C744: DATA MINING AND ANALYTICS II

Tyler Cabutto

Western Governors University

December 15th, 2019

**Tool Selection (Section A, B, and C)**

This section shall delve into the developing of the foundation of what tools we will use, defining our goals for the analysis of our telecommunications company data, and what method(s) we will use in order to accomplish these goals set forth.

**What language will be our weapon for success? (A)**

With the given scenario, the task proposes what would appear to be a simple question you can ask yourself. Which out of the three languages of Python, R, and SAS would be beneficial to use? If you know anything about these languages, you know they all have their special pros and cons that come with them. In this scenario, we are trying to identify why the telecommunication company is losing so many customers to the cable competitors.

Based off personal choice, I feel that the best option in this case will be R. The reason being is due to the pros outweighing the cons in this scenario in comparison to the other languages, excluding SAS. The main decision why to exclude the SAS language is due to the price being too much, therefore ruling it out. However, R is free, provides a ginormous library of packages for this request to be completed, and outputs our data in an extremely convenient manner in comparison to Python for example. Even though Python for the most part can be up to par with R, there just isn’t as many options in Python to perform the given task in my opinion.

**What goals are we to obtain from this analysis? (B)**

As for an approach for making goals for this project, it was decided to use the structure from IBM called the CRISP-DM. It will help with the understanding of the life cycle of the project and the best steps to take in the data mining methodology. With that said, first and foremost, our goal should be to take our data and get an understanding of what it is that we are trying to accomplish. In this case, we need to determine why the churn rate is so high and what can the company do to fix it? Next, we will need to take our data and put it in a suitable format that will help us not get stuck halfway through our evaluation. It is important to take our dataset and clean it to ensure that it can be easily imported and exported in the chosen formats. Once this is done, the main concept at this point is to serve our client to the best of our ability by identifying the issue they are wanting us to assist in fixing.

In this scenario, we are tasked with trying to determine why it is that their clientele’ are cancelling their services with them and going to cable competitors. How can we reach this goal? As noted in the book “Handbook of Statistical Analysis and Data Mining Applications”, we should develop and deploy a model that will generate a significant business value, and obtaining knowledge base of modeling, that we can use a leverage later to be more successful with our data mining task (Handbook of Statistical Analysis and Data Mining Applications, 2019). Nevertheless, we need to define some aims in order to reach this goal as we cannot just jump to a conclusion without the focus of some other aspects. In this case, we need to determine what is it that is causing these customers to leave? To take this even further, we can also try to predict what the outcome would be if a certain variable is not changed for the better once we have identified it as a possible reason as to why they are leaving. We can then propose a solution as to what the company can do to reverse this affect and retain their customers for as long as possible.

**What methods to use for analyzing this data? (C)**

For this, it will be best as recommended to use a descriptive method, and a predictive method. The best descriptive method that makes sense for this would be clustering our variables. The reason is whenever you consider using clustering, is whenever you have multiple different variables that can make things look as if they are super disorganized. Using this method will help us determine similarities between the independent variables and can help us group these people and determine if that group of people are churning or not. This will ensure the ability for us to see patterns in our data that we may not be able to depict alone by just examining it with the naked eye.

Next, we will choose a non-descriptive method to analyze the data. In this scenario, the best one that we are going to attempt to use will be a logistic regression. The reason this will be used is the fact that our target variable is binary (Yes or No), and against all the other dependent variables. This can be helpful as it will help us confirm exactly what is best to predict churn, and how accurately it lines up with what is actually occurring with our data.

**Data Exploration and Preparation (D, E, F, G, H)**

**What is our target variable? (D)**

Upon observation, it is obvious that are target variable in this scenario is going to be from the ‘churn’ feature in our dataset. This feature will pass two values, either ‘yes’ or ‘no’. The yes value is indicating that the customer did in fact churn and is no longer using the services. The no means that the customer is still active and is using the services and have not left the company. The data type in this scenario can be defined as **nominal qualitative** due to its values being of text, and not numerical as you can see in the picture located in appendix A.

**What is an independent predictor variable in our data? (E)**

An example of one of the independent predictor variables in this case could be the ‘Monthly Charges’ feature in our dataset. We can use this variable to possibly predict the effect of the target variable, for example, we could speculate that the churn rate is higher whenever the amount of monthly charges is higher. The data type that is used in this feature is a **continuous quantitative** type. The reason for this is because as you can see in appendix B, the value passed is a numerical type, being how much the customer is spending on a monthly basis, which can be infinite set of real numbers.

**The goal for data manipulation and data preparation aims (F)**

The goal for data manipulation in this case will be to ensure our data is clean to the best of our abilities without the use of importing it into R. This will include possible imputing data if necessary or removing data that may be missing if it is not extensive. We will also investigate the possibility of removing some fields if they seem to be unnecessary to our objective. Once this is done, we will load our data into R and see if there is anything that was overlooked that could be a little more difficult for us to tell by just observing the data, such as outliers hidden deep within our dataset, the need to develop new fields to help us reach our goal, or having an imbalance in our data, making it more difficult to perform a prediction on why a customer stays, or leaves. This can be assumed then, that our main data preparation goal is to ensure that we have the cleanest dataset possible before moving forward and building our models.

**Defining our data and cleaning it (G, H)**

The first thing that we can do is to look at the data and identify what each of the fields are. This can give us insight if we can clearly identify what each field is, or if we may need to do a little research in order to determine what it is our business is trying to say. First and foremost, I chose to identify any unnecessary features in our dataset that we may not need. Truly the only one that was necessary to get rid of was the ‘customer ID’ column. This value is unique for each observation given to us. This leaves us with 20 total columns in our dataset that we can observe. The target variable as mentioned earlier is the ‘churn’ field and is a nominal qualitative variable. The rest of the 19 fields in our dataset can be considered independent variables, as we are going to observe how they all could be affecting our target variable, churn.

As far as continuous values, there are a two, monthly charges, and total charge. The reason being is because they are infinite subset of the set of real numbers as explained in the course text in chapter 3. As for discrete data, there are two fields, the senior citizen, and tenure field. Senior citizen can only have a ‘0’ or ‘1’ value, basically defining if they are a senior citizen in a Boolean manner, where 1 means they are, and 0 meaning they are not a senior citizen.

As far as qualitative values, there are a few. There are 15 (excluding the target variable) nominal data types, being gender, 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. The reason is because none of these values require any type of order to them, otherwise they would be ordinal data types.

As you may already can tell, the phenomenon that is trying to be predicted, is which one of these independent fields listed above, could be causing this telecommunications churn rate to be so high? Therefore, the two methods above were chosen, variable clustering, and logistic regression, because we have different types of data that we are observing, and a good handful of them could be affecting the churn rate of this company.

Next, it was found that there were some values that would prove to be irrelevant to our cause of predicting why churn rate is high and could interfere with our prediction. It was noticed that the tenure field had some ‘0’ values in our dataset, which also caused the total charges field to be blank. It is presumed the reasoning for this is that the telecommunication company had some customers in the dataset that were going to subscribe to their services, but never did. Therefore, it was determined to remove them as they were never subscribed to their services in the first place. Next, instead of doing things manually from my observations, it was decided to go ahead and import the data into R to begin using some of the useful tools to our advantage to ‘pre-process’ our data even more so. The first task I wanted to do was to take the data and recode some of the values to allow for better grouping or readability. For example, I decided to change the senior citizen values to be ‘yes’ or ‘no’ values instead of ‘0’ or ‘1’ just to make things more consistent with the other values. You can see how this was done in Appendix C. Next, the same method was used to change Multiple Lines value “No phone service, to simply say “No”, grouping it with the other values. See Appendix D for the code used to accomplish this. The same will be accomplished on the other variable’s online security, backup, device protection, tech support, streaming tv, and streaming movies.

So far, we have gotten a little over half way of prepping our data to be analyzed, and you can look from the output of our data from what has been seen in the appendixes listed below, we have gotten a majority of our variables more simplified of which will help us group them better. However, we are left with three more variables that could use some work, total and monthly charges, and tenure. The question was brought up, how can I simplify these to allow for better results? It was decided after doing research on how to do summary and descriptive statistics in R, that using the function ‘’summary’ on the tenure and monthly variables to categorize them into three groups respectively; low, mid, and high. The name of the variable was also appended to his re-naming to help identify them more easily. This was determined based off the stats generated back to us and was also used in relation with the car package to ‘recode’ our data into our newfound format. You can see how this was performed in Appendix F and that the fields have been appropriately updated. We are now ready for the next step.

**Data Analysis (I, J, K, L, M)**

**Univariate statistics of our data (I)**

Now that are data has been prepared and cleaned, we can take this opportunity to get a visualization of each individual field to help us obtain a deeper understanding of what each value looks like from within, and will help make sense as to why some of our variables have been transformed into what they are now, versus what was given. First, we will do a univariate statistical analysis of each variable so we can se truly what is going on with each value within. All the charts that show the imbalance in some of the variables such as senior citizens or monthly charges can be found from within Appendix G with its associated variable name at the top of the chart.

The first thing I did was utilize ‘summary’ again to get a view of all the variables and their counts for each of them since they are all qualitative data at this point. This was recommended way of handling this data according to this courses text in chapter 3. To take it a little further, I decided to plot each induvial variable to get a better visualization as to what each one is doing. As you can see, there are a few variables that are evenly distributed amongst each other, which will help us. The churn is uneven with 5163 no values and 1869 yes values. You can find the summary table listed along with the charts under Appendix G.

**Bivariate analysis of our data (J)**

Now we will perform some bivariate analysis of our data to get a better correlation with our independent variables and our dependent variable, search for incompatibilities, and links between the independent variables. In order to accomplish this, it was decided to perform a contingency table on our data. We can look at Appendix I for the table for 5 of the variables. This was used from the ‘gmodels’ library found from ‘Categorical Data Analysis in R’ video listed in the references (2018). You can see the number of items in each category, the Chi-square contribution, and percentage of values from the row, column, and table totals. The chi square contribution is high for the month to month and two-year contracts in comparison to one year, which means that the frequencies obtained are different than what was normally expected. We can also use this to take senior citizens for example and see that the number of younger people that are not seniors’ citizens seem to be churning more than that of the senior citizens, about 75% of them to be exact.

**Evaluative and analytic method (K)**

The first step in this instance will be to get our data more so in a better group and help us better identify the independent variable(s) that is having the biggest impact on our target variable, and even more specifically why people are leaving. The first step in this case it to perform our analytic method, variable clustering. After some research, a package in R was found that seem to be very useful called ‘ClustOfVar which allows us to easily perform a cluster analysis and have the visual tools necessary to review our results. Prior to doing so, it was decided to train our data as you can see in the results of Appendix I. After adding the necessary library and training our data, we were able to produce the dendrogram seen in Appendix J. From here, we can observe and that it is suggesting about 8 groups of variables are closest in association with each other.

Now that this has been performed, we can use the stability function available in this library in order to cut our clusters into 8 groups for observation. In this case we will set the bootstrap to 60 as seen in the example on the documentation for this package on rdrr.io. If you look at Appendix K, you will see that it is confirming that we should stay around 8 clusters for our observations. We can now move on to splitting them into 8 groups. If you look at the chart in Appendix L, you will see how each variable can be grouped. This can be used for example, to see the correlation between internet service, monthly charges, payment method, and paperless billing. This can be used if the churn rate is high in these individuals, to better target their customers and determine if they should change how they offer services, or charge services to these individuals.

Next, we will use our predictive method in order to look at our data. In this scenario, it was decided to use logistic regression. First and foremost, we will take our data and use the ‘glm()’ function to build our model. As you can see in Appendix M, you will notice that we are using the ‘binomial’ family which defines it as a logistic regression. Once that is done, we obtain the summary of the model. You can see that our deviance residuals are close to zero which is a good sign. We can focus on the coefficients table and on our p-values to determine if we have good variables to use for the prediction of churn rate. We can see right off the bat in Appendix M that male gender is not a very good predictor as it is well above 0.05 which does not make it statistically significant. We can also clearly see that we have great predictor variables based off the “\*\*\*” symbol being passed next to some variables such as online security, fiber optic, and total charges being low to name a few.

It was then decided to generate a graph showing the probability of people in our data to churn. This is to show that our logistic regression has done a good job in reference to assigning these individuals based off the variables that have been given, in correlation with people who churned. It seems to line up accurately.

**Justification to our methods chosen (M)**

The reason that it was chosen to do hierarchical clustering, was because it will help us, and our client, the telecommunications company, have a better visual as to what is going on, and to help group items a little better with each other. This can be used to help the company with possible advertising campaigns for those group of people to possibly get them to switch to let’s say a 2-year contract versus monthly contracts to help them stay for longer. Logistic regression is used because at this point, it is well known and accepted throughout this field. It is also easy to interpret in comparison to most models, if not better. This will also help the company fully understand what is going on with the visual of what is a good predictor for why people are leaving, and to also back up the facts with the chart logistic regression chart generated to back up these claims. It was decided to take one of the highest coefficients

**Data summarization (N, O)**

**Discriminating or Non-Discriminating? That is the question (N, O)**

After review of our models, it clearly helped us identify what is causing some of this companies’ clients to churn. We can see from our cluster the groups of senior citizens, paperless billing, monthly charges, payment method, and internet services seem to be together more so than the other categories. This just so happened to be some of the higher amounts of people who were churning and was backed up as being a good predictor in our logistic regression, especially for paperless billing, and payment method. We were able to see as well with our bivariate analysis; we used the chi square table to help determine the exact number of people who were churning or not churning in this category to the exact number. This was also able to help us determine interactions between the different variables that have been given to us. It is clear that the best move for the telecommunications company would be to use the data given to determine what type of audience they need to broadcast to (Senior citizens for example), and to also consider who should they exactly aim for whenever it comes to obtaining new prospects to get their services.

References (P)

Data Mining and Statistics for Decision Making8.4 Data privacy. (2011) Retrieved from

<https://wiley.engagelms.com/learn/mod/lti/view.php?id=22778>

Sum Stats Analytics Example. (2009). Retrieved from

<https://www.r-bloggers.com/wp-content/uploads/2009/11/sumStatsAnalysisExample1.txt>

Dov, Tamara, & Duong, J. (n.d.). Recode Data in R with R Recode Examples. Retrieved from

<http://rprogramming.net/recode-data-in-r/>

James Dayhuff. R: Exploratory Data Analysis (EDA), Univariate analysis. (2017) Retrieved from

<https://www.youtube.com/watch?v=ONrGJF_8onw&t=222s>

Categorical Data Analysis in R (2018) Retrieved from

<https://www.youtube.com/watch?v=j9YFazcAjB4>

ClustOfVar: Clustering of Variables (2019) Retrieved from

<https://rdrr.io/cran/ClustOfVar/>

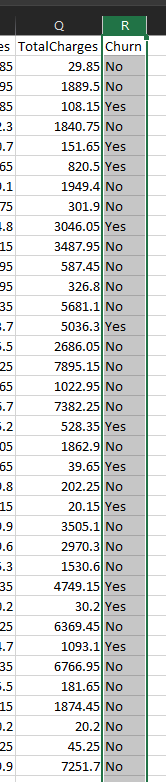
Josh Starmer. Logistic Regression in R, Clearly Explained!!!! (2018) Retrieved from

<https://www.youtube.com/watch?v=C4N3_XJJ-jU&t=542s>

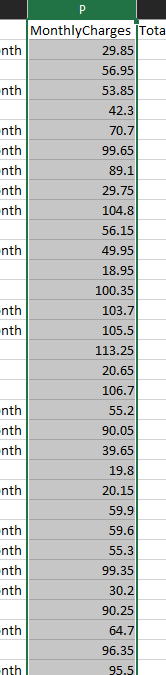
Aimee Backiel. Analytics Application: Churn Prediction (2013) Retrieved from

<https://www.youtube.com/watch?v=6yCxbzDjBDc>

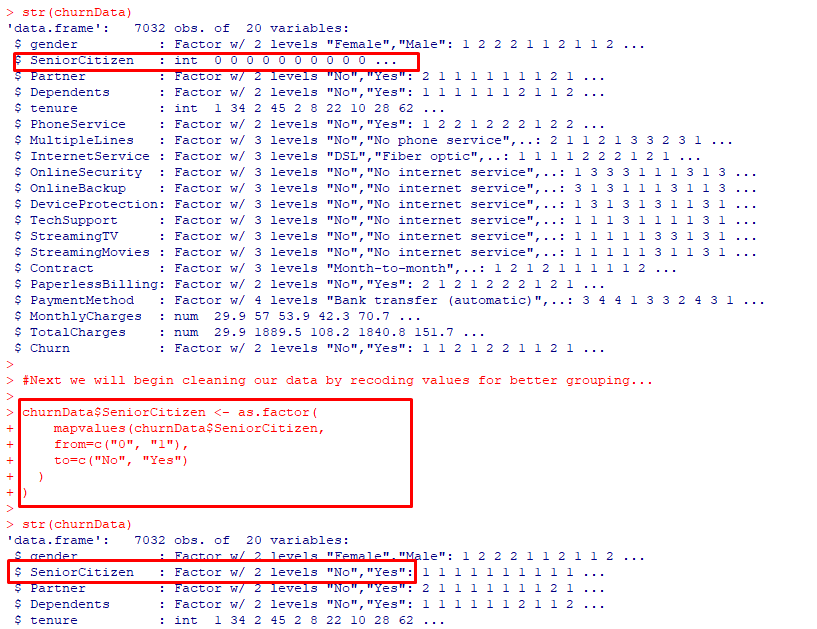
Appendix A



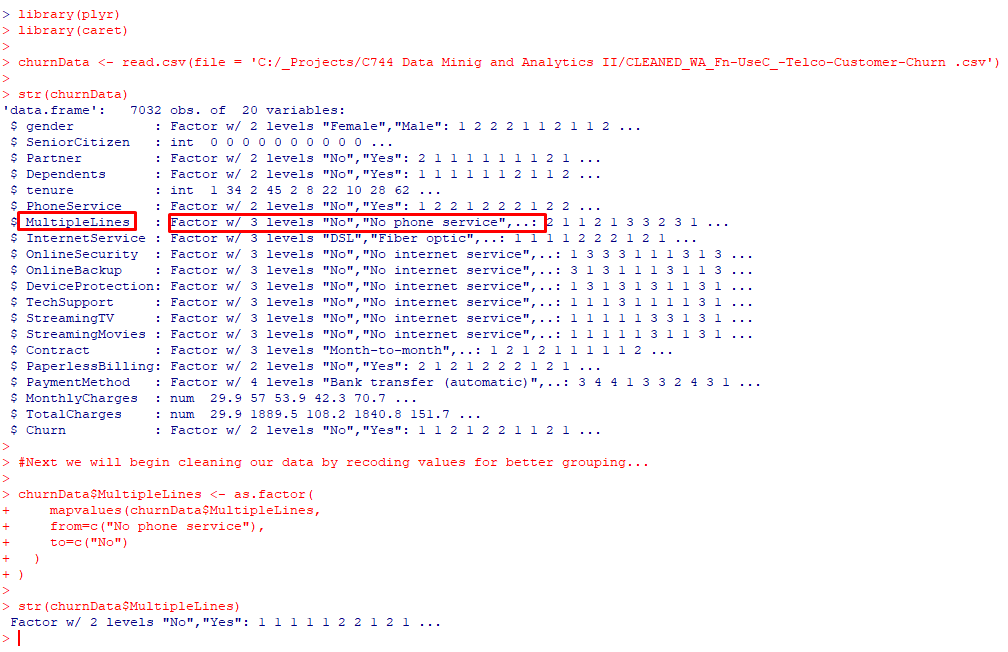
Appendix B



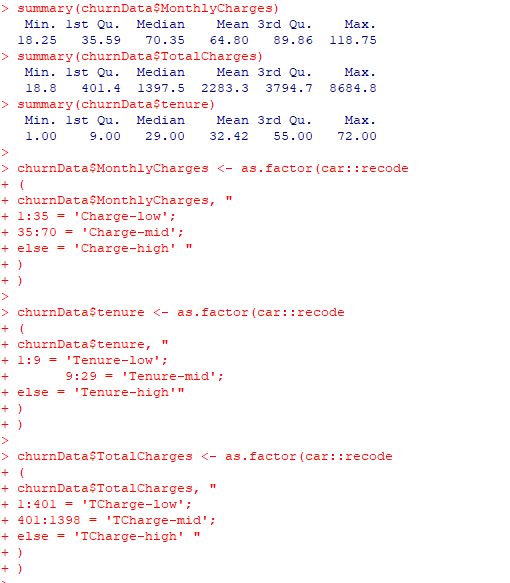
Appendix C



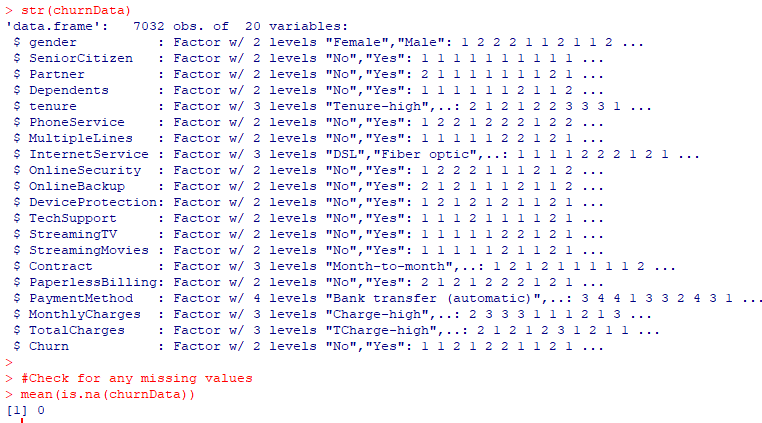
Appendix D



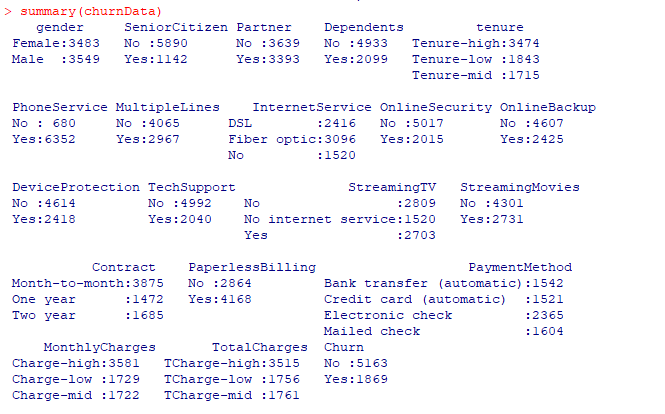
Appendix E

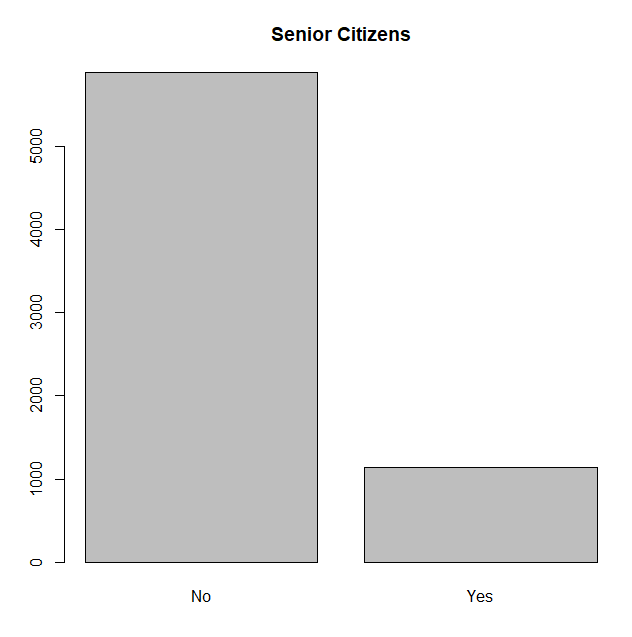


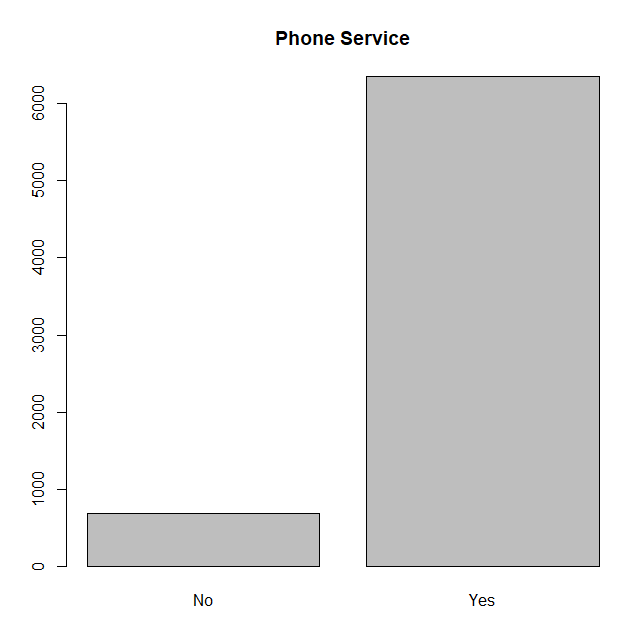
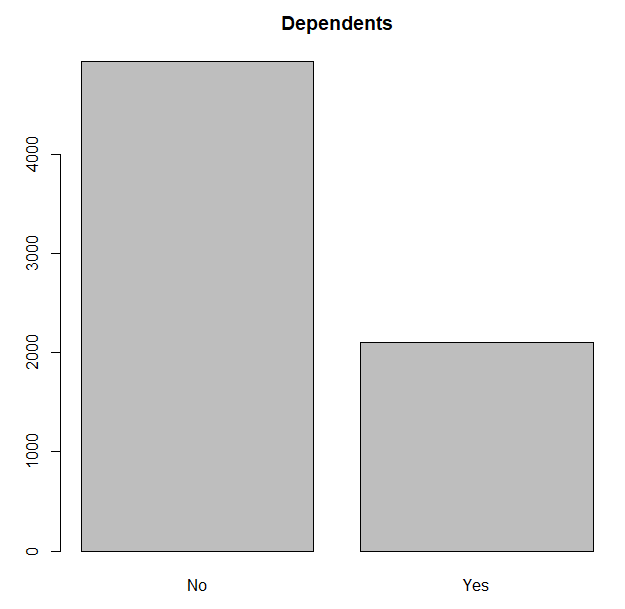
Appendix F

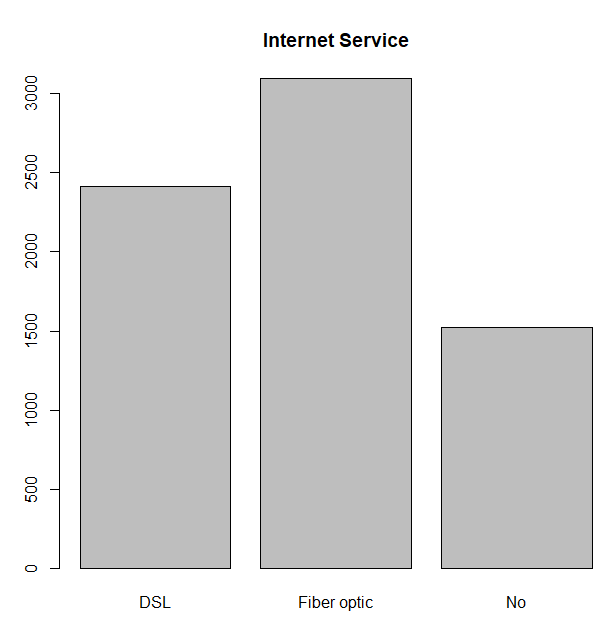


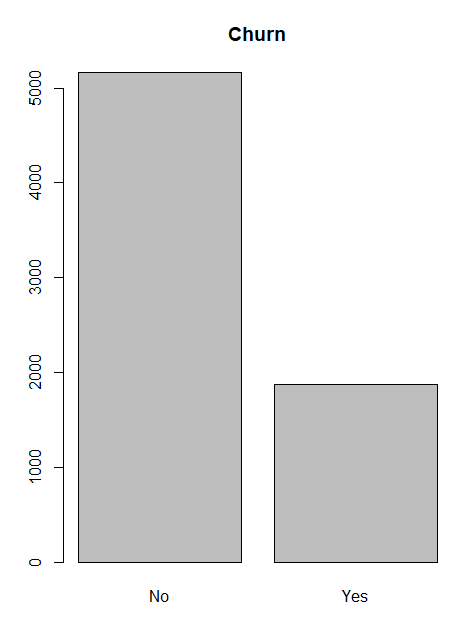
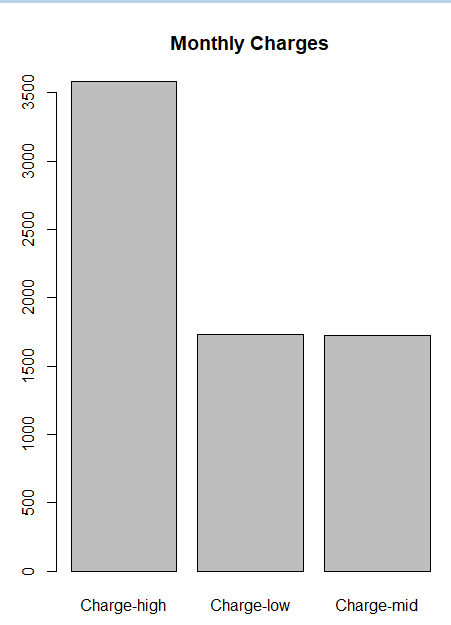
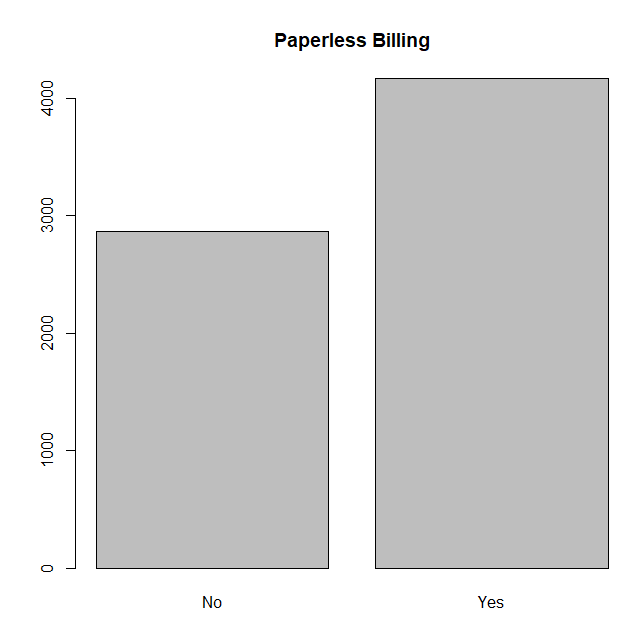
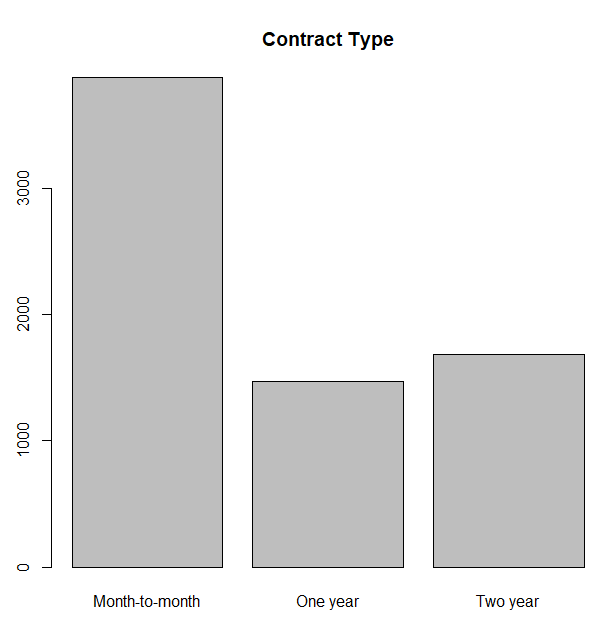
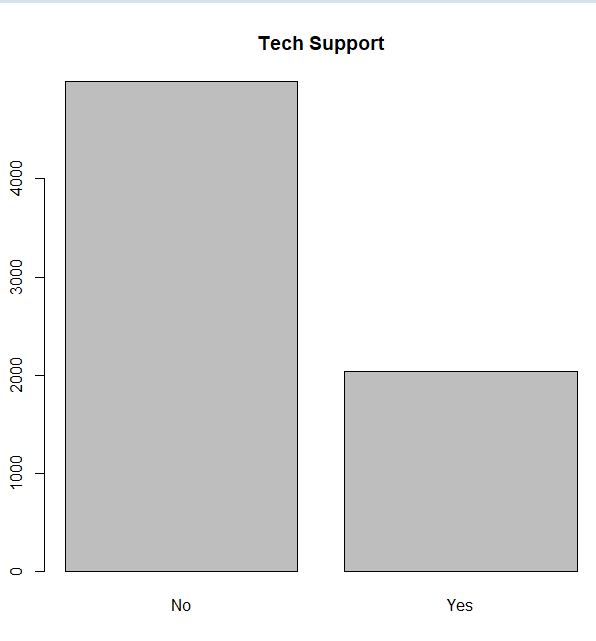
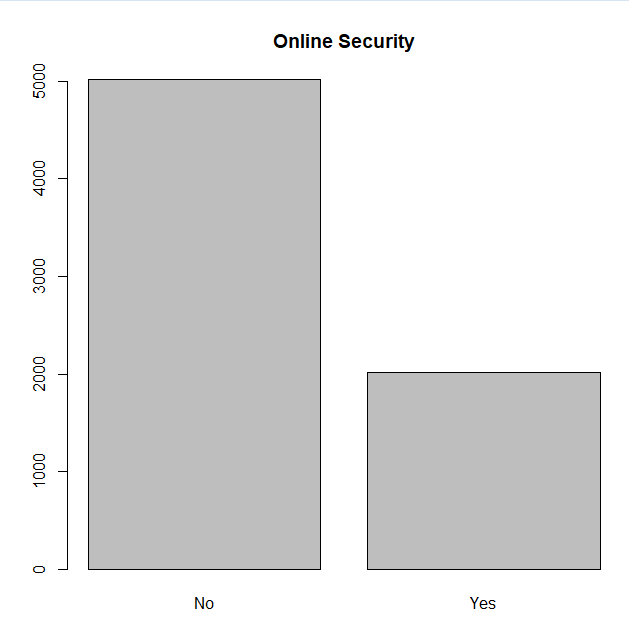
Appendix G



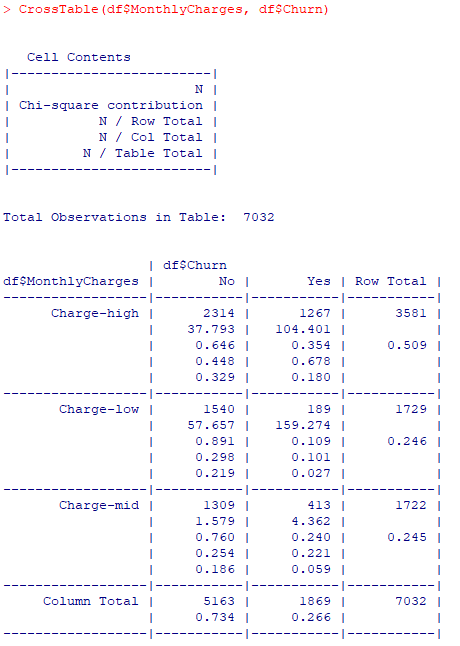


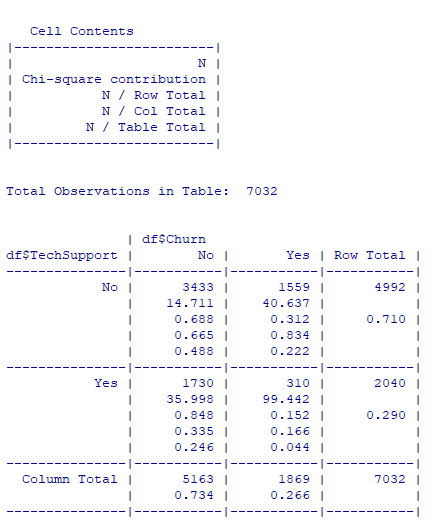


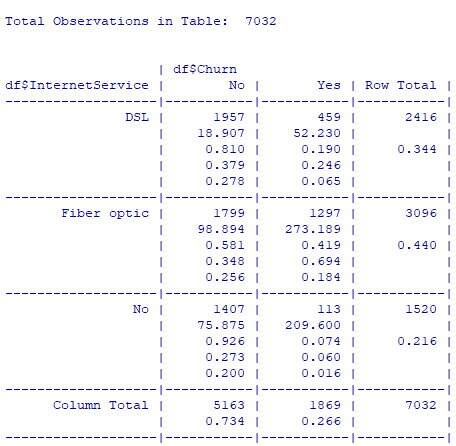
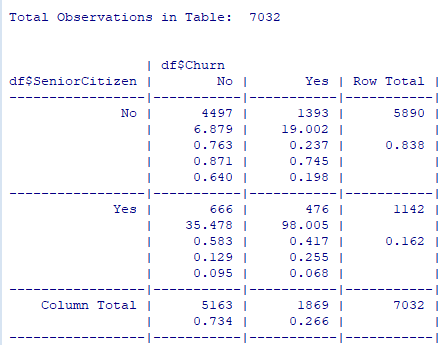
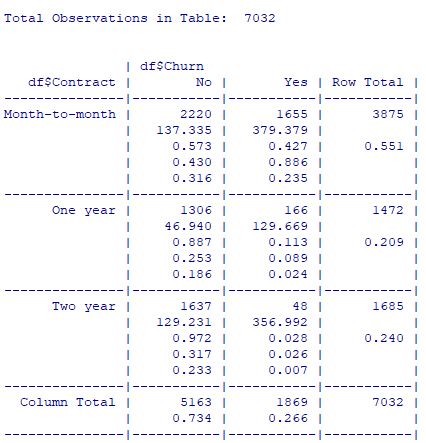
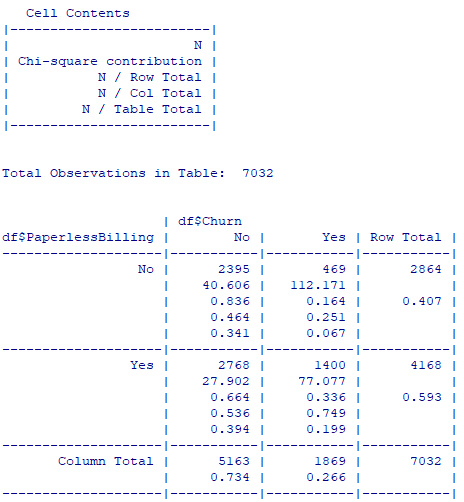


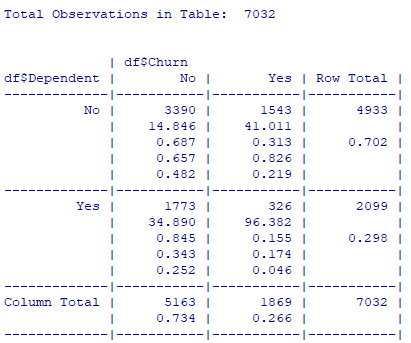


Appendix H

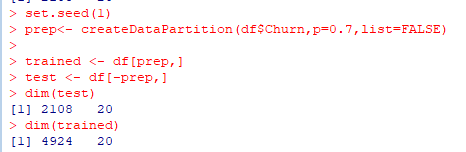




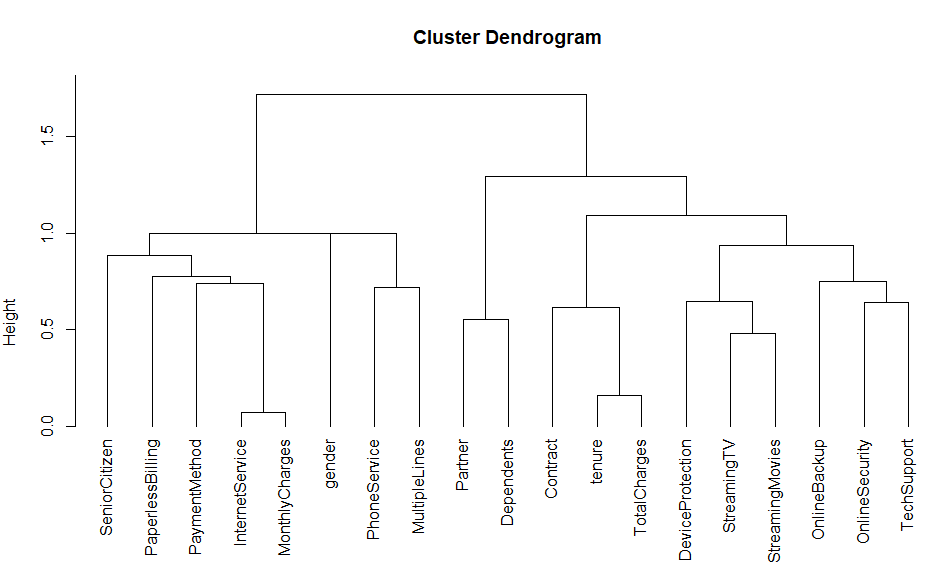




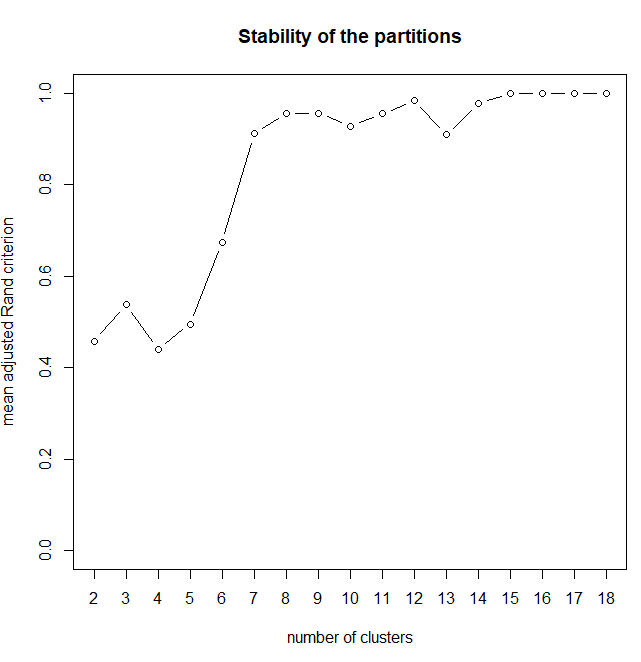
Appendix I



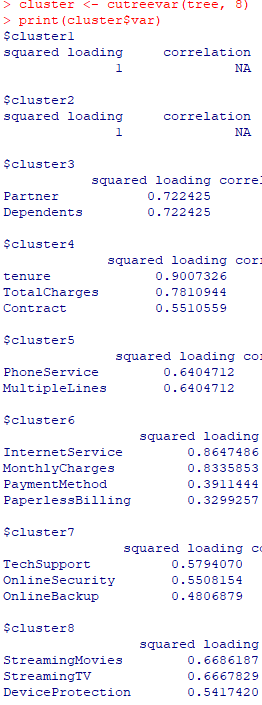
Appendix J



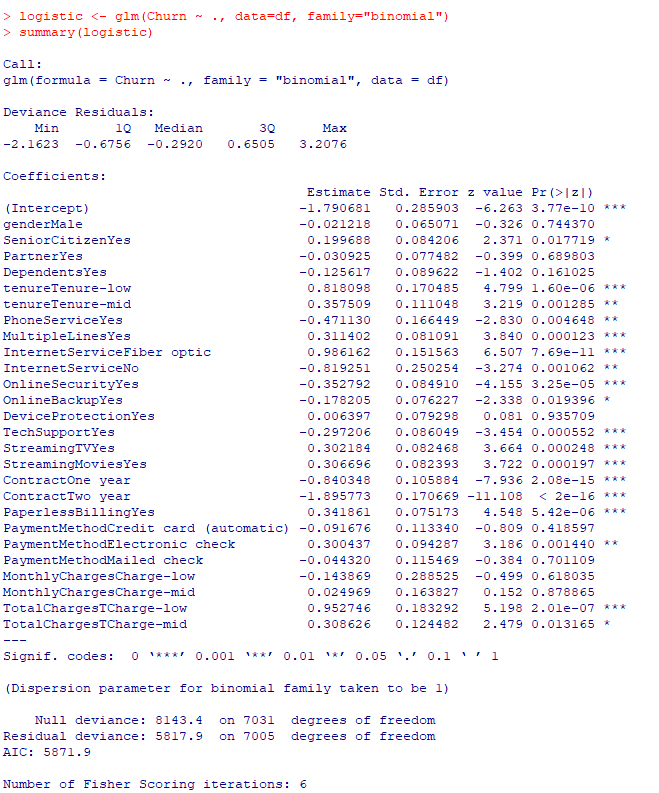
Appendix K



Appendix L



Appendix M



Appendix N

