2024

Logistic Regression Model

TASK 2: LOGISTIC REGRESSION MODELING

Part I: Research Question

A. Describe the purpose of this data analysis by doing the following:

1. **Summarize one research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using logistic regression.**

   The research question that is purposed with this analysis is as follows:

   Based on the available churn dataset this analysis will be using a logistic regression model to predict how variables within the dataset may affect a customer's churn rates.

2. **Define the goals of the data analysis.**

   One of the goals of a company is to maintain a low churn rate. A churn rate is characterized by the rate at which customers discontinue doing business with a company. By maintaining a low churn rate a company is likely to grow, increase profits, and preserve overall cost effectiveness (Frankenfield, 2022). By exploring the provided dataset, an analyst may predict which customers will most likely discontinue their services with a telecommunications company.

Part II: Method Justification

B. Describe logistic regression methods by doing the following:

1. **Summarize four assumptions of a logistic regression model.**

   Logistic regression models depend on assumptions for the dataset for regression to be viable. There are four main assumptions of a logistic regression model:

   a. Outliers – The logistic regression model assumes the data does not contain extreme outliers, or the data is free of any external observations that may influence the model's outcome (Voxco, 2023).
   b. Multicollinearity – Similar to multiple linear regression, logistic regression assumes that independent variables are not highly correlated with one another (Voxco, 2023). When multicollinearity is present, it is indicative of independent variables being too highly correlated with one another (Statistic Solutions, 2023).
   c. Independent Observations – Observations within the dataset should be independent of each other. Each observation within the dataset occurs without the influence of another observation. No observation should be dependent on another observation (Voxco, 2023).
   d. Large Sample Sizes – Logistic regression requires larger sample sizes. As a general guideline, a minimum of 10 cases with the least frequent outcome for each independent variable is needed within the model (Statistic Solutions, 2023).

2. **Describe <u>two</u> benefits of using Python or R in support of various phases of the analysis.**

    1. Statistical Focus – R Programming is designed with statistical computing and data analysis in mind. It is a very diverse and rich set of statistical packages that are specifically designed and functional for statistical analysis. Considering this assessment is asking for data analysis using logistic regression model of a dataset, R is an ideal choice (Statistics Solutions, 2023).
    2. Data Visualization – R Programming is a very powerful tool for creating dynamic data visualizations. This is especially true when using packages such as ggplot2 which will be used for this assessment. Visualizations are a good way to explore data but to also test the logistic regression model such as creating sigmoid curves to explore lines of best fit of the data (Simplilearn, 2023).

3. **Explain why logistic regression is an appropriate technique to analyze the research question summarized in part I.**

    Logistic regression is an appropriate analysis for discrete values such as binary (0, 1) data types from a set of independent variables. It is designed to predict the probability of an event by fitting the data into a logistical function. In this analysis, the research question is attempting to make a prediction on churn rates which is a categorical yes and no data type which will be converted to binary. In this case, logistic regression is an appropriate analysis (Simplilearn, 2023).

Part III: Data Preparation

C. Summarize the data preparation process for logistic regression by doing the following:

1. **Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including the annotated code.**

    The goals of data cleaning and preparation are to gain an understanding of the available data for analysis. To achieve this, an in-depth look at the data structure and summaries of the variables is necessary.

    My methodology to achieve the data goals are as follows:

    1. Make a copy of the data
    2. Import data into R programming.
    3. Examine the structure of the data to better understand the dataset.
    4. Examine and clean the data for potential missing data, renaming columns, duplications, data errors, anomalies, removal of unneeded variables, or anything else that might aid in the analysis.

5. Summarize data by discovering the distribution and potential outliers within the variables that might alter the statistical analysis of the dataset using both histograms and boxplots. Handle outliers as necessary.
6. Summarize and find relationships with the data using chi-square analysis.

2. Describe the dependent variable and *all* independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

The following process was executed in R to prepare and clean the data for analysis:

Using R, packages were imported to conduct analysis. Once the packages were imported, setwd() was used to create a working directory. Then, importing the .csv file was used using read.csv():

```
# Packages that will be used for regression:
    library(tidyverse)
    library(dplyr)
    library(plyr)
    library(readr)
    library(ggplot2)
    library(gridExtra)
    library(stats)
    library(gplots)
    library(tidycomm)
    library(AICcmodavg)

# Setting the working directory:

    setwd('C:/Users/agana/OneDrive/Desktop/WGU/D208/Datasets/Churn')

# Importing the dataset:

    churn_df <-read.csv('churn_data.csv')

# Renaming the dataset:

    mydata <- churn_df
```

Once the dataset was imported and the directory was set, to prep the data for cleaning, examining the structure of the data is extremely useful. The str() command was used first which is proceeded by renaming the dataset to “mydata” for easier navigation within coding:

```
# Summary/Structure of Data

    str(mydata)
    summary(mydata)
```

The str() command output revealed the dataset contains 10,000 observations. In addition, the dataset contained 50 variables:

```
> str(mydata)
'data.frame':	10000 obs. of  50 variables:
 $ CaseOrder           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Customer_id         : chr  "K409198" "S120509" "K191035" "D90850" ...
 $ Interaction         : chr  "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "344d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...
 $ UID                 : chr  "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a" "f1784cfa9f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...
 $ City                : chr  "Point Baker" "West Branch" "Yamhill" "Del Mar" ...
 $ State               : chr  "AK" "MI" "OR" "CA" ...
 $ County              : chr  "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...
 $ Zip                 : int  99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...
 $ Lat                 : num  56.3 44.3 45.4 33 29.4 ...
 $ Lng                 : num  -133.4 -84.2 -123.2 -117.2 -95.8 ...
 $ Population          : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
 $ Area                : chr  "Urban" "Urban" "Urban" "Suburban" ...
 $ TimeZone            : chr  "America/Sitka" "America/Detroit" "America/Los_Angeles" "America/Los_Angeles" ...
 $ Job                 : chr  "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Officer" "Solicitor" ...
 $ Children            : int  0 1 4 1 0 3 0 2 2 1 ...
 $ Age                 : int  68 27 50 48 83 83 79 30 49 86 ...
 $ Income              : num  28562 21705 9610 18925 40074 ...
 $ Marital             : chr  "Widowed" "Married" "Widowed" "Married" ...
 $ Gender              : chr  "Male" "Female" "Female" "Male" ...
 $ Churn               : chr  "No" "Yes" "No" "No" ...
 $ Outage_sec_perweek  : num  7.98 11.7 10.75 14.91 8.15 ...
 $ Email               : int  10 12 9 15 16 15 10 16 20 18 ...
 $ Contacts            : int  0 0 0 2 2 3 0 0 2 1 ...
 $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
 $ Techie              : chr  "No" "Yes" "Yes" "Yes" ...
 $ Contract            : chr  "One year" "Month-to-month" "Two Year" "Two Year" ...
 $ Port_modem          : chr  "Yes" "No" "Yes" "No" ...
 $ Tablet              : chr  "Yes" "Yes" "No" "No" ...
 $ InternetService     : chr  "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...
 $ Phone               : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Multiple            : chr  "No" "Yes" "Yes" "No" ...
 $ OnlineSecurity      : chr  "Yes" "Yes" "No" "Yes" ...
 $ OnlineBackup        : chr  "Yes" "No" "No" "No" ...
 $ DeviceProtection    : chr  "No" "No" "No" "No" ...
 $ TechSupport         : chr  "No" "No" "No" "No" ...
 $ StreamingTV         : chr  "No" "Yes" "No" "Yes" ...
 $ StreamingMovies     : chr  "Yes" "Yes" "Yes" "No" ...
 $ PaperlessBilling    : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ PaymentMethod       : chr  "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)" "Mailed Check" ...
 $ Tenure              : num  6.8 1.16 15.75 17.09 1.67 ...
 $ MonthlyCharge       : num  172 243 160 120 150 ...
 $ Bandwidth_GB_Year   : num  905 801 2055 2165 271 ...
 $ Item1               : int  5 3 4 4 4 3 6 2 5 2 ...
 $ Item2               : int  5 4 4 4 4 3 5 2 4 2 ...
 $ Item3               : int  5 3 2 4 4 3 6 2 4 2 ...
 $ Item4               : int  3 3 4 2 3 2 4 5 3 2 ...
 $ Item5               : int  4 4 4 5 4 4 1 2 4 5 ...
 $ Item6               : int  4 3 3 4 4 3 5 3 3 2 ...
 $ Item7               : int  3 4 3 3 4 3 5 4 4 3 ...
 $ Item8               : int  4 4 3 3 5 3 5 5 4 3 ...
```

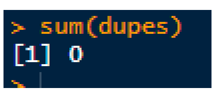
As previously stated, there are 50 variables consisting of 4 unique identifying attributes of the customers which are CaseOrder, Customer_id, Interaction, and UID. Additionally, there are 15 demographic variables: City, State, County, Zip Code, Longitude, Latitude, Population, Area, Income, Martial (Status), and Gender. One variable stating if the customer has left within the last month: Churn. There are 9 variables regarding customer services: internet services, phone, multiple (lines), online security, online backup, device protection, tech support, streaming TV, and streaming movies. There are 13 variables specifying customer account information: outage_sec_perweek (seconds per week), email, contacts, yearly_equip_failure, techie, contract, port_modem, table, paperlessbilling, paymentmethod, tenure, monthlycharge, and bandwidth_GB_year. Lastly, there are 8 variables concerning survey information: Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8.

The variables range from continuous, categorical, ordinal, etc. The several continuous variables are: Tenure, Outage_sec_perweek, MonthlyCharge, Bandwidth_GB_Year, CaseOrder, Population, Children, Age, Email, Contracts, Yearly_equip_failure, and Income. There are 20 categorical variables that range from yes/no such as Churn and Tablet, to more specified such as Area and TimeZone. They are the following: Area, TimeZone, Marital, Gender, Churn, Techie, Contract, Port_modem, Tablet, PaperlessBilling, PaymentMethod, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies. Additionally, there are 4 string variables: City, State, County, and Job. Also, 3 variables fall into the alphanumeric data type: Customer_id, Interaction, and UID. While it is debatable of what data types of geographic variables are, these 3 variables will be listed as “geographic”: Zip, Lat, Lng. Lastly, there are 8 ordinal variables of survey information: Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8.

To ensure the data is complete before proceeding, a quick check to ensure no duplicate records are in the dataset:

```
# Searching for Duplicates

        dupes <- duplicated(mydata)

# Summing to see if duplicates are present:

        sum(dupes)
```

The output for this check came back as 0 which concludes no records are duplications:



A further inspection of the data, there are a number of variables that not very meaningful for this analysis: CaseOrder, Customer_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Martial, and Payment Method.

These can be removed:

```
#Listing columns to be removed:
        columns_to_remove <- c('CaseOrder', 'Customer_id', 'Interaction', 'UID',
        'City', 'State', 'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone',
        'Job', 'Marital', ‘PaymentMethod’)

# Remove the specified columns:
        mydata <- mydata[, -which(names(mydata) %in% columns_to_remove)]
```

Before the categorical data is converted to binary, an understanding and summarization of the categorical data that expresses more than 2 values within the variable is recommended. In this case, 3 categorical variables express more than 2 values within the variable: InternetService, Gender, and Contract.

Their summaries are as follows:

Code:

```
# Categorical data summaries based on churn:

# Internet Service

cd1 <- ggplot(mydata, aes(x = Churn, fill = InternetService)) +
  geom_bar(position = "dodge", color = "black", show.legend = TRUE) +
  geom_text(stat = "count", aes(label =
scales::percent(..count../sum(..count..)),
                y = ..count.., group = InternetService),
      position = position_dodge(width = 0.9),
      vjust = -0.5) +
  labs(title = "Churn Distribution by Internet Service",
     x = "Churn",
     y = "Count") +
   scale_fill_manual(values = wes_palette('Royal2' , n = 3)) +
  theme_minimal()

# Gender

cd2 <- ggplot(mydata, aes(x = Churn, fill = Gender)) +
  geom_bar(position = "dodge", color = "black", show.legend = TRUE) +
  geom_text(stat = "count", aes(label =
scales::percent(..count../sum(..count..)),
                y = ..count.., group = Gender),
      position = position_dodge(width = 0.9),
      vjust = -0.5) +
  labs(title = "Churn Distribution by Gender",
     x = "Churn",
     y = "Count") +
  scale_fill_manual(values = wes_palette('Royal2' , n = 3)) +
  theme_minimal()

# Contract

cd3 <- ggplot(mydata, aes(x = Churn, fill = Contract)) +
  geom_bar(position = "dodge", color = "black", show.legend = TRUE) +
  geom_text(stat = "count", aes(label =
scales::percent(..count../sum(..count..)),
                y = ..count.., group = Contract),
      position = position_dodge(width = 0.9),
      vjust = -0.5) +
```

```
  labs(title = "Churn Distribution by Contract",
       x = "Churn",
       y = "Count") +
  scale_fill_manual(values = wes_palette('Royal2' , n = 3)) +
  theme_minimal()
```

Without Churn

Interenet Service

```
cd_noc1 <- ggplot(mydata, aes(x = InternetService, fill = InternetService))
+
  geom_bar(position = "dodge", color = "black", show.legend = TRUE) +
  geom_text(stat = "count", aes(label =
scales::percent(..count../sum(..count..)),
                                y = ..count.., group = InternetService),
            position = position_dodge(width = 0.9),
            vjust = -0.5) +
  labs(title = "Distribution by Internet Service",
       x = "InternetService",
       y = "Count") +
  scale_fill_manual(values = wes_palette('Royal2' , n = 3)) +
  theme_minimal()
```

Gender

```
cd_noc2 <- ggplot(mydata, aes(x = Gender, fill = Gender)) +
  geom_bar(position = "dodge", color = "black", show.legend = TRUE) +
  geom_text(stat = "count", aes(label =
scales::percent(..count../sum(..count..)),
                                y = ..count.., group = Gender),
            position = position_dodge(width = 0.9),
            vjust = -0.5) +
  labs(title = "Distribution by Gender",
       x = "Gender",
       y = "Count") +
  scale_fill_manual(values = wes_palette('Royal2' , n = 3)) +
  theme_minimal()
```

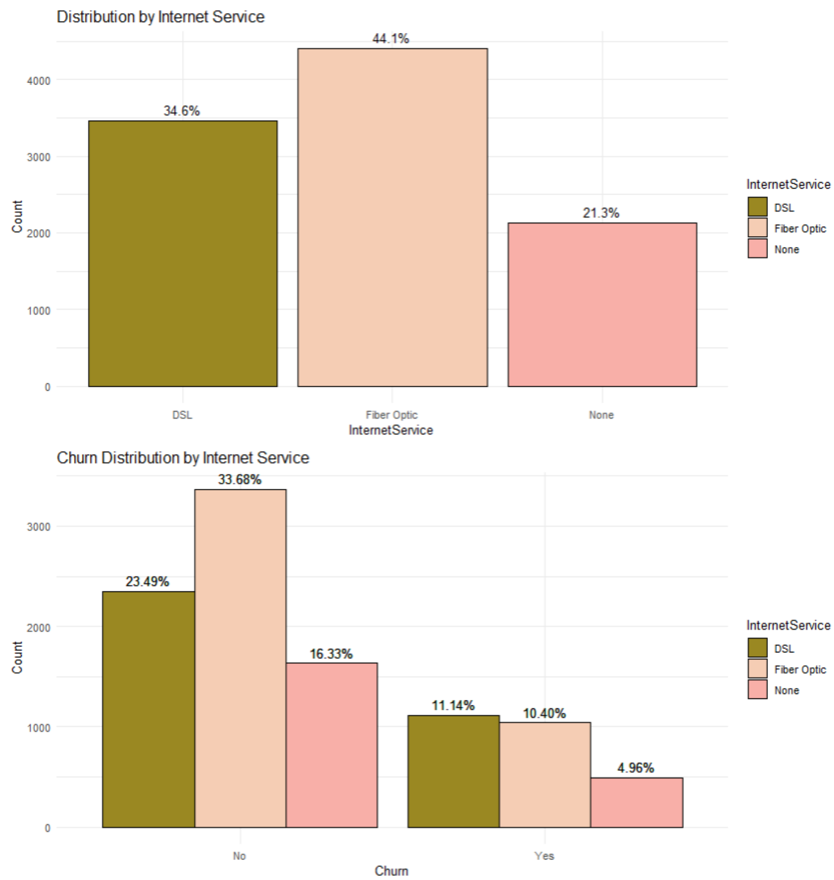
Contract

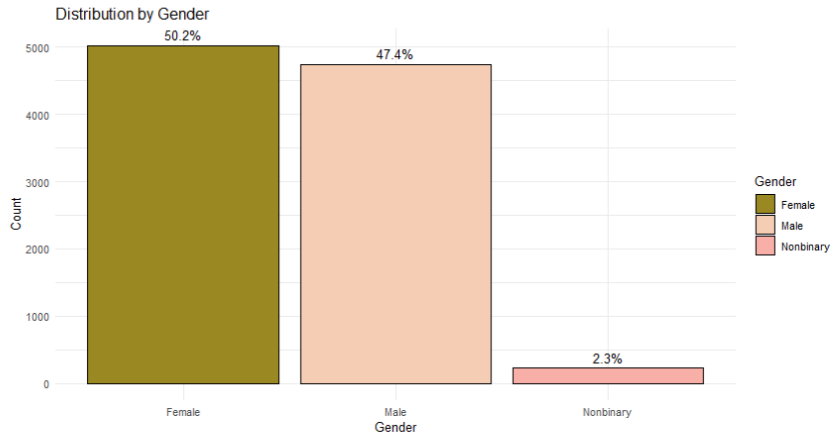
```
cd_noc3 <- ggplot(mydata, aes(x = Contract, fill = Contract)) +
  geom_bar(position = "dodge", color = "black", show.legend = TRUE) +
  geom_text(stat = "count", aes(label =
scales::percent(..count../sum(..count..)),
                                y = ..count.., group = Contract),
            position = position_dodge(width = 0.9),
            vjust = -0.5) +
  labs(title = "Distribution by Contract",
       x = "Contract",
       y = "Count") +
```

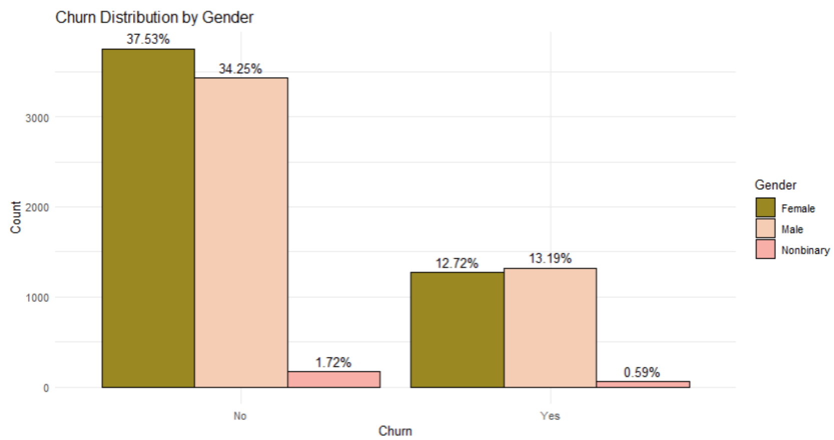
```
scale_fill_manual(values = wes_palette('Royal2' , n = 3)) +
theme_minimal()
```

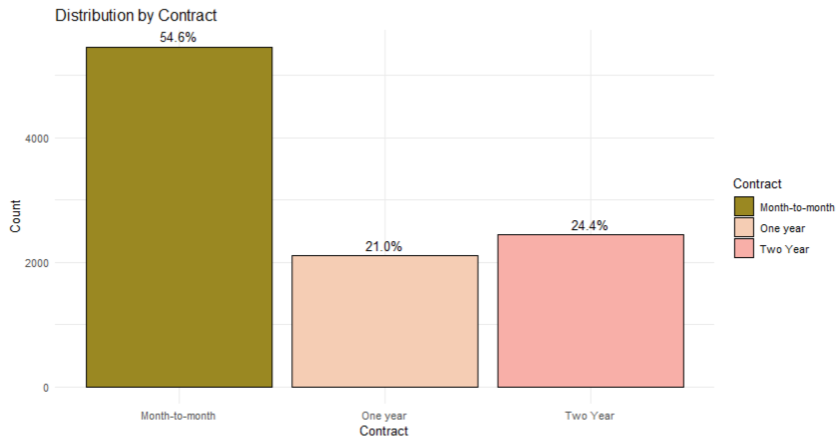
```
# Arranging the grids by variable:

grid.arrange(cd_noc1, cd1)
grid.arrange(cd_noc2, cd2)
grid.arrange(cd_noc3, cd3)
```

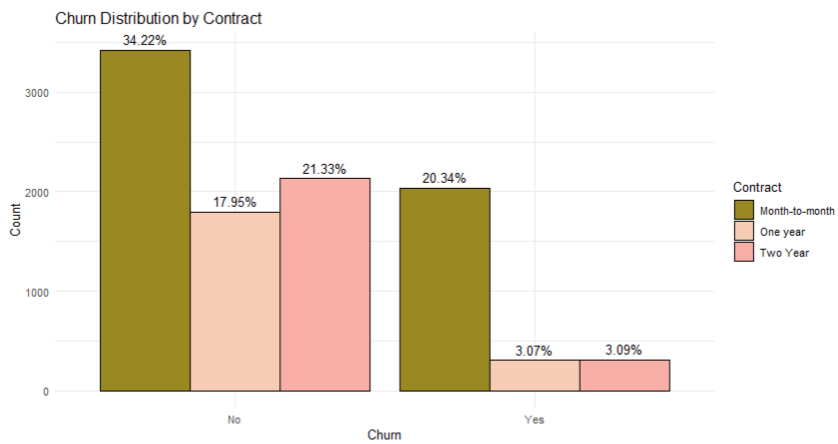
These are the visuals for each variable with and without visualizing their churn:




Distribution by Gender
50.2%
47.4%
2.3%
5000
4000
3000
2000
1000
0
Count
Female
Male
Nonbinary
Gender
Gender
Female
Male
Nonbinary


Churn Distribution by Gender
37.53%
34.25%
1.72%
12.72%
13.19%
0.59%
3000
2000
1000
0
Count
No
Yes
Churn
Gender
Female
Male
Nonbinary





Next, categorical data must be converted to numerical fields. To do this, a code is created to change all no's to 0 and all yes's to 1. The new variables will be known as Dummy variables.

The following is the code to convert to binary:

```
# Creating Dummy Variables for Categorical Data

mydata$DummyGender <- ifelse(mydata$Gender == 'Male', 1, 0)
mydata$DummyChurn <- ifelse(mydata$Churn == 'Yes', 1, 0)
mydata$DummyTechie <- ifelse(mydata$Techie == 'Yes', 1, 0)
mydata$DummyContract <- ifelse(mydata$Contract == 'Two Year', 1, 0)
mydata$DummyPort_modem <- ifelse(mydata$Port_modem == 'Yes', 1,
0)
mydata$DummyTablet <- ifelse(mydata$Tablet == 'Yes', 1, 0)
mydata$DummyInternetService <- ifelse(mydata$InternetService ==
'Fiber Optic', 1, 0)
mydata$DummyPhone <- ifelse(mydata$Phone == 'Yes', 1, 0)
mydata$DummyMultiple <- ifelse(mydata$Multiple == 'Yes', 1, 0)
mydata$DummyOnlineSecurity <- ifelse(mydata$OnlineSecurity == 'Yes',
1, 0)
mydata$DummyOnlineBackup <- ifelse(mydata$OnlineBackup == 'Yes',
1, 0)
```

```
mydata$DummyDeviceProtection <- ifelse(mydata$DeviceProtection ==
'Yes', 1, 0)
mydata$DummyTechSupport <- ifelse(mydata$TechSupport == 'Yes', 1,
0)
mydata$DummyStreamingTV <- ifelse(mydata$StreamingTV == 'Yes', 1,
0)
mydata$DummyStreamingMovies <- ifelse(mydata$StreamingMovies ==
'Yes', 1, 0)
mydata$DummyPaperlessBilling <- ifelse(mydata$PaperlessBilling ==
'Yes', 1, 0)
```

Now, all the original categorical variables will be removed from the dataset.

Code:

```
# Dropping all old categorical variables:

    remove_original_categories <- c('Gender', 'Churn', 'Techie', 'Contract',
    'Port_modem', 'Tablet', 'InternetService', 'Phone', 'Multiple',
    'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
    'TechSupport', 'StreamingTV', 'StreamingMovies',
    'PaperlessBilling')

    mydata <- mydata[, -which(names(mydata) %in%
    remove_original_categories)]
```

Rechecking the structure of the data to make sure variables were removed properly:

Code:

```
str(mydata)
```

Output:

```
> str(mydata)
'data.frame':	10000 obs. of  34 variables:
 $ Children             : int  0 1 4 1 0 3 0 2 2 1 ...
 $ Age                  : int  68 27 50 48 83 83 79 30 49 86 ...
 $ Income               : num  28562 21705 9610 18925 40074 ...
 $ Outage_sec_perweek   : num  7.98 11.7 10.75 14.91 8.15 ...
 $ Email                : int  10 12 9 15 16 15 10 16 20 18 ...
 $ Contacts             : int  0 0 0 2 2 3 0 0 2 1 ...
 $ Yearly_equip_failure : int  1 1 1 0 1 1 1 0 3 0 ...
 $ Tenure               : num  6.8 1.16 15.75 17.09 1.67 ...
 $ MonthlyCharge        : num  172 243 160 120 150 ...
 $ Bandwidth_GB_Year    : num  905 801 2055 2165 271 ...
 $ Item1                : int  5 3 4 4 4 3 6 2 5 2 ...
 $ Item2                : int  5 4 4 4 4 3 5 2 4 2 ...
 $ Item3                : int  5 3 2 4 4 3 6 2 4 2 ...
 $ Item4                : int  3 3 4 2 3 2 4 5 3 2 ...
 $ Item5                : int  4 4 4 5 4 4 1 2 4 5 ...
 $ Item6                : int  4 3 3 4 4 3 5 3 3 2 ...
 $ Item7                : int  3 4 3 3 4 3 5 4 4 3 ...
 $ Item8                : int  4 4 3 3 5 3 5 5 4 3 ...
 $ DummyGender          : num  1 0 0 1 1 0 1 0 0 0 ...
 $ DummyChurn           : num  0 1 0 0 1 0 1 1 0 0 ...
 $ DummyTechie          : num  0 1 1 1 0 0 1 1 0 0 ...
 $ DummyContract        : num  0 0 1 1 0 0 0 0 0 1 ...
 $ DummyPort_modem      : num  1 0 1 0 1 1 0 0 1 1 ...
 $ DummyTablet          : num  1 1 0 0 0 0 0 0 0 0 ...
 $ DummyInternetService : num  1 1 0 0 1 0 0 0 0 1 ...
 $ DummyPhone           : num  1 1 1 1 0 1 1 0 1 1 ...
 $ DummyMultiple        : num  0 1 1 0 0 1 0 0 0 0 ...
 $ DummyOnlineSecurity  : num  1 1 0 1 0 1 0 0 1 1 ...
 $ DummyOnlineBackup    : num  1 0 0 0 0 1 0 1 1 0 ...
 $ DummyDeviceProtection: num  0 0 0 0 0 1 0 0 0 1 ...
 $ DummyTechSupport     : num  0 0 0 0 1 0 1 0 0 0 ...
 $ DummyStreamingTV     : num  0 1 0 1 1 0 1 0 0 0 ...
 $ DummyStreamingMovies : num  1 1 1 0 0 1 1 0 0 1 ...
 $ DummyPaperlessBilling: num  1 1 1 1 0 0 0 1 1 1 ...
```

The columns have been removed, replaced, and categories are now binary.

Now, look at the summary of the data to see if there is any missing data is important:

Look for missing data points via summary()

Summary(mydata)

The output of this command:

```
> summary(mydata)
    Children           Age            Income         Outage_sec_perweek     Email          Contacts       Yearly_equip_failure
 Min.   : 0.000   Min.   :18.00   Min.   :   348.7   Min.   : 0.09975   Min.   : 1.00   Min.   :0.0000   Min.   :0.000
 1st Qu.: 0.000   1st Qu.:35.00   1st Qu.: 19224.7   1st Qu.: 8.01821   1st Qu.:10.00   1st Qu.:0.0000   1st Qu.:0.000
 Median : 1.000   Median :53.00   Median : 33170.6   Median :10.01856   Median :12.00   Median :1.0000   Median :0.000
 Mean   : 2.088   Mean   :53.08   Mean   : 39806.9   Mean   :10.00185   Mean   :12.02   Mean   :0.9942   Mean   :0.398
 3rd Qu.: 3.000   3rd Qu.:71.00   3rd Qu.: 53246.2   3rd Qu.:11.96949   3rd Qu.:14.00   3rd Qu.:2.0000   3rd Qu.:1.000
 Max.   :10.000   Max.   :89.00   Max.   :258900.7   Max.   :21.20723   Max.   :23.00   Max.   :7.0000   Max.   :6.000
     Tenure        MonthlyCharge    Bandwidth_GB_Year      Item1           Item2           Item3           Item4           Item5
 Min.   : 1.000   Min.   : 79.98   Min.   : 155.5   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.: 7.918   1st Qu.:139.98   1st Qu.:1236.5   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000
 Median :35.431   Median :167.48   Median :3279.5   Median :3.000   Median :4.000   Median :3.000   Median :3.000   Median :3.000
 Mean   :34.526   Mean   :172.62   Mean   :3392.3   Mean   :3.491   Mean   :3.505   Mean   :3.487   Mean   :3.498   Mean   :3.493
 3rd Qu.:61.480   3rd Qu.:200.73   3rd Qu.:5586.1   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :71.999   Max.   :290.16   Max.   :7159.0   Max.   :7.000   Max.   :7.000   Max.   :8.000   Max.   :7.000   Max.   :7.000
     Item6           Item7           Item8         DummyGender       DummyChurn     DummyTechie     DummyContract    DummyPort_modem
 Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:3.000   1st Qu.:3.00   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :3.000   Median :4.00   Median :3.000   Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :3.497   Mean   :3.51   Mean   :3.496   Mean   :0.4744   Mean   :0.265   Mean   :0.1679   Mean   :0.2442   Mean   :0.4834
 3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :8.000   Max.   :7.00   Max.   :8.000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
  DummyTablet     DummyInternetService   DummyPhone      DummyMultiple    DummyOnlineSecurity DummyOnlineBackup DummyDeviceProtection
 Min.   :0.0000   Min.   :0.0000       Min.   :0.0000   Min.   :0.0000   Min.   :0.0000      Min.   :0.0000    Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000       1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000    1st Qu.:0.0000
 Median :0.0000   Median :0.0000       Median :1.0000   Median :0.0000   Median :0.0000      Median :0.0000    Median :0.0000
 Mean   :0.2991   Mean   :0.4408       Mean   :0.9067   Mean   :0.4608   Mean   :0.3576      Mean   :0.4506    Mean   :0.4386
 3rd Qu.:1.0000   3rd Qu.:1.0000       3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000      3rd Qu.:1.0000    3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000       Max.   :1.0000   Max.   :1.0000   Max.   :1.0000      Max.   :1.0000    Max.   :1.0000
 DummyTechSupport DummyStreamingTV DummyStreamingMovies DummyPaperlessBilling
 Min.   :0.000   Min.   :0.0000   Min.   :0.000        Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000        1st Qu.:0.0000
 Median :0.000   Median :0.0000   Median :0.000        Median :1.0000
 Mean   :0.375   Mean   :0.4929   Mean   :0.489        Mean   :0.5882
 3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.000        3rd Qu.:1.0000
 Max.   :1.000   Max.   :1.0000   Max.   :1.000        Max.   :1.0000
```

The summary of the data shows none of the variables are missing any data (i.e., no blanks or NA's).

Lastly, to make the readability of the ordinal data easier, Item1 – Item8 will be renamed.

```
# Changing Column Names of Ordinal Data:

colnames(mydata)[colnames(mydata) == 'Item1'] <- 'Response'
colnames(mydata)[colnames(mydata) == 'Item2'] <- 'Fixes'
colnames(mydata)[colnames(mydata) == 'Item3'] <- 'Replacements'
colnames(mydata)[colnames(mydata) == 'Item4'] <- 'Reliability'
colnames(mydata)[colnames(mydata) == 'Item5'] <- 'Options'
colnames(mydata)[colnames(mydata) == 'Item6'] <- 'RespectfulResponse'
colnames(mydata)[colnames(mydata) == 'Item7'] <- 'CourtExchange'
colnames(mydata)[colnames(mydata) == 'Item8'] <- 'ActiveListening'
```

The next step involves investigating the remaining data further which will be to utilize both univariate and bivariate methods.

3. **Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.**

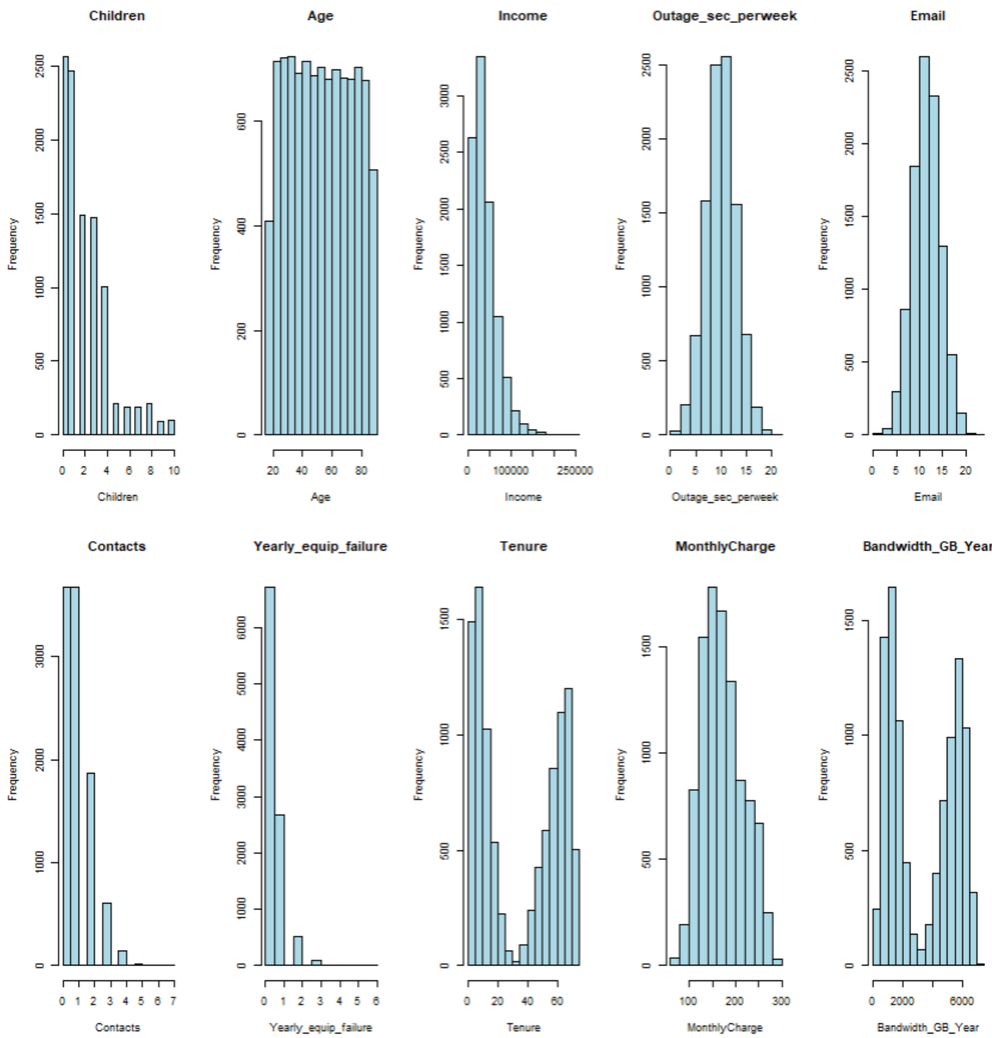
To begin, analyzing both continuous and categorical variables is required.

First, in the **univariate analysis** of continuous variable is necessary to ensure the data is accurate and does not interfere with the integrity of the analysis. Histograms will allow for the proper analysis of the data as will boxplots.

Histograms of Continuous Variables:

Code:

```
# Creating histograms for continuous variables by choosing variables first:

    selected_columns <- c('Children', 'Age', 'Income', 'Outage_sec_perweek',
    'Email', 'Contacts', 'Yearly_equip_failure', 'Tenure', 'MonthlyCharge',
    'Bandwidth_GB_Year')

# Create the layout for multiple histograms in a visualization (2 rows, 5 columns):

    par(mfrow = c(2, 5))

# Creating the histograms:

    for (col in selected_columns) {
      hist(churn_df[[col]], main = col, xlab = col, col = "lightblue")
    }
```



<u>Boxplot for each continuous variable:</u>

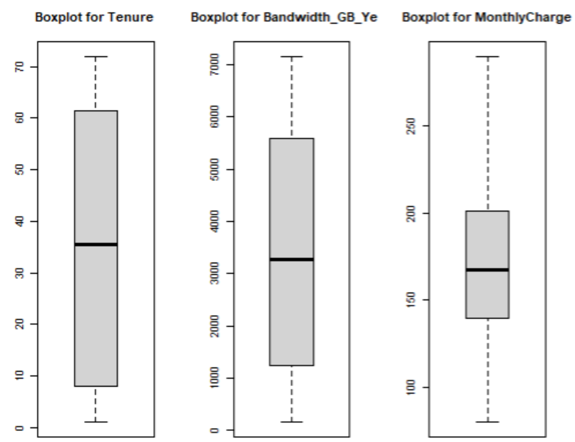
1. Tenure – No outliers present.

2. MonthlyCharges – No outliers present.

3. Bandwidth_GB_Year – No outliers present.

Boxplot for variables to check for outliers:

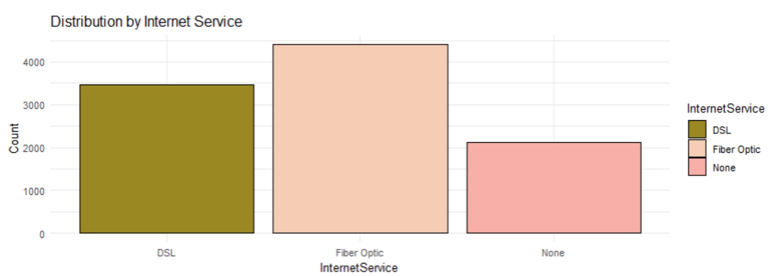
```
boxplot(mydata$Tenure, main = 'Boxplot for Tenure')$out
boxplot(mydata$Bandwidth_GB_Year, main = 'Boxplot for
Bandwidth_GB_Year')$out
boxplot(mydata$MonthlyCharge, main = 'Boxplot for MonthlyCharge')$out
```

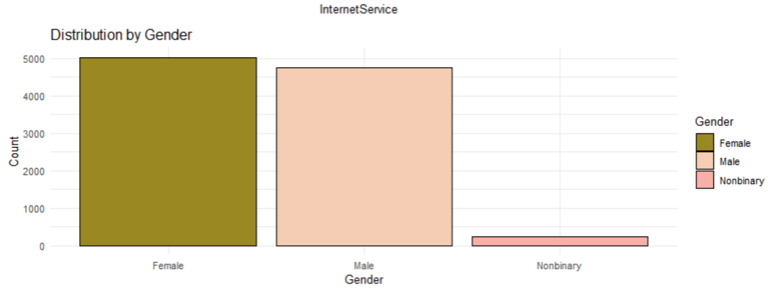


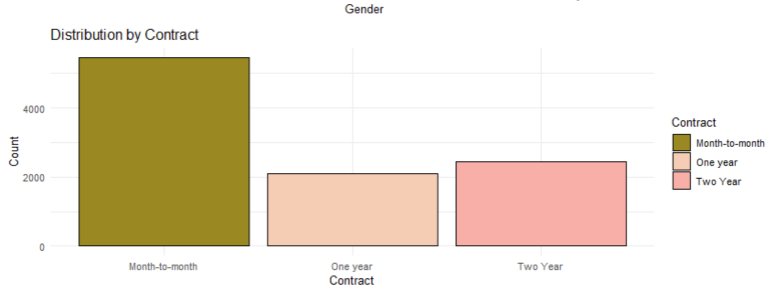
There does not appear to be any outliers in the modified dataset.

Next, several of the remaining independent variables are categorical. It is important to summarize these using univariate analysis.

As stated above, prior to converting the categorical data to binary a summarization was completed. Below are the visualizations for these:







More than half of the customers (>78%) have some form of internet service such as DSL or fiber optic. Over half of the customers are female (50.2%) and a very small percentage of customers are nonbinary (2.3%). Lastly, more than half of the customers are on a month-to-month contract (54.6%) while the other customers have either a one-year or two-year contract.

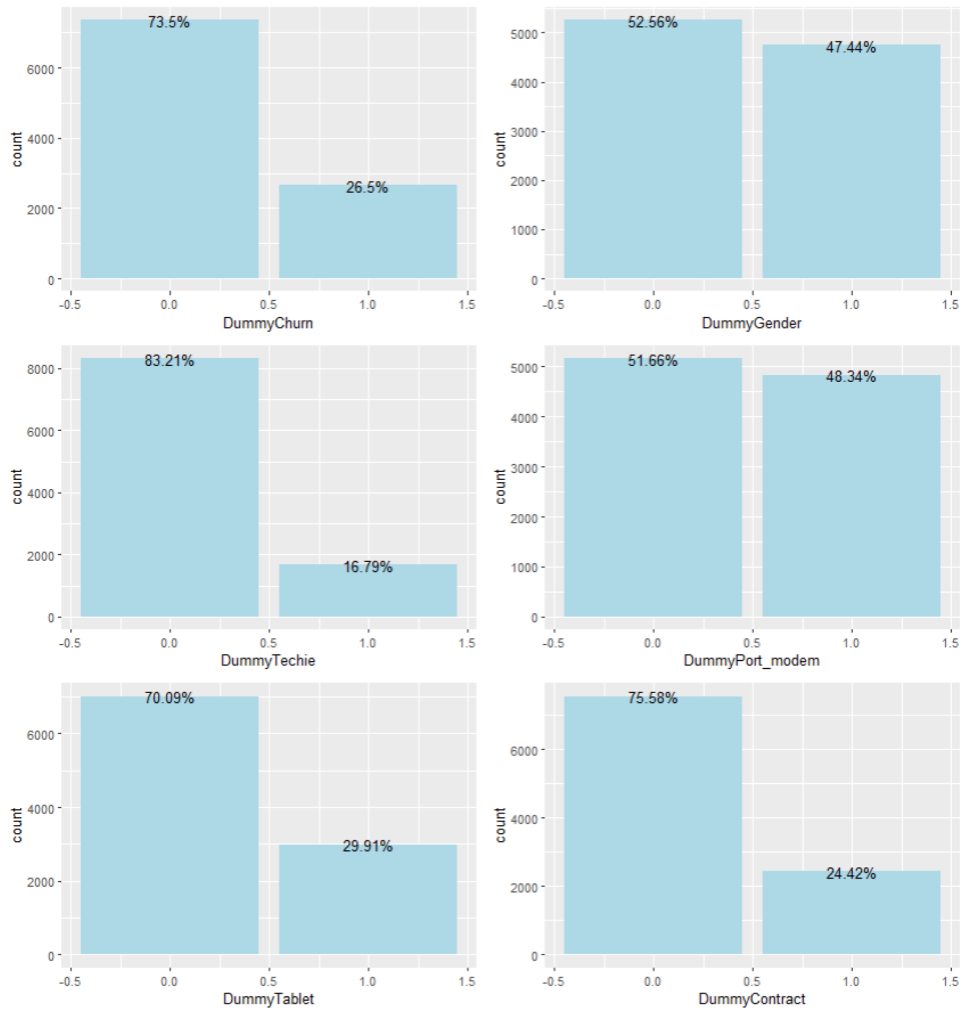
Once the categorical variables were converted to binary the following is their summaries.

This is the following code to create the visualizations that allow summaries of the categorical variables:

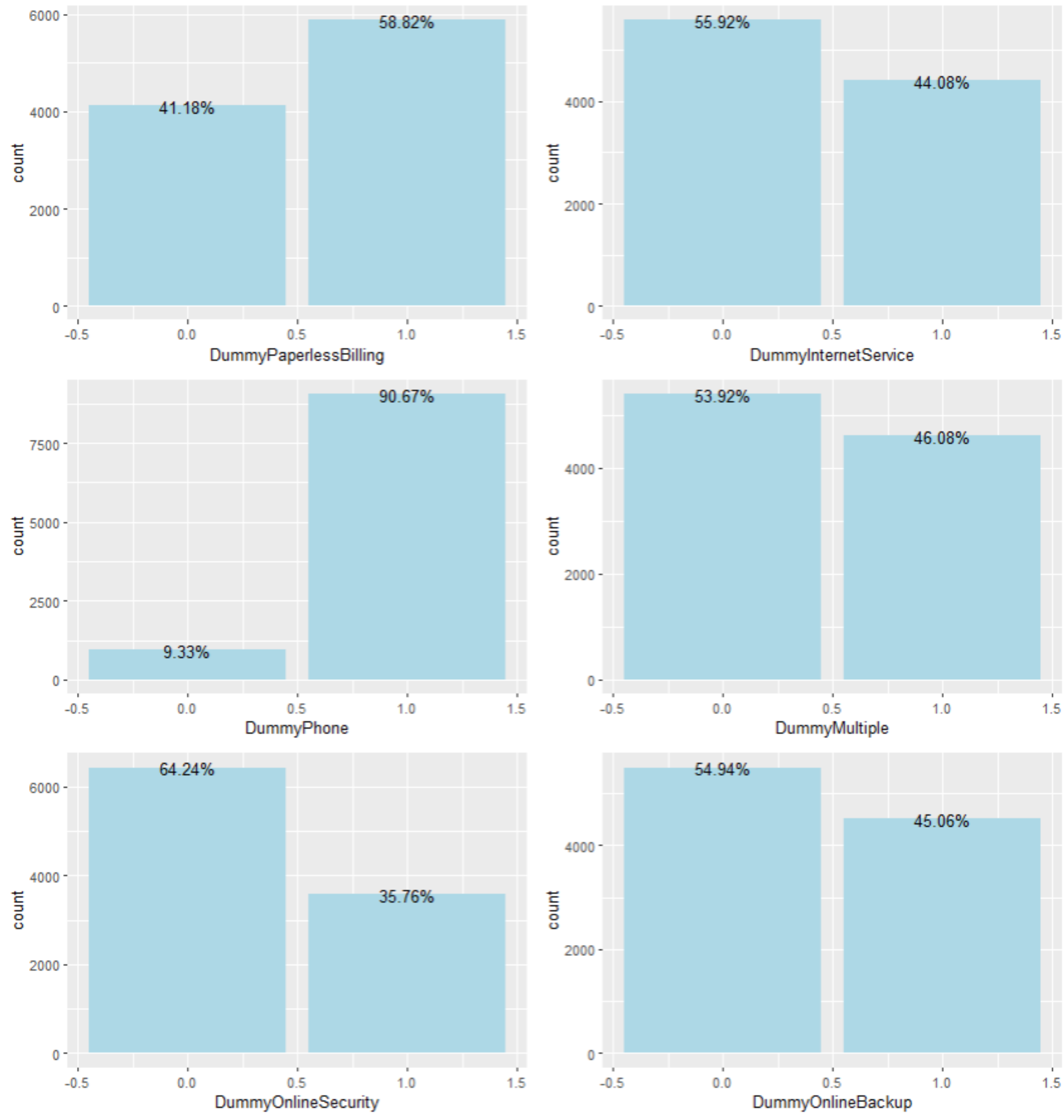
```
# Summary of Independent Variables

Churn_Summary <- ggplot(mydata, aes(x = DummyChurn)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Gender_Summary <- ggplot(mydata, aes(x = DummyGender)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Techie_Summary <- ggplot(mydata, aes(x = DummyTechie)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Port_modem_Summary <- ggplot(mydata, aes(x = DummyPort_modem)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Tablet_Summary <- ggplot(mydata, aes(x = DummyTablet)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Contract_Summary <- ggplot(mydata, aes(x = DummyContract)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')

PaperlessBilling_Summary <- ggplot(mydata, aes(x = DummyPaperlessBilling)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
InternetService_Summary <- ggplot(mydata, aes(x = DummyInternetService)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Phone_Summary <- ggplot(mydata, aes(x = DummyPhone)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
Multiple_Summary <- ggplot(mydata, aes(x = DummyMultiple)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
```

```
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
OnlineSecruity_Summary <- ggplot(mydata, aes(x = DummyOnlineSecurity)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
OnlineBackup_Summary <- ggplot(mydata, aes(x = DummyOnlineBackup)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')

DeviceProtection_Summary <- ggplot(mydata, aes(x =
DummyDeviceProtection)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
TechSupport_Summary <- ggplot(mydata, aes(x = DummyTechSupport)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
StreamingTV_Summary <- ggplot(mydata, aes(x = DummyStreamingTV)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')
StreamingMovies_Summary <- ggplot(mydata, aes(x =
DummyStreamingMovies)) +
  geom_bar(position = 'dodge', stat = 'count', fill = 'lightblue') +
  geom_text(aes(label = paste0(round(prop.table(after_stat(count)) * 100, 2),
'%')), stat = 'count')

grid.arrange(Churn_Summary, Gender_Summary, Techie_Summary,
Port_modem_Summary, Tablet_Summary, Contract_Summary)
grid.arrange(PaperlessBilling_Summary, InternetService_Summary,
Phone_Summary,
          Multiple_Summary, OnlineSecruity_Summary, OnlineBackup_Summary)
grid.arrange(OnlineBackup_Summary, DeviceProtection_Summary,
TechSupport_Summary, StreamingTV_Summary, StreamingMovies_Summary)
```

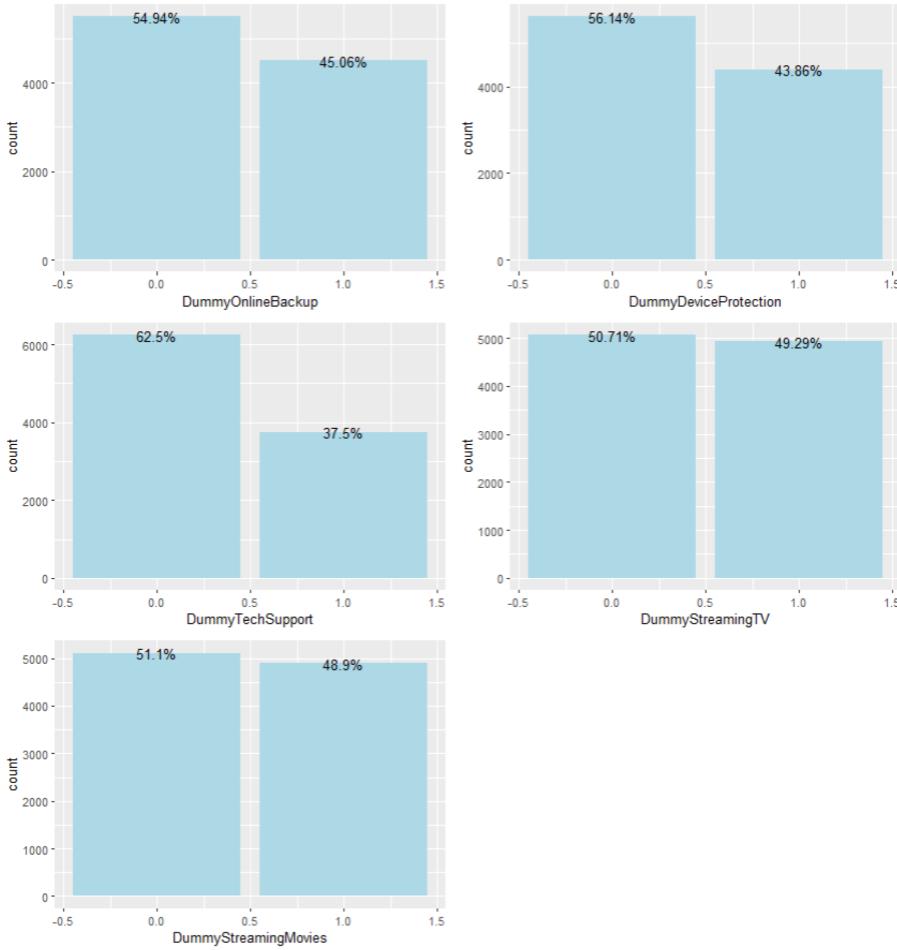
Once these were created, using grid.arrange() allowed the create the following visualizations (they were broken up into 3 grids to make it more readable):



As seen in the above visual, almost 75% of the customers have not churned with a little over 25% churning. More than half of the customers are Female/Binary (female/binary = 0 and male = 1). Many customers do not see themselves as technically inclined. Over half of the customers do not use a port modem. Over 70% of customers do not use tablets and over half of the customers are on a month-to-month contract with the company.



As seen in the above visualization, more customers (> 50%) have chosen paperless billing. More customers have fiber optics over DSL/none (fiber optics = 0, DSL/none = 1). In addition, over 90% of their customers use the phone service. Although over 50% do not have multiple lines. More customers have opted out of having online security (less than 40% have it).



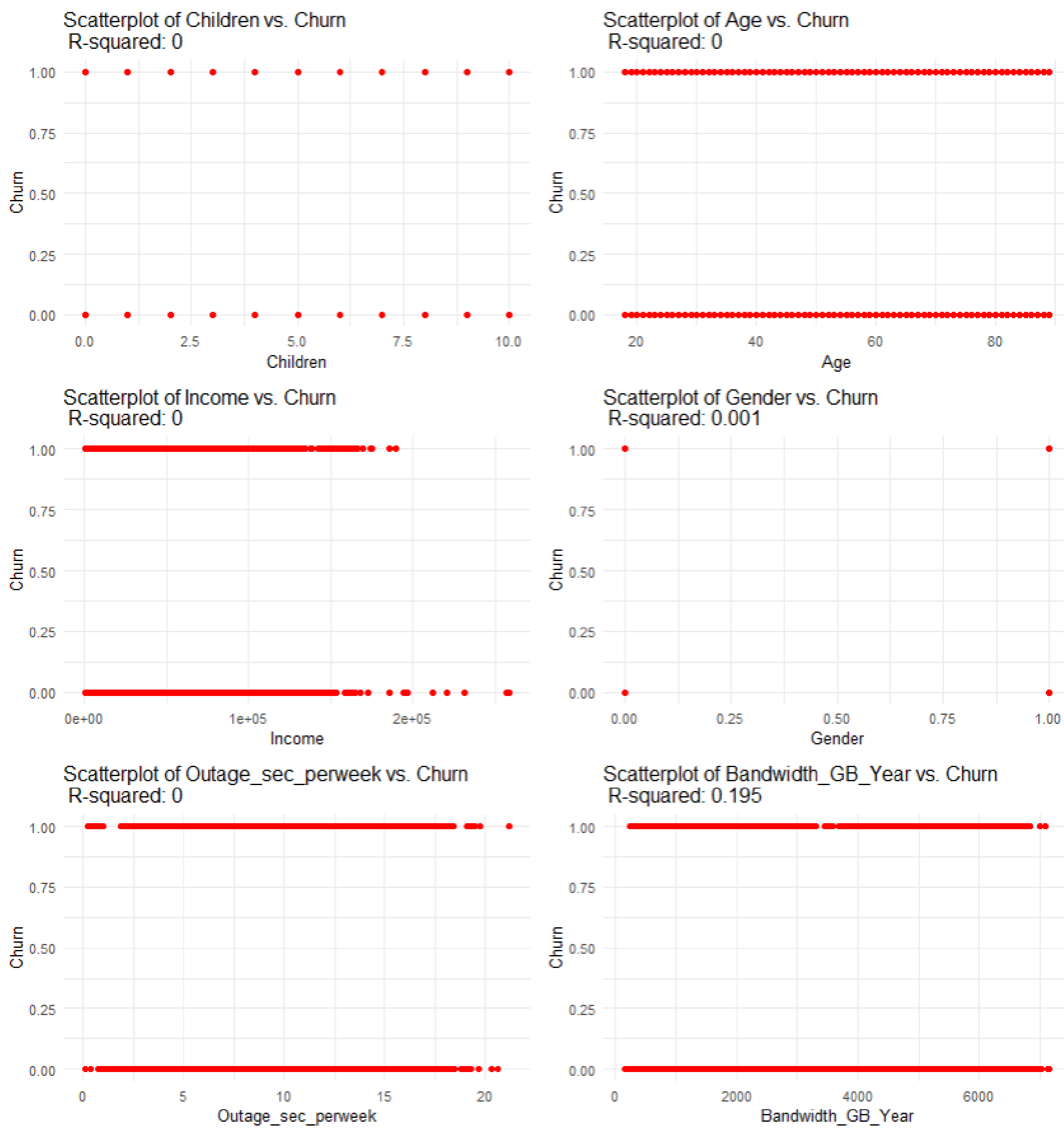
As seen in the above visualization, over 56% of their customers do not have online backups. More than half (> 56%) do not have device protection on their devices. Additionally, only 36% of customers have a technical support add-on. When it comes up Streaming TV, there is an almost 50/50 on customers that have it in comparison to customer that do not. Same with streaming movies (over 50% do not).

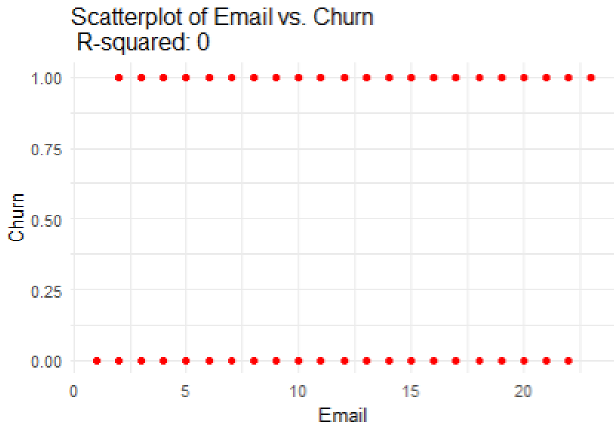
Next, **bivariate statistics** are conducted. This is a logistic regression model and binary values are necessary for analysis. One of the appropriate ways to see the relationships between Churn and the other variables is scatterplots with ggplot.

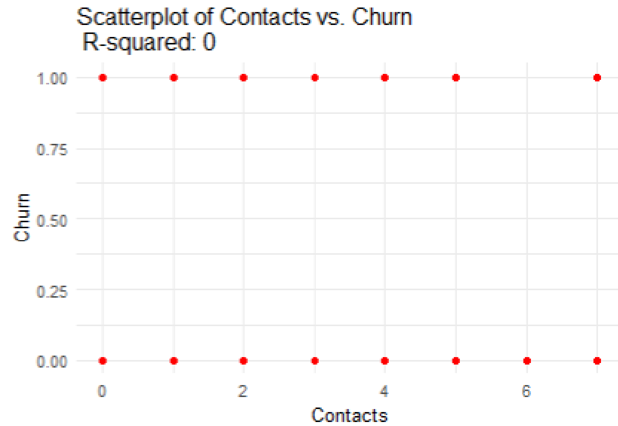
All code for the ggplot() is as follows but changing the **X** variable for each execute:

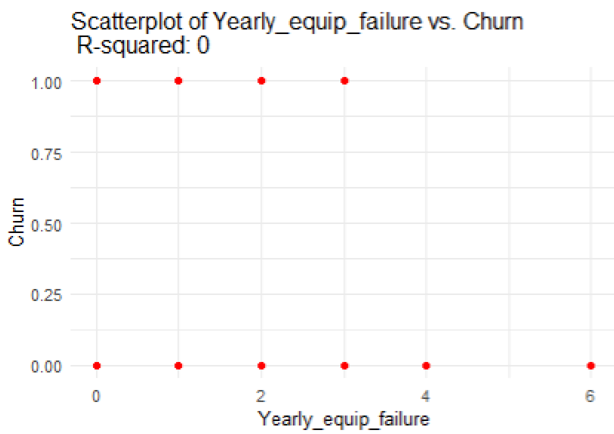
```
# Create scatterplot x = Children, y = Churn:

sp1 <- ggplot(mydata, aes(x = Children, y = DummyChurn)) +
  geom_point(color = 'red') +
  labs(title = paste('Scatterplot of Children vs. Churn\n',
```

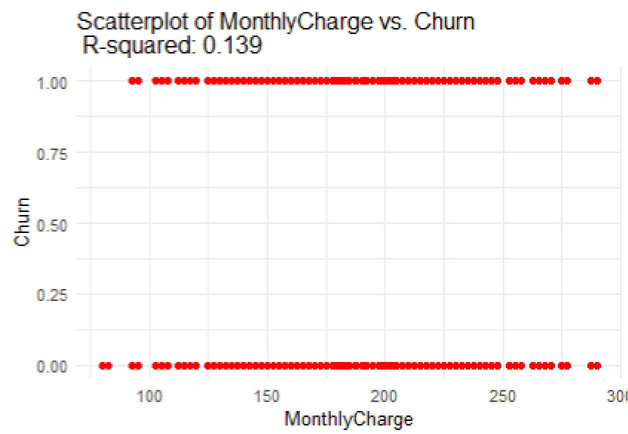
'R-squared:', round(cor(mydata$**Children**, mydata$DummyChurn)^2, 3)),
x = '**Children**',
y = 'Churn') +
theme_minimal()

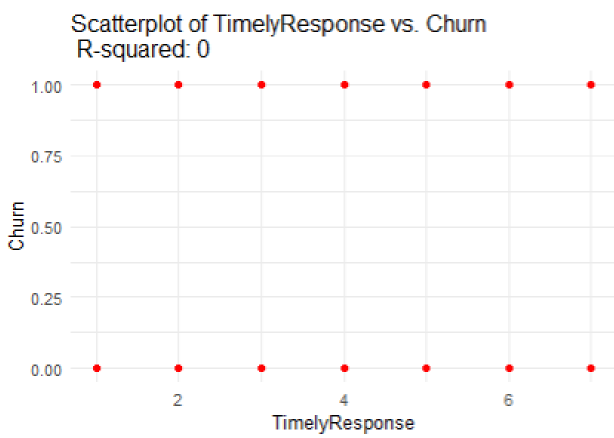


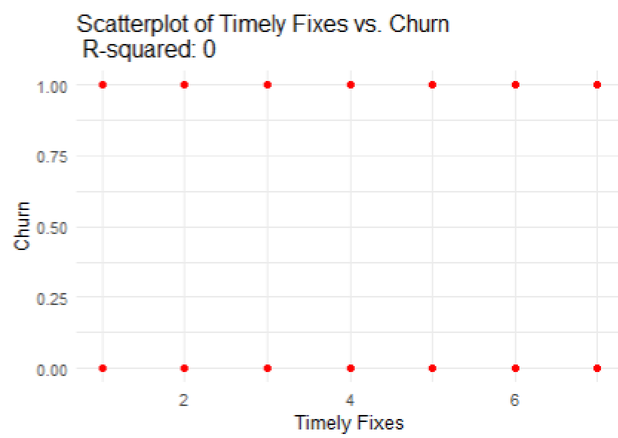

Scatterplot of Email vs. Churn
R-squared: 0
Churn
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
Email

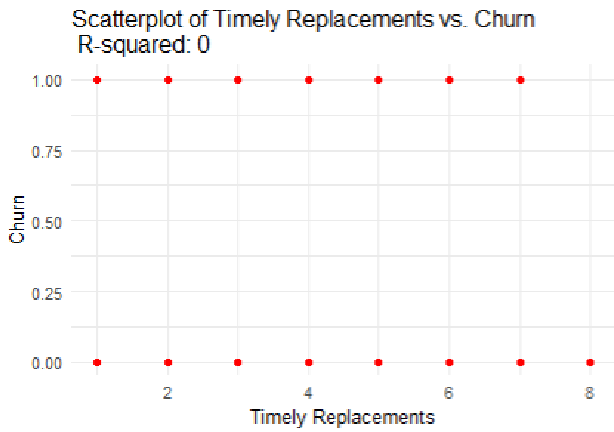

Scatterplot of Contacts vs. Churn
R-squared: 0
Churn
1.00
0.75
0.50
0.25
0.00
0
2
4
6
Contacts

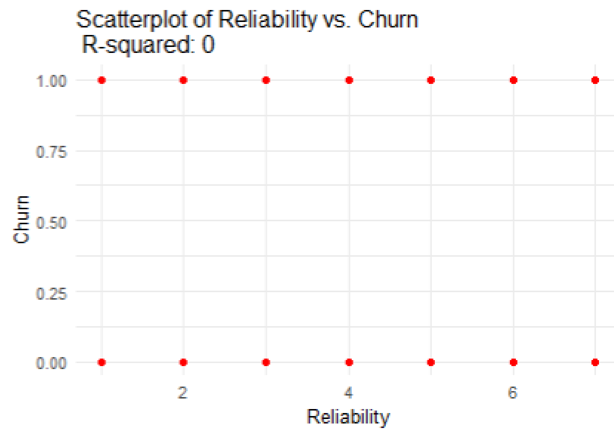

Scatterplot of Yearly_equip_failure vs. Churn
R-squared: 0
Churn
1.00
0.75
0.50
0.25
0.00
0
2
4
6
Yearly_equip_failure

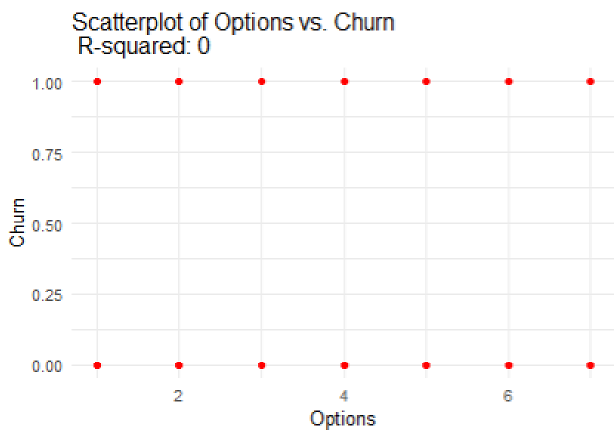

Scatterplot of MonthlyCharge vs. Churn
R-squared: 0.139
Churn
1.00
0.75
0.50
0.25
0.00
100
150
200
250
MonthlyCharge

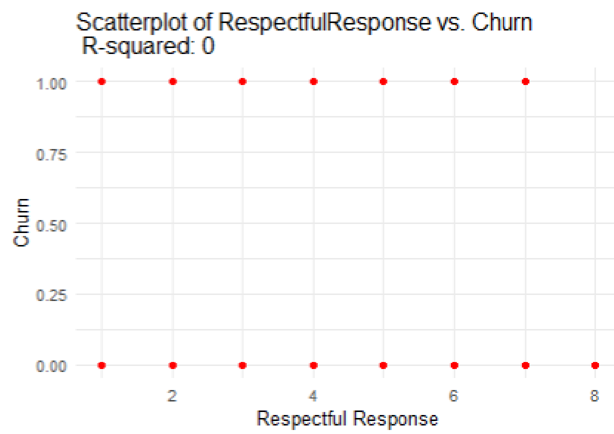

Scatterplot of TimelyResponse vs. Churn
R-squared: 0
Churn
1.00
0.75
0.50
0.25
0.00
2
4
6
TimelyResponse

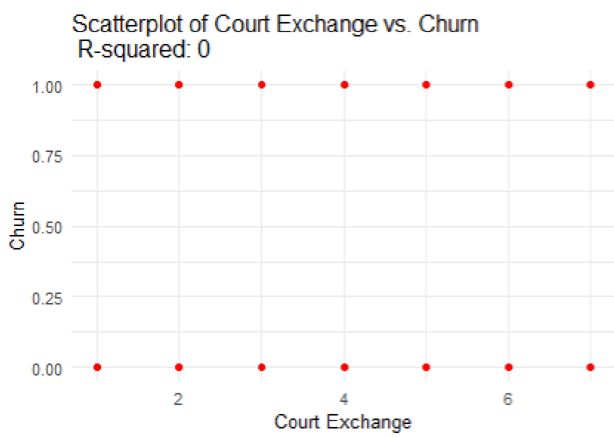

Scatterplot of Timely Fixes vs. Churn
R-squared: 0
Churn
1.00
0.75
0.50
0.25
0.00
2
4
6
Timely Fixes

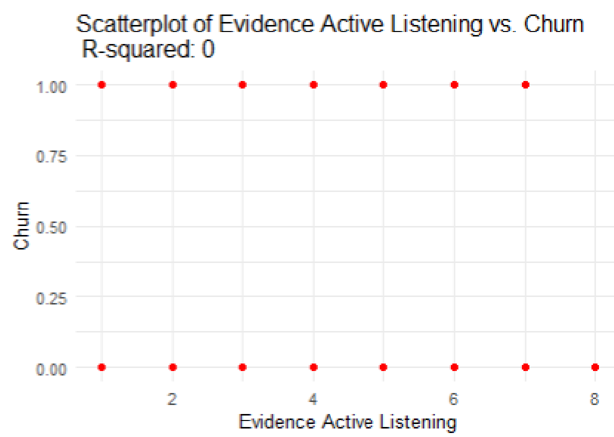












All scatterplots express a low R-Square except Churn vs. Bandwidth_GB_Year, MonthlyCharge and an extremely small R-Square with Gender. The R-Square Value for these scatterplots were 0.195, 0.139, and 0.001 respectively. These are considered to have a low correlation but, more analysis is needed to understand the relationship between Churn and the other independent variables. An R-square value does not necessarily mean causation.

Further analysis will help determine if the mentioned variables help to predict higher churn rates of customers such as running a Logistic Regression Analysis.

4. **Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.**

   My data transformation goals were to ensure the data was properly cleaned. Also, I wished to address any data error, anomalies, null or blank data, etc. None were found within the dataset. Outliers were not detected in the selected continuous variables.

   The steps to transform the data, including the annotated code, can be found in the previous questions answered above. To further achieve the goals of the study, an investigation using multiple linear regression will be conducted.

5. **Provide the prepared data set as a CSV file.**

```
# .csv of data transformation

    write.csv(mydata, file = 'modified_dataset.csv', row.names = FALSE)
```

   This will be uploaded with the assessment.

Part IV: Model Comparison and Analysis

D. Compare an initial and a reduced logistic regression model by doing the following:

1. **Construct an initial logistic regression model from *all* independent variables that were identified in part C2.**

   The logistic regression model was performed to include all independent variables with Churn being the dependent variable.

   Code:

```
# Fit a logistic regression model with all predictors with Churn being the
dependent

    logistic_model_all <- glm(DummyChurn ~ ., data = mydata, family =
    binomial)


# Printing out the results:

    Print(logistic_model_all)
```

```
> print(logistic_model_all)

Call:  glm(formula = DummyChurn ~ ., family = binomial, data = mydata)

Coefficients:
          (Intercept)              Children                   Age                Income     Outage_sec_perweek
           -4.876e+00            -5.036e-02             8.181e-03             2.976e-07              5.539e-04
                Email              Contacts   Yearly_equip_failure                Tenure          MonthlyCharge
           -1.768e-03             2.894e-02            -3.326e-02            -2.354e-01              2.901e-02
     Bandwidth_GB_Year              Response                 Fixes          Replacements            Reliability
            1.721e-03            -1.759e-02             2.167e-02            -1.820e-02             -2.012e-02
              Options            Respectful          CourtExchange       ActiveListening            DummyGender
           -3.007e-02            -3.442e-02             5.353e-03            -8.250e-03              1.092e-01
          DummyTechie         DummyContract        DummyPort_modem           DummyTablet   DummyInternetService
            8.157e-01            -2.288e+00             1.536e-01            -7.525e-02             -9.108e-01
           DummyPhone         DummyMultiple    DummyOnlineSecurity     DummyOnlineBackup  DummyDeviceProtection
           -3.291e-01             2.553e-01            -3.132e-01            -1.576e-01             -2.319e-01
     DummyTechSupport      DummyStreamingTV   DummyStreamingMovies  DummyPaperlessBilling
           -1.220e-01             6.961e-01             9.203e-01             1.127e-01

Degrees of Freedom: 9999 Total (i.e. Null);  9966 Residual
Null Deviance:      11560
Residual Deviance: 5419        AIC: 5487
```

A total of 34 variables (Including Churn): Churn = -4.876 (intercept) - 5.036e-02 (Children) + 8.181e-03 (Age) + 2.976e-07 (Income) + 5.539e-04 (Outage_sec_perweek) - 1.768e-03 (Email) + 2.894e-02 (Contacts) - 3.326e-02 (Yearly_equip_failure) – 0.2354 (Tenure) + 2.901e-02 (MonthlyCharge) + 1.721e-03 (Bandwidth_GB_Year) - 1.759e-02 (Response) + 2.167e-02 (Fixes) - 1.820e-02 (Replacements) - 2.012e-02 (Reliability) – 3.007e-02 (Options) - 3.442e-02 (Respectful) + 5.353e-03 (CourtExchange) - 8.250e-03 (ActiveListening) + 0.1092 (DummyGender) + 0.8157 (DummyTechie) - 2.288 (DummyContract) 0.1536 (DummyPort_Modem) - 7.525e-02 (DummyTablet) – 0.9108 (DummyInternetService) – 0.3291 (DummyPhone) + 0.2553 (DummyMultiple) – 0.3132 (DummyOnlineSecruity) – 0.1576 (DummyOnlineBackup) – 0.2319 (DummyDeviceProtection) – 0.1220 (DummyTechSupport) + 0.6961 (DummyStreamingTV) + 0.9203 (DummyStreamingMovies) + 0.1127 (DummyPaperlessBilling)

To further the understanding of the model, a summary of the model is important:

Code:

```
summary(logistic_model_all)
```

```
Call:
glm(formula = DummyChurn ~ ., family = binomial, data = mydata)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -4.876e+00  4.982e-01  -9.788  < 2e-16 ***
Children               -5.036e-02  1.818e-02  -2.770 0.005600 **
Age                     8.181e-03  1.944e-03   4.208 2.58e-05 ***
Income                  2.976e-07  1.223e-06   0.243 0.807781
Outage_sec_perweek      5.539e-04  1.155e-02   0.048 0.961758
Email                  -1.768e-03  1.138e-02  -0.155 0.876515
Contacts                2.894e-02  3.467e-02   0.835 0.403863
Yearly_equip_failure   -3.326e-02  5.429e-02  -0.613 0.540087
Tenure                 -2.354e-01  2.404e-02  -9.791  < 2e-16 ***
MonthlyCharge           2.901e-02  4.747e-03   6.112 9.83e-10 ***
Bandwidth_GB_Year       1.721e-03  2.920e-04   5.893 3.79e-09 ***
Response               -1.759e-02  4.890e-02  -0.360 0.718988
Fixes                   2.167e-02  4.614e-02   0.470 0.638618
Replacements           -1.820e-02  4.202e-02  -0.433 0.664854
Reliability            -2.012e-02  3.731e-02  -0.539 0.589751
Options                -3.007e-02  3.904e-02  -0.770 0.441138
Respectful             -3.442e-02  3.998e-02  -0.861 0.389294
CourtExchange           5.353e-03  3.813e-02   0.140 0.888362
ActiveListening        -8.250e-03  3.611e-02  -0.228 0.819283
DummyGender             1.092e-01  7.116e-02   1.535 0.124803
DummyTechie             8.157e-01  8.946e-02   9.117  < 2e-16 ***
DummyContract          -2.288e+00  1.028e-01 -22.247  < 2e-16 ***
DummyPort_modem         1.536e-01  6.870e-02   2.235 0.025395 *
DummyTablet            -7.525e-02  7.466e-02  -1.008 0.313482
DummyInternetService   -9.108e-01  1.884e-01  -4.834 1.34e-06 ***
DummyPhone             -3.291e-01  1.171e-01  -2.811 0.004941 **
DummyMultiple           2.553e-01  1.585e-01   1.610 0.107303
DummyOnlineSecurity    -3.132e-01  7.403e-02  -4.230 2.33e-05 ***
DummyOnlineBackup      -1.576e-01  1.144e-01  -1.378 0.168249
DummyDeviceProtection  -2.319e-01  8.370e-02  -2.770 0.005602 **
DummyTechSupport       -1.220e-01  9.265e-02  -1.317 0.187772
DummyStreamingTV        6.961e-01  1.850e-01   3.762 0.000169 ***
DummyStreamingMovies    9.203e-01  2.282e-01   4.034 5.49e-05 ***
DummyPaperlessBilling   1.127e-01  6.985e-02   1.613 0.106741
---
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:  5419.3  on 9966  degrees of freedom
AIC: 5487.3

Number of Fisher Scoring iterations: 7
```

Additionally, the McFadden Pseudo-R^2 was used to help measure the goodness of fit.

Code:

McFadden R^2:

pscl::pR2(logistic_model_all)[“McFadden”]

Output:

```
> pscl::pR2(logistic_model_all)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.5313805
```

The logistic regression model is being built to predict Churn (DummyChurn) based on 33 independent variables. There are a few indicators in both the logistic regression model summary and the McFadden Pseudo-R^2 that show there might be goodness of fit to the model. First, the difference between the null deviance and the residual deviance is substantial. The null deviance is 11564.4 and the residual deviance is substantially lower at 5416.3. This suggests that the logistical model with all independent variables is a better fit for the data and the predictions compared to a model with no predictors. Secondly, the McFadden R^2 is equal to 0.5313805. While it is hard to judge what is considered “good” with the McFadden R^2, the 0.531 does suggest the initial model might be a better fit compared to a null model. Further investigation is necessary with a reduced model.

2. Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.

R provides a function that allows for a stepwise regression:

Code:

Reduce the model backwards:

```
reduced_model <- step(logistic_model_all, direction = 'backward')

summary(reduced_model)
```

The results of this reduced model from 34 variables to 18 variables. Of the 18, DummyChurn (the dependent variable) is included with 17 independent variables: Children, Age, Tenure, MonthlyCharge, Bandwidth_GB_Year, DummyGender, DummyTechie, DummyContract, DummyPort_modem, DummyInternetService, DummyPhone, DummyMultiple, DummyOnlineSecurity, DummyDeviceProtection, DummyStreamingTV, DummyStreamingMovies, and DummyPaperlessBilling.

The following is the summary:

```
> summary(reduced_model)

Call:
glm(formula = DummyChurn ~ Children + Age + Tenure + MonthlyCharge + 
    Bandwidth_GB_Year + DummyGender + DummyTechie + DummyContract + 
    DummyPort_modem + DummyInternetService + DummyPhone + DummyMultiple + 
    DummyOnlineSecurity + DummyDeviceProtection + DummyStreamingTV + 
    DummyStreamingMovies + DummyPaperlessBilling, family = binomial, 
    data = mydata)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.9073304  0.2790125 -17.588  < 2e-16 ***
Children              -0.0555893  0.0178103  -3.121   0.0018 ** 
Age                    0.0088862  0.0018978   4.682 2.84e-06 ***
Tenure                -0.2513729  0.0218603 -11.499  < 2e-16 ***
MonthlyCharge          0.0228811  0.0026537   8.622  < 2e-16 ***
Bandwidth_GB_Year      0.0019253  0.0002627   7.330 2.31e-13 ***
DummyGender            0.1010476  0.0706033   1.431   0.1524    
DummyTechie            0.8192597  0.0893956   9.164  < 2e-16 ***
DummyContract         -2.2777383  0.1020382 -22.322  < 2e-16 ***
DummyPort_modem        0.1502209  0.0686230   2.189   0.0286 *  
DummyInternetService  -0.7080214  0.1373011  -5.157 2.51e-07 ***
DummyPhone            -0.3316537  0.1167373  -2.841   0.0045 ** 
DummyMultiple          0.4423126  0.1021625   4.329 1.49e-05 ***
DummyOnlineSecurity   -0.3154812  0.0738713  -4.271 1.95e-05 ***
DummyDeviceProtection -0.1679686  0.0737938  -2.276   0.0228 *  
DummyStreamingTV       0.9163470  0.1167132   7.851 4.12e-15 ***
DummyStreamingMovies   1.2024474  0.1358343   8.852  < 2e-16 ***
DummyPaperlessBilling  0.1102530  0.0697279   1.581   0.1138    
---
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:  5426.4  on 9982  degrees of freedom
AIC: 5462.4

Number of Fisher Scoring iterations: 7
```

The McFadden Pseudo-R^2 value was taken as well:

```
> pscl::pR2(reduced_model)["McFadden"]
fitting null model for pseudo-r2
 McFadden 
0.5307661 
```

As seen in the above outputs, there are some variables that show significance as their p-values are indicated with the significance codes: ***, **, *, . , and blank. There is also a 53% variance within the model. Additionally, the AIC is smaller with the reduced model (AIC 5462.4) compared to the initial model (AIC 5487.3) which suggests the reduced model does have a better fit. The significance codes to pay attention to are those of ***, **, and * since they represent the p-values of less than

0.05 which shows either very high significance (***), high significance (**), or significant (*). The following 15 variables that express these low p-values are:

Continuous (5 Variables):

- Children, Age, Tenure, MonthlyCharge, and Bandwidth_GB_Year

Categorical (8 Variables):

- DummyTech, DummyContract, DummyPort_modem, DummyDeviceProtection, DummyInternetService, DummyPhone, DummyMultiple, DummyOnlineSecurity, DummyStreamingTV, and DummyStreamingMovies.

From these 13 variables another reduced model was conducted.

Code:

```
# Specifying the variables in the new reduced model
    selected_variables_rm <- c('Children', 'DummyPort_modem',
    'DummyDeviceProtection', 'Age', 'Tenure', 'MonthlyCharge',
    'Bandwidth_GB_Year', 'DummyTechie', 'DummyContract',
                     'DummyInternetService', 'DummyPhone',
    'DummyMultiple', 'DummyOnlineSecurity','DummyStreamingTV',
                     'DummyStreamingMovies')
# Creating the reduced model with specific variables
    reduced_model_2 <- glm(DummyChurn ~ .,
                  data = mydata[, c("DummyChurn", selected_variables_rm)],
                  family = binomial)

    summary(reduced_model_2)
    pscl::pR2(reduced_model_2)["McFadden"]
```

Output:

```
Call:
glm(formula = DummyChurn ~ ., family = binomial, data = mydata[,
    c("DummyChurn", selected_variables_rm)])

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -4.7837625  0.2723296 -17.566  < 2e-16 ***
Children             -0.0577960  0.0177141  -3.263  0.00110 ** 
Age                   0.0092313  0.0018836   4.901 9.54e-07 ***
Tenure               -0.2583929  0.0212744 -12.146  < 2e-16 ***
MonthlyCharge         0.0222948  0.0026201   8.509  < 2e-16 ***
Bandwidth_GB_Year     0.0020116  0.0002553   7.880 3.28e-15 ***
DummyTechie           0.8181014  0.0893410   9.157  < 2e-16 ***
DummyContract        -2.2722793  0.1018958 -22.300  < 2e-16 ***
DummyInternetService -0.6679445  0.1346253  -4.962 6.99e-07 ***
DummyPort_modem       0.1512066  0.0685897   2.205  0.02749 *  
DummyDeviceProtection -0.1635263  0.0737324  -2.218  0.02657 *  
DummyPhone           -0.3359204  0.1166664  -2.879  0.00399 ** 
DummyMultiple         0.4539448  0.1017794   4.460 8.19e-06 ***
DummyOnlineSecurity  -0.3185779  0.0738053  -4.316 1.59e-05 ***
DummyStreamingTV      0.9196173  0.1166018   7.887 3.10e-15 ***
DummyStreamingMovies  1.2144094  0.1355971   8.956  < 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  on 9999  degrees of freedom
Residual deviance:  5431  on 9984  degrees of freedom
AIC: 5463

Number of Fisher Scoring iterations: 7

> pscl::pR2(reduced_model_2)["McFadden"]
fitting null model for pseudo-r2
 McFadden 
0.5303732 
> 
```

With the selected variables within the reduced model the variance was almost the same as the second iteration of the reduced model is 53.03% instead of 53.07% of the first reduced model. Additionally, the AIC of the first reduced model was 5462.4 and the selected variable reduced model very slightly higher at 5463. This suggests the second reduced model conducted has less of a best fit compared to the first reduced model despite it being extremely close.

To take it a step further, analyzing the three models may help:

Create a list of models

model_list <- list(logistic_model_all, reduced_model, reduced_model_2)
model_names <- c('all.mod', 'reduced.mod', ‘reduced.mod2’)

Run aictab to compare models

aictab_result <- aictab(model_list, modnames = model_names)

# Print the result
print(aictab_result)

```
Model selection based on AICc:

              K    AICc Delta_AICc AICcWt Cum.Wt       LL
reduced.mod  18 5462.50       0.00   0.57   0.57 -2713.21
reduced.mod2 16 5463.03       0.53   0.43   1.00 -2715.49
all.mod      34 5487.56      25.07   0.00   1.00 -2709.66

>
```

With the AICcmodavg package, the aictab() function was used to compare the models. The best fit model is always listed first (Bevans, 2023). Between the initial model and the 2 reduced models, the original reduced model is considered the best fit with only a 0.53 AIC discrepancy between the two reduced models.

3. Provide a reduced logistic regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

   The logistic regression model is as follows with 18 variables: DummyChurn = -4.907330 - 0.055589 (Children) + 0.008886 (Age) - 0.251373 (Tenure) + 0.022881 (MonthlyCharge) + 0.001925 (Bandwidth_GB_Year) + 0.101048 (DummyGender) + 0.819260 (DummyTechie) - 2.277738 (DummyContract) + 0.150221 (DummyPort_modem) - 0.708021 (DummyInternetService) - 0.331654 (DummyPhone) + 0.442313 (DummyMultiple) - 0.315481 (DummyOnlineSecurity) - 0.167969 (DummyDeviceProtection) + 0.916347 (DummyStreamingTV) + 1.202447 (DummyStreamingMovies) + 0.110253 (DummyPaperlessBilling)

   Screenshots for each model are pasted above with the reduced model showing the best fit.

E. Analyze the data set using your reduced logistic regression model by doing the following:

1. **Explain your data analysis process by comparing the initial logistic regression model and reduced logistic regression model, including the following element:**

   <u>A model evaluation metric:</u>

   The initial logistic regression model as previously shown and again shown below, consists of all variables selected.

```
> summary(logistic_model_all)

Call:
glm(formula = DummyChurn ~ ., family = binomial, data = mydata)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.876e+00  4.982e-01  -9.788  < 2e-16 ***
Children              -5.036e-02  1.818e-02  -2.770 0.005600 ** 
Age                    8.181e-03  1.944e-03   4.208 2.58e-05 ***
Income                 2.976e-07  1.223e-06   0.243 0.807781    
Outage_sec_perweek     5.539e-04  1.155e-02   0.048 0.961758    
Email                 -1.768e-03  1.138e-02  -0.155 0.876515    
Contacts               2.894e-02  3.467e-02   0.835 0.403863    
Yearly_equip_failure  -3.326e-02  5.429e-02  -0.613 0.540087    
Tenure                -2.354e-01  2.404e-02  -9.791  < 2e-16 ***
MonthlyCharge          2.901e-02  4.747e-03   6.112 9.83e-10 ***
Bandwidth_GB_Year      1.721e-03  2.920e-04   5.893 3.79e-09 ***
Response              -1.759e-02  4.890e-02  -0.360 0.718988    
Fixes                  2.167e-02  4.614e-02   0.470 0.638618    
Replacements          -1.820e-02  4.202e-02  -0.433 0.664854    
Reliability           -2.012e-02  3.731e-02  -0.539 0.589751    
Options               -3.007e-02  3.904e-02  -0.770 0.441138    
Respectful            -3.442e-02  3.998e-02  -0.861 0.389294    
CourtExchange          5.353e-03  3.813e-02   0.140 0.888362    
ActiveListening       -8.250e-03  3.611e-02  -0.228 0.819283    
DummyGender            1.092e-01  7.116e-02   1.535 0.124803    
DummyTechie            8.157e-01  8.946e-02   9.117  < 2e-16 ***
DummyContract         -2.288e+00  1.028e-01 -22.247  < 2e-16 ***
DummyPort_modem        1.536e-01  6.870e-02   2.235 0.025395 *  
DummyTablet           -7.525e-02  7.466e-02  -1.008 0.313482    
DummyInternetService  -9.108e-01  1.884e-01  -4.834 1.34e-06 ***
DummyPhone            -3.291e-01  1.171e-01  -2.811 0.004941 ** 
DummyMultiple          2.553e-01  1.585e-01   1.610 0.107303    
DummyOnlineSecurity   -3.132e-01  7.403e-02  -4.230 2.33e-05 ***
DummyOnlineBackup     -1.576e-01  1.144e-01  -1.378 0.168249    
DummyDeviceProtection -2.319e-01  8.370e-02  -2.770 0.005602 ** 
DummyTechSupport      -1.220e-01  9.265e-02  -1.317 0.187772    
DummyStreamingTV       6.961e-01  1.850e-01   3.762 0.000169 ***
DummyStreamingMovies   9.203e-01  2.282e-01   4.034 5.49e-05 ***
DummyPaperlessBilling  1.127e-01  6.985e-02   1.613 0.106741    
---
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:  5419.3  on 9966  degrees of freedom
AIC: 5487.3

Number of Fisher Scoring iterations: 7
```

A backwards step regression was conducted which is a stepwise regression that takes a fully saturated model as seen above and gradually eliminates variables from the regression model to find the reduced model that best explains the data (Analyst Soft, 2024). In other words, it reduces the model to the best-fit model. This is also known as the backward elimination regression. The coding and results of this backwards elimination is as follows:

```
#Reducing the model backwards
reduced_model <- step(logistic_model_all, direction = 'backward')
```

Results:

```
> summary(reduced_model)

Call:
glm(formula = DummyChurn ~ Children + Age + Tenure + MonthlyCharge +
    Bandwidth_GB_Year + DummyGender + DummyTechie + DummyContract +
    DummyPort_modem + DummyInternetService + DummyPhone + DummyMultiple +
    DummyOnlineSecurity + DummyDeviceProtection + DummyStreamingTV +
    DummyStreamingMovies + DummyPaperlessBilling, family = binomial,
    data = mydata)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.9073304  0.2790125 -17.588  < 2e-16 ***
Children             -0.0555893  0.0178103  -3.121   0.0018 **
Age                   0.0088862  0.0018978   4.682 2.84e-06 ***
Tenure               -0.2513729  0.0218603 -11.499  < 2e-16 ***
MonthlyCharge         0.0228811  0.0026537   8.622  < 2e-16 ***
Bandwidth_GB_Year     0.0019253  0.0002627   7.330 2.31e-13 ***
DummyGender           0.1010476  0.0706033   1.431   0.1524
DummyTechie           0.8192597  0.0893956   9.164  < 2e-16 ***
DummyContract        -2.2777383  0.1020382 -22.322  < 2e-16 ***
DummyPort_modem       0.1502209  0.0686230   2.189   0.0286 *
DummyInternetService -0.7080214  0.1373011  -5.157 2.51e-07 ***
DummyPhone           -0.3316537  0.1167373  -2.841   0.0045 **
DummyMultiple         0.4423126  0.1021625   4.329 1.49e-05 ***
DummyOnlineSecurity  -0.3154812  0.0738713  -4.271 1.95e-05 ***
DummyDeviceProtection -0.1679686  0.0737938  -2.276   0.0228 *
DummyStreamingTV      0.9163470  0.1167132   7.851 4.12e-15 ***
DummyStreamingMovies  1.2024474  0.1358343   8.852  < 2e-16 ***
DummyPaperlessBilling 0.1102530  0.0697279   1.581   0.1138
---
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:  5426.4  on 9982  degrees of freedom
AIC: 5462.4

Number of Fisher Scoring iterations: 7
```

To ensure the reduced model is the best fit compared to the initial model, model metrics can be evaluated in comparison to the initial model. The model metrics being used for evaluation between the initial model and the reduced model(s) are the AIC and the McFadden R^2. The AIC is a statistical method that helps to evaluate how well a regression model fits the data. Comparing AIC values between both initial and reduced model(s) can help determine which model is the best fit for the data. A low AIC indicates a better fit while a high AIC value indicates a lesser fit model (Bevans, 2023). The McFadden R^2 is a statistical measurement that shows how well the data fits the regression and it also reveals the variability (percentage) of the target variable is explained by the regression model. While having a high R^2 is ideal, other factors may present a better fit model such as the AIC (Taylor, 2024).

The AIC and the R^2 values for the initial model are as follows:

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564.4  on 9999  degrees of freedom
Residual deviance:  5419.3  on 9966  degrees of freedom
AIC: 5487.3

Number of Fisher Scoring iterations: 7
```

```
> pscl::pR2(logistic_model_all)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.5313805
```

The initial model presents an AIC of 5487.3 and an R^2 value of 0.5313 or 53.13%. After the backwards elimination was conducted, the reduced model's AIC and R^2 values are as follows:

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11564.4  on 9999  degrees of freedom
Residual deviance:  5426.4  on 9982  degrees of freedom
AIC: 5462.4
```

```
> pscl::pR2(reduced_model)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.5307661
```

As seen above, the reduced model has a lower AIC in comparison to the initial model. This indicates the reduced model is the best fit and the variables within the model are optimal to make predictions on the remaining coefficients. The reduced model in comparison to the initial model has the following variables removed during the backwards elimination regression: Income, Outages_sec_perweek, Email, Contacts, Yearly_equip_failure, Response, Fixes, Replacements, Reliability, Options, Respectful, CourtExchange, ActiveListening, DummyTablet, DummyOnlineBackup, and DummyTechSupport. All variables listed that were a part of the saturated initial model that were removed during the backwards elimination showed no significant as per their significance code (p-value = 1). The only two variables from the initial model that showed no significance remained in the reduced model are DummyGender and DummyPaperlessBilling.

Out of curiosity, running three more additional reduced models may show a best fit model based on their AIC values. The three models will be the following: 1) Include DummyGender and Exclude DummyPaperlessBilling, 2) Exclude DummyGender and Include DummyPaperlessBilling, and 3) last is to remove both DummyGender and DummyPaperlessBilling.

1) For the first additional reduced model which is to include DummyGender but exclude DummyPaperlessBilling:

```
> summary(reduced_model_2)

Call:
glm(formula = DummyChurn ~ ., family = binomial, data = mydata[, 
    c("DummyChurn", selected_variables_rm_1)])

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.8422452  0.2755333 -17.574  < 2e-16 ***
Children              -0.0552912  0.0177988  -3.106  0.00189 ** 
Age                    0.0089088  0.0018970   4.696 2.65e-06 ***
Tenure                -0.2511229  0.0218553 -11.490  < 2e-16 ***
MonthlyCharge          0.0228793  0.0026525   8.625  < 2e-16 ***
Bandwidth_GB_Year      0.0019227  0.0002626   7.321 2.46e-13 ***
DummyGender            0.1008122  0.0705900   1.428  0.15325    
DummyTechie            0.8207615  0.0893740   9.183  < 2e-16 ***
DummyContract         -2.2751832  0.1019867 -22.309  < 2e-16 ***
DummyInternetService  -0.7063385  0.1372533  -5.146 2.66e-07 ***
DummyPort_modem        0.1502378  0.0686060   2.190  0.02853 *  
DummyDeviceProtection -0.1653763  0.0737500  -2.242  0.02494 *  
DummyPhone            -0.3345752  0.1165815  -2.870  0.00411 ** 
DummyMultiple          0.4424685  0.1021088   4.333 1.47e-05 ***
DummyOnlineSecurity   -0.3149397  0.0738657  -4.264 2.01e-05 ***
DummyStreamingTV       0.9140295  0.1166525   7.835 4.67e-15 ***
DummyStreamingMovies   1.2036155  0.1357827   8.864  < 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:  5428.9  on 9983  degrees of freedom
AIC: 5462.9

Number of Fisher Scoring iterations: 7

> pscl::pR2(reduced_model_2)["McFadden"]
fitting null model for pseudo-r2
 McFadden 
0.5305496 
```

Reminder:
Initial AIC: 5487.3
Backwards Reduced AIC: 5462.4

Removing Paperless Billing and including Gender, the AIC of this reduced model is 5462.9. The backwards reduced model still shows a better and lower AIC value of 5462.4. Therefore, this model is not the best fit.

2) For the second additional reduced model which is to include DummyPaperlessBilling but exclude Dummy Gender:

```
Call:
glm(formula = DummyChurn ~ ., family = binomial, data = mydata[,
    c("DummyChurn", selected_variables_rm_2)])

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.8484786  0.2758118 -17.579  < 2e-16 ***
Children              -0.0580966  0.0177256  -3.278  0.00105 **
Age                    0.0092089  0.0018844   4.887 1.02e-06 ***
Tenure                -0.2586519  0.0212807 -12.154  < 2e-16 ***
MonthlyCharge          0.0222953  0.0026212   8.506  < 2e-16 ***
Bandwidth_GB_Year      0.0020144  0.0002554   7.888 3.06e-15 ***
DummyTechie            0.8165963  0.0893630   9.138  < 2e-16 ***
DummyContract         -2.2748035  0.1019457 -22.314  < 2e-16 ***
DummyInternetService  -0.6695373  0.1346725  -4.972 6.64e-07 ***
DummyPort_modem        0.1512475  0.0686066   2.205  0.02748 *
DummyDeviceProtection -0.1661226  0.0737771  -2.252  0.02434 *
DummyPhone            -0.3330330  0.1168260  -2.851  0.00436 **
DummyMultiple          0.4538142  0.1018336   4.456 8.33e-06 ***
DummyOnlineSecurity   -0.3191285  0.0738111  -4.324 1.54e-05 ***
DummyStreamingTV       0.9219174  0.1166624   7.902 2.73e-15 ***
DummyStreamingMovies   1.2132864  0.1356473   8.944  < 2e-16 ***
DummyPaperlessBilling  0.1100322  0.0697104   1.578  0.11447
---
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:  5428.5  on 9983  degrees of freedom
AIC: 5462.5

Number of Fisher Scoring iterations: 7
```

```
> pscl::pR2(reduced_model_3)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.5305889
```

Reminder:
Initial AIC: 5487.3
Backwards Reduced AIC: 5462.4

Including Paperless Billing but exclusing Gender, the AIC of this reduced model is 5462.5. The backwards reduced model is still marginally better with a lower AIC value of 5462.4. Therefore, this model is not the best fit.

3) For the last additional reduced model which it to exclude both the DummyGender and the DummyPaperlessBilling:

```
Call:
glm(formula = DummyChurn ~ ., family = binomial, data = mydata[,
    c("DummyChurn", selected_variables_rm_3)])

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.7837625  0.2723296 -17.566  < 2e-16 ***
Children              -0.0577960  0.0177141  -3.263  0.00110 **
Age                    0.0092313  0.0018836   4.901 9.54e-07 ***
Tenure                -0.2583929  0.0212744 -12.146  < 2e-16 ***
MonthlyCharge          0.0222948  0.0026201   8.509  < 2e-16 ***
Bandwidth_GB_Year      0.0020116  0.0002553   7.880 3.28e-15 ***
DummyTechie            0.8181014  0.0893410   9.157  < 2e-16 ***
DummyContract         -2.2722793  0.1018958 -22.300  < 2e-16 ***
DummyInternetService  -0.6679445  0.1346253  -4.962 6.99e-07 ***
DummyPort_modem        0.1512066  0.0685897   2.205  0.02749 *
DummyDeviceProtection -0.1635263  0.0737324  -2.218  0.02657 *
DummyPhone            -0.3359204  0.1166664  -2.879  0.00399 **
DummyMultiple          0.4539448  0.1017794   4.460 8.19e-06 ***
DummyOnlineSecurity   -0.3185779  0.0738053  -4.316 1.59e-05 ***
DummyStreamingTV       0.9196173  0.1166018   7.887 3.10e-15 ***
DummyStreamingMovies   1.2144094  0.1355971   8.956  < 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  on 9999  degrees of freedom
Residual deviance:  5431  on 9984  degrees of freedom
AIC: 5463

Number of Fisher Scoring iterations: 7
```

```
> pscl::pR2(reduced_model_4)["McFadden"]
fitting null model for pseudo-r2
 McFadden
0.5303732
```

Reminder:
Initial AIC: 5487.3
Backwards Reduced AIC: 5462.4

Removing both Paperless Billing and Gender, the AIC of this reduced model is 5463 which is the highest AIC value of all the reduced models. The backwards reduced model still shows a better and lower AIC value of 5462.4. Therefore, this model is not the best fit.

Ultimately, the first backwards elimination reduced model is considered the best fit model for this analysis and the coefficients will explain how each variable will help predict the churn rates of customers which is in section F.

All McFadden R^2 values for the reduced showed marginal differences and did not affect the outcome of which reduced model was best fit.

2. Provide the output and *all* calculations of the analysis you performed, including the following elements for your reduced logistic regression model:

<u>Confusion Matrix</u>

A confusion matrix was done to provide a comprehensive view of a model's performance.

Code:

Confusion Matrix
Predicted probabilities of Churn

predicted_probabilities <- predict(reduced_model, newdata = mydata, type = "response")

Convert probabilities to class labels (0 or 1) based on a threshold (e.g., 0.5)
This needs to be done since Churn is in class 0,1

predicted_labels <- ifelse(predicted_probabilities > 0.5, 1, 0)

Creating the actual labels for Churn

actual_labels <- mydata$DummyChurn

Both actual and predicted labels into the confusion matrix:

conf_matrix <- table(actual_labels, predicted_labels)
print(conf_matrix)

```
> print(conf_matrix)
             predicted_labels
actual_labels    0    1
            0 6844  506
            1  724 1926
```

<u>Accuracy Calculation for Optimal Reduced Model:</u>

Code:

Getting Calculations of Accuracy from matrix
Explicitly using the table function from the caret package (was having problems doing this so had to call on it specifically)

cm_data <- as.matrix(caret::confusionMatrix(conf_matrix)$table)

Calculate metrics

accuracy <- sum(diag(cm_data)) / sum(cm_data) * 100
precision <- cm_data[2, 2] / sum(cm_data[, 2])

```
recall <- cm_data[2, 2] / sum(cm_data[2, ])
specificity <- cm_data[1, 1] / sum(cm_data[1, ])
```

```
# Print the metrics

cat("Accuracy:", accuracy, "% \n")
cat("Precision: ", precision, "\n")
cat("Recall: ", recall, "\n")
cat("Specificity: ", specificity, "\n")
```

```
> cat("Accuracy:", accuracy, "% \n")
Accuracy: 87.7 %
> cat("Precision: ", precision, "\n")
Precision:  0.7919408
> cat("Recall: ", recall, "\n")
Recall:  0.7267925
> cat("Specificity: ", specificity, "\n")
Specificity:  0.9311565
>
```

The accuracy of the model is 87.7%.

3. Provide an executable error-free copy of the code used to support the implementation of the logistic regression models using a Python or R file.

Code will be provided in an R Source file and a .txt file attached to this assessment: task_2_code_R.txt. In addition, coding has been provided above.

Part V: Data Summary and Implications

F. Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following elements:

<u>A regression equation for the optimal reduced model:</u>

The logistic regression model is as follows with 18 variables: DummyChurn = -4.907330 - 0.055589 (Children) + 0.008886 (Age) - 0.251373 (Tenure) + 0.022881 (MonthlyCharge) + 0.001925 (Bandwidth_GB_Year) + 0.101048 (DummyGender) + 0.819260 (DummyTechie) - 2.277738 (DummyContract) + 0.150221 (DummyPort_modem) - 0.708021 (DummyInternetService) - 0.331654 (DummyPhone) + 0.442313 (DummyMultiple) - 0.315481 (DummyOnlineSecurity) - 0.167969 (DummyDeviceProtection) + 0.916347 (DummyStreamingTV) + 1.202447 (DummyStreamingMovies) + 0.110253 (DummyPaperlessBilling)

```
> summary(reduced_model)

Call:
glm(formula = DummyChurn ~ Children + Age + Tenure + MonthlyCharge +
    Bandwidth_GB_Year + DummyGender + DummyTechie + DummyContract +
    DummyPort_modem + DummyInternetService + DummyPhone + DummyMultiple +
    DummyOnlineSecurity + DummyDeviceProtection + DummyStreamingTV +
    DummyStreamingMovies + DummyPaperlessBilling, family = binomial,
    data = mydata)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.9073304  0.2790125 -17.588  < 2e-16 ***
Children             -0.0555893  0.0178103  -3.121   0.0018 **
Age                   0.0088862  0.0018978   4.682 2.84e-06 ***
Tenure               -0.2513729  0.0218603 -11.499  < 2e-16 ***
MonthlyCharge         0.0228811  0.0026537   8.622  < 2e-16 ***
Bandwidth_GB_Year     0.0019253  0.0002627   7.330 2.31e-13 ***
DummyGender           0.1010476  0.0706033   1.431   0.1524
DummyTechie           0.8192597  0.0893956   9.164  < 2e-16 ***
DummyContract        -2.2777383  0.1020382 -22.322  < 2e-16 ***
DummyPort_modem       0.1502209  0.0686230   2.189   0.0286 *
DummyInternetService -0.7080214  0.1373011  -5.157 2.51e-07 ***
DummyPhone           -0.3316537  0.1167373  -2.841   0.0045 **
DummyMultiple         0.4423126  0.1021625   4.329 1.49e-05 ***
DummyOnlineSecurity  -0.3154812  0.0738713  -4.271 1.95e-05 ***
DummyDeviceProtection -0.1679686  0.0737938  -2.276   0.0228 *
DummyStreamingTV      0.9163470  0.1167132   7.851 4.12e-15 ***
DummyStreamingMovies  1.2024474  0.1358343   8.852  < 2e-16 ***
DummyPaperlessBilling  0.1102530  0.0697279   1.581   0.1138
---
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:  5426.4  on 9982  degrees of freedom
AIC: 5462.4

Number of Fisher Scoring iterations: 7
```

<u>The statistical and practical significance of the reduced model:</u>

An interpretation of the coefficients of the reduced model is necessary to find the statistical and practical significance of the reduced model.


```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.9073304  0.2790125 -17.588  < 2e-16 ***
Children             -0.0555893  0.0178103  -3.121   0.0018 **
Age                   0.0088862  0.0018978   4.682 2.84e-06 ***
Tenure               -0.2513729  0.0218603 -11.499  < 2e-16 ***
MonthlyCharge         0.0228811  0.0026537   8.622  < 2e-16 ***
Bandwidth_GB_Year     0.0019253  0.0002627   7.330 2.31e-13 ***
DummyGender           0.1010476  0.0706033   1.431   0.1524
DummyTechie           0.8192597  0.0893956   9.164  < 2e-16 ***
DummyContract        -2.2777383  0.1020382 -22.322  < 2e-16 ***
DummyPort_modem       0.1502209  0.0686230   2.189   0.0286 *
DummyInternetService -0.7080214  0.1373011  -5.157 2.51e-07 ***
DummyPhone           -0.3316537  0.1167373  -2.841   0.0045 **
DummyMultiple         0.4423126  0.1021625   4.329 1.49e-05 ***
DummyOnlineSecurity  -0.3154812  0.0738713  -4.271 1.95e-05 ***
DummyDeviceProtection -0.1679686  0.0737938  -2.276   0.0228 *
DummyStreamingTV      0.9163470  0.1167132   7.851 4.12e-15 ***
DummyStreamingMovies  1.2024474  0.1358343   8.852  < 2e-16 ***
DummyPaperlessBilling  0.1102530  0.0697279   1.581   0.1138
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Of the 17 independent coefficients, 12 are binary (categorical) dummy variables.

Each coefficient represents the log-odds of the outcome variable which is Churn. For numerical variables, every one-unit change (increase or decrease) of a coefficient while holding all other variables constant, the log-odds of churning are either increased or decreased. For every categorical variable, for every 1 (yes) or 0 (no) of a coefficient while holding all other variables constant, the log-odds of churning is either increased or decreased.

Numerical Coefficients

In the above equation it can be stated for all coefficients that are numerical, positive, and express a significance of $p < 0.05$ (Children, Age, Tenure, MonthlyCharge, Bandwidth_GB_Year) can increase the log-odds of a customer churning. For example, for every one-unit increase of Age (0.0088862) while holding all other variables constant increases the log-odds of churning by 0.009. Or, to calculate the percentage change in odd we can calculate using the following equation:

$$\%\, Change\ in\ Odds = (\exp(coeffcient) - 1) * 100$$

For Age:

$$\%\, Change\ in\ Odds = (\exp(0.0088862) - 1) * 100 = 0.89\%$$

In other words, for every one-unit increase in age, the odds of a customer churning are increased by approximately 0.89%.

Conversely, for coefficients that are numerical, negative, and express a significance of $p < 0.05$ (Tenure) can decrease the log-odds of a customer churning. For every one-unit increase of Tenure (-0.2513729) while holding all other variables constant decreases the log-odds of customer churning by 0.26.

For Tenure:

$$\%\, Change\ in\ Odds = (\exp(-0.2513729) - 1) * 100 = -22.2\%$$

In other words, for every one-unit increase in Tenure, the odds of a customer churning are decreased by approximately 22.8%.

Categorical Coefficients

Categorical coefficients follow the same interpretation as numerical except instead of an increase in one-unit, the increase or decrease is based on if the variable is present or not. For example, DummyTechie would indicate that the customer is either a techie or they are not. In this case, DummyTechie (0.8192597) is a positive coefficient and expresses a significant p-value. This indicates, if a customer is considered a Techie while holding all other variables constant, the log-odds of the customer churning is 0.82.

For Techie:

$$\%\, Change\ in\ Odds = (\exp(0.8192597) - 1) * 100 = 127\%$$

In other words, being a techie while holding all other variables constant, compared to not being a techie, increases the odds of the customer churning by approximately 127%.

The following is for all numerical, positive/negative, and express a significant p-value:

1. Children (Decrease)

   For every one-unit increase of Children (-0.0555893) while holding all other variables constant decreases the log-odds of churning by 0.26.

   $$\% \, Change \; in \; Odds = (\exp(-0.0555893) - 1) * 100 = -5.4\%$$

   For every one-unit increase in Children, the odds of a customer churning are decreased by approximately 5.4%.

2. Age (Increase)

   For every one-unit increase of Age (0.0088862) while holding all other variables constant increases the log-odds of churning by 0.009.

   $$\% \, Change \; in \; Odds = (\exp(0.0088862) - 1) * 100 = 0.89\%$$

   For every one-unit increase in Age, the odds of a customer churning are increased by approximately 0.89%.

3. **Tenure (Decrease)**

   For every one-unit increase of Tenure (-0.2513729) while holding all other variables constant decreases the log-odds of churning by 0.26.

   $$\% \, Change \; in \; Odds = (\exp(-0.2513729) - 1) * 100 = -22.2\%$$

   For every one-unit increase in Tenure, the odds of a customer churning are decreased by approximately 22.2%.

4. **MonthlyCharge (Increase)**

   For every one-unit increase of Monthly Charge (0.0228811) while holding all other variables constant increases the log-odds of churning by 0.023.

   $$\% \, Change \; in \; Odds = (\exp(0.0228811) - 1) * 100 = 2.31\%$$

   For every one-unit increase in Monthly Charge, the odds of a customer churning are increased by approximately 2.31%.

5. Bandwith_GB_Year (Increase)

   For every one-unit increase of Bandwidth_GB_Year (0.0019253) while holding all other variables constant increases the log-odds of churning by 0.002.

   $$\%\ Change\ in\ Odds = (\exp(0.0019253) - 1) * 100 = 0.19\%$$

   For every one-unit increase in Bandwidth_GB_Year, the odds of a customer churning are increased by approximately 0.19%.

The following is for all categorical, positive/negative, and express a significant p-value:

1. DummyTechie (Increase**)**

   If a customer is considered a Techie (0.8192597) while holding all other variables constant, the log-odds of the customer churning is 0.82. 0.8192597

   $$\%\ Change\ in\ Odds = (\exp(0.8192597) - 1) * 100 = 127\%$$

   If a customer is considered a Techie while holding all other variables constant, compared to not being a Techie, increases the odds of the customer churning by approximately 127%.

2. **DummyContract (Decrease)**

   If a customer has a Contract (-2.2777383) while holding all other variables constant, the log-odds of the customer churning is 2.3.

   $$\%\ Change\ in\ Odds = (\exp(-2.2777383) - 1) * 100 = -89.7\%$$

   If a customer has a contract while holding all other variables constant, compared to not having a contract, decreases the odds of the customer churning by approximately 89.7%.

3. DummyPort_Modem (Increase)

   If a customer has a Port Modem (0.1502209) while holding all other variables constant, the log-odds of the customer churning is 0.15.

   $$\%\ Change\ in\ Odds = (\exp(0.1502209) - 1) * 100 = 16.2\%$$

   If a customer has a contract while holding all other variables constant, compared to not having a contract, increases the odds of the customer churning by approximately 16.2%.

4. DummyInternetService (Decrease)

   If a customer has Internet Services (-0.7080214) while holding all other variables constant, the log-odds of the customer churning is 0.71.

$$\%\ Change\ in\ Odds = (\exp(-0.7080214) - 1) * 100 = -50.7\%$$

   If a customer has Internet Services while holding all other variables constant, compared to not having Internet Services, decreases the odds of the customer churning by approximately 50.7%.

5. DummyPhone (Decrease)

   If a customer has a Phone service (-0.3316537) while holding all other variables constant, the log-odds of the customer churning is 0.33.

$$\%\ Change\ in\ Odds = (\exp(-0.3316537) - 1) * 100 = -28.2\%$$

   If a customer has a Phone service while holding all other variables constant, compared to not having a Phone service, decreases the odds of the customer churning by approximately 28.2%.

6. DummyMultiple (Increase)

   If a customer has Internet Services (0.4423126) while holding all other variables constant, the log-odds of the customer churning is 0.44.

$$\%\ Change\ in\ Odds = (\exp(0.4423126) - 1) * 100 = 55.6\%$$

   If a customer has Multiple services while holding all other variables constant, compared to not having Multiple services, increases the odds of the customer churning by approximately 55.6%.

7. DummyOnlineSecurity (Decrease)

   If a customer has Online Security (-0.3154812) while holding all other variables constant, the log-odds of the customer churning is 0.32.

$$\%\ Change\ in\ Odds = (\exp(-0.3154812) - 1) * 100 = -27.1\%$$

   If a customer has Online Security while holding all other variables constant, compared to not having Online Security, decreases the odds of the customer churning by approximately 27.1%.

8. DummyDeviceProtection (Decrease)

   If a customer has Device Protection (-0.1679686) while holding all other variables constant, the log-odds of the customer churning is 0.17.

   $$\%\ Change\ in\ Odds = (\exp(-0.1679686) - 1) * 100 = -15.5\%$$

   If a customer has Device Protection while holding all other variables constant, compared to not having Device Protection, decreases the odds of the customer churning by approximately 15.5%.

9. DummyStreamingTV (Increase)

   If a customer has a Streaming TV (0.9163470) while holding all other variables constant, the log-odds of the customer churning is 0.82.

   $$\%\ Change\ in\ Odds = (\exp(0.9163470) - 1) * 100 = 127\%$$

   If a customer has a Streaming TV while holding all other variables constant, compared to not having Streaming TV, increases the odds of the customer churning by approximately 127%.

10. **DummyStreamingMovies (Increase)**

    If a customer has Streaming Movies (1.2024474) while holding all other variables constant, the log-odds of the customer churning is 1.2.

    $$\%\ Change\ in\ Odds = (\exp(1.2024474) - 1) * 100 = 233\%$$

    If a customer has Streaming Movies while holding all other variables constant, compared to not having Streaming Movies, increases the odds of the customer churning by approximately 233%.

DummyGender and DummyPaperlessBilling were not used for the odd-logs and the % Change in Odds because they did not show significance in their p-values in the reduced model.

After examining the coefficients and both their log-odds and % change in odds, the two numerical variables that stand out are Tenure and Monthly Charge. It seems the longer a customer stays with the company the less likely they churn, or for every one-unit increase in Tenure, while all other variables within the model stay constant, the customer is less likely to churn by 22.2%. Conversely, an increase of every one-unit in their Monthly Charge increases the likelihood of a customer churning by 2.31% if all other variables remain constant.

Concerning coefficients that are categorical, the variables that show the largest increase and decrease in churn rates are Streaming Movies and Contracts. If a customer has

Streaming Movies compared to customers that do not while holding all other variables constant have increase odds of churning by 233%. Conversely, customers that have contracts compared to customers that do not have contracts while holding all other variables constant have decreased odd of churning by 89.7%

The limitations of the data analysis:

There are some limitations that can be considered for a logistic regression analysis. For this analysis, the data was not in binary format and some variables had to be converted using logic such as Internet Services. In addition, based on a previous analysis (task 1) it was discovered that multicollinearity was likely present in the data. With multicollinearity, variables having high correlation may make it difficult to assess the effects of each predictor in the analysis, or the coefficients examined previously may show high sensitivity to changes in the model such as reducing the model (Bhandari, 2024). These are only a few examples of limitations and there are likely many more, but these are a few that can impact the analysis.

2. **Recommend a course of action based on your results.**

The recommended course of action based on the results of this analysis is for the telecommunications company to consider their services to their customers. Churn will always happen within a company as customers will move on to other products or business that may better suit their needs. After examining the models and checking the log-odds of the coefficients, it is recommended that the company investigate their Streaming Movie services and their Monthly Charges to their customers. Both variables seem to show higher churn rates with their customers compared to other variables. Also, the second highest churn rate is present with both Techies and Streaming TV. Their % change in odds is 127% for both variables. This is rather high and can signify that techies might be more likely to investigate other companies to suit their needs. Additionally, the company's streaming services for both TV and Movies may not be adequate for customers to maintain the subscriptions with the company.

Part VI: Demonstration

Video Link: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=562966cf-6036-4706-b219-b1c50032c9bd

Works Cited

Analyst Soft. (2024, March). *Analyst Soft StatPlus.* Retrieved 2024, from Backward Stepwise Regression: https://www.analystsoft.com/en/products/statplus/content/help/pdf/analysis_regression_backward_stepwise_elimination_regression_model.pdf

Bevans, R. (2023, June 22). *Akaike Information Criterion | When & How to Use It (Example)*. Retrieved from Scribbr: https://www.scribbr.com/statistics/akaike-information-criterion/

Bhandari, A. (2024, 13 June). *Analytics Vidhya*. Retrieved from What is Multicollinearity: Causes and Detection using VIF: https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/

Frankenfield, J. (2022, May 18). *Churn Rate: What It Means, Examples, and Calculations*. Retrieved from Investopia: https://www.investopedia.com/terms/c/churnrate.asp

Simplilearn. (2023, January 12). *Logistic Regression in R: The Ultimate Tutorial with Examples*. Retrieved from Simplilearn: https://www.simplilearn.com/tutorials/data-science-tutorial/logistic-regression-in-r#:~:text=Logistic%20regression%20is%20used%20to,This%20is%20called%20logistic%20regression.

Solumaths. (2024). *Solumaths.* Retrieved from Exponential Calculator: https://www.solumaths.com/en/calculator/calculate/exp/1.2024474#google_vignette

Statistic Solutions. (2023). *Assumptions of Logistic Regression*. Retrieved from Complete Dissertation by Statistics Solutions: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

Statistics Solutions. (2023). *Assumptions of Multiple Linear Regression*. Retrieved from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/

Taylor, S. (2024). *Corporate Finance Institute*. Retrieved from R-Squared: https://corporatefinanceinstitute.com/resources/data-science/r-squared/#:~:text=R%2DSquared%20(R%C2%B2%20or%20the,explained%20by%20the%20independent%20variable.

Voxco. (2023). *From Basics to Brilliance: Insights into Logistic Regression Assumptions*. Retrieved from Voxco: https://www.voxco.com/blog/logistic-regression-assumptions/#:~:text=Logistic%20regression%20assumes%20the%20observations,an%20order%20for%20the%20observations.