Customer Retention

Christopher Flowers

College of Information Technology, Western Governors University

C744 Data Analytics and Mining II

Dr. Smith

September 02, 2021

I chose to use the programming language R for extracting data for this scenario. R has several benefits to using it and the biggest is its versatility. R can collect and store a lot of data but also can run any statistic and graphical analysis needed. Also, it is easier to clean and prep the data better than python. The visual representation in R is native compared to in python having to install additional packages, however; packages can be added for R. The statistical advantages are given to R because of its ability to run deep statistical analysis and heavy statistical models, which can lead to the better observation of the data, all this with very few lines of code. For example, by just using the summary method it categories columns to give the count for each string value and give the min, max, mean and median. The following are the goals:

Does customer contract type depend on the customer's age?

Is churn affected by the loyalty of the customer?

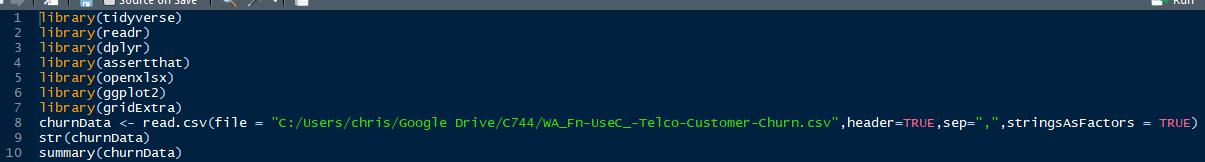
Do paperless billings affect the churn?

Does having internet service affect the churn?

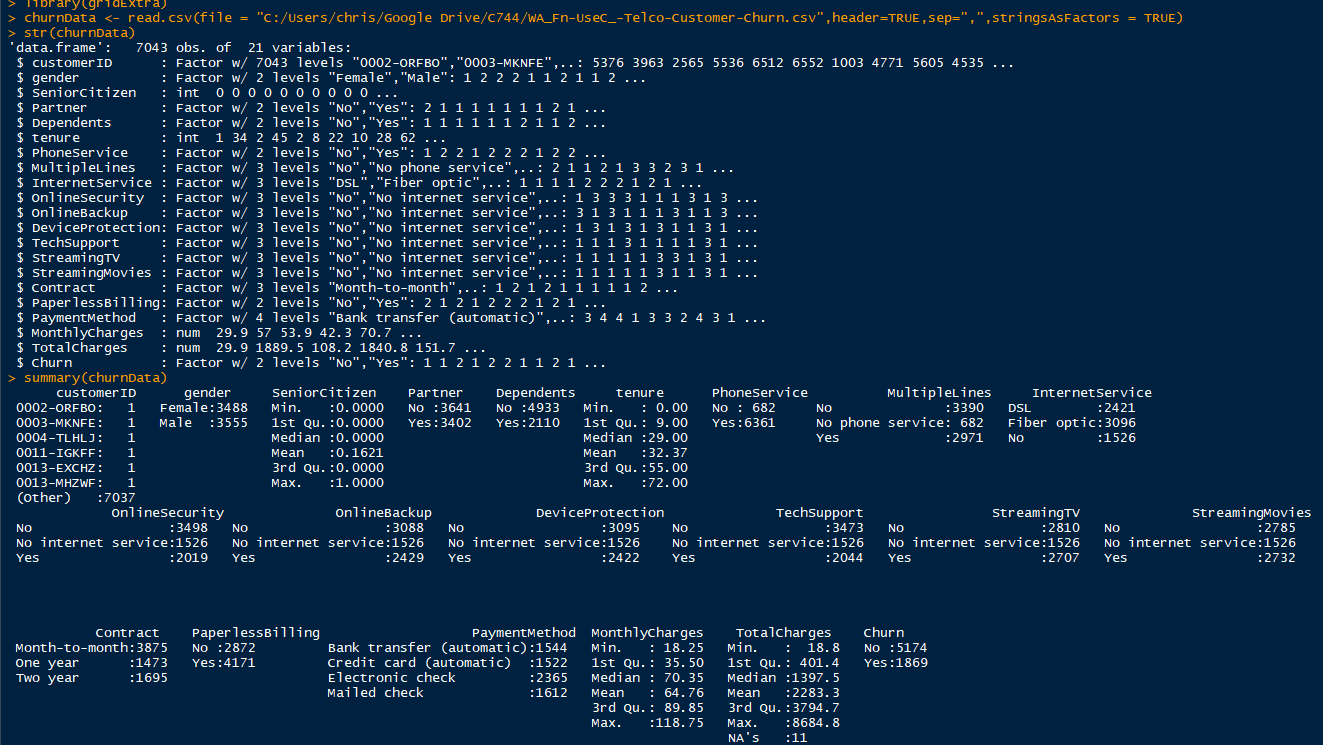
What variables give the greatest gauge of what customers will leave? The goal is to figure out what categories need to be focused on to mitigate the number of customers lost. Also what categories have the highest effect on the churn.

The descriptive method picked to use is Multiple Correspondence Analysis because it deals with qualitative data versus the principle component analysis which deals with quantitative data. This along with FactoMineR helps determine which categories have the most effect on the churn and by helping highlight the categories which have the least effect on churn. The nondescriptive method chose will be logistic regression because of its binomial response variables in how it predicts. This also helps with giving a probability of the churn outcome.

Now we will talk about the extraction of the data from the comma-separated values datasheet. I used the R packages of tidyverse and readr for their ability to extract data. I read the CSV file, where it is setting strings as factors and delimiting data by a comma. The data will be extracted into a variable to use to clean, explore, manipulate and visualize the data. After extracting the data the next thing to do is use the str function to give an idea about the data types and data. Then we summarize the data with the summary function to get an idea of the minimum, maximum, quarter values, counts for categories, and not available data as well.



Code showing how the data is being read into a variable with strings as factors, checking the structure of the data and the summary of the data.



Results of the structure-function and summary function.

Now that we have the data extracted and summarized we need to prepare and explore the data. The target variable in the data set is the churn variable which contains nominal data. Churn either a value of Yes or No as its categories, also meaning this variable is a string or factor. It is known as the target variable because it is the emphasis of the data scenario. Churn is the only variable in the data set, which is qualitative versus quantitative. Below is a screenshot of a glimpse of the data set showing Churn is a factor with two levels or categories and is either Yes or No as stated above.



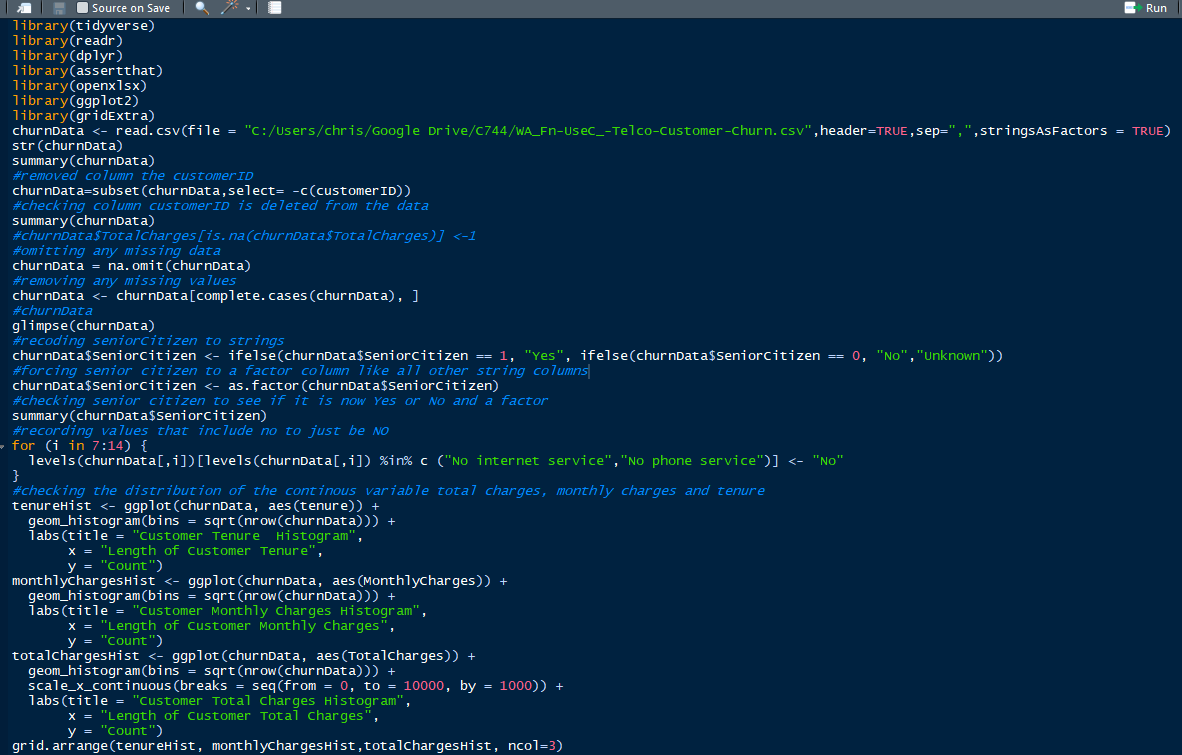
There are three independent predictor variables located in the churn data set and these variables are continuous quantitative data. The first of the three independent variables is tenure. Tenure is how long the customer has been a customer of the telecommunication company. The values of tenure range from zero to a max of seventy-two with a median of twenty-nine and a mean of 32.37. Another independent predictor variable is monthly charges. Monthly charges are what the customer pays the telecommunication company for its services. The values of monthly charges range from 18.25 to a max of 118.75 with a median of 70.65 and a mean of 64.76. The last independent predictor variable int the data set is called total charges. Total charges are total about the telecommunication company has charged the customer. Depending on how long the customer has been a customer of the company varies the amount. For example, the higher the total charge amount indicates a long-time customer versus a small amount will indicate a new customer. The values range from 18.8 to a max of 86884.8 with a median of 1397.5 and a mean of 2283.3.

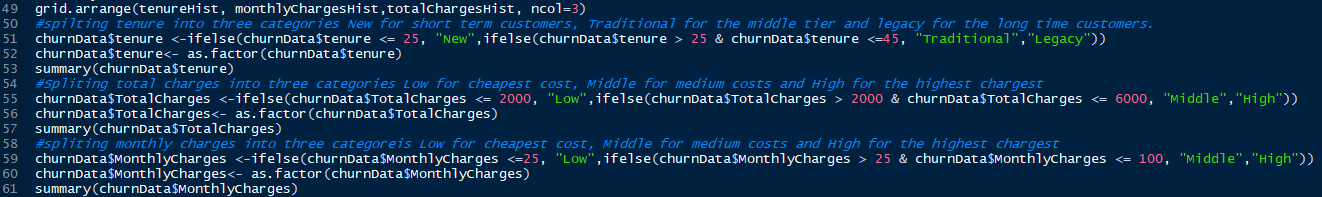
The next process to be completed next for the churn data set is to manipulate it. The goal in manipulating the data is to have a trimmed data, meaning rows and columns exits for the variables and their observation. Another goal of data manipulation is to have cleaned data set. A cleaned data set is removing unneeded columns and missing or duplicated data. Unneeded columns can be for a multitude of reasons for example if a column is a derivative of other columns in the data set or if the columns just don’t have any need to be in the data set for the task. Then change any columns whose data type doesn’t match the needed data type or has too many After the data has been cleaned the next step would be to see if the data set needs to be split into separate data sets if needed. If the data set is very large it could lead to misleading results, also if the data set is too small it could be too small of sample size and not lead to good results as well.

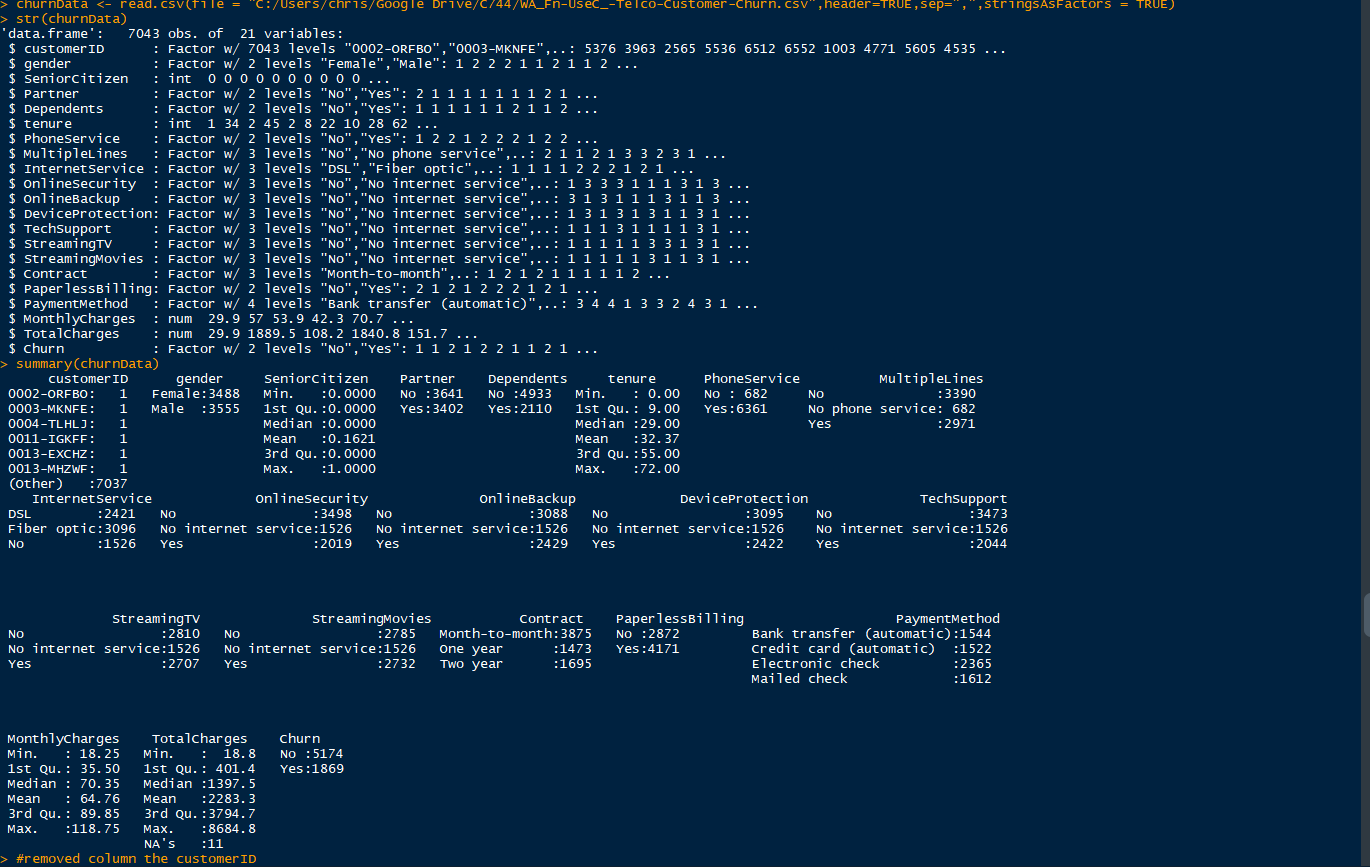
The data set contains three quantitative continuous independent variables. Those three variables are Monthly Charges, Total Charges, and tenure. All three of these fit this because they have numeric values. There are sixteen categorical qualitative independent variables in this data set and those columns are gender, senior citizens, partner, dependents, phone services, multiple lines, internet services, online security, online backup, device protection, tech support, streaming tv, streaming movies, contract, paperless billings and payment methods. Senior citizens data is different than all other categorical independent variables, it is not a string variable but instead is listed in the glimpse as integer, however; that is correct it technically is a Boolean or bit data type which is a numerical way of saying “Yes or No” or “True or False”. In the cleaning process, the senior citizens will be changed to a string column and a factor like the other 15 categorial qualitative independent variables. There is only a single categorical dependent variable in the data set and this variable is the churn variable. It is also the target variable in the data set. There is a single identifier in the data set which is the customer id column. By knowing the customer id you can identify all other variables for that customer.

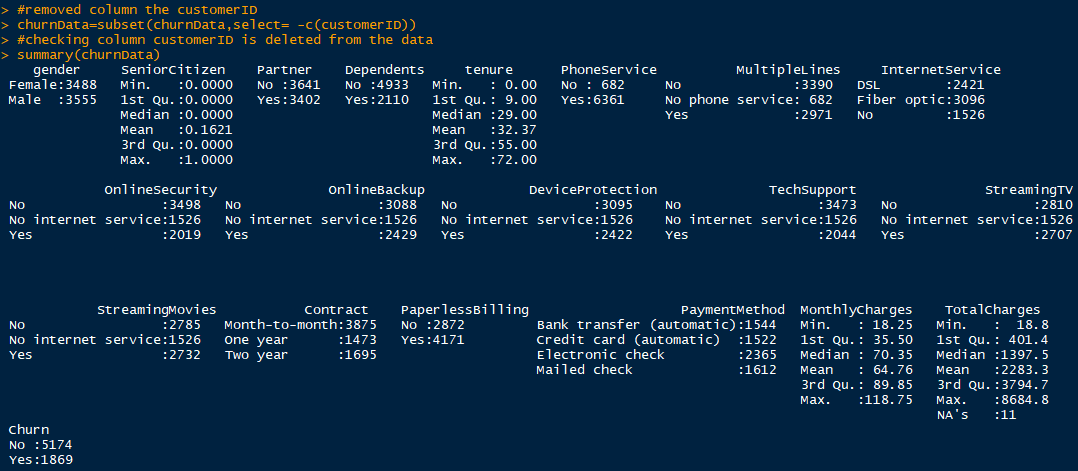
The first thing in cleaning the data was to remove the customerID column because it does not need to determine the likelihood of who would leave the company. We need to check for missing data to do that we will remove it by using an omit function on the data and use the complete.cases function to remove any missing values in the data. Next thing is to change the values of columns that are not in the right data type and the first column to do this to is the senior citizens. Then any column that has a value that includes ‘No phone service’ or ‘No internet service’ to just be ‘No’. All columns need to have a qualitative value for Multiple correspondence Analysis, to do this is to convert three columns. The three columns being converted are tenure, monthly charges, and total charges and all columns will be split into three categories. Tenure was split into categories called New, Traditional, and Legacy. Total Charges and Monthly Charges were split into categories called low, middle, and high. Now the data is cleaned and ready to be split into two data sets. After the data is cleaned we need to spilt it into two data sets, named test, and train at random to a 75% and 25% split. Now that the data is split and cleaned we will export the data sets to an excel spreadsheet.

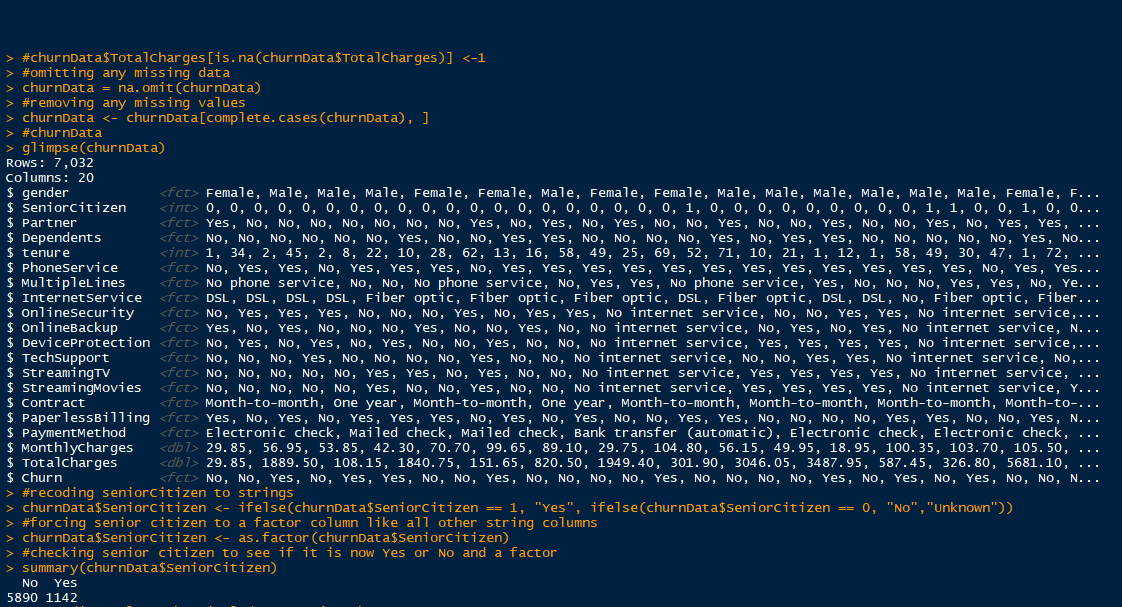
We will also take a look at how all the variables compared to the dependent variable churn. This will be done through bar graphs of the variables with the colored values being the dependent variable churn. One of the objectives stated earlier was to find out if customer loyalty played an effect on the churn. Churn is the lowest in the Legacy and Traditional categories of the tenure variable but highest in the New category despite it also having a high churn number as well. This meaning Legacy customers didn’t affect the churn because they were the majority not leaving the company. The lowest monthly charges churn is in the Low category, with the highest being in the middle, however; it also has the highest values of not churning. Total charges have a very low number of customers churned in the high category meaning customers who have higher total charges are not leaving. Senior citizens have low numbers who have churned and who have not left the company, however; the highest in not churning is not senior citizens meaning age doesn’t play a fact in the churn. There isn’t much of a difference amongst the different genders in who leaves and who stays with the company. A single person is more likely to churn and leave the company with someone who has a partner and a single person without dependents is more likely to leave the company. If the customer has no phone services there is the likelihood to churn is super low but it is also low that they will not churn, which is a higher amount. Multiple lines churn around the same about whether they have more than one line or not, meaning to test churn is not a good indicator. Customers without internet are the least likely to churn than if they have DSL or Fiber optic internet services. Staying with the internet theme online security not churning is higher for those without online security, which are probably the same customers who don’t have online backups. Adding to the internet theme customers who do not stream tv or movies have a higher number of customers who do not churn and leave. The objective of how does internet services affect the churn looking at the overall three categories inside internet services customers are most prone to not churning the company. Now looking at devices and support, device protection and tech support are mostly prone to not churning on the company. Customers who are on contract for more than a year at a time have churned the least. Let's see if the payment method affects the churn, those who do not churn are about the same across the categories but the electronic check customers are most prone to leaving. Paperless billing is the highest for those who have paperless billing set up but it doesn’t affect those who stay, which was an observation we wanted to know. The leads to another objective which does age affect the customer's contract type. Regardless of age, most customers chose to have a month-to-month contract type versus a long contract type. Senior citizens rarely chose to have a contract type.



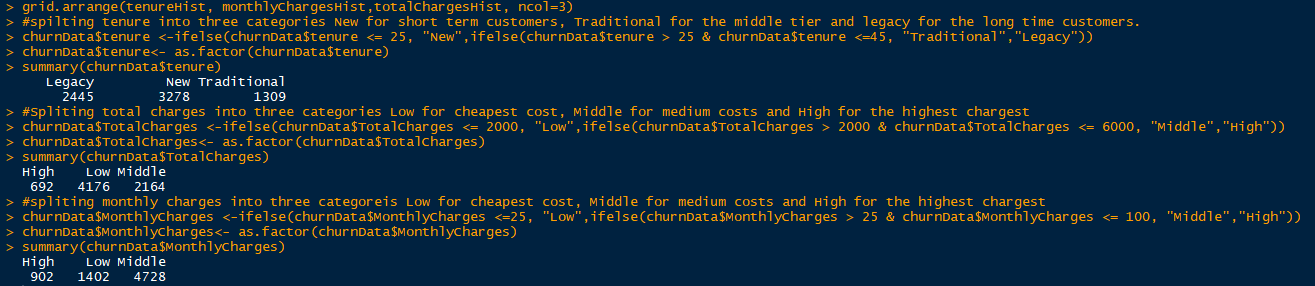


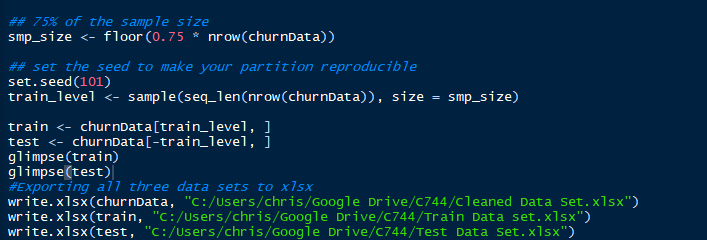


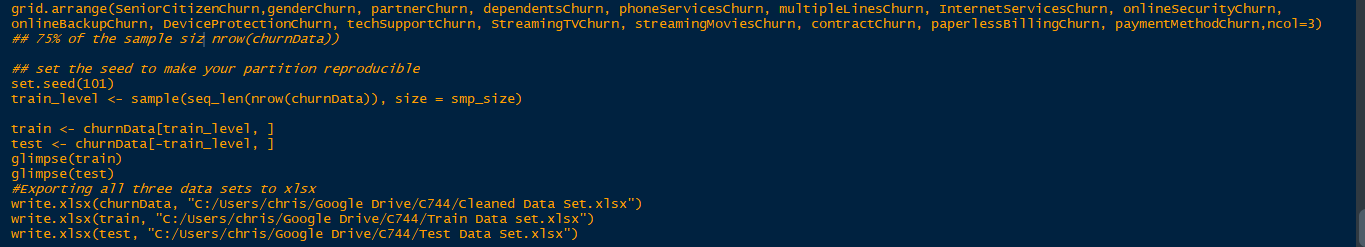


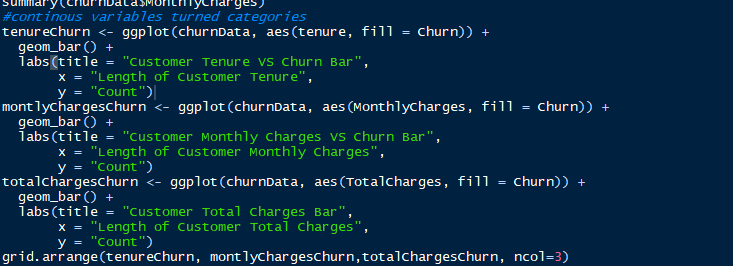


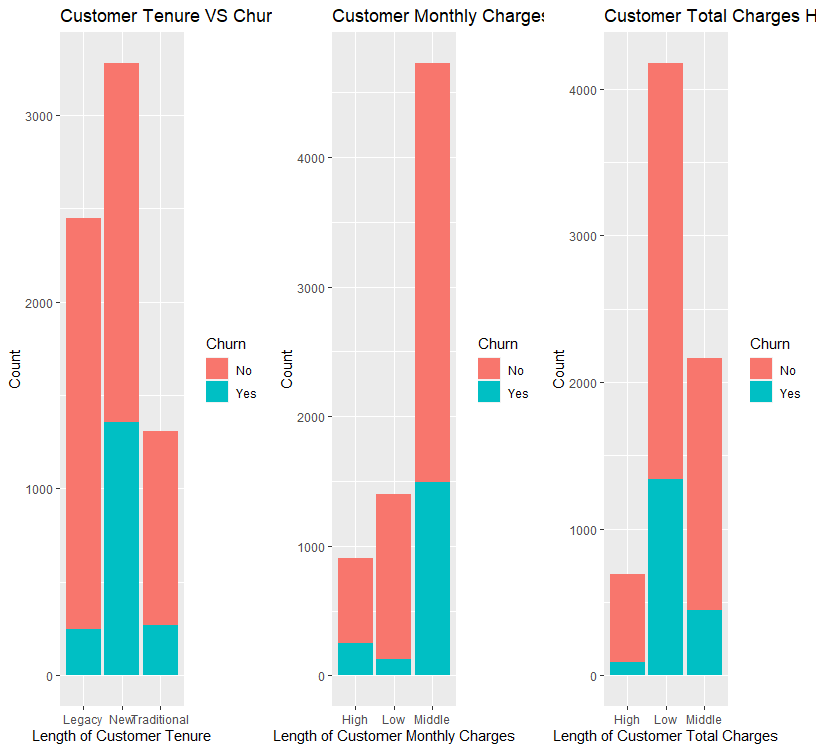


******

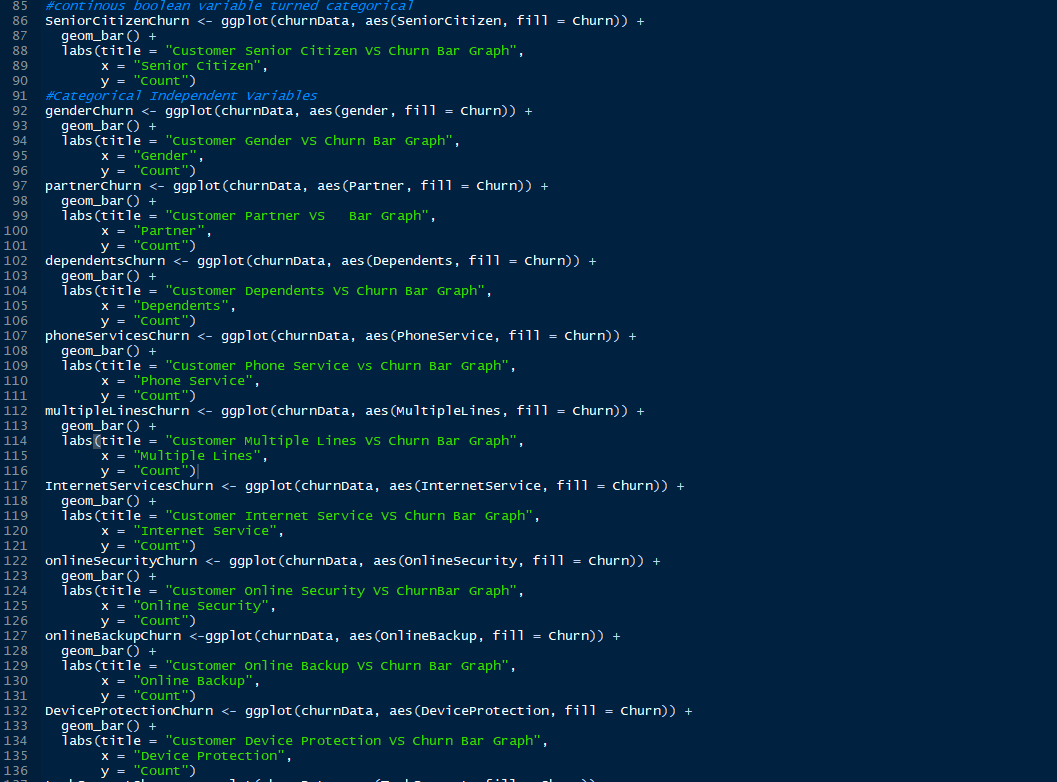


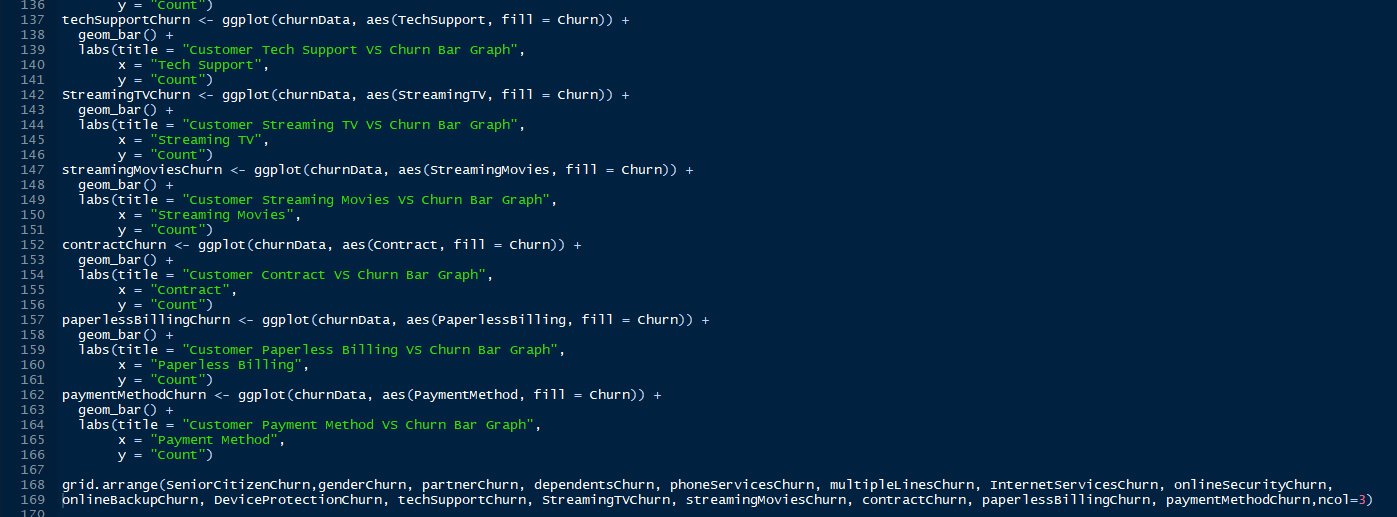


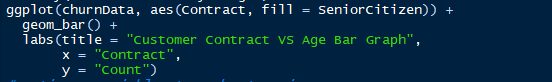


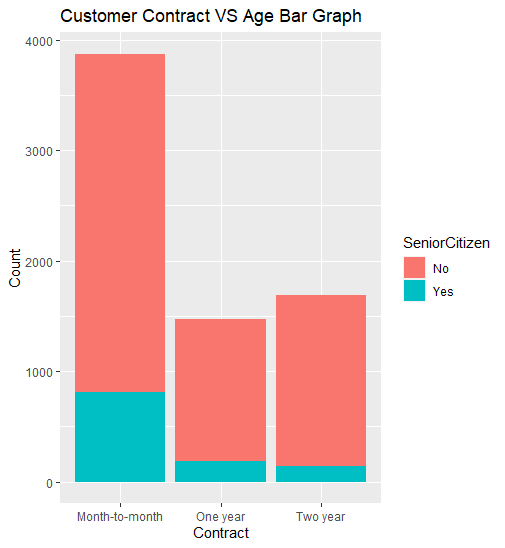




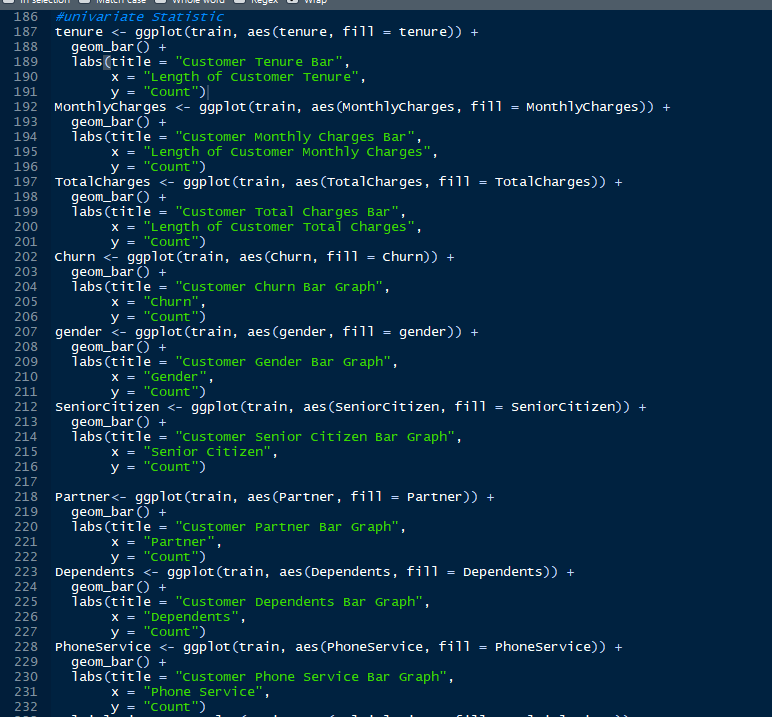
******

******

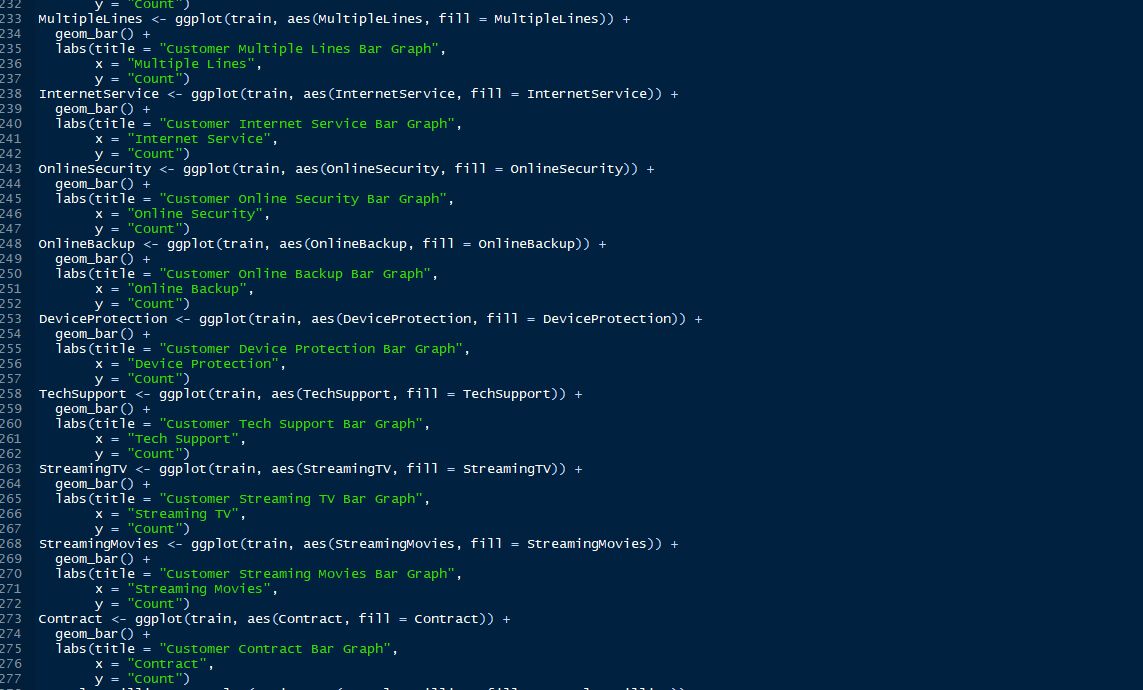
******



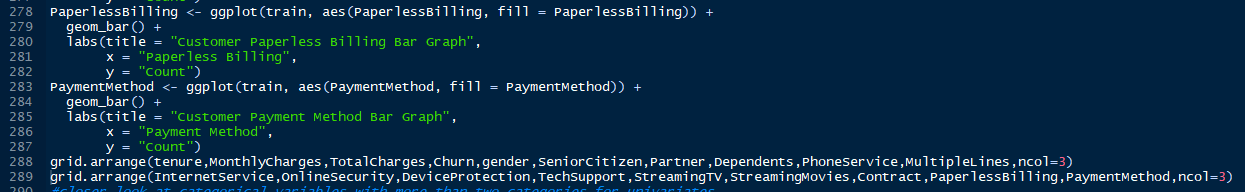
Next, we need to look at the data analysis of this data set. We have already done a distribution on all the variables in the previous section but that was their distribution versus the churn or different variables. We will look at the univariate statistic of each variable including the dependent variable churn. Univariate is only looking at one variable at a time. The variables used bar charts to show the distributions amongst the subset of each variable. Below are the screenshots of code to chart the univariate statistic and the graphs as well, with annotations below each screenshot.



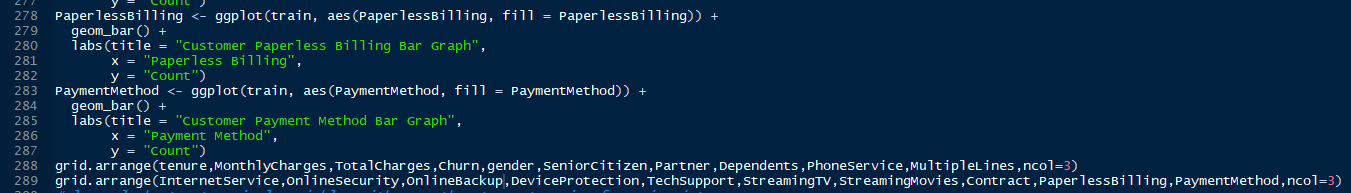
Code part one because all variables charts wouldn’t fit in a single screenshot. Each chart is set to be a bar chart with a fill of the variable being plotted to give the chart's color.



Second screenshot of code to show univariate statistics through charts.



The last screenshot of the code to chart the univariate graphs than to arrange all the charts to be placed on two pages with three columns.



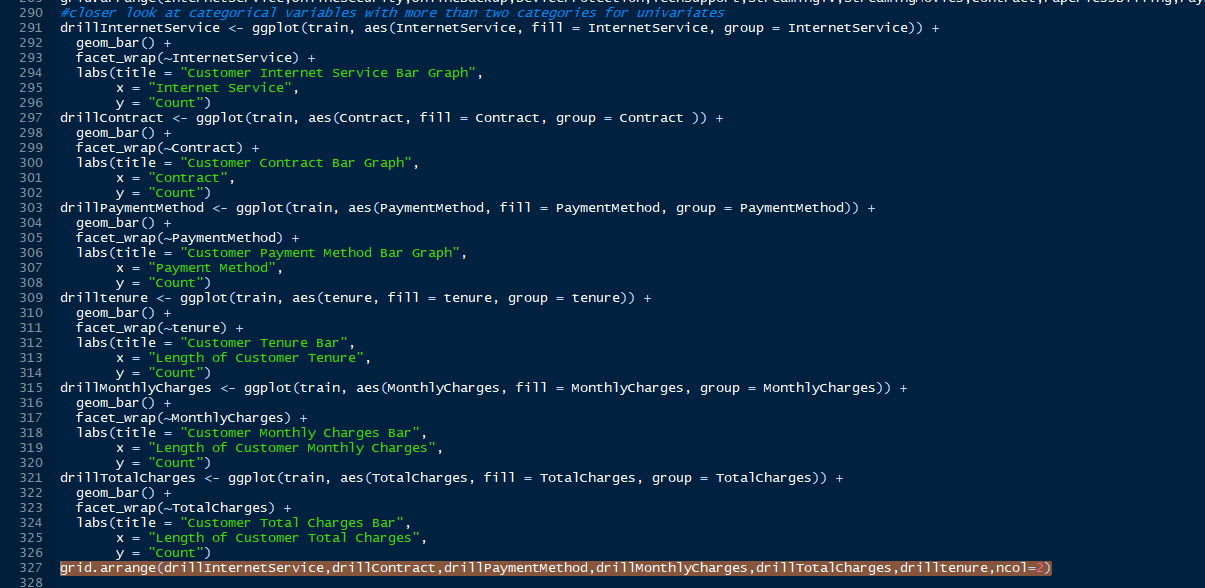
The two screenshots below this are graphically showing the univariate statistic, also is the visual side of the previous code screenshots.

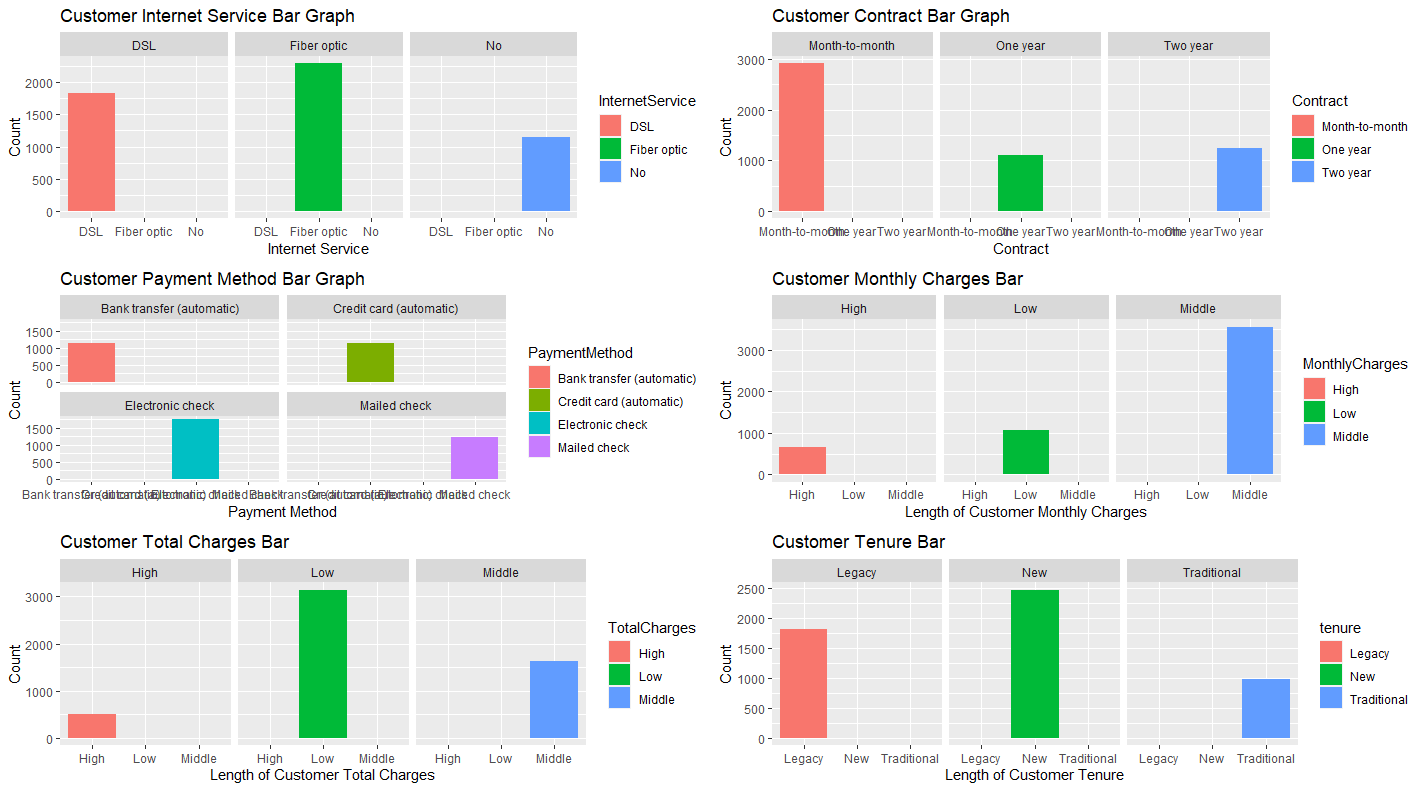
The company has more new customers than legacy or traditional customers in the tenure variable. The middle level in the monthly charges variable is by far more than the high or low levels. Most customers are at the low level for total charges, meaning the company has a lot of newer customers. The company has had fewer customers churn than they have churned. There is not much difference between genders but males do have a slight advantage. The majority of the company’s customers are not senior citizens and in the train, data set there are less than a thousand senior citizens in the data set. In the partner variable, there are less than 500 hundred customers difference but the company has more single customers than partnered customers. The majority of the customers have no dependents with this observation it can be said the majority of the customers are single people. A vast number of customers have phone services than those who do not. More customers have multiple lines than those who only have one line.



More customers have internet services than those who do not. Most customers of the company do not have online security or online backup by a decent margin. Also, most customers have do not have device protection nor do they have tech support, both of these are true by a majority of the sample size. Streaming tv or streaming movies are options customers have and more than half of them. Based on the univariate the customers intend not to use the internet for more than casual use. Most customers are not on a contract than those who are and more customers like to use paperless billing than those who do not. Customers like to make out electric checks than any other payment type.

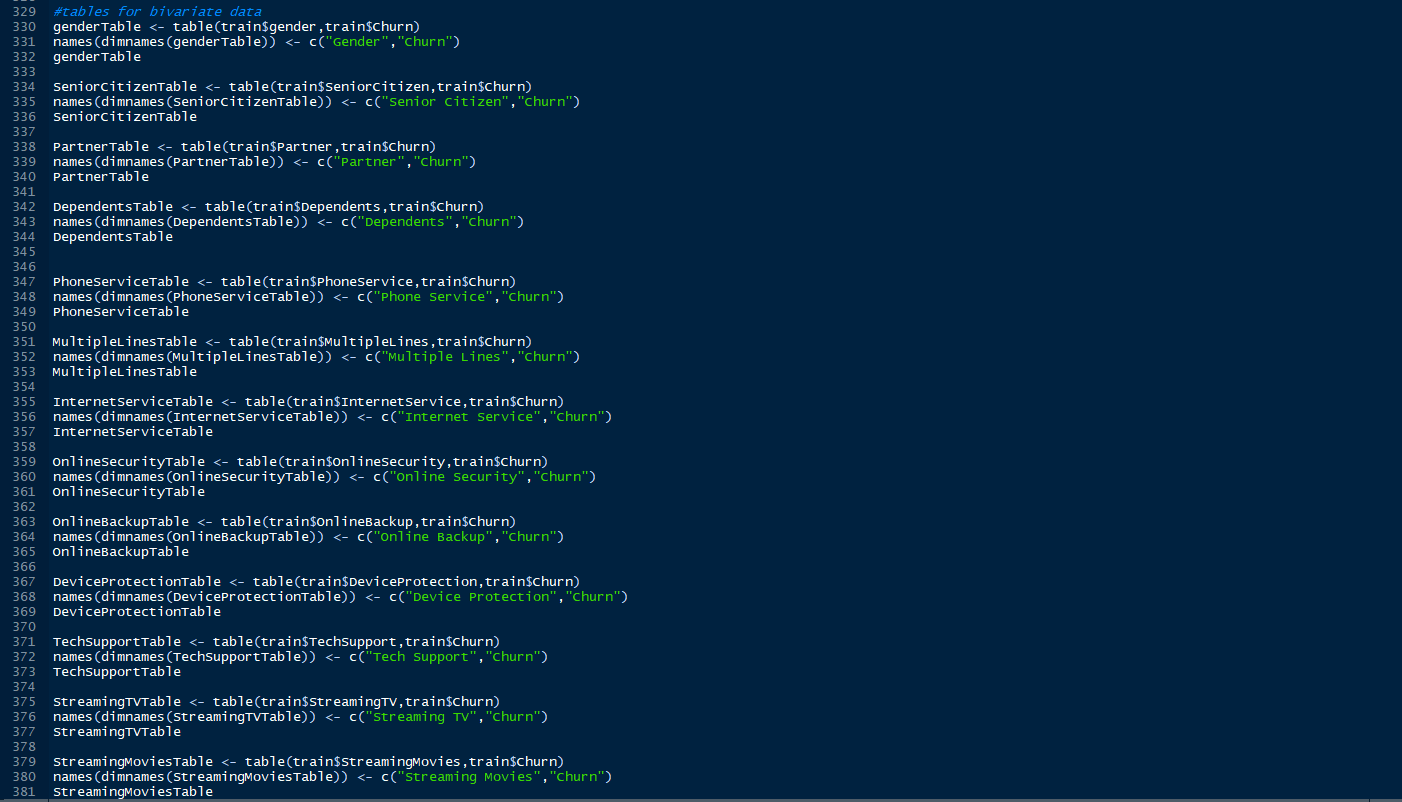
Now let's look at the 3 level variables closer with screenshots of the code and then visual graphs of the code.

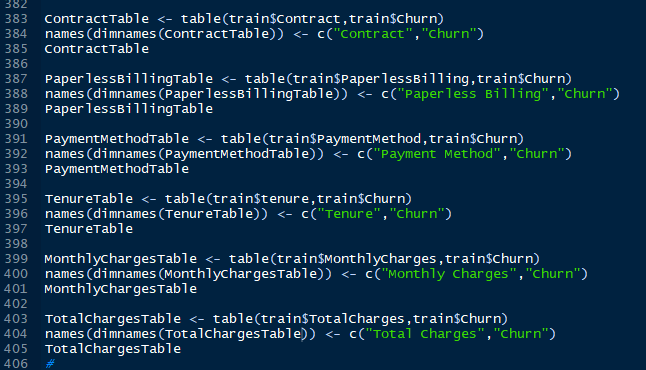




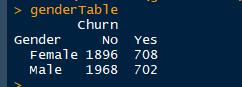
Most customers have fiber optic for their internet service. Customers far less have chosen to have a contract longer than a month. All payment methods are about the same as the other except for electronic check. Monthly charges are far less in the high and low levels than in the middle level. High-level customers have the least amount of customers and when this is compared with tenure’s legacy level being its second-highest level, leads to legacy customers aren’t paying more than newer customers.

Now we will continue to analyze the data in the datasheet. We have already looked at the univariate of each variable and its graphical representation. Also, we have annotated the graphs for every chart. Now we will look at bivariate statistics, which is looking at two variables together. We have already compared the data graphically and annotated the graphs. We will now look at the bivariate through tables instead. Below are the screenshots of code to chart the univariate statistic and the graphs as well, with annotations below each screenshot.

******

******

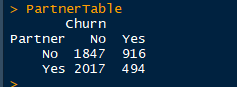
We create a variable with the table function and both variables to compare. Then used the names and dimnames function to add the columns to the table then called the variable.



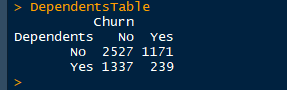
When comparing gender and churn there is no difference. Those churn the different genders were 6 customers.



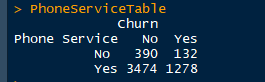
The table above shows a huge correlation between not senior citizens and not churned customers.



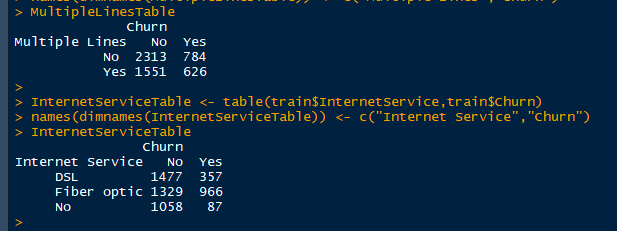
The table above shows a difference between these variables with churned and having a partner being over 2000 customers.



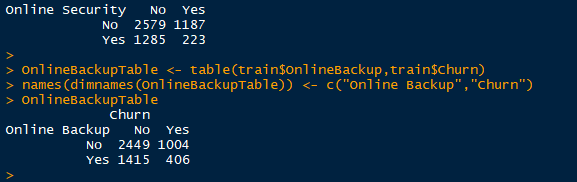
Churned and having dependents by a vast majority less than those who have not churned and dependents.



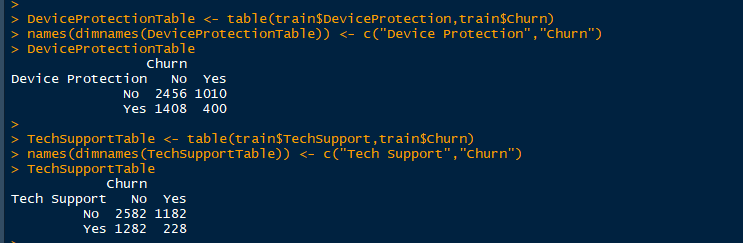
Most customers have phone services but have not churned. The least has churned and no phone services.



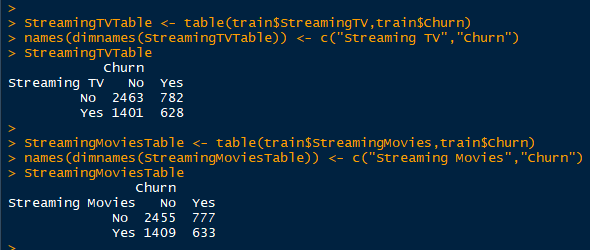
Not churning and no multiple lines are characteristics of most customers. There are comparable results in churning with multiple lines. A small number of churned customers have no internet with most having not churned with DSL has their internet service type. The number for the churned customers having no internet services 87.

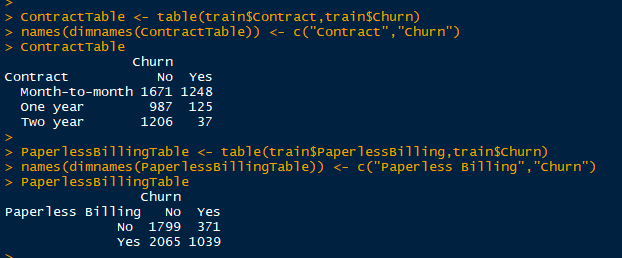


223 customers have churned with having online security with the majority being a combination of the two variables. The greatest is not churned and no online security with a value of 2579. Online Backup and churned customer combination are by var the least with less than 500 customers at 406. The most being not churned and no online backup with 2449 customers. It can be stated that most customers have not churned and no online security or online backup.

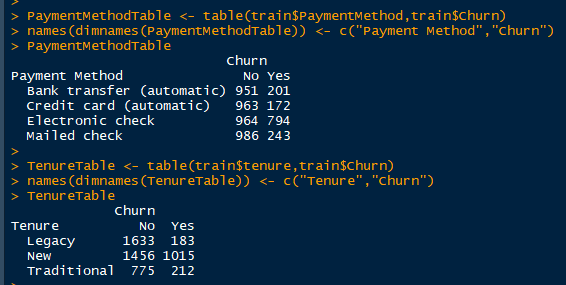


Most customers have not churned and have no device protection with 2456 and the least being 400 having churned and having device protection. Tech Support and churned customers by far the lowest at 228 customers. Most existing customers have not used tech support and this is evident by the high number of 2582.

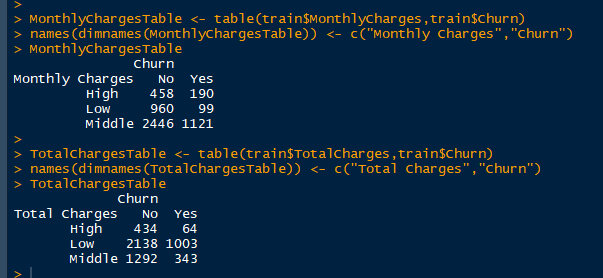


Customers not streaming tv and churned are the most common with 2463. Churned and not churned customers are less than two hundred customers. Also churned and not churned customers are comparable but the most favorable is churned and not streaming tv with 2455 customers. Most customers have not churned and do not stream movies or tv’s. 

The rarest customer combination is churned with a two-year contract with less than 50 customers at 37. This is a huge difference compared to those who have not churned and are month to month at 1671. The most common theme is not churned and the variable as yes but churned and paperless billing is the most common for this table, at 2065 customers.

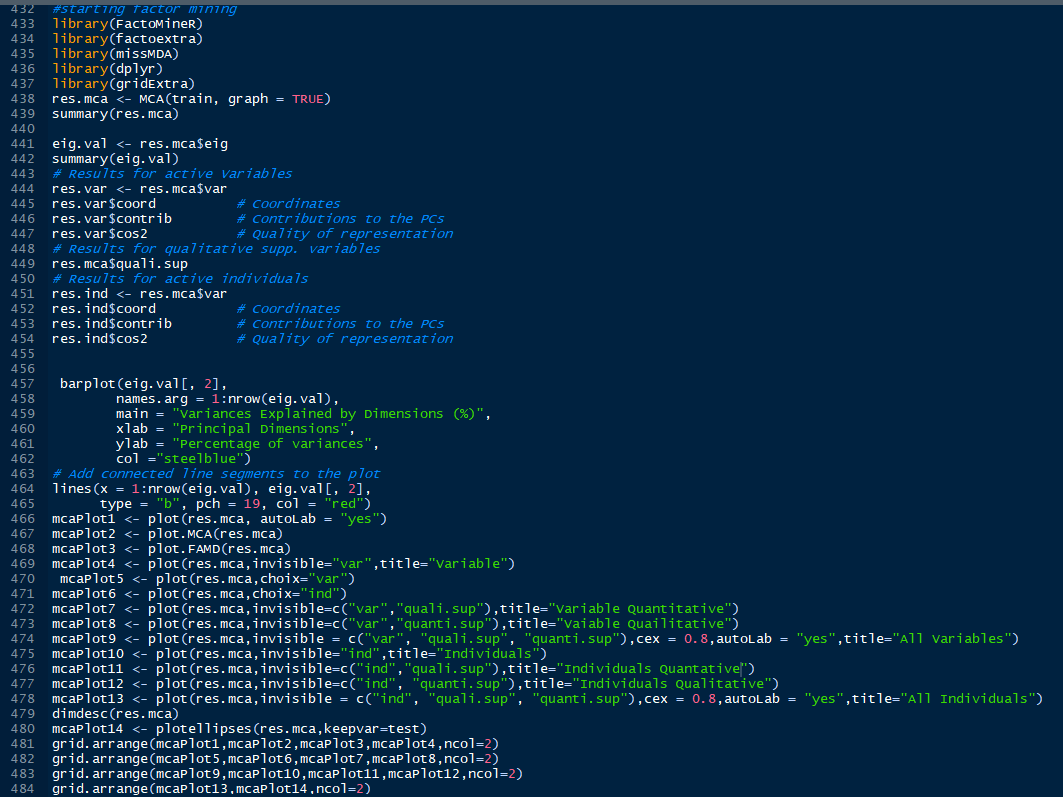


Payment method not churned numbers are less than 50 customers between them all. All of them are over 950 customers. Churned customers are different with most under 250 customers expect electronic check at 794. Not churned customers have higher by about double than the traditional tenured customers. Churned customers which have newer customers are far more common than other customers with 1015.

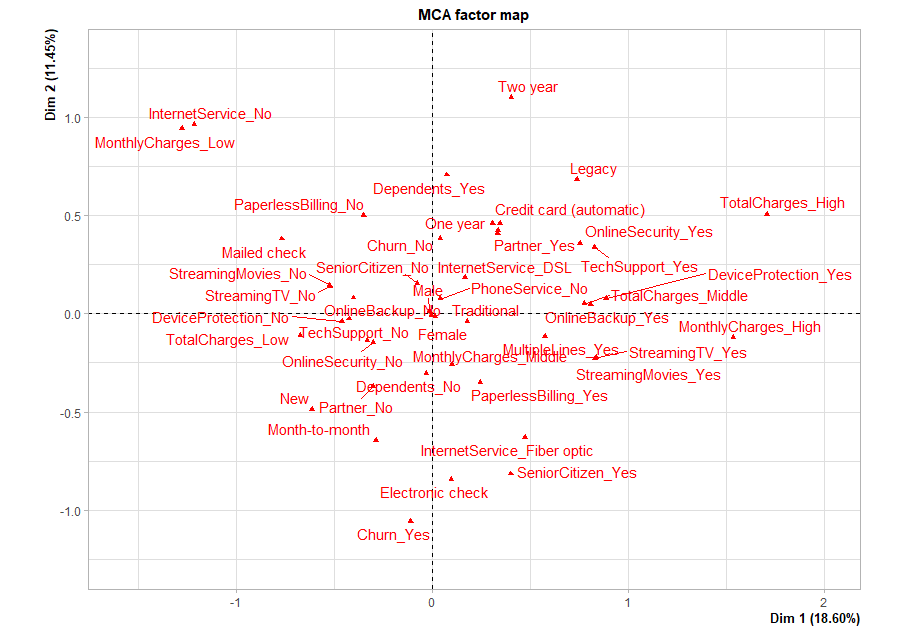


Just at a glance, it is easy to tell most customers to fall in the middle level with not churn having over double the amount of churn with 2446 customers. The same could be said about total charges and churn with most falling in the low level. Not churned and low is the most common amongst customers at 2138. Churned and high total charged customers is rare with under 100 customers at 64.

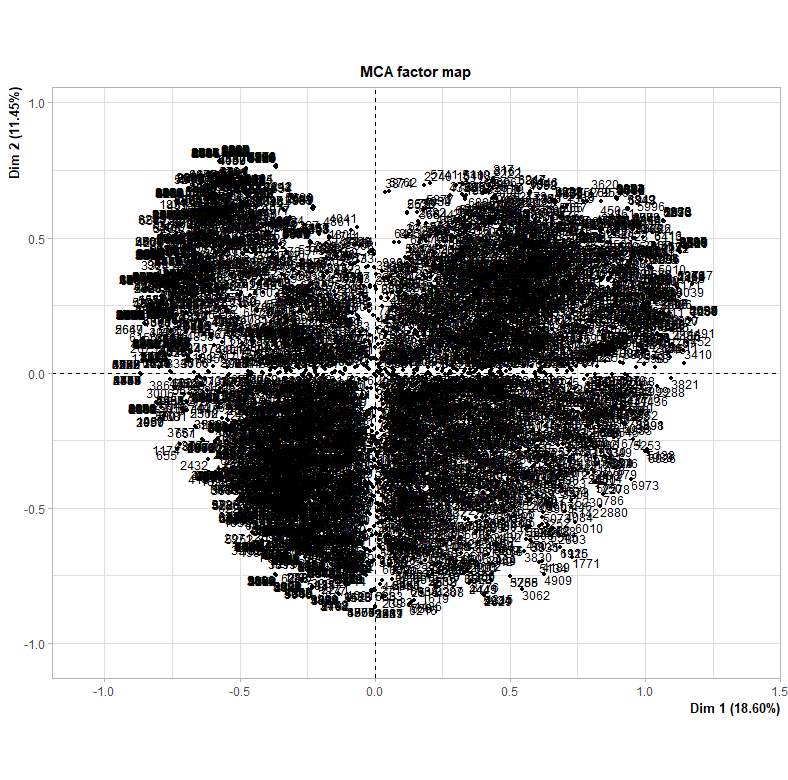
Now we will start applying analytic methods to the datasheet and evaluate the methods used. As stated in the beginning the descriptive statistic method going to be used is Multiple Correspondence Analysis or MCA for short. MCA is a type of Correspondence Analysis(CA) just like Principal Component Analysis is another statistic. The reason for choosing MCA is because it is meant to use two or more categorical data, which is what we have in this instance once the data is clean.



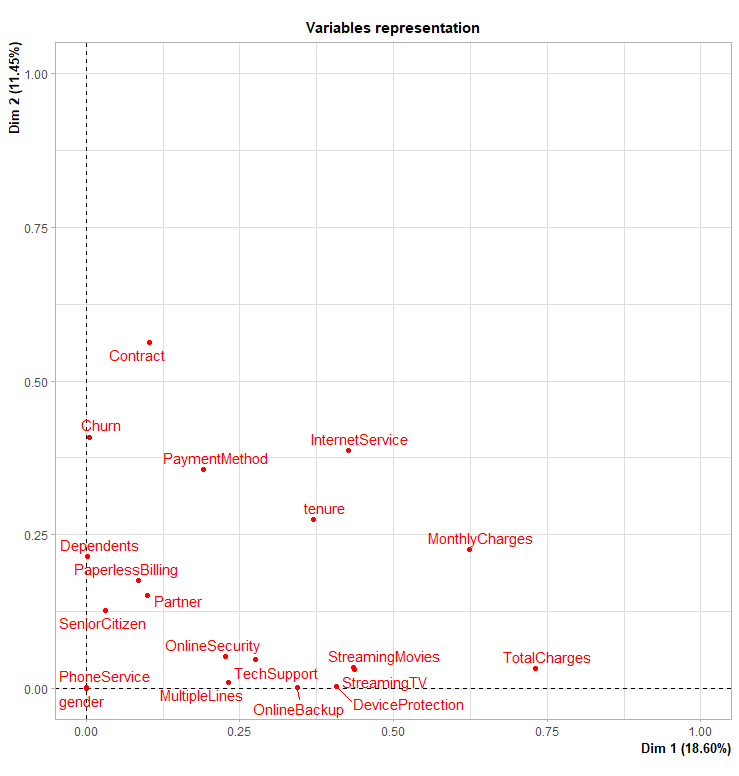
As stated earlier in this report we will use the FactoMineR to create our MCA data. This is done by using the MCA function in the package. We use the MCA function with the argument of graphing the MCA as true with all this saved into a variable then summarize the variable. The next thing to do is find the eigenvalues in the MCA variable and summarize the findings. Next find the individual, active, and qualitative supplementary variables. Then plot MCA in different ways showing different variables.



The plot above it shows two dimensions with all the factors and how they correlate to one another, also how strong the factors are. Dimension one runs east to west or across the x-axis and dimension two runs north to south or across the y-axis. The factor Churn\_Yes will have a negative correlation on dimension two because of its factor value of -1.0 and it will barely have a negative correlation dimension one because it is close to zero. Another example would be female it is in the center so it doesn’t have any correlation on either dimension because of this it is not a priority factor.

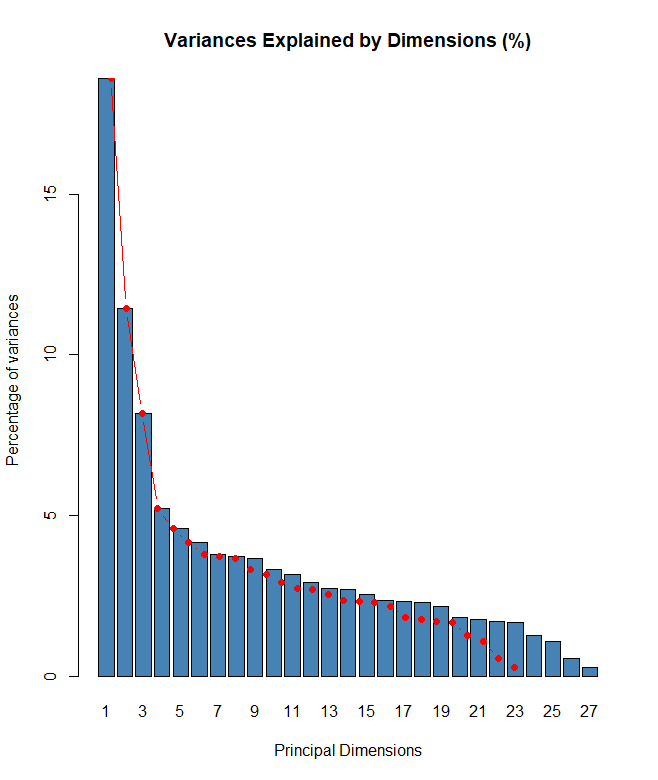


The plot above shows the individual variables and active variables. Other than showing what percent of the data set is in the dimension this chart does not help evaluate the data.

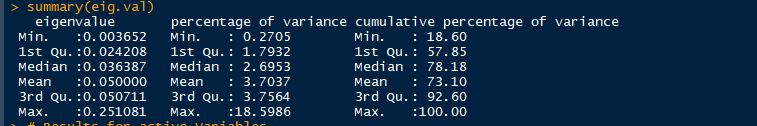


In the plot above we are only looking at the variables instead of the factors. As shown gender and phone services do not correlate with either dimension. Contract and Churn have a strong correlation to dimension than the other variables based on this plot. Total and monthly charges have the strongest correlation to dimension two.

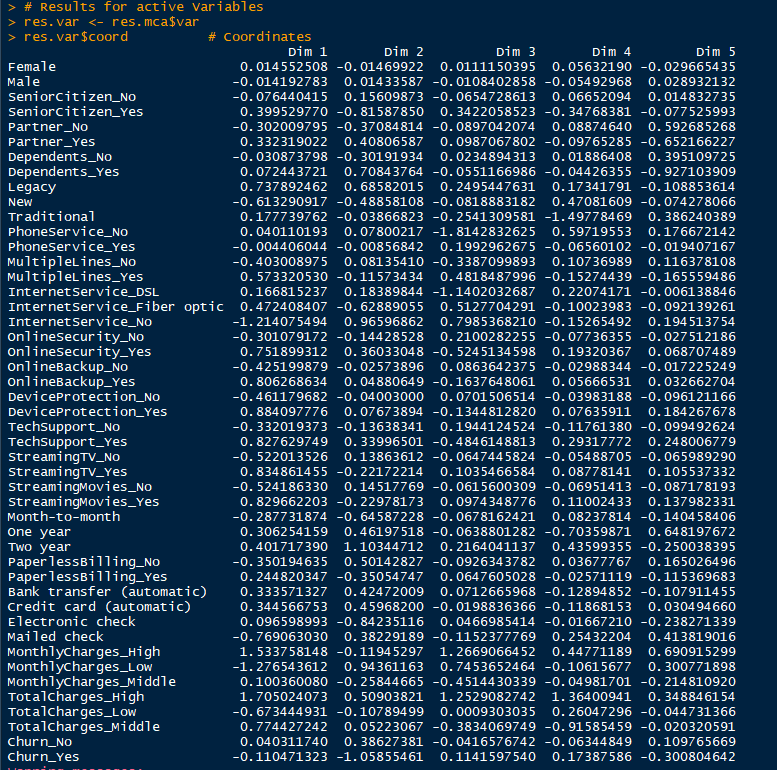
The screenshot above looks at the results of the MCA function in table format. Looking at the percentages variance of the dimensions the only dimensions that carry the most weight are the top 3 dimensions. This shows the first ten categories with a statistic of it’s in the first 3 dimensions, also it gives ten categories in table format.



Graphical implementation of why the top three dimensions are the only ones looked at. After the third dimension, the line of best fit slowly curves down and shows not much changing.



This shows the summary of the eigenvalues and shows that dimension one was the highest percentage of variance.



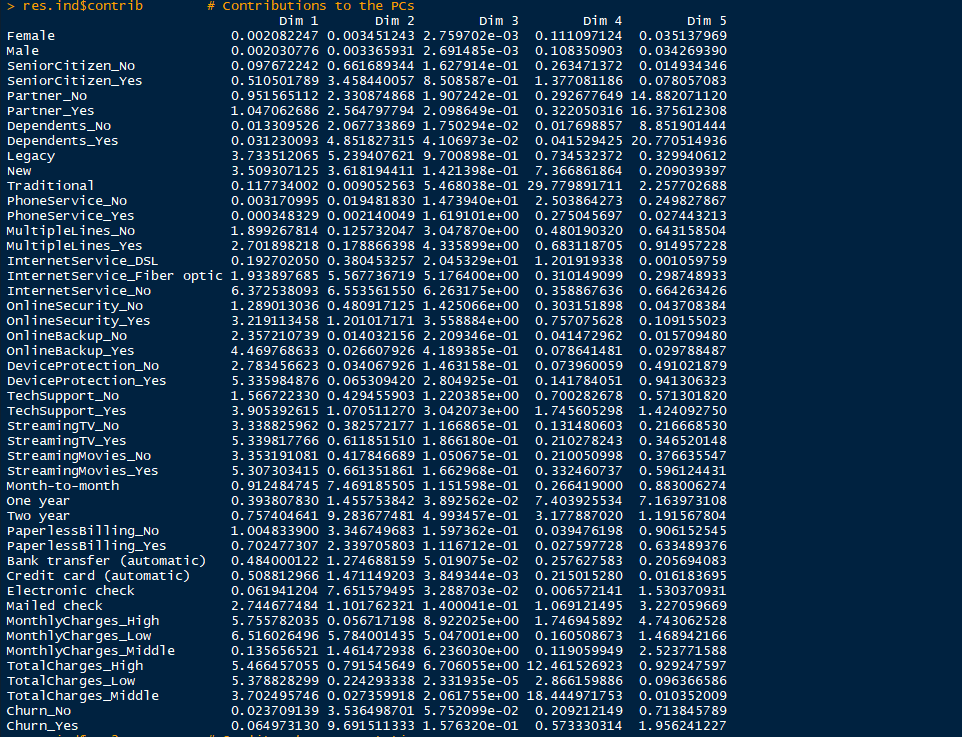
Shows the coordinates of all variables.

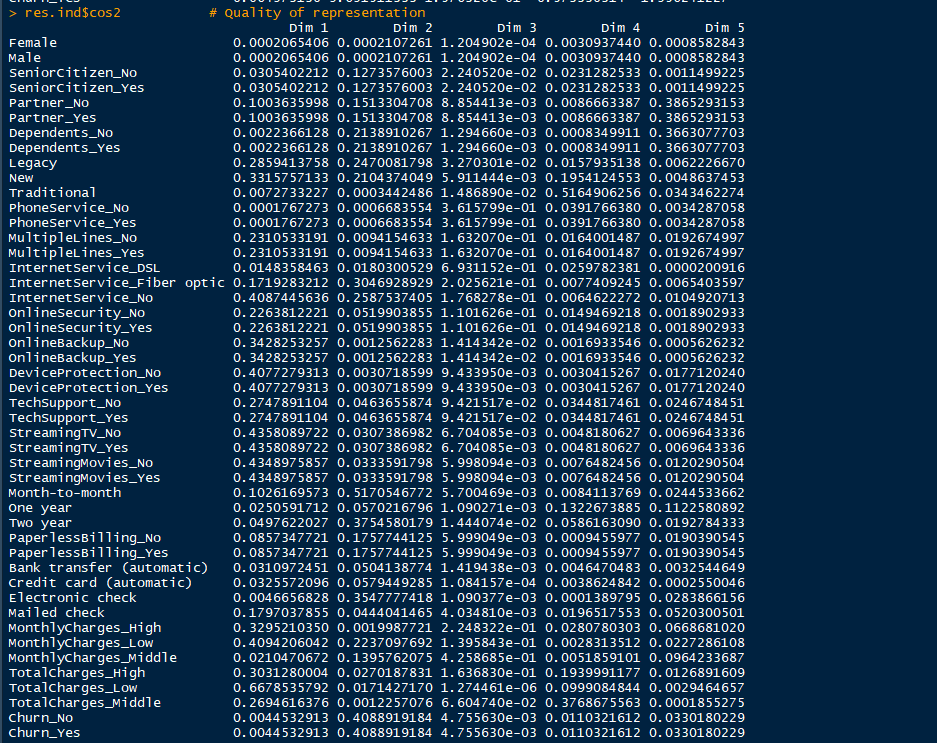


Shows the contribution of all categories.

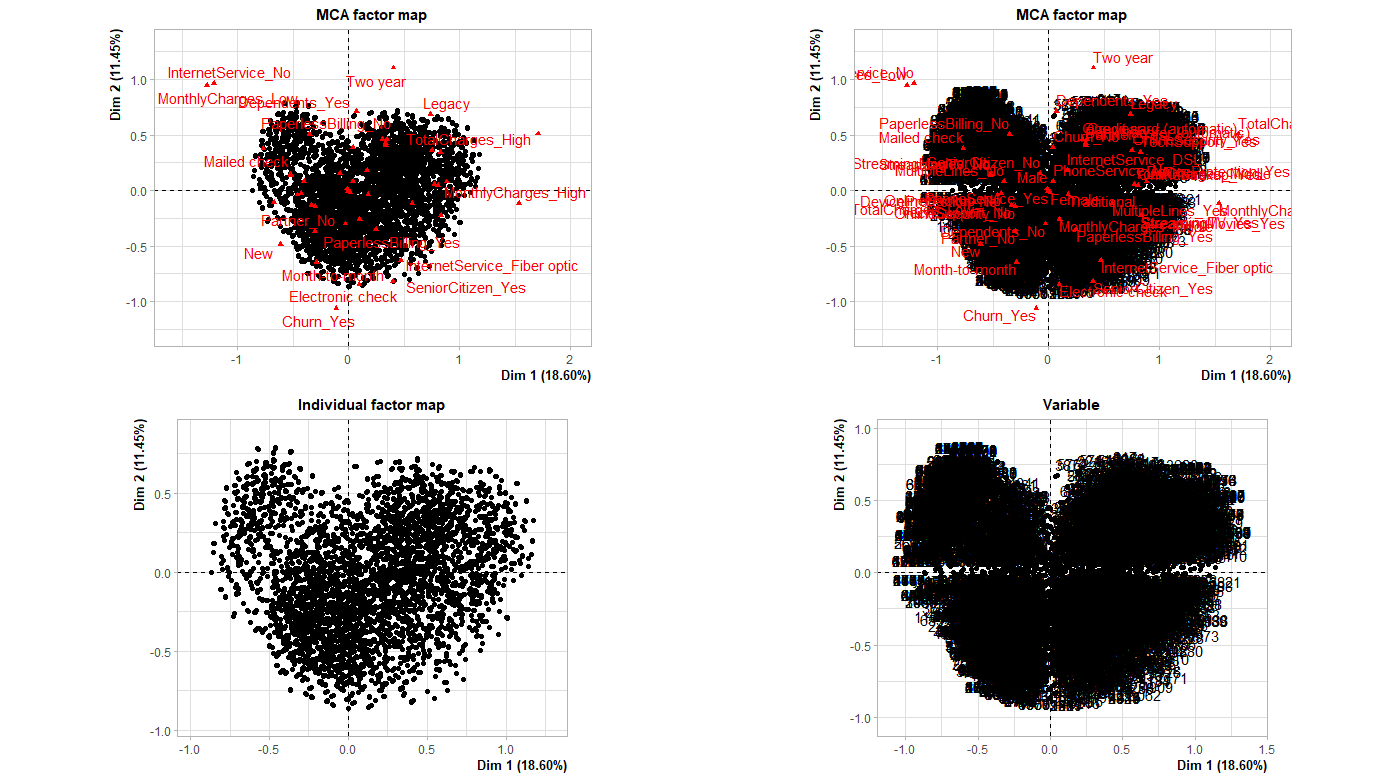
Shows the quality of representation



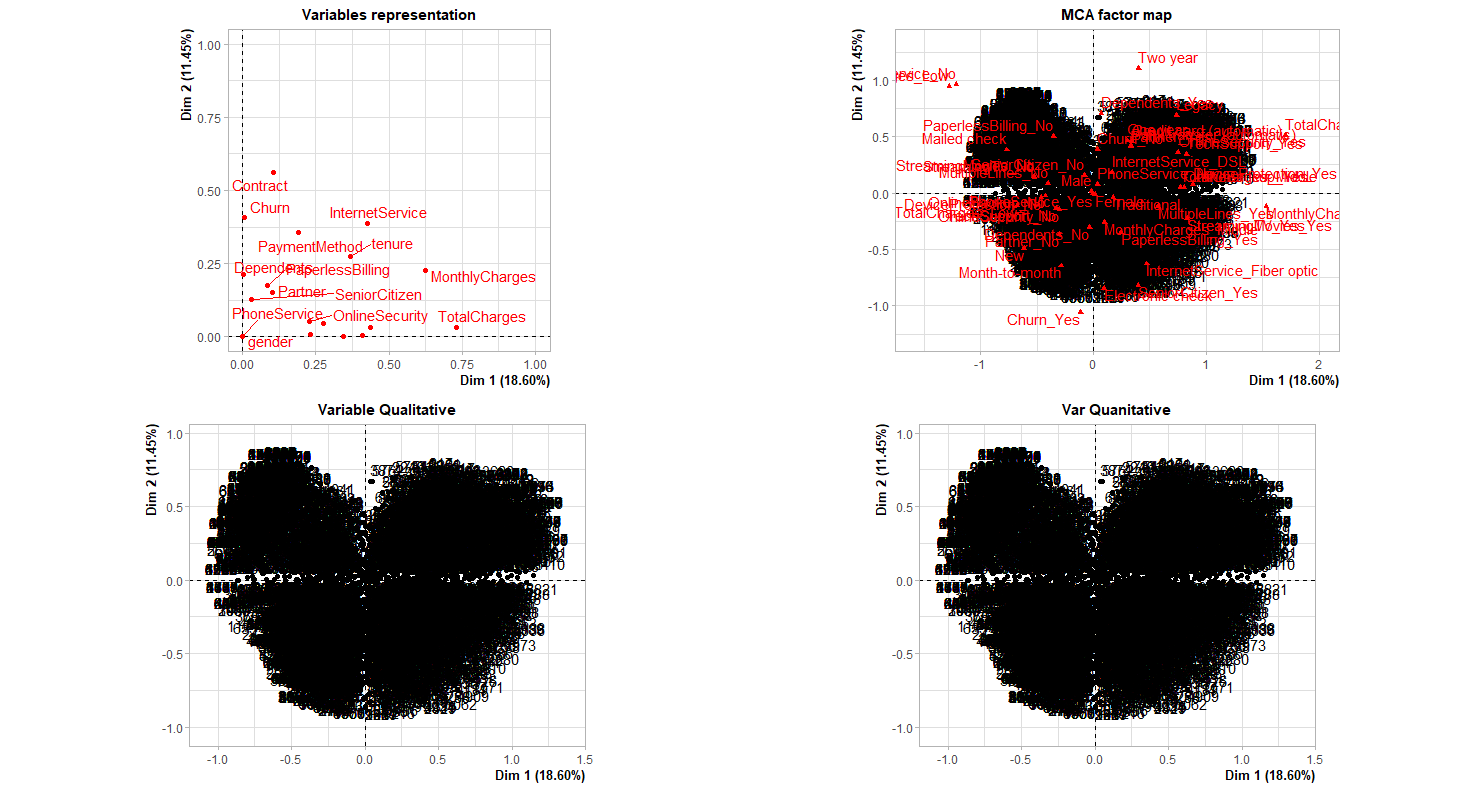
Individual coordinates for the categories

Contributions of the individuals 

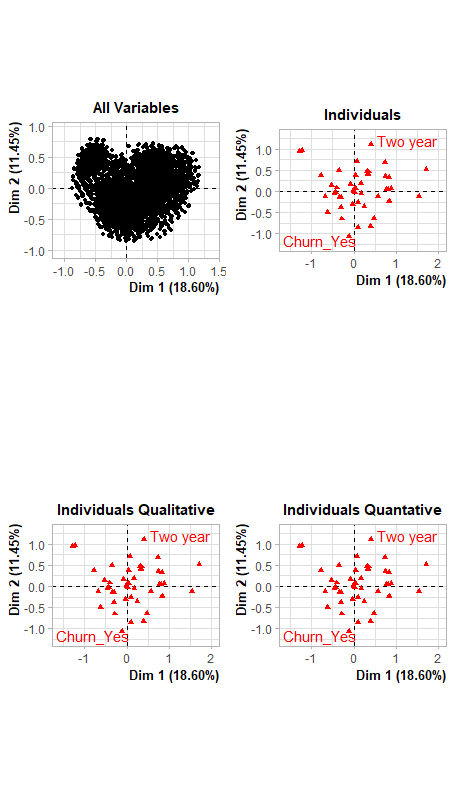
Quality of representation for the individuals.



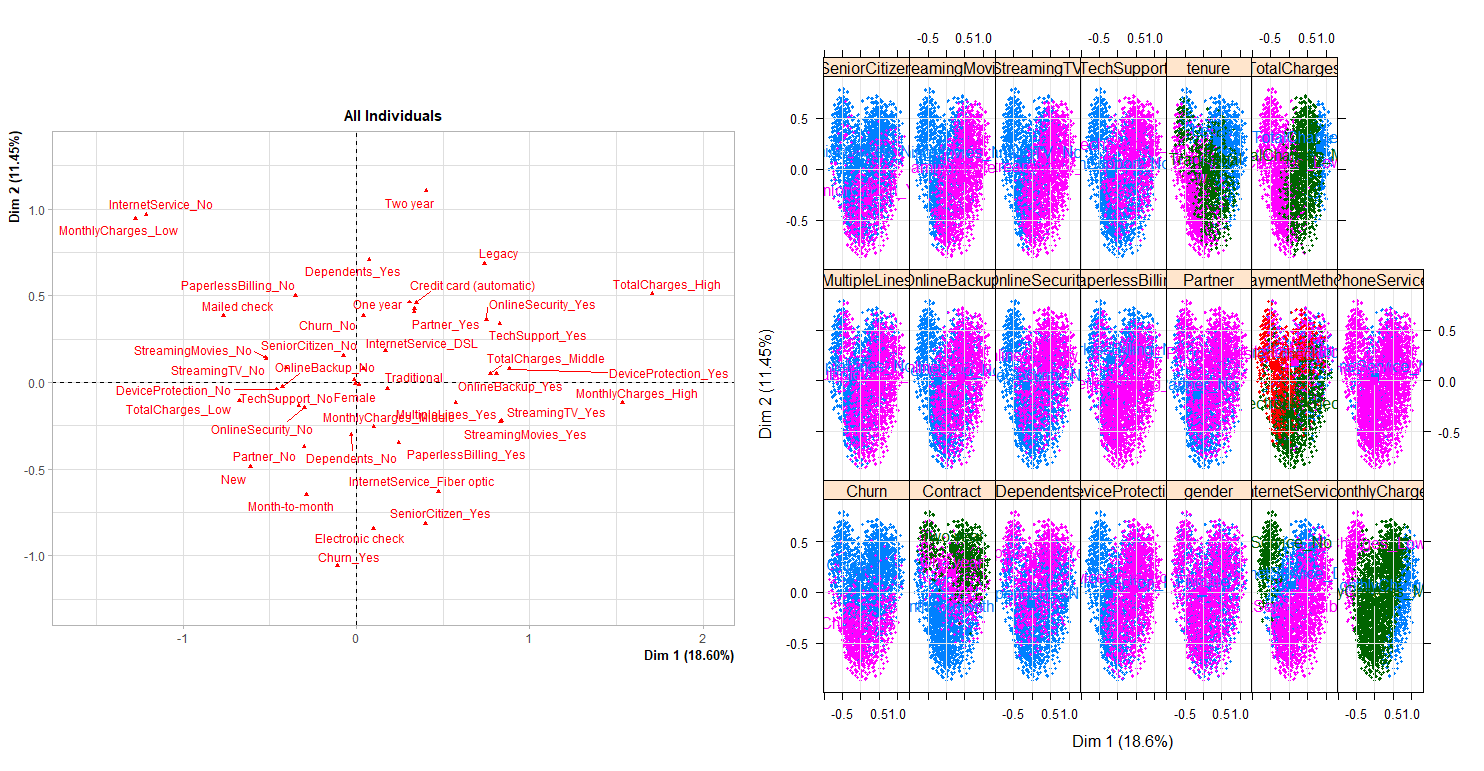
The top left is the res.mca plot showing the individuals and categories. The top left shows the same but using the plot.MCA function labels all categories unlike the plot to its left. The variables are red and the individuals are black. The bottom left shows a graphical option of all the individual factors. The bottom right shows all variables.

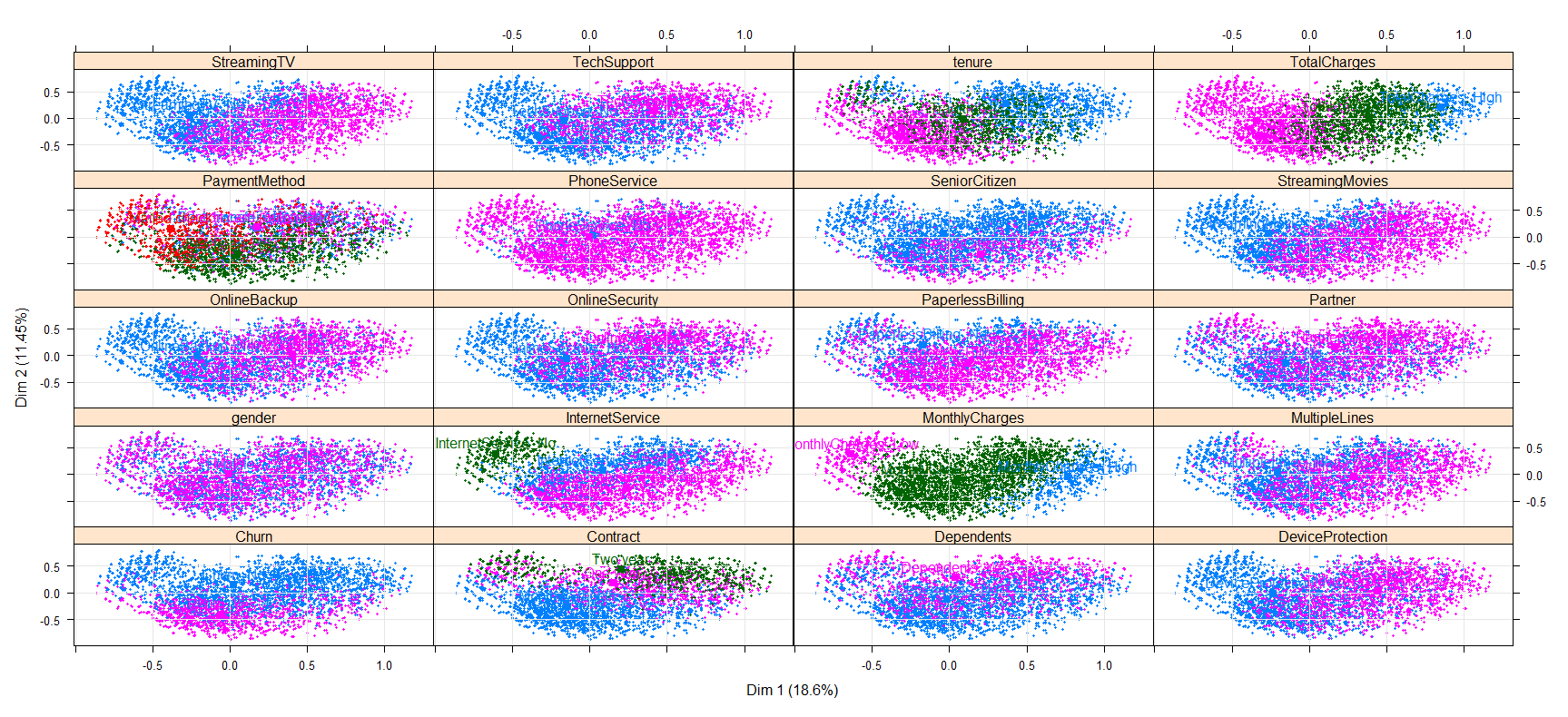


The top left shows the variables in a single quadrant because it’s the variable representation. The top right shows the individuals with the variables. The bottom charts show the quantitative and qualitative supplementary of the variables.



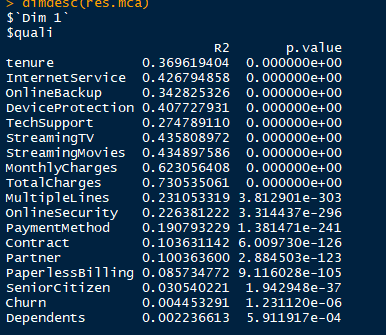
The top left shows only all the variables with quantitative and qualitative variables. The top right shows the correlation of the individual's variables. The bottom charts are the individual variables qualitative and quantitative correlation to the dimensions. It can be seen that monthly charges are lower in these three charts compared to the MCA Factors Map chart.

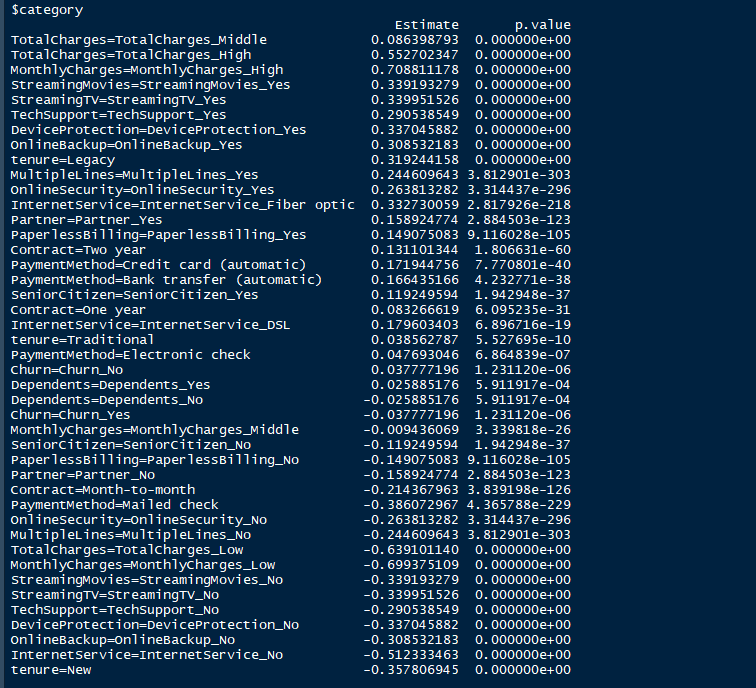


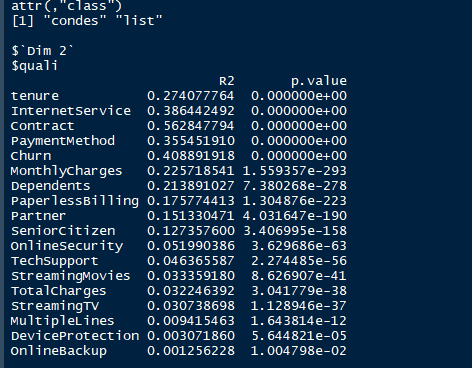


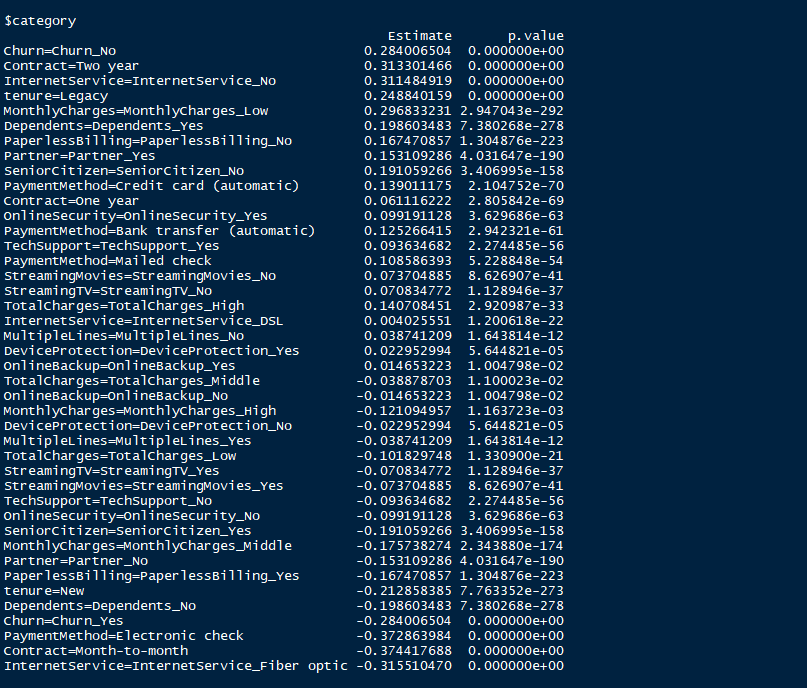
The chart to the left shows all the individual categories graphically to show their correlation to the dimensions. The chart to chart is a breakdown of each category in each variable and how it correlates to the dimensions. The last chart is a zoomed-in version of the screenshot above it on the right but it is hard to see, so we included another version of it. This chart graphically shows the differences between each category in each variable and all different colors.

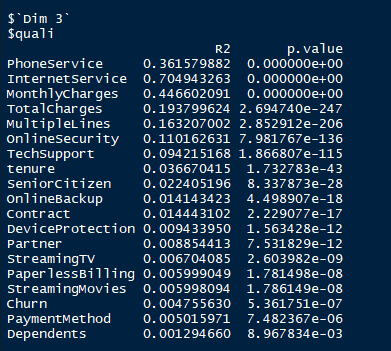
Below are the R2 and p values for each variable and its categories for qualitative and quantitative variables.





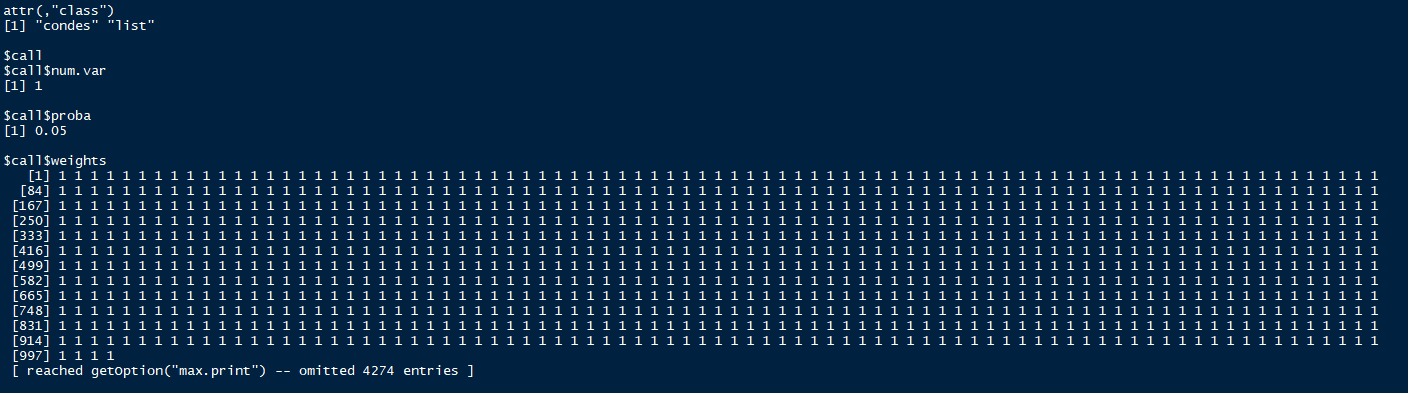


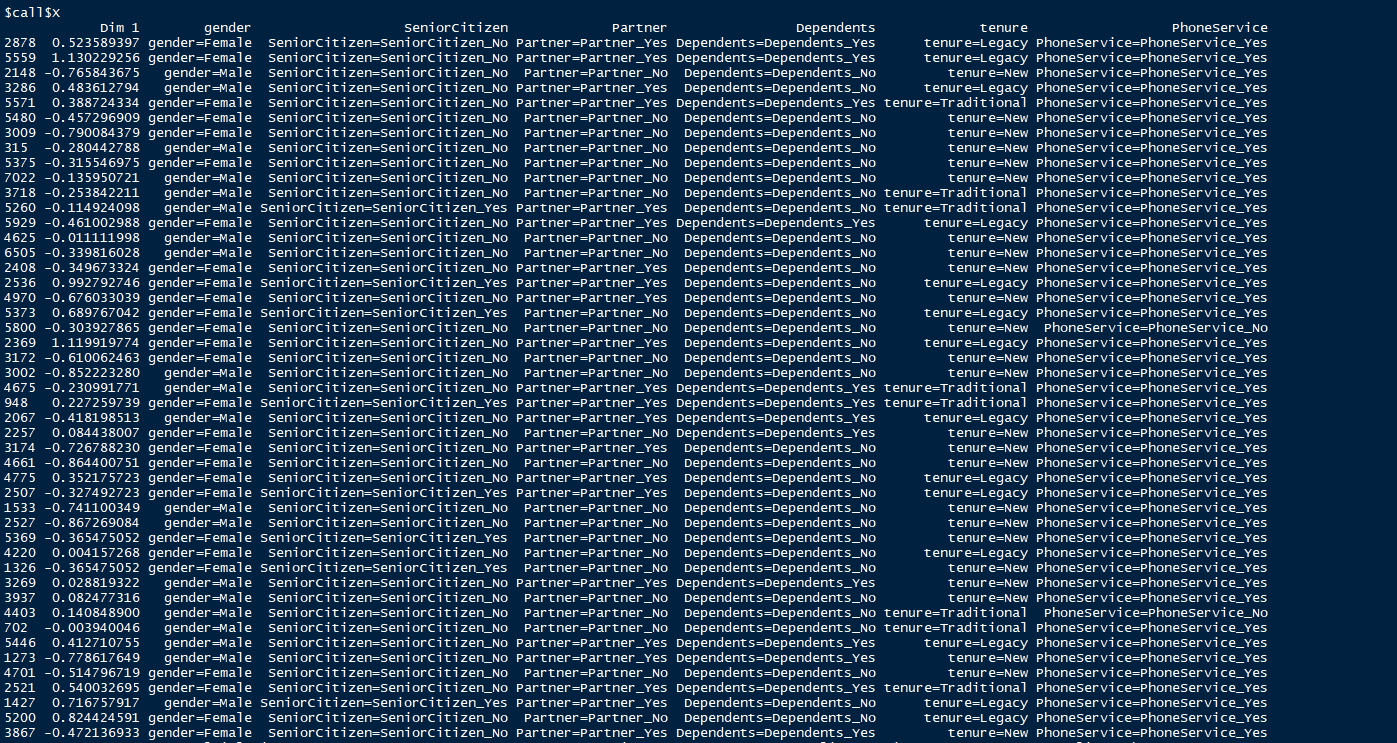


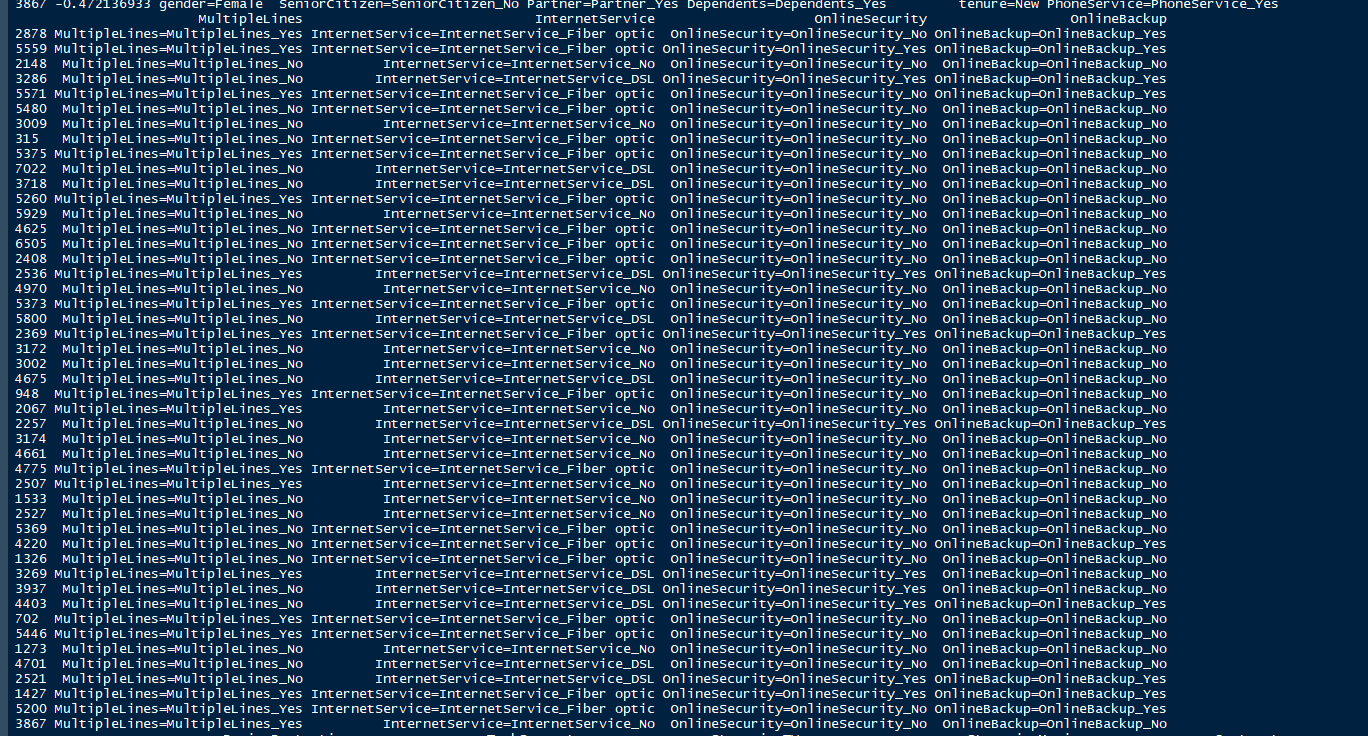


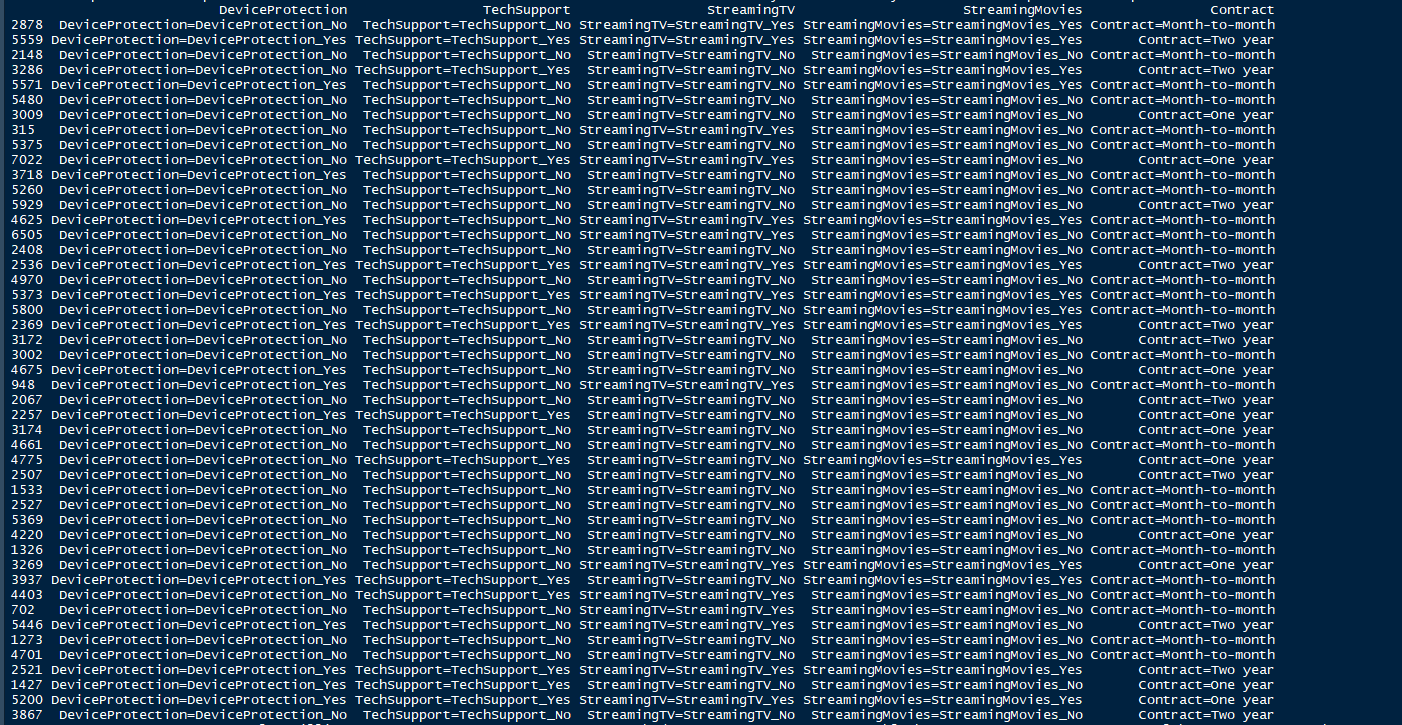


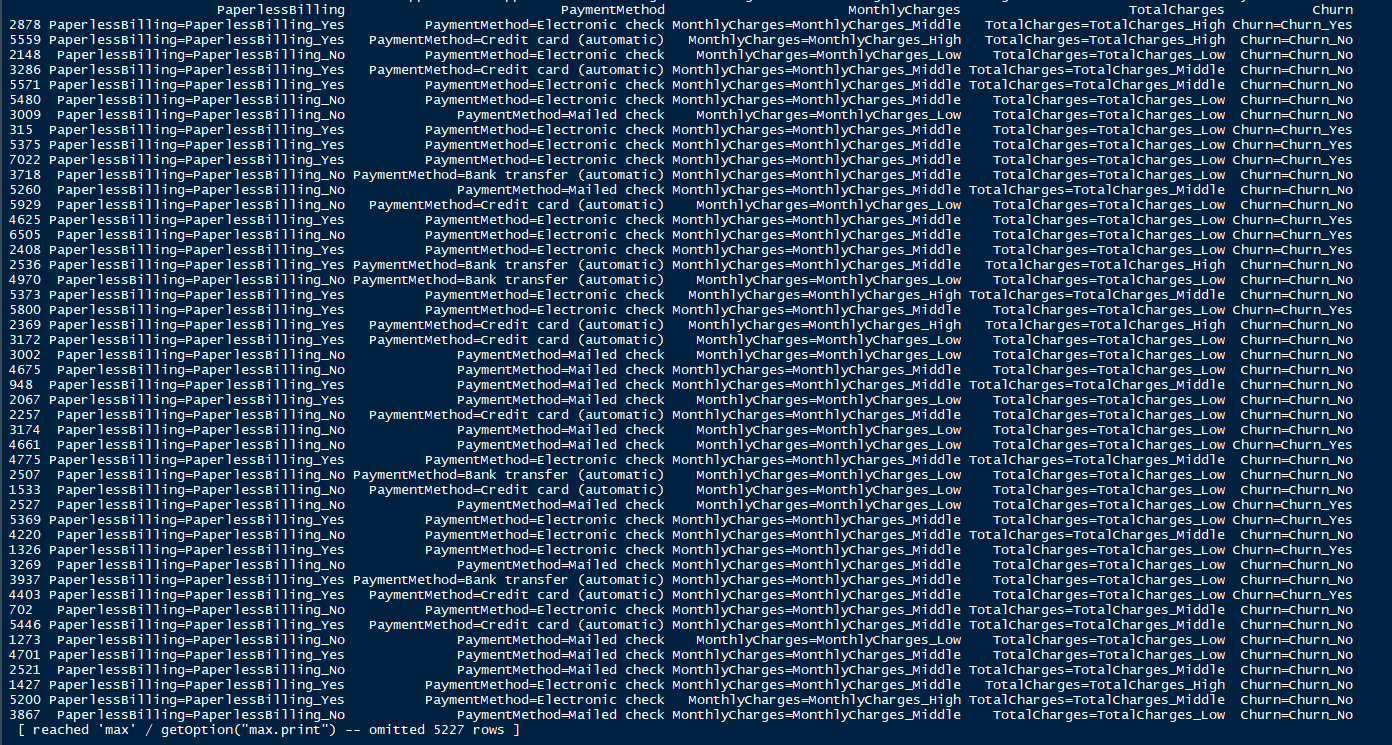
Below shows the weights of specific records.





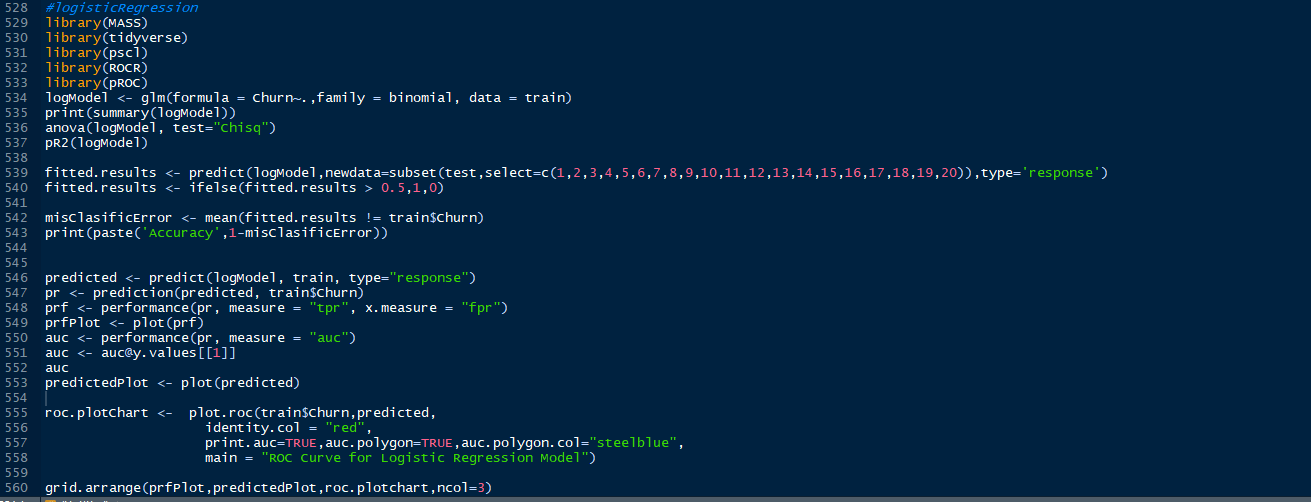


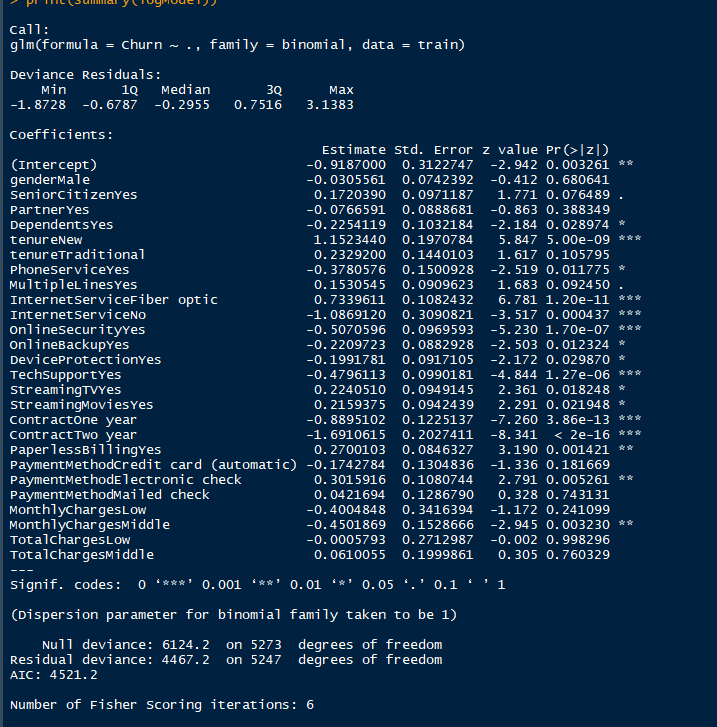




Multiple Correspondence Analysis was chosen because of its ability to handle both qualitative and quantitative data in a single datasheet. Also because in multiple ways you can visualize the categories. With the way to break down the data to quantitative and qualitative supplementary variables. This method can hide the active variables and active individuals to give to show the data is correlated and impacted with just that data. With the extra ability to hide columns, you would like. For instance, you can graphically represent the individual's qualitative variables and hide everything else. How to do this and visualize this is done early in this section. Other methods would have given the ability to use mixed data and see how the variables relate to one another. Clustering could have been used but it would have focused on the clusters, where this allowed the ability to look at clusters and non-clusters to see how they correlate with one another. Also allows the ability to see patterns amongst the data and how the data can be seen in table form as well. As mentioned earlier you can see the correlation amongst factors as wells as the correlation and impact on dimensions. This method helped reduce the factors to be viewed based on the impact and correlation this leads to the findings on this method is how to better reduce the data set and or add data that will be told what the company needs to do to reduce churn.

Now for the nondescriptive method chose to evaluate this data set. The method chosen was logistic regression because of its ability to evaluate multiple variables and its ability to predict. This will be evaluated by charts and tables of results. The data will have the ANOVA statistic, prediction, prediction of fitness, and R2 performed. The screenshot below is all the logistic regression code. With logistic regression, we used pscl and pRPC packages on top of the tidyverse and gridextra packages to perform the evaluation.

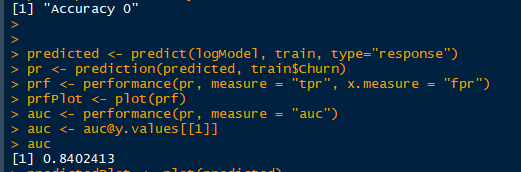




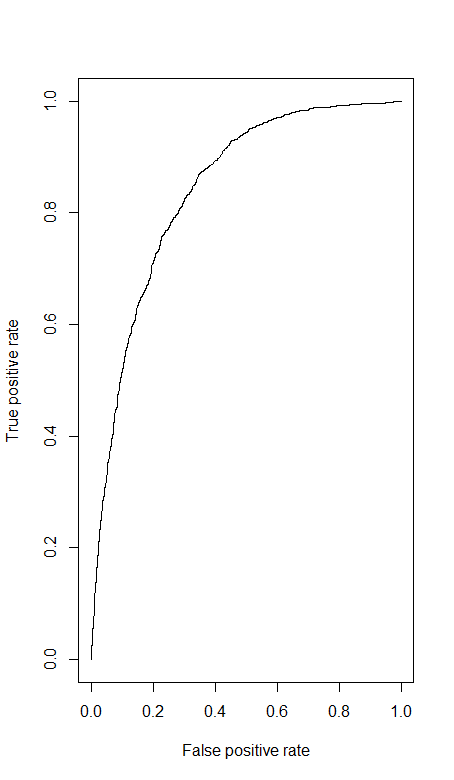
Let’s look at the significances of the coefficients of the categories. Six of the categories have 0.001 significances based on the coefficients table. Four categories have 0.01 has its significances code. Also, four categories have a significance code of 0.05, also two categories have a significances code of 0.1. No categories had 1 has its significances code and eight had zero significances at all. This has an AIC value of 4521.2 and a Fisher’s scoring iteration of 6.



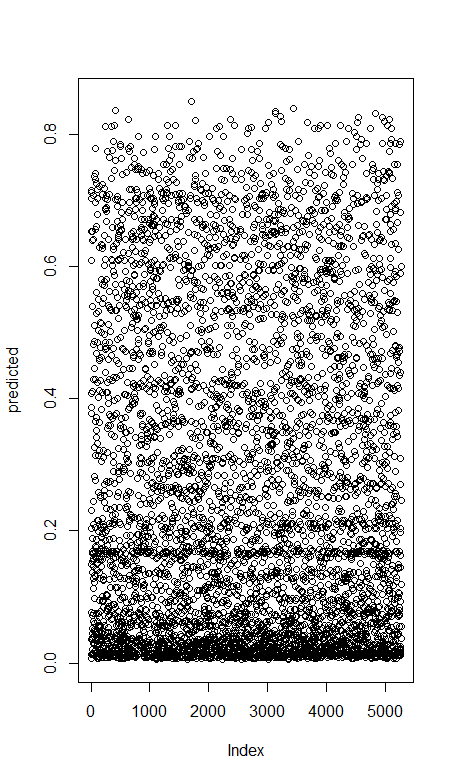
Eleven variables have a 0.001 significances code in the ANOVA Statistic, also four have a code of 0.01. Only one variable monthly charges has a 0.05 significances code which would. Of all the variables only three have zero significance. Those variables are gender, phone services, and total charges. The R2 values are 0.2696232 for the r2ML and r2cu with a value of 0.3925261.

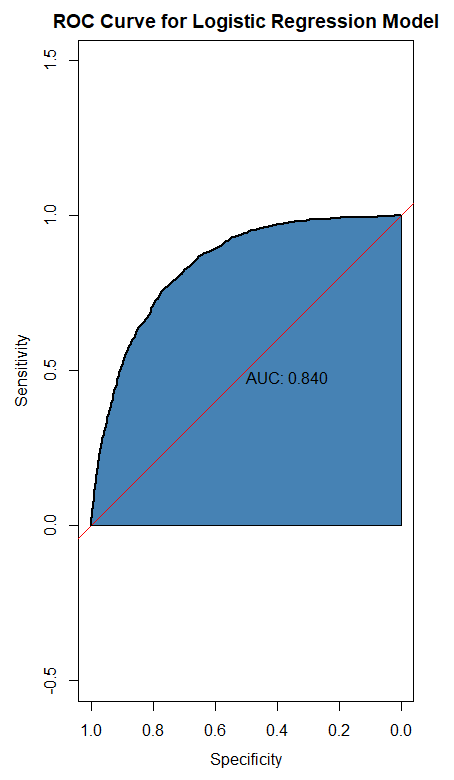


Accuracy being one minus the misClasificError is zero meaning the data is correct. The auc value of this is 0.8402413.



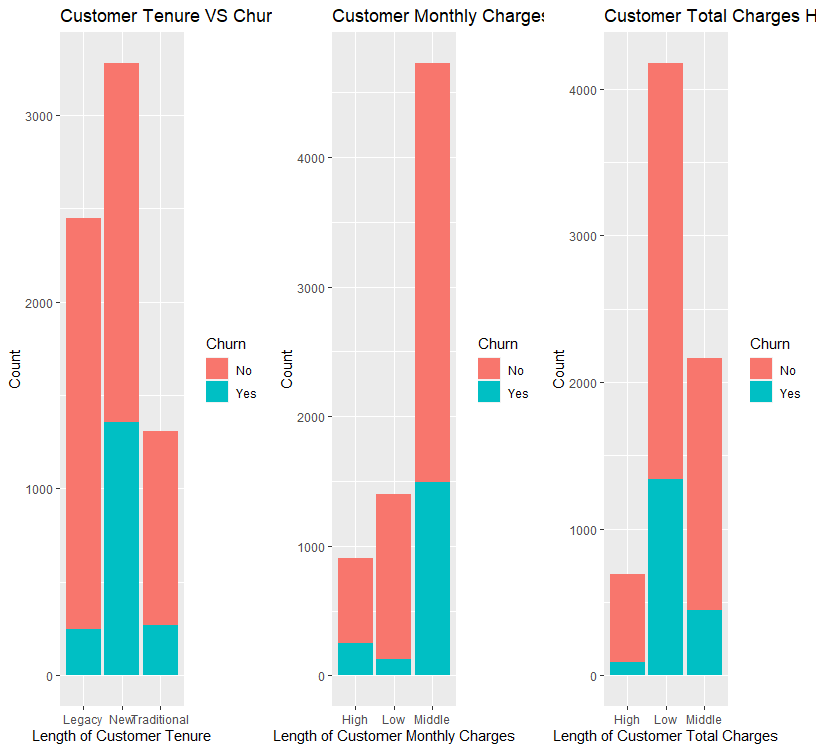
The fitted plot shows a true positive trend with the line curving around the true positive rate of 0.6.

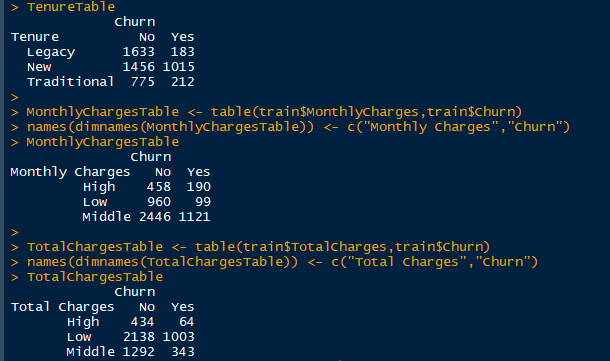
This shows the predicted values of the response and it shows a heavy number of values around the 0.0 predicted and 0 of the index.



Here is the auc graph which shows the trend line going through the regression model with what the area under the curve is. The logistic regression model further helped what variables and an impact on the data set towards the scenario. For example, ANOVA proved the gender variable does not have an impact on churn. This was expected after running the MCA because it was centrally located in dimensions one and two. This was the right model to use because of having it can predict values and trends, which would not be able to be done in other methods.

We have used two different types of data visualization methods used in this report. The two different types are tables or results charts and graphs. Below is an example of a graph used and table use. These were used to compare the tenure, monthly charges, and total charges to the churn variable. The graphical option helps to give a physical visual of what the data says. The table gives the combination of the two variables with one being the column and the other being a row. Both of these help show the variables correlate with one another and how much impact they have on the churn variable. This will help with proving or disproving how the attrition of the churn in customers is happening.





This data does not completely discriminate itself. The data does show how important or unneeded some variables are needed to help with the attrition of the churn. Based on the data provide for the churn it looks those who have churned are those who are new customers. This assumption can be made based on the churn variable vs the total charges and tenure leaning heavily towards the low and new categories respectfully. To help keep these customers from churning month-to-month contract type promotions attract them to stay because they are looking for the “best bang for the buck” to say. The data shows longer-tenured customers are not churning at the moment so offering the same deals these customers has new or newer customers will help to minimize the attrition before it starts. Also not appealing to one gender over another will help in attrition because the logistic regression and the Multiple Correspondence Analysis genders don’t have a correlation or impact on either male or female because they are about equal in the number of times they appear in the data. This helps to answer our goal of what variables help mitigate the churn and which lead to the churn, as stated the total charges and tenure give an example of the variables to help mitigate the churn but an example of having the most effect on the churn rate would be monthly charges the middle category was the highest. To help lower this effect look into a way to lower the monthly charges for the customers. The ANOVA statistic was one of the methods used to detect interactions amongst the variables. The significances codes give an analytical way to support interactions seen plotted in the Multiple Correspondence Analysis. For example, the gender variable and how it correlates with other variables, also looking at the univariate for gender will show male and female are so close in observations as well it wouldn’t show any importance. For selecting the most important predictor was the bivariate and comparing tenure, total charges, and monthly charges based on the charts it shows clearly what categories in the category to focus on. Monthly charges had an ANOVA Statistic significance of 0.5 which is less to show it is significant and in the bivariate charts middle category had the highest churn.

References

Alice, M. (2020, July 5). *How to perform a logistic regression in r: R-bloggers*. R. https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/.

Christopher DuBoisChristopher DuBois 39k2323 gold badges6868 silver badges9191 bronze badges, David LeBauerDavid LeBauer 29.1k2727 gold badges106106 silver badges180180 bronze badges, Claus WilkeClaus Wilke 14.4k77 gold badges4646 silver badges8484 bronze badges, DeenaDeena 4, Dirk EddelbuettelDirk Eddelbuettel 336k5353 gold badges601601 silver badges682682 bronze badges, ThierryThierry 17.2k55 gold badges4343 silver badges6565 bronze badges, TungTung 21k66 gold badges7171 silver badges8787 bronze badges, baptistebaptiste 71.7k1515 gold badges183183 silver badges270270 bronze badges, Jeromy AnglimJeromy Anglim 31.1k2424 gold badges110110 silver badges167167 bronze badges, shinyshiny 2, IVIMIVIM 1, Mayank AgrawalMayank Agrawal 2744 bronze badges, & timtim 3. (1958, January 1). *Side-by-side plots with ggplot2*. Stack Overflow. https://stackoverflow.com/questions/1249548/side-by-side-plots-with-ggplot2.

Houee-Bigot, M., & Husson, F. (n.d.). Multiple Correspondence Analysis(MCA) with FactoMineR (hobbies dataset). Husson Github.

Kassambara, & Visitor. (2017, August 17). *MCA in R USING FACTOMINER: Quick scripts and videos*. STHDA. http://www.sthda.com/english/articles/22-principal-component-methods-videos/71-mca-in-r-using-factominer-quick-scripts-and-videos/.

Kassambara, Kimosh, Hibernatus, Visitor, Anjos, M. C., Ding, Nira, Kassambara, & Kim, S.-K. (2017, September 24). *MCA - multiple Correspondence analysis in R: Essentials*. STHDA. http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/.

Li, S. (2017, November 20). *Predict customer Churn – logistic Regression, decision tree and Random forest: R-bloggers*. R. https://www.r-bloggers.com/2017/11/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/.

*MCA: Multiple Correspondence Analysis (MCA)*. RDocumentation. (n.d.). https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/MCA.

Prabhakaran, S. (n.d.). *Logistic regression*. Logistic Regression With R. http://r-statistics.co/Logistic-Regression-With-R.html.

Prabhakaran, S. (n.d.). *Logistic regression*. Logistic Regression With R. http://r-statistics.co/Logistic-Regression-With-R.html.