Customer Retention

Christopher Flowers

College of Information Technology, Western Governors University

D208 Predictive Modeling Task 1

Dr. Straw

October 7, 2021

**Part I: Research Question**

A

1. Can bandwidth use determine whether a customer churns?

2. Does age affect customer churn?

Is a customer’s streaming use affected by the churn?

Do technically inclined customers to churn?

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.

**Part II: Method Justification**

B.

1.

* Has a linear relationship
* Has Multivariate normality
* Does not have multicollinearity
* Does not have auto-correlation
* Is homoscedasticity

2.

For this course, I have chosen to use the R programming language. Using R has benefits over other programming languages such as python. One of the benefits of using it is that it has a built-in graphical library, statistical analysis, and the ability to clean the data. It also can install and use additional libraries. R can’t just run statistical analysis but the ability to run a deep statistical analysis with fewer lines of code than other languages. Additionally, R has some shortcuts for traditional programming concepts like its shortcut for if-else.

3.

Multiple regression is the appropriate technique to use to analyze this research question because it will show which variables have a linear relationship. Also, it will show how correlated each variable is to another. By using this technique it allows us to show how the multivariate distribution is amongst the variables. Based on the regression we will be able to predict if bandwidth usage affects which customers churn. By doing this we can try to reduce the number of customers who churn.

**Part III: Data Preparation**

C.

1.

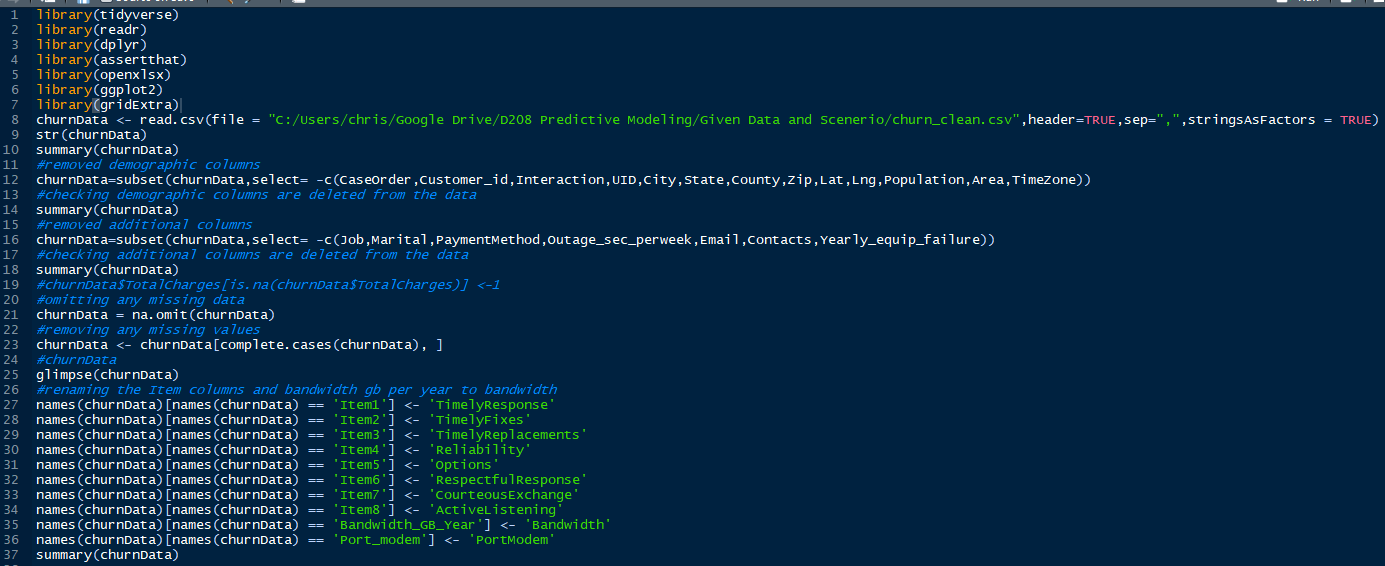
The data preparation goals are to load the data set into a variable then clean and narrow the data set. This will be done by reading in the CSV file to import the datasheet into R. At this point, the data can continue to be prepared and can be manipulated as well. But continuing with the preparation we will summarize the data to look for missing data and what data columns are relative to the research question. Then remove the unnecessary columns and change any data types needed. This will help in welding the best results of the statistical methods. Also to have all variables in the correct data types needed. The goal of data manipulation is to see how the variables are distributed and compare to the dependent variable. Also in manipulating the data we will split it into two data sets to give a better sample size of the data. The showing of the bivariate and univariate data will also help manipulate the data.

2.

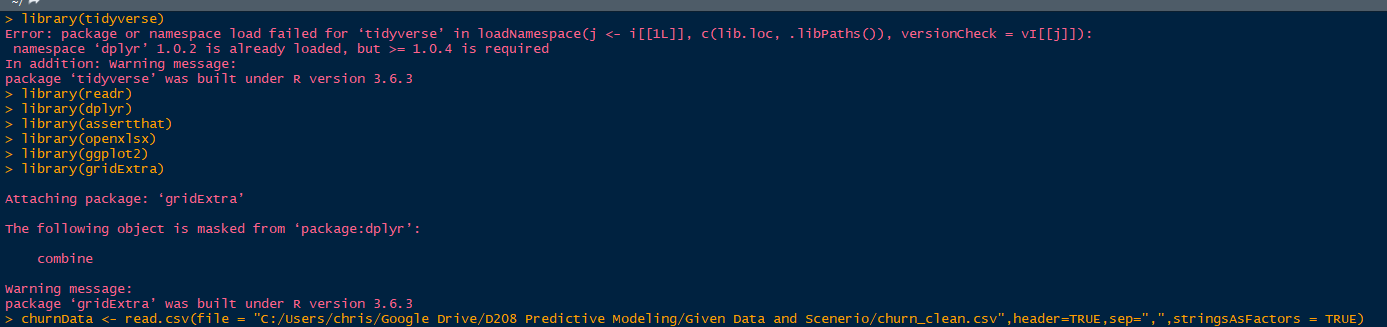
This dataset has several continuous independent variables but not all of these variables are needed for the multiple linear regression. But the variables needed are Children, Age, Income, Tenure, Monthly Charge, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8. All these variables are different than the other variables except for the dependent variable will also be continuous. These values majority of the time will be different from each other. There are several categorical independent variables which most of which containing only two categories, which are either ‘Yes’ or ‘No’. These variables are Gender, Churn, Techie, Contract, Port\_modem, Tablet, InternetServices, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling. The last continuous variable is going to be the continuous dependent variable is the Bandwidth\_GB\_Year and to use this to determine whether Bandwidth usage will contribute to the churn of customers. Yes, there is an identifier for all records because there are 10,000 records in the dataset and there are no NA’s in the dataset as seen in the summary. Whether the bandwidth usage in a year will play an effect on whether the customer will Churn and have other factors contribute to that or do not contribute.

3.

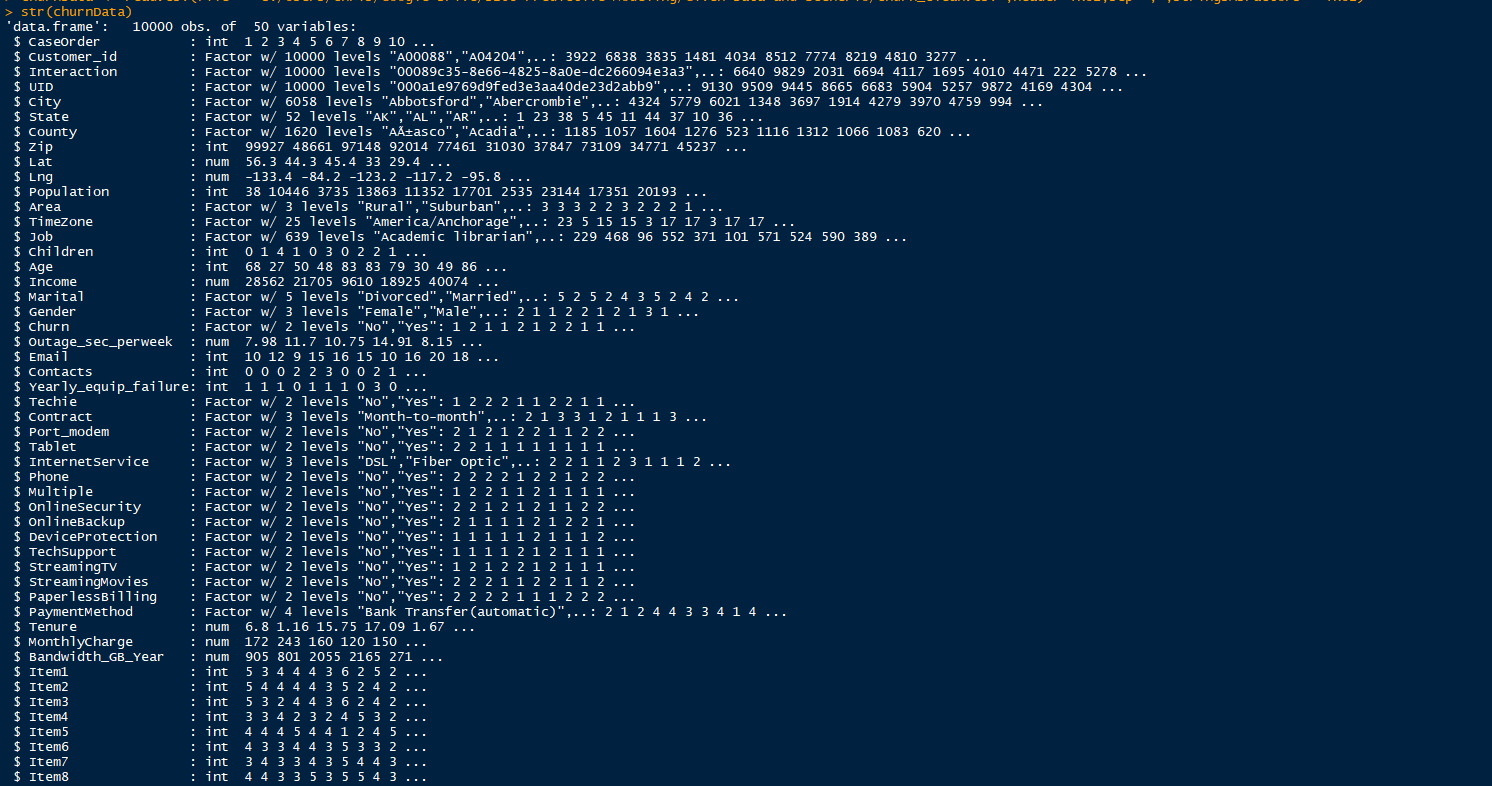
To prepare this data for analysis the first thing to do is to load the library used such as ggplot2, openxlsx, assertthat, dplyer, readr, and tidyverse. These libraries give the use of functions to be able to do certain things like openxlsx allows for the creation of an excel file. After the libraries are loaded the next thing to do is to read the excel file into R and place it into a variable, which can be manipulated, in this case, the variable is called churnData. Next is to use the str() to give us an idea of the data and types present in the data. Now let’s take a look at the summary for the churnData dataset to give a total of each category for categorical data and continuous data will have minimums and maximums for each variable. Next, the thing to happen will be to remove any unneeded variables from the dataset. Most removed variables are going to be those in which are considered demographic data like City and County. But some other variables are not needed that will need to be removed as well to help reduce outliers in the data. After removing the unneeded columns, the next thing needed is to remove any N/A observations or other missing data. This is done by using the na.omit function and the complete.case function followed by using the glimpse function to take a look at the data to see if any data was removed by looking at the total row count. For readability reasons, we change the Item numbered columns to named columns. Also changed any columns with an underscore to something without the underscore like Bandwidth\_GB\_Year and Port\_Modem will change respectfully to Bandwidth and PortModem. Now that variables have been removed and or renamed, we can alter the categorical data variables to continuous data variables. For example, Gender will be changed from Female, Male, and Nonbinary to one for Male, two for Female, and three for nonbinary. After each variable check each variable by running a summary on it to make sure the attended result is achieved. The last thing done is to split the data into two different datasets test and train with the train being the larger set. This is to get a better sample size than the original sample size and then export all new datasets to excel.



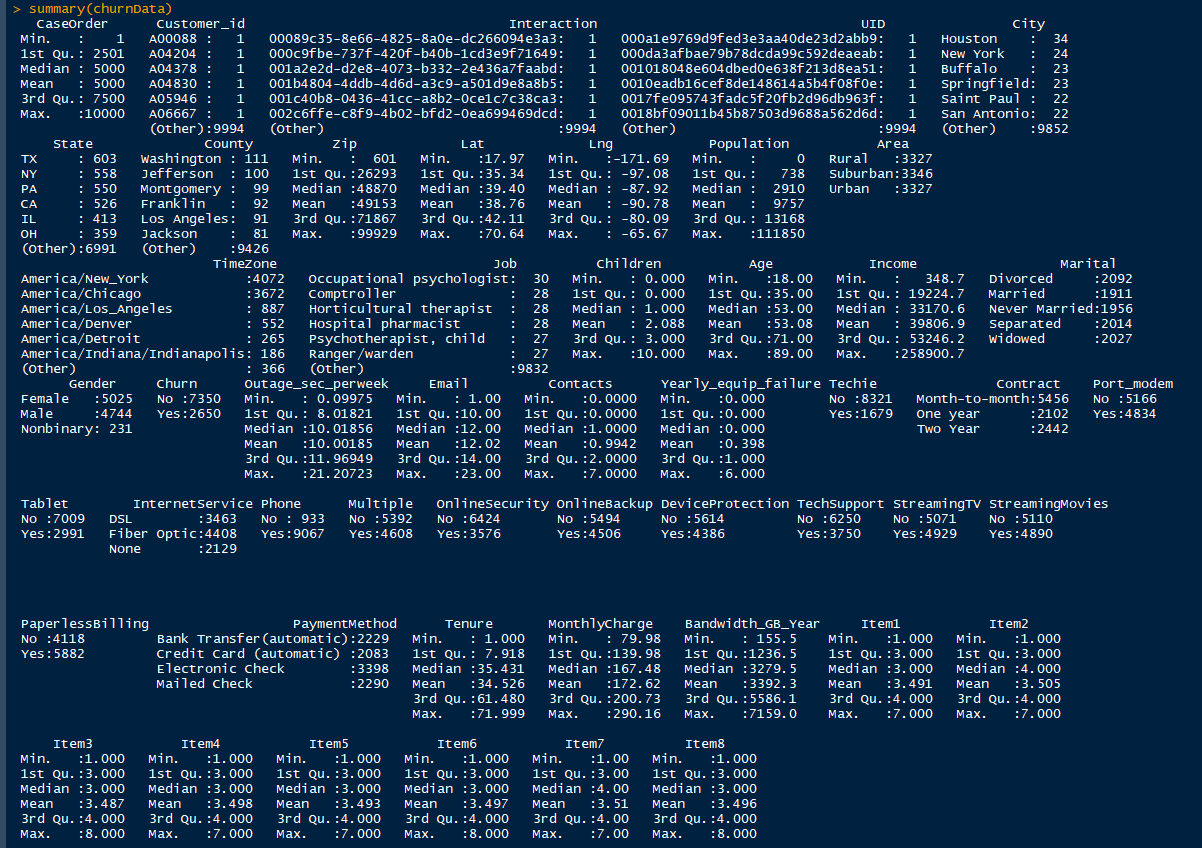
Above is the loading of the libraries, reading the CSV file, running the str and summary function. Also removing the variables that are not needed then summarizing the data to check these variables have been removed. Checking for missing data then checking the row count and renaming some variables then checking the renaming.



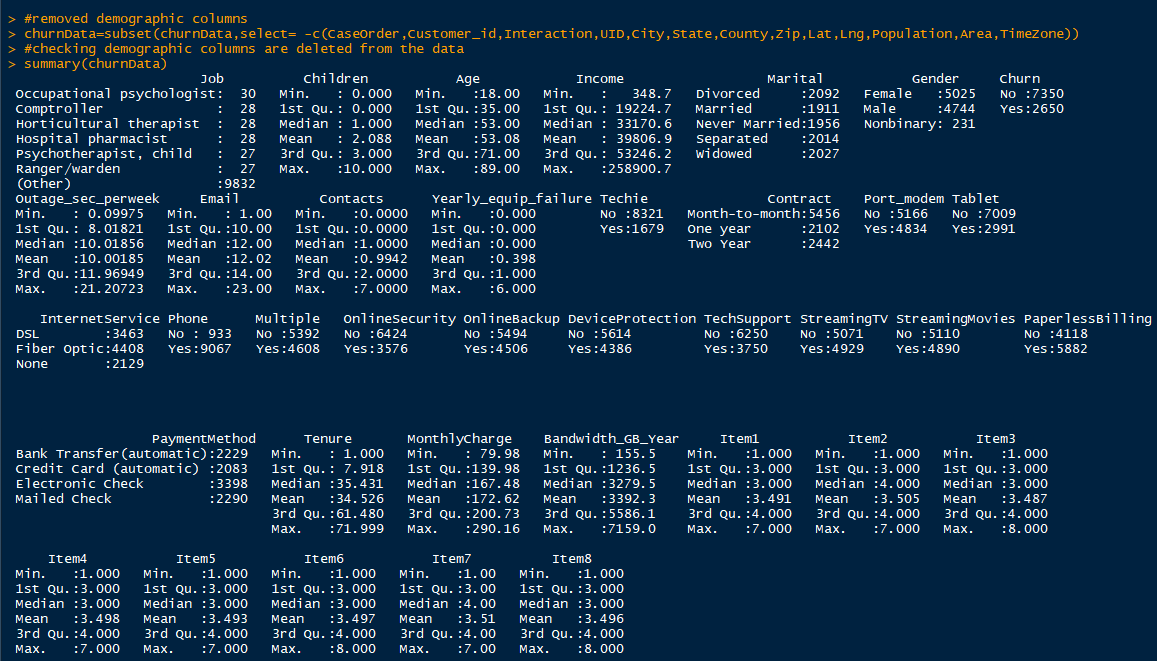
Results of loading libraries and reading the CSV file in.



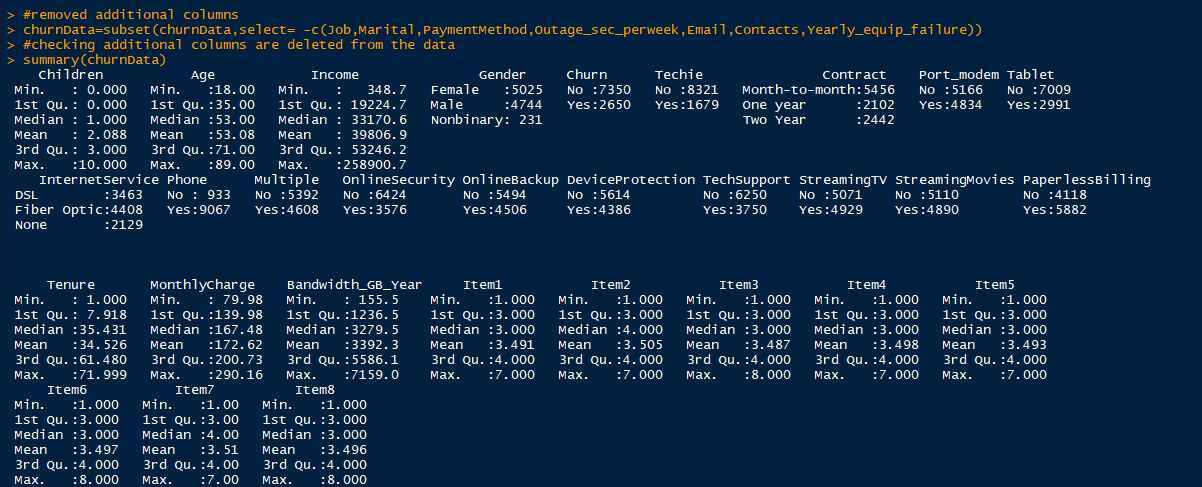
Results of the Str function.



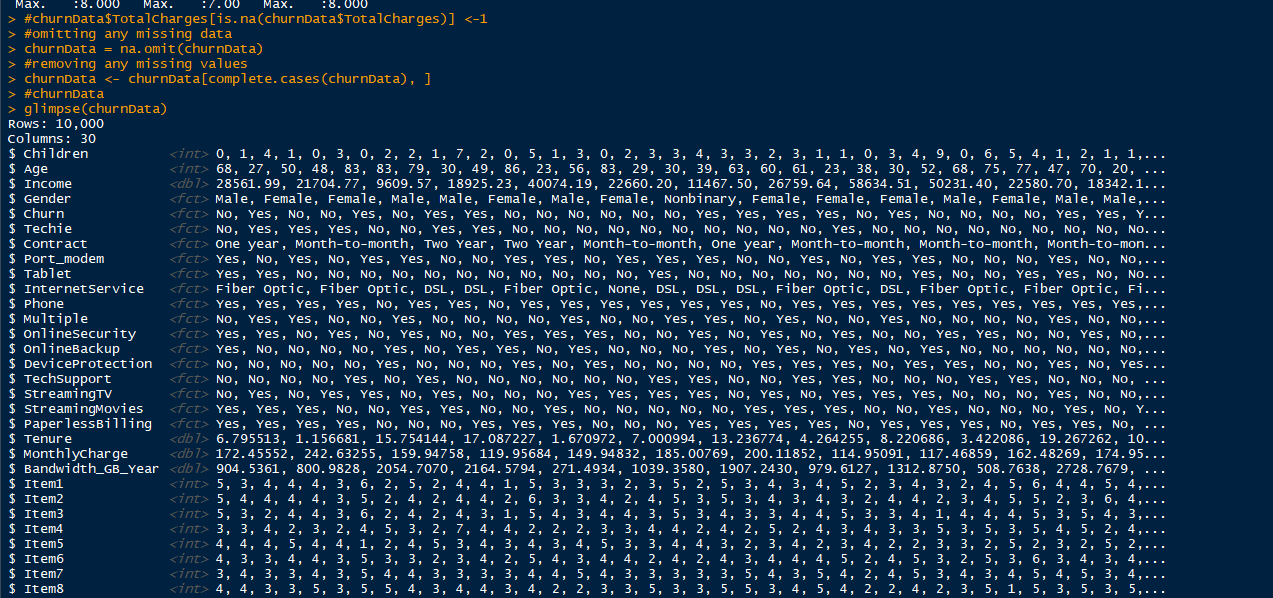
Results of the summary of the churnData datasets.



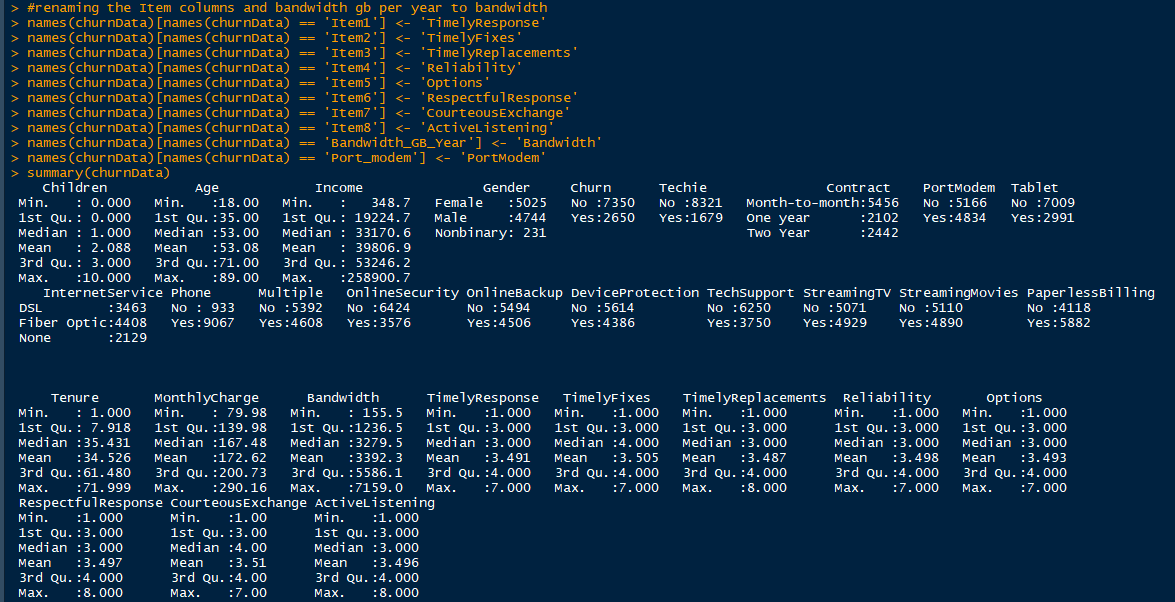
Summary of the churnData datasets after removing the demographic variables.



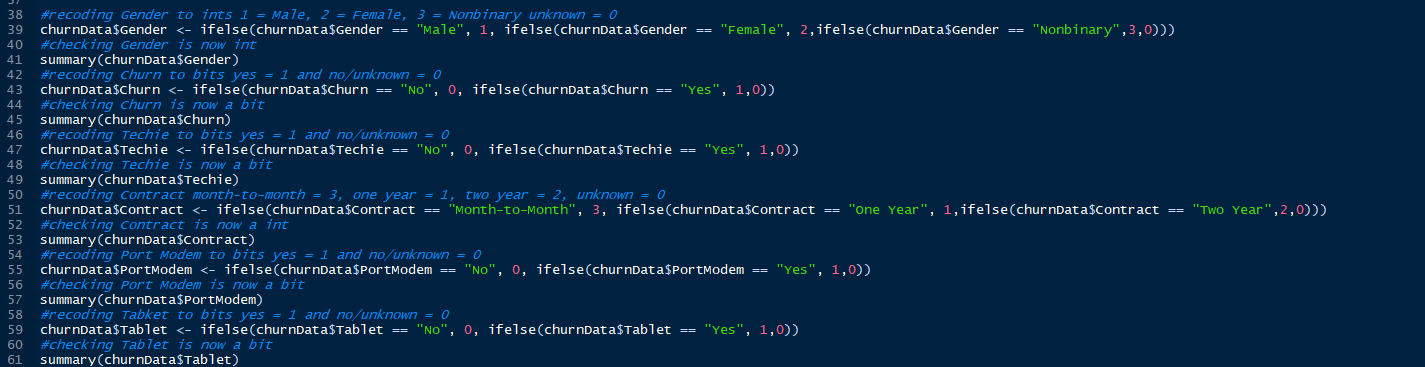
Summary of churnData dataset after removing the other variables.



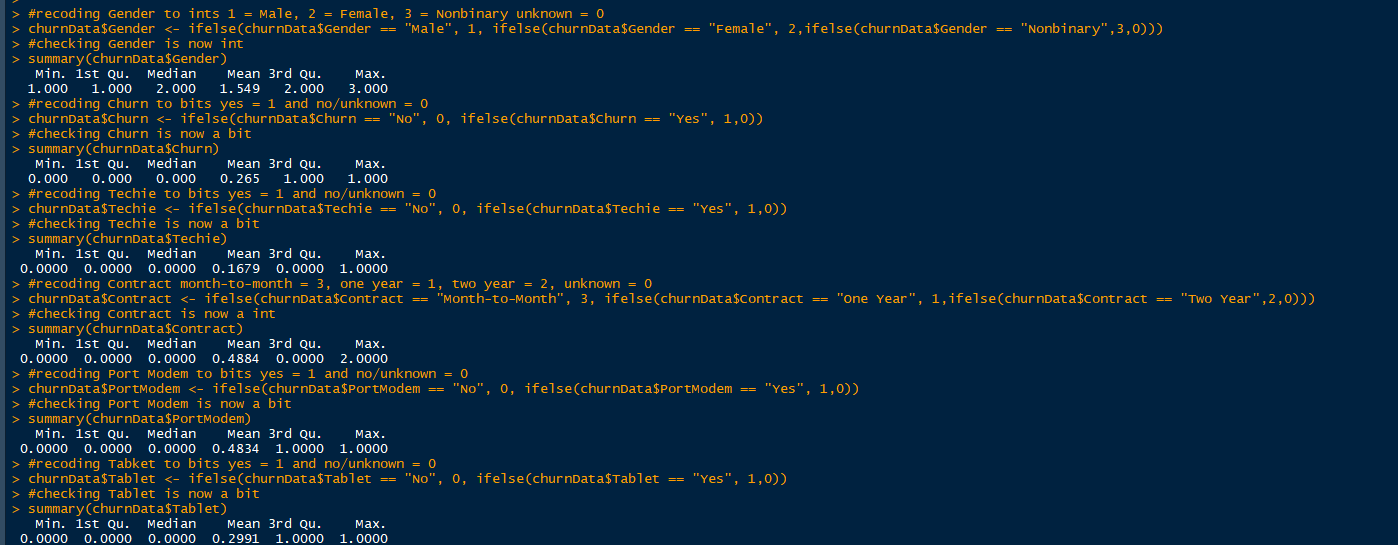
Results of checking for missing data or N/A observations.



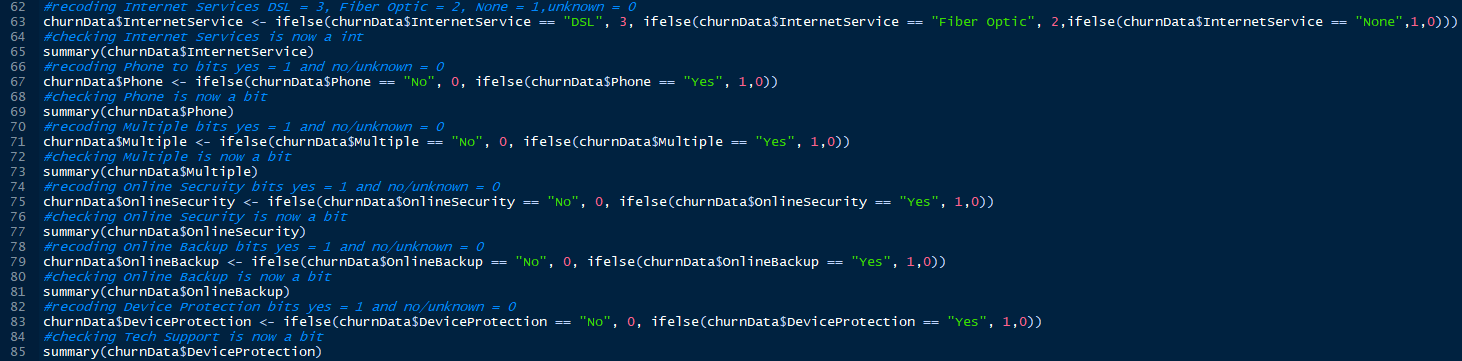
Results of renaming variables



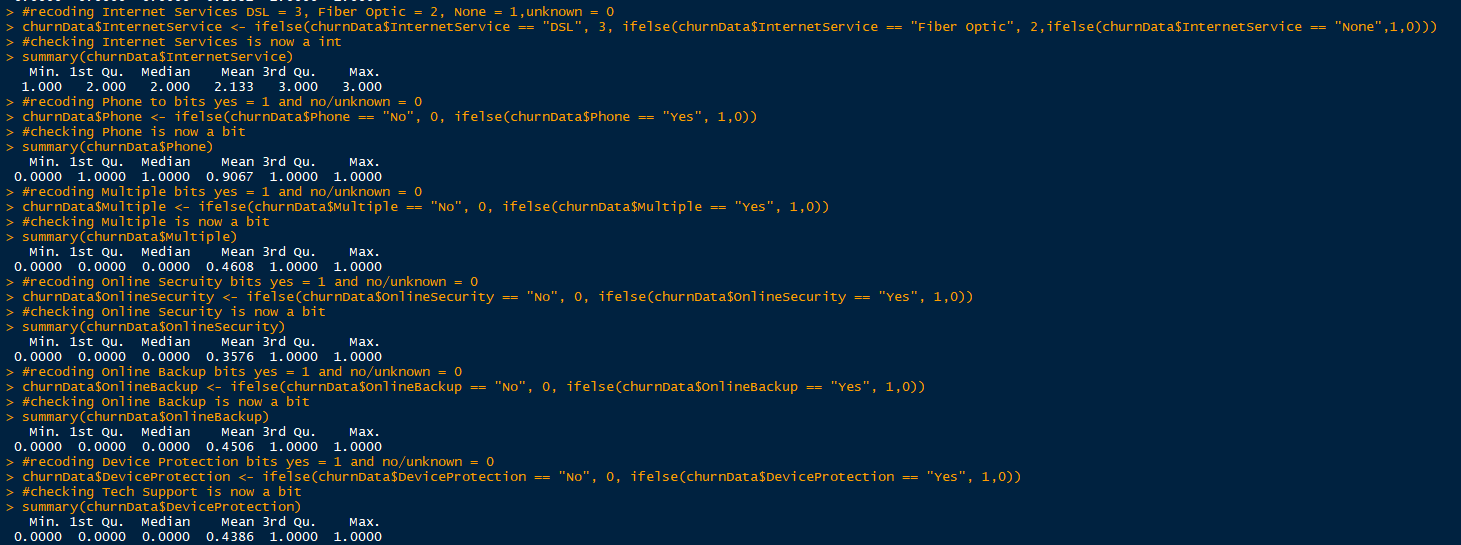
Changing the data to continuous Gender, Churn, Techie, Contract, PortModem, and Tablet.



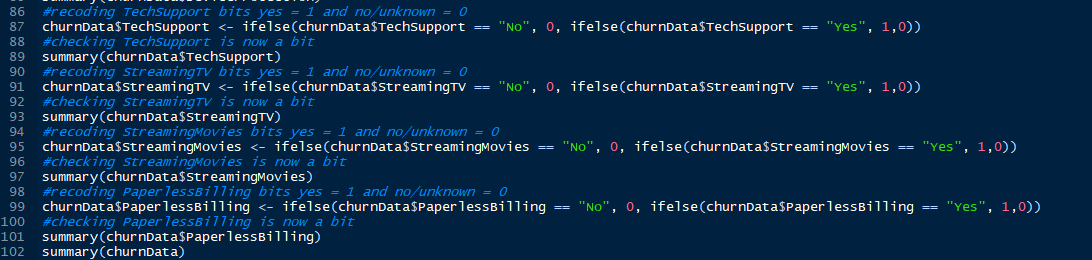
Results of changing Gender, Churn, Techie, Contract, PortModem, and Tablet to continuous variables with the minimum and maximum.



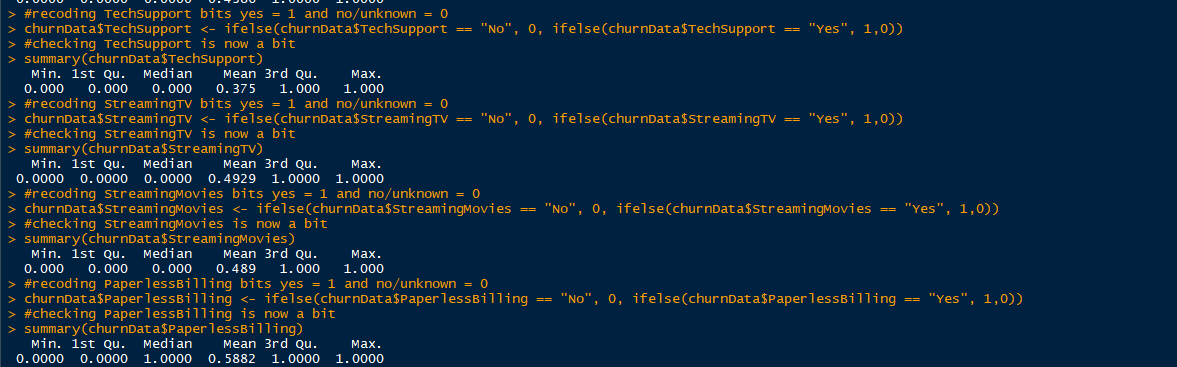
Changing the data to continuous InternetServices, Phone, Multiple, OnlineSecurtiy, OnlineBackup, and DeviceProtection.



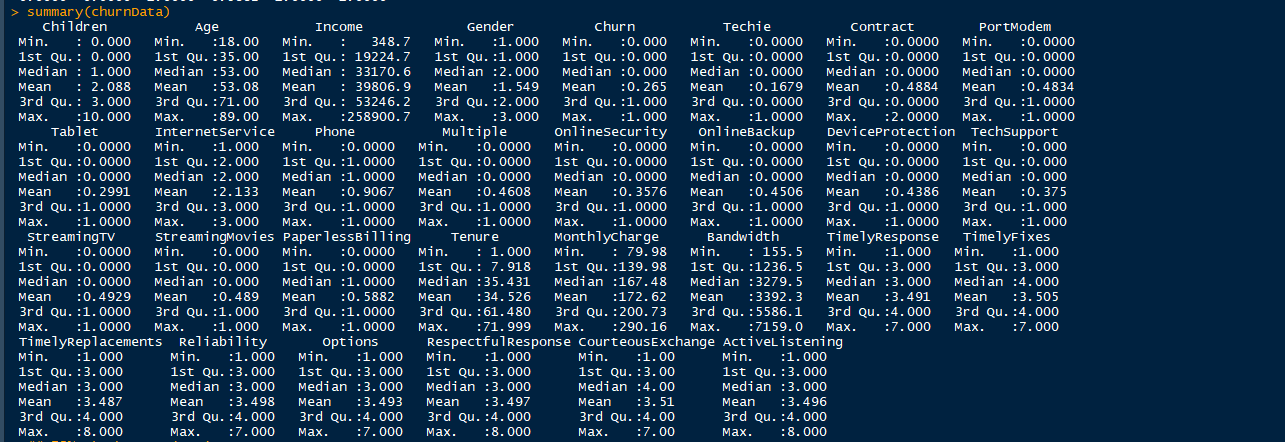
Results of changing InternetServices, Phone, Multiple, OnlineSecurtiy, OnlineBackup, and DeviceProtection to continuous variables with the minimum and maximum.



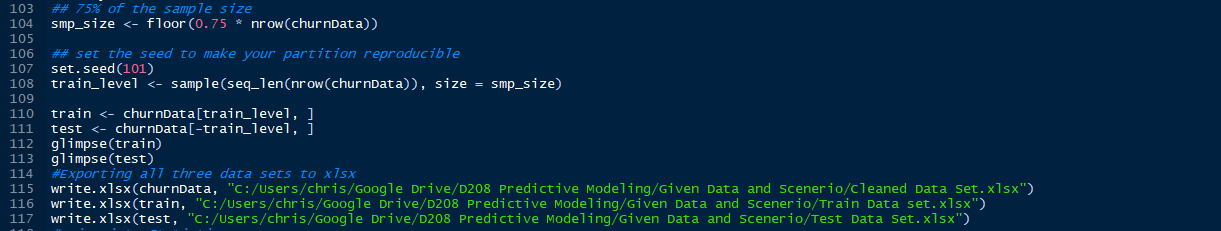
Changing the data to continuous TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling.



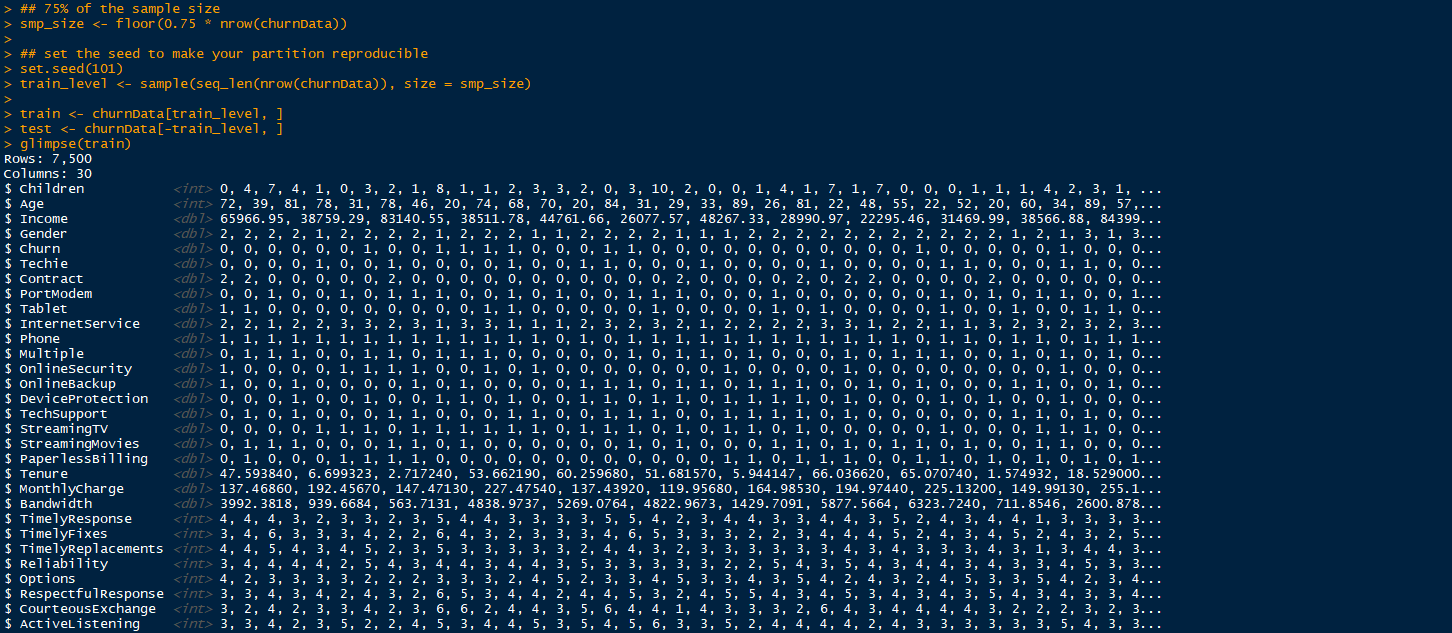
Results of changing TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling to continuous variables with the minimum and maximum.

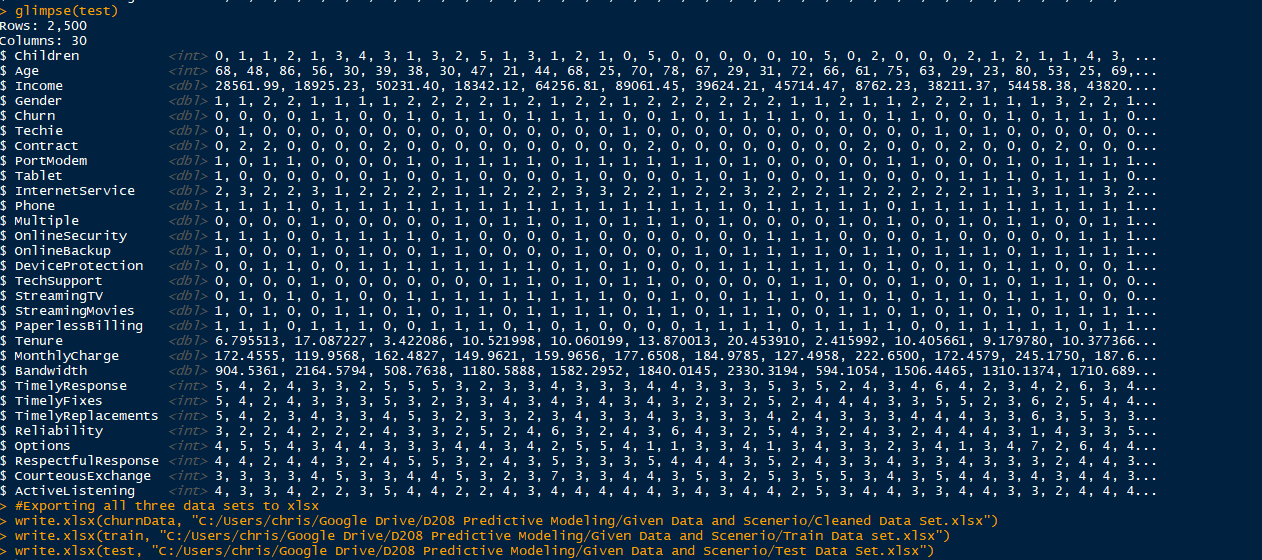


Summary of the dataset after most of the cleaning



Splitting the dataset into two and looking at the two different datasets then exporting all datasets to excel.

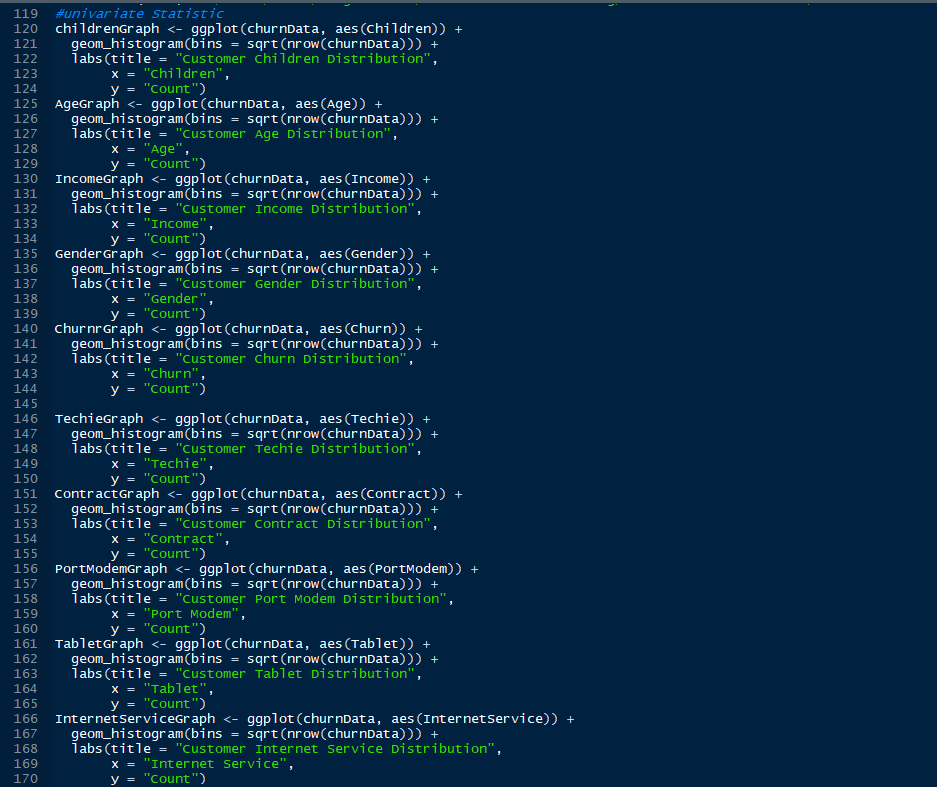
Results of splitting the dataset and the results of the glimpse for the train.

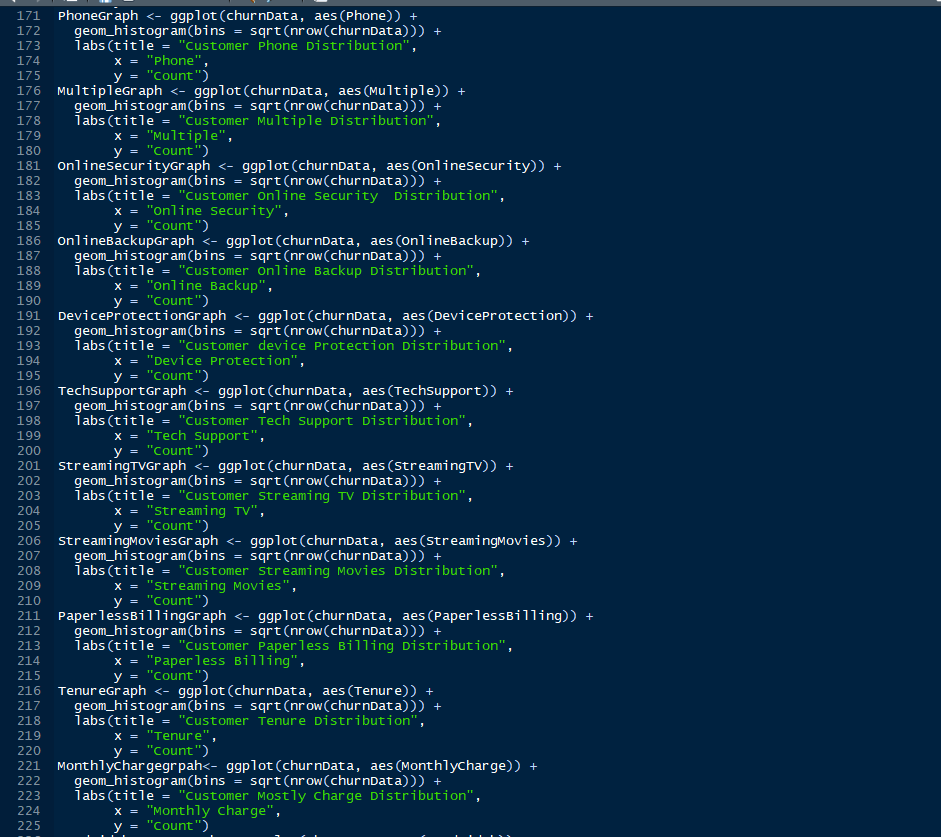


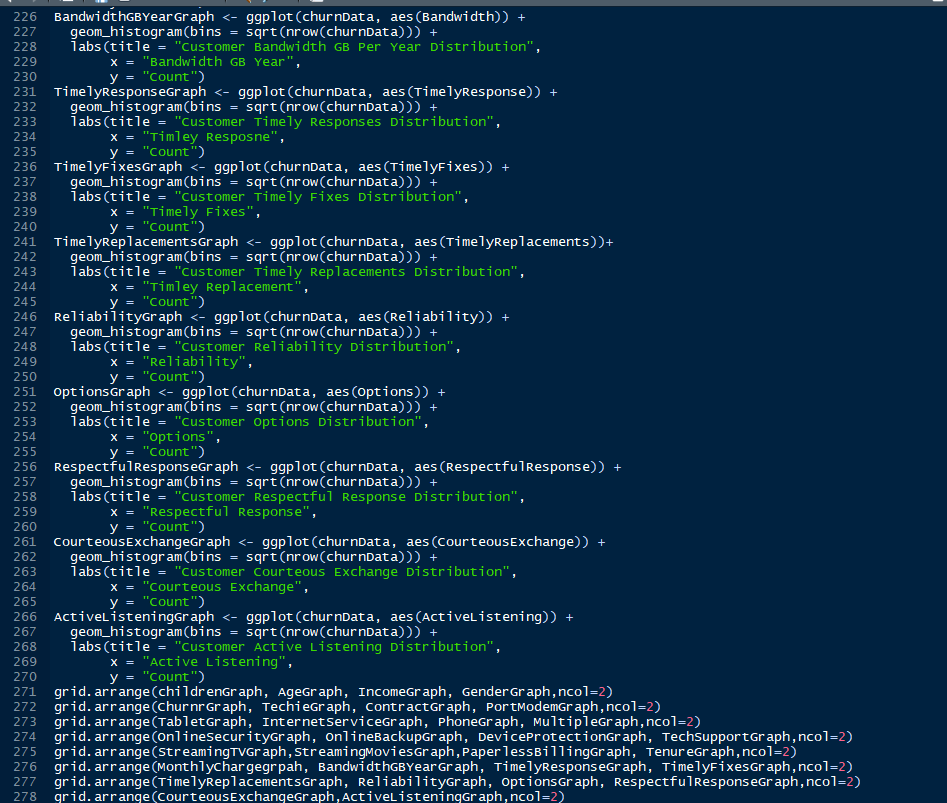
A glimpse of test dataset and exporting the datasets to excel

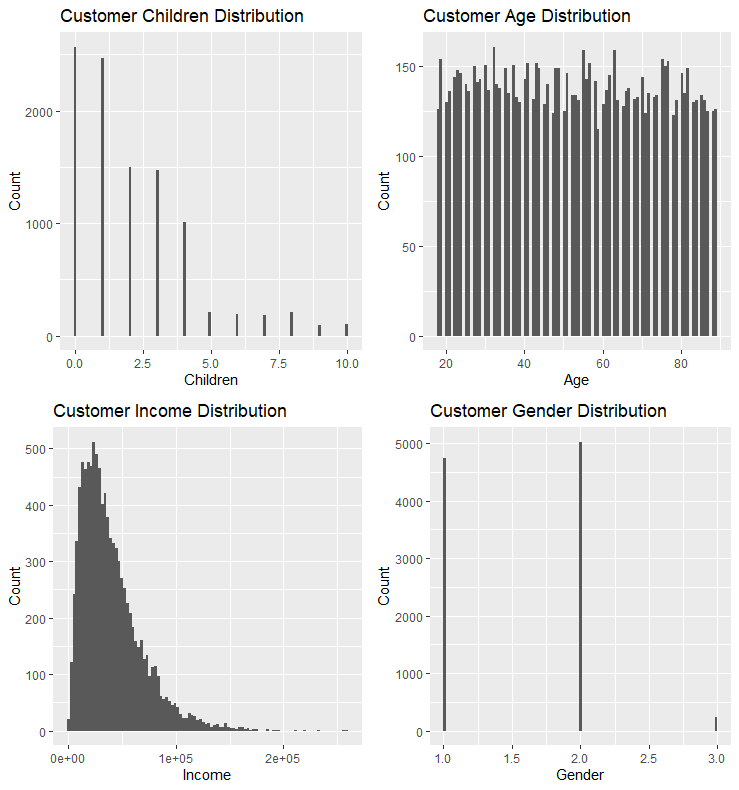
4.

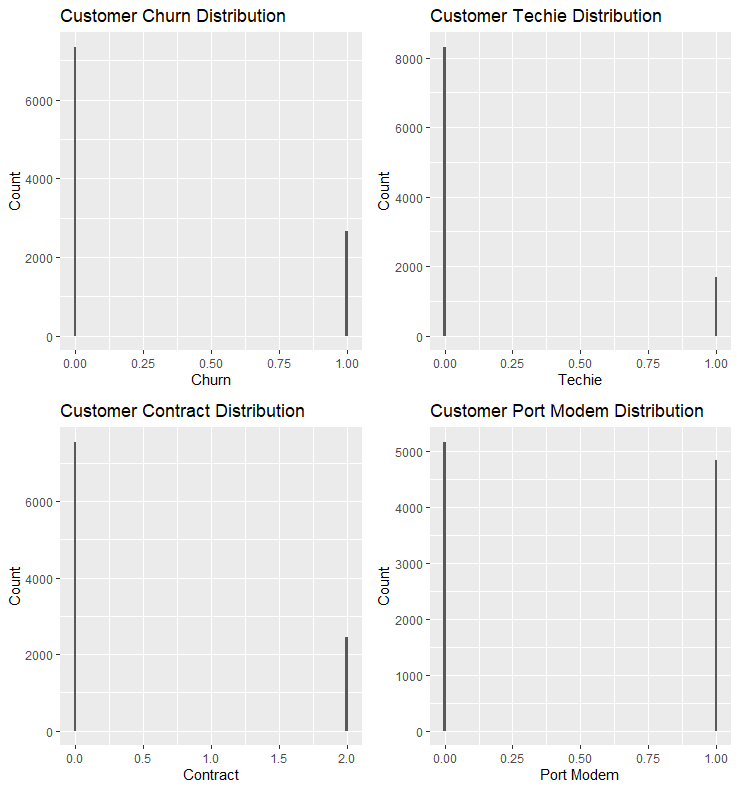
Univariates statistics

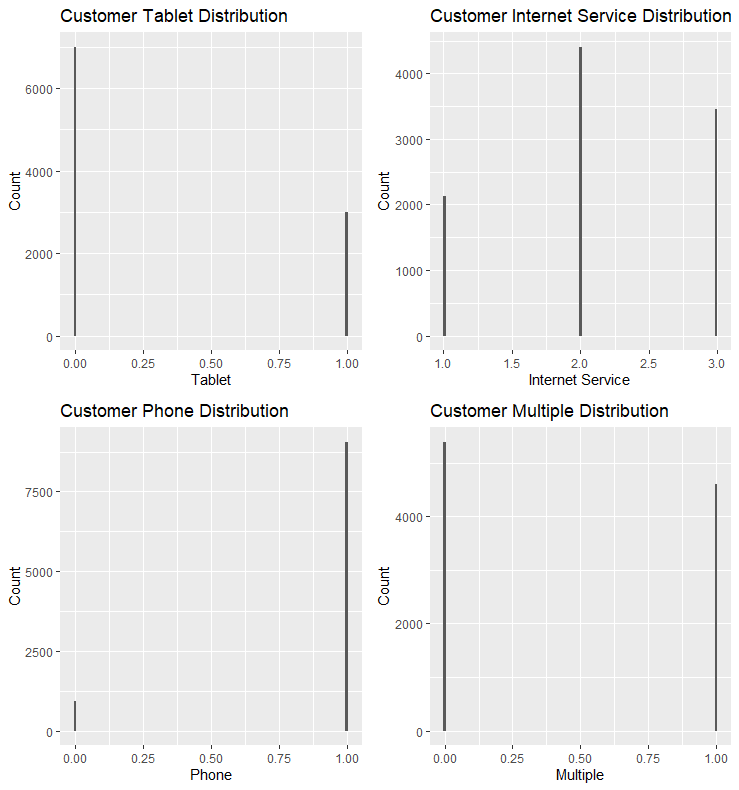


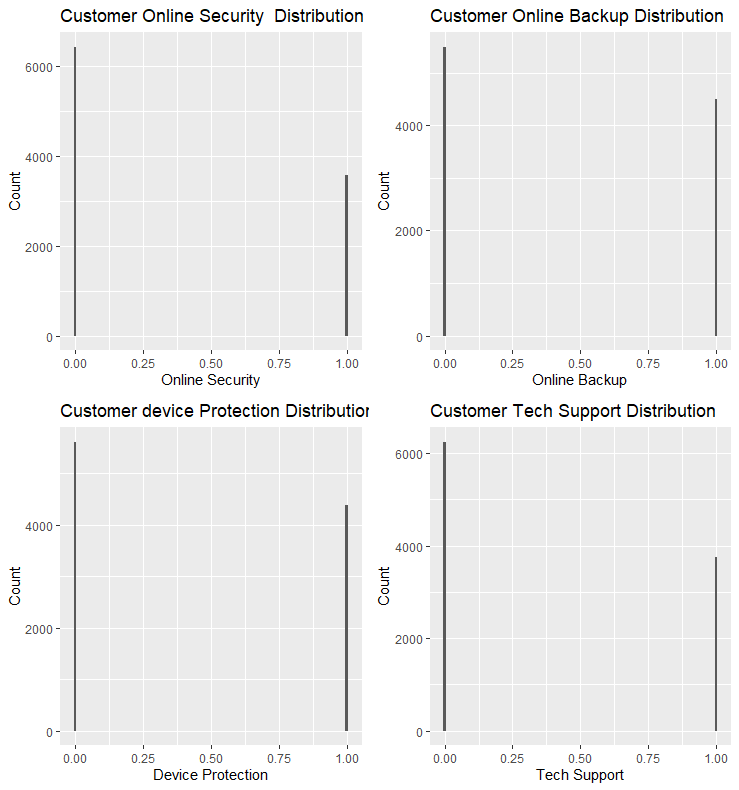


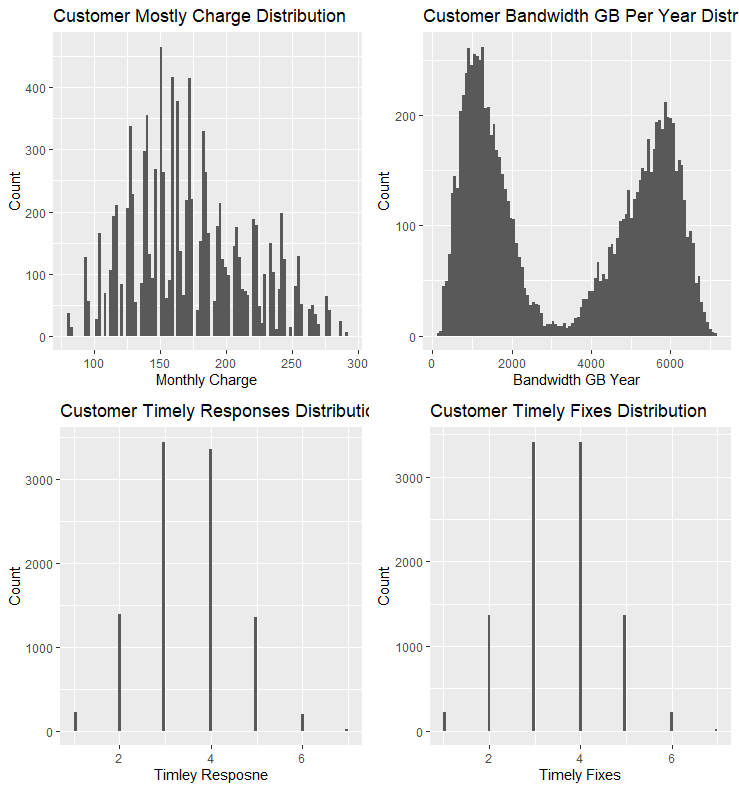


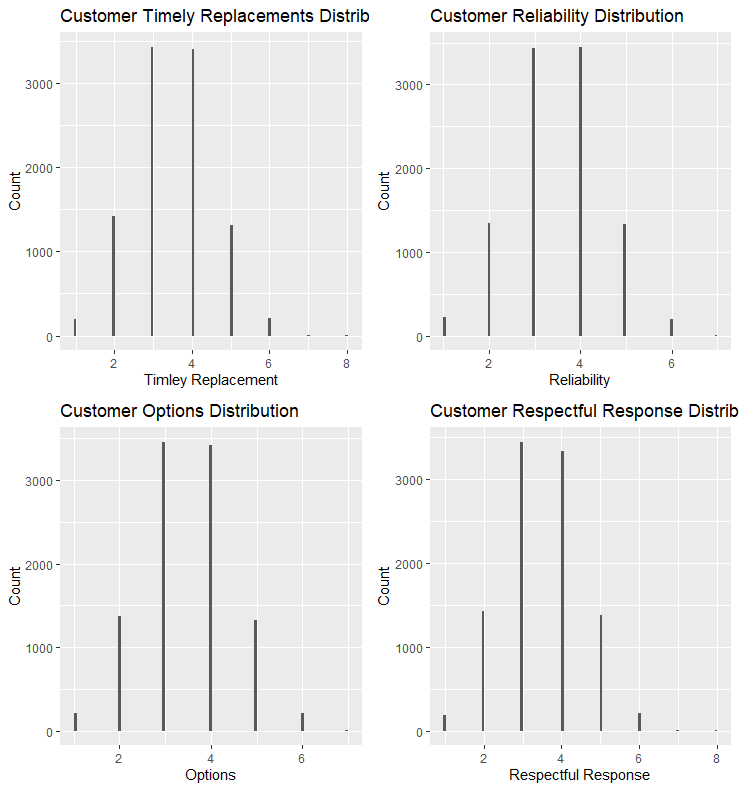


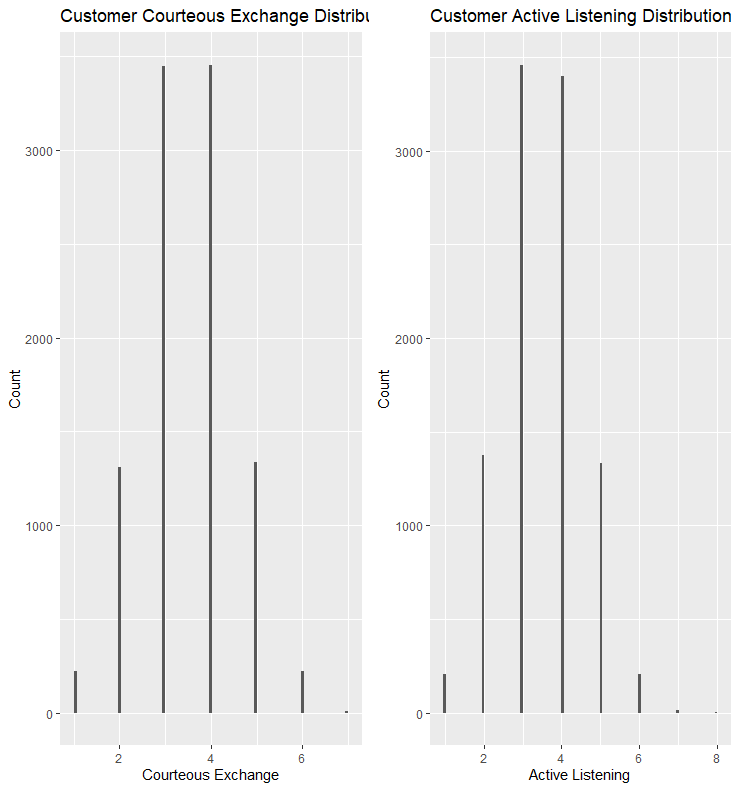




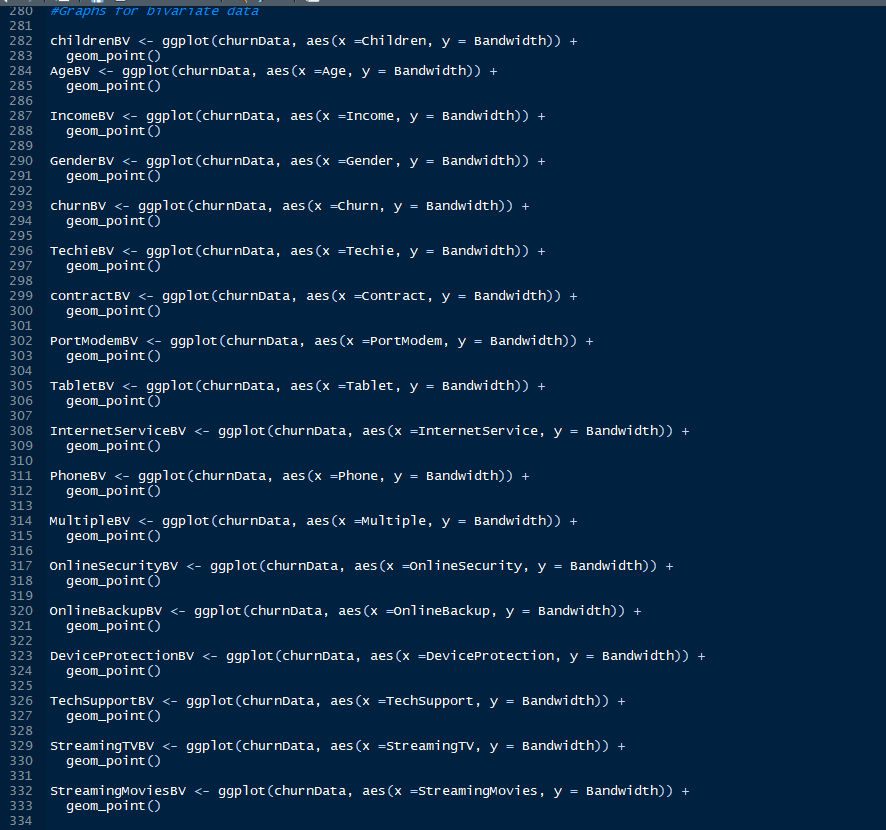
  

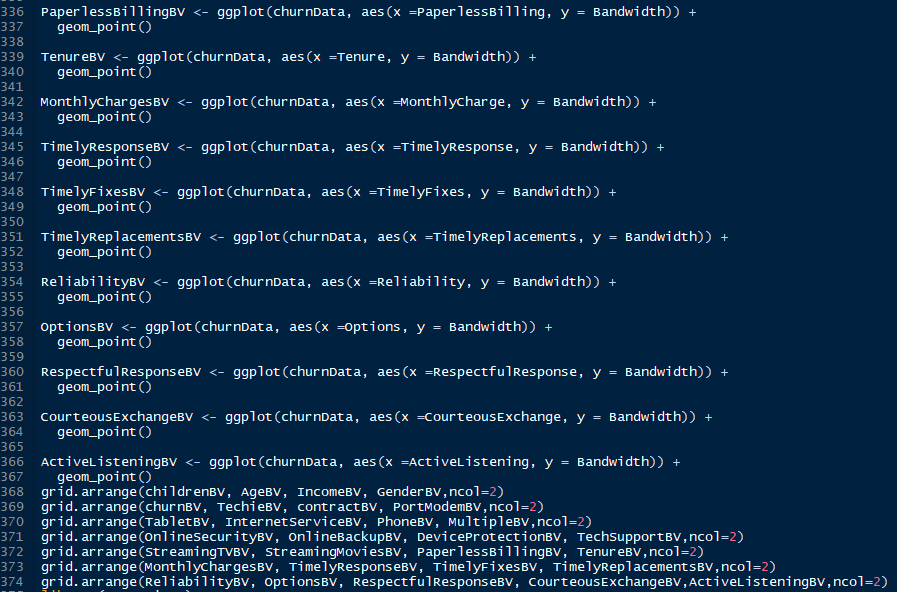



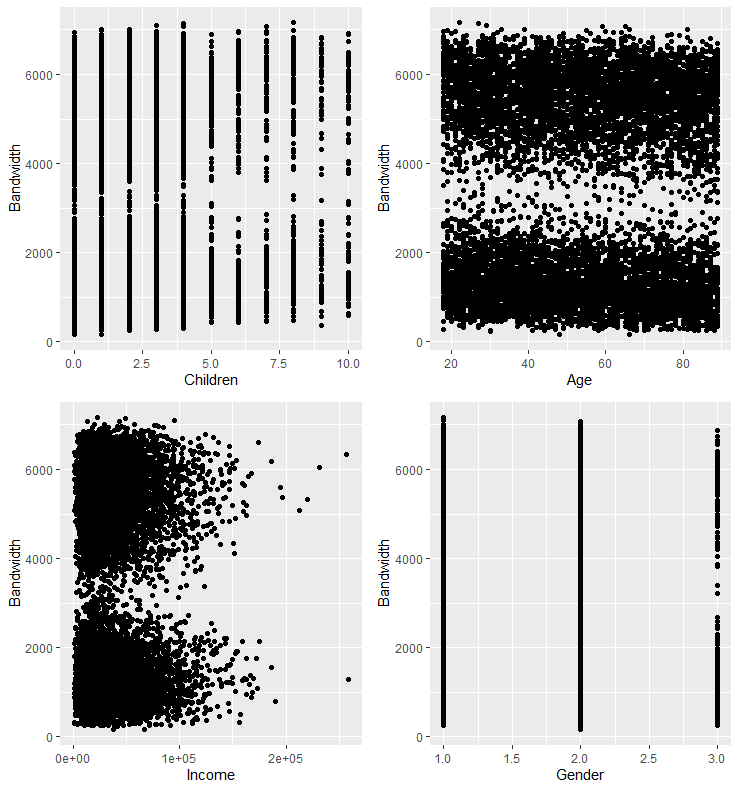


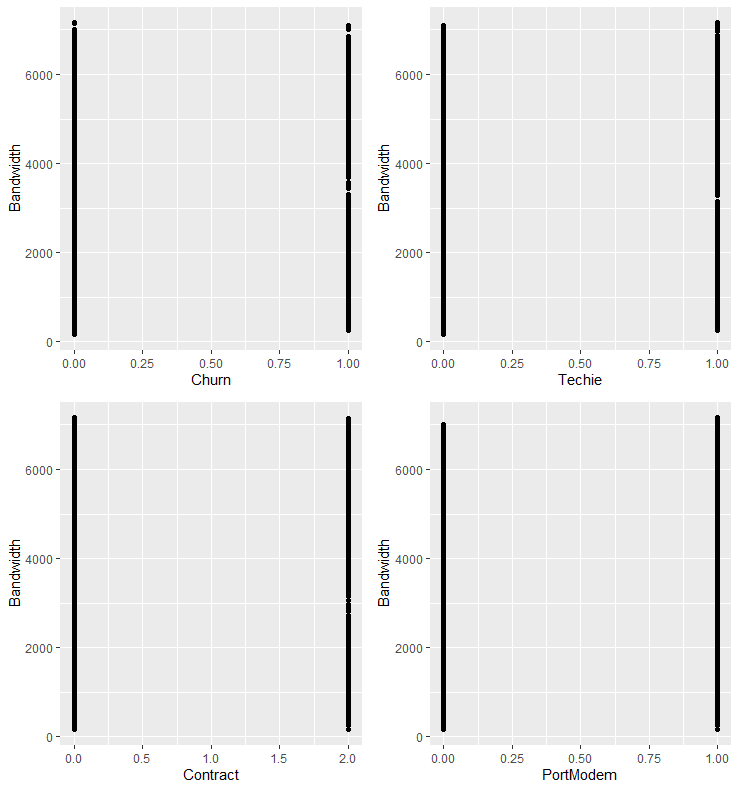


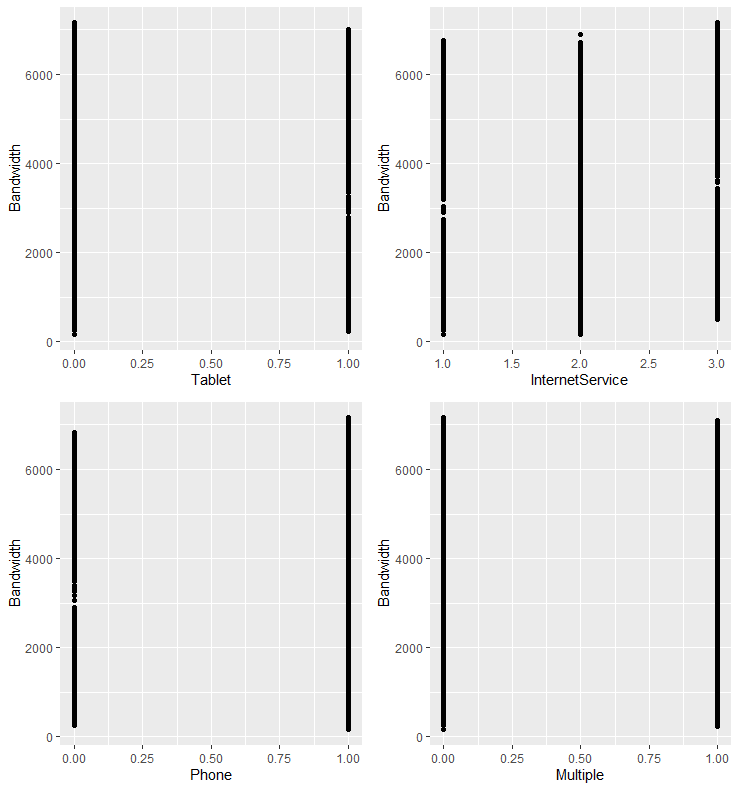
Bivariate statistics

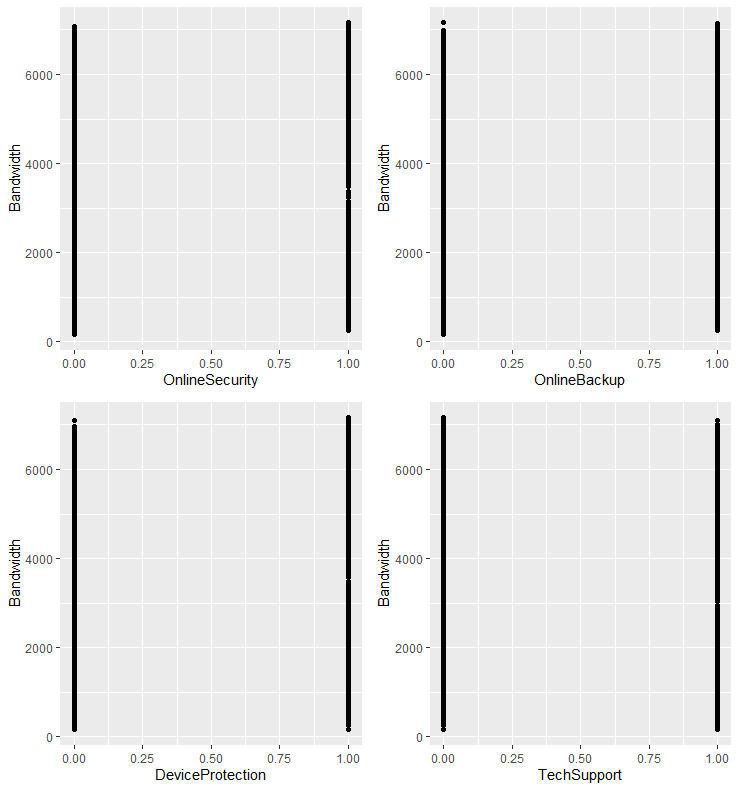


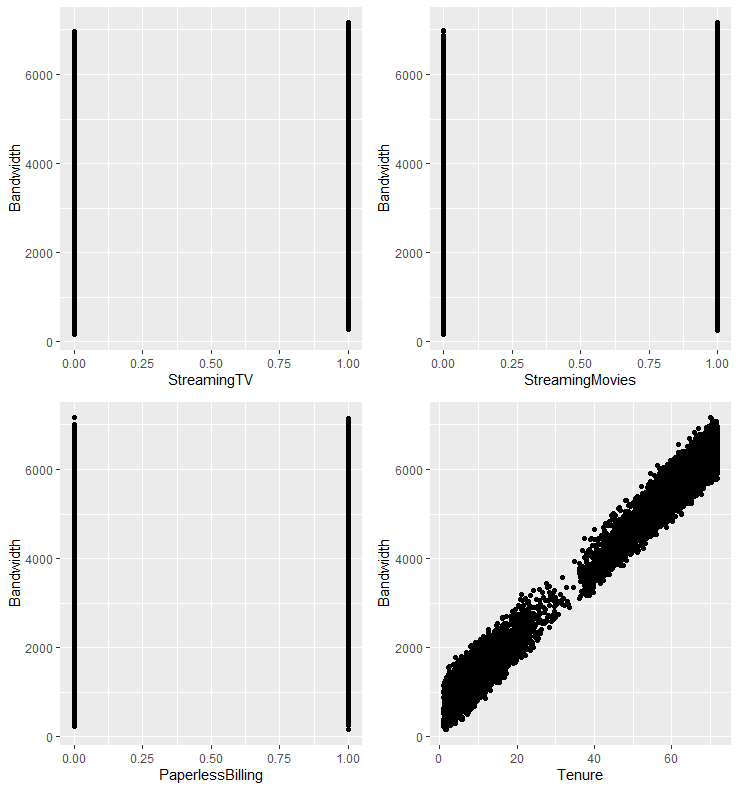


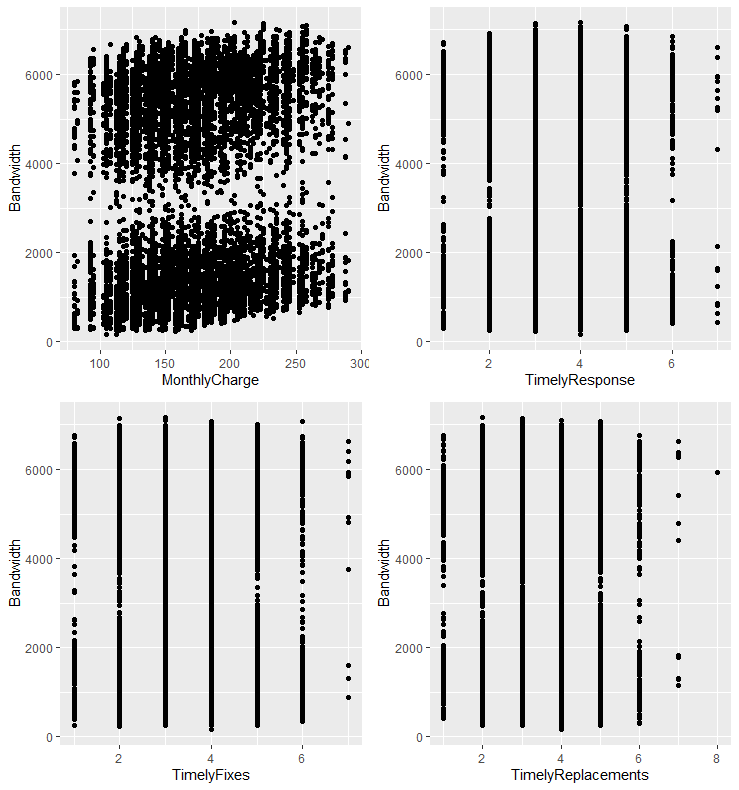


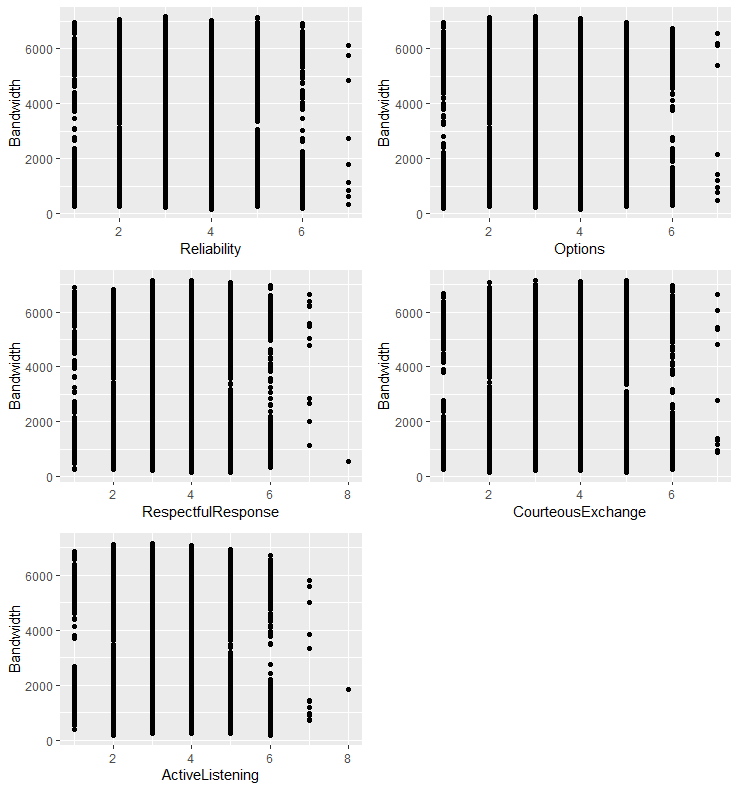




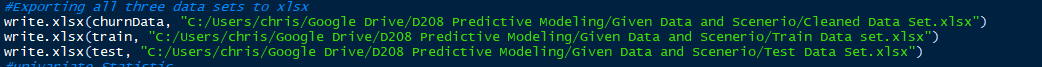








5.  Provide a copy of the prepared data set.

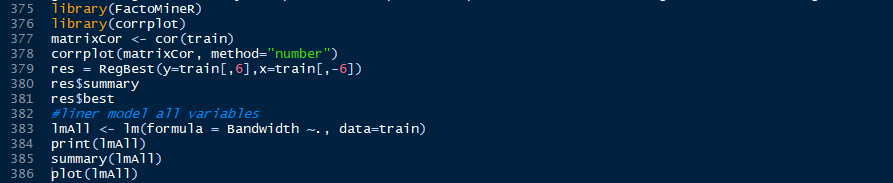


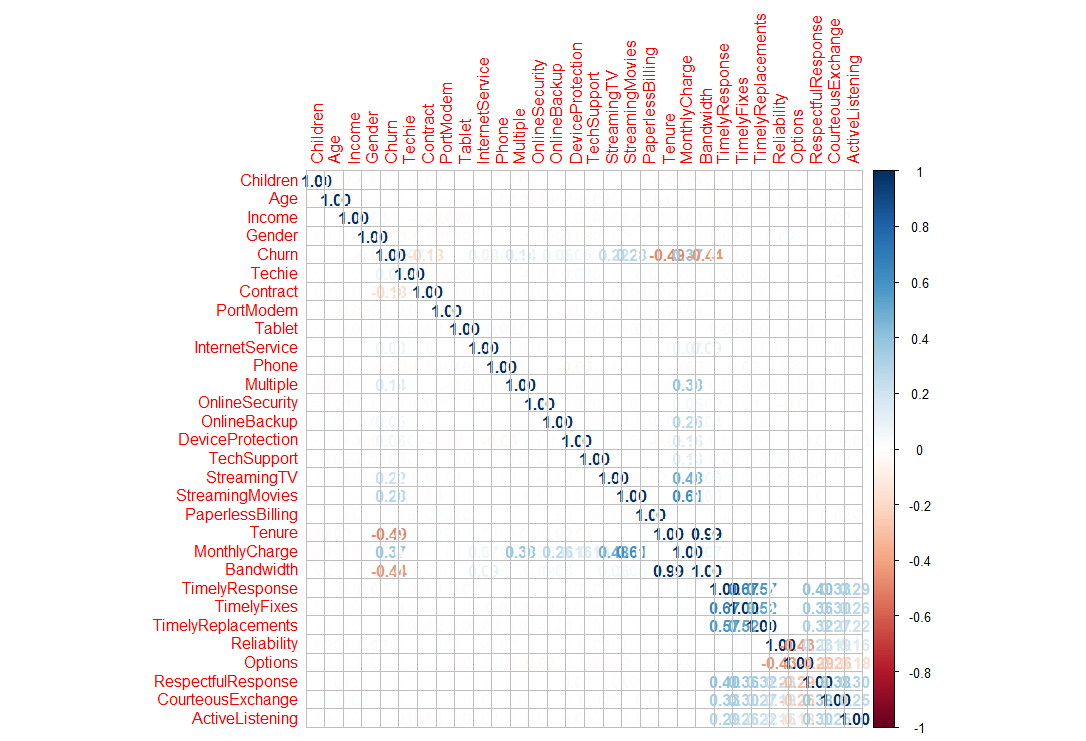
Excel files are included in the submission.

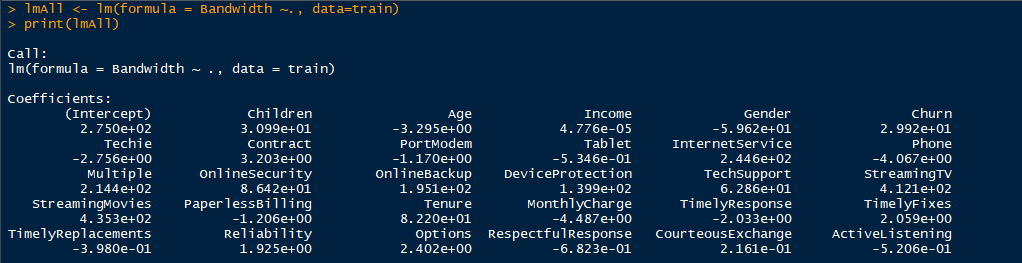
**Part IV: Model Comparison and Analysis**

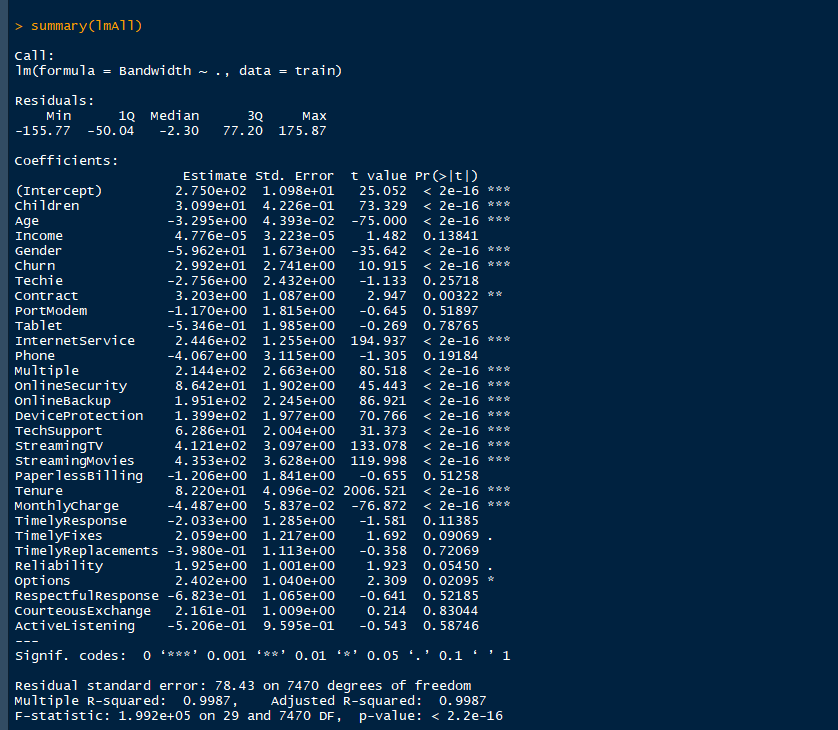
D.

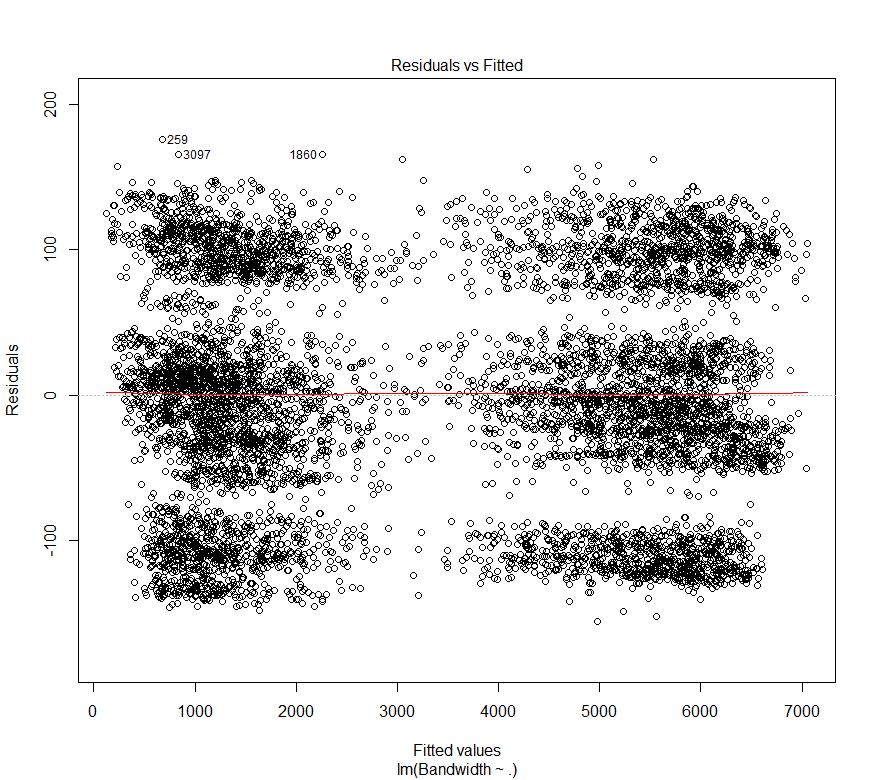
1.

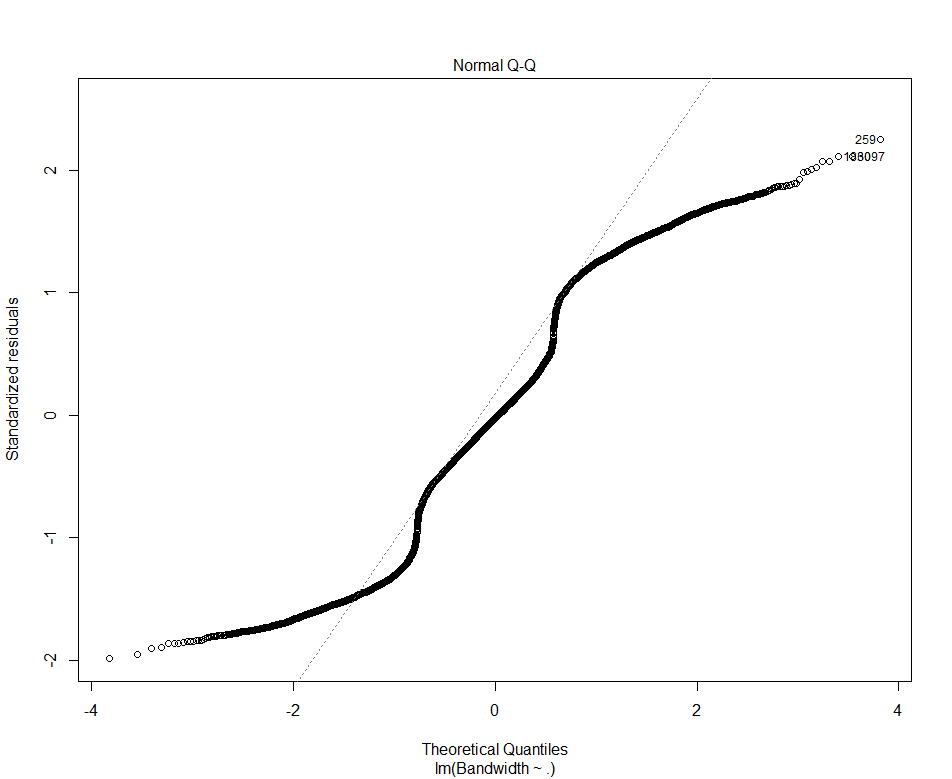


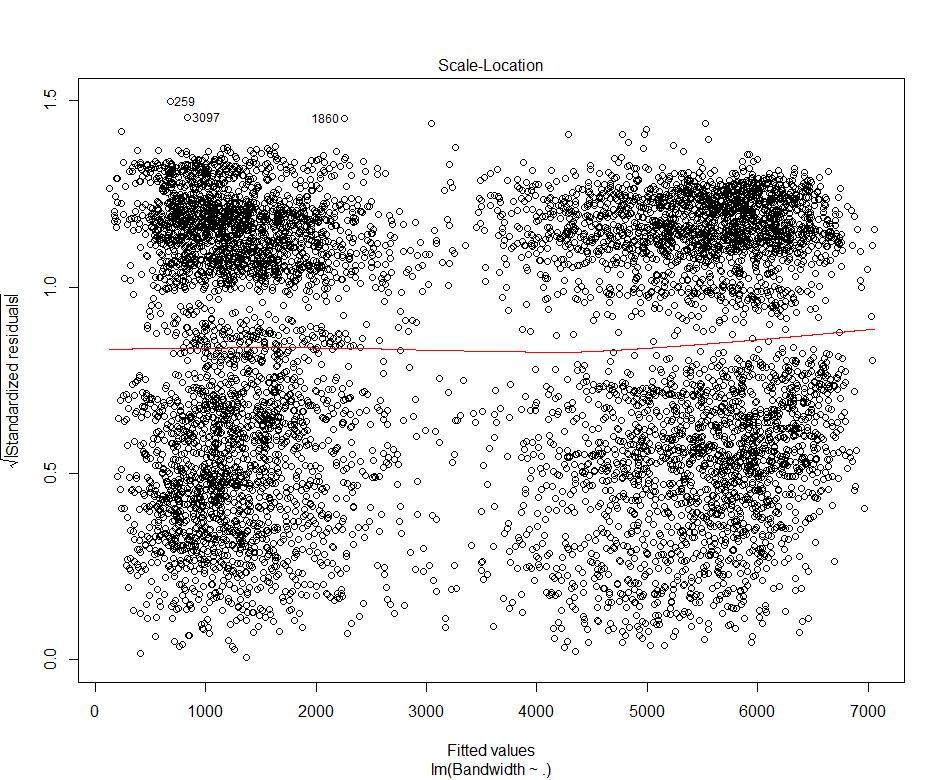


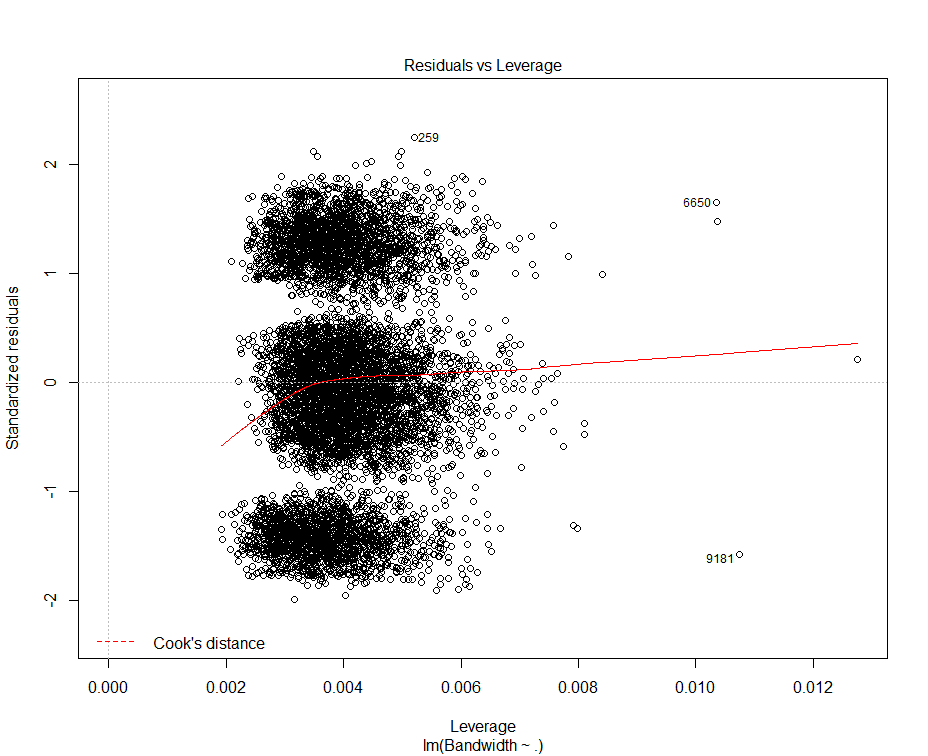






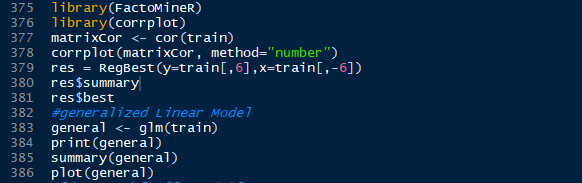


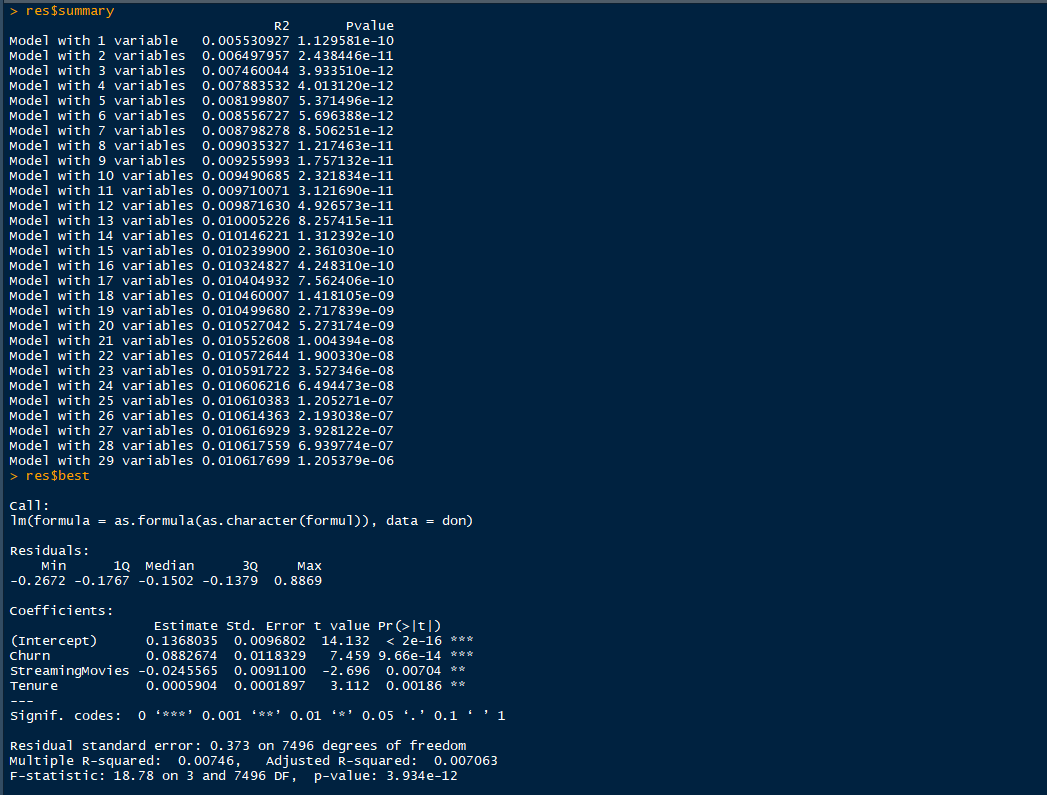




The residual of all predictor variables minimum is -155.77 with a maximum of 175.87. The intercept and 14 other variables have significant codes of 0.001. The estimate compared to the std error for the most part are comparable for most variables. But there other variables that’s std error and estimate are completely noncomparable. With an R squared value close to one we can say there is a large proportion of variability found in the outcome and this will lead to strong multicollinearity. This also tells us that we have more variables needed for this regression. The formula of the regression is y = 2.750+ 3.099 \* children – 3.295\* Age + 4.776\* Income – 5.962\* Gender + 2.992\* Churn – 2.756\* Techie + 3.203 \* Contract – 1.170 \* PortModem – 5.346\* Tablet + 2.446\* InternetService – 4.067\* Phone + 2.144 \* Multiple + 8.642\* OnlineSecurity + 1.951\* OnlineBackup + 1.399\* DeviceProtection + 6.286\* TechSupport+ 4.121\* StreamingTV+ 4.353\* StreamingMovies- 1.206\* PaperlessBilling +8.220\* Tenure – 4.487\* MonthlyCharge – 2.033\* TimelyResponse + 2.059\* TimelyFixes – 3.980\* TimelyReplacements + 1.925\* Reliability + 2.402\* Options – 6.823\* RespectfulResponse + 2.161\* CourteuousExchange – 5.206\*ActiveListening.

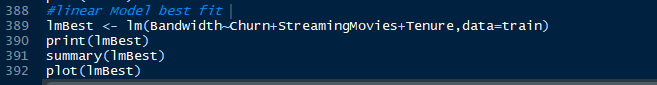
2.

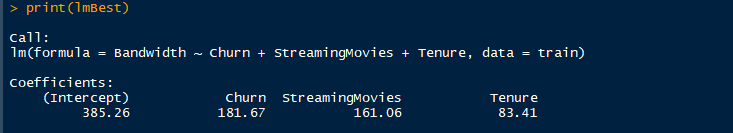


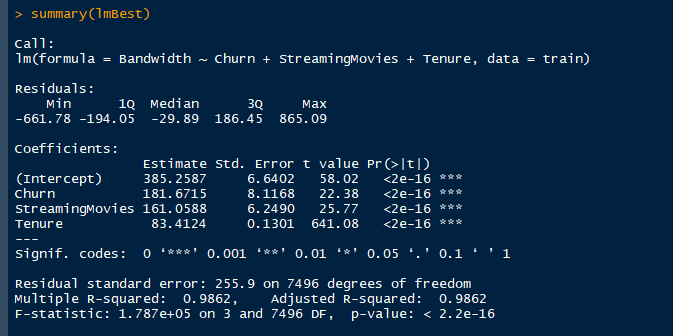


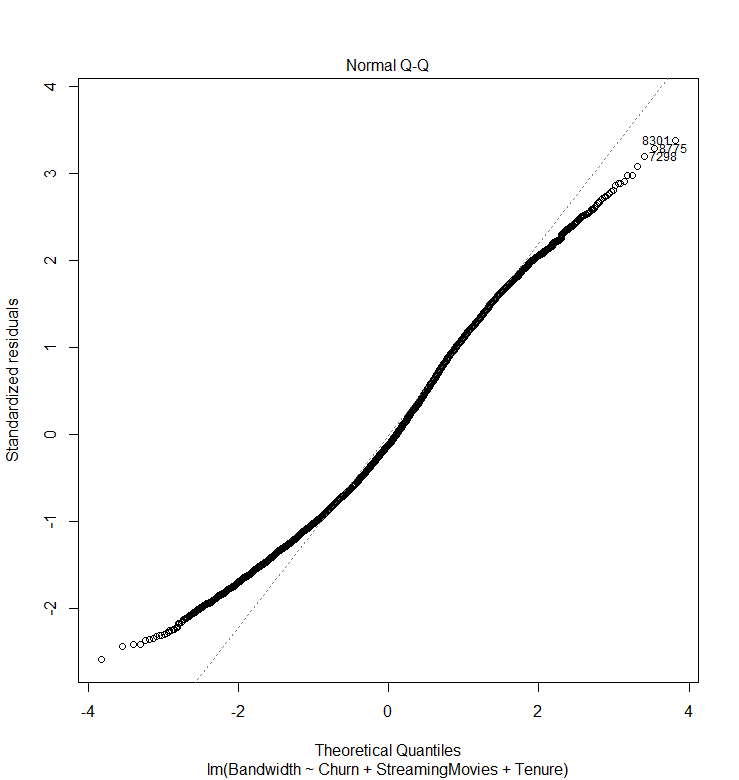
Using the FactoMineR function RegBest we can determine the best variables to use in the reduced multiple regression model. Based on this the best statistical variables are Churn, StreamingMovies, and Tenure. Tenure and StreamingMovies have both p-value significant codes 0.01 then Churn and the intercept of p values significant codes of 0.001.

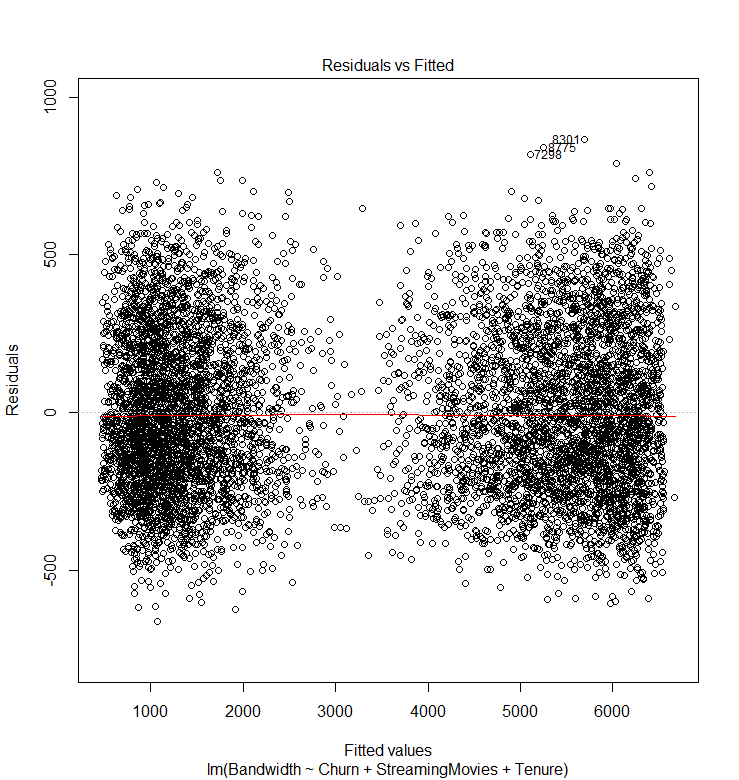
3.

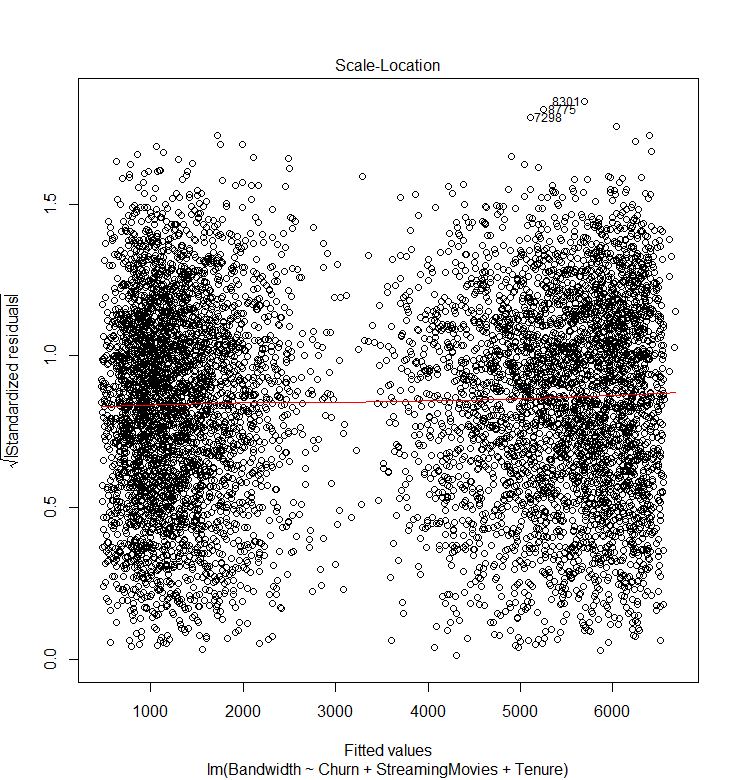


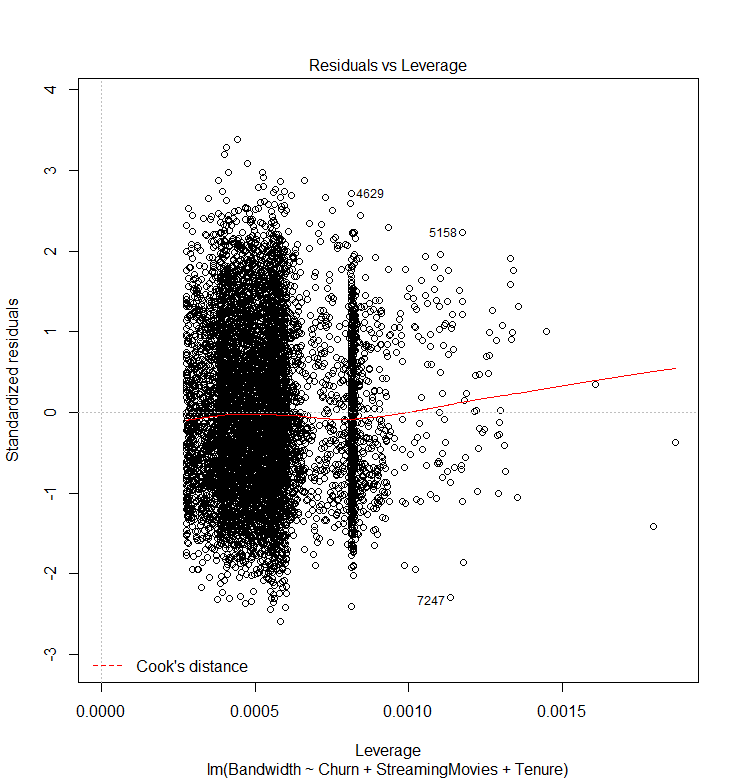












The R squared variable for this model explains over 98 percent of the variables are explained in the model. The equation of the model is y = 385.26 + 181.67\*Churn +161.06\*StreamingMovies+83.41\*Tenure.

E.

1.

Refer to D1 and D3 for residuals of each model. The logical variable selection was done by using FactoMineR’s RegBest function which statistically calculates the best variable to be used. The model evaluation metric was completed by looking at the R Squared value for both models separately but the result of these models was the model explained 98 percent of the variables.

2.

Refer to the above to see any calculations and code output.

3.

All code is included above.

**Part V: Data Summary and Implications**

F.

1.

Based on the reduced multiple linear regression model using the following independent variables Churn, StreamingMovies, and Tenure. The R squared statistic of this model was 0.9862, which would be approximately 98 percent. This means that 98 percent of the model variables are correlated in the model. The equation for this regression model is y = 385.26 + 181.67\*Churn +161.06\*StreamingMovies+83.41\*Tenure. In other terms coefficients mean Churn will increase 181.67 units times Bandwidth, StreamingMovies will increase 161.06 units times Bandwidth and Tenure will increase 83.41 units times Bandwidth. All variables are statistically significant to 0.001 and all have the same p-values of <2e-16. There could be some limitations to the analysis could the dataset used be too small and whether not splitting the dataset would have fixed these issues. Also having different additional variables on particular details like bandwidth per month for example.

2.

For this course, we were asked to help a telecommunication company with the churn of its customers. In this report, we looked at how bandwidth gigabits per year used determined if a customer was likely to churn from the company. Based on the results from the reduced regression model using Churn, StreamingMovies, and Tenure. All of these variables have the same significant codes and p values, meaning they have a direct linear relationship to the dependent variable of Bandwidth. The recommendation to the company would be to focus on the customers who stream movies and have a higher bandwidth usage. So offering a package with a higher bandwidth cap or that has a lower fee for going over the bandwidth cap. Also, the company should look at giving discounts to customers who have been tenured longer with the company. In conclusion, the company needs to focus on customers streaming movies and who have longer tenure with high bandwidth because these customers are valuable to the company.

**Part VI: Demonstration**

G.  Provide a Panopto video recording that includes all of the following elements:

•  a demonstration of the functionality of the code used for the analysis

•  an identification of the version of the programming environment

•  a comparison of the **two** multiple regression models you used in your analysis

•  an interpretation of the coefficients.

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.

To submit your recording, upload it to the Panopto drop box titled “Multiple Regression Modeling – NBM2 | D208.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

*Assumptions of linear regression*. Statistics Solutions. (2021, August 11). Retrieved October 8, 2021, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/.

Francois Husson, J. J. (2021, January 8). *RegBest: Select variables in multiple linear regression in factominer: Multivariate exploratory data analysis and Data Mining*. RegBest: Select variables in multiple linear regression in FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. Retrieved from https://rdrr.io/cran/FactoMineR/man/RegBest.html.

J.  Demonstrate professional communication in the content and presentation of your submission.