**D208: Predictive Modeling**

**Performance Assessment**

**Task 2**

Logan Rosemeyer

Western Governor's University

D208: Predictive Modeling

Dr. Eric Straw

January 2, 2023

## Contents

A1: Research Question 3

A2: Objectives and Goals 3

B1: Summary of Assumptions 3

B2: Tool Benefits 3

B3: Appropriate Technique 3

C1: Data Goals 3

C2: Summary Statistics 3

C3: Steps to Prepare the Data 4

C4: Visualizations 5

C5: Prepared Data Set 7

D1: Initial Model 14

D2: Justification of Model Reduction 15

D3: Reduced Multiple Regression Model 15

E1: Model Comparison 15

E2: Output and Calculations 16

E3: Code 18

F1: Results 18

F2: Recommendations 19

G: Panopto Demonstration 19

H: Sources of Third-Party Code 19

I: Sources 19

## A1: Research Question

My research question is: what factors influence whether a customer churns?

## A2: Objectives and Goals

The goal of the data is to see what factors influence whether a customer churns. This will allow us to see which factors play into a customer churning. By knowing which factors play into a customer churning, we can better predict which customers will churn and do our best to prevent it.

## B1: Summary of Assumptions

There are a few assumptions of a logistic regression model. The first is that there is little or no multicollinearity within the independent variables. Another assumption is that logistic regression assumes linearity of independent variables and log odds. Another assumption is that you have a large sample size to run a logistic regression model. (Assumptions of Logistic Regression, n.d.)

## B2: Tool Benefits

I chose to use R for all phases of my analysis. The reason for this is because many functions I will need to use are already built into R as opposed to Python. I had to run a logistic regression, and R already has a stats package with the glm function loaded into it. Another reason that I used R is because it creates better graphics and data visualizations. Although this task is about logistic regression, I had to make a numerous amount of univariate and bivariate visualizations. This is much easier in R rather than Python.

## B3: Appropriate Technique

Logistic regression is the appropriate technique to analyze whether a customer churns or not. Churn is a categorical variable, so logistic regression is the appropriate technique.

## C1: Data Goals

My first goal is to make a data frame of variables that I might use. This includes most of the variables except the last eight (survey responses), unique IDs, and some location information. After I did this, my second goal was to make sure my data would give me accurate information. I looked for duplicates and null values. After that, I looked for any outliers in the data that I would need to treat.

## C2: Summary Statistics

The population variable had a median of 2910 and a mean of 9757 which is skewed right. The area attributes are uniformly distributed, with all values appearing around 3330 times. The children variable had a median of 1 and a mean of 2.1, which is skewed right. The income variable had a median of 33170.6 and a mean of 39806.9. The marital variable is uniformly distributed, with all values appearing around 2000 times. The gender variable has a similar amount of female and male values, and 231 nonbinary. There are 7350 customers who did not churn, and 2650 who did. The outage per week variable had a median of 10 and a mean of 10, which is normally distributed. The email variable had a median of 12 and a mean of 12, which is normally distributed. The contacts variable had a median of 1 and a mean of 1, which is normally distributed. The yearly equipment failure variable has a median of 0 and a mean of 0.4 which is skewed right. There are 8321 customers who do not consider themselves a techie, and 1679 who do. 5456 customers had a month-to-month contract, 2102 had a one year contract, and 2442 had a two year contract. 4408 customers had fiber optic internet, 3463 had DSL internet, and 2129 had no internet. 9067 customers had a phone, while 933 did not. 5392 customers did not have multiple lines, and 4608 had multiple lines. 6250 customers did not have a technical support add-on, and 3750 did. 5071 does not have streaming tv, while 4929 do. 5110 do not have streaming movies, and 4890 do. 5882 customers have paperless billing, while 4118 do not. 3398 customers pay with electronic checks, 2290 pay with a mailed check, 2229 pay with an automatic bank transfer, and 2083 pay automatically with a credit card. The tenure variable had a median of 35.4 and a mean of 34.5, which is normally distributed. The monthly charge variable had a median of 167.5 and a mean of 172.6, which is slightly skewed right. The bandwidth per year variable has a median of 3279.5 and a mean of 3392.3, which is normally distributed.

## C3: Steps to Prepare the Data

The first thing I did was to narrow down the columns I might use for the data.

```
    log_churn_data <- churn %>% select(-CaseOrder:-Lng, -Item1:-
    Item8, -TimeZone, -Job, -Port_modem, -Tablet, -OnlineSecurity, -
    OnlineBackup, -DeviceProtection)
```

After this I checked for duplicates and null values.

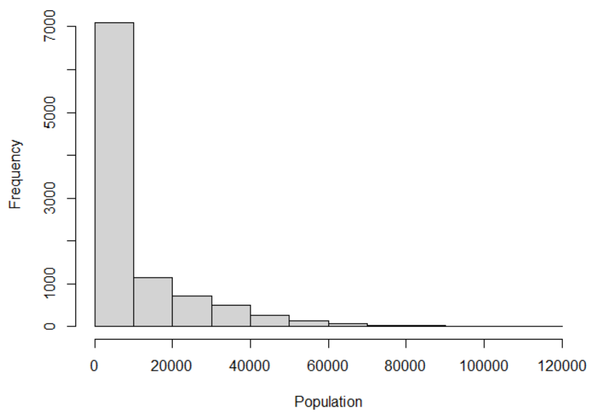
```
sum(duplicated(log_churn_data))

colSums(is.na(log_churn_data))
```

Then I checked for outliers of numeric values.

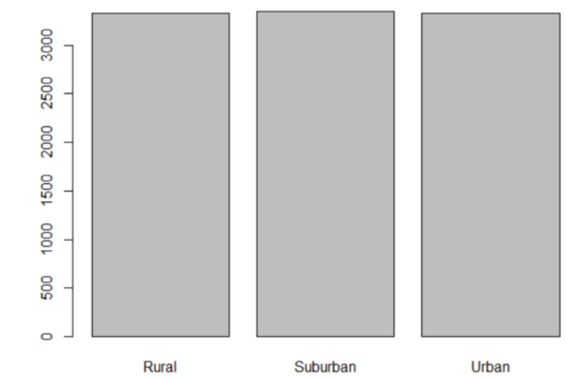
```
log_churn_data_zscore <- scale(log_churn_data %>% select(Population,
Children, Age, Income, Outage_sec_perweek, Email, Contacts,
Yearly_equip_failure, Tenure, MonthlyCharge, Bandwidth_GB_Year))

head(log_churn_data_zscore)

log_churn_data_zscore <- as.data.frame(log_churn_data_zscore)

hist.data.frame(log_churn_data_zscore)
```

# C4: Visualizations

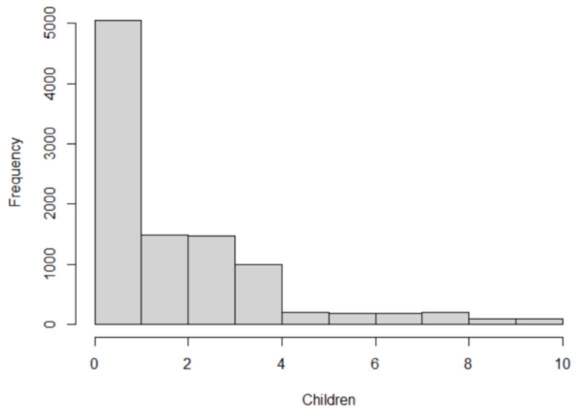
**Population**



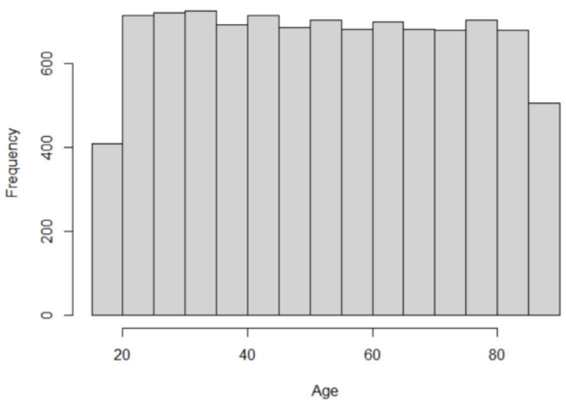
**Area**



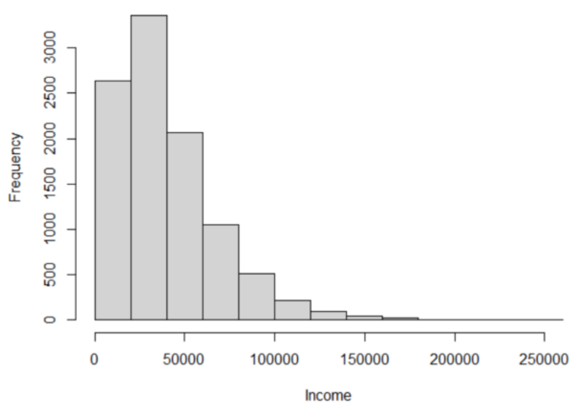
**Children**



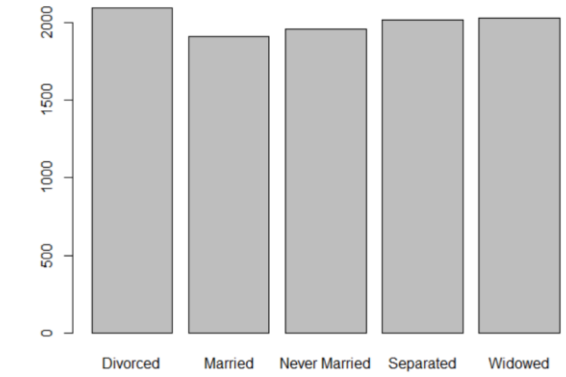
**Age**

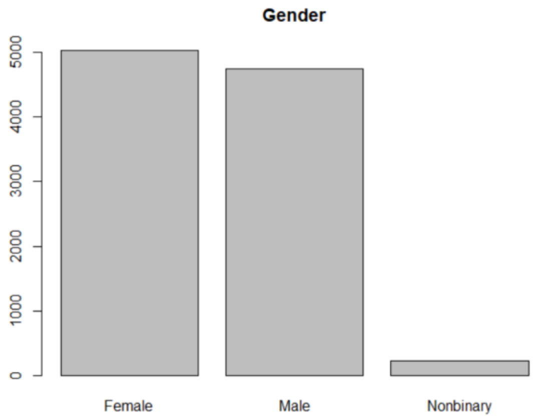


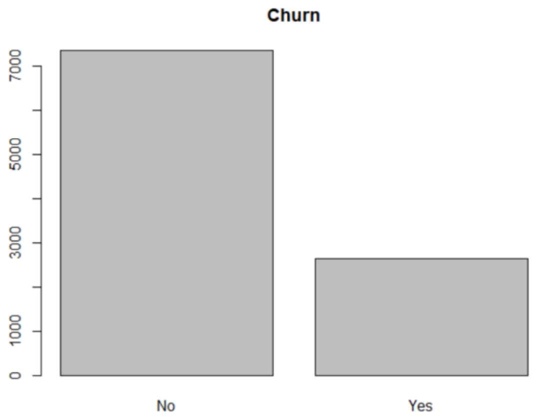
**Income**

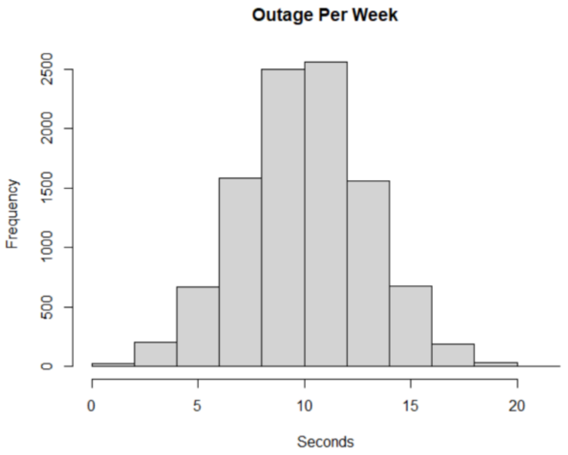


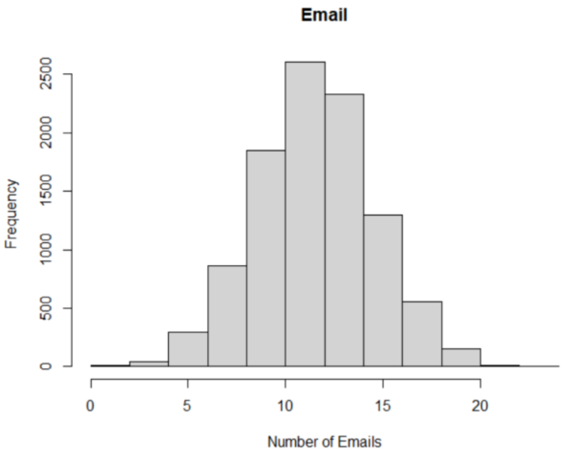
**Marital Status**

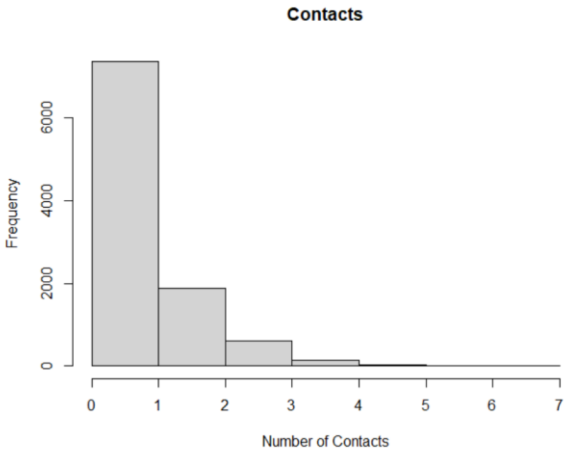


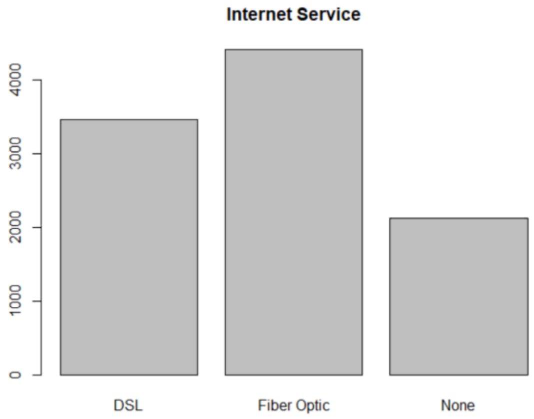

Gender
0
1000
2000
3000
4000
5000
Female
Male
Nonbinary

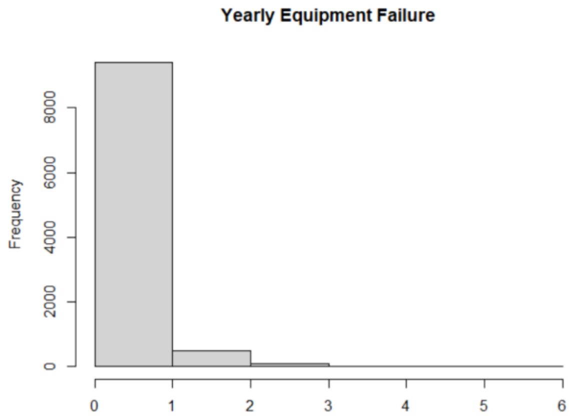

Churn
0
1000
3000
5000
7000
No
Yes


Outage Per Week
Frequency
0
500
1000
1500
2000
2500
0
5
10
15
20
Seconds

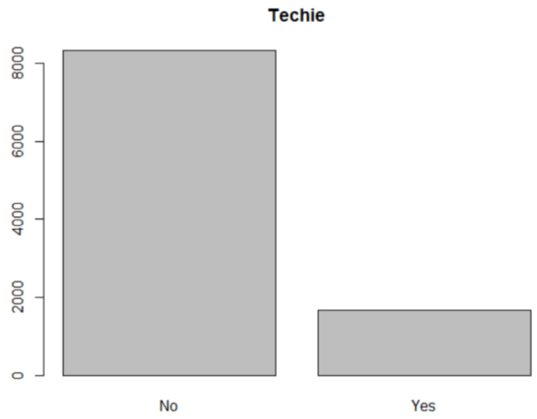

Email
Frequency
0
500
1000
1500
2000
2500
0
5
10
15
20
Number of Emails


Contacts
Frequency
0
2000
4000
6000
0
1
2
3
4
5
6
7
Number of Contacts

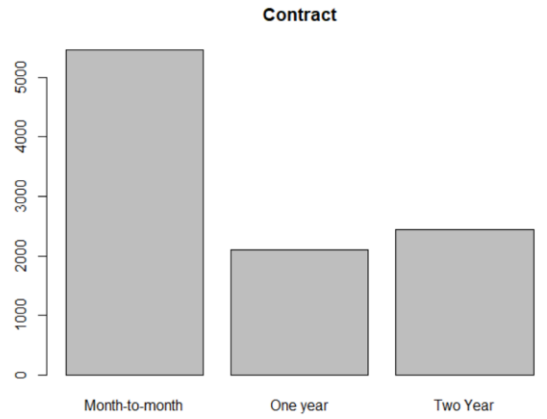

Internet Service
0
1000
2000
3000
4000
DSL
Fiber Optic
None



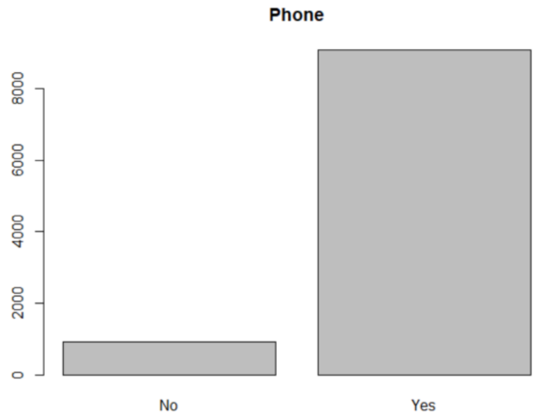
Yearly Equipment Failure



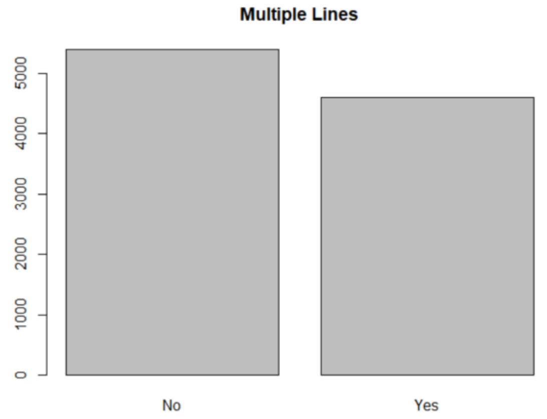
Techie



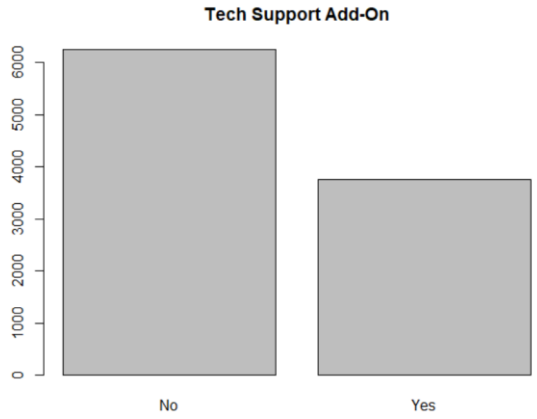
Contract



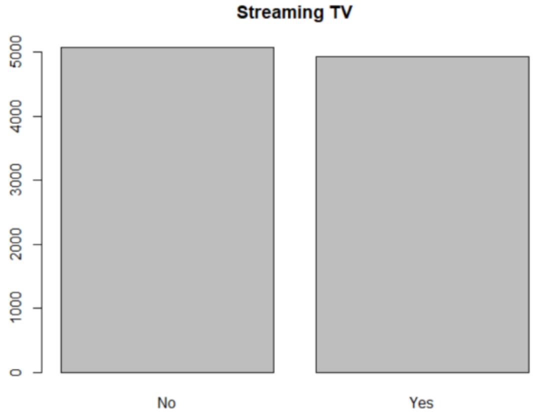
Phone

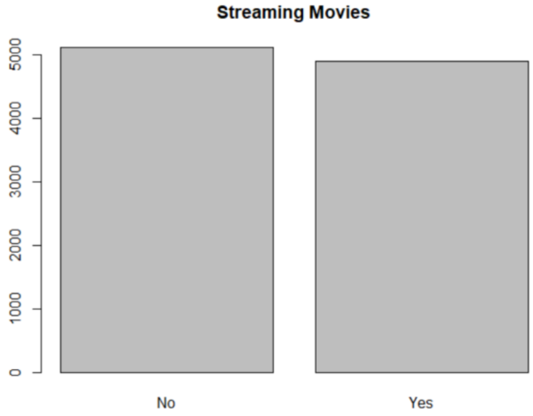


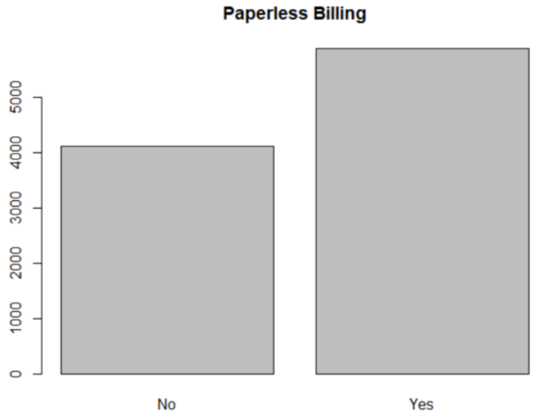
Multiple Lines

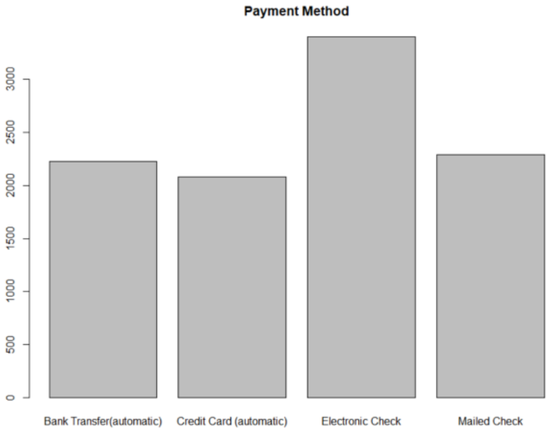


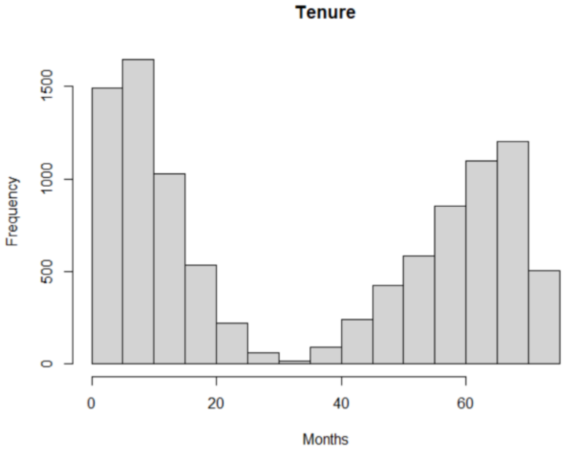
Tech Support Add-On

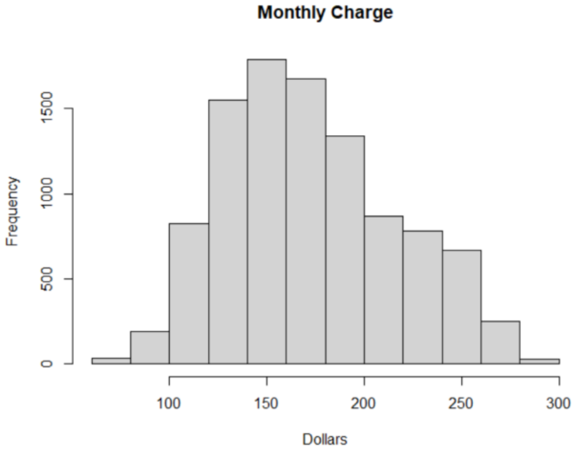

Streaming TV
0
1000
2000
3000
4000
5000
No
Yes

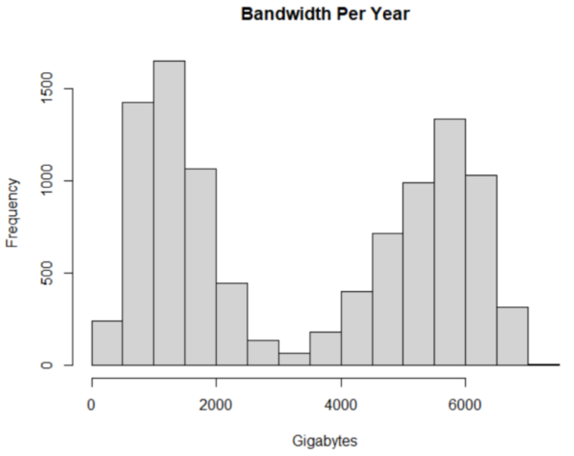

Streaming Movies
0
1000
2000
3000
4000
5000
No
Yes


Paperless Billing
0
1000
2000
3000
4000
5000
No
Yes

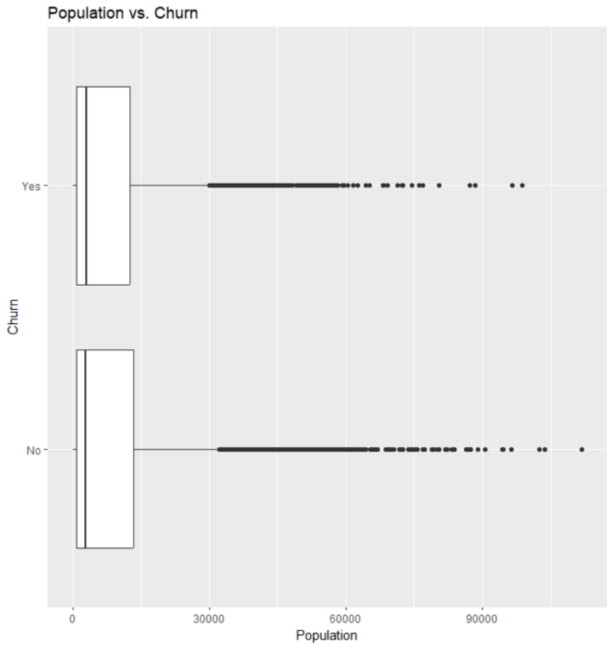

Payment Method
0
500
1000
1500
2000
2500
3000
Bank Transfer(automatic)
Credit Card (automatic)
Electronic Check
Mailed Check


Tenure
Frequency
0
500
1000
1500
0
20
40
60
Months

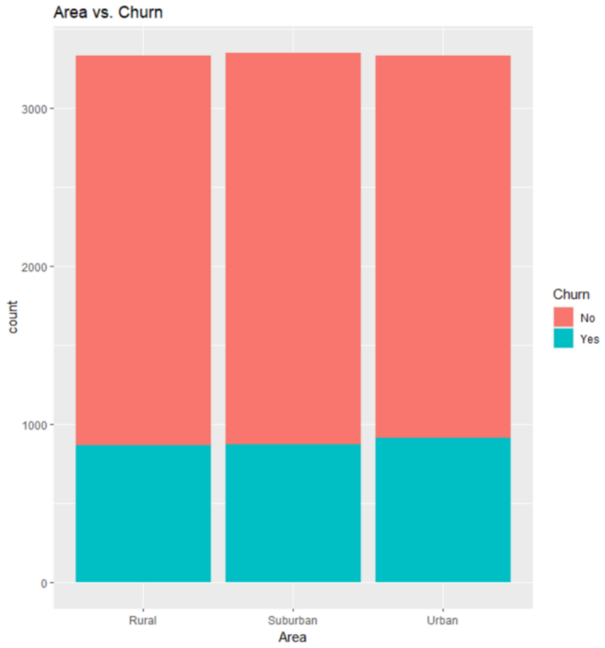

Monthly Charge
Frequency
0
500
1000
1500
100
150
200
250
300
Dollars



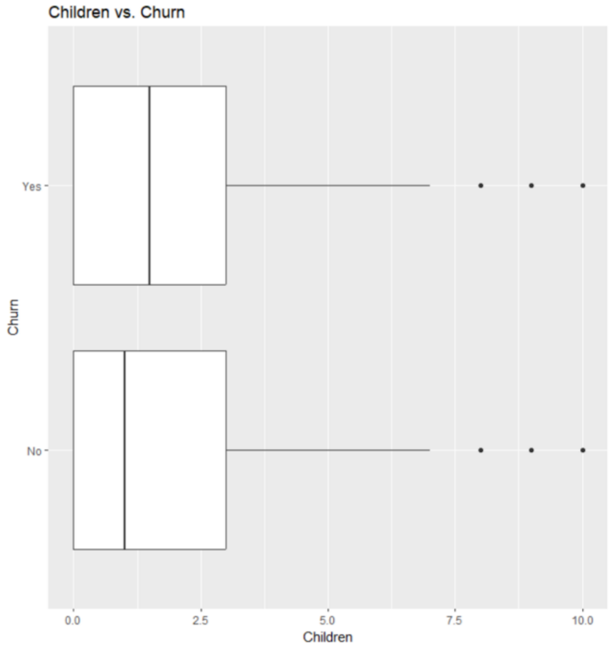
Bandwidth Per Year



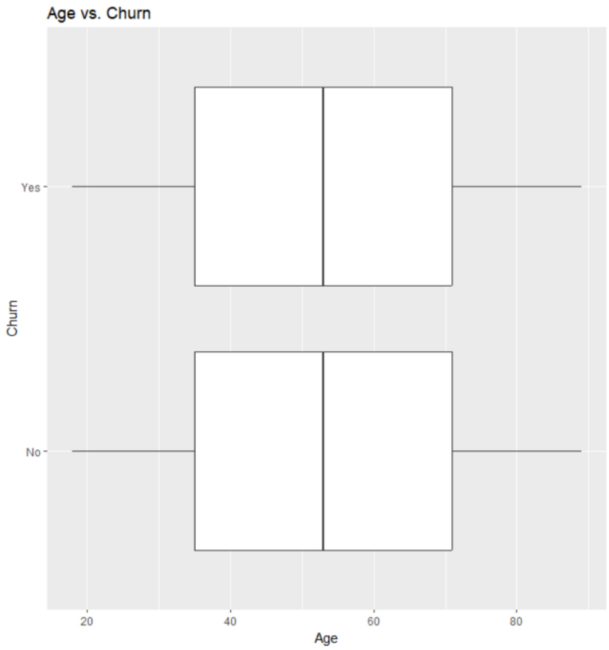
Population vs. Churn

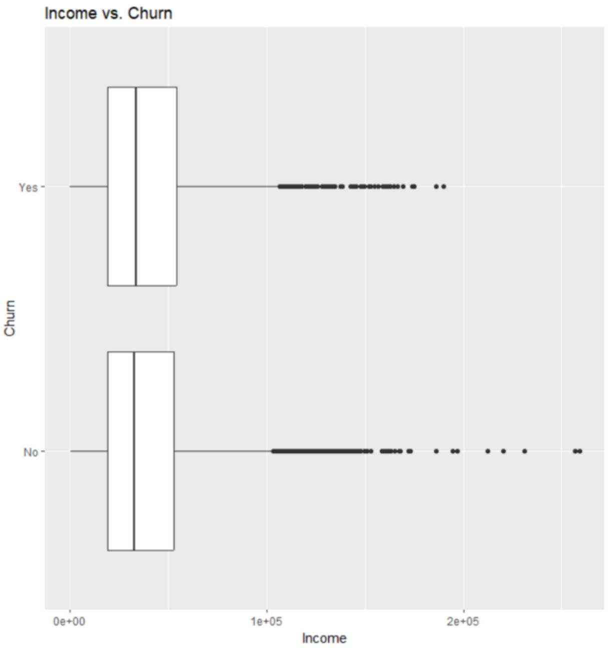


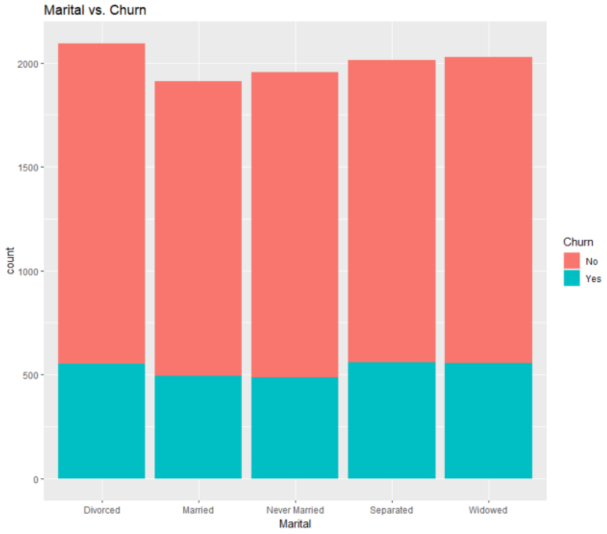
Area vs. Churn

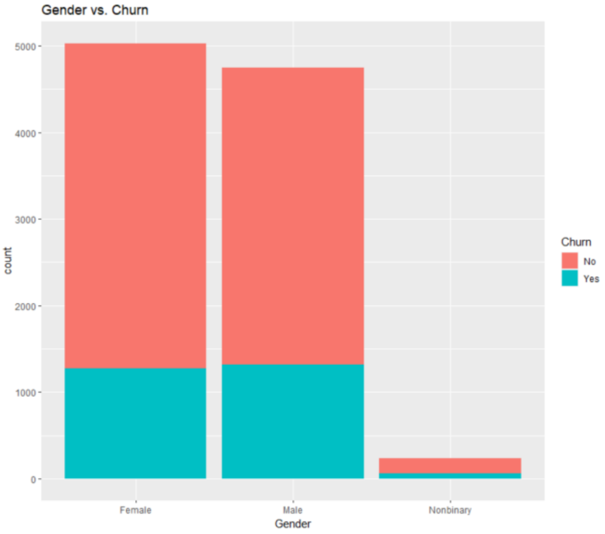


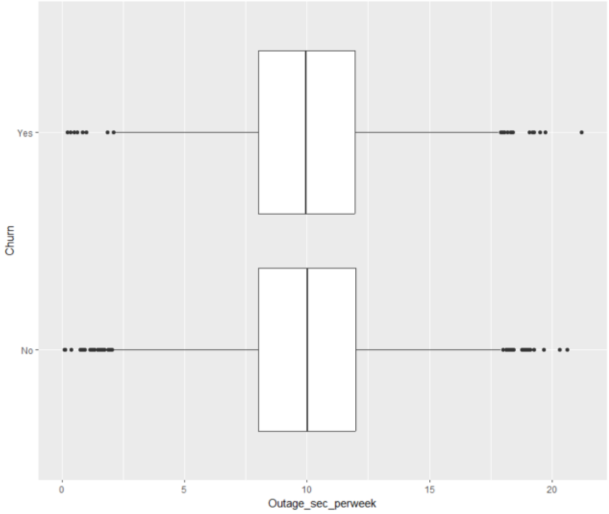
Children vs. Churn


Age vs. Churn

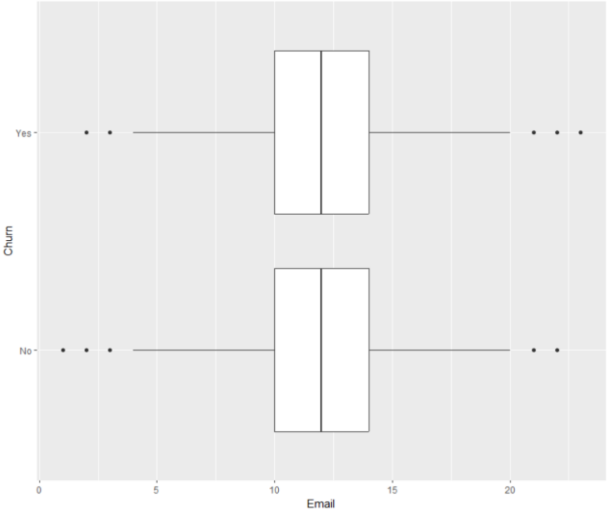

Income vs. Churn


Marital vs. Churn

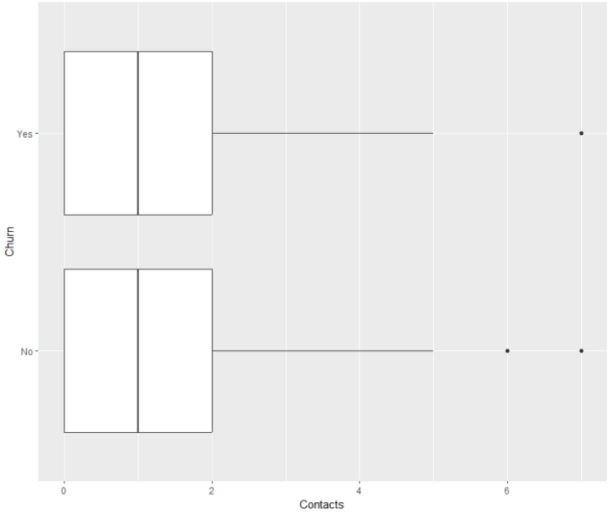

Gender vs. Churn



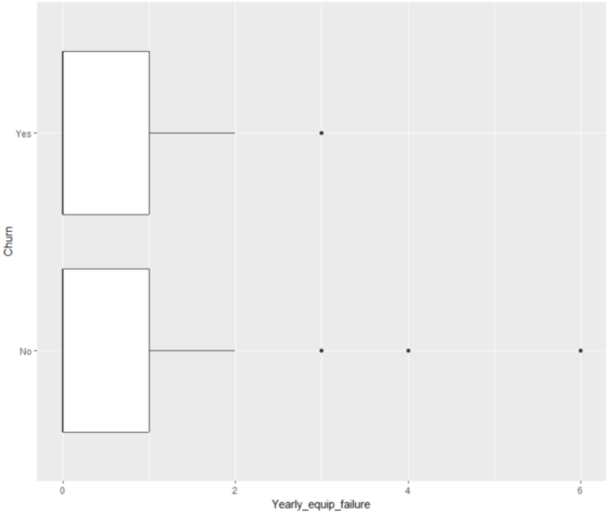
Outage Per Week vs. Churn



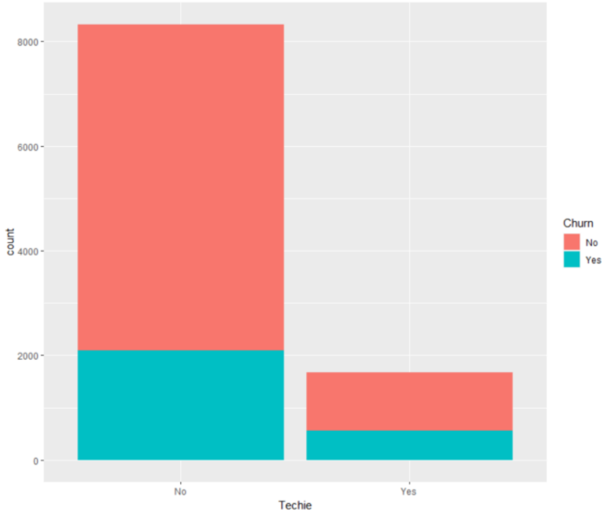
Email vs. Churn



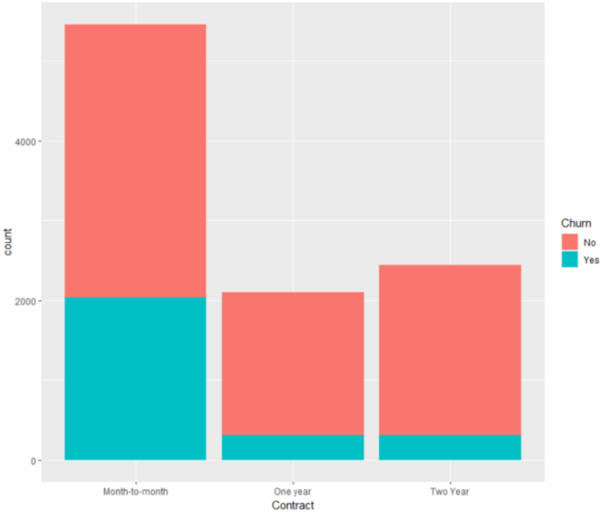
Contacts vs. Churn



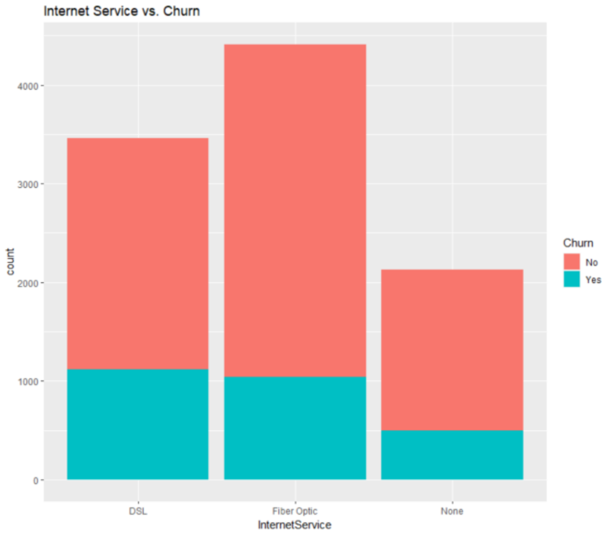
Yearly Equipment Failure vs. Churn

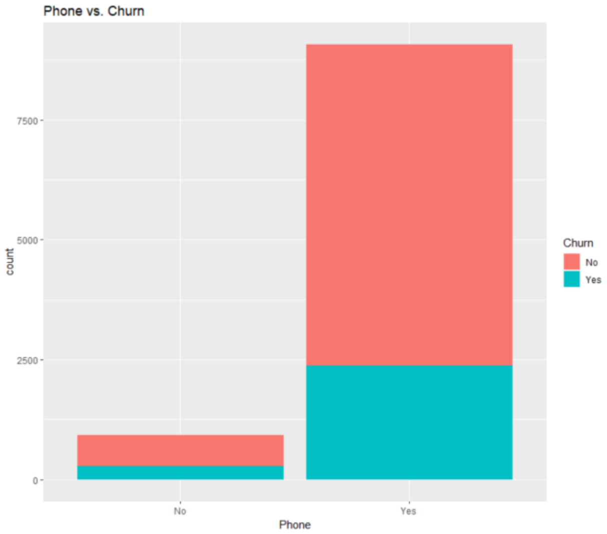


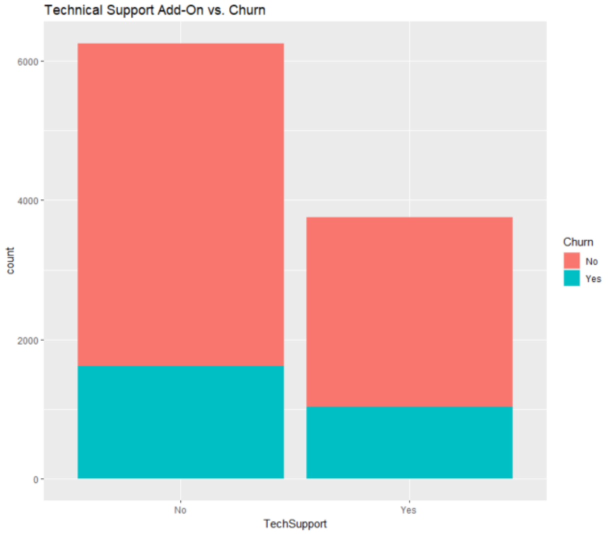
Techie vs. Churn

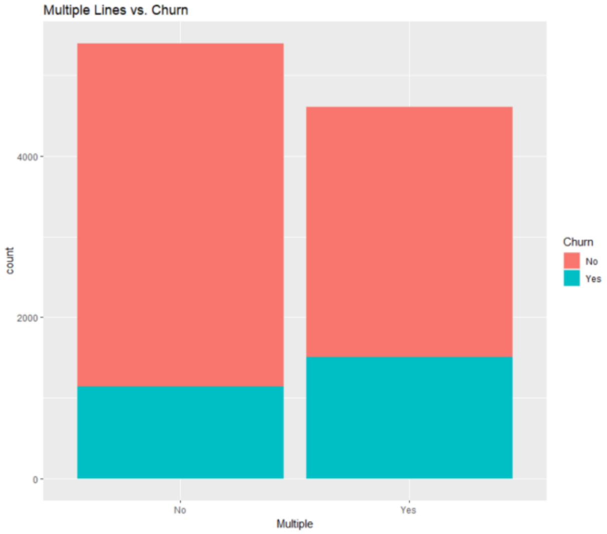


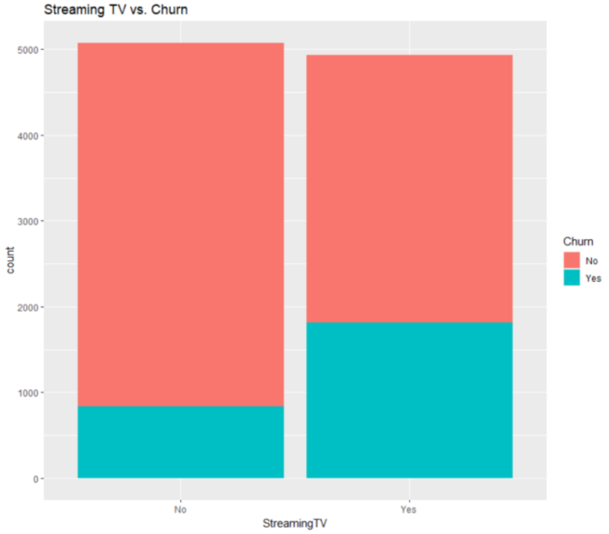
Contract vs. Churn

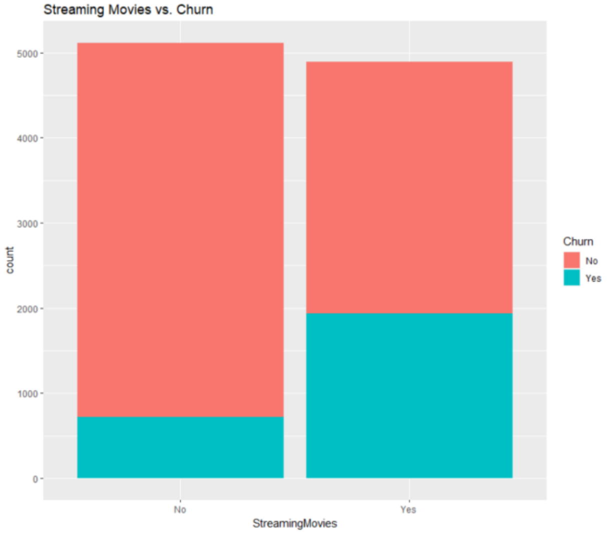

Internet Service vs. Churn

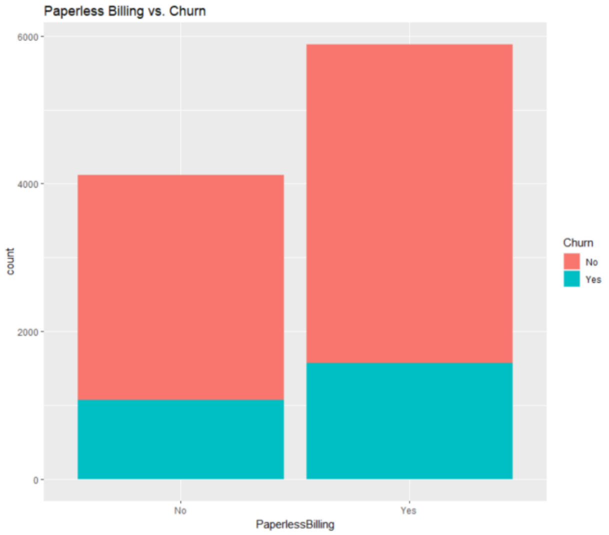

Phone vs. Churn

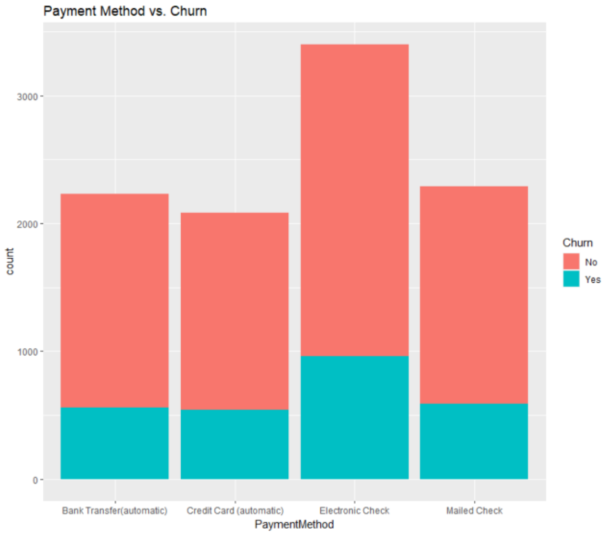

Technical Support Add-On vs. Churn

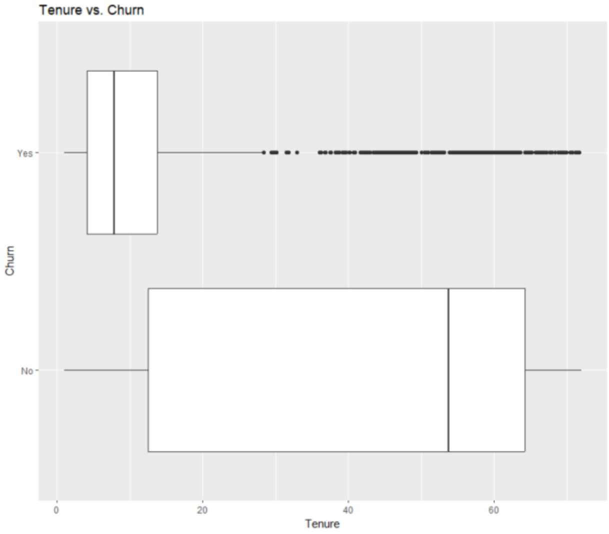

Multiple Lines vs. Churn

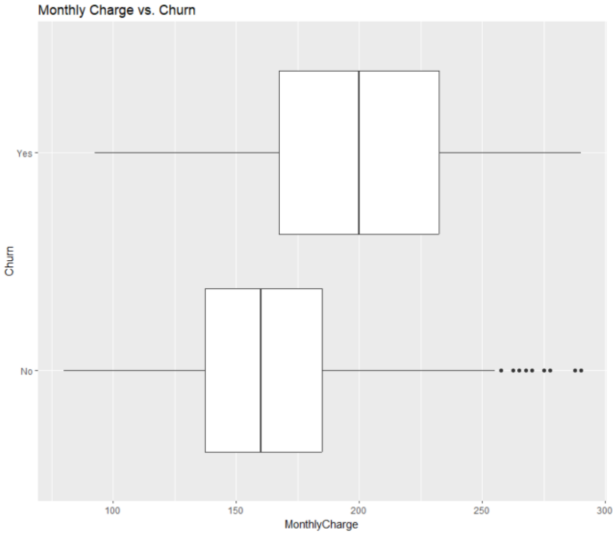

Streaming TV vs. Churn

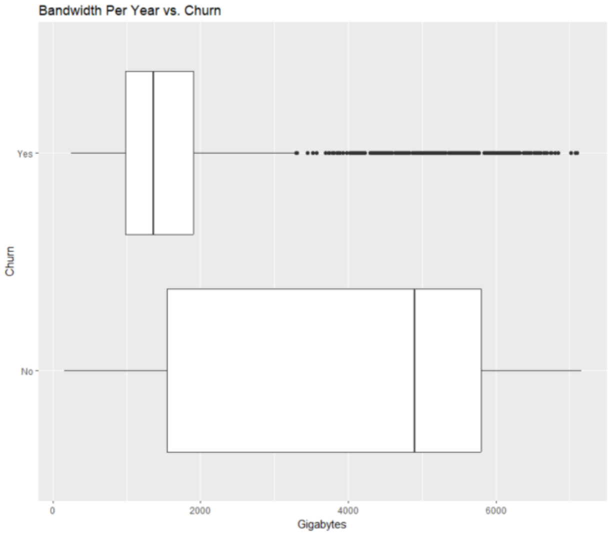

Streaming Movies vs. Churn


Paperless Billing vs. Churn
count
6000
4000
2000
0
No
Yes
PaperlessBilling
Churn
No
Yes


Payment Method vs. Churn
count
3000
2000
1000
0
Bank Transfer(automatic)
Credit Card (automatic)
Electronic Check
Mailed Check
PaymentMethod
Churn
No
Yes


Tenure vs. Churn
Churn
Yes
No
0
20
40
60
Tenure


Monthly Charge vs. Churn
Churn
Yes
No
100
150
200
250
300
MonthlyCharge

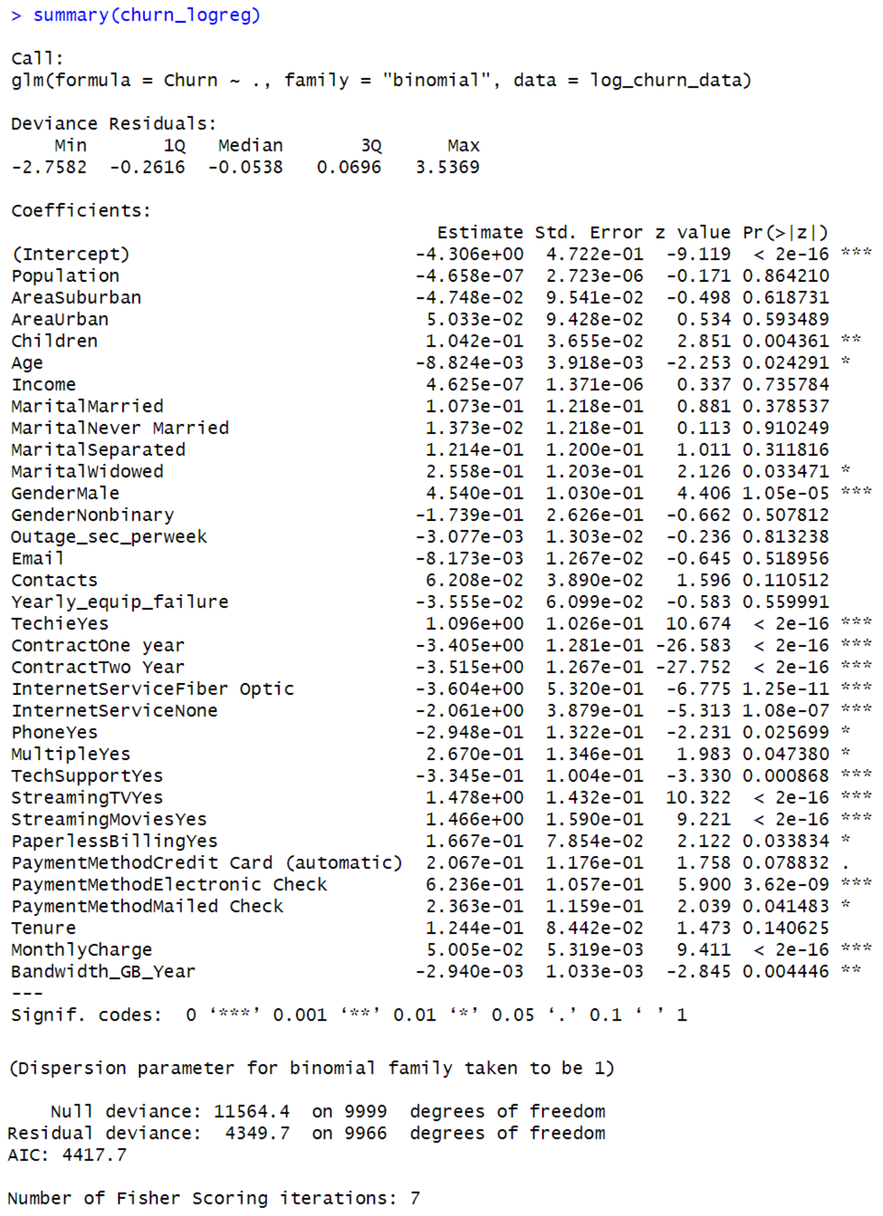

Bandwidth Per Year vs. Churn
Churn
Yes
No
0
2000
4000
6000
Gigabytes

## C5: Prepared Data Set

Prepared data set is attached in the submission.

## D1: Initial Model

My initial model had a target variable of churn and explanatory variables of population, area, children, age, income, marital status, gender, outage per week in seconds, email, contacts, yearly equipment failure, techie, contract, internet service, phone, multiple, tech support, streaming TV, streaming movies, paperless billing, payment method, tenure, monthly charge, and bandwidth per year in GBs.



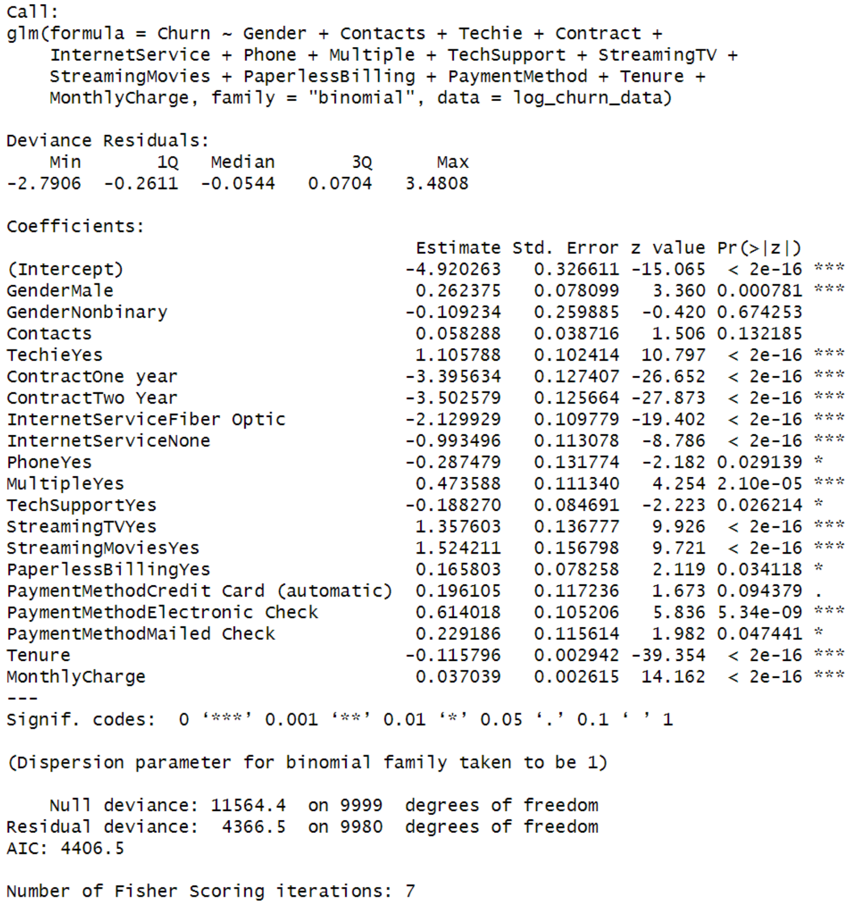
```
> summary(churn_logreg)

Call:
glm(formula = Churn ~ ., family = "binomial", data = log_churn_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7582  -0.2616  -0.0538   0.0696   3.5369  

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          -4.306e+00  4.722e-01  -9.119  < 2e-16 ***
Population                           -4.658e-07  2.723e-06  -0.171 0.864210    
AreaSuburban                         -4.748e-02  9.541e-02  -0.498 0.618731    
AreaUrban                             5.033e-02  9.428e-02   0.534 0.593489    
Children                              1.042e-01  3.655e-02   2.851 0.004361 ** 
Age                                  -8.824e-03  3.918e-03  -2.253 0.024291 *  
Income                                4.625e-07  1.371e-06   0.337 0.735784    
MaritalMarried                        1.073e-01  1.218e-01   0.881 0.378537    
MaritalNever Married                  1.373e-02  1.218e-01   0.113 0.910249    
MaritalSeparated                      1.214e-01  1.200e-01   1.011 0.311816    
MaritalWidowed                        2.558e-01  1.203e-01   2.126 0.033471 *  
GenderMale                            4.540e-01  1.030e-01   4.406 1.05e-05 ***
GenderNonbinary                      -1.739e-01  2.626e-01  -0.662 0.507812    
Outage_sec_perweek                   -3.077e-03  1.303e-02  -0.236 0.813238    
Email                                -8.173e-03  1.267e-02  -0.645 0.518956    
Contacts                              6.208e-02  3.890e-02   1.596 0.110512    
Yearly_equip_failure                 -3.555e-02  6.099e-02  -0.583 0.559991    
TechieYes                             1.096e+00  1.026e-01  10.674  < 2e-16 ***
ContractOne year                     -3.405e+00  1.281e-01 -26.583  < 2e-16 ***
ContractTwo Year                     -3.515e+00  1.267e-01 -27.752  < 2e-16 ***
InternetServiceFiber Optic           -3.604e+00  5.320e-01  -6.775 1.25e-11 ***
InternetServiceNone                  -2.061e+00  3.879e-01  -5.313 1.08e-07 ***
PhoneYes                             -2.948e-01  1.322e-01  -2.231 0.025699 *  
MultipleYes                           2.670e-01  1.346e-01   1.983 0.047380 *  
TechSupportYes                       -3.345e-01  1.004e-01  -3.330 0.000868 ***
StreamingTVYes                        1.478e+00  1.432e-01  10.322  < 2e-16 ***
StreamingMoviesYes                    1.466e+00  1.590e-01   9.221  < 2e-16 ***
PaperlessBillingYes                   1.667e-01  7.854e-02   2.122 0.033834 *  
PaymentMethodCredit Card (automatic)  2.067e-01  1.176e-01   1.758 0.078832 .  
PaymentMethodElectronic Check         6.236e-01  1.057e-01   5.900 3.62e-09 ***
PaymentMethodMailed Check             2.363e-01  1.159e-01   2.039 0.041483 *  
Tenure                                1.244e-01  8.442e-02   1.473 0.140625    
MonthlyCharge                         5.005e-02  5.319e-03   9.411  < 2e-16 ***
Bandwidth_GB_Year                    -2.940e-03  1.033e-03  -2.845 0.004446 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564.4  on 9999  degrees of freedom
Residual deviance:  4349.7  on 9966  degrees of freedom
AIC: 4417.7

Number of Fisher Scoring iterations: 7
```

## D2: Justification of Model Reduction

I decided to use p-values for my variable selection and AIC for my model evaluation metric. I ran my initial model, then I selected the variable with the highest p-value and removed it, and ran the model again. If the AIC was lower, I repeated the process of selecting the variable with the highest p-value and removed it. If the AIC was higher, I kept the variable in the model. This allowed me to lower my AIC from 4417.7 to 4406.5. Almost all of my variables are also significant at the 0.05 level.

## D3: Reduced Logistic Regression Model

My reduced model includes the categorical variables gender, techie, contract, internet service, phone, multiple lines, tech support add-on, streaming TV, streaming movies, paperless billing, and payment method. It also included the continuous variables contacts, tenure, and monthly charge.



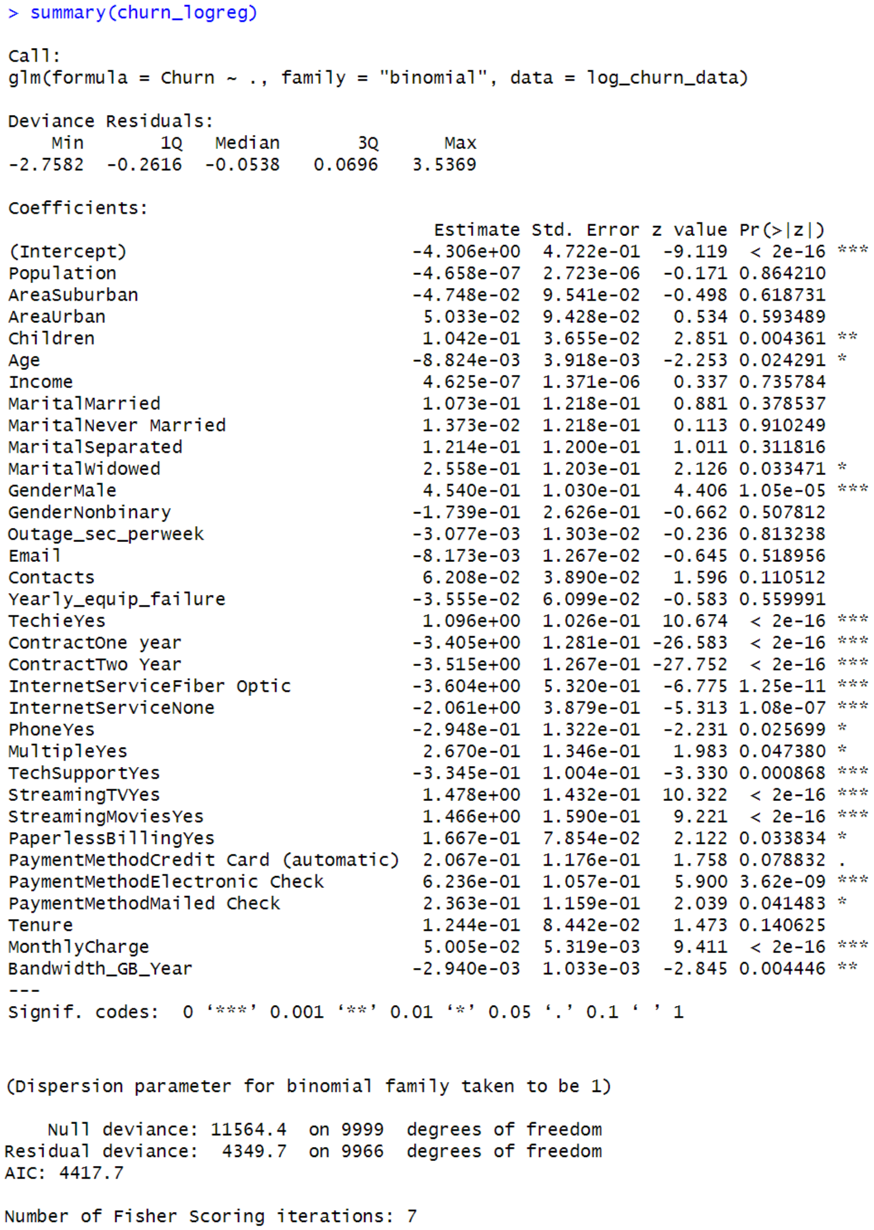
```
Call:
glm(formula = Churn ~ Gender + Contacts + Techie + Contract + 
    InternetService + Phone + Multiple + TechSupport + StreamingTV + 
    StreamingMovies + PaperlessBilling + PaymentMethod + Tenure + 
    MonthlyCharge, family = "binomial", data = log_churn_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7906  -0.2611  -0.0544   0.0704   3.4808  

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          -4.920263   0.326611 -15.065  < 2e-16 ***
GenderMale                            0.262375   0.078099   3.360 0.000781 ***
GenderNonbinary                      -0.109234   0.259885  -0.420 0.674253    
Contacts                              0.058288   0.038716   1.506 0.132185    
TechieYes                             1.105788   0.102414  10.797  < 2e-16 ***
ContractOne year                     -3.395634   0.127407 -26.652  < 2e-16 ***
ContractTwo Year                     -3.502579   0.125664 -27.873  < 2e-16 ***
InternetServiceFiber Optic           -2.129929   0.109779 -19.402  < 2e-16 ***
InternetServiceNone                  -0.993496   0.113078  -8.786  < 2e-16 ***
PhoneYes                             -0.287479   0.131774  -2.182 0.029139 *  
MultipleYes                           0.473588   0.111340   4.254 2.10e-05 ***
TechSupportYes                       -0.188270   0.084691  -2.223 0.026214 *  
StreamingTVYes                        1.357603   0.136777   9.926  < 2e-16 ***
StreamingMoviesYes                    1.524211   0.156798   9.721  < 2e-16 ***
PaperlessBillingYes                   0.165803   0.078258   2.119 0.034118 *  
PaymentMethodCredit Card (automatic)  0.196105   0.117236   1.673 0.094379 .  
PaymentMethodElectronic Check         0.614018   0.105206   5.836 5.34e-09 ***
PaymentMethodMailed Check             0.229186   0.115614   1.982 0.047441 *  
Tenure                               -0.115796   0.002942 -39.354  < 2e-16 ***
MonthlyCharge                         0.037039   0.002615  14.162  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564.4  on 9999  degrees of freedom
Residual deviance:  4366.5  on 9980  degrees of freedom
AIC: 4406.5

Number of Fisher Scoring iterations: 7
```

## E1: Model Comparison

I used p-values and AIC to narrow down my model. I did this by running the model with all variables, then I removed one variable at a time based on p-values. I used AIC as a guide for if my model is improving or not. If my model improved, I removed another variable. I continued this process until I could get the AIC as low as possible. I ended up lowering my AIC from 4417.7 to 4406.5.

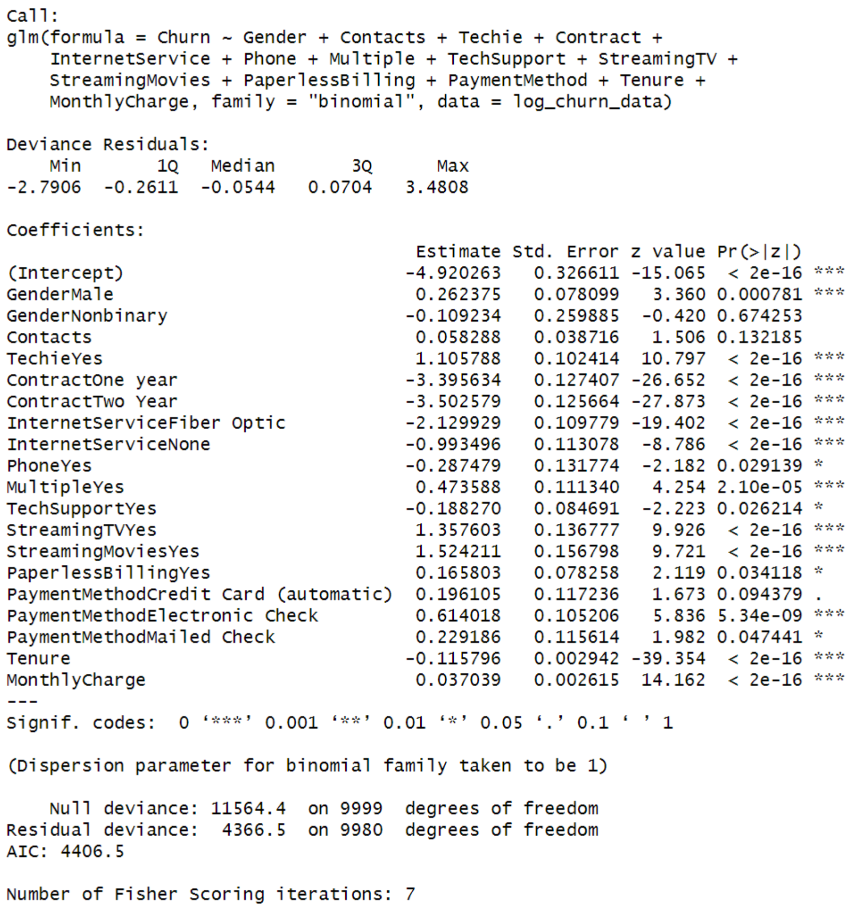
## E2: Output and Calculations

Here is my initial model with all calculations shown.



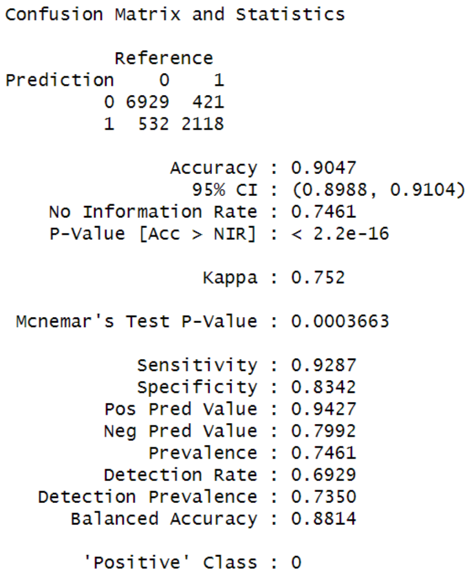
```
> summary(churn_logreg)

Call:
glm(formula = Churn ~ ., family = "binomial", data = log_churn_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7582  -0.2616  -0.0538   0.0696   3.5369  

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          -4.306e+00  4.722e-01  -9.119  < 2e-16 ***
Population                           -4.658e-07  2.723e-06  -0.171 0.864210    
AreaSuburban                         -4.748e-02  9.541e-02  -0.498 0.618731    
AreaUrban                             5.033e-02  9.428e-02   0.534 0.593489    
Children                              1.042e-01  3.655e-02   2.851 0.004361 ** 
Age                                  -8.824e-03  3.918e-03  -2.253 0.024291 *  
Income                                4.625e-07  1.371e-06   0.337 0.735784    
MaritalMarried                        1.073e-01  1.218e-01   0.881 0.378537    
MaritalNever Married                  1.373e-02  1.218e-01   0.113 0.910249    
MaritalSeparated                      1.214e-01  1.200e-01   1.011 0.311816    
MaritalWidowed                        2.558e-01  1.203e-01   2.126 0.033471 *  
GenderMale                            4.540e-01  1.030e-01   4.406 1.05e-05 ***
GenderNonbinary                      -1.739e-01  2.626e-01  -0.662 0.507812    
Outage_sec_perweek                   -3.077e-03  1.303e-02  -0.236 0.813238    
Email                                -8.173e-03  1.267e-02  -0.645 0.518956    
Contacts                              6.208e-02  3.890e-02   1.596 0.110512    
Yearly_equip_failure                 -3.555e-02  6.099e-02  -0.583 0.559991    
TechieYes                             1.096e+00  1.026e-01  10.674  < 2e-16 ***
ContractOne year                     -3.405e+00  1.281e-01 -26.583  < 2e-16 ***
ContractTwo Year                     -3.515e+00  1.267e-01 -27.752  < 2e-16 ***
InternetServiceFiber Optic           -3.604e+00  5.320e-01  -6.775 1.25e-11 ***
InternetServiceNone                  -2.061e+00  3.879e-01  -5.313 1.08e-07 ***
PhoneYes                             -2.948e-01  1.322e-01  -2.231 0.025699 *  
MultipleYes                           2.670e-01  1.346e-01   1.983 0.047380 *  
TechSupportYes                       -3.345e-01  1.004e-01  -3.330 0.000868 ***
StreamingTVYes                        1.478e+00  1.432e-01  10.322  < 2e-16 ***
StreamingMoviesYes                    1.466e+00  1.590e-01   9.221  < 2e-16 ***
PaperlessBillingYes                   1.667e-01  7.854e-02   2.122 0.033834 *  
PaymentMethodCredit Card (automatic)  2.067e-01  1.176e-01   1.758 0.078832 .  
PaymentMethodElectronic Check         6.236e-01  1.057e-01   5.900 3.62e-09 ***
PaymentMethodMailed Check             2.363e-01  1.159e-01   2.039 0.041483 *  
Tenure                                1.244e-01  8.442e-02   1.473 0.140625    
MonthlyCharge                         5.005e-02  5.319e-03   9.411  < 2e-16 ***
Bandwidth_GB_Year                    -2.940e-03  1.033e-03  -2.845 0.004446 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564.4  on 9999  degrees of freedom
Residual deviance:  4349.7  on 9966  degrees of freedom
AIC: 4417.7

Number of Fisher Scoring iterations: 7
```

Here is my final reduced model with all calculations shown.



```
Call:
glm(formula = Churn ~ Gender + Contacts + Techie + Contract + 
    InternetService + Phone + Multiple + TechSupport + StreamingTV + 
    StreamingMovies + PaperlessBilling + PaymentMethod + Tenure + 
    MonthlyCharge, family = "binomial", data = log_churn_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7906  -0.2611  -0.0544   0.0704   3.4808  

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          -4.920263   0.326611 -15.065  < 2e-16 ***
GenderMale                            0.262375   0.078099   3.360 0.000781 ***
GenderNonbinary                      -0.109234   0.259885  -0.420 0.674253    
Contacts                              0.058288   0.038716   1.506 0.132185    
TechieYes                             1.105788   0.102414  10.797  < 2e-16 ***
ContractOne year                     -3.395634   0.127407 -26.652  < 2e-16 ***
ContractTwo Year                     -3.502579   0.125664 -27.873  < 2e-16 ***
InternetServiceFiber Optic           -2.129929   0.109779 -19.402  < 2e-16 ***
InternetServiceNone                  -0.993496   0.113078  -8.786  < 2e-16 ***
PhoneYes                             -0.287479   0.131774  -2.182 0.029139 *  
MultipleYes                           0.473588   0.111340   4.254 2.10e-05 ***
TechSupportYes                       -0.188270   0.084691  -2.223 0.026214 *  
StreamingTVYes                        1.357603   0.136777   9.926  < 2e-16 ***
StreamingMoviesYes                    1.524211   0.156798   9.721  < 2e-16 ***
PaperlessBillingYes                   0.165803   0.078258   2.119 0.034118 *  
PaymentMethodCredit Card (automatic)  0.196105   0.117236   1.673 0.094379 .  
PaymentMethodElectronic Check         0.614018   0.105206   5.836 5.34e-09 ***
PaymentMethodMailed Check             0.229186   0.115614   1.982 0.047441 *  
Tenure                               -0.115796   0.002942 -39.354  < 2e-16 ***
MonthlyCharge                         0.037039   0.002615  14.162  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564.4  on 9999  degrees of freedom
Residual deviance:  4366.5  on 9980  degrees of freedom
AIC: 4406.5

Number of Fisher Scoring iterations: 7
```

And here is my confusion matrix



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6929  421
         1  532 2118
                                          
               Accuracy : 0.9047          
                 95% CI : (0.8988, 0.9104)
    No Information Rate : 0.7461          
    P-Value [Acc > NIR] : < 2.2e-16       
                                          
                  Kappa : 0.752           
                                          
 Mcnemar's Test P-Value : 0.0003663       
                                          
            Sensitivity : 0.9287          
            Specificity : 0.8342          
         Pos Pred Value : 0.9427          
         Neg Pred Value : 0.7992          
             Prevalence : 0.7461          
         Detection Rate : 0.6929          
   Detection Prevalence : 0.7350          
      Balanced Accuracy : 0.8814          
                                          
       'Positive' Class : 0               
```

## E3: Code

```
log_churn_data$Churn <- as.factor(log_churn_data$Churn)

churn_logreg <- glm(Churn ~ ., data=log_churn_data, family="binomial")

summary(churn_logreg)

vif(churn_logreg)


reduced_log_churn <- glm(Churn~. -Bandwidth_GB_Year -Population -
Outage_sec_perweek -Income -Age -Email -Marital -Area -
Yearly_equip_failure -Children, data=log_churn_data,
family="binomial")

summary(reduced_log_churn)

vif(reduced_log_churn)


reduced_log_churn <- glm(Churn ~ Gender + Contacts + Techie + Contract
+ InternetService + Phone + Multiple + TechSupport + StreamingTV +
StreamingMovies + PaperlessBilling + PaymentMethod + Tenure +
MonthlyCharge, data=log_churn_data, family="binomial")

summary(reduced_log_churn)

vif(reduced_log_churn)


predicted <- predict(reduced_log_churn, type="response")

predicted

predicted <- ifelse(predicted>=0.5, "1", "0")

levels(as.factor(predicted))

actual <- ifelse(log_churn_data$Churn == "No","0","1")

levels(as.factor(actual))

confusionMatrix(as.factor(actual), as.factor(predicted))
```

## F1: Results

My logistic regression equation is: $Churn = 0.26(GenderMale) - 0.11(GenderNonbinary) + 0.06(Contacts) + 1.11(TechieYes) - 3.4(ContractOne\ Year) - 3.5(ContractTwo\ Year) - 2.13(InternetServiceFiber\ Optic) - InternetServiceNone - 0.29(PhoneYes) - 0.47(MultipleYes) - 0.19(TechSupportYes) + 1.36(StreamingTVYes) + 1.52(StreamingMoviesYes) + 0.17(PaperlessBillingYes) +$

$$0.2\left(PaymentMethodCredit\ Card\ (automatic)\right) + 0.61(PaymentMethodElectronic\ Check) + 0.23(PaymentMethodMailed\ Check) - 0.12(Tenure) + 0.04(MonthlyCharge) - 4.92$$

My model found that being male, number of contacts, being a techie, having multiple lines, streaming TV, streaming movies, paperless billing, payment methods of credit card, electronic check, and mailed check, and monthly charge all increased chances of churn. Being non-binary, having a one or two year contract, having fiber optic internet or no internet, phone service, tech support add-on, and tenure all decreased chances of churn.

I do think this model is strong enough to accurately predict whether someone will churn. Over half of the variables are significant at the 0.001 level. The confusion matrix statistics showed that the model was over 90% accurate. (CN, 8/3/22)

## F2: Recommendations

I think the organization would be smart to implement this regression equation to predict if a customer will churn. If anything, looking at the values with the lowest p-values and highest estimated coefficients would help as well. Contracts decrease the chances of churn the most, so offering discounts for customers that sign up for a one or two year contract could keep customers for a longer period of time. Lowering churn rate is obviously important for an organization, and keeping customers for a longer period of time will actually lower the odds that they eventually churn, which is also shown in this model.

## G: Panopto Demonstration

The panopto video is attached in the submission and can be found in the link below:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f1081c4a-2257-4797-b903-af8900362dd6

## H: Sources of Third-Party Code

*Glm: Fitting Generalized Linear Models*. RDocumentation. Retrieved 1/11/23. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

Kassambara. *Stepwise Logistic Regression Essentials in R.* Statistical Tools for High-Throughput Data Analysis. 11/3/18. Retrieved 1/11/23. https://www.sthda.com/english/articles/36-classification-methods-essentials/150-stepwise-logistic-regression-essentials-in-r/#perform-stepwise-variable-selection

## I: Sources

*Assumptions of Logistic Regression.* Statistics Solutions. Retrieved 1/12/23. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

CN, Prajwal. *Confusion Matrix in R: A Complete Guide.* Digital Ocean. 8/3/22. Retrieved 1/10/23. https://www.digitalocean.com/community/tutorials/confusion-matrix-in-r