**Project Report**

**Case Study:**

**Utilizing Logistics Regression and Decision Tree Model**

**to Predict Customer Churn for a Telecommunication Company**

Group #

Names :

- A

- B

- C

- D

# **Introduction**

Businesses rely heavily on their ability to retain customers. Losing customers would mean a reduction in overall customer lifetime value, and it directly affects profitability. Moreover, the loss would be translated into higher customer acquisition costs since more new customers should be acquired to balance off the number of leaving customers. It is imperative that this issue has raised major concerns in various industries. While the issue is relevant in every industry, it is particularly a vital issue in the telecommunication industry. Companies in this industry can have an average annual churn rate to be as high as 67%, and 75% of the new customers are churners of other companies (Hughes). Preventing churn can save millions of dollars, and hence predicting churn before it happens is crucial. Developing a solid model to predict churn will generate insight into the customer profile that has a higher possibility to churn. Therefore, the company will be able to create quick prevention measures.

The target of this project is to create a logistic regression model and the decision tree model to predict customer churn. This project will split the data into a training set, which the model will be built from, and a test set, which will be utilized to test the model’s accuracy. The project will also compare the two models and determine which model gave the more accurate result.

# **Data Overview**

## **Variables, Data Types, and Pre-processing**

The data for the case consists of 2,114 observations with 14 variables, ranging from customers’ profiles, such as ID and gender, to the profile of their interaction with the company, such as tenures, payment method, and monthly charges. There is also the target variable Churn, which will be valued as yes when the customer has churned or no when the customer has not This target variable should also be made as a factor and numerical 0 and 1 for the predictive model.

Most of the variables are nominal data with character data type, except the SeniorCitizen variable and tenure variable being integer, and MonthlyCharges and TotalCharges being numerical. Character variables should be converted to factors for further analysis. The SeniorCitizen is comprised of 1 if the customer is a senior citizen, and 0 if the customer is not. The data type should be converted to factor with 2 levels, as well as converted to “yes” and “no” to ensure consistency of categorical data. After the conversion, the data type for each variable in the dataset can be shown as follows.

|  |  |
| --- | --- |
| The data type for every variable obtained from R | Type |
| $ customerID : chr "5575-GNVDE" "7795-CFOCW" "9237-HQITU" ...  $ gender : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 2 1 ...  $ SeniorCitizen : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  $ Partner : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...  $ Dependents : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 2 1 ...  $ tenure : int 34 45 2 8 22 28 25 52 71 21 ...  $ PhoneService : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...  $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 2 2 2 2 2 ...  $ Contract : Factor w/ 3 levels "Month-to-month",..: 2 2 1 1 1 1 1 2 ...  $ PaperlessBilling : Factor w/ 2 levels "No","Yes": 1 1 2 2 2 2 2 1 1 2 ...  $ PaymentMethod : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 1 3 ...  $ MonthlyCharges : num 57 42.3 70.7 99.7 89.1 ...  $ TotalCharges : num 1890 1841 152 820 1949 ...  $ Churn : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 2 1 1 1 1 ... | Nominal  Nominal  Nominal  Nominal  Numerical  Nominal  Nominal  Nominal  Nominal  Nominal  Numerical  Numerical  Nominal |

Tabel 1 Data type for every variable after the conversion into factors

To ensure data quality, the program should check whether there are missing values and eliminate unnecessary variables that will not be used in the model. A preliminary check has found that there are two missing values in TotalCharges. These two observations can be considered insignificant compared to the total of 2,114 observations, so the analysis will eliminate both and get the total of 2,112 observations.

The next part is to eliminate unnecessary variables. In this case, the customerID has nothing to do with determining whether a customer will churn or not. Therefore, it can be eliminated. The program can also check if there is a correlation between numeric variables. If there is a strong correlation between two numeric variables, the analysis can eliminate one of them in the logistic regression analysis. However, this project will include all numeric variables first before making adjustments. Thus, the project will analyze 2,112 observations, using 12 predictor variables to predict the target variable Churn.

## **Exploratory Data Analysis (EDA)**

To better understand the dataset, the project will execute a thorough exploratory data analysis (EDA). The EDA will be executed for categorical variables first and then numerical variables. For categorical variables, the EDA will provide bar graphs for each variable and with an overlay to target variable Churn. The bar graph will also provide the normalized view to better understand how a variable may be useful to predict Churn. Contingency tables will also be provided to better visualize the categorical variables and how it relates to Churn.

For numerical variables, the descriptive statistics will be provided along with a histogram that shows how the variables may relate to Churn. The EDA will look for outliers if there are any. The project will also examine the correlation between these numerical variables.

### **EDA for categorical variables**

The bar graphs that show how each categorical variable relates to the target variable Churn are shown in **Appendix 1**, along with the normalized bar graphs and contingency tables. There are multiple variables that may prove to be a significant predictor of Churn. For instance, in the variable Contract, a month-to-month contract may have a higher chance of churning, compared to one year and two years contract. It can be shown by the overlay bar graph and contingency table below.

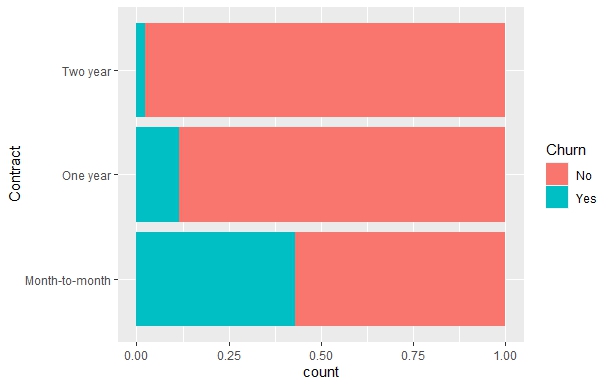
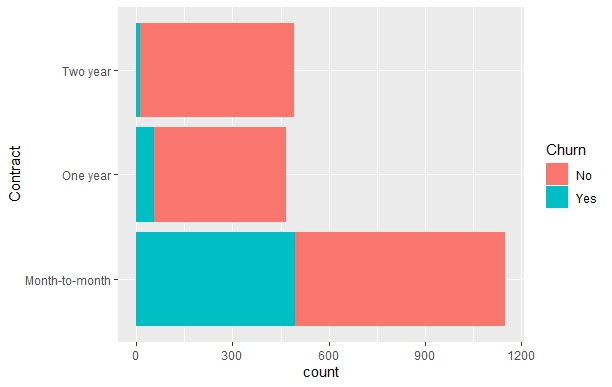


Figure 1. Bar graph overlay for variable Contract and Churn, and its normalized version.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Month-to-month | One year | Two year | Total |
| No | 656 | 413 | 482 | 1551 |
| Yes | 496 | 54 | 11 | 561 |
| Total | 1152 | 467 | 493 | 2112 |

Table 2. Contract vs Churn Contingency Table

While the dataset has more observations for month-to-month contracts, the normalized bar graph still shows that almost half of the customers with a month-to-month contract have churned. It can be understood logically that customers with a month-to-month contract have lower switching costs to leave the company’s service and choose another company. It makes the decision to churn easier for these customers compared to those with one-year or two-year contracts. This logical reasoning can become initial proof that the difference is not due to random chances.

Executing the same analysis for each variable, the project obtained a bar graph overlay for each variable related to Churn, the normalized bar graph, and the contingency table, which are shown in **Appendix 1**. Using these graphs, EDA can see some potential variables to predict Churn, such as InternetService and PaymentMethod. There are also variables showing only moderate differences, such as SeniorCitizen, Partner, Dependent, and PaperlessBilling. While variable gender and PhoneService may not provide significant differences in Churn. The logistic regression will have a deeper analysis of how the variable Contract may or may not be useful to predict Churn.

### **EDA for numerical variables**

The descriptive statistics for the numerical variables of the dataset are shown in **Appendix 2**, along with the histogram overlay and its normalized version. The summary of the numerical variables shows that there is no outlier in the dataset. The histogram overlay and the normalized version can provide a preliminary analysis of how each numerical variable may influence Churn. For instance, the histogram overlay for variable tenure and its normalized version is shown below.

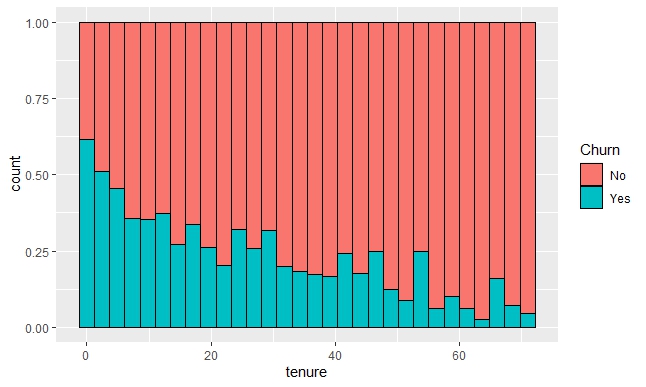
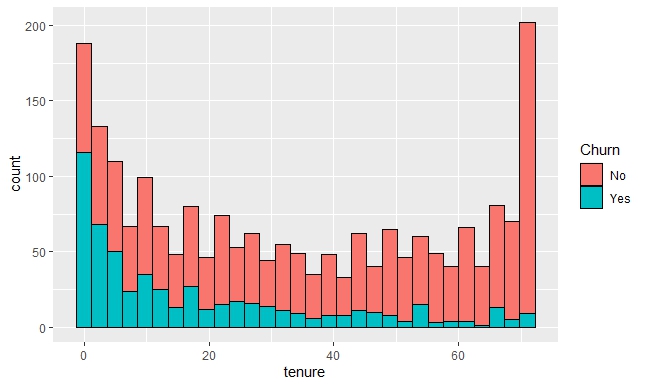


Figure 2. Histogram overlay for variable tenure and Churn, and its normalized version.

By default, R divides the histogram into 30 bins. While there are a different number of observations in each bin, the normalized version of the histogram shows that customers with shorter tenure have a higher chance to churn. It is logical that customers who perceive the company’s service positively will stay in the company’s service for a longer time, and hence they have a lower churn rate. Newer customers may find other companies' offers more appealing and look at the churn decision as easier than older customers. This logical reasoning can become initial proof that the difference is not due to random chances.

Histograms for two other numerical variables are shown in **Appendix 2**, along with the summary of the data. Both histograms show that the other two numerical variables MonthlyCharges and TotalCharges may not be as significant as variable tenure. The logistic regression will have a deeper analysis of how each of these numerical variables may or may not be useful to predict Churn.

Aside from how each numerical variable relates to churn, a correlation between these numerical variables can also be examined to better understand the data. Logically, tenure and TotalCharges should have a strong positive association, since customers with longer tenure paid more cumulatively than those with shorter tenure. The correlation for each variable can be shown through the correlation matrix below.

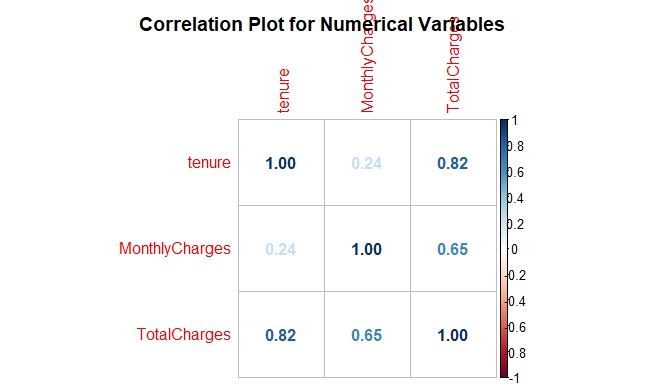


Figure 3. Correlation matrix for numerical variables tenure, MonthlyCharges, and TotalCharges.

As predicted above, tenure and TotalCharges have a strong positive association with a correlation value of 0.82. There is also a strong correlation between MonthlyCharges and TotalCharges with 0.65 correlation values. It may suggest that logistic regression will be sufficient using only variable tenure and MonthlyCharges. However, the project will incorporate all numerical variables into the analysis first and see how the model accuracy will be. The model can then be adjusted using less number of variables by eliminating variables that have low significance in predicting Churn.

# **Results and Analysis**

The model to predict Churn will be developed using logistic regression and a decision tree. Both are also the most common technique used by data analysts to model binary response prediction since both are well-developed and well-studied methods. In this case, the binary response would be Yes or No in the variable Churn, describing whether a customer has churned or not.

## **Data Partition**

Both techniques are supervised methods, which means that the dataset should be divided into a training dataset and a testing dataset. The training dataset will be used to develop the model and the test dataset will be used to test the accuracy of the model. Using R, the data partition will be created using random seed value 1027, so that users of the model can recreate the result.

## **Logistic Regression**

The analysis will provide three logistic regression models. The first model will use all of the variables to predict Churn and evaluate which variables have a high significance in the model. The second model will use variables with higher significance that was found in the first model to predict Churn. The third model will not use any variables to predict Churn and let it be decided by random. The analysis will create a comparison of the three models using the analysis of variance (ANOVA). It will also compare the accuracy of the model using the test dataset.

Using *glm* function in R with the family set to binomial, the logistic regression summary can be obtained and it is shown below in figure 4. The coefficients table shows the coefficient for each variable, as well as its standard error, z value, and p-value. The stars at the end of each row depict the significance of the variables. Variables with three stars have higher significance than variables with two or lower numbers of stars. The analysis found that there are various significant variables, such as tenure, internet service, contract, payment method, and total charges. Almost all of these variables have a significant coefficient estimate value and lower than a 0.05 p-value which means that it is statistically significant. Variable total charges have lower estimate value since the variable contains large numerical values (further analysis can standardize variables with large numerical values for a better depiction of their significance).

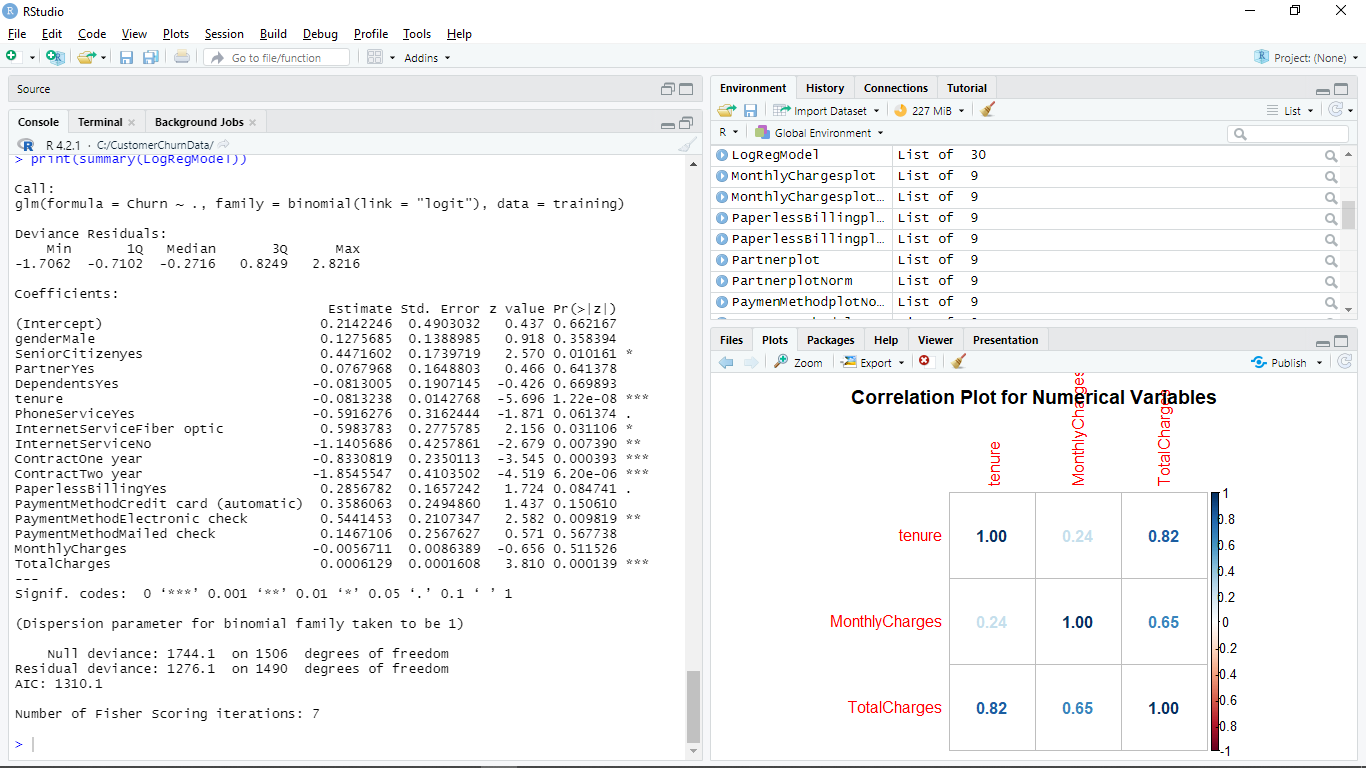
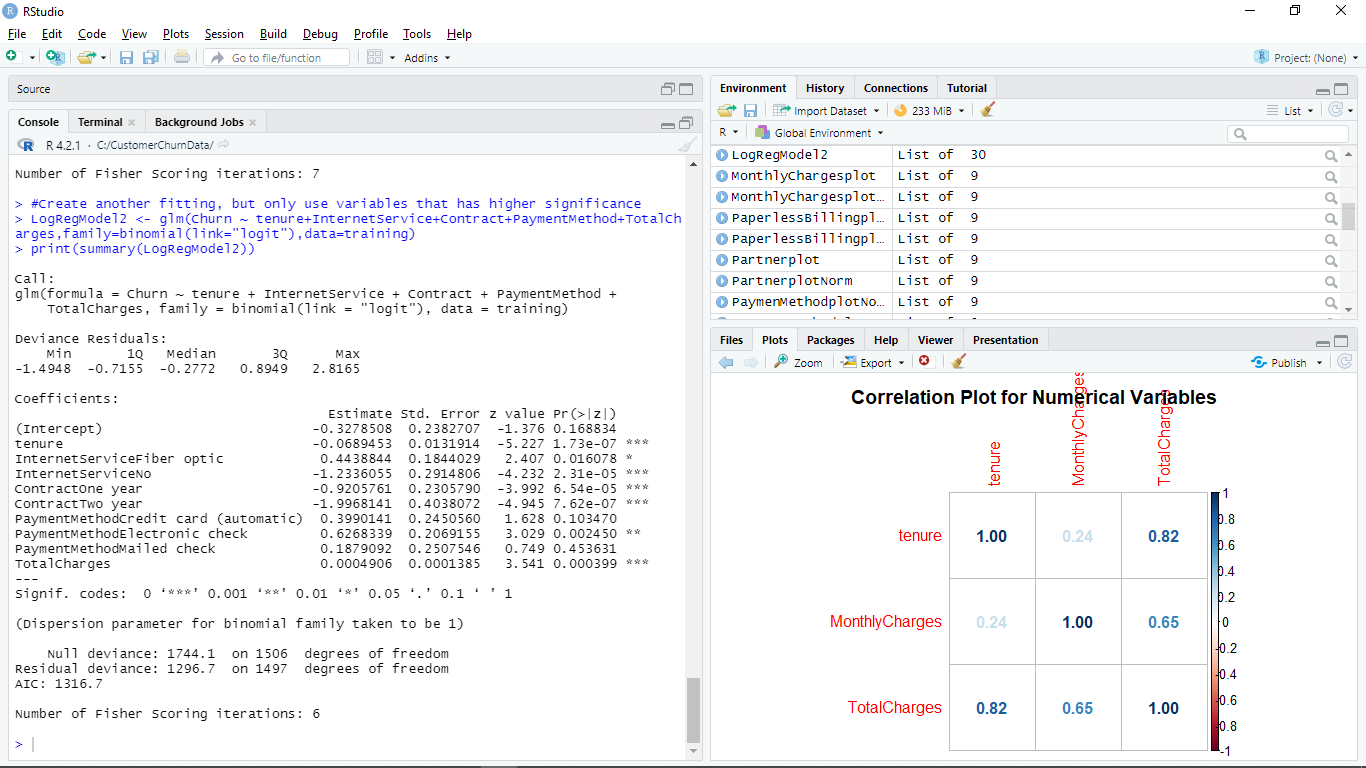


Figure 4. the result of the first logistic regression for training dataset in R.

Logistic regression will provide a linear model to predict the log odds of churn, with the estimated value above being the coefficient value of the variable. By solving the logit, the model can provide the churn odds prediction equation. Thus the first model can be written as an equation below.

The same process can be executed to create the second model. The model will use five variables that are the most significant from the first model, which are tenure, InternetService, Contract, PaymentMethod, and TotalCharges. The logistic regression summary result is shown below.

  
Figure 5. the result of the second logistic regression for training dataset in R.

Notice that most of the variables are statistically significant with a p-value less than 0.05. Using the same analysis as the previous model, the linear logit equation can be written using the estimated value above as the coefficient for each variable. Thus, the equation can be written as follow.

### **ANOVA Test**

ANOVA for the first model and second model shows a significant p-value of 0.0043 which is lower than 0.05. It means that the second model provides less variance and hence, it is significantly better in modeling the training data to predict Churn. ANOVA of the first and second models also shows a significant p-value when each of the models is compared to a model without predictor variables. It means that both are significantly better models than random chances.

The ANOVA will provide the model with less variance when modeling the training dataset. However, there is still a possibility of overfitting, in which a model can not be consistently good in predicting the target variable. In supervised methods, the model should also be tested using a test dataset to determine the accuracy and see whether overfitting exists.

### **Model Accuracy Test**

The accuracy of the first and second models above can be tested using the test dataset. Using R, the test will use variable values of each customer in the test dataset and predict whether the customer will churn or not. Then, the program will compare the prediction and the actual data to determine the accuracy level of the model.

For example, the first observation in the test dataset is a female with no partner, no dependents, and not a senior citizen, also with a tenure of 8 months, using phone services, fiber optic internet service, and month-to-month contract, using paperless billing, have an electric check as a payment method, with 99.65 monthly charges, and 820.5 total charges. Using these data as an input for the first model, the odd of the customer churning can be written as follow.

The model predicts that the customer has a 0.6 chance of churning. Using 0.5 as a threshold, the model will classify this customer as churning and put “Yes” in the prediction. The testing dataset already has the churn data of the customer, and it turns out that the customer has churned. Therefore, in this case, the model predicted the right outcome for the customers.

Iterating this process for each customer in the testing dataset, the program will get the overall accuracy of the model. For the first model, the accuracy is 0.809 while for the second model, the accuracy is 0.803. The confusion matrix for both models is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| Model 1 Confusion matrix | | Predicted data | |
| No | Yes |
| Actual data | No | 412 | 32 |
| Yes | 83 | 78 |

|  |  |  |  |
| --- | --- | --- | --- |
| Model 2 Confusion matrix | | Predicted data | |
| No | Yes |
| Actual data | No | 407 | 37 |
| Yes | 82 | 79 |

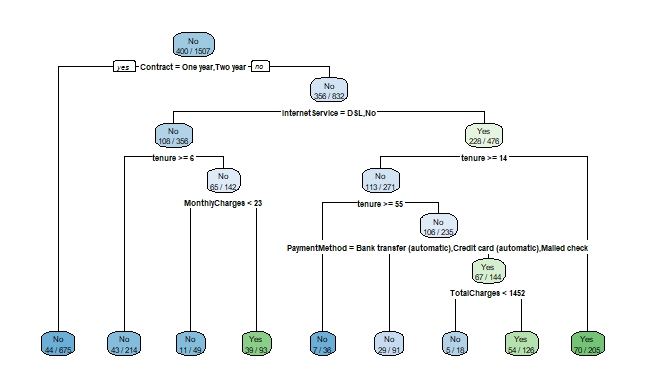
Figure 4. the result of the first logistic regression for training dataset in R.

The matrix shows that the first model accurately predicts 412 cases of not churning and 78 cases of churning customers. The accuracy calculated earlier can be obtained by dividing the number of accurate predictions by the number of inaccurate predictions. For instance, the first model accuracy is 0.809 = (412+78)/(412+78+32+83). The same calculation can be provided for the second model, using the second confusion matrix, which will result in an accuracy value of 0.803.

The result shows that the first model did a slightly better job in predicting the test dataset. While the second model is better in terms of variance of the training dataset, its performance is not better than the first model. However, the difference is small. Hence, the project may choose either model since both have a similar level of accuracy. For this project, the chosen model will be the second model, since it provides a simpler model using fewer variables that are proven significant. Providing a larger dataset to test can be a good way to show which model is better in accurately predicting customer churn.

## **Decision Tree**

Using R, the decision tree can be established from the training dataset. The result of the decision tree diagram can be shown below. The root node used the variable Contract, which means that the variable is the most important one in predicting customer churn. If a customer has one year or two years contract, the model will directly categorize the customer as not churning. While for customers with a month-to-month contract, the decision should be followed by the next branch, which is InternetService, and then tenure.

 Figure 5. the decision tree model extracted from R.

The model consists of 9 leaf nodes and 7 branch nodes. Note that the model only uses 6 variables out of the 12 variables in the dataset. Here, the most important variables to predict churn are Contract, InternetService, tenure, MonthlyCharges, PaymentMethod, and TotalCharges. This result is similarly consistent with the second model of the logistic regression model which only uses the same 5 most important variables. Only here does it includes variable MonthlyCharges.

|  |  |  |  |
| --- | --- | --- | --- |
| Decision Tree Confusion matrix | | Predicted data | |
| No | Yes |
| Actual data | No | 384 | 57 |
| Yes | 60 | 104 |

Figure 6. confusion matrix of how the prediction accuracy of the model in R.

The confusion matrix for the decision tree is shown above. The matrix shows that the decision tree model accurately predicts 384 cases of not churning and 104 cases of churning customers. The accuracy of this model can be obtained by dividing the number of accurate predictions by the number of inaccurate predictions. Using the same calculation as above, the accuracy of this model is obtained as 0.806. The decision tree model is proven to have a similar accuracy level to the logistic regression model with about 80% accuracy.

# **Discussion**

The application of logistic regression and decision tree technique primarily found that several variables are the most important to predicting customer churn. Those variables are Contract, tenure, InternetService, PaymentMethod, and TotalCharges. It is important to find logical reasoning for how these variables can become the most significant variables that determine churn. The discussion will be elaborated on in the next section.

## **Contract and tenure**

The most important variables in all models are related to how long the customers have used the company’s service. Customers with month-to-month contracts and shorter tenure tend to churn since they have low switching costs if they want to switch and use a competitor’s service. This issue will be increasingly vital in an industry with intense competition. Without a clear strategy, the company will fight and struggle for customer acquisition instead of customer retention.

From the second model of logistic regression, the increase in tenure will decrease to log odds of the customer churning by 0.069. The longer the tenure of a customer, the less likely for the customer to churn. The second model also shows that customers with a month-to-month contract have -0.92 fewer log odds of churning compared to customers with one year contract, and -1.99 less likely when compared to customers with two years contract. This result points out the importance of managing newer customers so that they have a lower churn rate.

## **Internet Service**

Customers with DSL or no internet service tend to stay, compared to customers with fiber optic internet service. It is logical that customers with only phone service and DSL, without fiber optic internet service, may have no urgency to churn and use similar services from other companies. However, the higher churn rate from the customers using fiber optic service may provide a sign that there are competitors with better fiber optic internet services. The company may want to evaluate its fiber optic internet service and how it is perceived by the customers compared to similar services from the competitor. The company may also want to include the complaint rate in the dataset.

## **Payment Method**

It is not surprising that customers’ payment method is also considered a significant factor in the model. The EDA process found out that the company has more customers with the electronic check payment method. It is an alarm for the company that customers with electronic check payment methods have a higher chance of churning compared to customers with other payment methods. Typically, electronic check payment is relatively less expensive than other payment methods (Kagan). Thus, customers with electronic check payment methods will also have lower switching costs to jump over to competitors’ services. The company may also want to provide a better payment experience or encourage customers to use other payment methods.

## **Predictive Model to Be Selected**

The decision to choose which predictive model to use is determined by various criteria, such as the model’s accuracy, bias and variance, and ease of use. The first and second logistic regression models as well as the decision tree model have similar accuracy of 80%. However, ANOVA shows that the second logistic regression model is significantly better than the first model. It also uses fewer variables and focuses on the more important variables. Hence, the option is only between the second logistic regression model and the decision tree model.

The decision tree is relatively easier to understand and easier to execute. However, there are several limitations of the decision tree. The first is that it tends to oversimplify the impact of a variable. For instance, in this case, a customer with one year and two years contract will be directly categorized as not churning customer. In reality, customers with one-year and two years contracts should still have the probability of churning due to various other factors. The second limitation is that the decision tree may have potential misinterpretation, especially for variables with high correlation.

Meanwhile, the logistic regression is more difficult to understand and it requires some statistical and mathematical knowledge to correctly interpret the result. It can lead to wrong interpretation and causes an overall wrong model application. However, if it can be executed properly, it usually provides a better depiction of how a variable will influence the target variable.

Bock stated that for an academic study aiming to provide solid conclusions about a predictive model, logistic regression is typically better than a decision tree, but, if the goal is to describe the data or solely to create a prediction, then a decision tree is usually better. In this case, logistic regression can provide more insights for further study. Thus, the recommendation is to use the second logistic regression model. However, if the company has more computing power, it can keep both model, logistic regression, and decision tree for future prediction problems.

# **Conclusion and Recommendation**

The first recommendation is for the company to give more focus on the customers with shorter contracts and tenure. The company can create a strong marketing campaign that can encourage newer customers to stay in the company’s services for a longer period of time. The company can also provide better incentives that can encourage customers to switch from a month-to-month contract to one year or two years contract. The incentive can be provided as a charge discount or free additional services. The company should strengthen its competitive advantage so that it can have a strong position in the industry, instilling the perception of higher switching costs.

The second recommendation is to focus on customers with fiber optic internet service and electronic check payment methods. The churn rate of the customers with fiber optic internet is high, which may signal that the company’s fiber optic internet service may not the best in the industry. The company can also encourage customers to switch from electronic checks, credit cards, and mailed check payment methods to automatic bank transfers which will result in the lowest coefficient for InternetService variable. Evaluating customers' experiences related to these two variables is essential.

The next recommendation is to use the second model of logistic regression to predict customer churn until a better model can be developed. The model uses fewer variables than other models, and hence it involves simpler calculations. It is better than the decision tree model in depicting the reality of variable impact to churn variable. However, the company should take extra care and provide thorough calculations in establishing the model, since the technique is more difficult to interpret and prone to mistakes.

Another recommendation is to increase the quality of the model by adding more data into the training and testing data set. The company can also consider using more sophisticated machine learning techniques to better predict churn. Nevertheless, with an accuracy as high as 80%, the model established in this project is sufficient.

# **References**

Bock, Tim. “Decision Trees Are Usually Better Than Logistic Regression.” *Displayr*, 9 June 2021, [www.displayr.com/decision-trees-are-usually-better-than-logistic-regression](http://www.displayr.com/decision-trees-are-usually-better-than-logistic-regression).

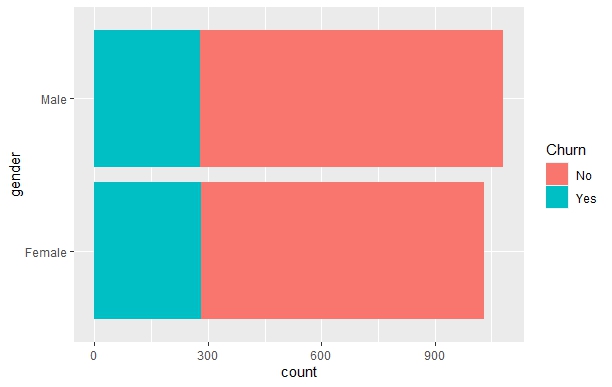
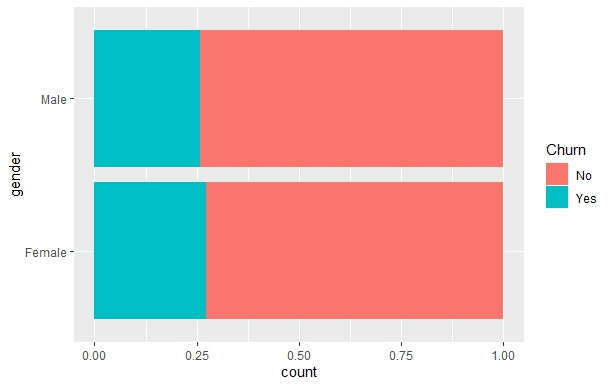
Hughes, A. “Churn Reduction in the Telecom Industry.” *Database Marketing Institute*, Mar. 2010, [www.dbmarketing.com/2010/03/churn-reduction-in-the-telecom-industry](http://www.dbmarketing.com/2010/03/churn-reduction-in-the-telecom-industry).

Kagan, Julia. “Electronic Checks: Understanding the Basics.” *Investopedia*, 19 May 2020, [www.investopedia.com/terms/e/electroniccheck.asp](http://www.investopedia.com/terms/e/electroniccheck.asp).

# **Appendix 1**

**Exploratory Data Analysis for Categorical Variables**

1. **Gender**

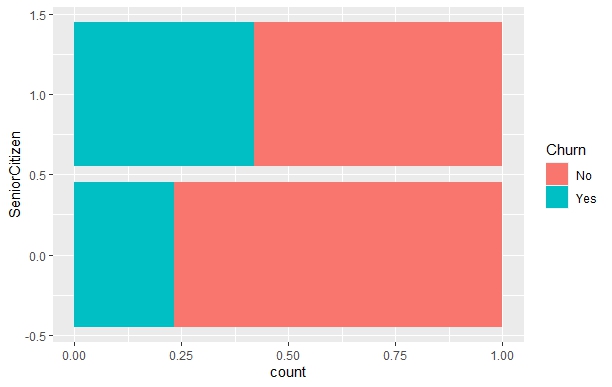
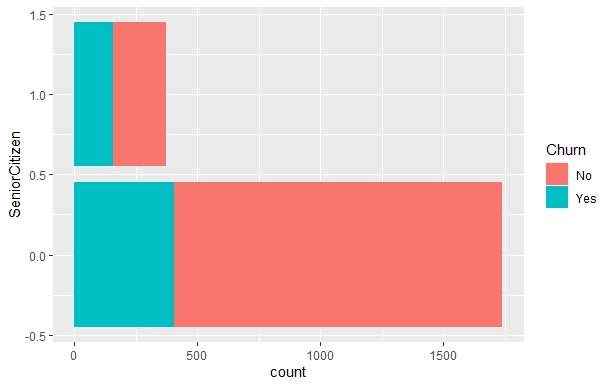
Female Male total

No 750 801 1551

Yes 281 280 561

total 1031 1081 2112

1. **Senior Citizen**



Senior Citizen

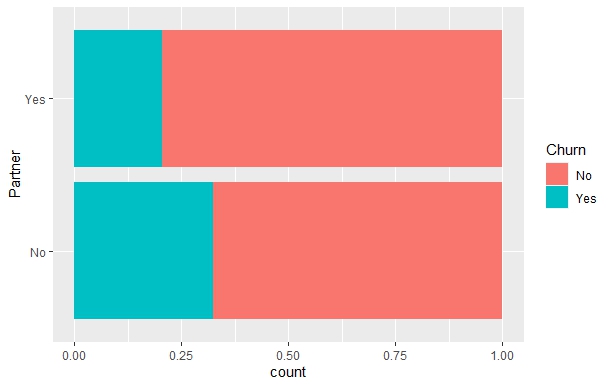
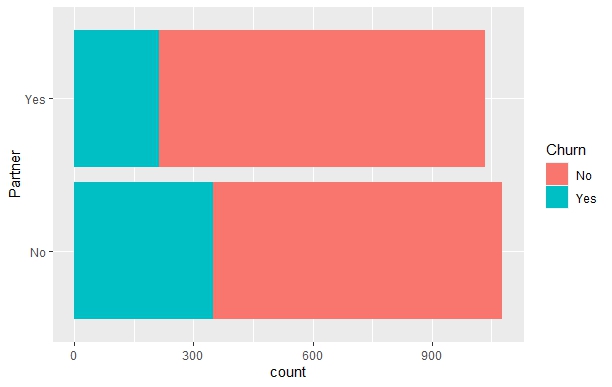
0 1 total

No 1335 216 1551

Yes 404 157 561

total 1739 373 2112

1. **Partner**



Partner

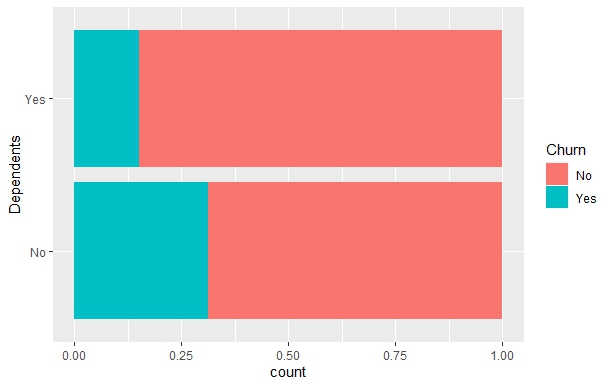
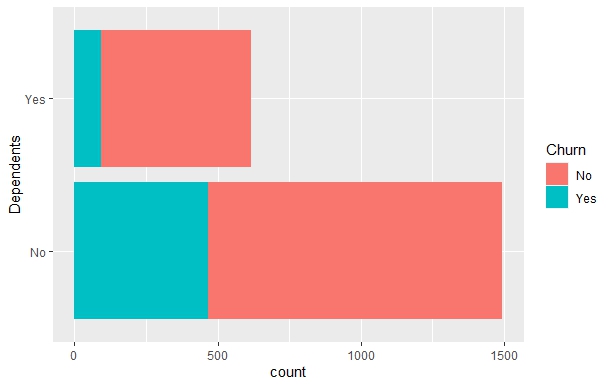
No Yes total

No 729 822 1551

Yes 349 212 561

total 1078 1034 2112

1. **Dependents**



Dependents

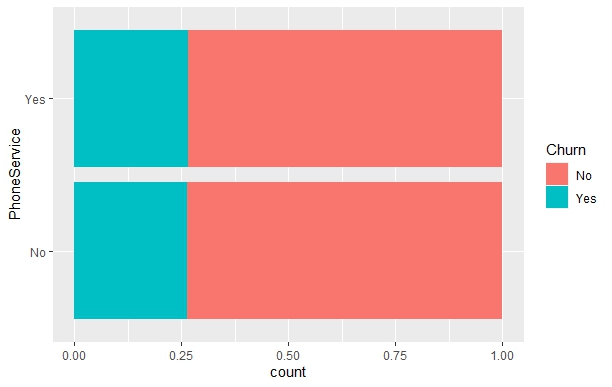
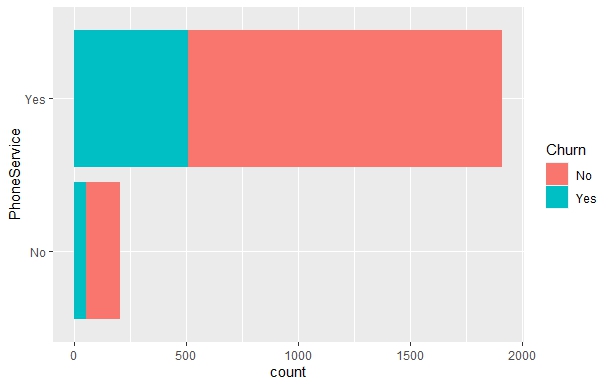
No Yes total

No 1027 524 1551

Yes 468 93 561

total 1495 617 2112

1. **PhoneService**



PhoneService

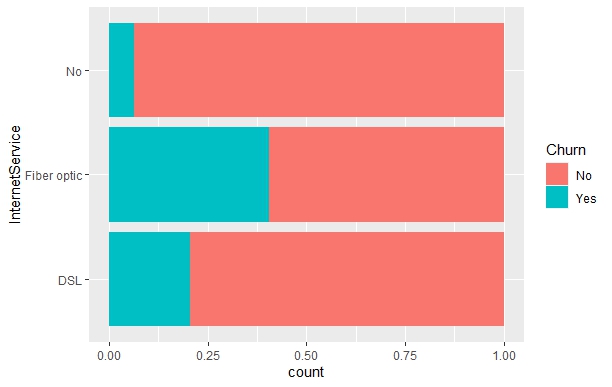
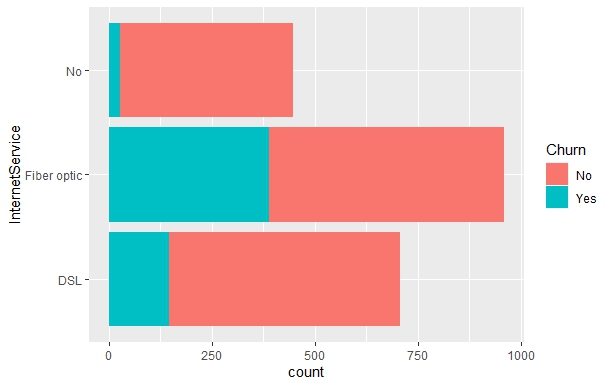
No Yes total

No 149 1402 1551

Yes 53 508 561

total 202 1910 2112

1. **InternetService**



InternetService

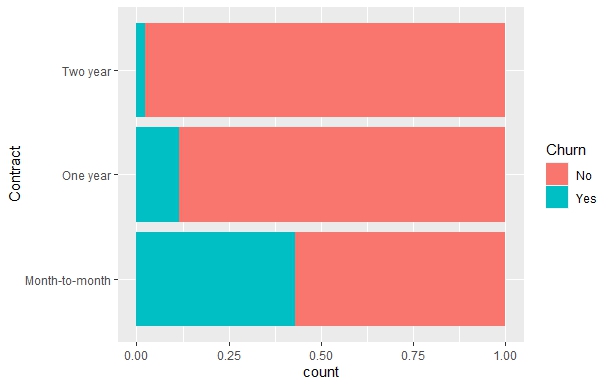
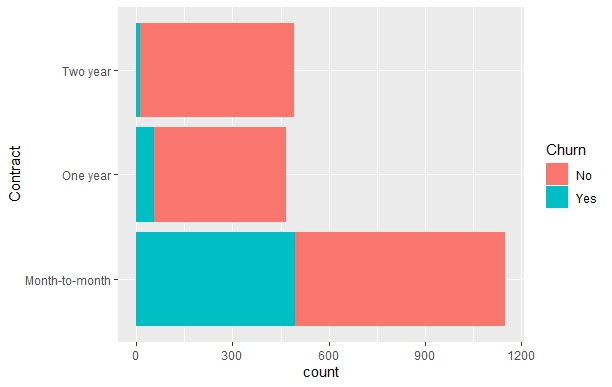
DSL Fiber optic No total

No 561 571 419 1551

Yes 145 388 28 561

total 706 959 447 2112

1. **Contract**



Contract

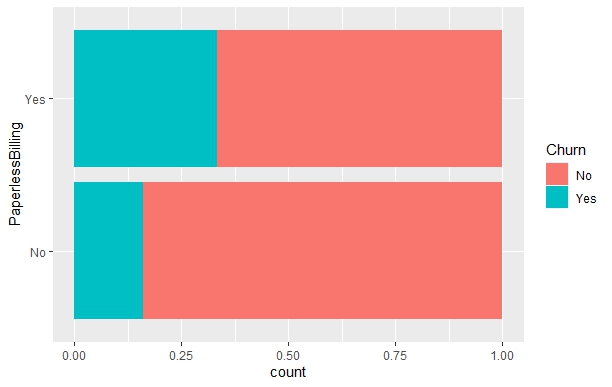
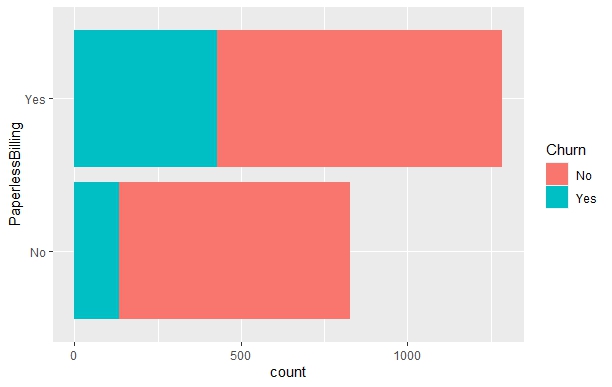
Month-to-month One year Two year total

No 656 413 482 1551

Yes 496 54 11 561

total 1152 467 493 2112

1. **PaperlessBilling**



PaperlessBilling

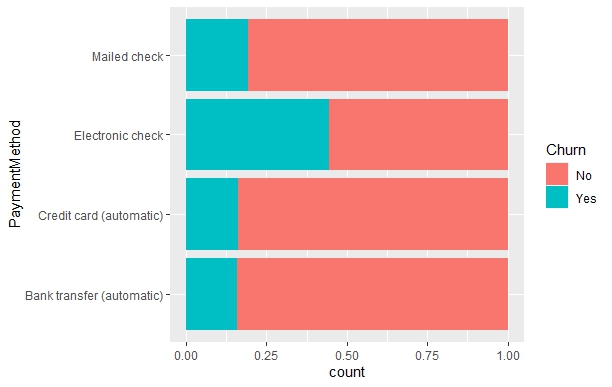
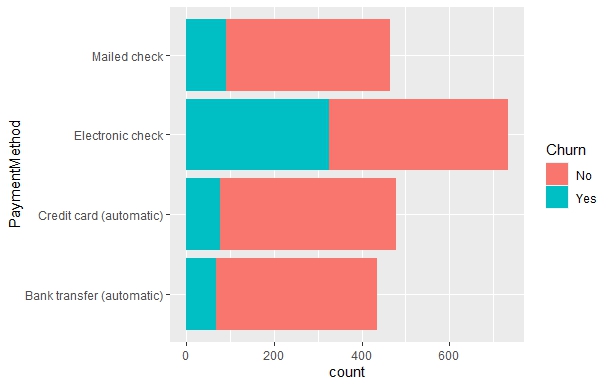
No Yes total

No 694 857 1551

Yes 133 428 561

total 827 1285 2112

1. **PaymentMethod**



PaymentMethod

Bank transfer (auto) Credit card (auto) Electronic check Mailed check total

No 367 402 408 374 1551

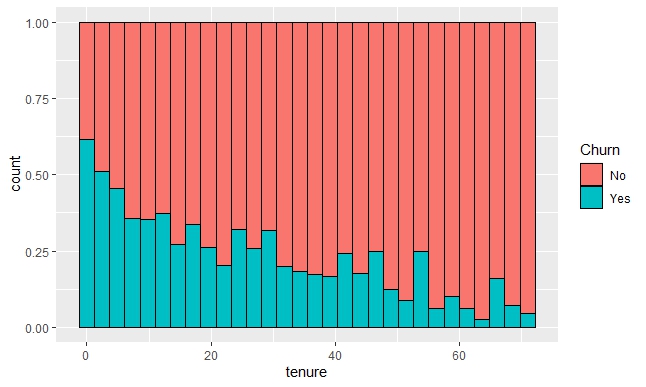
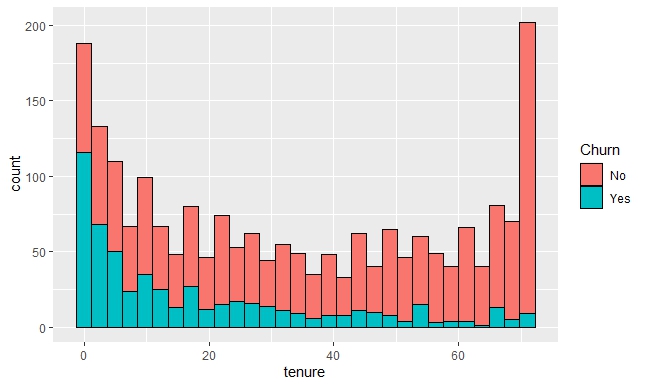
Yes 69 77 325 90 561

total 436 479 733 464 2112

# **Appendix 2**

**Exploratory Data Analysis for Categorical Variables**

1. **tenure**

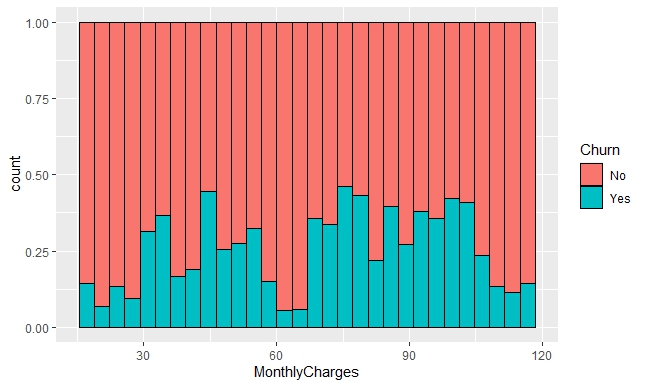
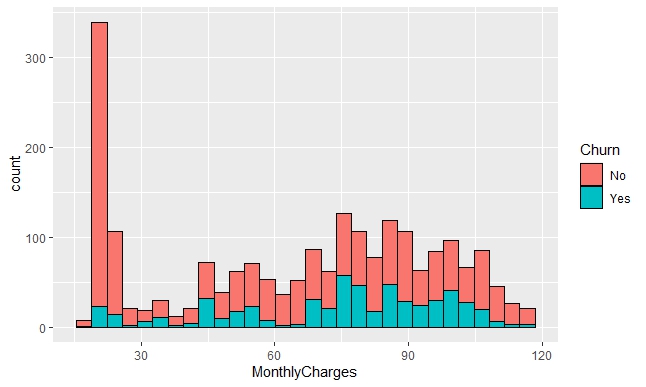


tenure

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 9.00 30.00 33.08 56.00 72.00

1. **MonthlyCharges**

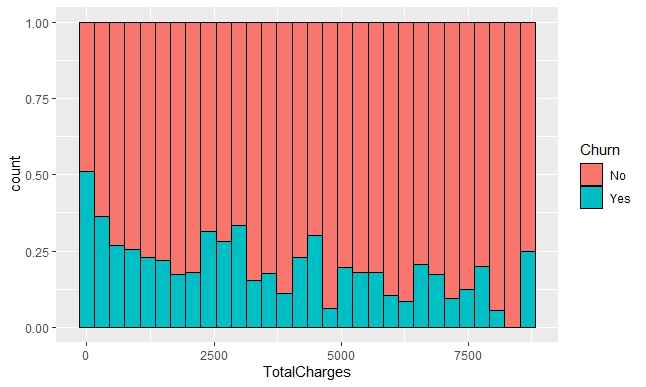
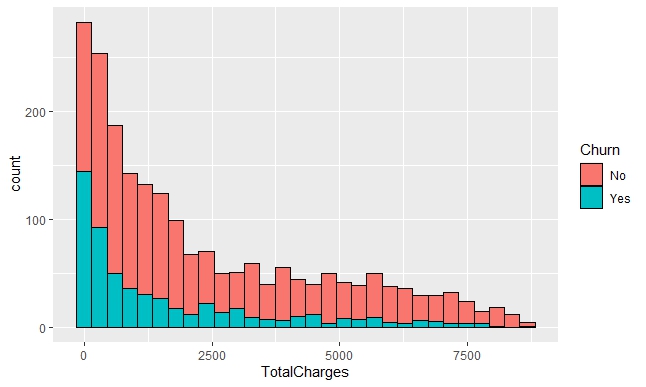


MonthlyCharges

Min. 1st Qu. Median Mean 3rd Qu. Max.

* 1. 38.21 71.30 65.43 89.96 117.80

1. **TotalCharges**



TotalCharges

Min. 1st Qu. Median Mean 3rd Qu. Max.

18.8 434.0 1490.1 2337.3 3921.9 8684.8

talhlyCharges 733nd Tota is sufficient.t servicgthe cha