**Part I: Research Question**

Companies who lose customers obviously lose the money that comes with the customers. When a customer leaves a company, it is called churning. By doing a churn analysis the company can predict what factors are involved and use them to reduce churn (Martin, 2020). The goal of this project is to use the telecommunication data and logistic regression to show the most valuable variables and predict whether a customer will churn or not.

**Part II: Method Justification**

In logistic regression there a few assumptions that should be met. First, the outcome is binary. This means the outcome can only be one of two answers, yes or no. It could also be positive or negative, or male or female for example. Another assumption is that the observations are independent of each other. This means they should not be related to each other in any way. There should also be no multicollinearity among the explanatory variables. There should not be any extreme outliers in the data set. If there are outliers they can be removed, replaced, or kept in the model, noting how they could affect the results. Logistic regression also assumes there is a linear relationship between each variable and the response variable. The last assumption is the sample size being sufficiently large (Awasthi, Mahajan, Pisal, & Srivastava, 2020).

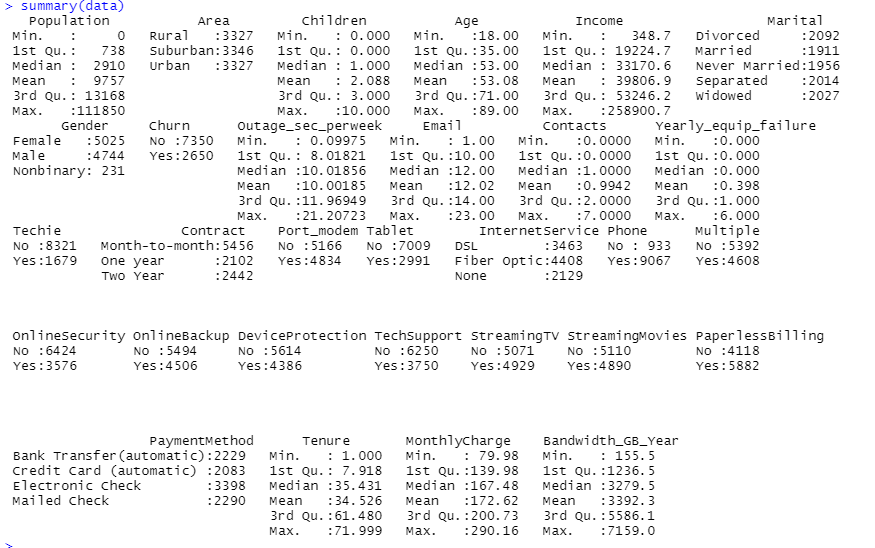
R is a valuable tool for statistical analysis. The coding is straightforward and graphs to support modeling are easily made and easy to understand. There are many powerful libraries with commands to help make predictive modeling easier (Brownlee, 2019). R’s coding is easier to understand than other statistical languages, such as Python. R is also a little faster at processing, making it the choice for most data science professionals (Admin, Haemorrhoids, & Francis, 2021).

Logistic regression is appropriate for this dataset because the outcome variable has only two possibilities. Churn is a yes or no response.

**Part III: Data Preparation**

The goal in data preparation is to make the data usable for R and predictive modeling. Before the data can be used for the logistic regression it will need to be imported, and a new data set will be created using only the variables needed for analysis. The categorical variables are in the character format, they will be changed to factors. This will enable R to use them for logistic regression.

The summary stats show the spread of the continuous data. It also shows the amount of each value for each column of categorical data.



The initial data in imported with the line of code:

.

The new data set with variables to be analyzed is the created:

.

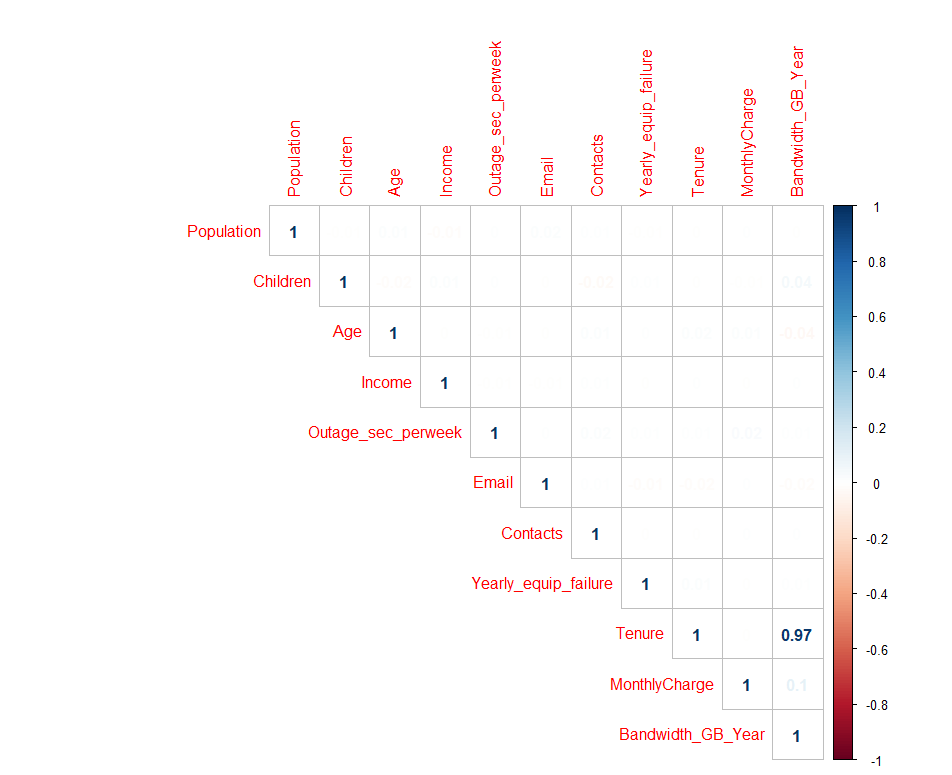
The categorical data is transformed into factors by using the code below:

.

A data set with only continuous variables is created to make checking the correlations easier.

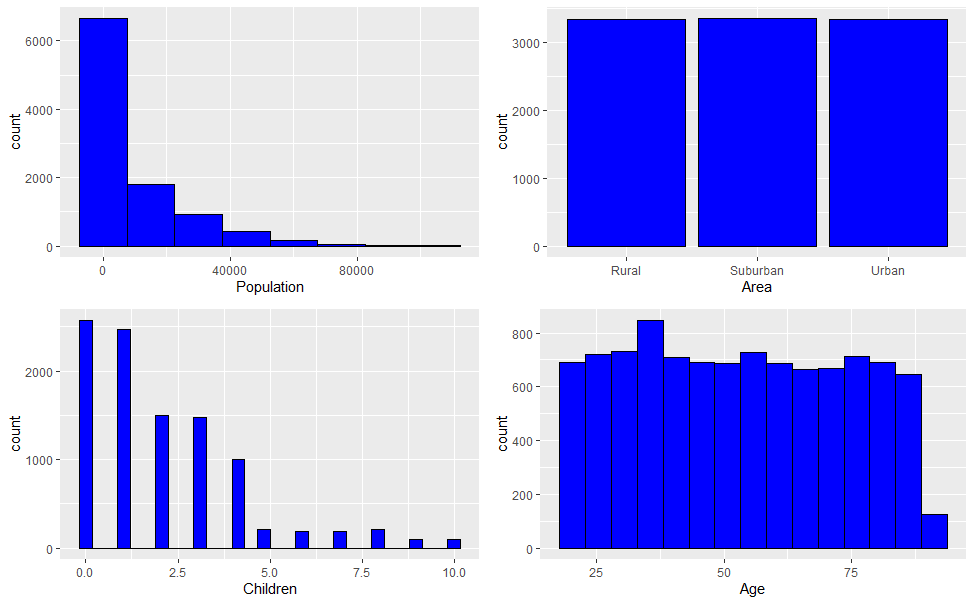
. A correlation plot shows the results:

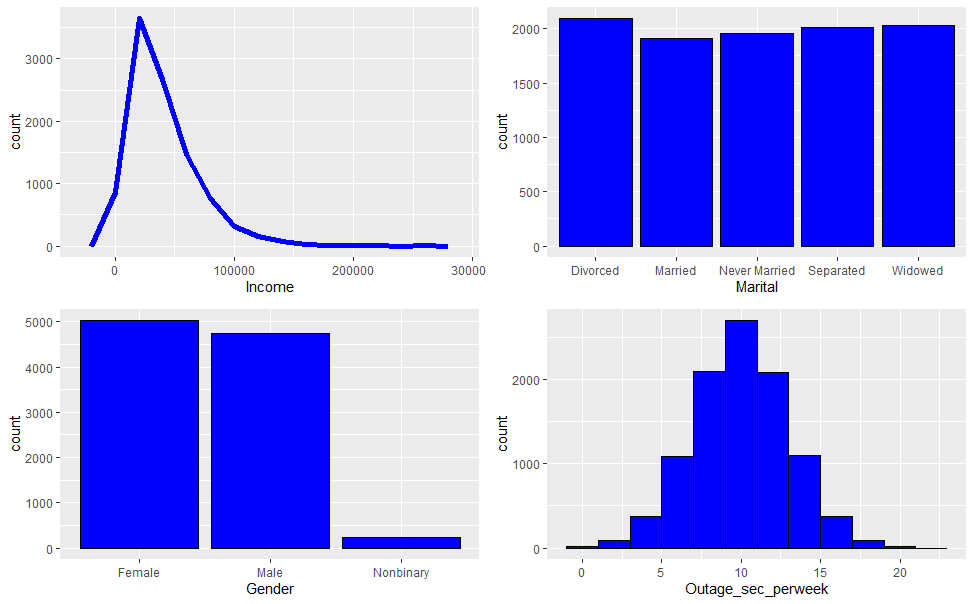


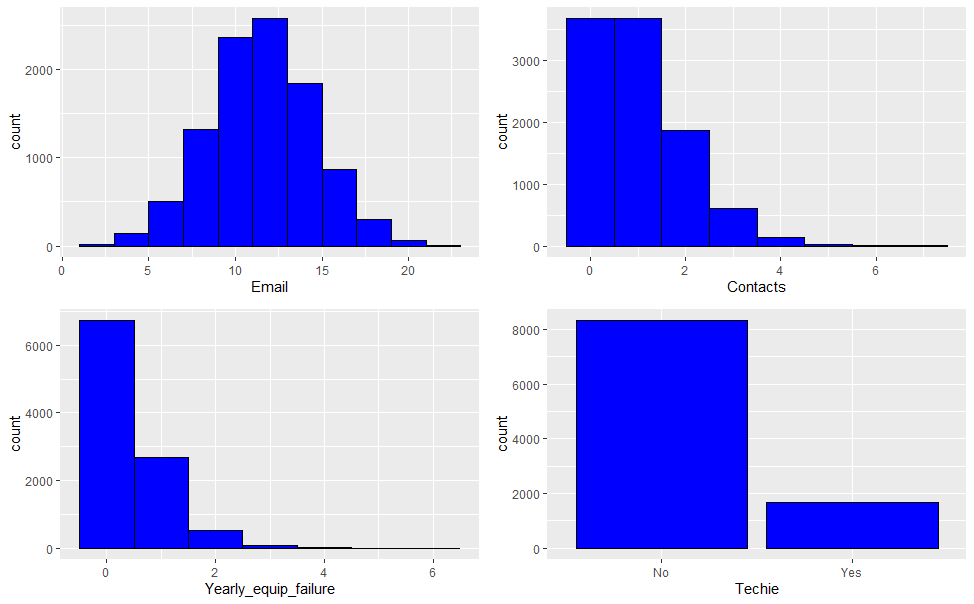
.

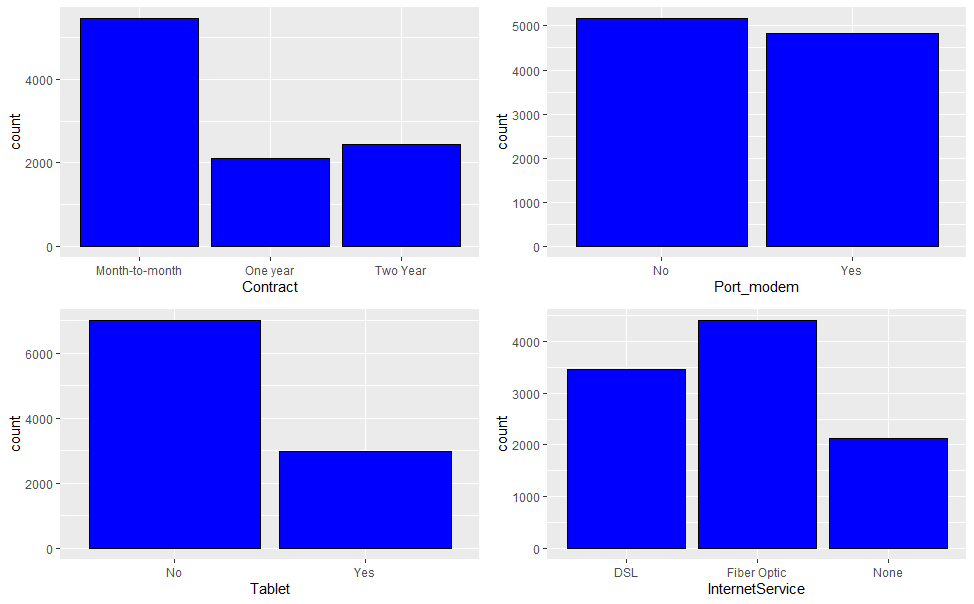
The variables Tenure and Bandwidth\_GB\_Year have a very high correlation. The variable Bandwidth\_GB\_Year will be dropped before creating the model.

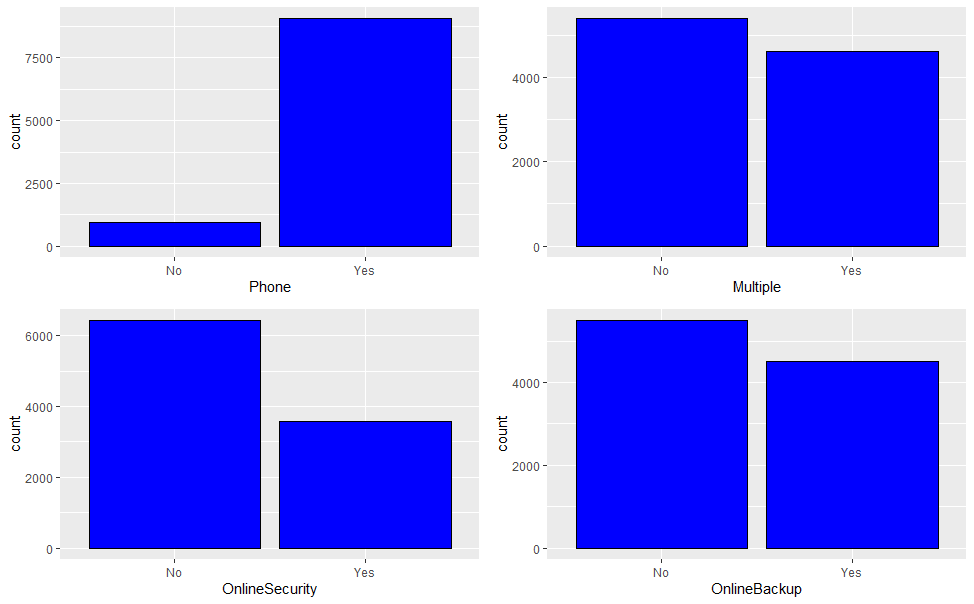
The univariate statistics of the data are presented in the graphs below:

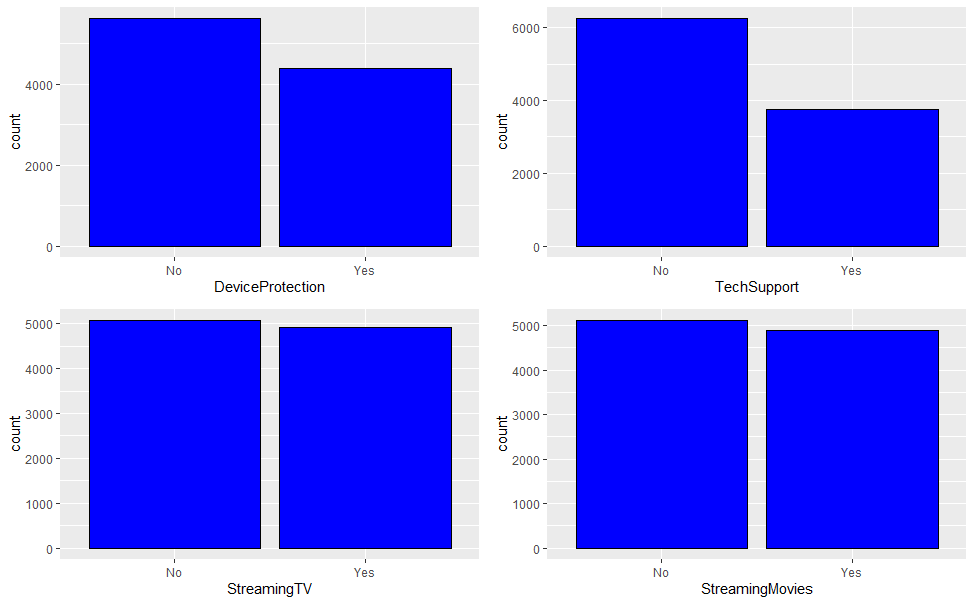


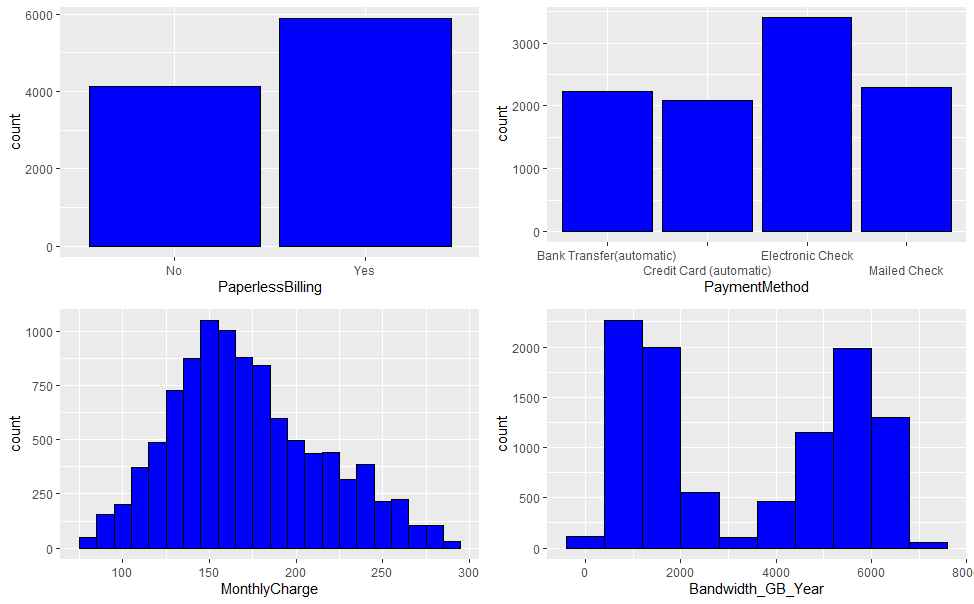


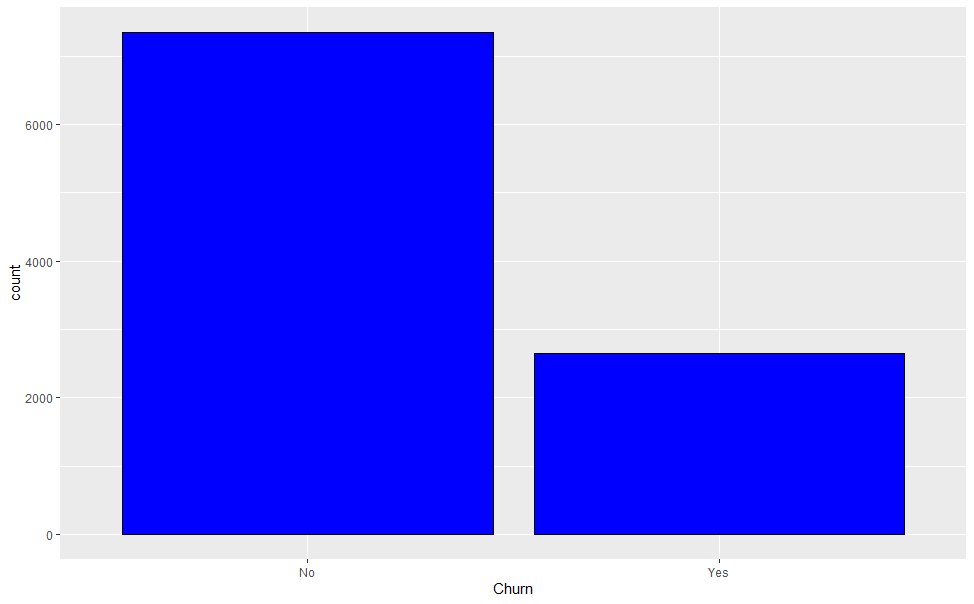






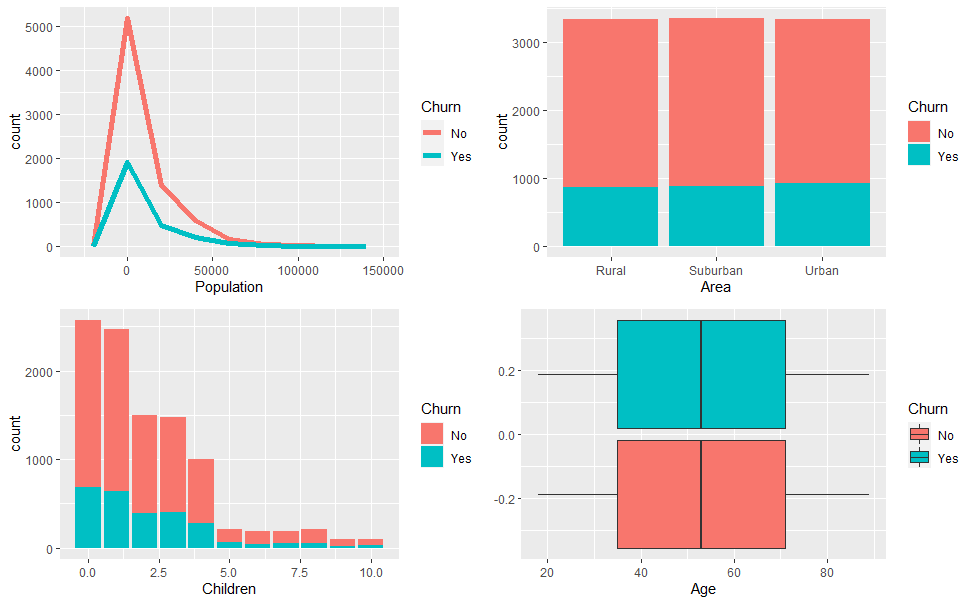


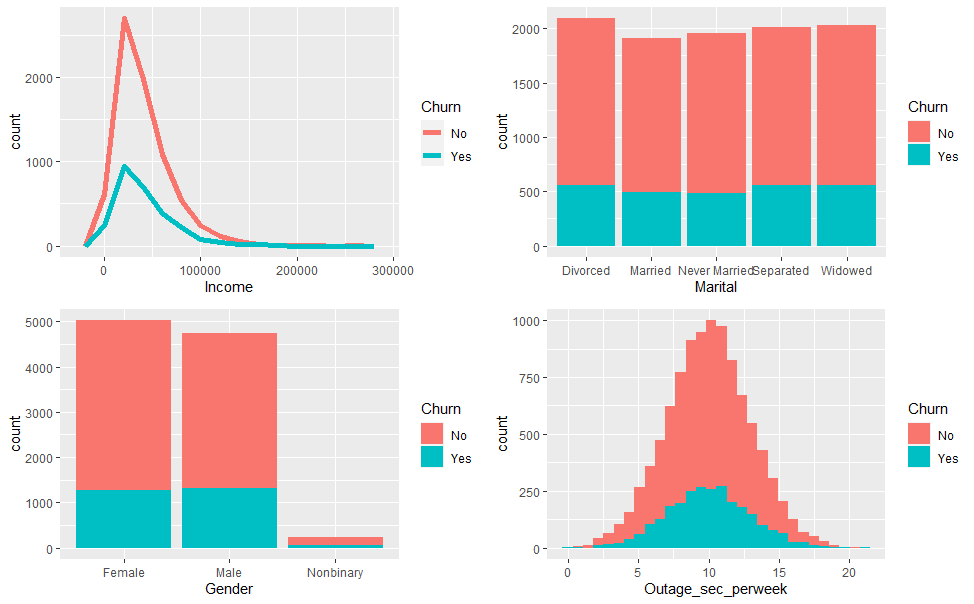


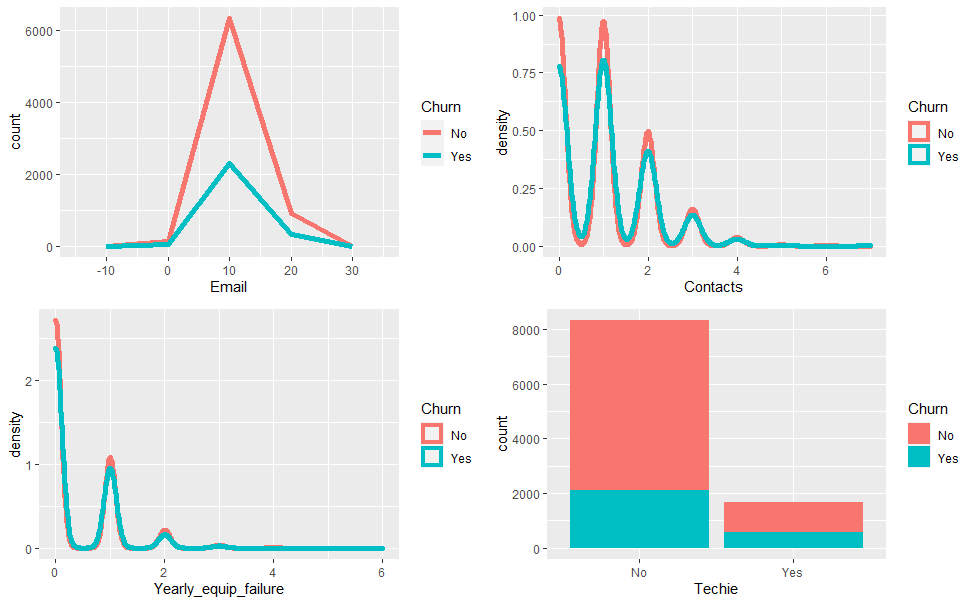


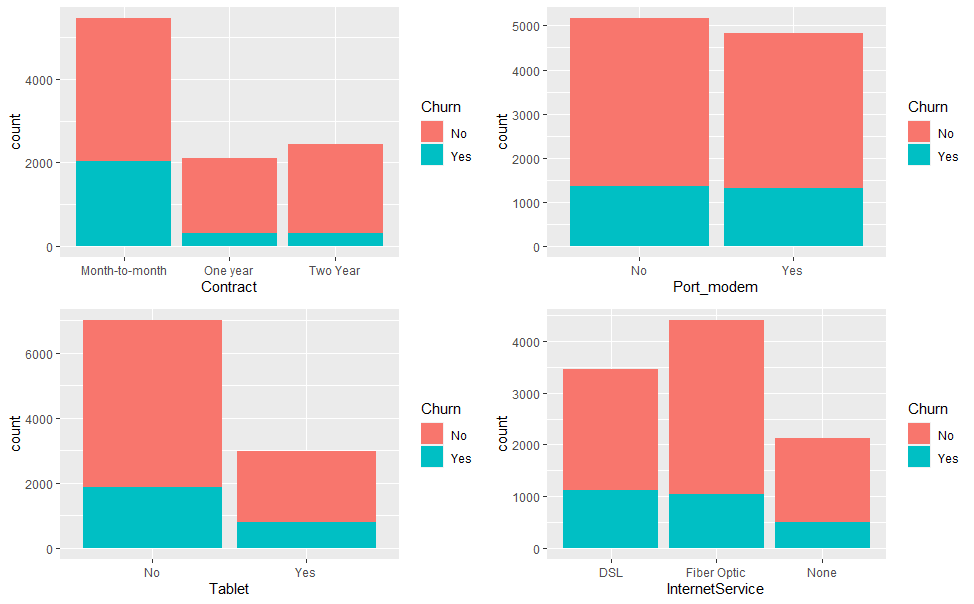
The last graph shows the amount of people who have churned.

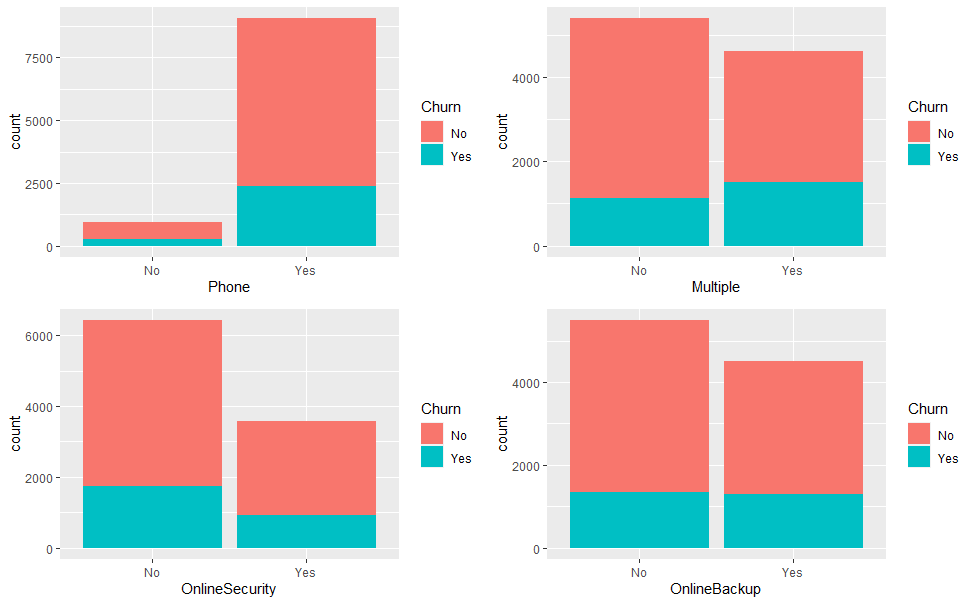
The bivariate graphs below show the same information as it relates to churn.

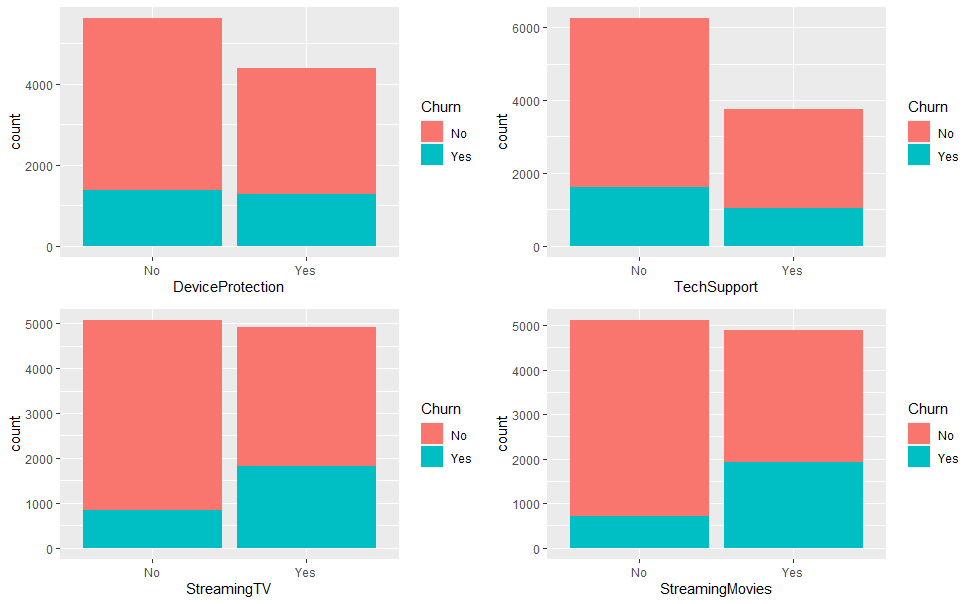


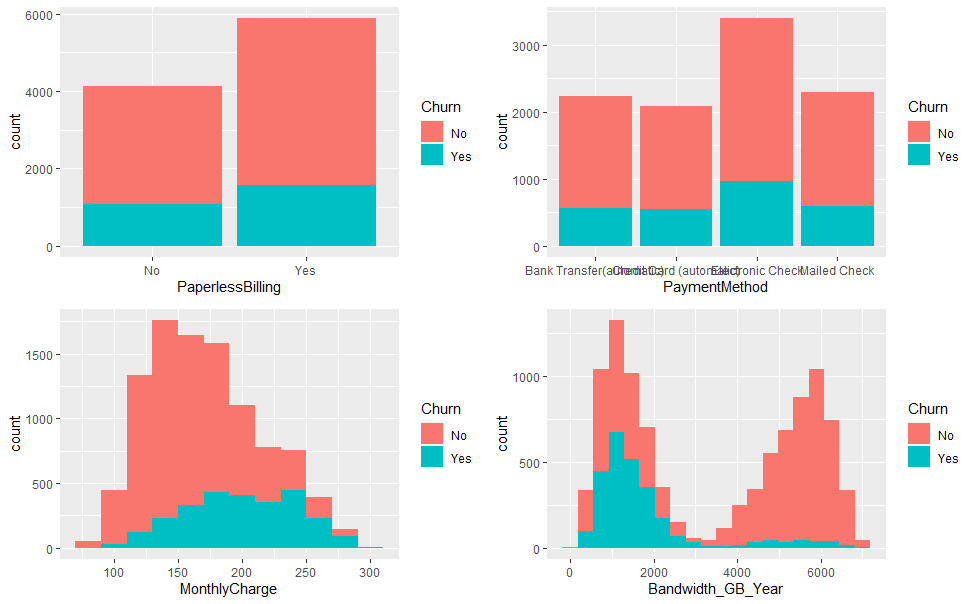






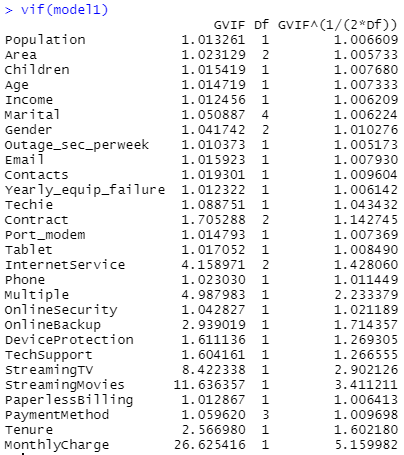






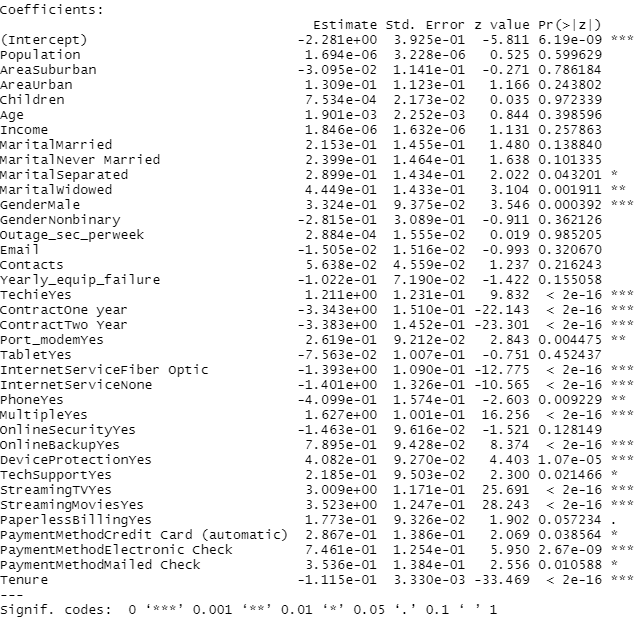
**Part IV: Model Comparison and Analysis**

The initial model is created with all the variables minus the Bandwidth\_GB\_Year because of correlation. After the initial model is created, a VIF command is used to check for multicollinearity. The results are shown below:

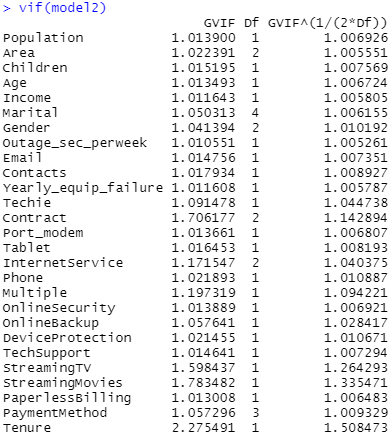
.

The MonthlyCharge variable has the highest score and will be dropped from the next model.

The next model is produced, and the summary shown below:

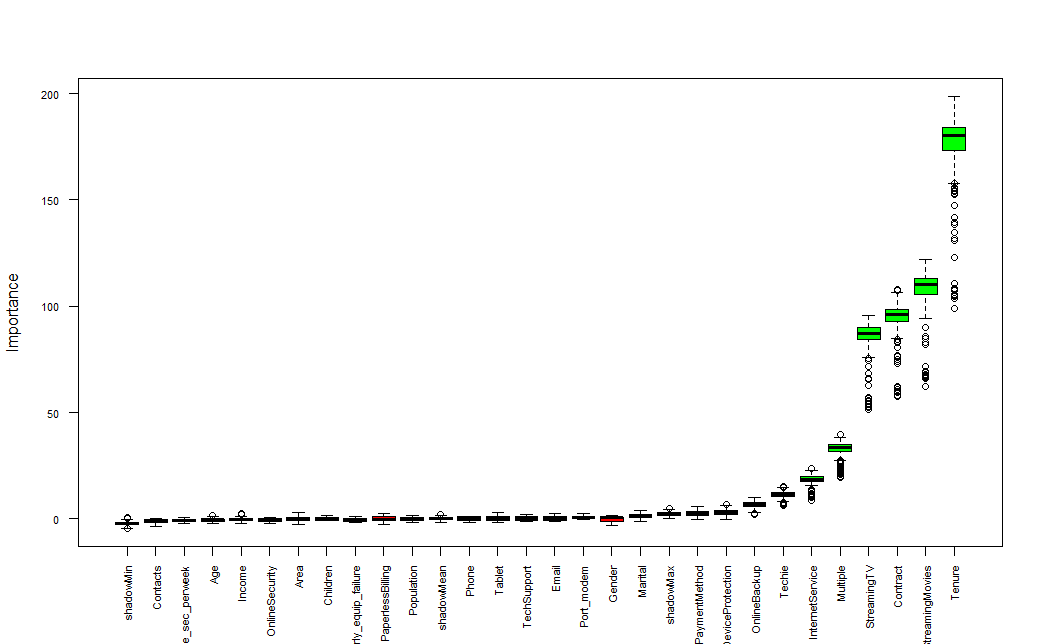
.

The VIF scores below look good, so the above output will be all of the variables used for the initial model.

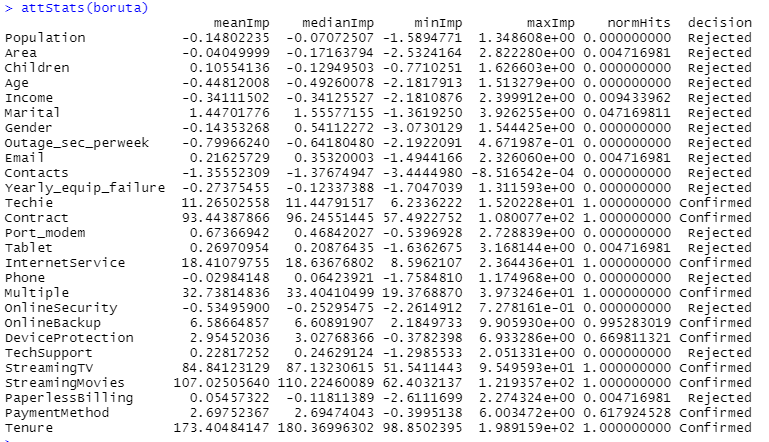


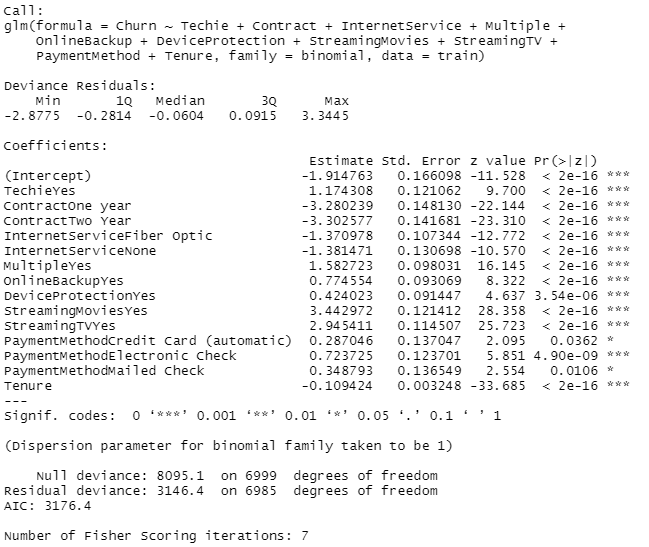
Variable selection is important for modeling because it removes redundant variables which improves accuracy. Having too many variables can result in overfitting and slow computation. This can take more memory and add strain to hardware. The Boruta Package in R picks the variables based on importance. It is an improvement of random forest, a very popular method for variable selection (Rasool, 2019). Using Boruta takes more time to run in R but will help modeling and computer resource usage in the long term. To run Boruta, the code below is used:

. The results are plotted:



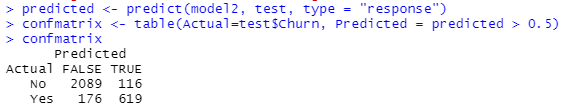
The plot shows the most important variables in green, and how important they are. Using the “attStats” command, it lists the variables and their statistics.

  
Using this, a reduced model will be created as seen below:



The final model now has all the most statistically significant variables and is ready for making predictions.

A confusion matrix is a good way to evaluate the performance of the initial model and the reduced final model. The initial model used to make predictions on the test data set with the code below:



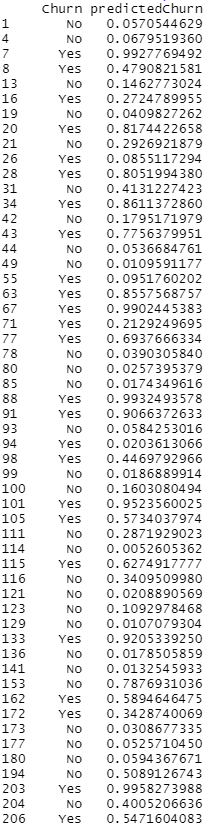
The accuracy of the model is calculated with the following formula: (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives). Filling the information, the output is a rate of 0.9027. The sensitivity is the proportion of actual positives which are correctly identified. The sensitivity for this model is 0.842. The specificity is the proportion of negatives that are correctly predicted. The specificity of the initial model is 0.922. These values will be compared with the reduced model. The code and output below show the confusion matrix from the reduced model:



Using this model, the accuracy is calculated at 0.907. The sensitivity is 0.849. The specificity of the reduced model comes in at 0.925.

The reduced model provides results that are better than the original model. This shows the variable selection process worked as the much-reduced model can accurately predict churning with a fewer number of variables. This is important in machine learning to keep computer resources free and save time in prediction making process.

Some of the predicted results versus actual results are posted below:



The first column shows if a customer churned or not, and the second column shows the percentage of a customer churning or not. For example, the first entry shows a customer did not churn and the predicted probability of the customer churning to be about 6 percent. The third entry shows a customer did churn and the probability of the customer churning to be 99 percent.

**Part V: Data Summary and Implication**

From the results of the variable reduction, the final model equation is:

y = -1.915 = (1.174\*TechieYes) + (-3.280\*ContractOne-Year) + (-3.302\*ContractTwo-Year) + (-1.371\*InternetServiceFiberOptic) + (-1.381\*InternetServiceNone) + (1.583\*Multiple) + (0.775\*OnlineBackupYes) + (0.424\*DeviceProtectionYes) + (3.443\*StreamingMoviesYes) + (2.945\*StreamingTVYes) + (0.287\*PaymentMethodCreditCard) + (0.724\*PaymentMethodElectronicCheck) + (0.349\*PaymentMethodMailedCheck) + (-0.109\*Tenure).

The Techie coefficient tells us that if the customer is a Techie, they are more likely to churn. Customers with a one- or two-year contract are less likely to churn than customers with a month-to-month contract as shown with a negative coefficient. Customers with Fiber Optic or No Internet Service are less likely to churn than customers with DSL. If have customer has Multiple services, Online Backup or Device Protection, they are more likely to churn given their coefficient is positive. Customer who streams TV and Movies have a high positive coefficient, showing they are more likely to churn than customers who do not have those options. For Tenure, the longer the customer is with the company, the less likely they are to leave since the coefficient is negative. Telecommunications customers have many options in the present time, the quality of service and content available could be a key factor in keeping or losing customers.

The practical significance of the model is to determine what customers would churn given the options they subscribe too. Being able to predict the factors that cause a customer to churn, a company could put time and resources into meeting the needs most used by customers to reduce churn.

From the data collected and model created customers who have a longer Tenure are more likely to stay. To increase tenure, customers with short term month-to-month contracts should be focused on to get them on longer term contracts. Customers with DSL are more likely to churn as well. The quality of DSL service should be investigated further to see if changes in quality can be made. Customers who have streaming services are more likely to churn, and since many options are available to customers, this should be an area of focus.

**References**

Martin, J. (2020, November 29). What is churn, how to calculate it, and how to prevent it in 3 ways. *CloudApp*. Retrieved March 13, 2021, from <https://www.getcloudapp.com/blog/what-is-churn>

Brownlee, J. (2019, August 22). Use r for machine learning.

Retrieved March 29, 2021, from <https://machinelearningmastery.com/use-r-for-machine-learning/>

Rasool, R. (2019, November 12). Logistic regression in r: A classification technique to predict credit card default: R-bloggers. Retrieved March 29, 2021, from <https://www.r-bloggers.com/2019/11/logistic-regression-in-r-a-classification-technique-to-predict-credit-card-default/>

Posted by Deepanshu Bhalla on June 1, 2. (n.d.). Select important variables USING Boruta Algorithm. Retrieved March 29, 2021, from <https://www.datasciencecentral.com/profiles/blogs/select-important-variables-using-boruta-algorithm#:~:text=%20Select%20Important%20Variables%20using%20Boruta%20Algorithm%20,in%20which%20it%20considers%20all%20features...%20More%20>

Zach. (2020, October 29). How to perform logistic regression in r (step-by-step). Retrieved March 29, 2021, from <https://www.statology.org/logistic-regression-in-r/>

UCertify. (n.d.). Sign in. Retrieved March 29, 2021, from <https://wgu.ucertify.com/?func=ebook&chapter_no=15#top>

Awasthi, S., Mahajan, P., Pisal, S., & Srivastava, S. (2020, December 30). Assumptions of logistic regression. Retrieved March 31, 2021, from <https://datamahadev.com/assumptions-of-logistic-regression/#:~:text=Assumptions%20of%20Logistic%20Regression%201%20Logistic%20Regression%20Algorithm.,5%20Large%20sample%20size.%20...%206%20Conclusion.%20>

Admin, Haemorrhoids, & Francis. (2021, March 24). R vs python: Which programming language is better for you? Retrieved April 01, 2021, from https://statanalytica.com/blog/r-vs-python/#:~:text=R%20is%20slightly%20faster%20than%20Python%20to%20perform,of%20the%20slowest%20programming%20languages%20in%20the%20world.