C744 Proficiency Assessment

Jonathan DiQuattro

Data Mining and Analytics II C744

Daniel Smith, PhD

December 23, 2020

Abstract

This paper is being submitted as a final assessment for the C744 Data Mining and Analytics class, and demonstrates the skills learned throughout the course via a presentation to a telecommunications company that wants to mitigate customer churn.

C744 Proficiency Assessment

**You are an analyst for a telecommunications company that is concerned about the number of customers leaving their landline business for cable competitors. The company needs to know which customers are leaving and attempt to mitigate continued customer loss. You have been asked to analyze customer data to identify why customers are leaving and potential indicators to explain why those customers are leaving so the company can make an informed plan to mitigate further loss.**

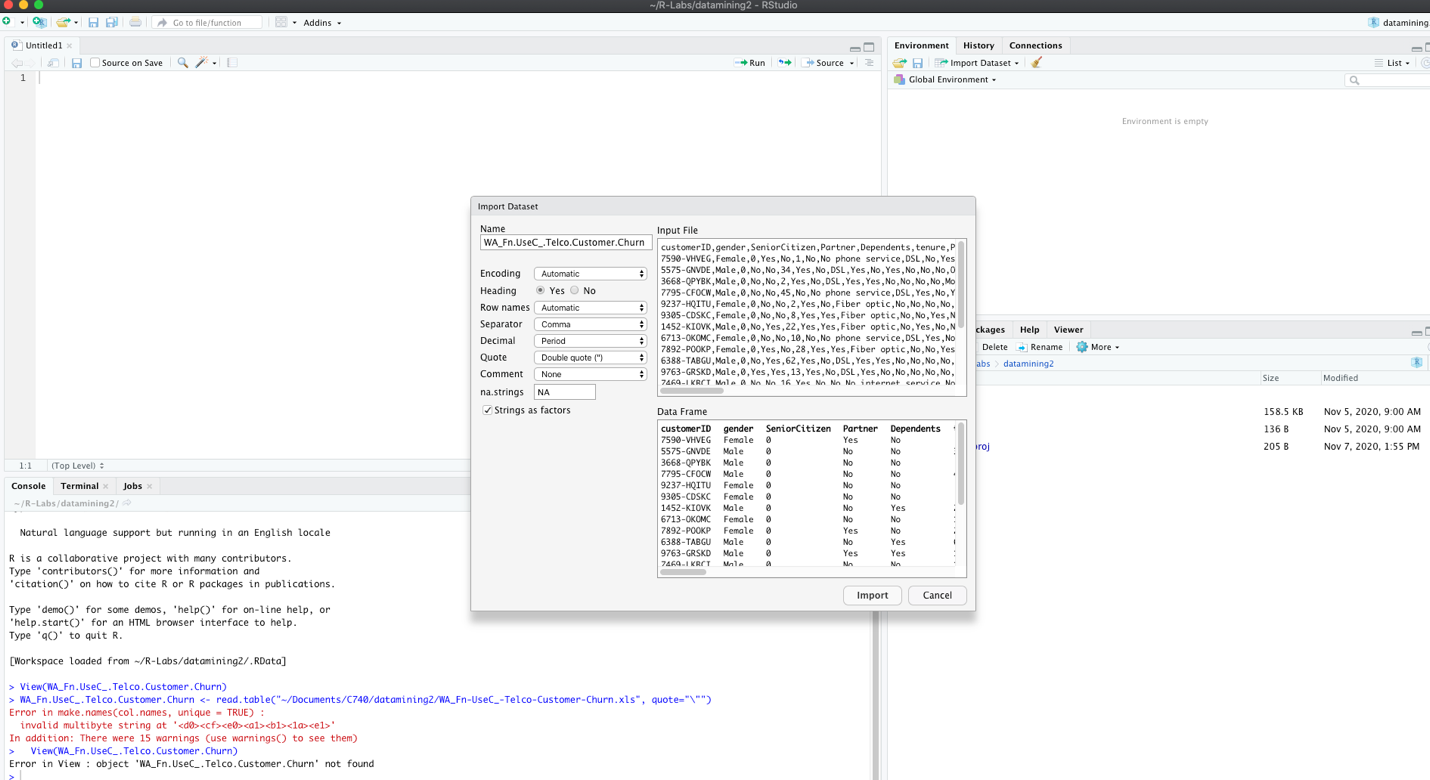
**I. Tool Selection:**

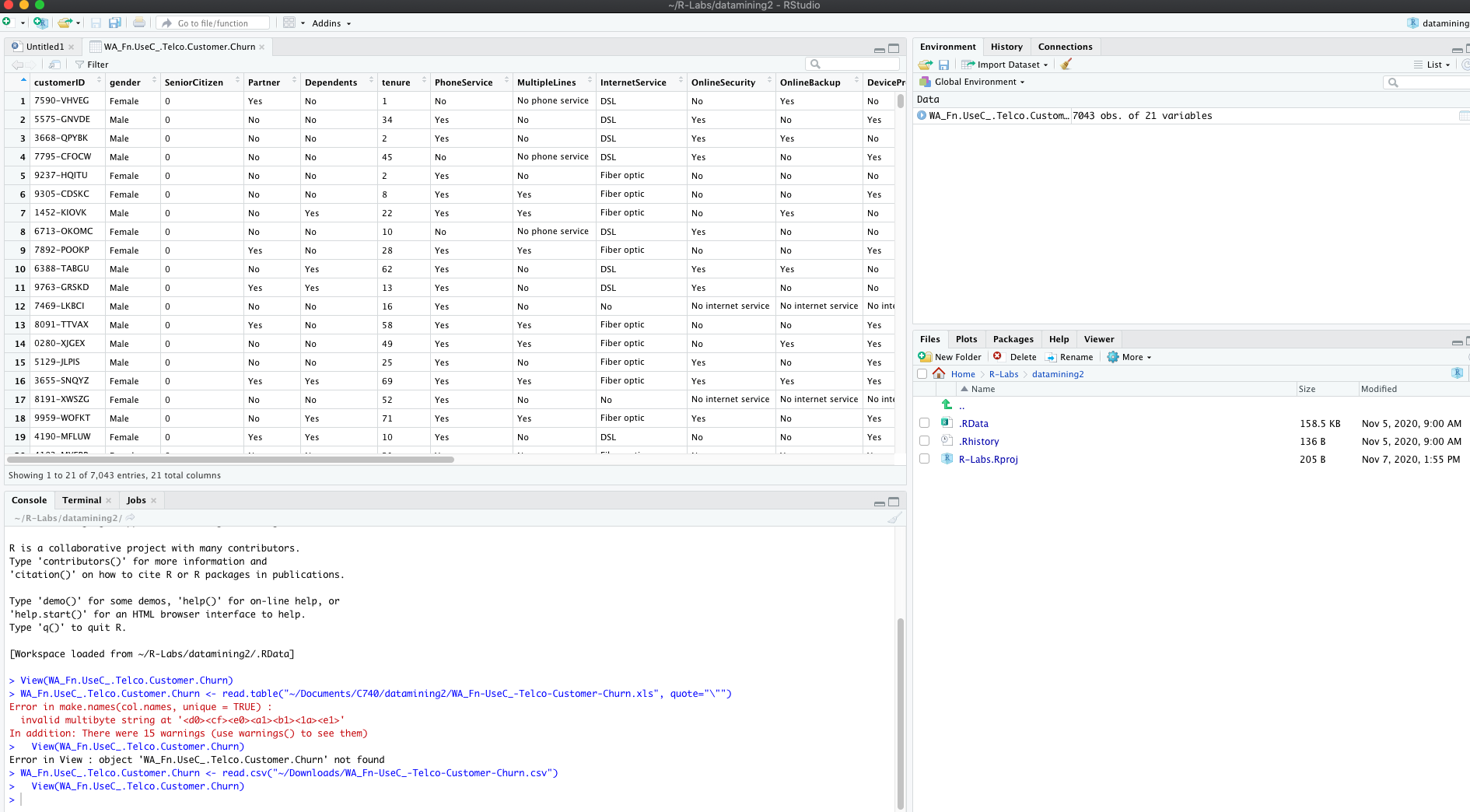
**Execute data extraction from the “Customer Data” web link using data mining software (Python, R, or SAS). Provide a screen shot of the code you have written and its successful application with a copy of all the extracted data.**

1. **Describe the benefits of using the tool you have chosen (Python, R, or**

**SAS) for extracting data in this scenario.**

I have decided on using R for this presentation because it provides a very easy to use graphical interface called R studio that allows for quick and painless importing of data. By selecting the import dataset drop down you can navigate to the csv file and import it into R studio within a few button clicks. IBM SAS allows for quick importing of data, but I find that the software is very clunky and difficult to navigate, and since R is open source I can rapidly import R packages that I need on the fly giving me much more control without having to install a lot of software that I don’t need.



****

1. **Define the objectives or goals of the data analysis. Ensure that your**

**objectives or goals are reasonable within the scope of the scenario and**

**are represented in the available data.**

The objectives of this analysis are to find which factors most influence whether a customer will “churn” (leave the company) or stay a costumer, and to create an algorithm that will accurately predict which customers are likely to churn, as to target them with potential promotional offers to keep the customer from leaving for a competitor.

1. **Select a descriptive method and a nondescriptive method (i.e.,**

**predictive, classification, or probabilistic techniques) you will use to**

**analyze the data, and explain how the methods you have selected are**

**appropriate for the objectives or goals you have defined.**

I will use Primary Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) to show how the different variables in the data set are related (Kassambra, 2017), and how much they contribute to the overall results of the data. This will help to visualize the underlying relationships between the variables, and is perfect for my goals of determining which customers are likely to churn, because it will give me a visual representation of the relationship between the variables. For a predictive technique I will use logistic regression to infer which individual is likely to churn, this will be appropriate for my goal of predicting which individuals will churn because the Churn variable is a binary variable and logistic regression classifies individuals into binary groups.

**II. Data Exploration and Preparation:**

**Clean the data you have extracted and save as .xls or .xlsx format for submission. Be sure to address all necessary formatting, converting, and missing data.**

**D. Describe the target variable in the data and indicate the specific type of data the target variable is using, including examples that support your claims.**

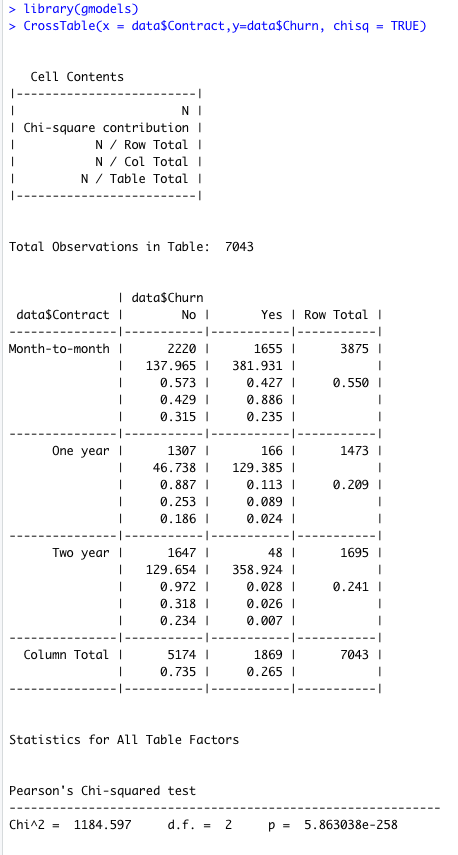
The target variable is Churn, which is a categorical binary variable that it is comprised of two levels: “No” and “Yes”. By running the summary(churn$Churn) function:



we are able to see that Churn has 5174 individuals who have not churned, and 1869 individuals who have already churned. This data can be used in a logistic regression algorithm to determine the change in the target variable churn based on the independent variables, and classify each individual in either a yes or no category.

**E. Describe an independent predictor variable in the data and indicate the specific type of data being described. Use examples from the data set that support your claims.**

The Contract variable is an independent predictor variable, as it changes there is a predictable change in the target variable Churn. Contract is a 3 factor categorical variable consisting of Month-to-month, Annual, and Two year. Through Multiple Correspondence Analysis (MCA) I will show how contract contributes to the over result of the data, and input it into a logistic regression algorithm to predict which individuals are likely to churn. Another way of determining if there is a relationship between the Contract variable and Churn is by calculating the chi square statistic which can be done with the CrossTable() function from the gmodels r package.



The low p-value indicates that there is a relationship with the churn variable, which will be investigated further through Multiple Correspondence Analysis.

1. **Propose the goal in manipulation of the data and define your data preparation aims.**

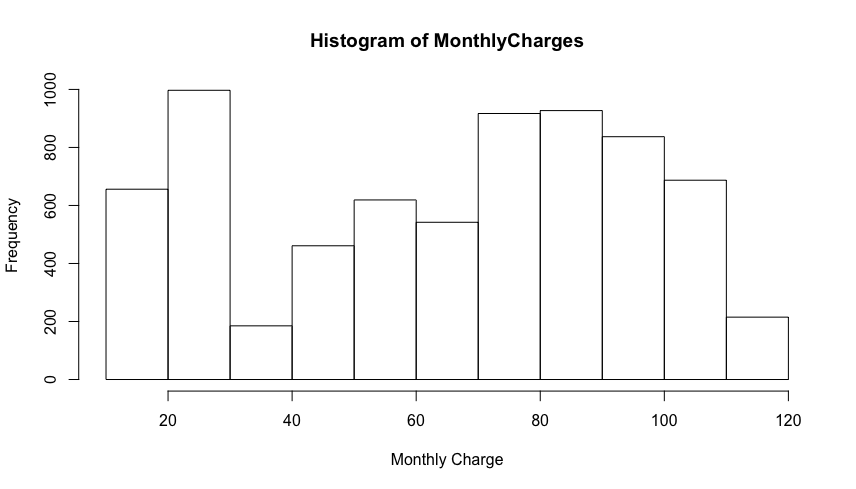
The goals of manipulation in this scenario would be to create uniform data that can be passed into Multiple Correspondence Analysis and Primary Component Analysis using the FAMD function from the DatamineR package, and to either convert all “NA” values to 0 or remove them, so that they can be used in the FAMD function, and to make a copy of the dataset called “data” which will be a working copy for all of the scripts, so the original dataset will remain unaltered. The aims for data preparation are to have reliable data in a format that can be used in the FAMD function and a logistic regression algorithm, which will also require converting variables like MonthlyCharges to make them easier to work with. For instance when MonthlyCharges are plotted in a histogram there are two peaks and some week candles, it might be more appropriate to combine the values into ranges which will make it perfect for Multiple Correspondence Analysis.

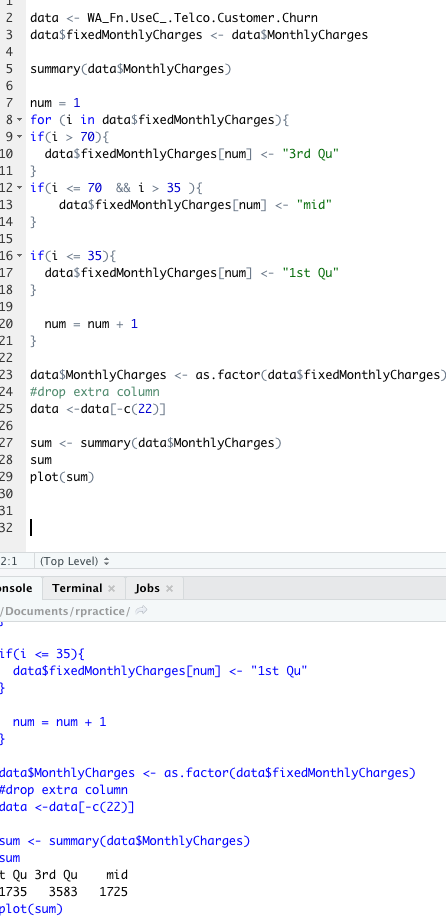
Example:

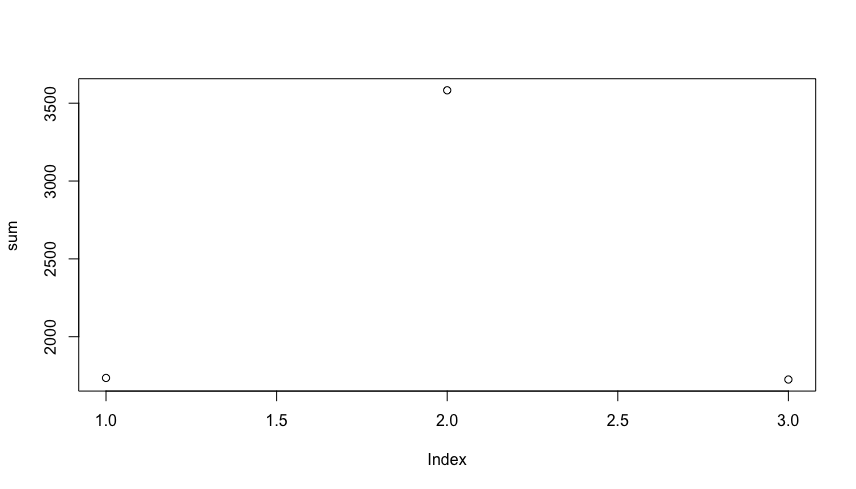
By running the summary() function on MonthlyData you get the following output:



To deal with the odd distribution I will add each individual to one of three categories representing the min -1st Qu, 1st Qu.-3rd Qu, and 3rd Qu – Max.



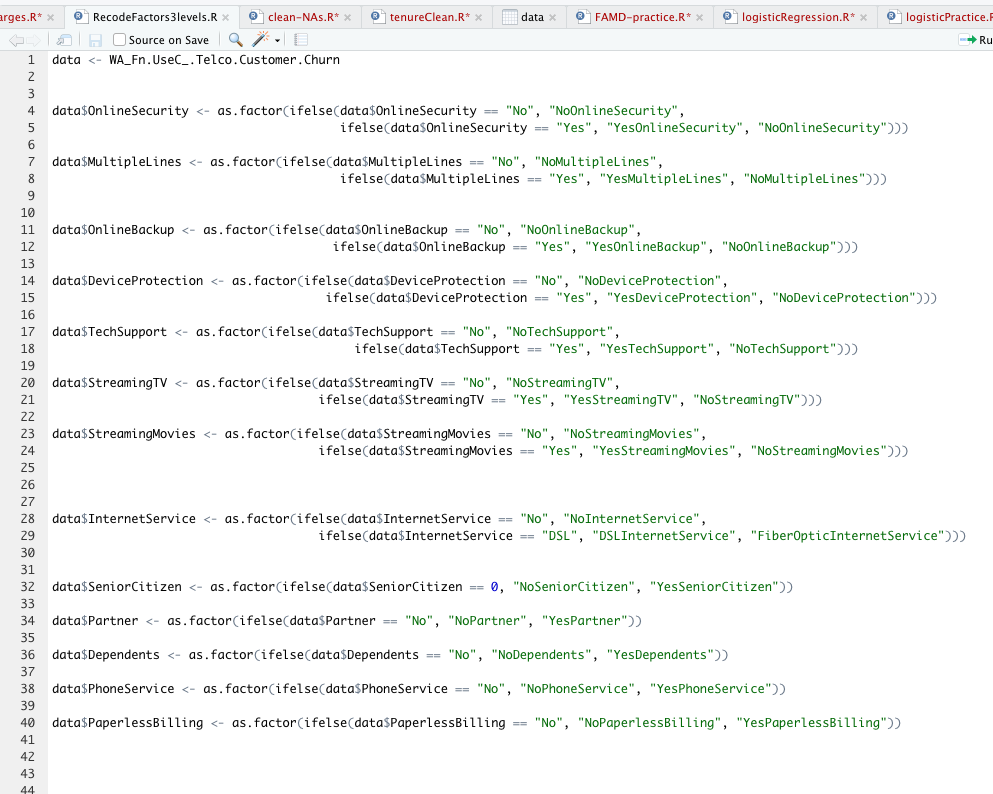




1. **Define the statistical identity of the data, including the essential**

**criteria and phenomenon to be predicted.**

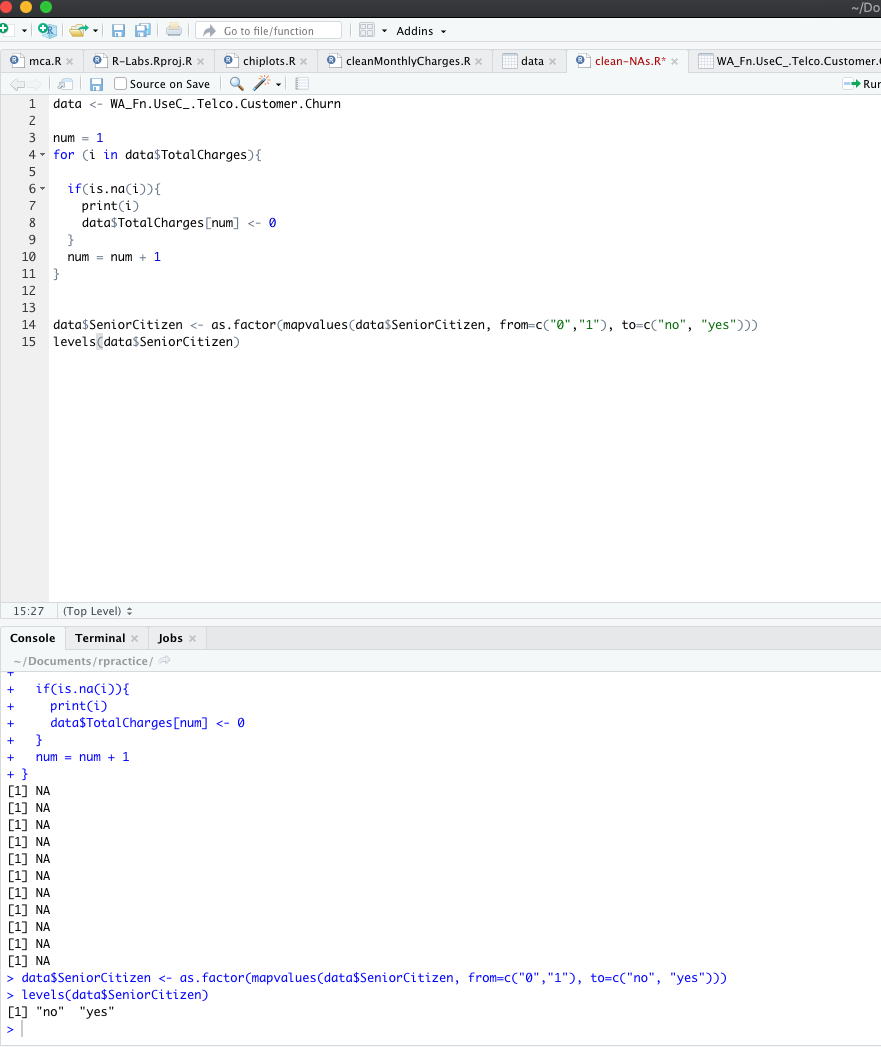
By running the summary() command the identity of the data is printed to the screen. When it is run on the data set you can see that there are 7 categorical variables with 2 factors to include: gender, Partner, Dependents, tenure, PhoneService, PaperlessBilling, and Churn. There are 9 categorical variables with 3 factors: MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and Contract. All of the 3 factor variables will be converted to 2 factor variables, because the third factor in each variable is described by another 2 factor variable, for example MultipleLines has a “no phone service” category which is already described by the phoneService variable. There is 1 four factor variable (PaymentMethod), and 1 binary categorical variable (Senior Citizen) which will be converted from “0” and “1” to “yes” and “no” for uniformity. There is a single factor customer ID variable which will be dropped because it provides no statistical significance. Finally, there are 3 ratio variables that describe tenure as an integer in months, and MonthlyCharges and TotalCharges in continuous monetary values. The phenomenon to be predicted will be which customers are likely to churn.



1. **Explain the steps used to clean the data and how you addressed**

**any anomalies or missing data.**

To clean the material it is a good idea to run the summary() function first and analyze the output of each variables to see what each one is comprised of and if there are missing values in the form of “N/A”. In the TotalCharges column there are 11 “N/A” values, and under further investigation each of the “N/A” entries are for individuals with less than 1 month of tenure indicating that the value should be 0.00 since they have not made a payment yet. The attached screen shot demonstrates the script used to correct the “N/A” values. After running the summary() function it is important to check for extreme values and outliers using a combination of boxplot() and hist() functions on the three ratio variables (tenure, MonthlyCharges, and TotalCharges). There were no values from those variables that needed correcting save the “N/A” values that have already been discussed, however due to the non-uniform distribution both an exponential and logarithmic transformation on MonthlyCharges and TotalCharges were performed, which did not result in a better distribution, but did bring to light the need to convert both variables to a 3 factor category based on the box plot of each variable. This is discussed in the next section with the graphic displays. For uniformity the binary categorical variable SeniorCitizen was changed from “0” and “1” to “yes” and “no”. The script for both removing “N/A” values and converting SeniorCitizen can be found in the following image.

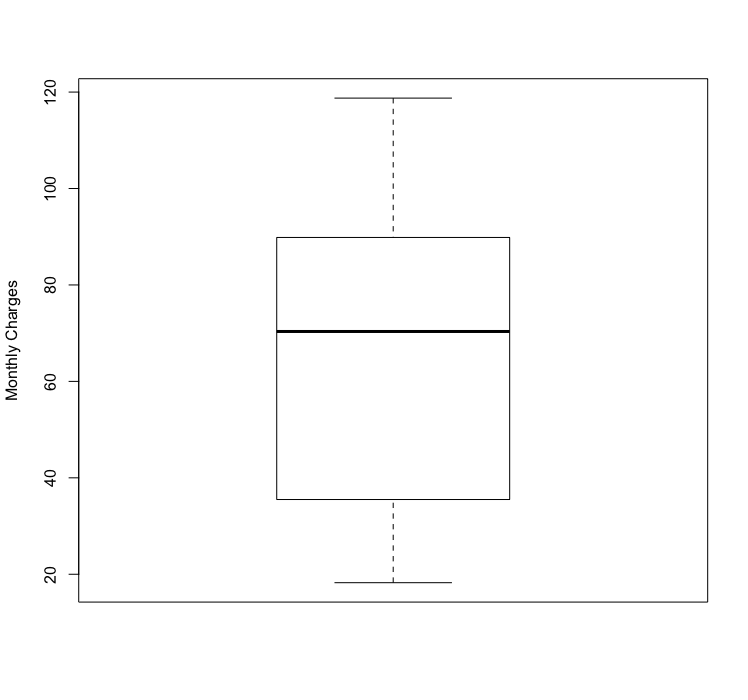
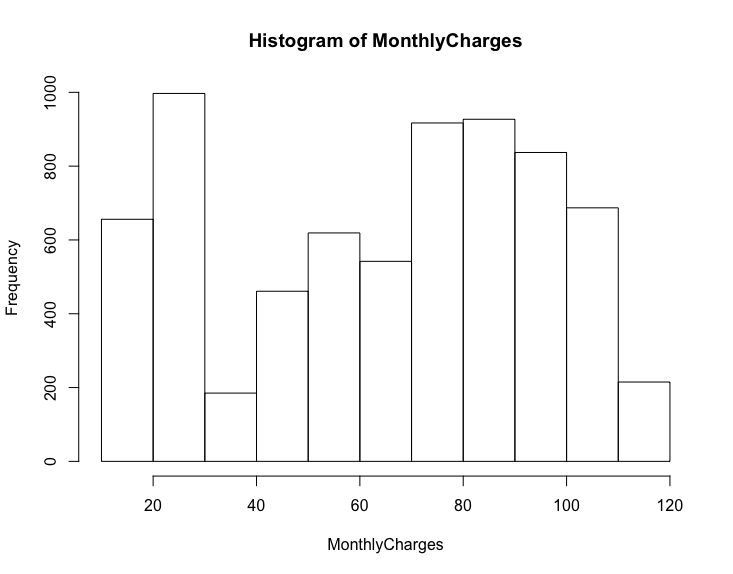


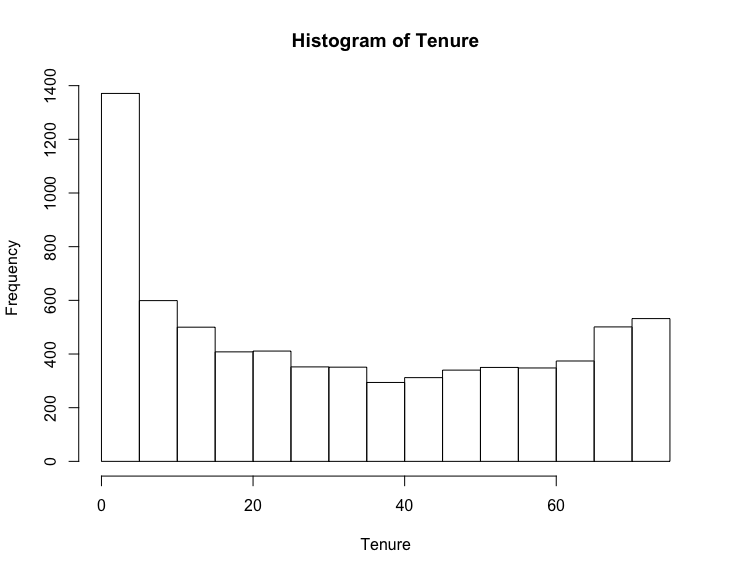
**III: Data Analysis**

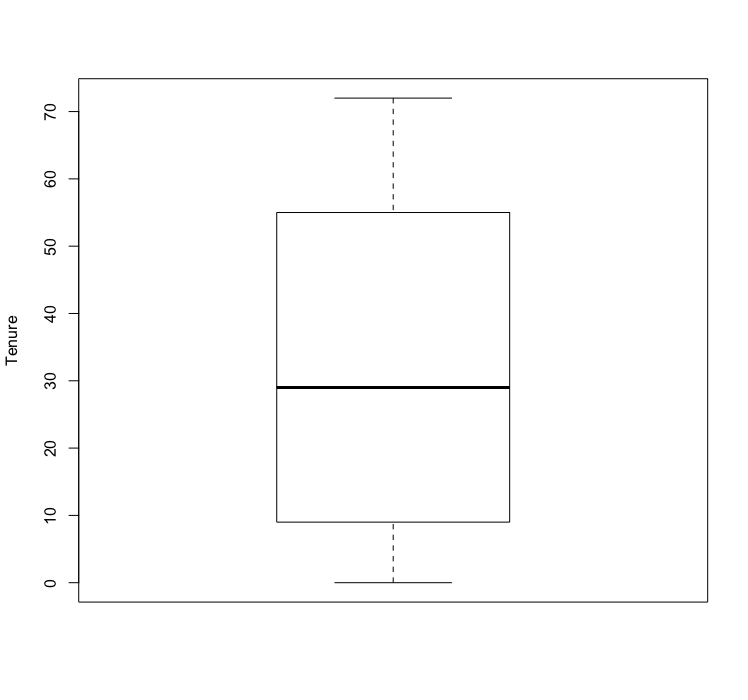
**For each of the following steps, be sure to clearly indicate each step within your data sheet with a screen shot and annotations in your final submission. All algorithms used need to be clearly identified in the screen shot and submission.**

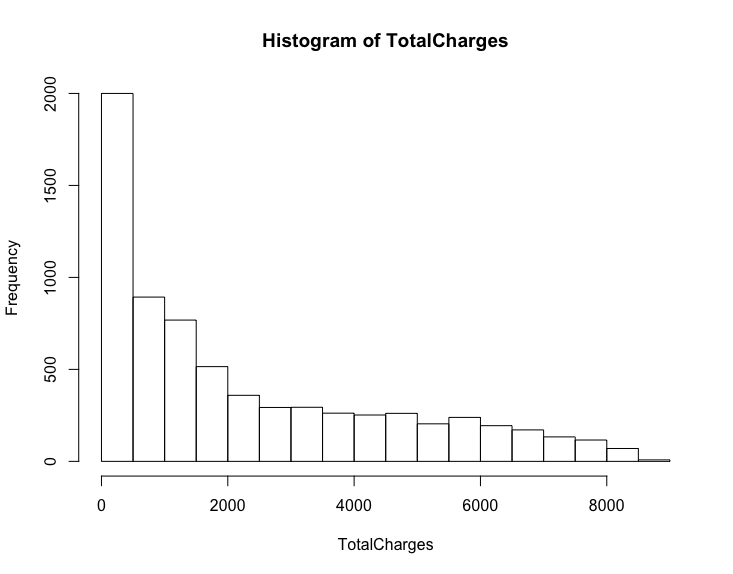
1. **Identify the distribution of variables using univariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.**

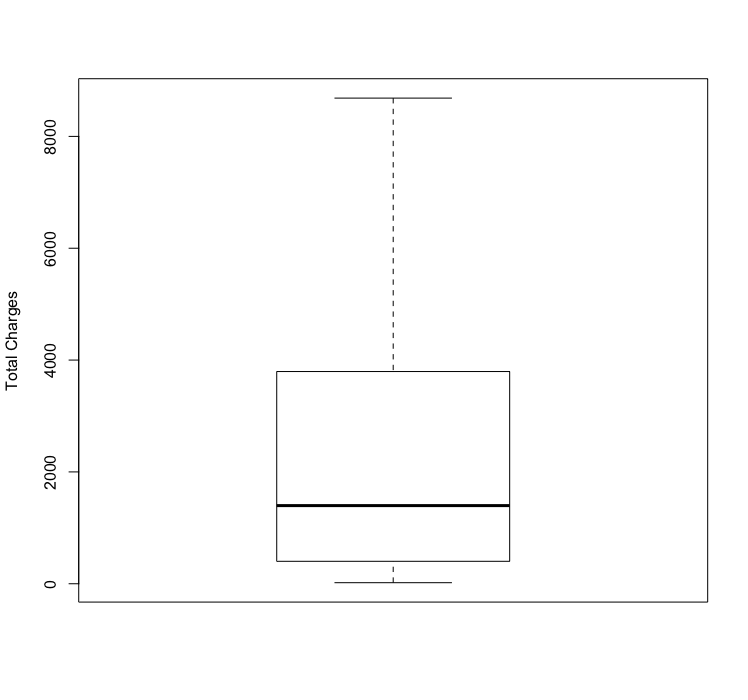
As suggested in chapter 3 of the reading material (Tuffery, 2011) both box plots and histograms where used to visualize the distribution of each of the continuous variables. By using the summary() and boxplot() functions on the data set you can see the distribution of each continuous variable, and by plotting the variables using hist(), the distribution can also be easily visualized. From the fore mentioned plots, we can see that the variables do not have “normal” or parametric distributions, and will need to be transformed for further analysis. For instance, the tenure values can be changed from a ratio integer to a categorical nominal variable that represent short, medium, and long tenures, and monthly charges can be converted to categorical nominal variables that represent each quartile the value falls into.





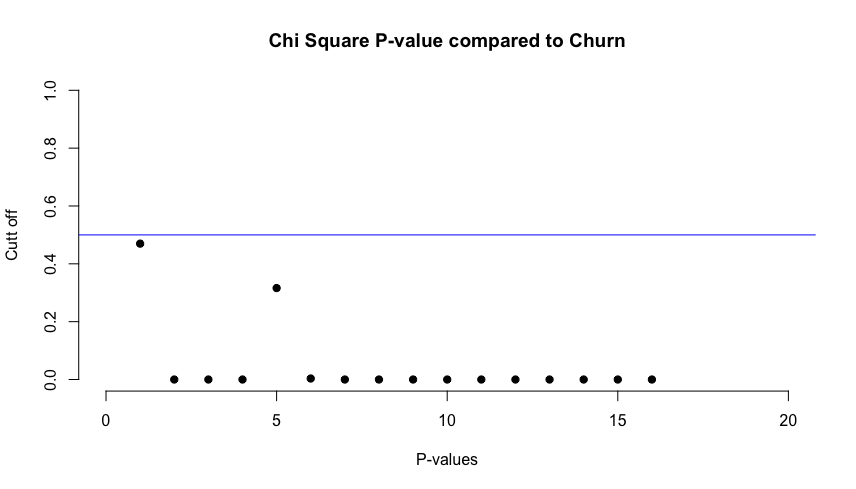


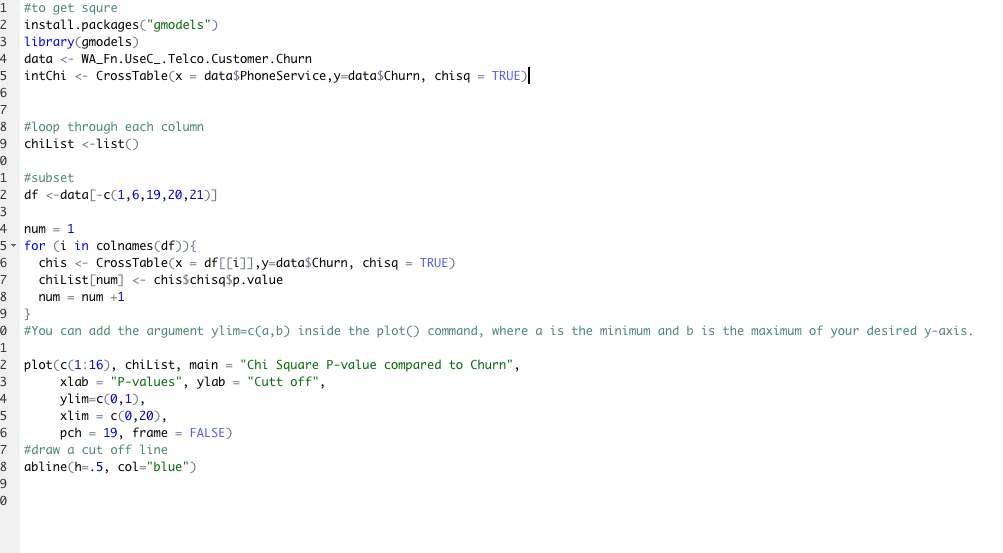




1. **Identify the distribution of variables using bivariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.**

Chapter 3 of the reading material (Tuffery, 2011**)** says to handle categorical data using bivariate statistics to detect links between each variable and the dependent variable. When the chi squared values for each variables in the data set compared with the dependent variable (churn) has been calculated an object containing the chi square value is returned, these values can be graphed to show each value. To keep the reader from having to look at several outputs, this graph was used in leu of an image and summary for each variable in the data set. As you can see by the blue cutoff line when each variable was compared to the churn variable the resulting p-values are very low aside from gender which is close enough to the standard .05 cutoff to be removed and will require further analysis. The other values indicate a correlation between each variable and the churn variable. I have also included an image of the code to re-create the graph.

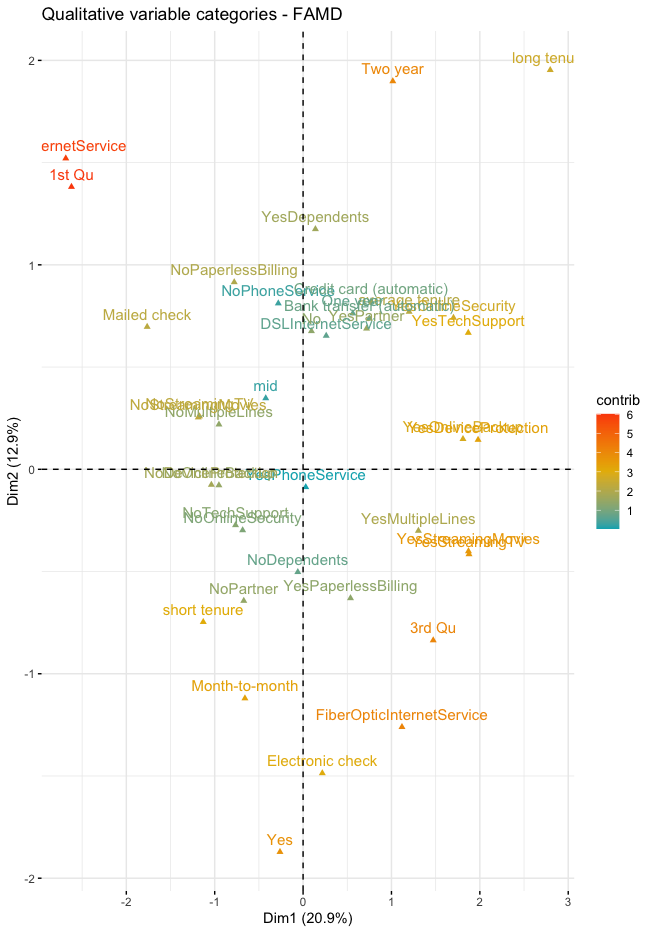




To further investigate the relationship between the variables it has been determined that a combination of Multiple Component Analysis (MCA) which is used for categorical variables and Primary Component Analysis (PCA) used for continuous variables via the FAMD function from the FactoMineR package will provide a visual insight into what the underlying relationships between the variables.

1. **Apply an analytic method and an evaluative method. Annotate the data showing both methods and your findings.**

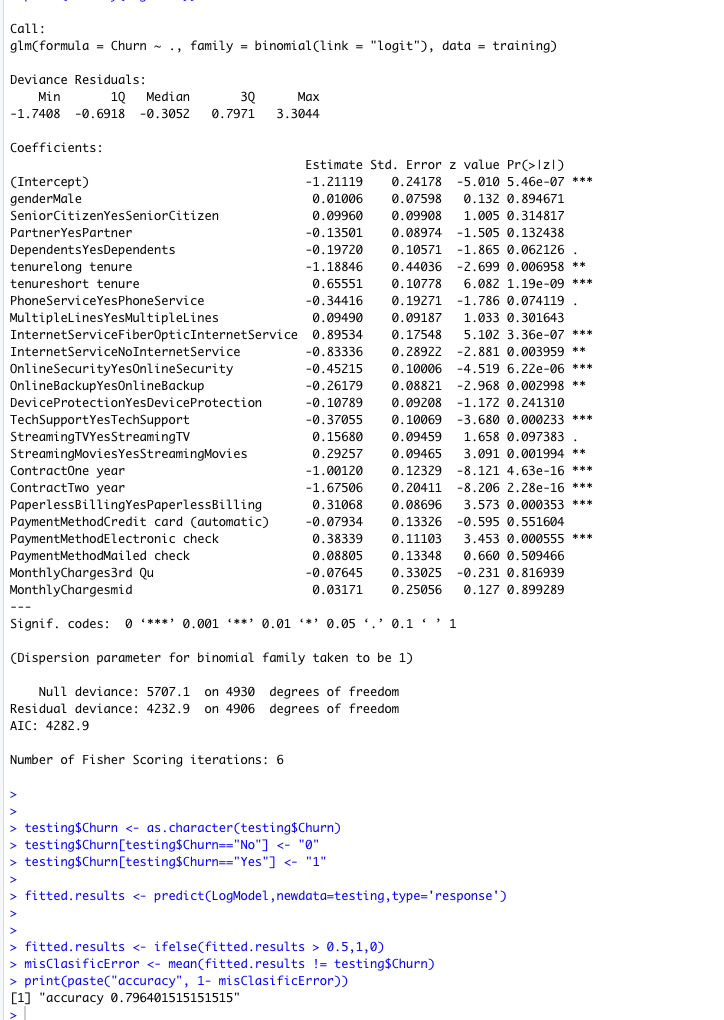
The analytic method will be a combination of MCA and PCA using the FAMD function from the FactoMineR package. After conducting the analysis it is clear that a short tenure is correlated with a month to month contract, individuals with no partner, and no tech support or internet security. The attached qualitative variable categories FAMD plot show how month to month contract contributes strongly to a short tenure, and the inverse a two year contract contributes to a long tenure.

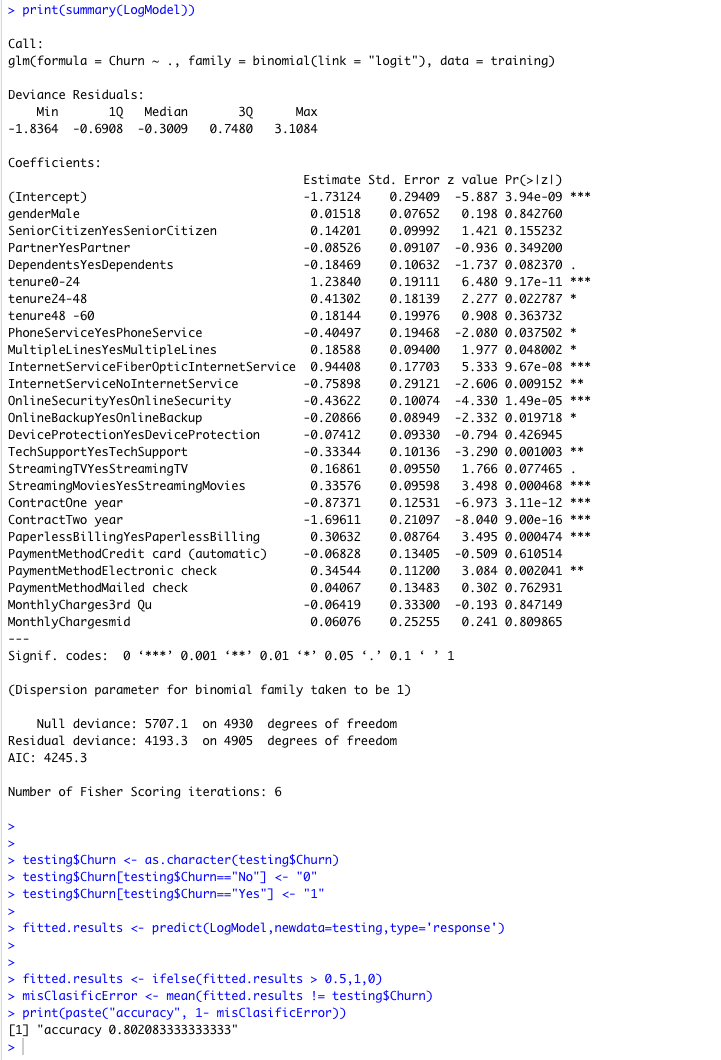


The evaluative method will build on this analysis and show that the primary indicator for a long tenure will be the contract type.

The evaluative method will be a logistic regression algorithm that classifies individuals as either “Yes” or “No” for churn, after creating a model and printing the summary we can look at how each variable contributes to the overall result of the data:

As you can see the accuracy of the model in the first image is ~79% and the tenure variable consists of short, long, and medium. The medium tenure is not contributing to the results, so to improve the model it was decided to change tenure to a categorical variable representing 0-12, 12-24, and 24-48 months, which resulted in the second image which has a slightly better accuracy of ~.80% and would be considered effective for the task at hand of predicting which individuals are likely to churn. There is also a print out of how each variable contributes to the overall result, and by looking at the Pr(>|z|) column we can see that gender does not actually have a relationship with churn as described above in the chi square plot because the p value would be considered over the acceptable value of .5. The other pieces of significance are the Estimate values of Contract and tenure. A two year contract is the largest predictor for determining if a customer will churn or not, and a tenure of 0-24 months shows that customers who are in that window might be influenced with a promotional offer, decreasing how many customers churn.





1. **Justify the methods you have chosen to analyze your data. Be**

**sure to include details about how the methods you have chosen**

**better represents your findings than other methods.**

The combination of MCA and PCA work well with the data because it is of mixed type, and are great for data sets with a lot of variables because they reduce the clutter of a lot of variables resulting in a description of how each variable is related, and chi square is great for determining the relationship between categorical variables and the target variable churn. Cramers V does a better job of correlating variables (Tuffery, 2011) and results in more accurate correlations where chi square can inflate the correlation, but for the purposes of finding a relationship it works very well, and because we are not using the chi square to determine inputs to our logistic regression algorithm it serves its purpose of alerting us to relationships that can be further investigated. Logistic regression is almost perfect for representing the findings because it classifies individuals into binary buckets which represents the churn data, but it only results in an ~80% accuracy, there may be a more accurate methods such as a random forest or decision tree algorithm that might produce more accurate results.

1. **Justify the methods you have chosen to visually present your**

**data. Be sure to include details about how the presentation**

**methods you chose better represents your findings than other**

**presentation methods.**

The decision to visually present the data with a heat plot of categorical variables and a plot of the chi square values for categorical data did a good job of demonstrating the data and relationship between the categorical variables, by using a heat map for the FAMD results it is easier to distinguish how much the variables contribute to the results of the dataset, this is much better than a plain graph depicting each variable in the same color and provides the customer with a clear visual of the underlying relationships. The chi square plot saved a lot of space in the presentation because it allowed for all the values to be seen in one place, and annotates where the “cutoff” is which reduces confusion and gets straight to the point without losing the information amongst several tables for each variable. For the quantitative variables there are multiple plots for each one showing the distribution and quartile values for each variable, these plots are great for the data because it gives full detail to the reader and does not sacrifice readability by attempting to group them into one image for each plot.

**IV: Data Summary**

**Summarize the findings of your data evaluation. Provide the final findings dataset, including evaluation measures.**

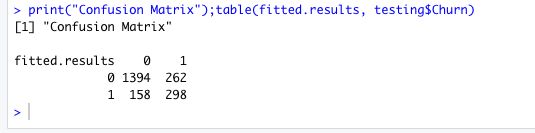
1. **Explain how your data shows that it was discriminating or not**

**and whether the phenomenon you wanted to detect was present in your findings. Provide specific examples from the data to support your claims. (read chapter 3.8)**

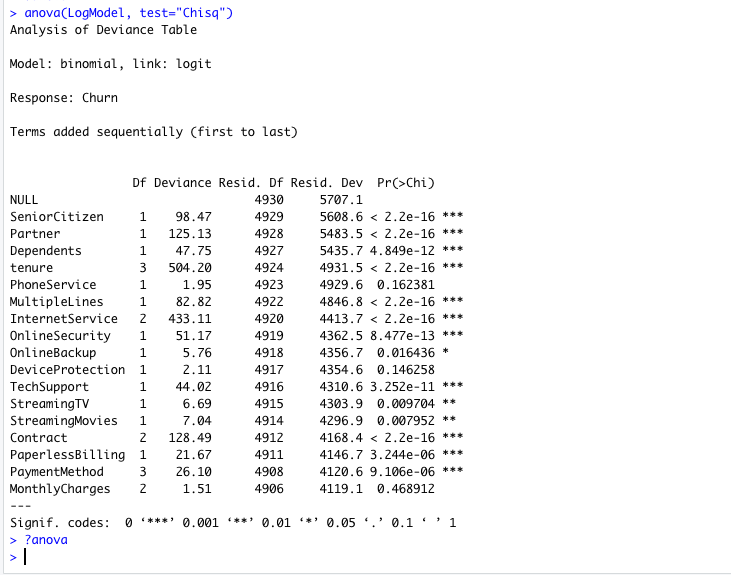
Per the results of the logistic regression algorithm, I would say that the data was in fact discriminating and did detect the phenomenon of which customers are likely to churn, and which customers can be targeted with a promotional offer. By splitting the data into training and testing with an 80, 20 split we are able to train the algorithm on 20% of the data and use 80% of the data for determining how accurate it is.



We can also determine how many of the actual individuals the algorithm predicted would churn vs how many actually churned by running:



This shows that the algorithm predicted that 1394 individuals would not churn while the actual number is 1656, and that 158 individuals would churn but the actual number was 298, which gives us an accuracy of about 80%. For determining which individuals are likely to churn we need look no further than the anova of the model:



This produces an output that computes Analysis of variance for a given model object, and shows that InternetService, tenure, and Contract are driving how the model predicts each churn value (Li, 2017). This combined with the information from PCA and the summary print out of the model indicates that by targeting individuals with less than 24 months of tenure, with internet service, and a contract less than 2 years for promotional offers, could potentially reduce the churn rate.

1. **Describe the methods you used for detecting interactions and**

**for selecting the most important predictor variables. Include the specific interactions you detected and the most important predictor variables that you found.**

By running the FAMD function from the factominer package a combination of PCA and MCA is done on the data for qualitative and quantitative variables respectively. After conducting this analysis, it is apparent that the significant variables are contract length, tenure, and Internet service, which can be verified via the summary of the model object which shows that Contract, tenure, and InternetService are the most important predictor variables.

1. **Acknowledge sources, using in-text citations and references, for**

**content that is quoted, paraphrased, or summarized.**

The citations for this paper were generated using the APA citation wizard from the writing center.

References

Kassambara, A. (2017, September 24). FAMD = Factor Analysis of Mixed Data in R: Essentials.

Retrieved from http://www.sthda.com/english/articles/31-principal-component-methods-

in-r-practical-guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/

Li, S. (2017, November 20). Predict Customer Churn – Logistic Regression, Decision Tree and

Random Forest. Retrieved from https://www.r-bloggers.com/2017/11/predict-customer-

churn-logistic-regression-decision-tree-and-random-forest/

Tuffery, S. (2011). Chapter 3 Data Exploration and Preparation. Data mining and statistics for

decision making. (pp. 44-59). West Sussex, United Kingdom: John Wiley & Sons, Ltd.