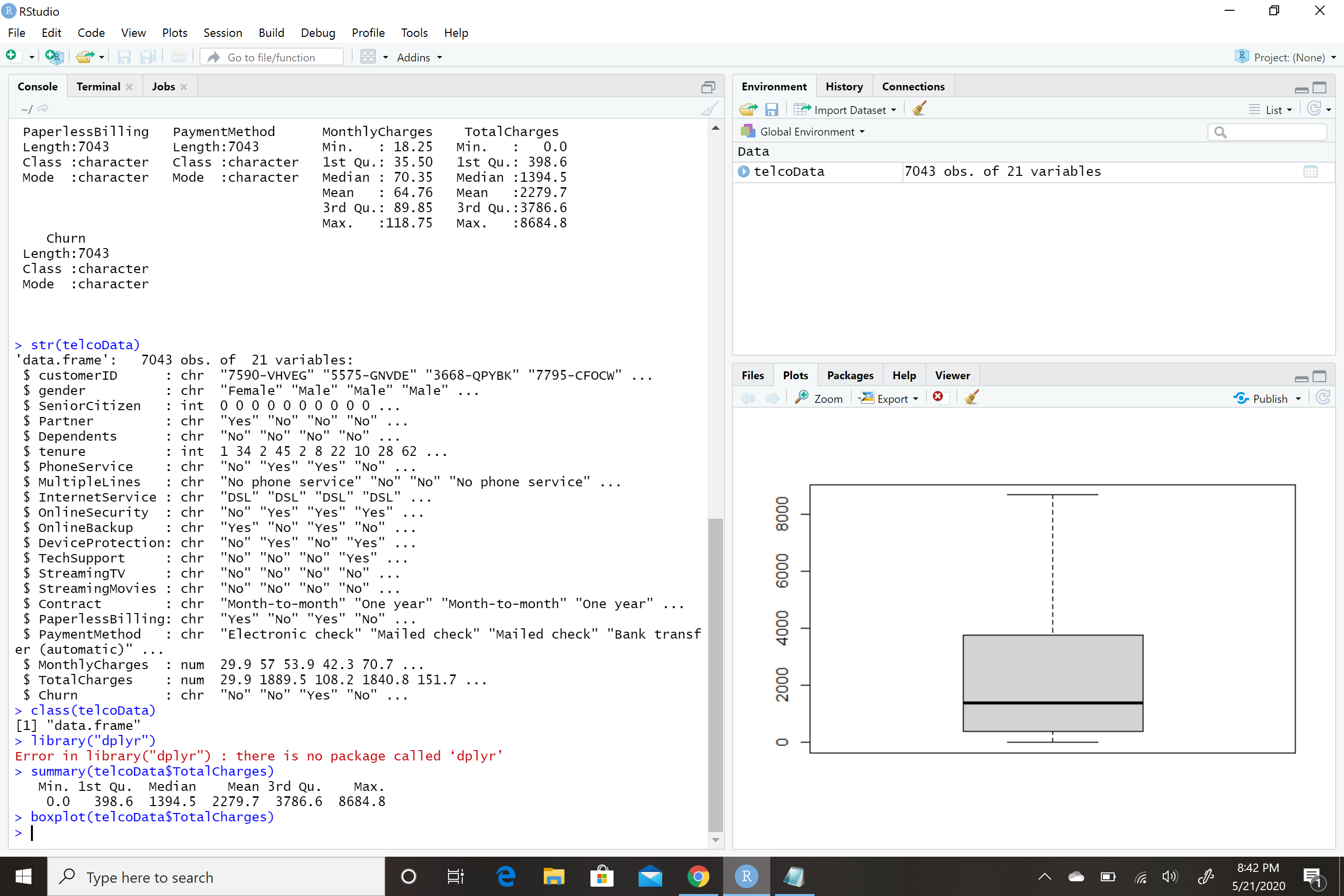
C. Descriptive and Nondescriptive Methods

The descriptive method to be used in this project is factor analysis. This allows the user to better understand the customer makeup of the telecommunications company as well as the current customers who have left and the factors that possibly affected their decision. This is descriptive because it is describing the data that we already have; these customers have already left. The predictive method to be used is a decision tree which will allow the company to predict if a customer will be retained or loss. By having this knowledge, the company may decide to offer certain promotions like free online security or a cheap package for Internet, Phone, and Streaming.

II. Data Exploration and Preparation

D. Target Variable

The target variable in this data is churn, which is whether or not a customer has left the telecommunications company. Churn initially was a character data type that classified the customer into one of two groups: yes or no. Keeping yes and no, I converted it into a factor type. By doing so, I was able to see about 2,000 customers who had left the landline business.



E. Independent Predictor Variable

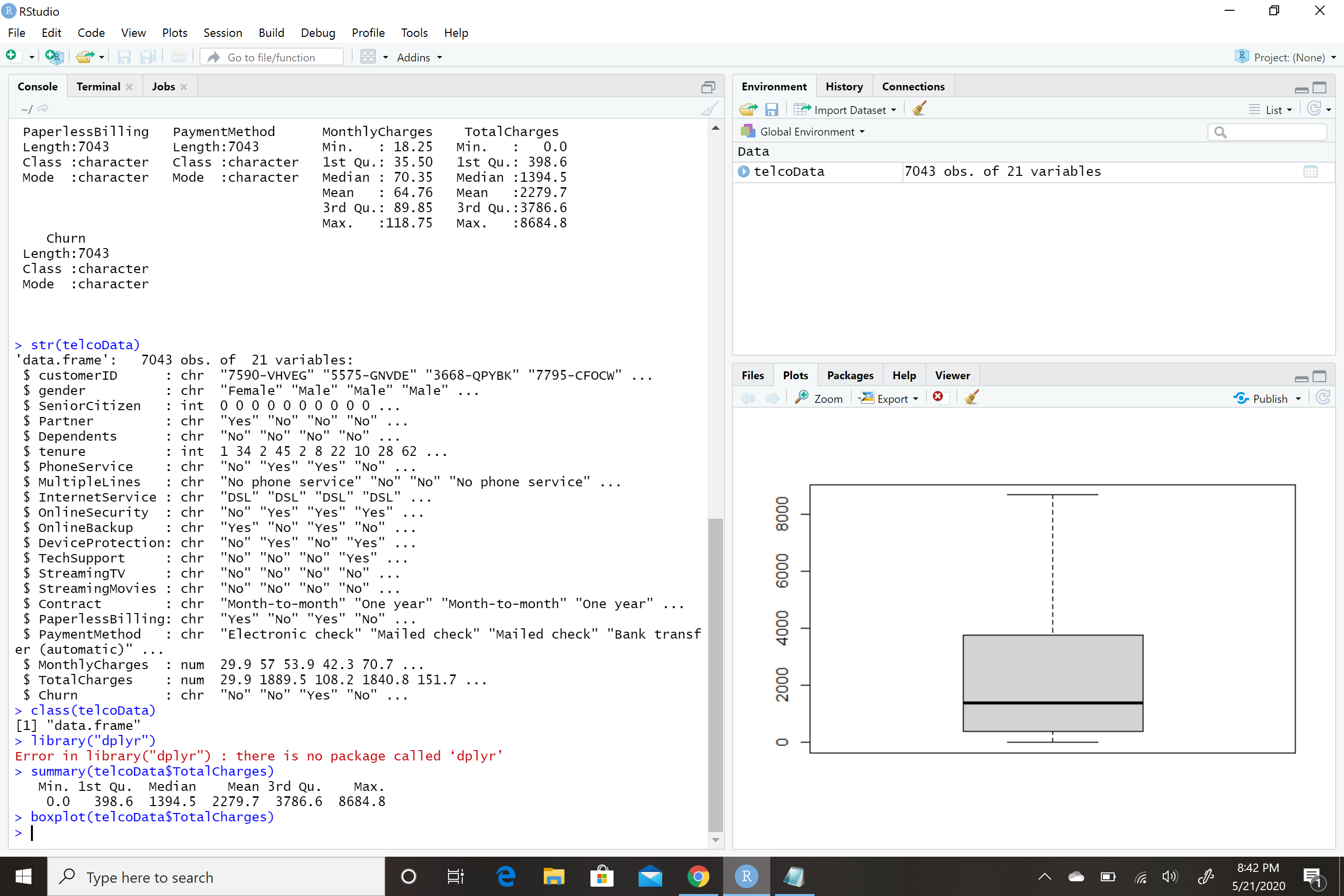
There are numerous predictor variables within this data set. Of the numerical type are tenure, monthly charges, and total charges. Of the factor type are gender, senior citizen, partner, dependents, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming tv, streaming movies, contract, paperless billing, and payment method. Each of these has the possibility to be related to customer attrition. For instance, it is possible that senior citizens are more likely to leave the company or that a customer who pays a higher monthly charge may be more likely to leave the company in favor of another.

F. Goal in Manipulation

In terms of data manipulation, the important tasks were to check for and remove any extraneous data as well as fill in missing values. I also observed data types and decided for this project it was best to convert the character data types to factors as this would best allow me to achieve the goal of the data analysis.

G. Statistical Identity

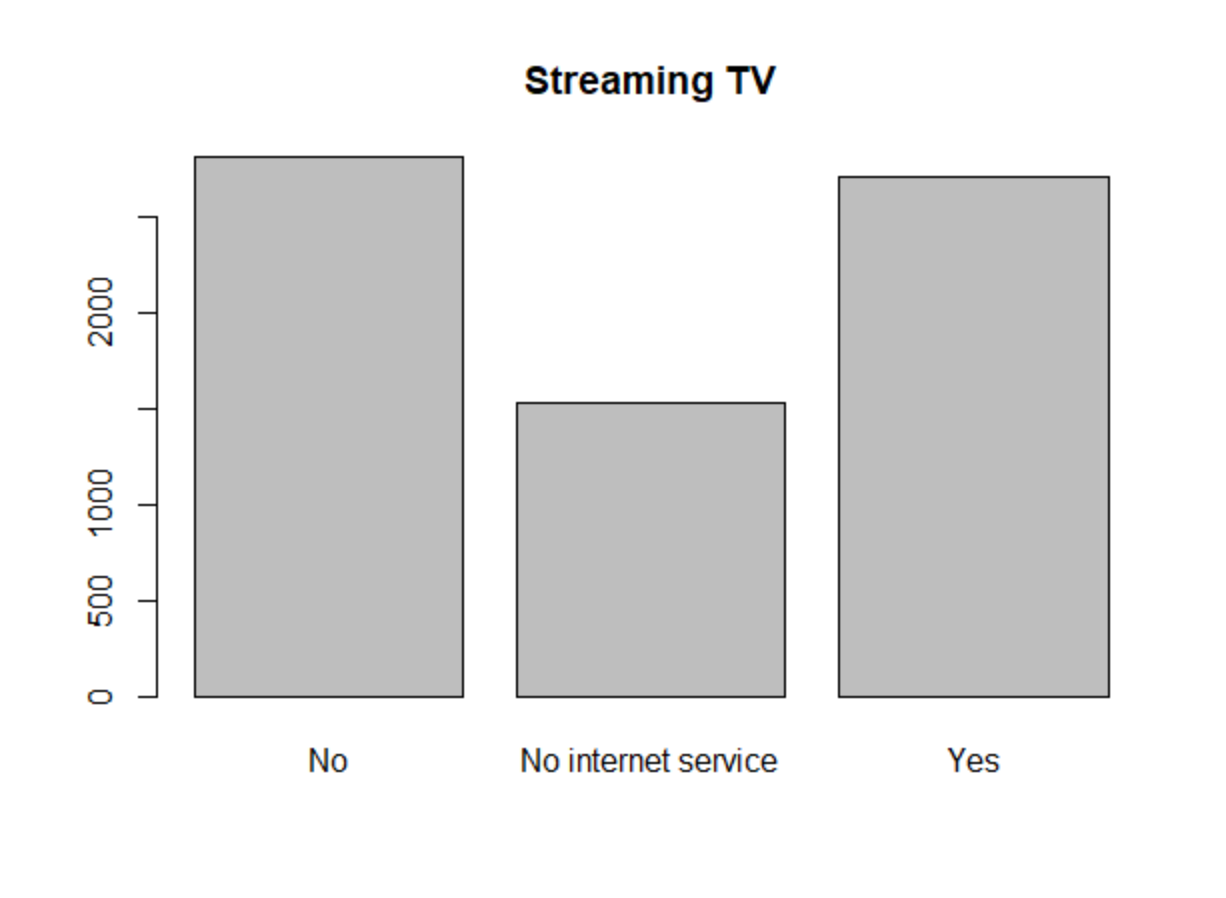
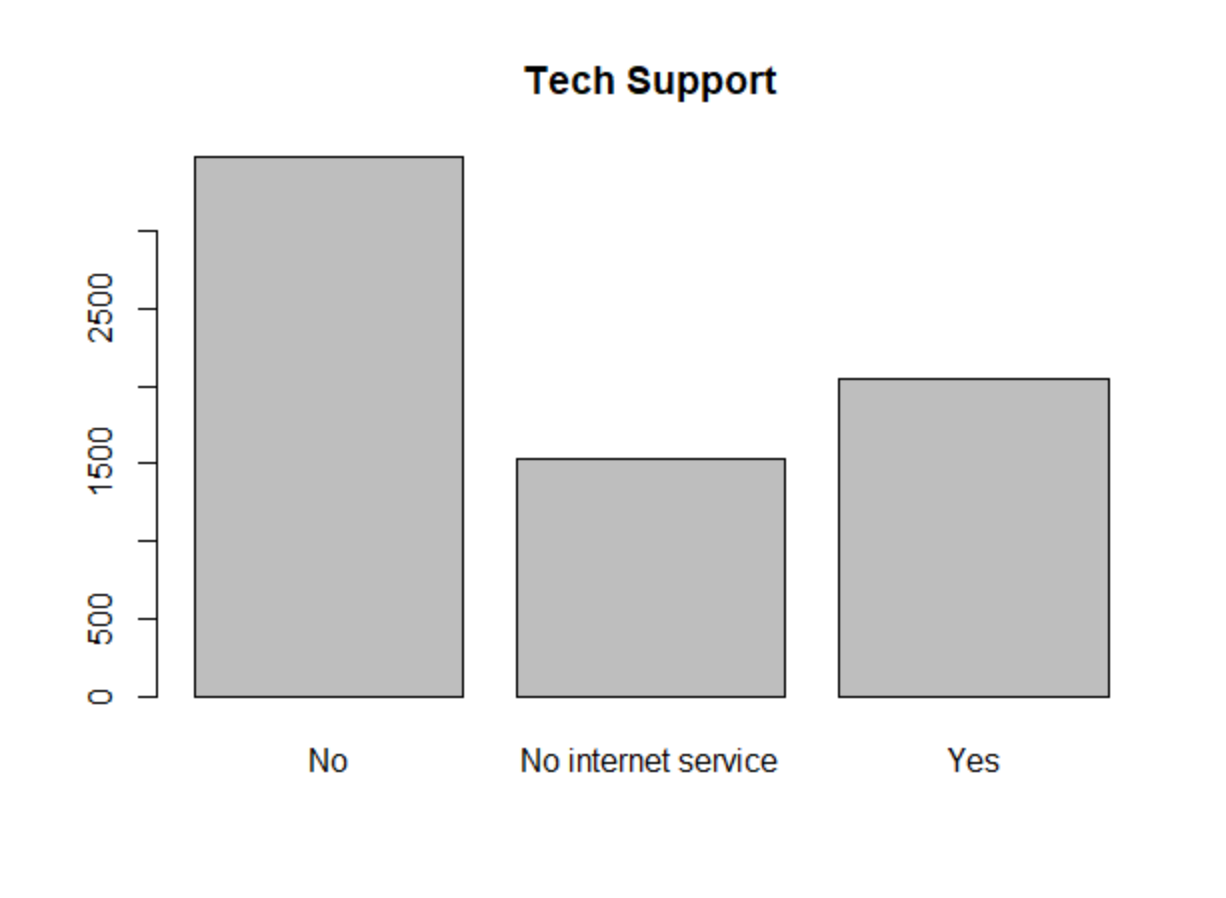
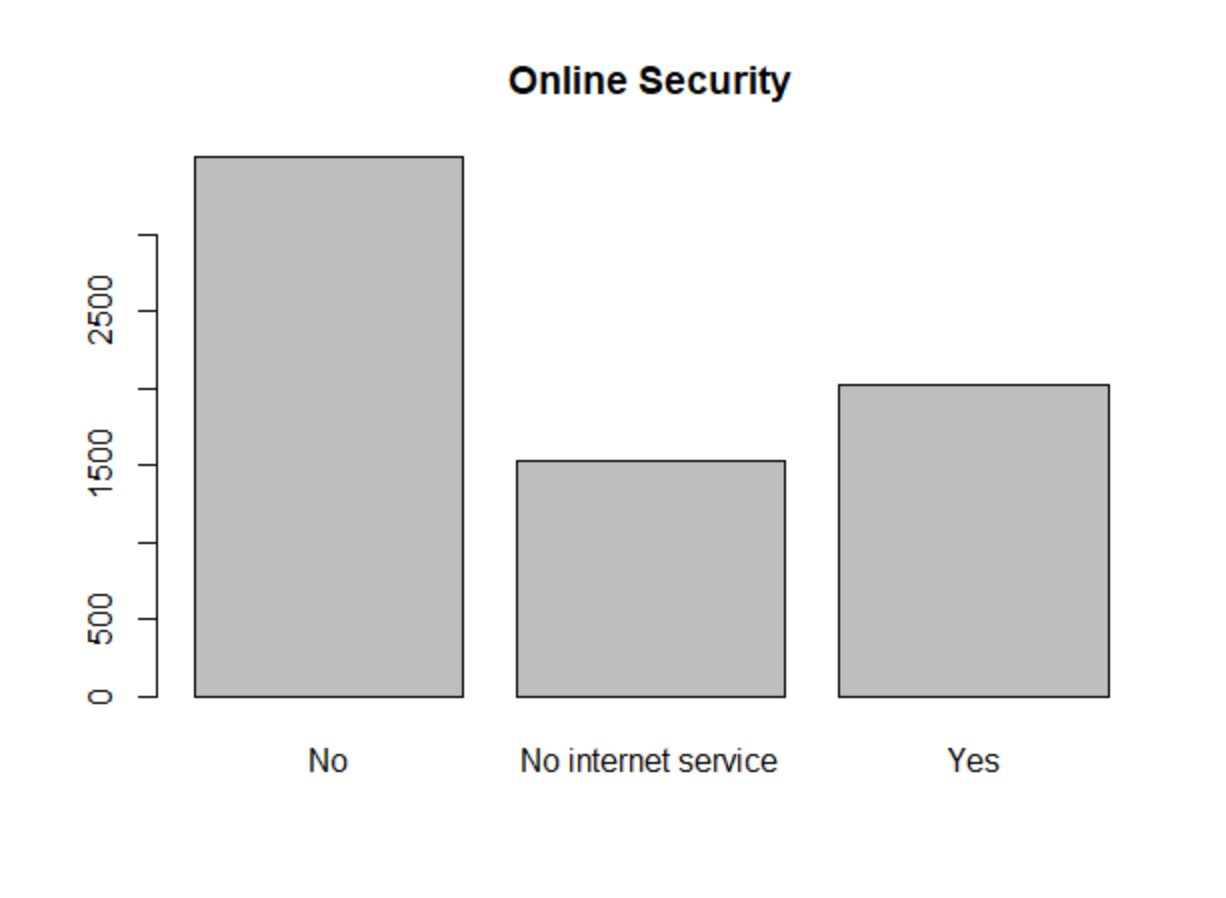
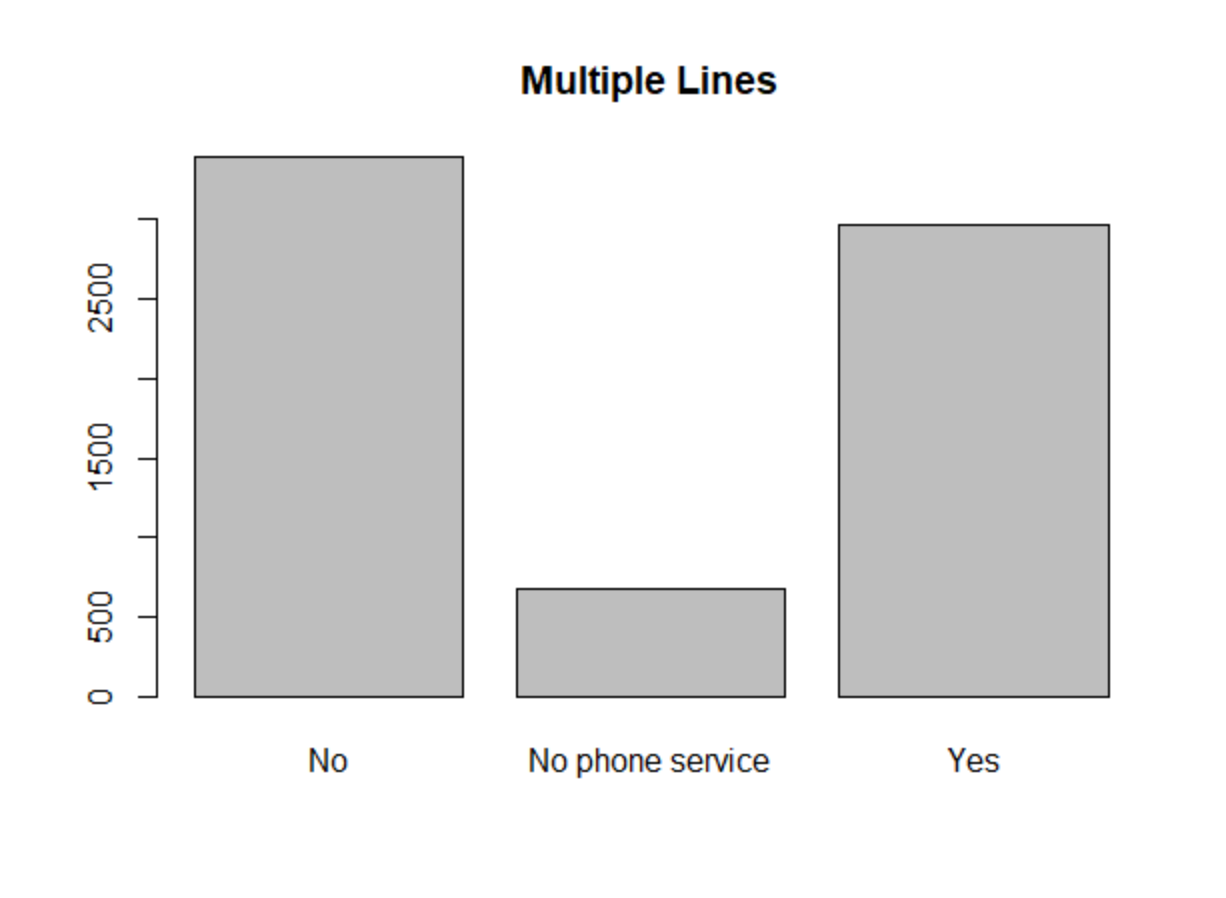
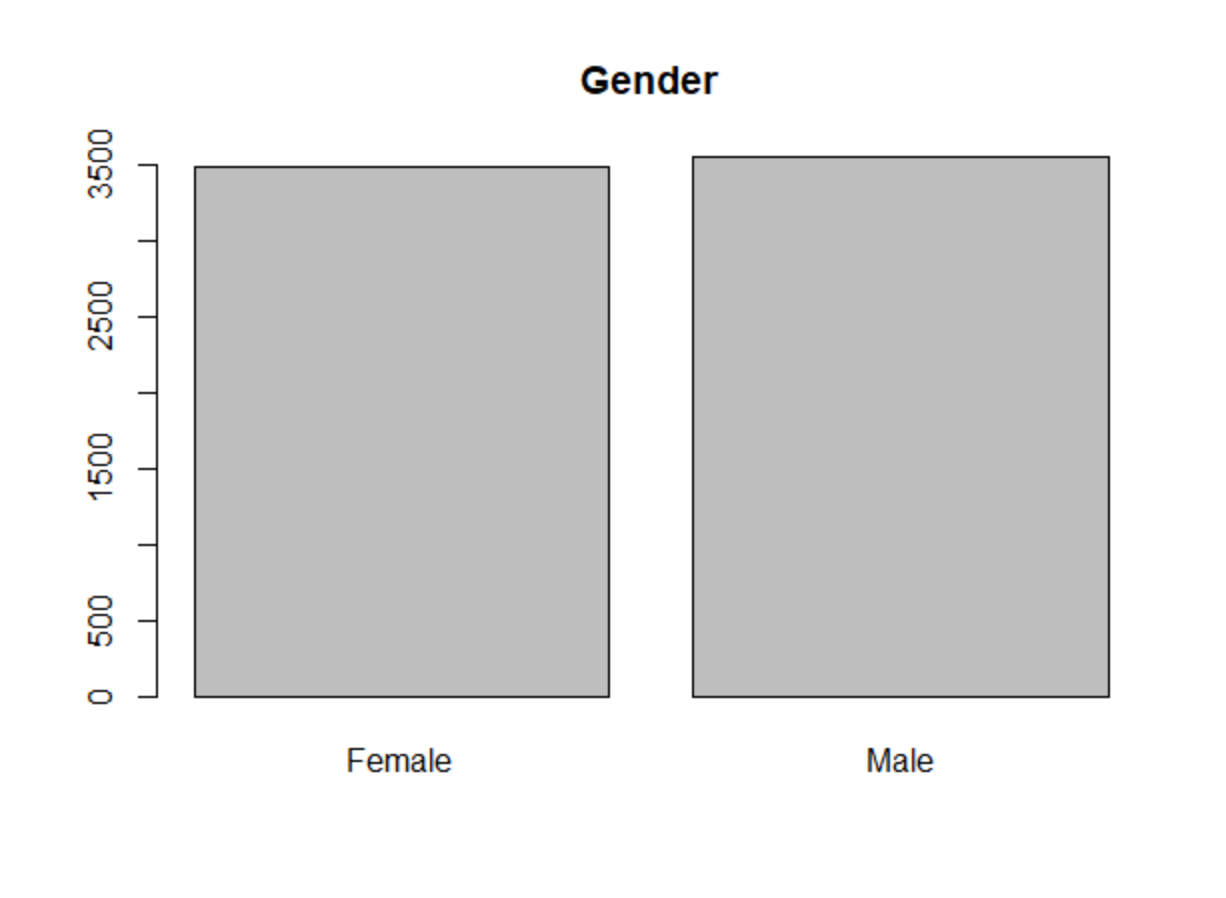
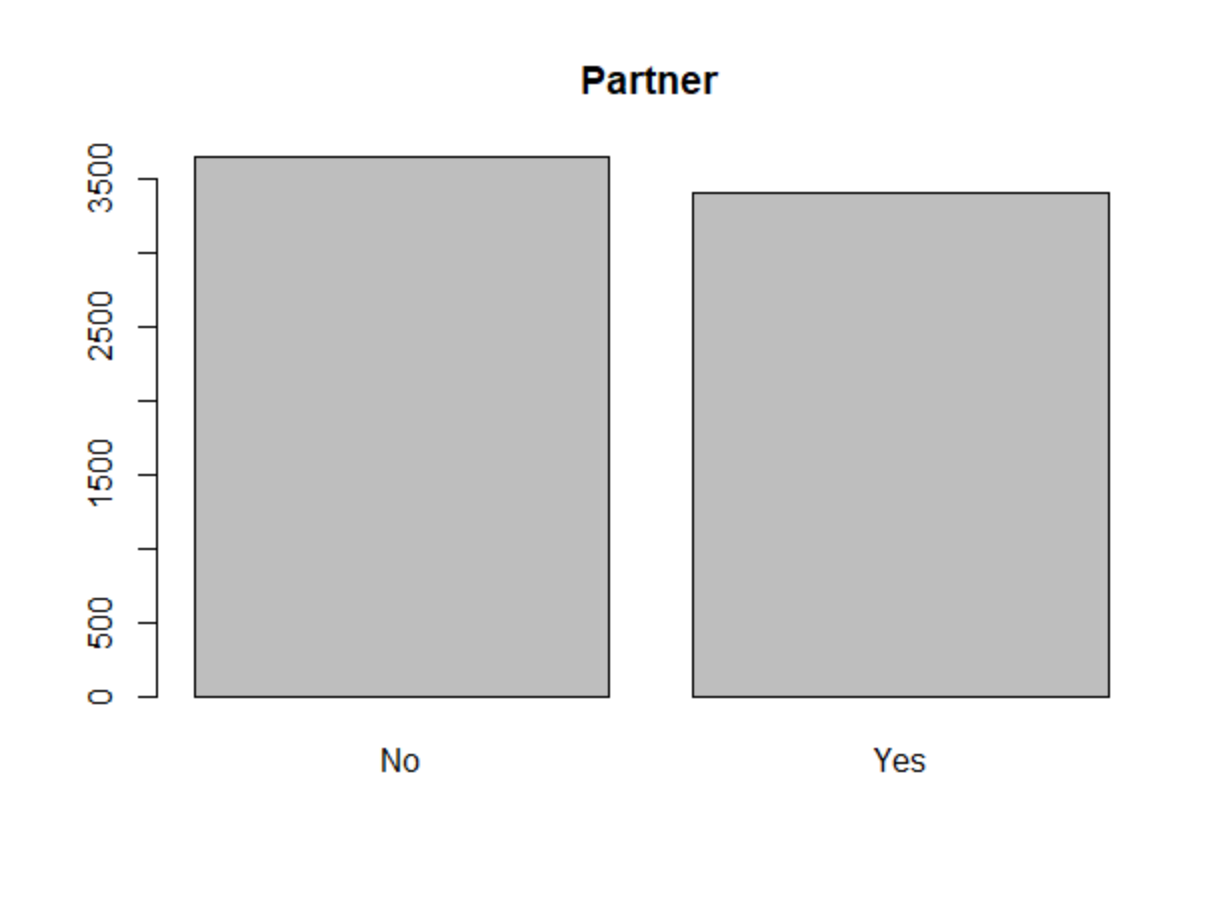
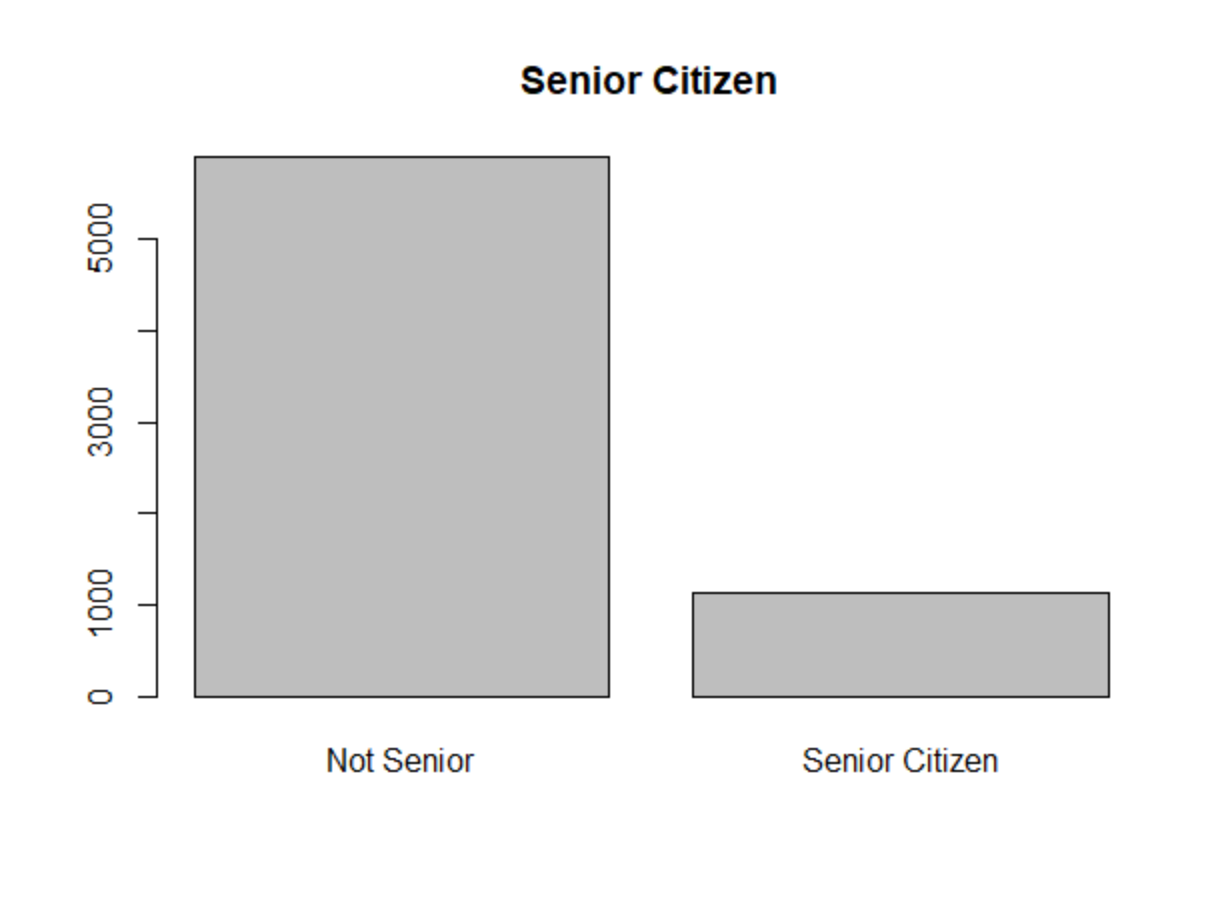
The data within this particular set is mainly qualitative data of the categorical type. Within each category is a nominal result. In addition, there is numerical data with the tenure, monthly payment, and total payments. The phenomenon to be predicted is whether the categorical data within churn is affected by other factors such as dependents or tenure.

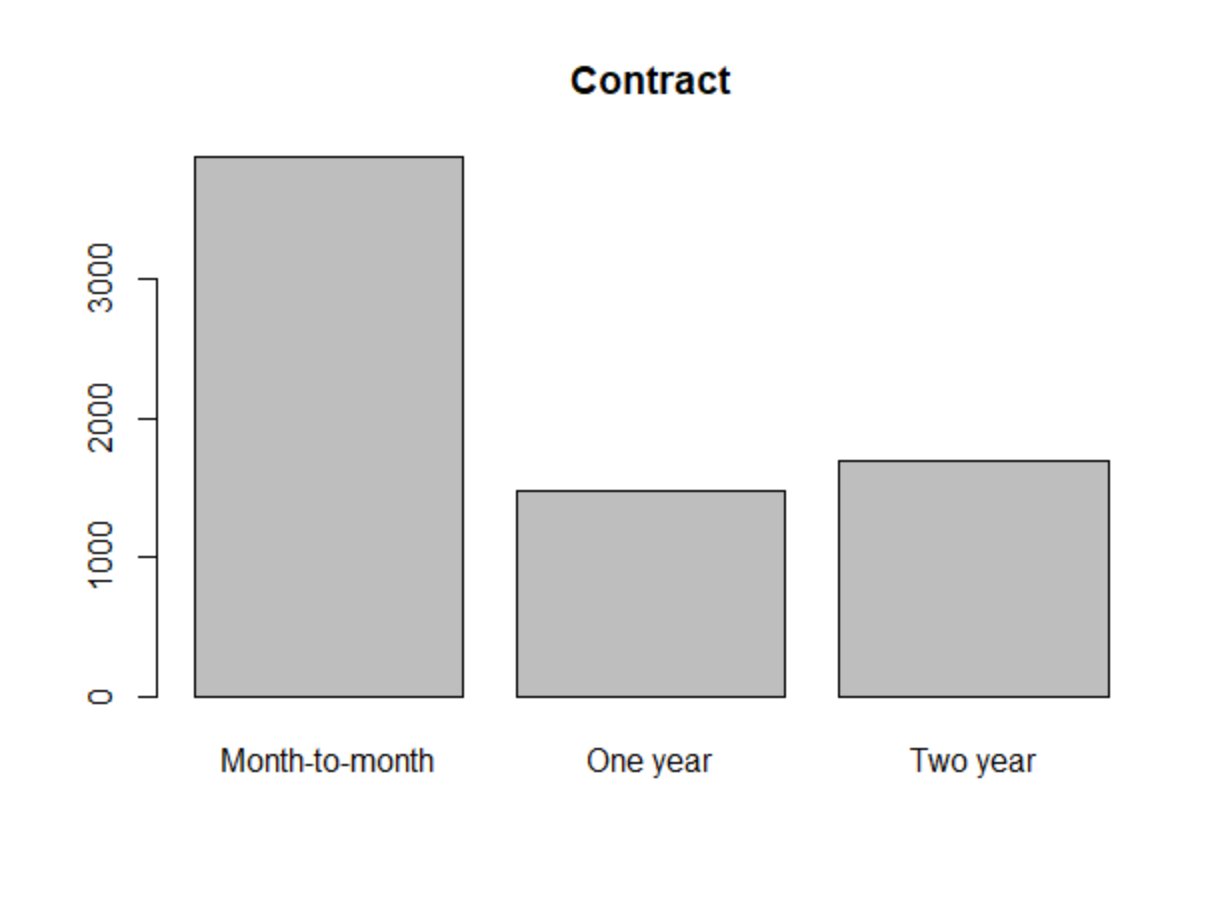
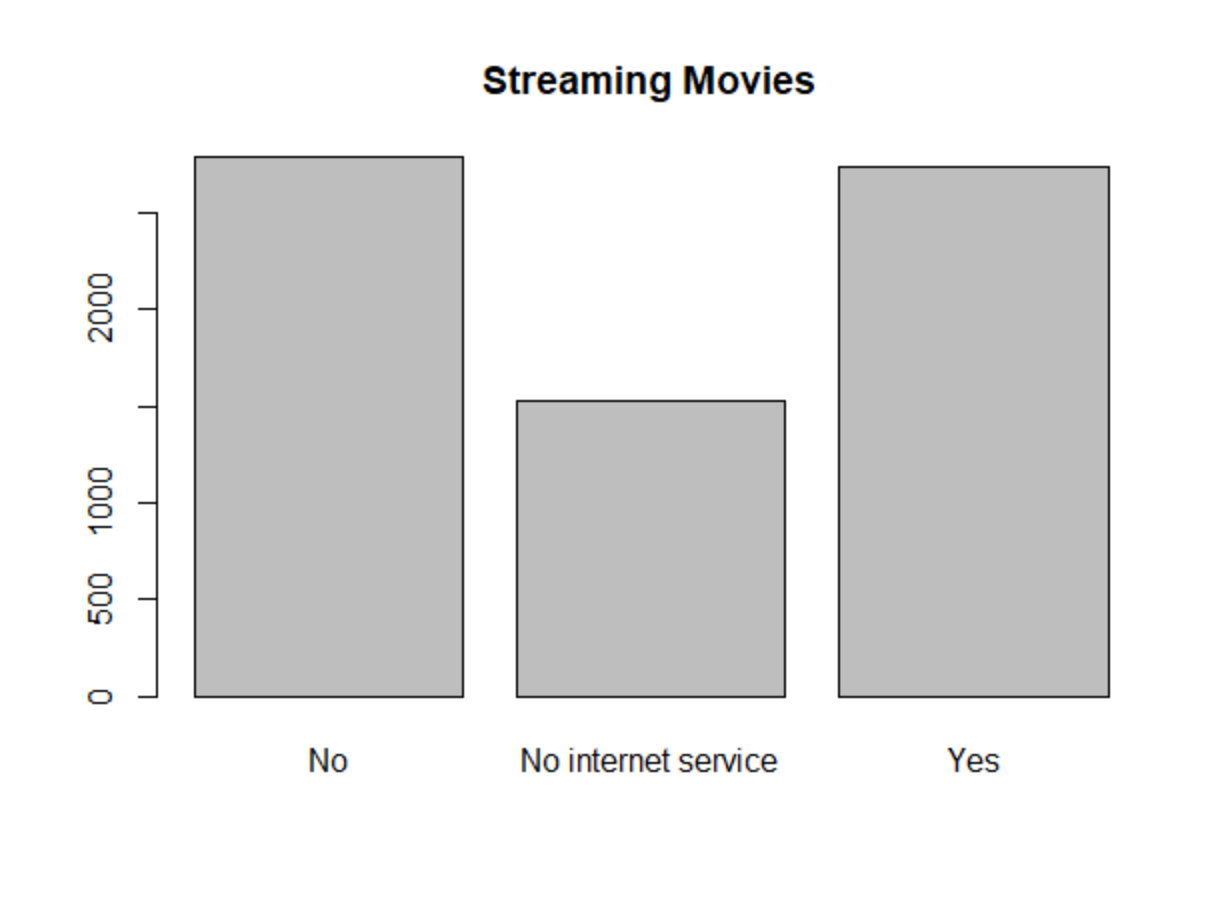


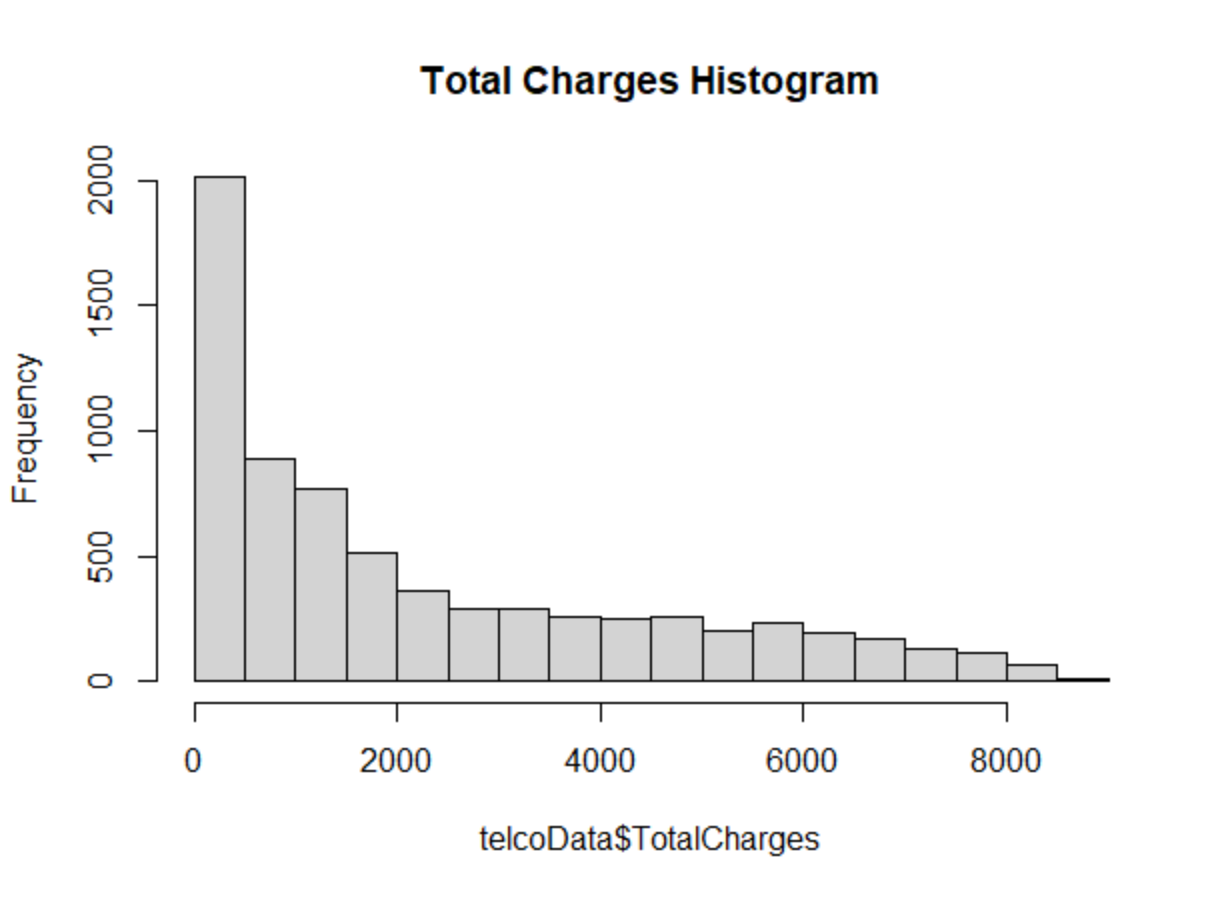
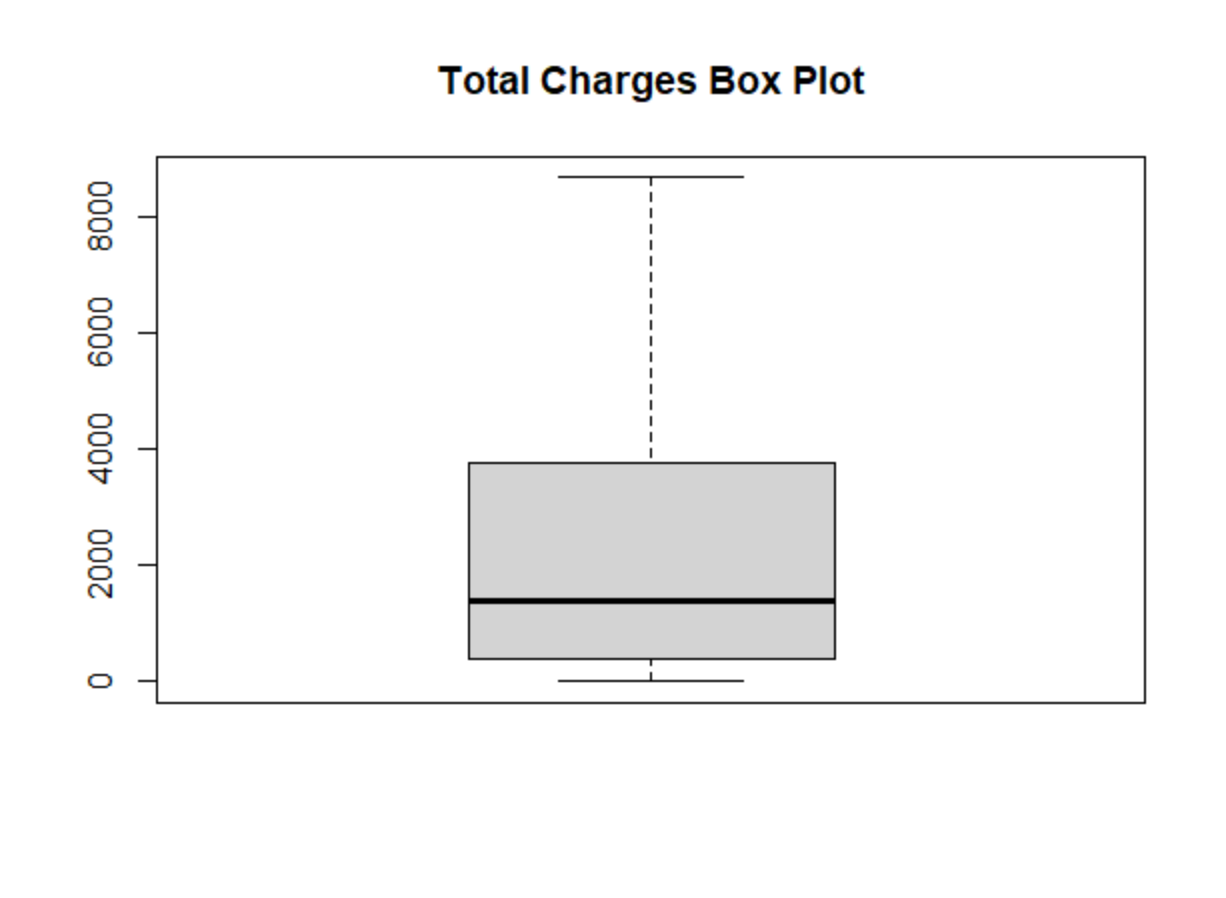
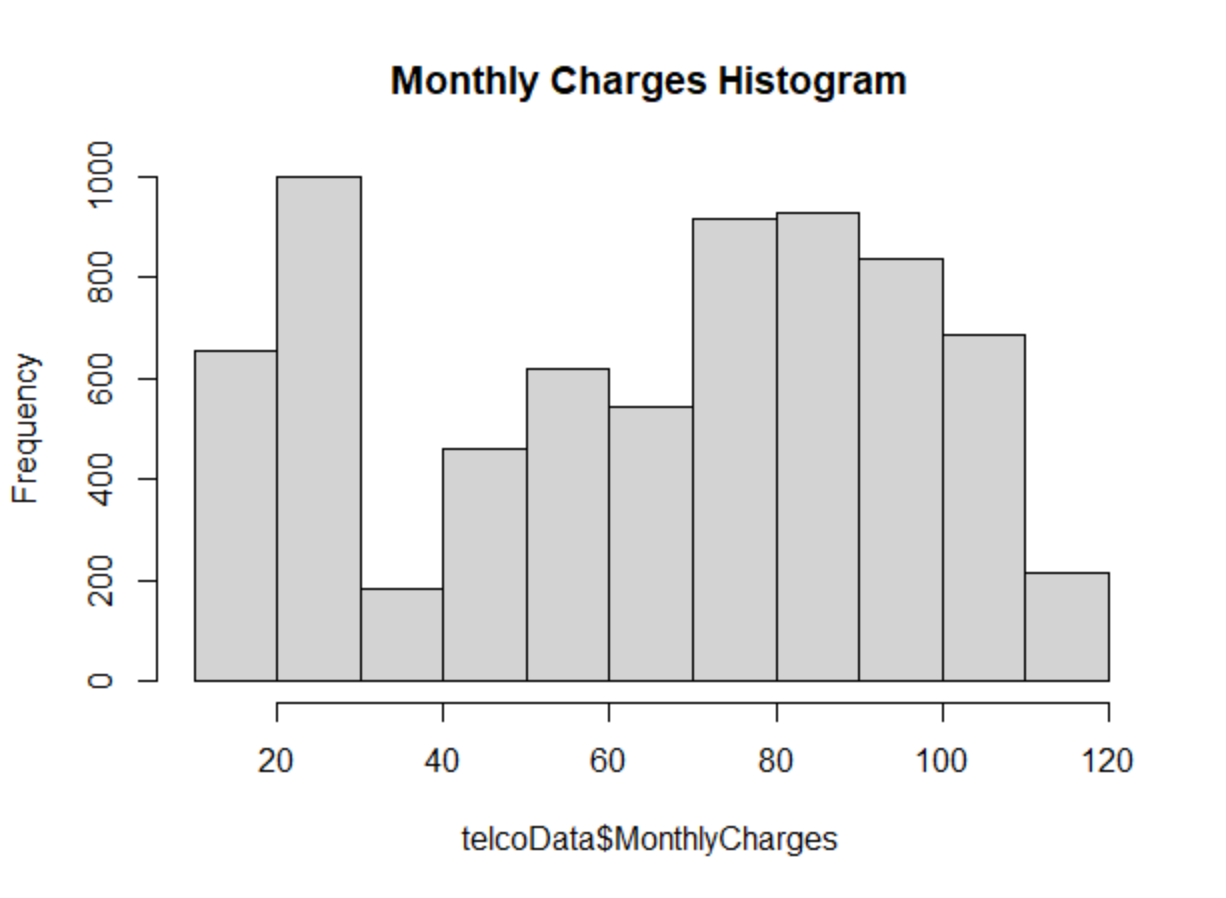
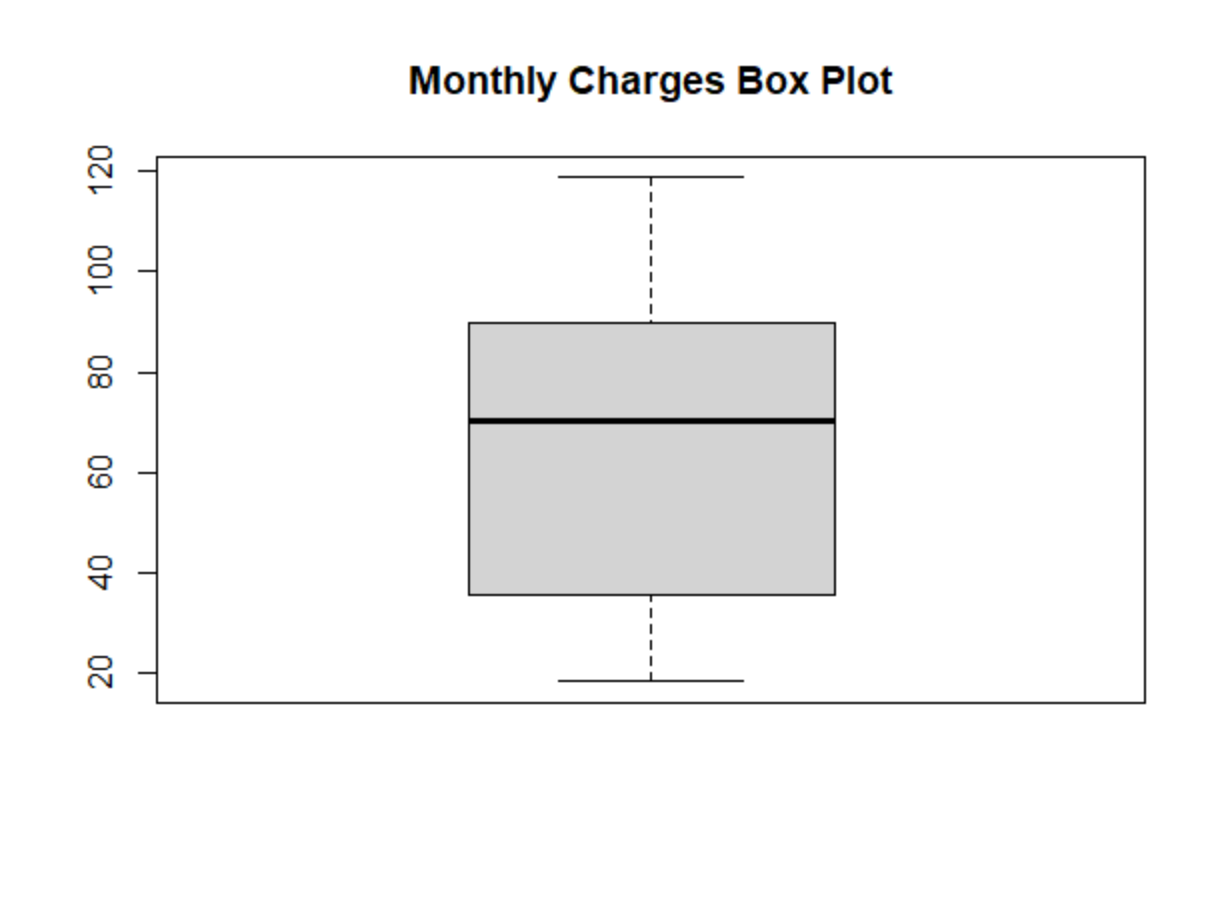
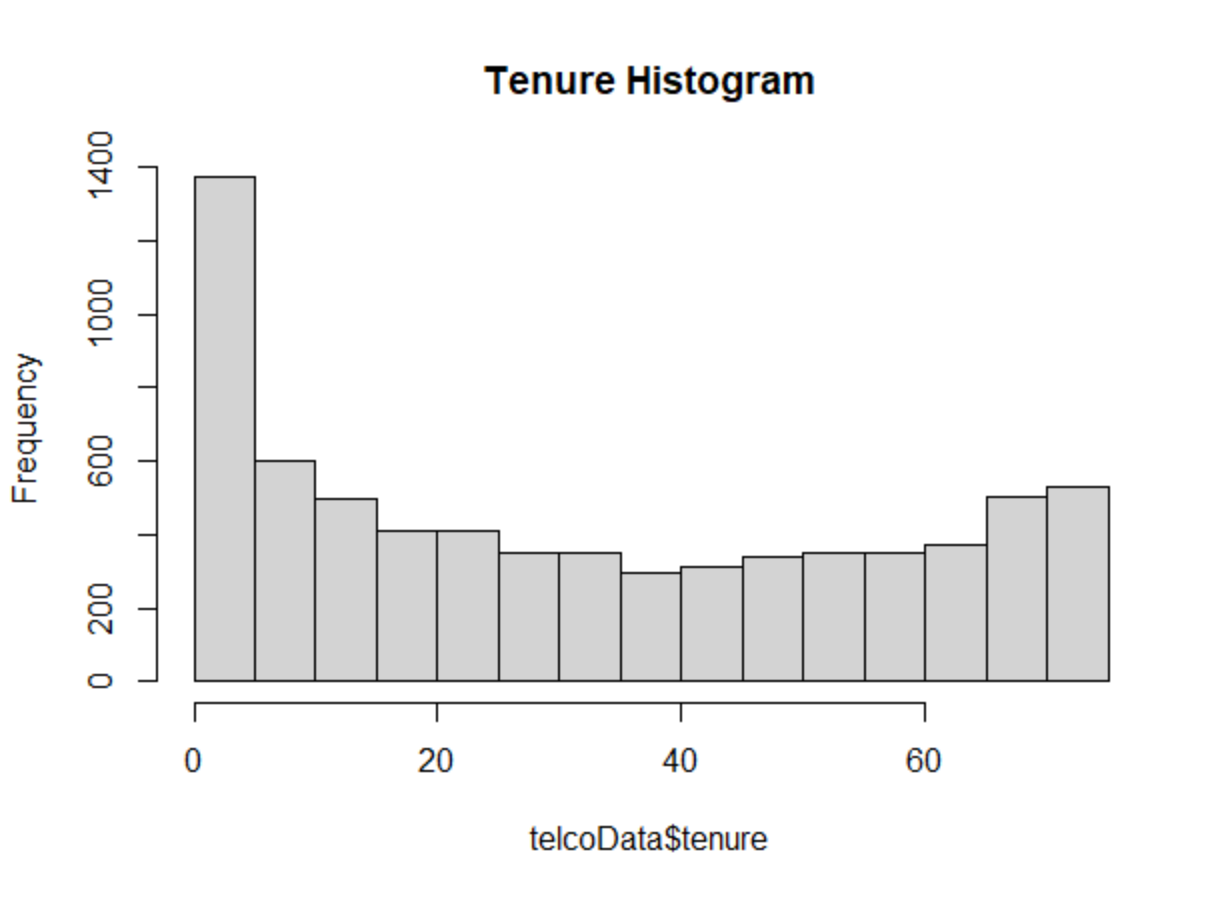
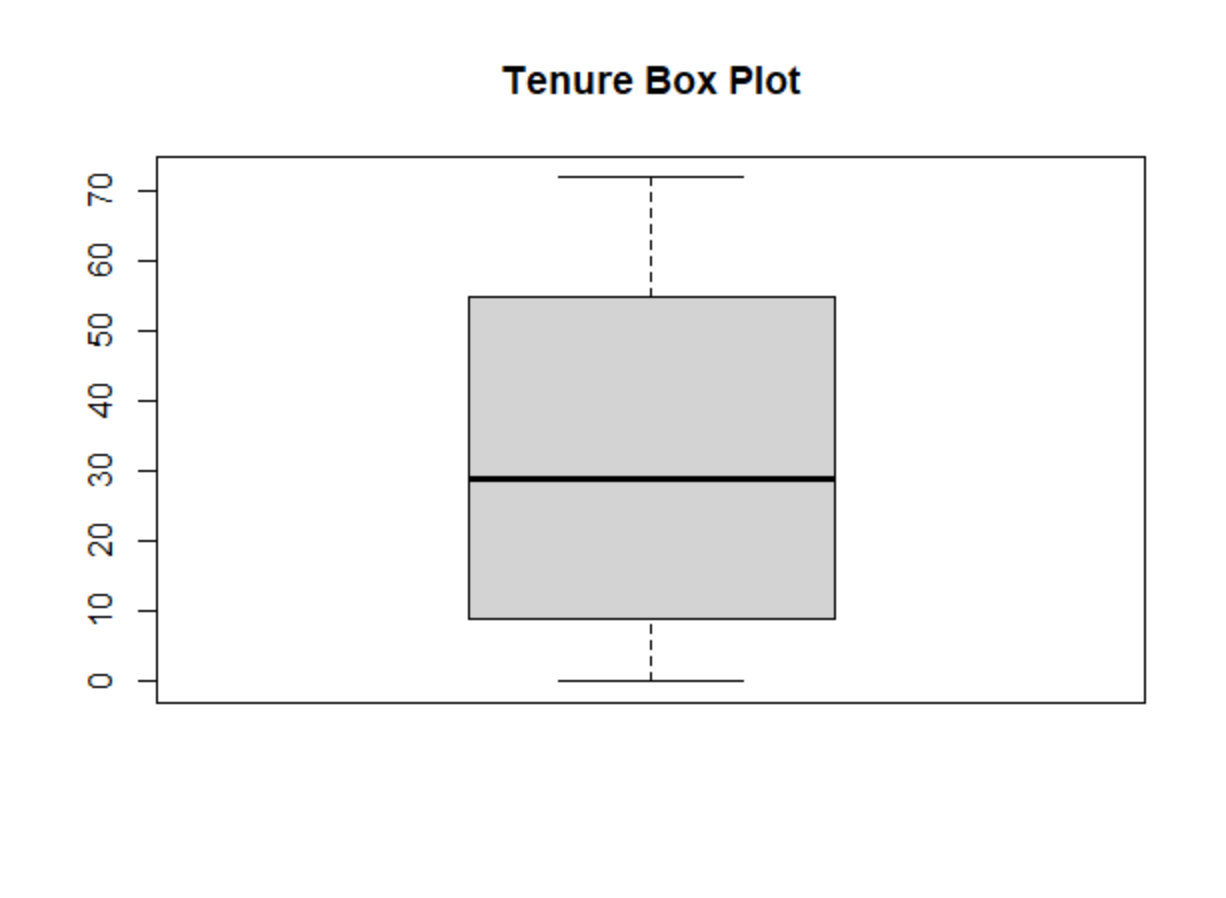
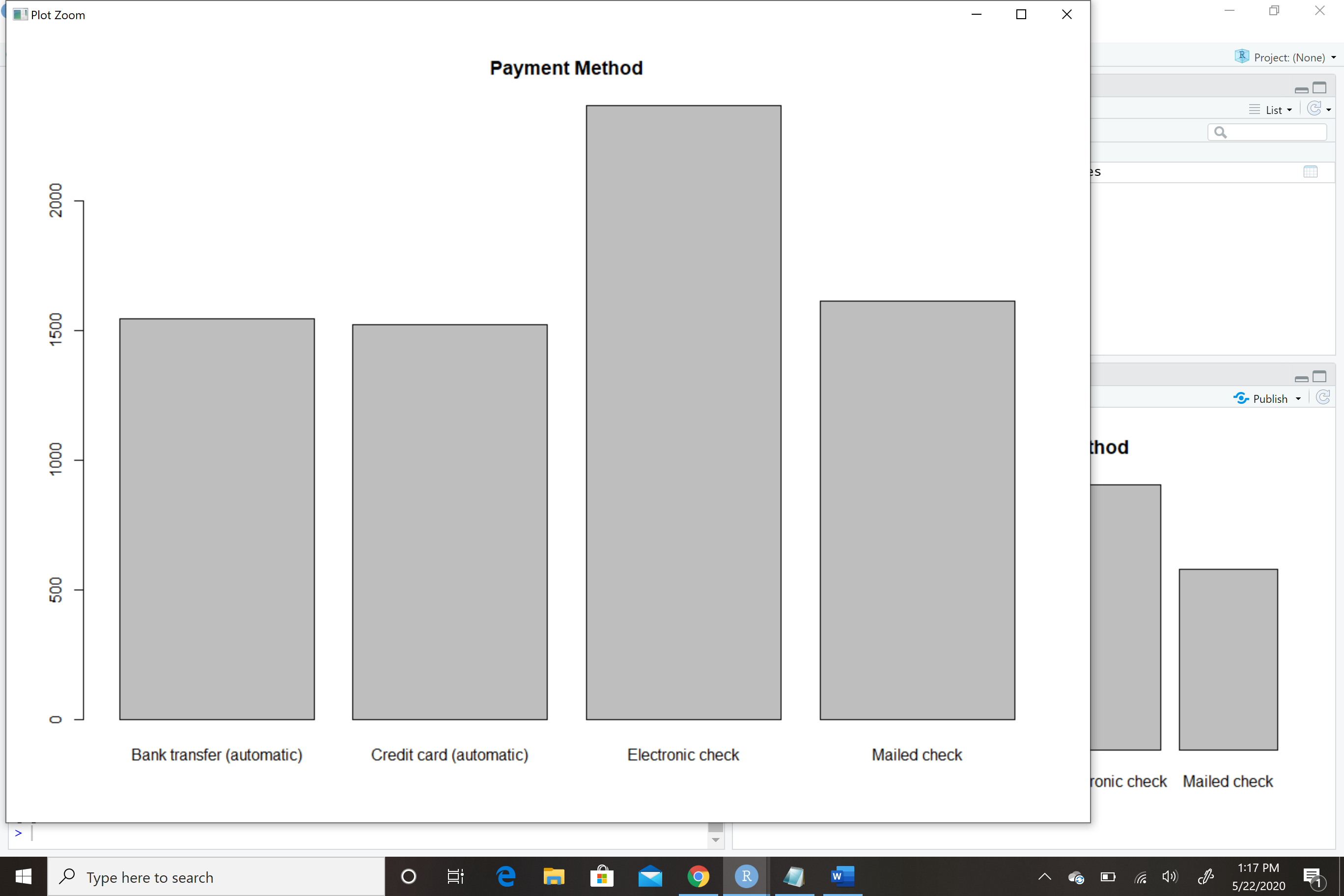
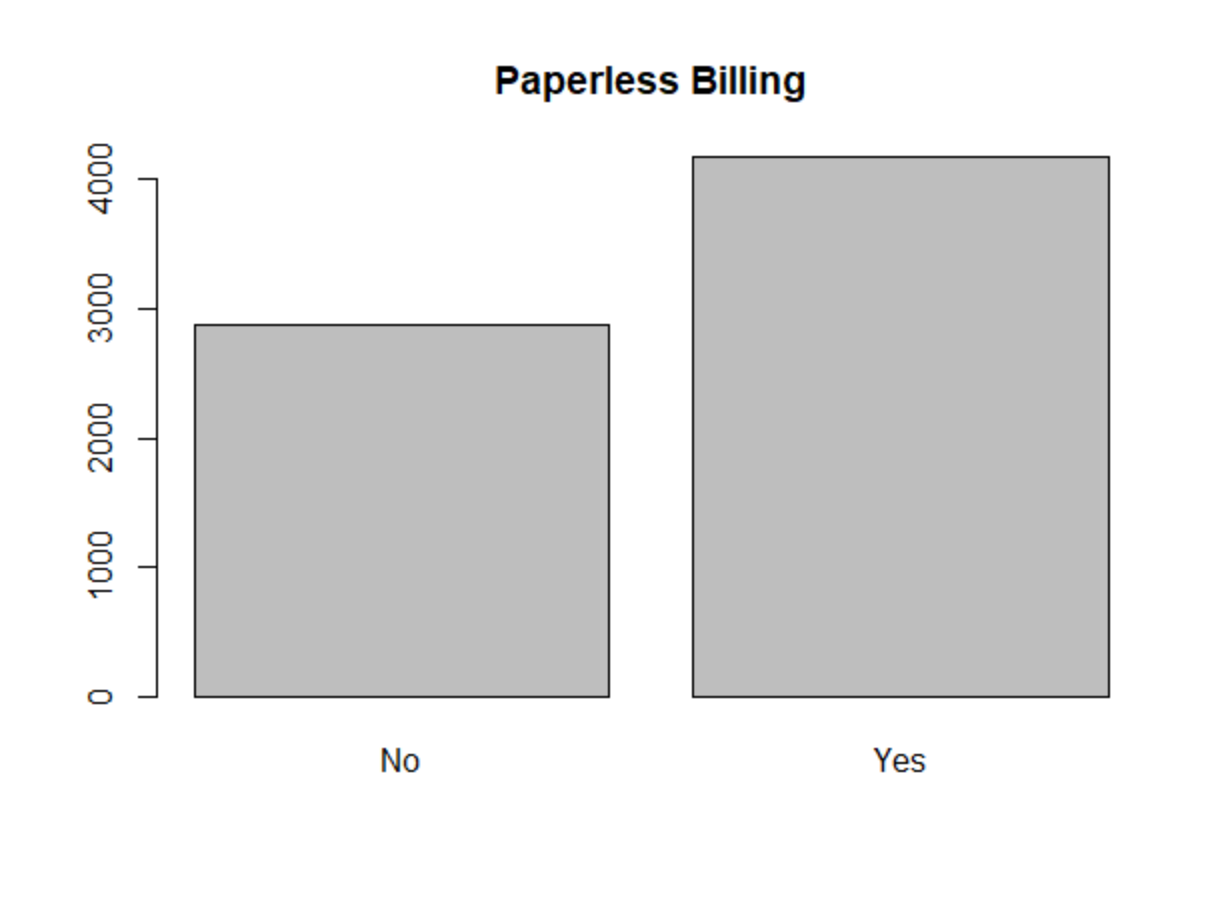
H. Cleaning Data

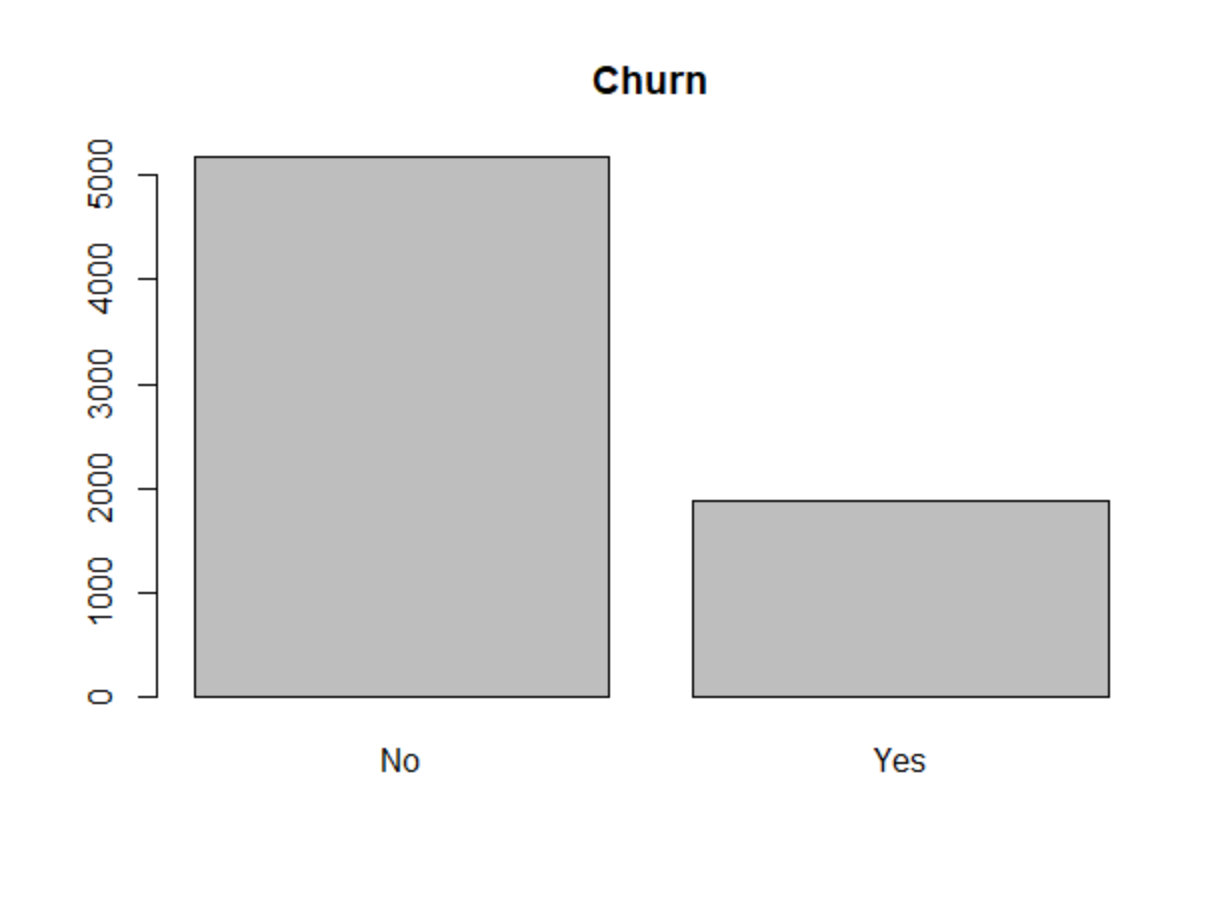
I first observed the data in its unaltered state using str(telcoData) and summary(telcoData). The summary revealed that there were 11 NA values within Total Charges. Upon inspection of this, I saw that the NA values belonged to customers who had recently signed up with the telecommunications company. I changed these values to 0 values. Furthermore, for this particular analysis, the customerID is not necessary; however, the company might decide to use this information later to target and retain or regain customers. I also used summary, boxplot, hist, and plot to observe the various data in order to determine if there were any erroneous values that might skew the analysis.

III. Data Analysis

I. Distribution of Variables Using Univariate Statistics   
By observing the univariate plots, one is able to better understand the customer makeup of the telecommunications company. Most customers are not senior citizens, there is almost a 50/50 split on gender, and most of the current customers are relatively new as seen in the tenure plot.   
  


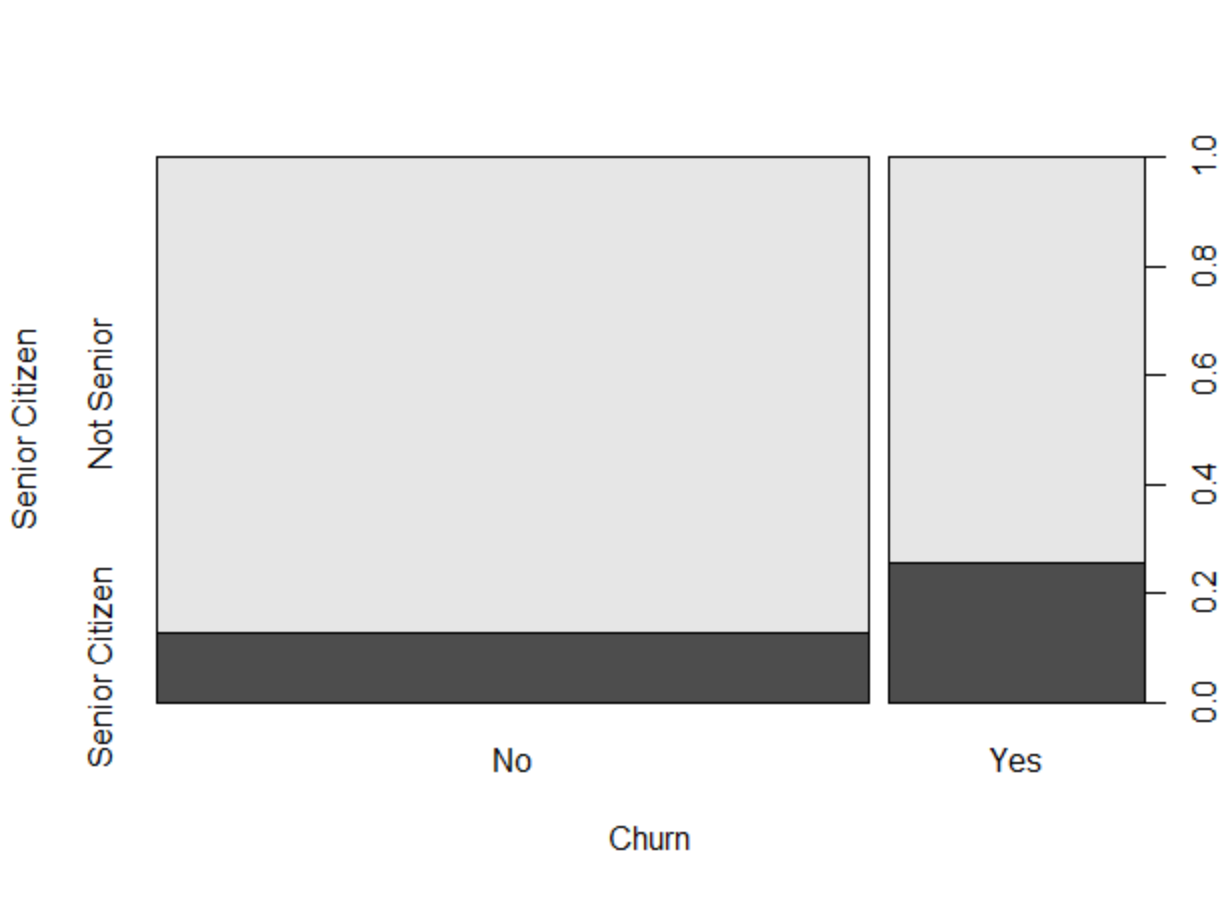
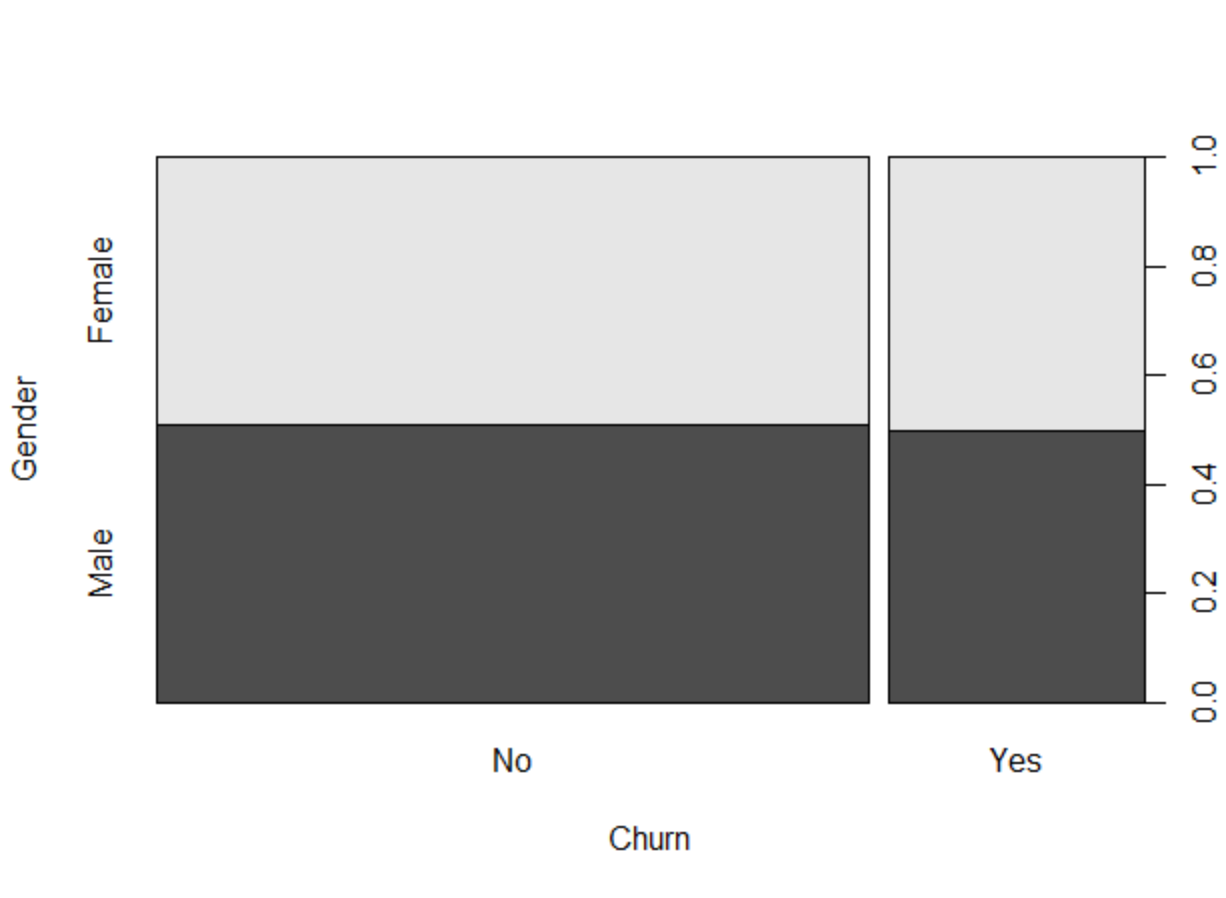


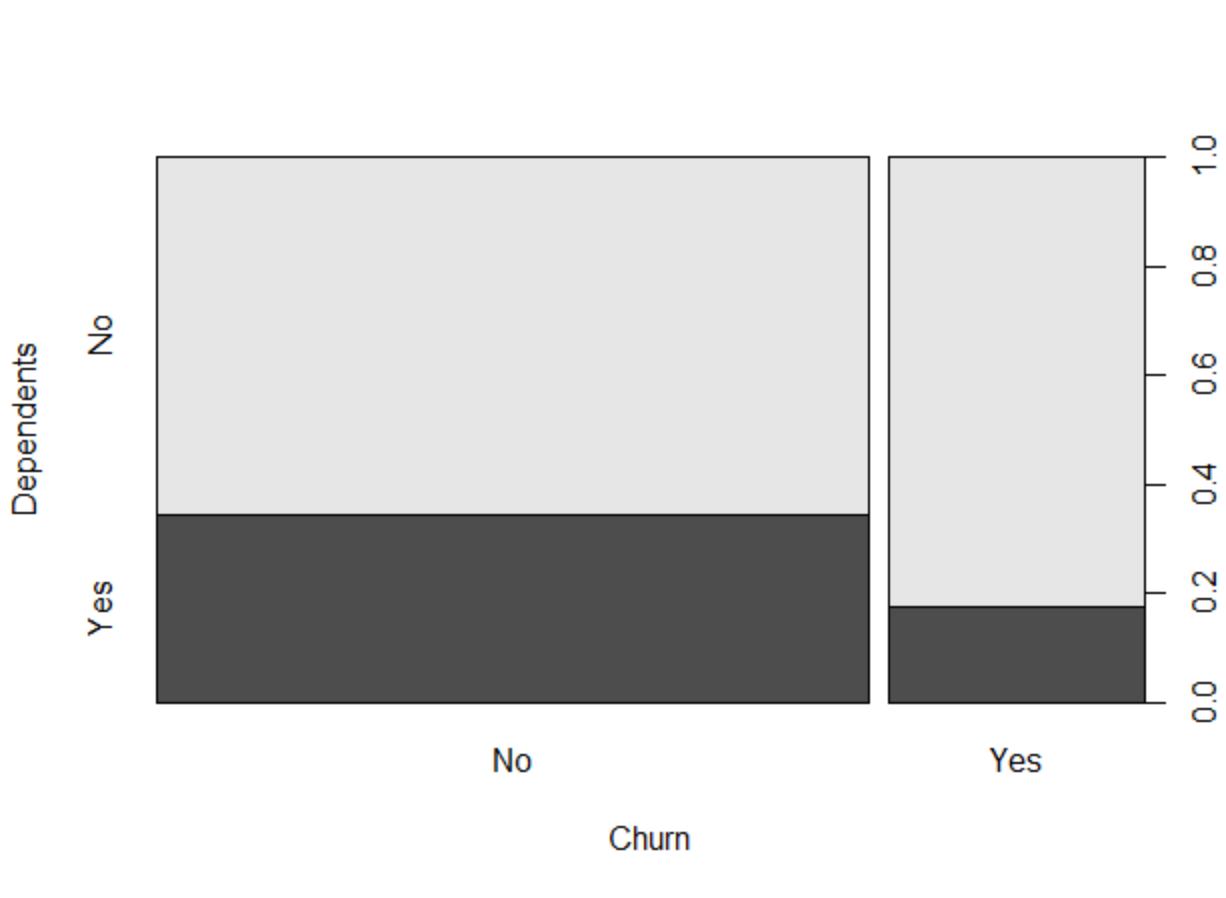
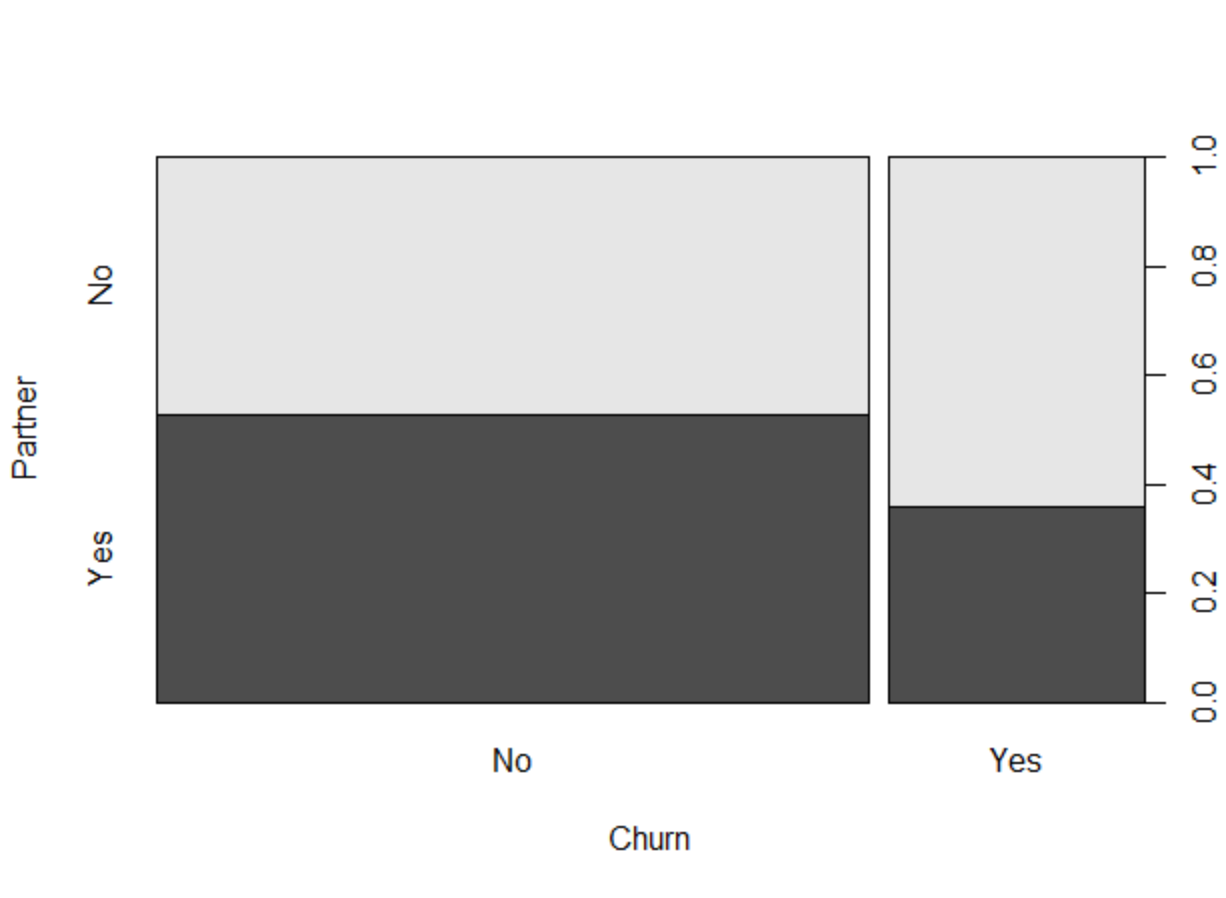


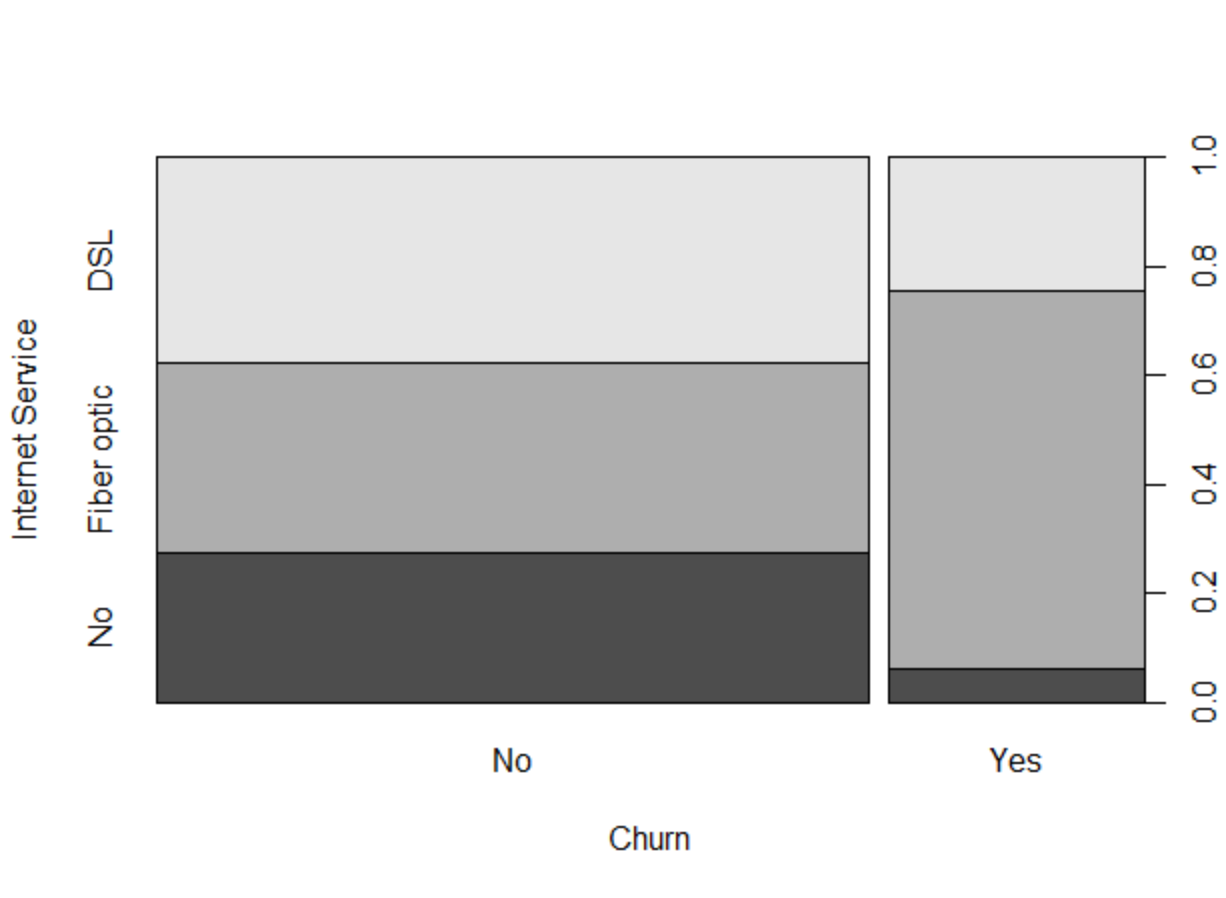
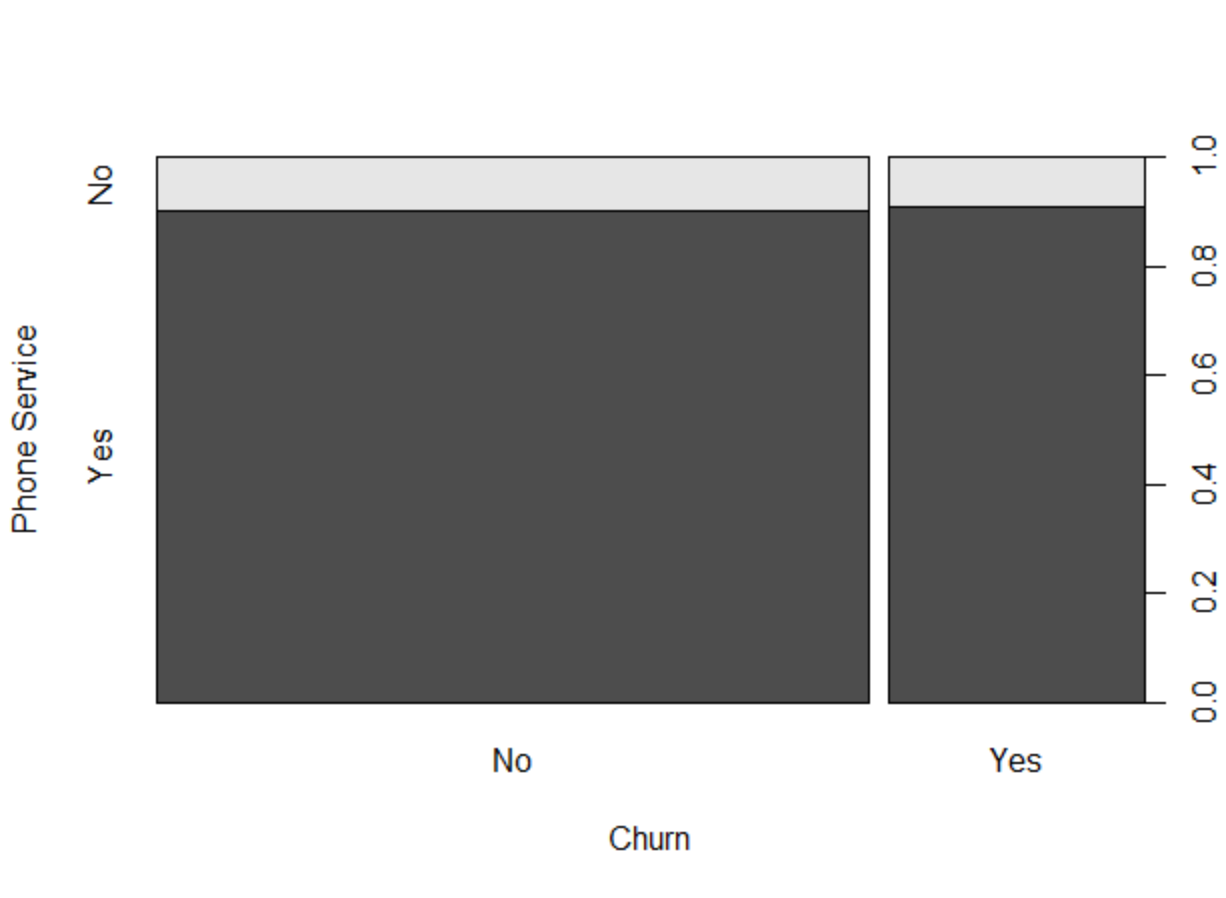


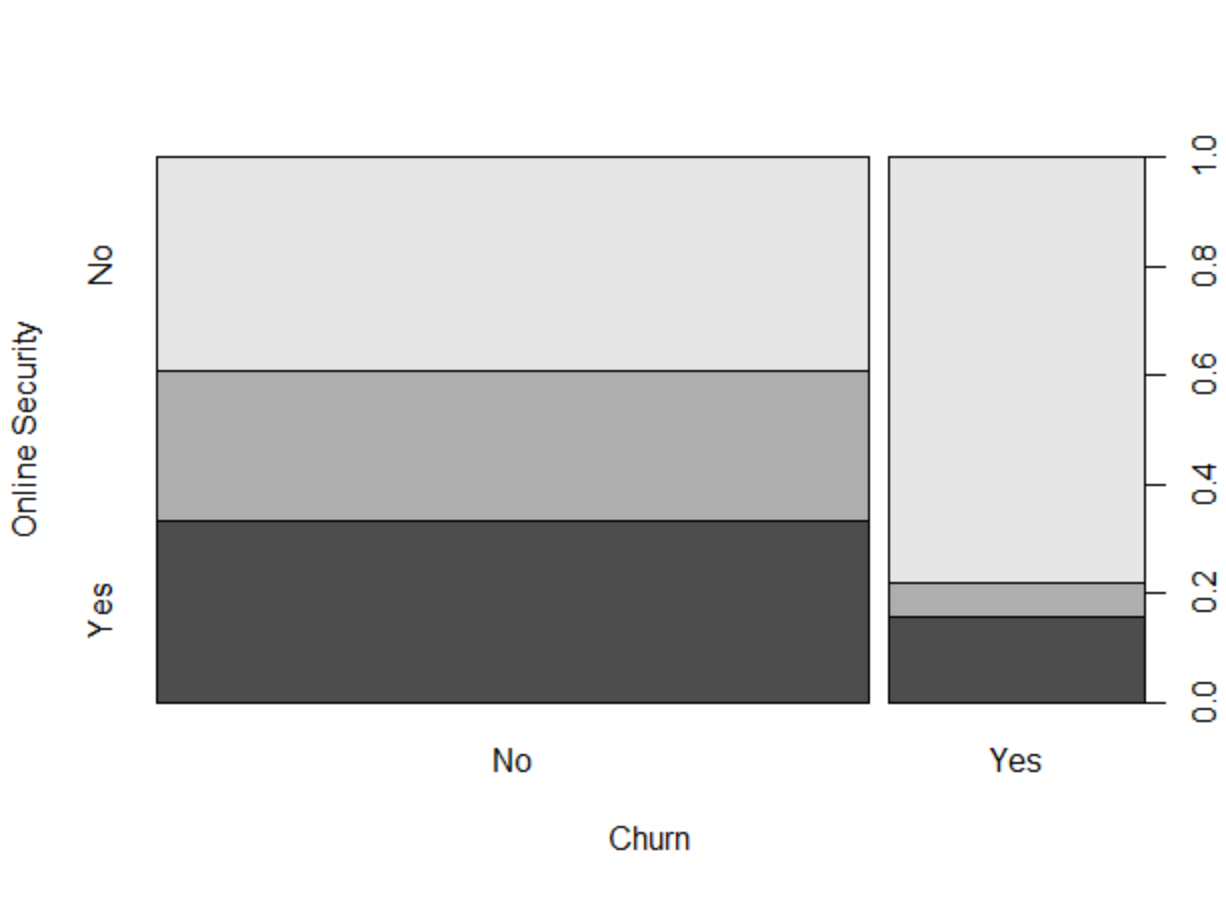
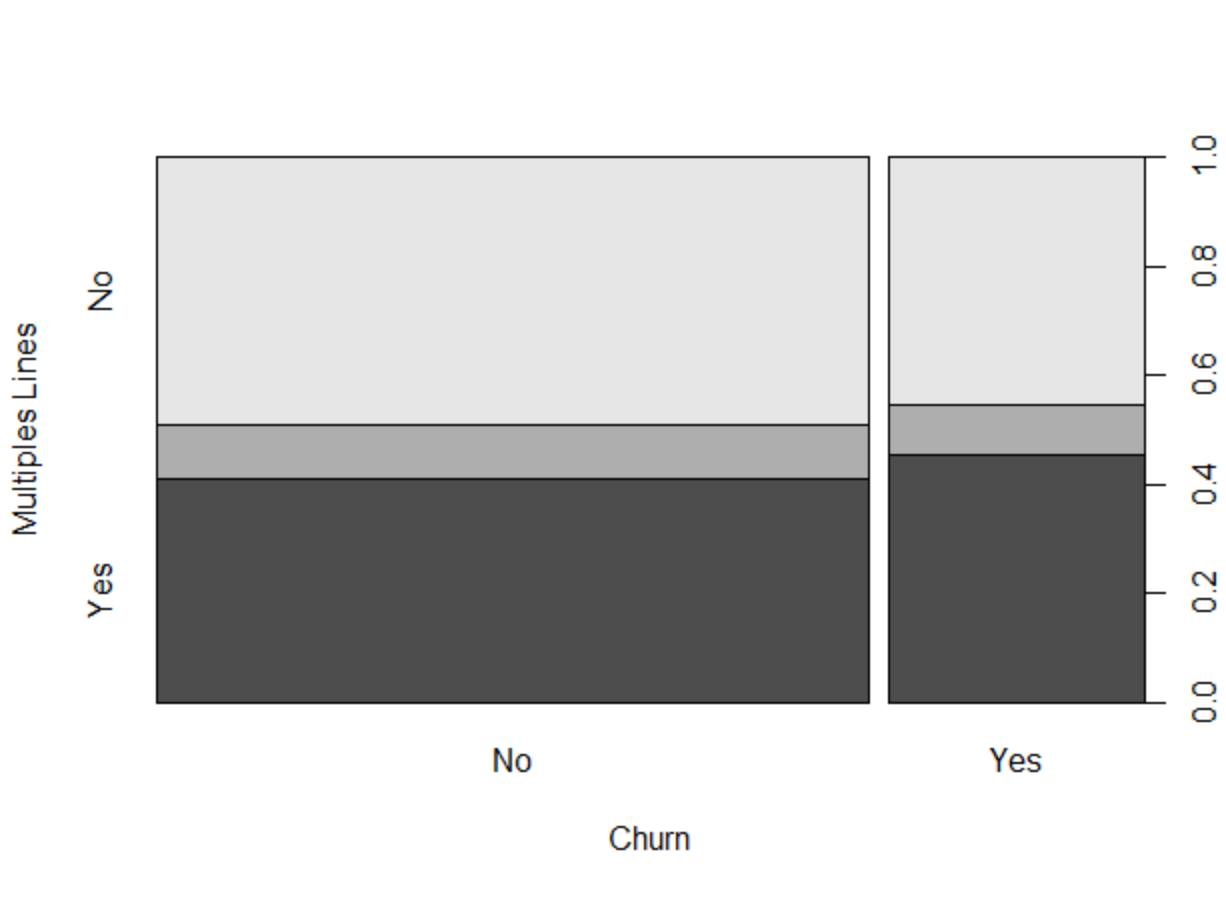
J. Distribution of Variables Using Bivariate Statistics

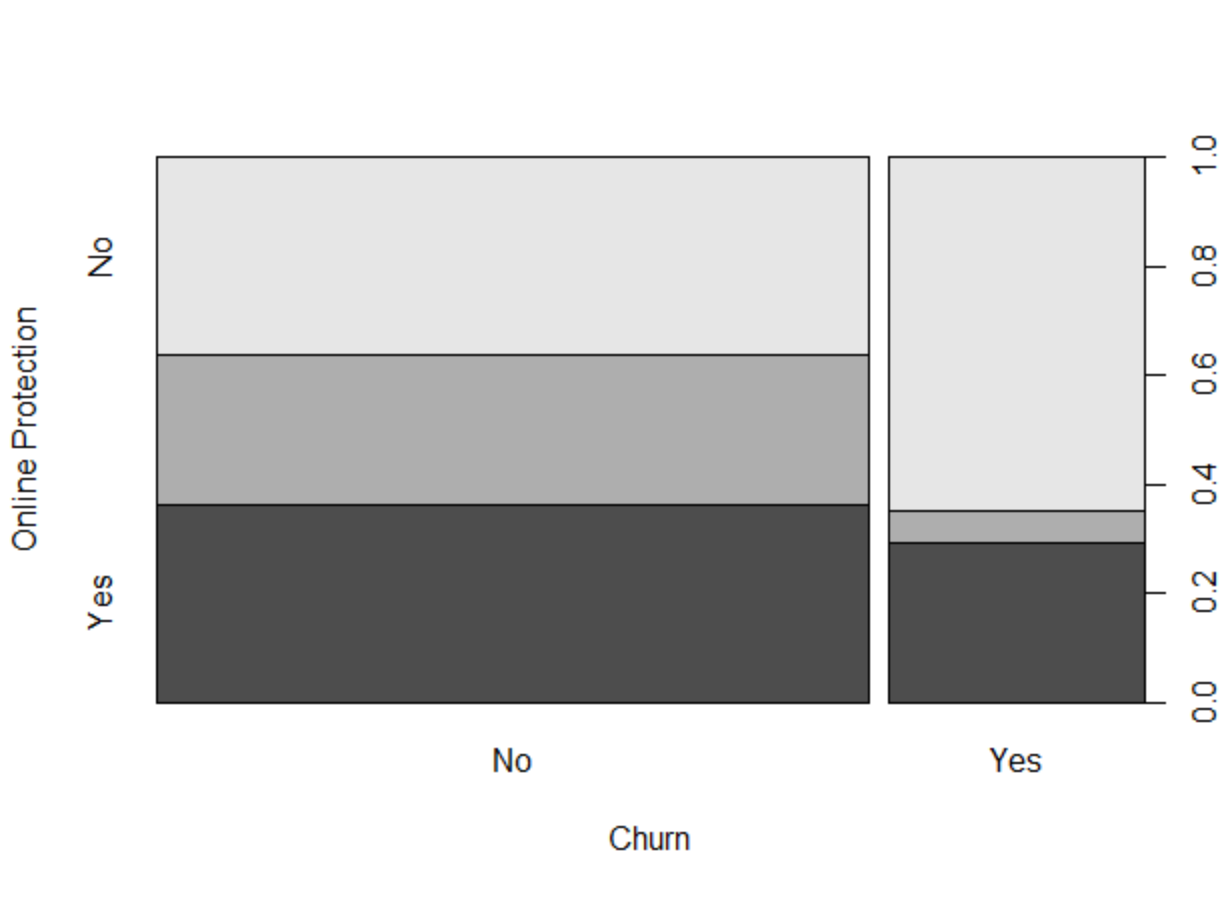
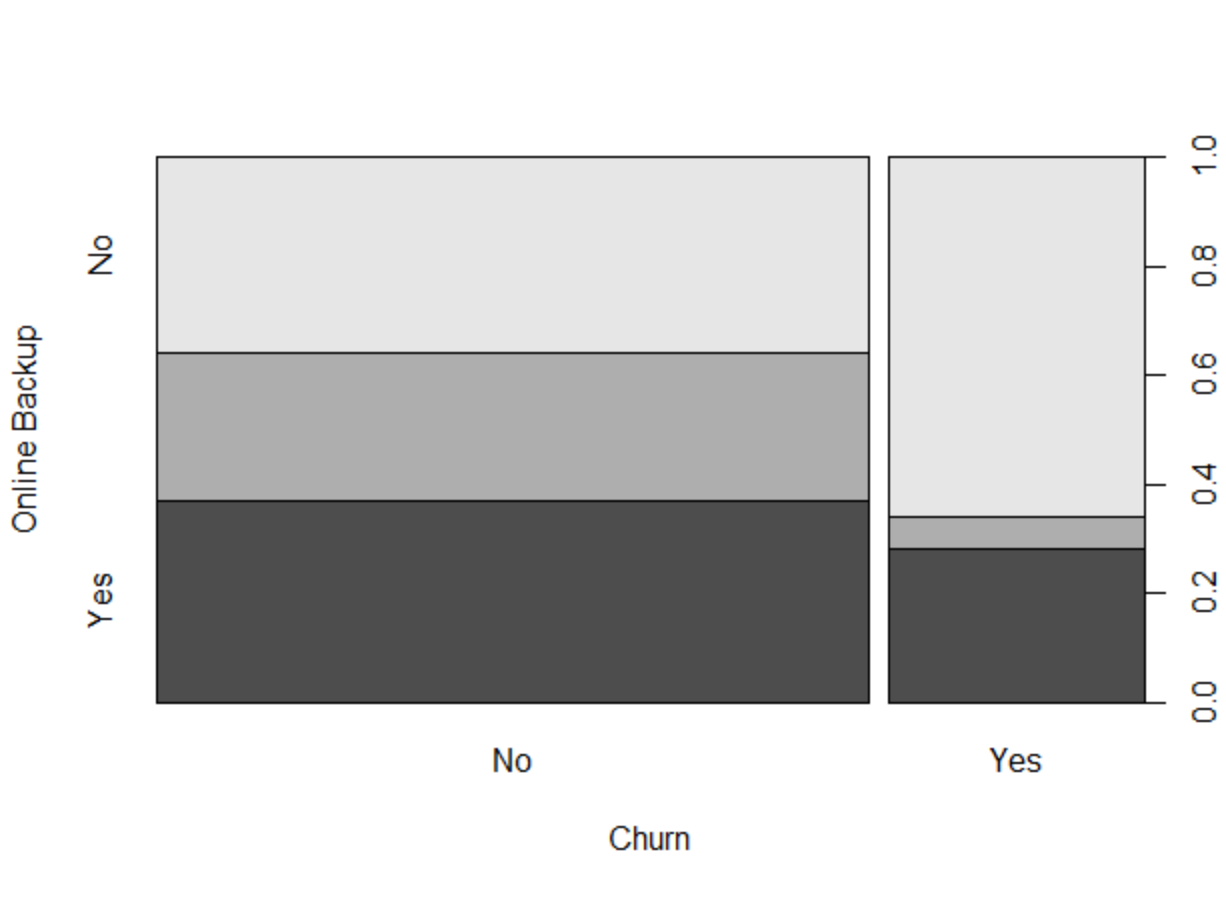
Some things I noticed while looking at the bivariate plots were that there seems to be a large discrepancy between those who have fiber optic internet service and did not leave versus those who have fiber optic internet service and did leave. This is something I would need to investigate further. It could potentially be caused be a higher cost of fiber optic internet service. This would also explain the higher mean monthly cost for those who stopped service versus those who are still with the company. Another notable difference within the plots is that those with month-to-month contracts appear far more likely to stop service than those with contracts. Again, this is something to investigate further to determine if any true correlation exists.

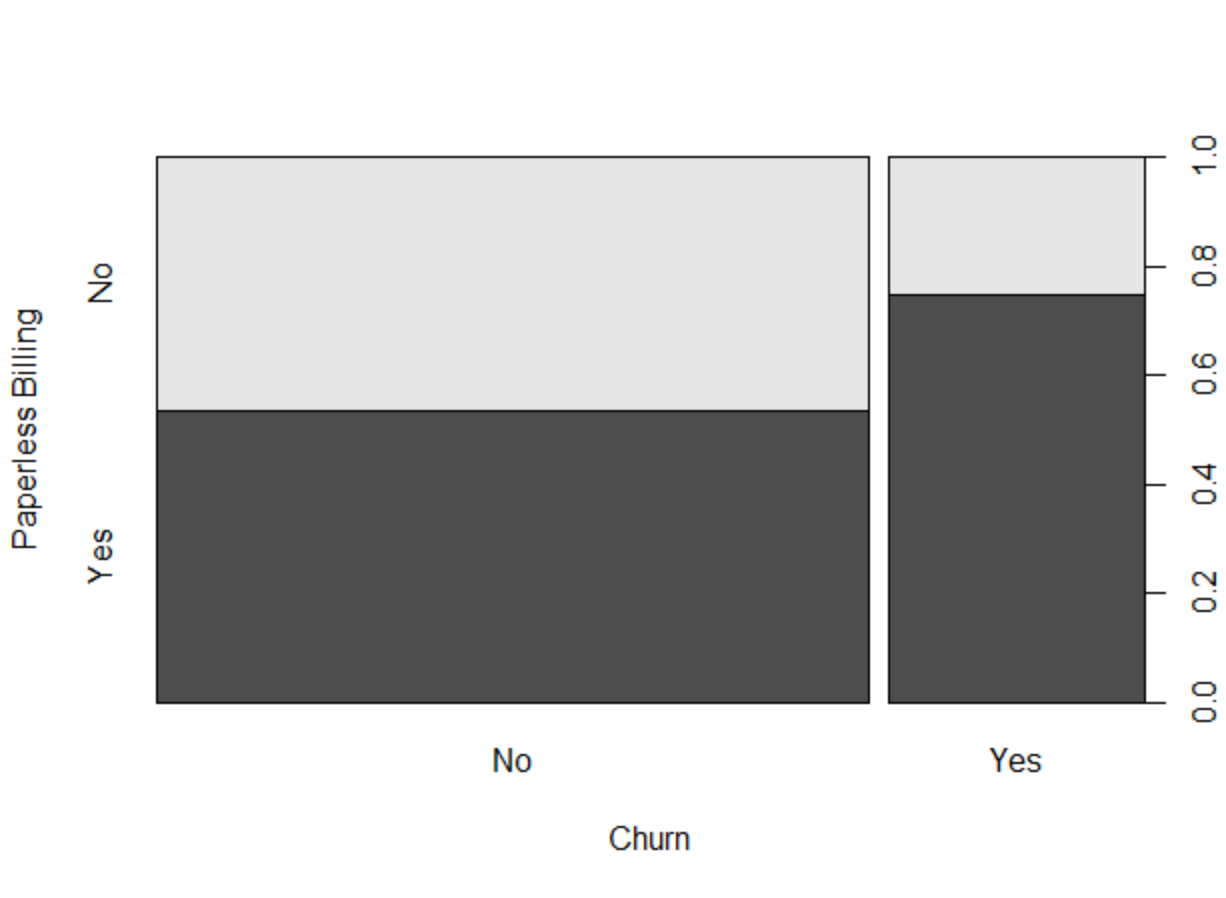
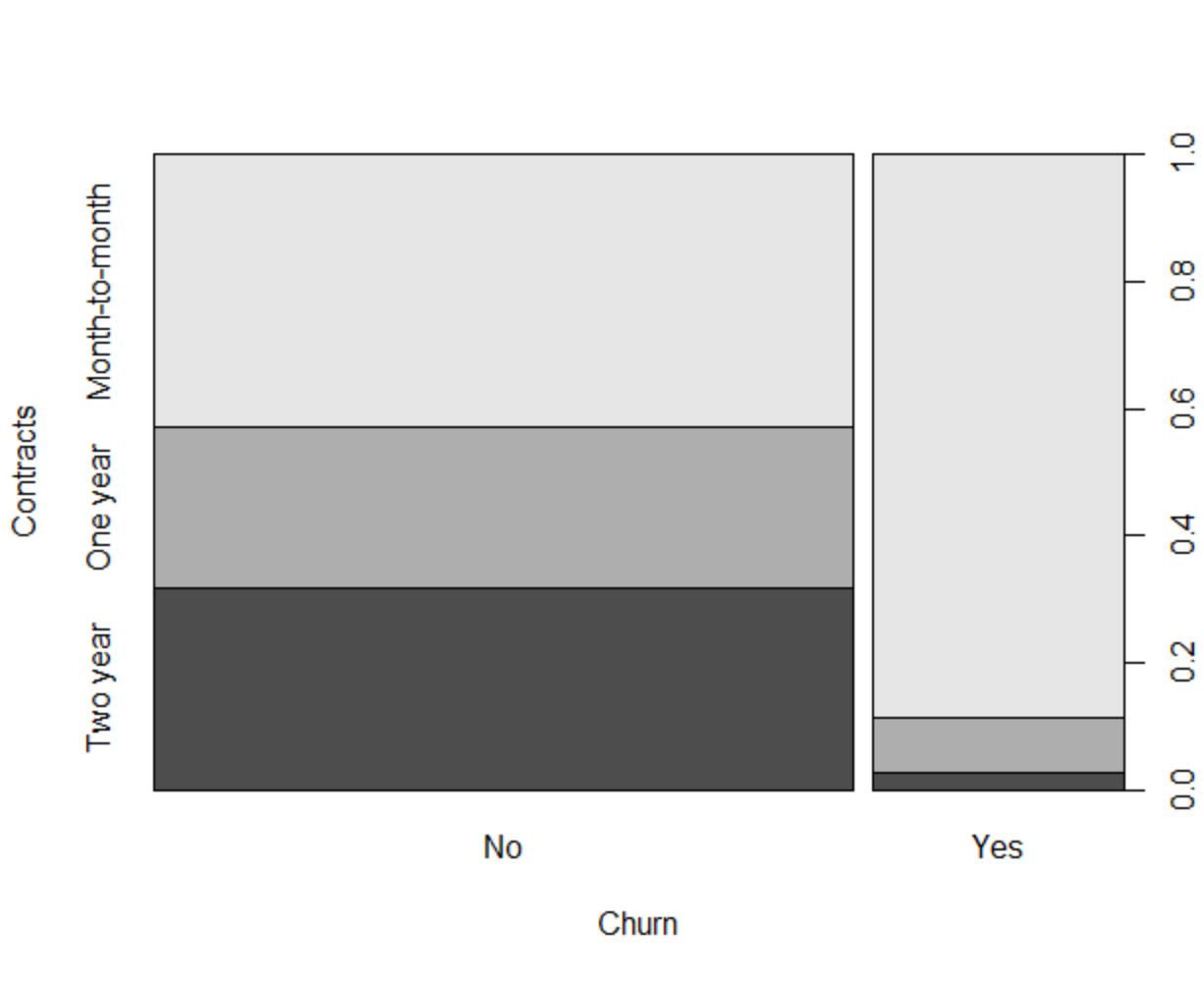
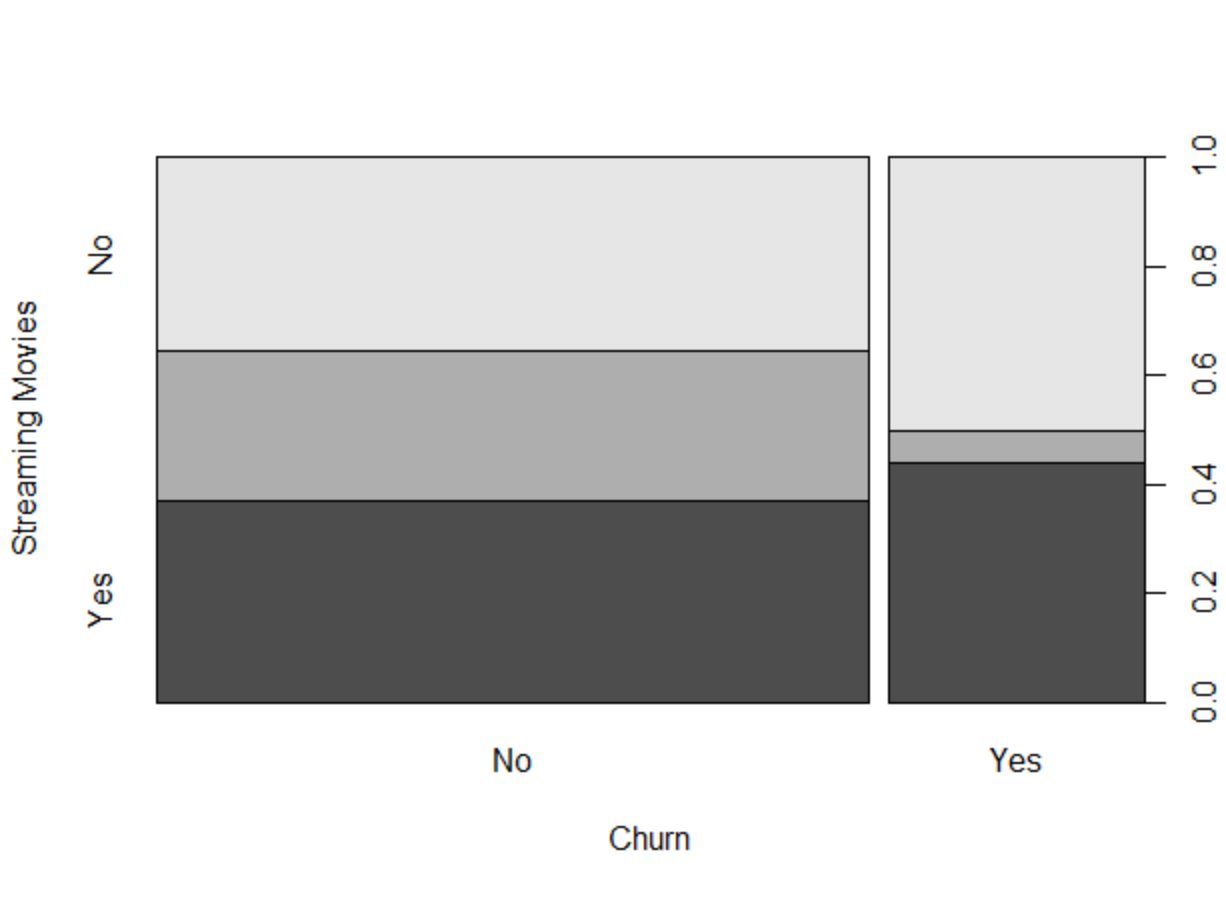
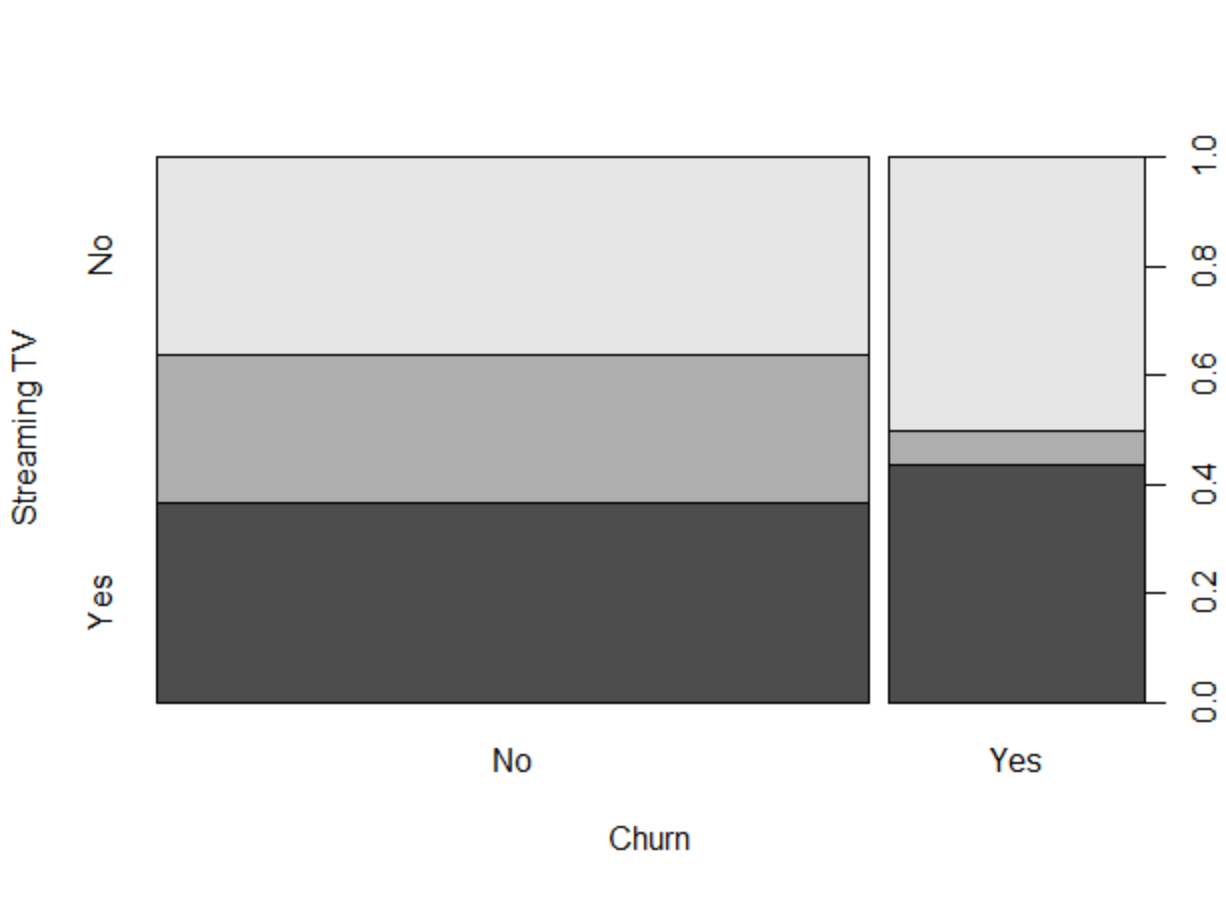
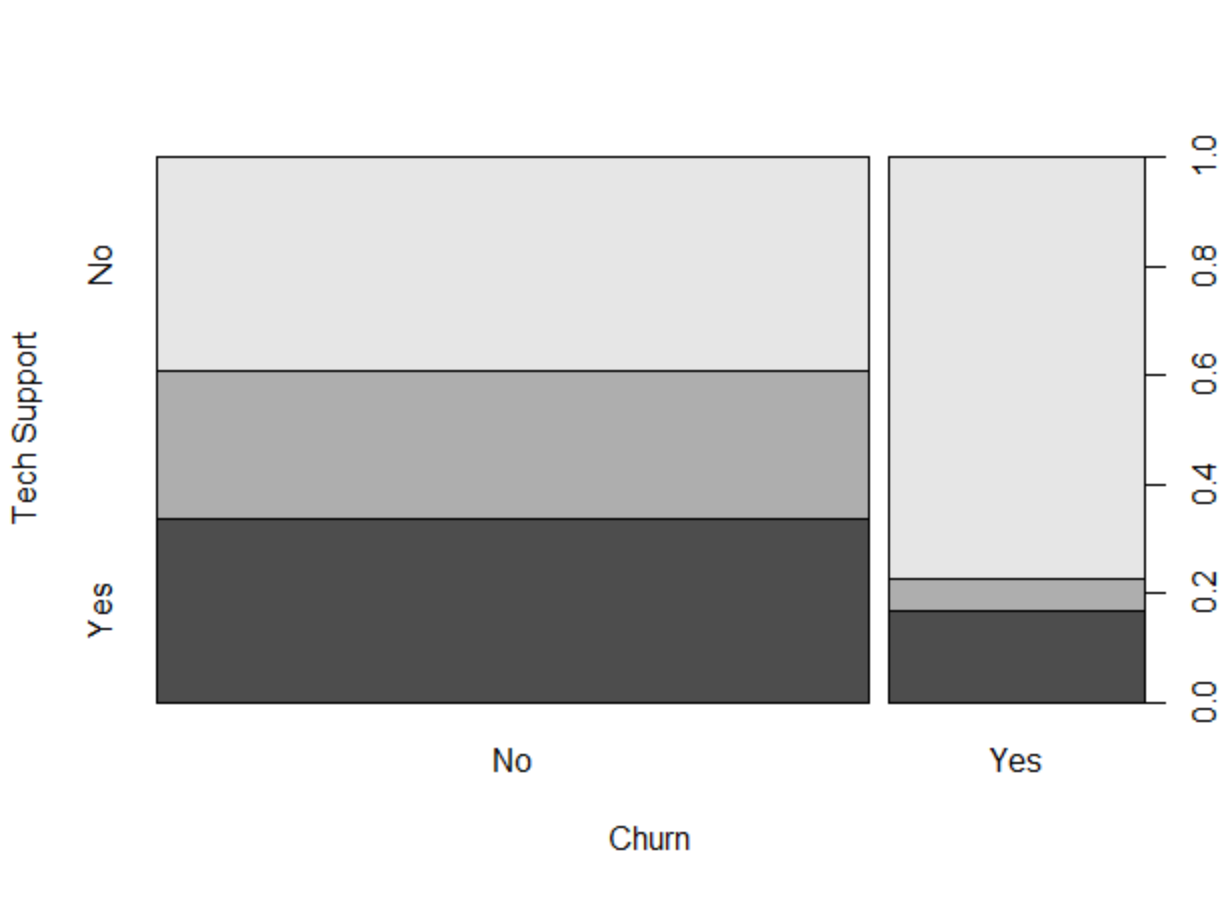
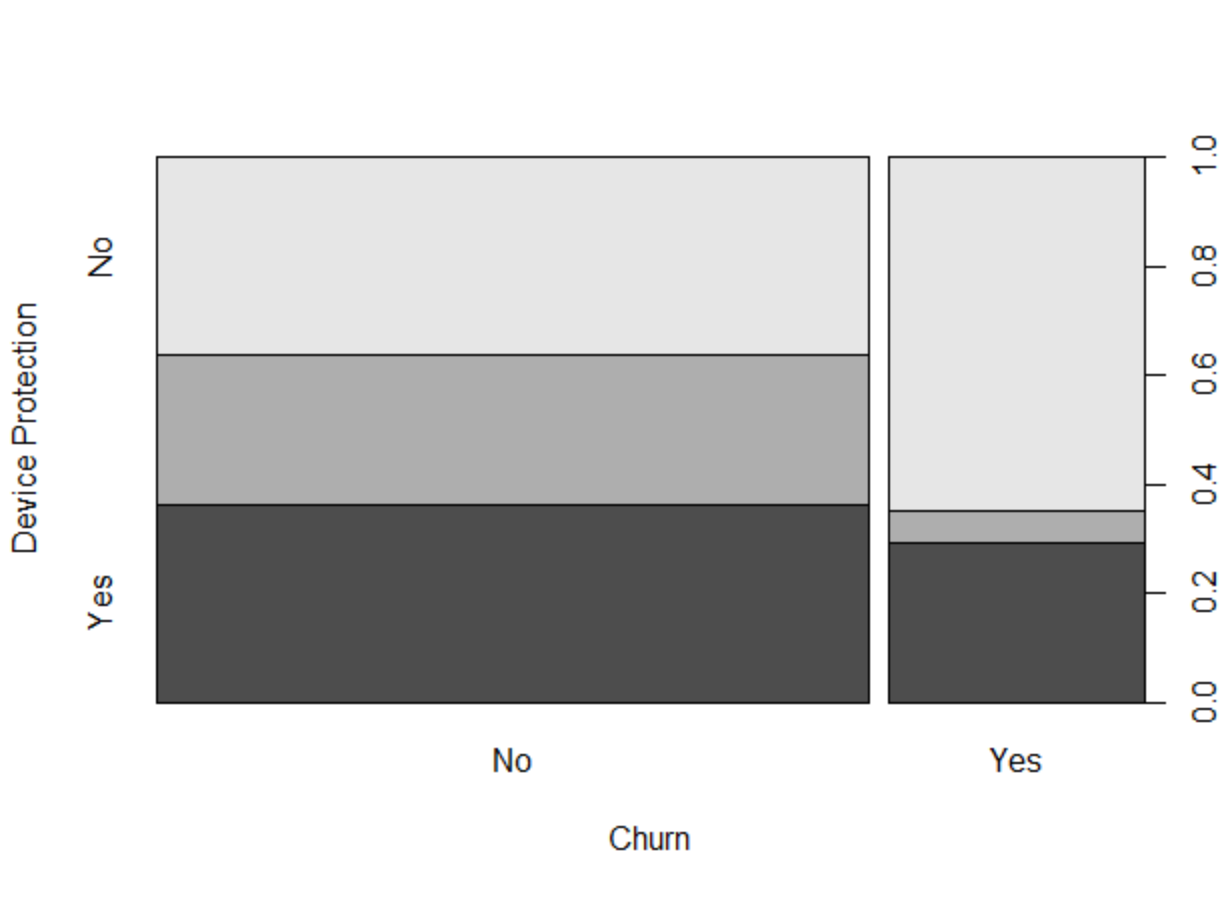


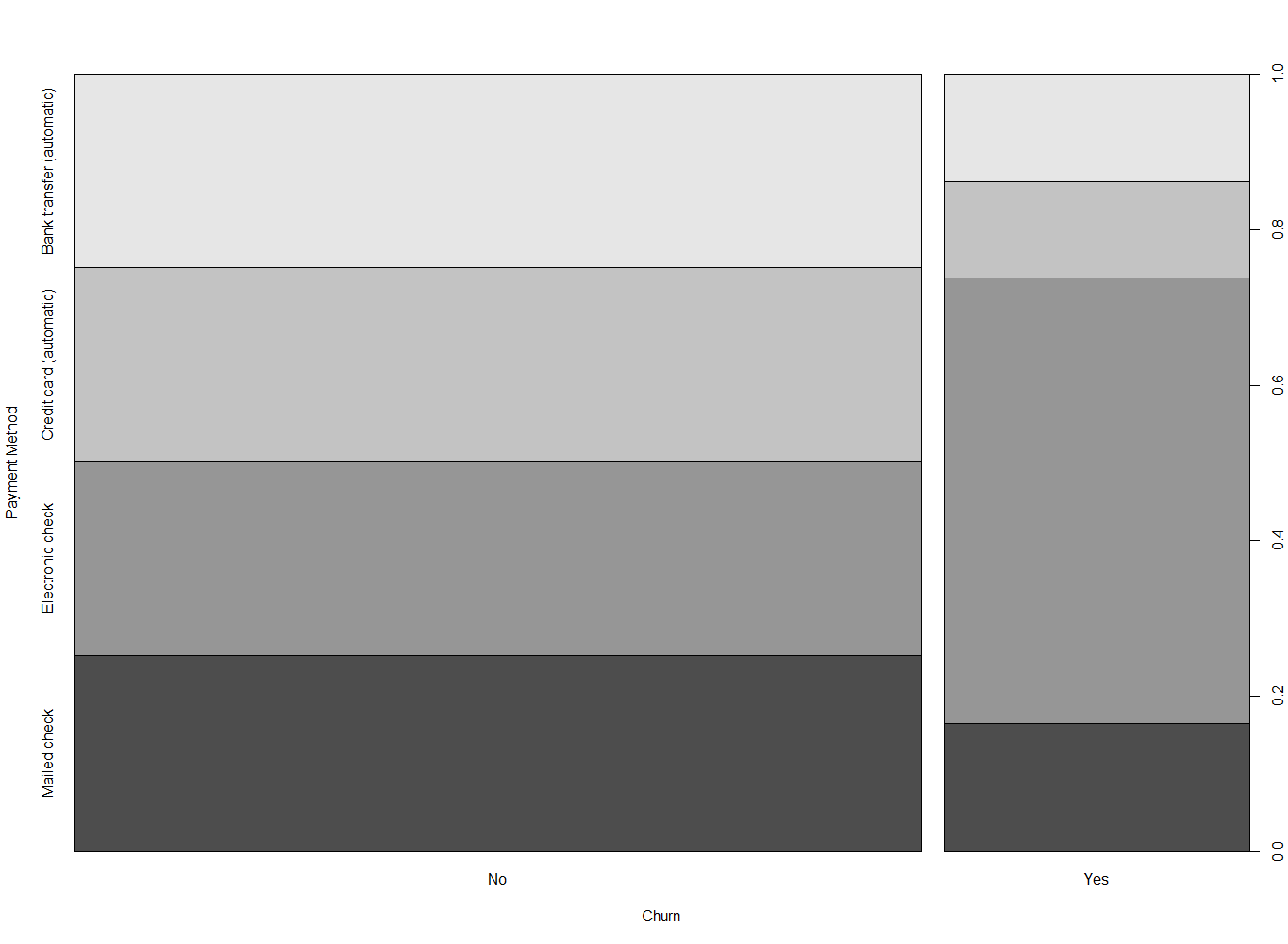


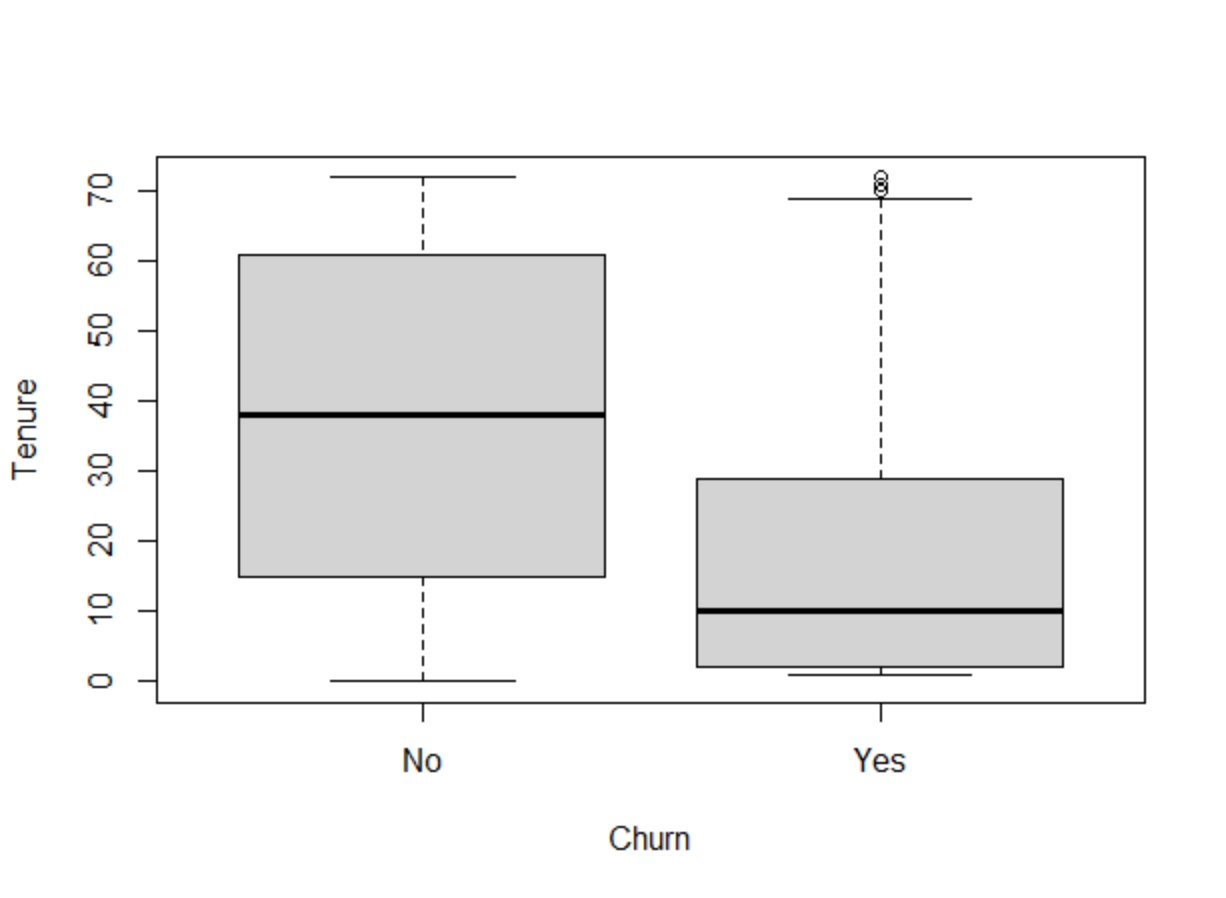
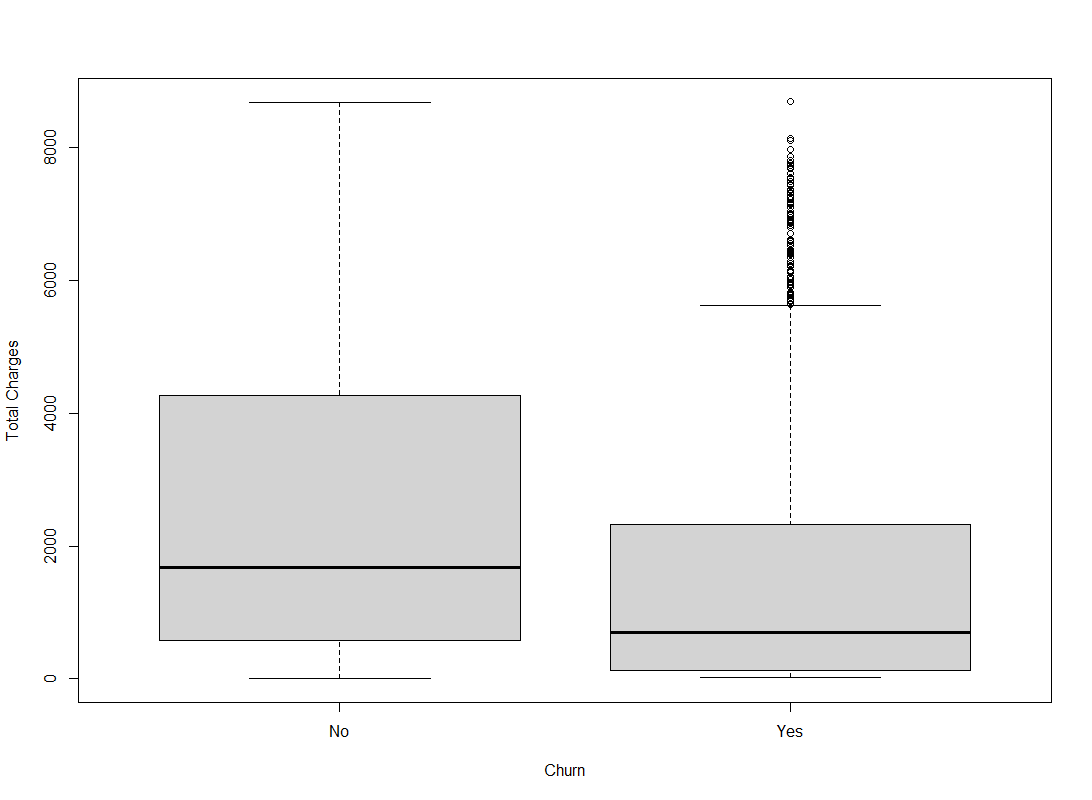




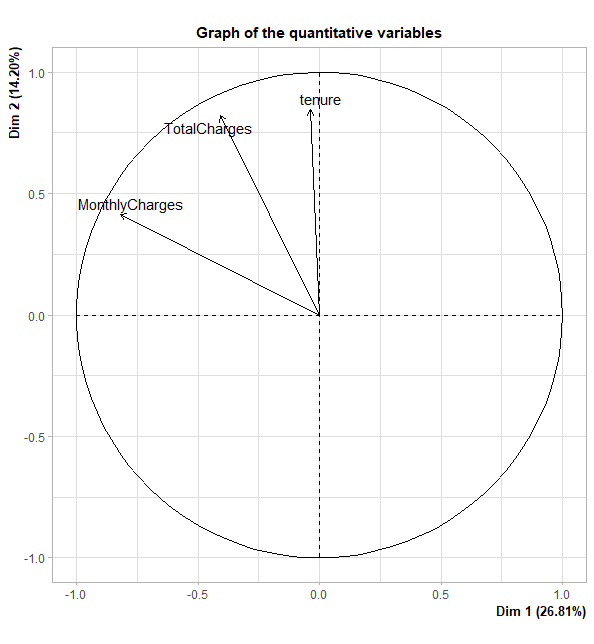




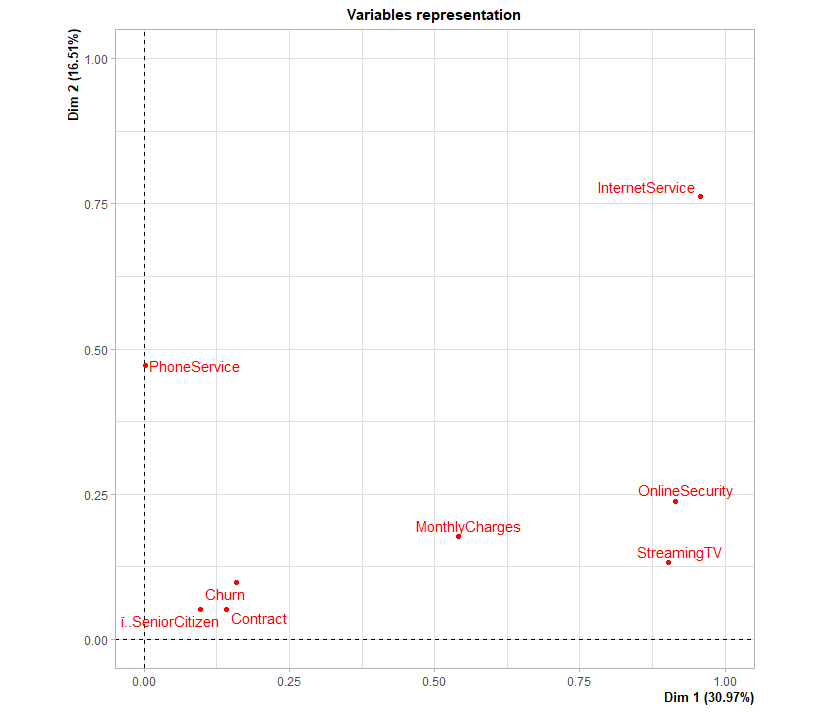




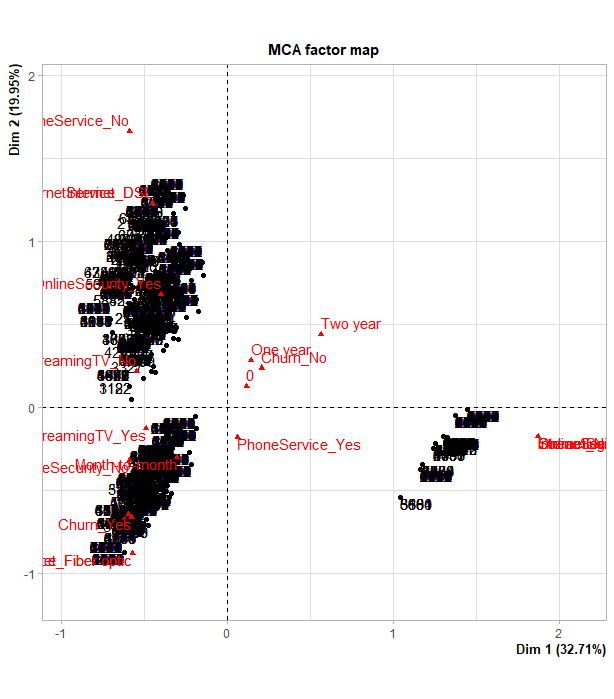
K. Analytic & Evaluative Methods



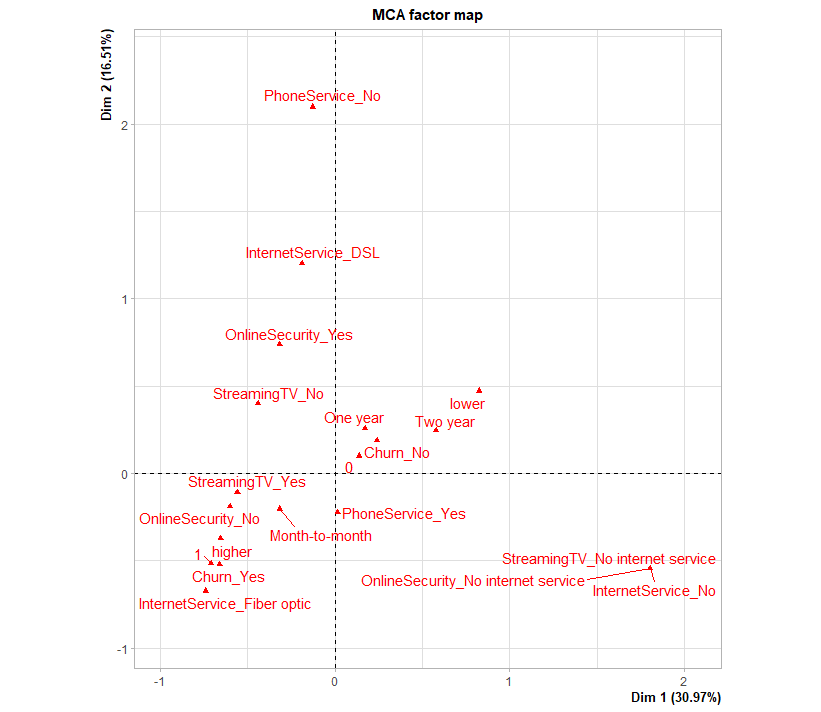
This plot shows the quantitative variables within the data. Monthly charges appear to pull the data towards churning; that is, if a customer pays higher monthly charges, they are more likely to leave the company.



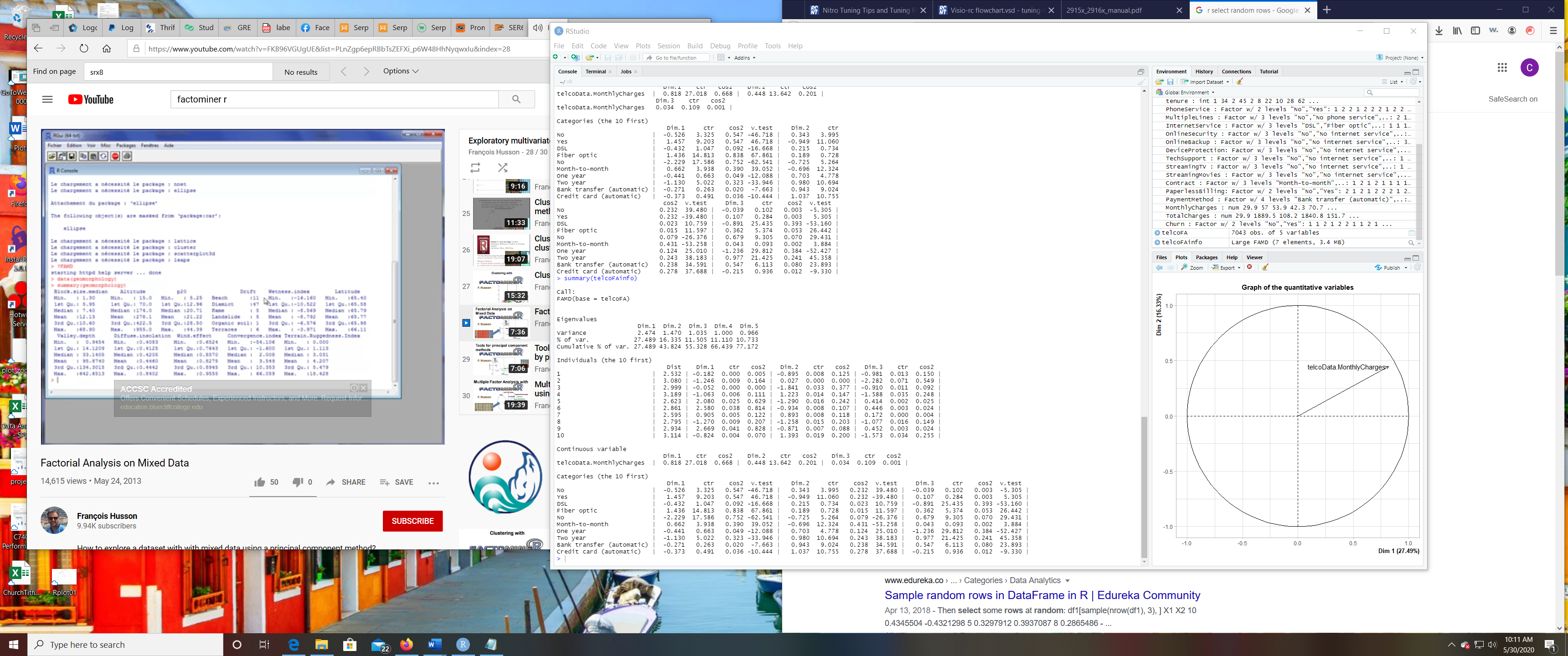
This plot shows the qualitative variables within the data. More will be explained in the graphs below.



Although this chart is somewhat difficult to read, I was able to see that Churn – Yes, a higher than average monthly payment, and Internet -Fiber Optic are in the same coordinate plane and close to each other. Also nearby and thus closely related are Internet Security - No and Month to Month Contract. On the other hand, DSL internet is much further to the churn – yes factor as are the yearlong or two-year-long contracts.

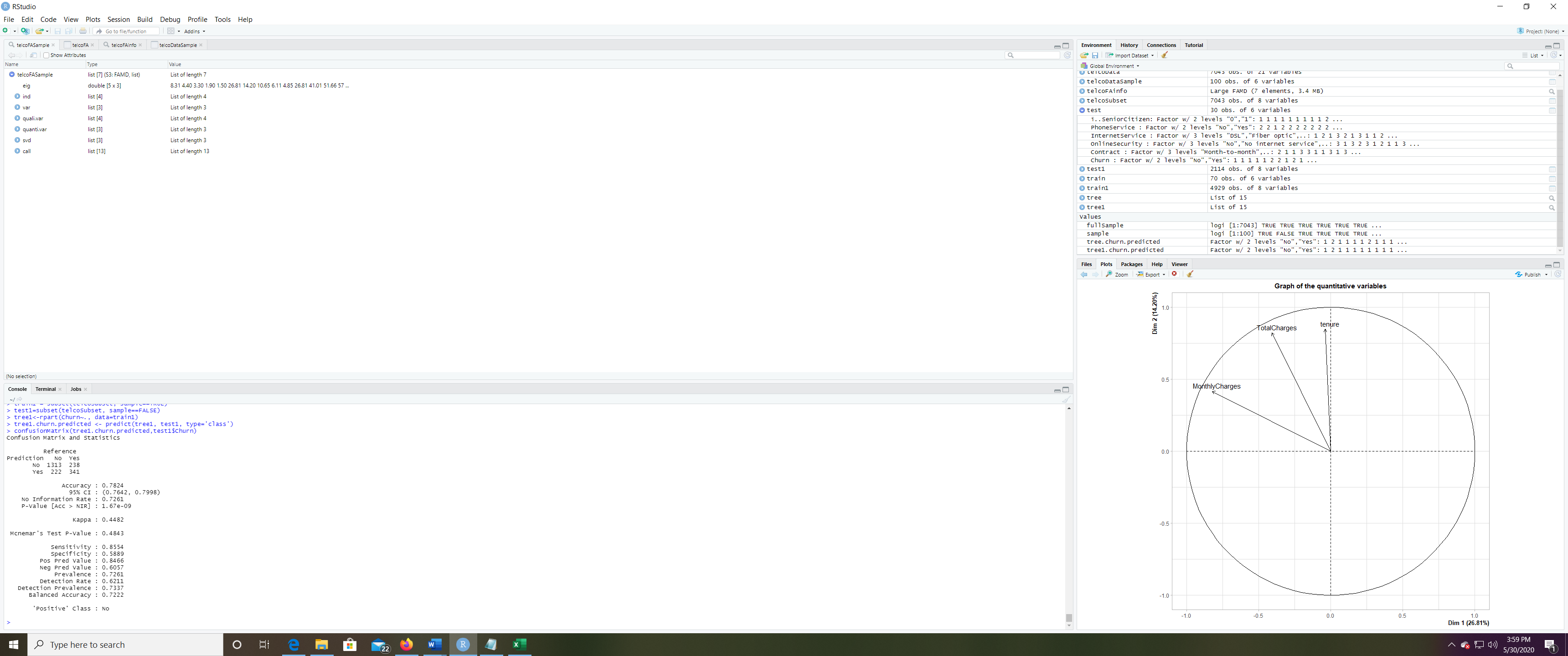


This MCA factor map shows similar results, although it is much cleaner.

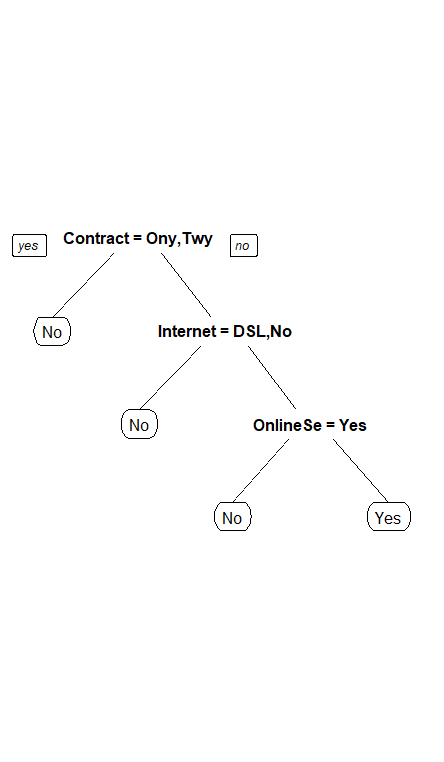


In observing some of this data, it is apparent that DSL results in a lower churn rate while fiber optic internet results in a higher churn rate. Similarly, month-to-month contracts result in a higher churn rate, while one to two-year contracts result in a lower churn rate. The type of payment doesn’t appear to have a huge impact on the churn rate.

Confusion Matrix on Churn



Based on this chart, for our tree, it predicted 1313 would not churn, and that was accurate with the data tested. Likewise, 341 were predicted to churn, and that was accurate with the data tested. The accuracy is about 78% with a p-value much smaller than 0.05 making it statistically significant.



Above is the decision tree created. In essence, if the customer has a contract it predicts they will not churn. If they do not have a contract and have DSL internet or no internet, it predicts they will not churn. If they do not have a contract, do not have DSL internet but are paying for some internet service, and have online security, they are predicted not to churn. However, if the customer does not have a contract or online security for their Internet service (excluding DSL), then they are predicted to leave the company.   
  
L. Justification of Methods for Analysis  
I chose to use the factor analysis for mixed data function (FAMD()) because the data set that we had included variables that were both integers and factors/character strings. Later, I chose to use MCA, multiple correspondence analysis as it focuses on categorical variables and would best let me see the correlation within factors. In addition, a decision tree was used as a predictive method which would allow the telecommunications company to best identify those individuals its at risk of losing. The confusion matrix helped to illustrate the effectiveness of the decision tree by predicting customer churn with 78% accuracy.

M. Justification of Methods for Visuals

The bar graphs used with the univariate/bivariate analysis of the data allowed for a quick visual comparison of the data. Immediately, I was able to see large gaps between current customers and former customers in categories like Internet Service and Contract. The factor maps also allowed for the quick ability to pick out closely related variables such as the fiber optic service and month to month contracts. It allowed me to isolate variables and their results. By using the factor maps, I was able to filter the data and represent both “individuals and categories simultaneously on the same plane”, one of the many benefits of this easily read visual (Tufféry, S., Retrieved 2020). Finally, the decision tree was used as a simple way to allow the telecommunications company to make predictions on customers who would possibly leave the company. Thus, they could make some type of corrective action in order to retain the largest number of customers possible. This is in contrast to using possibly a neural network if the categorical variables were altered to be binary which in addition to being difficult to code would also not be an easily followed visual.   
  
IV. Data Summary  
N. Discrimination of Data

According to the text, “a decision tree… detects the most discriminating variables automatically,” so thus by following its nodes, one can determine the factors that best identify data churn. This means that the objective of the research has been accomplished: we have successfully identified certain traits of customers who are likely to churn. Customers who do not have a contract, use fiber optic internet service, and do not have internet security are likely to churn. Furthermore, using the confusion matrix, it is shown that the decision tree making predictions on this customer churn is fairly accurate in its predictions.   
  
O. Methods for Detection of Interactions

Again, thankfully the decision tree worked to remove those values which might not have had a strong impact on churn (non-discriminating data), and it worked to remove variables that were closely related (interactions). For instance, although there are numerous factors that affect the monthly cost, it was removed from the decision tree because of its proximity to the internet service variable.   
  
P. References

Tufféry, S. *Data Mining and Statistics for Decision Making*. [Western Governors University]. Retrieved from https://wgu.vitalsource.com/#/books/9780470979280/