Tucker Atwood
WGU MSDA
D208 Task 2
3/31/24

A1. Research Question

Using data from the "churn" dataset, I will use multiple logistic regression to answer the question, "What factors affect customer churn?"

A2. Analysis Objectives

The goal of this analysis is to understand the relationship between various factors and whether or not a customer stays with the service. These relationships may then be used to inform the areas of focus that will ensure customers stay with the service more often.

B1. Regression Assumptions

In order to perform this multiple logistic regression and make validated conclusions based on the results, the following assumptions must be made regarding the relevant data:

- **Linearity**: each independent variable has a linear relationship with the logarithmic odds of the dependent variable.
- **Multicollinearity**: independent variables are not highly correlated with each other.
- **Independence**: all observations were collected independently of each other.
- **Sample size**: the dependent variable has at least 10 observations for each outcome.

B2. Programming Language

This research question will be answered using the programming language R in the RStudio environment. This language was chosen for its accessibility and efficiency in descriptive analysis such as univariate and bivariate visualizations and summary statistics such as means, medians, and quartiles. It is also ideal for predictive statistical modeling processes like multiple logistic regression. Within R, the following packages will be utilized: naniar, plyr, dplyr, ggplot2, fastDummies, car, and yardstick. These packages were chosen to provide several functions that will make the predictive modeling process more effective.

B3. Regression Justification

Multiple logistic regression will be used because the response variable, churn, is categorical, and the research question aims to find explanatory variables that can be used to predict it.

C1. Data Cleaning

To ensure the results of the multiple logistic regression will be accurate as a predictive modeling tool, the dataset will be checked for the following data quality issues, which will be cleaned if necessary:

- **Full duplicates**, defined as observations for which every variable is a match with another observation, will be detected and removed as needed.
- **Partial duplicates**, defined as observations for which a subset of variables match another observation, will be detected by searching for matches on the "Customer_id" variable, and removed as needed.
- **Missing values**, defined as entries of "NA" in any variable, will be investigated further and either imputed or removed as needed.
- **Outliers**, defined as values significantly higher or lower than other established values within the variable, will be investigated further and either retained, imputed, or removed as needed.

The results of this detection and treatment were as follows:

- **Zero full duplicates** were found.
- **Zero partial duplicates** were found.
- **Zero missing values** were found.
- **Outliers**:
    - An outlier check was conducted for all quantitative variables: Latitude, Longitude, Population, Children, Age, Income, Outage_sec_perweek, Email, Contacts, Tenure, Yearly_equip_failure, MonthlyCharge, Bandwidth_GB_Year, and Items 1-8.
    - **Latitude** and **Longitude**:
        - A scatterplot was determined to be the best visualization of potential outliers, as these factors are inherently linked and refer to physical locations easiest to view together.
        - In the scatterplot, clear groups of data were found on each end of Latitude and the low end of Longitude. Filtering by the apparent cutoffs for these groups revealed 75 low Latitude values, 77 high Latitude values, and 112 low Longitude values.
        - All low longitude values were found to be from Alaska (77 values) or Hawaii (35 values). All low latitude values were found to be from Hawaii (35 values) or Puerto Rico (40 values). All high latitude values were found to be from Alaska (77 values). These values were determined to be reasonable based on the geographic location of these regions. All outliers were retained.
    - **Items 1-8**:
        - For items 1-8, values were presented on a scale from 1 to 8, so any values outside that range were treated as outliers. Zero outliers were found.

- For **all other quantitative variables**, histograms provided a first look at the shape of the data, including whether or not a Normal distribution shape was apparent (which would inform the following outlier detection method) and whether there were values far away from all others.
- Boxplots were then used to further explore the shape and spread of the data, and were particularly useful for clearly identifying outliers using the IQR method.
- If the data appeared to follow a Normal distribution, outliers were determined by normalizing the data and filtering for z-scores less than -3 or greater than 3.
- Otherwise, outliers were determined using the Interquartile Range (IQR) method: by multiplying the IQR by 1.5, subtracting this value from Q1 (first quartile) and adding it to Q3 (third quartile), and finding values below or above these boundaries.
- **Population**:
  - 937 outliers were found, ranging from 31,816 to 111,850. This was determined to be reasonable and acceptable. All outliers were retained.
- **Children**:
  - 401 outliers were found, ranging from 8 to 10. This was determined to be reasonable and acceptable. All outliers were retained.
- **Age**:
  - Zero outliers were found.
- **Income**:
  - 336 outliers were found, ranging from 104,363 to 258,901. This was determined to be reasonable and acceptable. All outliers were retained.
- **Outage_sec_perweek**:
  - 76 outliers were found, ranging from 0.1-21.2. This was determined to be reasonable and acceptable. All outliers were retained.
- **Email**:
  - 12 outliers were found, ranging from 1 to 23. This was determined to be reasonable and acceptable. All outliers were retained.
- **Contacts**:
  - 8 outliers were found, ranging from 6 to 7. This was determined to be reasonable and acceptable. All outliers were retained.
- **Yearly_equip_failure**:
  - 94 outliers were found, ranging from 3 to 6. This was determined to be reasonable and acceptable. All outliers were retained.
- **Tenure**:
  - Zero outliers were found.
- **MonthlyCharge**:
  - Zero outliers were found.
- **Bandwidth_GB_Year**:
  - Zero outliers were found.

The following code executes the detection and treatment of data quality issues as described. An executable version of this code can be found in the attached file: Atwood_D208_Task2_Code.R.

```
sum(duplicated(churn)) # checks for full duplicates (0)

library(plyr) # using plyr package
library(dplyr) # using dplyr package

churn %>%
  count(Customer_id) %>%
  filter(n > 1) # checks for partial duplicates with matching Customer_id (0)

library(naniar) # using naniar package
n_miss(churn) # total missing values (0)

library(ggplot2) # using ggplot2 package
churn %>%
  ggplot(aes(x=Lat, y=Lng)) + geom_point()
# scatterplot of Latitude and Longitude values; outliers present on each end of Latitude, low end
of Longitude

churn %>%
  filter(Lng < -125) %>%
  count(State)
# 112 low Longitude values, 77 from Alaska, 35 from Hawaii

churn %>%
  filter(Lat < 24) %>%
  count(State)
# 75 low Latitude values, 35 from Hawaii, 40 from Puerto Rico

churn %>%
  filter(Lat > 50) %>%
  count(State)
# 77 high Latitude values, all from Alaska

item_outliers <- churn %>%
  filter(Item1 < 1 | Item1 > 8
         | Item2 < 1 | Item2 > 8
```

```
        | Item3 < 1 | Item3 > 8
        | Item4 < 1 | Item4 > 8
        | Item5 < 1 | Item5 > 8
        | Item6 < 1 | Item6 > 8
        | Item7 < 1 | Item7 > 8
        | Item8 < 1 | Item8 > 8)
# items 1-8 listed as scale from 1 to 8; finds values outside range
count(item_outliers) # confirms no values outside range

hist(churn$Population) # visualization of Population data; skewed right
boxplot(churn$Population) # many outliers present
pop_outliers <- churn %>%
  filter(Population < quantile(churn$Population, 0.25) - IQR(churn$Population) * 1.5
        | Population > (quantile(churn$Population, 0.75) + IQR(churn$Population) * 1.5))
# find outliers using IQR method
count(pop_outliers) # 937 outliers
summary(pop_outliers$Population) # outlier range is 31,816-111,850

hist(churn$Children) # visualization of Children data; skewed right
boxplot(churn$Children) # 3 outlier values appear
chi_outliers <- churn %>%
  filter((Children < quantile(churn$Children, 0.25, na.rm = TRUE) - IQR(churn$Children, na.rm
= TRUE) * 1.5)
        | (Children > quantile(churn$Children, 0.75, na.rm = TRUE) + IQR(churn$Children, na.rm
= TRUE) * 1.5))
# find outliers using IQR method
count(chi_outliers) # 401 outliers
summary(chi_outliers$Children) # outlier range is 8-10

hist(churn$Age) # visualization of Age data; relatively uniform
boxplot(churn$Age) # no outliers present
age_outliers <- churn %>%
  filter((Age < quantile(churn$Age, 0.25, na.rm = TRUE) - IQR(churn$Age, na.rm = TRUE) *
1.5)
        | (Age > quantile(churn$Age, 0.75, na.rm = TRUE) + IQR(churn$Age, na.rm = TRUE) *
1.5))
# find outliers using IQR method
count(age_outliers) # confirms zero Age outliers

hist(churn$Income) # visualization of Income data; skewed right
```

```
boxplot(churn$Income) # many outliers present
inc_outliers <- churn %>%
  filter((Income < quantile(churn$Income, 0.25, na.rm = TRUE) - IQR(churn$Income, na.rm =
TRUE) * 1.5)
        | (Income > quantile(churn$Income, 0.75, na.rm = TRUE) + IQR(churn$Income, na.rm =
TRUE) * 1.5))
# find outliers using IQR method
count(inc_outliers) # 336 outliers
summary(inc_outliers$Income) # outlier range is 104,363-258,901

hist(churn$Outage_sec_perweek) # visualization of Outage_sec_perweek data; skewed right
boxplot(churn$Outage_sec_perweek) # many outliers present
outage_outliers <- churn %>%
  filter((Outage_sec_perweek < quantile(churn$Outage_sec_perweek, 0.25) -
IQR(churn$Outage_sec_perweek) * 1.5)
        | (Outage_sec_perweek > quantile(churn$Outage_sec_perweek, 0.75) +
IQR(churn$Outage_sec_perweek) * 1.5))
# find outliers using IQR method
count(outage_outliers) # 76 outliers
summary(outage_outliers$Outage_sec_perweek) # outlier range is 0.1-21.2

hist(churn$Email) # visualization of Email data; normal distribution
boxplot(churn$Email) # 6 outlier values appear
email_outliers <- churn %>%
  mutate(email_z = scale(churn$Email)) %>%
  filter(email_z > 3 | email_z < -3)  # find outliers using z-score method
count(email_outliers) # 12 outliers
summary(email_outliers$Email) # outlier range is 1-23

hist(churn$Contacts) # visualization of Contacts data; skewed right
boxplot(churn$Contacts) # 2 outlier values appear
con_outliers <- churn %>%
  filter((Contacts < quantile(churn$Contacts, 0.25) - IQR(churn$Contacts) * 1.5)
        | (Contacts > quantile(churn$Contacts, 0.75) + IQR(churn$Contacts) * 1.5))
# find outliers using IQR method
count(con_outliers) # 8 outliers
summary(con_outliers$Contacts) # outlier range is 6-7

hist(churn$Yearly_equip_failure) # visualization of Yearly_equip_failure data; skewed right
boxplot(churn$Yearly_equip_failure) # 3 outlier values appear
```

```
yef_outliers <- churn %>%
  filter((Yearly_equip_failure < quantile(churn$Yearly_equip_failure, 0.25) -
IQR(churn$Yearly_equip_failure) * 1.5)
        | (Yearly_equip_failure > quantile(churn$Yearly_equip_failure, 0.75) +
IQR(churn$Yearly_equip_failure) * 1.5))
# find outliers using IQR method
count(yef_outliers) # 94 outliers
summary(yef_outliers$Yearly_equip_failure) # outlier range is 3-6

hist(churn$Tenure) # visualization of Tenure data; bimodal
boxplot(churn$Tenure) # no outliers present
ten_outliers <- churn %>%
  filter((Tenure < quantile(churn$Tenure, 0.25, na.rm = TRUE) - IQR(churn$Tenure, na.rm =
TRUE) * 1.5)
        | (Tenure > quantile(churn$Tenure, 0.75, na.rm = TRUE) + IQR(churn$Tenure, na.rm =
TRUE) * 1.5))
# find outliers using IQR method
count(ten_outliers) # confirms zero Tenure outliers

hist(churn$MonthlyCharge) # visualization of MonthlyCharge data; normal distribution
boxplot(churn$MonthlyCharge) # 5 outlier values appear
mon_outliers <- churn %>%
  mutate(mon_z = scale(churn$MonthlyCharge)) %>%
  filter(mon_z > 3 | mon_z < -3)  # find outliers using z-score method
count(mon_outliers) # confirms zero MonthlyCharge outliers

hist(churn$Bandwidth_GB_Year) # visualization of Bandwidth_GB_Year data; bimodal
boxplot(churn$Bandwidth_GB_Year) # no outliers present
bgy_outliers <- churn %>%
  filter((Bandwidth_GB_Year < quantile(churn$Bandwidth_GB_Year, 0.25, na.rm = TRUE) -
IQR(churn$Bandwidth_GB_Year, na.rm = TRUE) * 1.5)
        | (Bandwidth_GB_Year > quantile(churn$Bandwidth_GB_Year, 0.75, na.rm = TRUE) +
IQR(churn$Bandwidth_GB_Year, na.rm = TRUE) * 1.5))
# find outliers using IQR method
count(bgy_outliers) # confirms zero Bandwidth_GB_Year outliers
```

C2. Data Exploration

To answer the research question with multiple logistic regression, an exploration of the dependent variable and all relevant independent variables was conducted. The following understandings are

important in communicating the results of the data exploration, which will include an analysis of summary statistics and table summaries:

- A **minimum** value is the smallest observation in a dataset.
- A **maximum** value is the largest observation in a dataset.
- The minimum and maximum are often referred to as the **extrema** (plural of **extremum**) of the dataset.
- The **mean** value of a dataset is calculated by adding all values together and dividing by the number of observations. This is also referred to as the **average**.
- The **median** value of a dataset represents the middle value; if all values were set in order from smallest to largest, the value with an equal number of observations lower than and higher than it would be the median. If the median is significantly closer to the minimum or the maximum, this could indicate an uneven bunching of values between the median and its nearest extremum.
- The mean and median are both called **measures of central tendency** and are often close in value, especially if the dataset has a distribution that is **uniform** (values spread out relatively evenly), **bimodal** (two large clumps of values on either side of the middle of a dataset, such that the data is not considered skewed), or **Normal** (values follow a bell curve: large cluster in the middle, symmetric decreases on each end).
- If the mean is greater than the median, this is often an indication that the data is **skewed right**: a large cluster of values exists on the lower end of the set, with a longer tail extending to the right than the left. Similarly, if the mean is less than the median, this is often an indication that the data is **skewed left**: a large cluster of values on the higher end of the set, with a longer tail extending to the left than the right.
- The **first quartile** is the median of the lower half of values in a dataset. When a median is determined, the observations can be thought of as being split in half, with 50% of values below the median and 50% above. The same process to find the median is repeated with only the first 50% of values to calculate the first quartile. Thus, the first quartile is greater than 25% of all observations in the dataset, and less than 75% of all observations.
- The **third quartile** is the median of the upper half of values, and is calculated the same way as the first quartile, except with the latter 50% of values. Thus, the third quartile is greater than 75% of all observations in the dataset, and less than 25% of all observations.
- As with the note on the median above, if the first or third quartile is significantly closer to the minimum, median, or maximum, this could indicate an uneven bunching of values.
- These concepts are essential in understanding the data exploration of the response variable, Tenure, and all factors that have been determined to be relevant in answering the research question as predictor variables: Population, Children, Age, Income, Outage_sec_perweek, Email, Yearly_equip_failure, MonthlyCharge, Bandwidth_GB_Year, Churn, Contract, DeviceProtection, TechSupport, and InternetService.
- **Tenure** represents the number of months a customer has continued with the service.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	7.918	35.431	34.526	61.480	71.999

- The minimum value, 1.000, represents a customer who has been with the service for 1 month.
- The maximum value, 71.999, represents a customer who has been with the service for almost 6 years (72 months).
- The mean value, 34.526, signifies that the average customer has continued with the service for about 34.5 months (almost 3 years).
- The median value, 35.431, signifies that the middle value of all Tenure values is about 35.4 months.
- The mean and median are relatively close, indicating the data likely follows a uniform, bimodal, or Normal distribution.
- The first quartile, 7.918, signifies that the middle value of the lower half of the data is about 7.9 months. This is noticeably closer to the minimum than the median, potentially indicating a larger grouping of values between the minimum and the first quartile than between the first quartile and the median.
- The third quartile, 61.48, signifies that the middle value of the upper half of the data is about 61.5 months. This is noticeably closer to the maximum than the median, potentially indicating a larger grouping of values between the third quartile and the maximum than between the median and the third quartile.

- **Population** represents the census population for customers' area of residence.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	738	2910	9757	13168	111850

  - The minimum represents a customer whose residence has a population of 0.
  - The maximum represents a customer whose residence has a population of 111,850.
  - The mean signifies that the average population for all customers is about 10,000.
  - The median signifies that the middle population for all customers is about 3,000.
  - The median is significantly lower than the mean, which indicates the distribution may be skewed right.
  - The first quartile signifies that the middle value of the lower half of the data is about 750. This is noticeably closer to the minimum than the median, potentially indicating a larger grouping of values between the minimum and the first quartile than between the first quartile and the median.
  - The third quartile signifies that the middle value of the upper half of the data is about 13,000. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.
- **Children** represents the number of children each customer has.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	1.000	2.088	3.000	10.000

  - The minimum indicates that the least number of children a customer has is 0.
  - The maximum indicates that the greatest number of children a customer has is 10.
  - The mean signifies that the average number of children per customer is about 2.
  - The median signifies that the middle number of children a customer has is 1.

- The median is slightly lower than the mean, which indicates the distribution may be skewed right.
- The first quartile signifies that the middle value of the lower half of the data is 0. This is the same value as the minimum, which indicates that at least 25% of customers in the data have 0 children.
- The third quartile signifies that the middle value of the upper half of the data is 3. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.

- **Age** represents the customer's age in years.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   35.00   53.00   53.08   71.00   89.00
```

  - The minimum indicates that the youngest customer is 18 years old.
  - The maximum indicates that the oldest customer is 89 years old.
  - The mean signifies that the average age for all customers is about 53.
  - The median signifies that the middle age for all customers is 53.
  - The mean and median are relatively close, indicating the data likely follows a uniform, bimodal, or Normal distribution.
  - The first quartile signifies that the middle value of the lower half of the data is about 35. This is about the same distance from the minimum as from the median, which suggests that the first 50% of the data is spread relatively evenly.
  - The third quartile signifies that the middle value of the upper half of the data is about 71. This is exactly the same distance from the median as from the maximum, which suggests that the last 50% of the data is spread relatively evenly.
- **Income** represents the customer's annual income in US dollars.

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  348.7  19224.7  33170.6  39806.9  53246.2 258900.7
```

  - The minimum represents a customer whose annual income is about $350.
  - The maximum represents a customer whose annual income is about $260,000.
  - The mean signifies that the average income for all customers is about $40,000.
  - The median signifies that the middle income for all customers is about $33,000.
  - The median is significantly lower than the mean, which indicates the distribution may be skewed right.
  - The first quartile signifies that the middle value of the lower half of the data is about $19,000. This is slightly closer to the median than the minimum, potentially indicating a larger grouping of values between the first quartile and the median than between the minimum and the first quartile.
  - The third quartile signifies that the middle value of the upper half of the data is about $53,000. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.

- **Outage_sec_perweek** represents the average seconds of system outages per week in a customer's area of residence.

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  0.09975   8.01821 10.01856 10.00185 11.96949 21.20723
```

  - The minimum represents a customer whose area experiences an average of about 0.1 seconds of system outages per week.
  - The maximum represents a customer whose area experiences an average of about 21.2 seconds of system outages per week.
  - The mean signifies that the average seconds of system outages in the areas of all residences per week is about 10.
  - The median signifies that the middle seconds of system outages in the areas of all residences per week is about 10.
  - The mean and median are relatively close, indicating the data likely follows a uniform, bimodal, or Normal distribution.
  - The first quartile signifies that the middle value of the lower half of the data is about 8. This is noticeably closer to the median than the minimum, potentially indicating a larger grouping of values between the first quartile and the median than between the minimum and the first quartile.
  - The third quartile signifies that the middle value of the upper half of the data is about 12. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.
- **Email** represents the number of emails that were sent to a customer in the past year.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   10.00   12.00   12.02   14.00   23.00
```

  - The minimum indicates the least emails a customer received in the past year was 1.
  - The maximum indicates the greatest number of emails a customer received in the past year was 23.
  - The mean signifies that the average number of emails sent to all customers was about 12.
  - The median signifies that the middle number of emails sent to all customers was 12.
  - The mean and median are relatively close, indicating the data likely follows a uniform, bimodal, or Normal distribution.
  - The first quartile signifies that the middle value of the lower half of the data is 10. This is noticeably closer to the median than the minimum, potentially indicating a larger grouping of values between the first quartile and the median than between the minimum and the first quartile.
  - The third quartile signifies that the middle value of the upper half of the data is 14. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.

- **Yearly_equip_failure** represents the number of times a customer's equipment failed and/or needed to be replaced in the past year.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   0.000   0.000   0.398   1.000   6.000
```

  - The minimum indicates the least equipment failures in the past year was 0.
  - The maximum indicates the greatest equipment failures in the past year was 6.
  - The mean signifies that the average number of equipment failures in the past year for all customers was about 0.4.
  - The median signifies that the middle number of equipment failures in the past year for all customers was 0.
  - The median is slightly lower than the mean, which indicates the distribution may be skewed right.
  - The first quartile signifies that the middle value of the lower half of the data is 0. This is the same value as the minimum and the median, which indicates that at least 50% of customers did not have any equipment failures in the past year.
  - The third quartile signifies that the middle value of the upper half of the data is 1. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.
- **MonthlyCharge** represents the average amount a customer has been charged per month.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
79.98  139.98  167.48  172.62  200.73  290.16
```

  - The minimum represents a customer who was charged about $80 per month.
  - The maximum represents a customer who was charged about $290 per month.
  - The mean signifies that the average amount a customer was charged per month was about $173.
  - The median signifies that the middle amount a customer was charged per month was about $167.
  - The median is close to the mean but slightly lower, indicating the data may follow a uniform, bimodal, or Normal distribution, but with a slight skew right.
  - The first quartile signifies that the middle value of the lower half of the data is about $140. This is noticeably closer to the median than the minimum, potentially indicating a larger grouping of values between the first quartile and the median than between the minimum and the first quartile.
  - The third quartile signifies that the middle value of the upper half of the data is about $200. This is noticeably closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.
- **Bandwidth_GB_Year** represents the average amount of data a customer uses per year in gigabytes. If a customer has been with the service for less than a year, it is approximated.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
155.5  1236.5  3279.5  3392.3  5586.1  7159.0
```

- The minimum represents a customer who uses about 155 gigabytes of data per year.
- The maximum represents a customer who uses about 7160 gigabytes of data per year.
- The mean signifies that the average amount of data used by customers in a year is about 3390.
- The median signifies that the middle amount of data used by customers in a year is about 3280.
- The mean and median are relatively close, indicating the data likely follows a uniform, bimodal, or Normal distribution.
- The first quartile signifies that the middle value of the lower half of the data is about 1240. This is noticeably closer to the minimum than the median, potentially indicating a larger grouping of values between the minimum and the first quartile than between the first quartile and the median.
- The third quartile signifies that the middle value of the upper half of the data is about 5590. This is slightly closer to the median than the maximum, potentially indicating a larger grouping of values between the median and the third quartile than between the third quartile and the maximum.

- The remaining variables relevant to answering the research question are categorical, so it is not possible to calculate summary statistics like those above for these variables. Instead, a table of values will be presented for each.
- **Churn** represents whether or not a customer stopped using the service in the past month.

No	Yes
7350	2650

  - The "No" column indicates that 7350 customers continued using the service.
  - The "Yes" column indicates that 2650 customers stopped using the service.
- **Contract** represents the contract length for a customer's service plan.

Month-to-month	One year	Two Year
5456	2102	2442

  - The "Month-to-month" column indicates that 5456 customers are on a plan that can be continued or discontinued on a monthly basis.
  - The "One year" column indicates that 2102 customers are on a plan that can be continued or discontinued on a yearly basis.
  - The "Two Year" column indicates that 2442 customers are on a plan that can be continued or discontinued on a bi-yearly basis.
- **DeviceProtection** represents whether or not a customer has signed up for a device protection add-on to their service.

No	Yes
5614	4386

  - The "No" column indicates that 5614 customers do not have a device protection add-on to their service.
  - The "Yes" column indicates that 4386 customers do have a device protection add-on to their service.

- **TechSupport** represents whether or not a customer has signed up for a technical support add-on to their service.

```
  No  Yes
6250 3750
```

  - The "No" column indicates that 6250 customers do not have a technical support add-on to their service.
  - The "Yes" column indicates that 3750 customers do have a technical support add-on to their service.
- **InternetService** represents the customer's internet service provider, or indicates that a customer does not have an internet service provider.

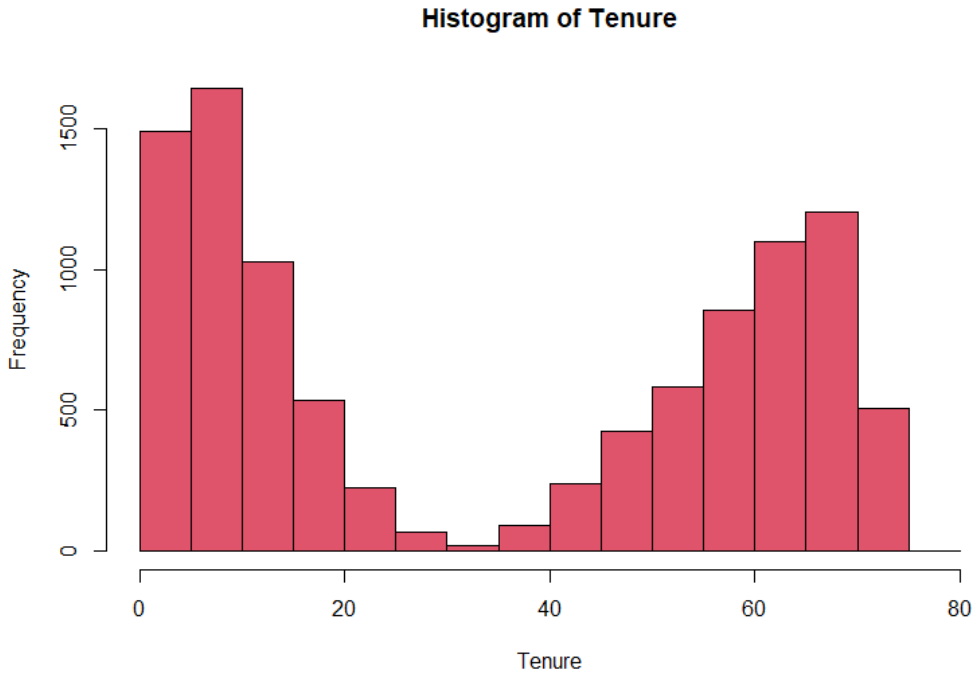
```
 DSL Fiber Optic        None
3463        4408        2129
```

  - The "DSL" column indicates that 3463 customers have DSL as their internet service provider.
  - The "Fiber Optic" column indicates that 4408 customers have Fiber Optic as their internet service provider.
  - The "None" column indicates that 2129 customers do not have an internet service provider.

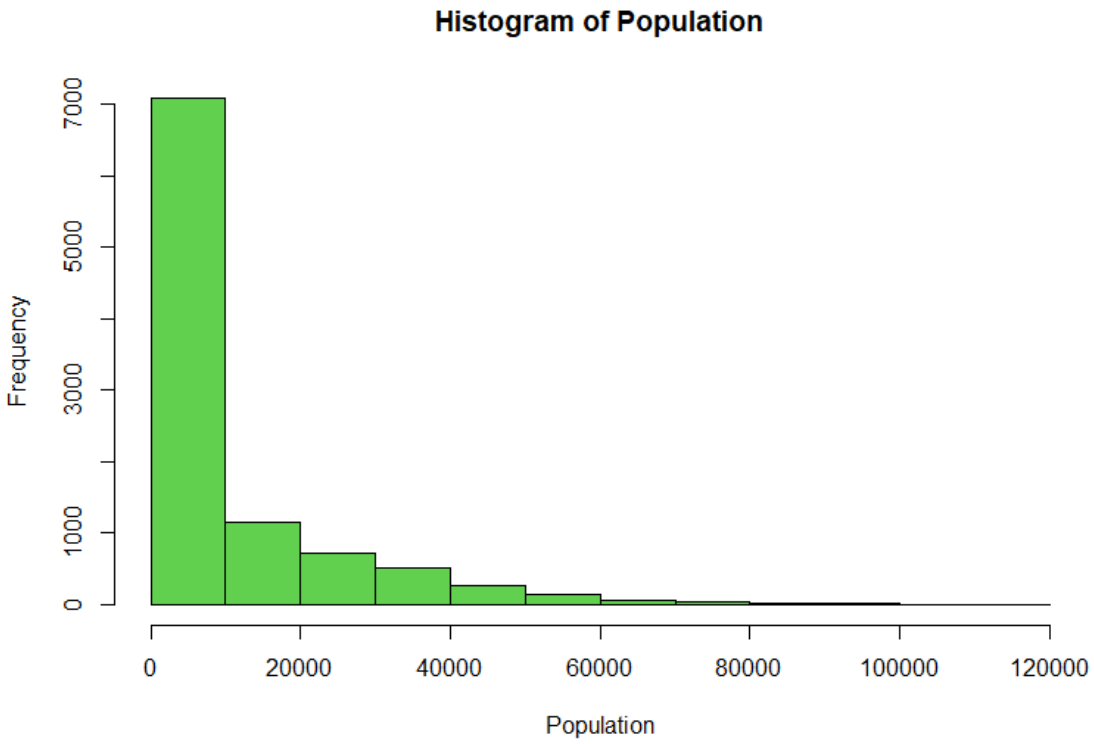
C3. Visualizations

The following graphs provide further insight into the **univariate distributions** of each individual variable referred to in the previous section; i.e., those relevant to answering the research question.

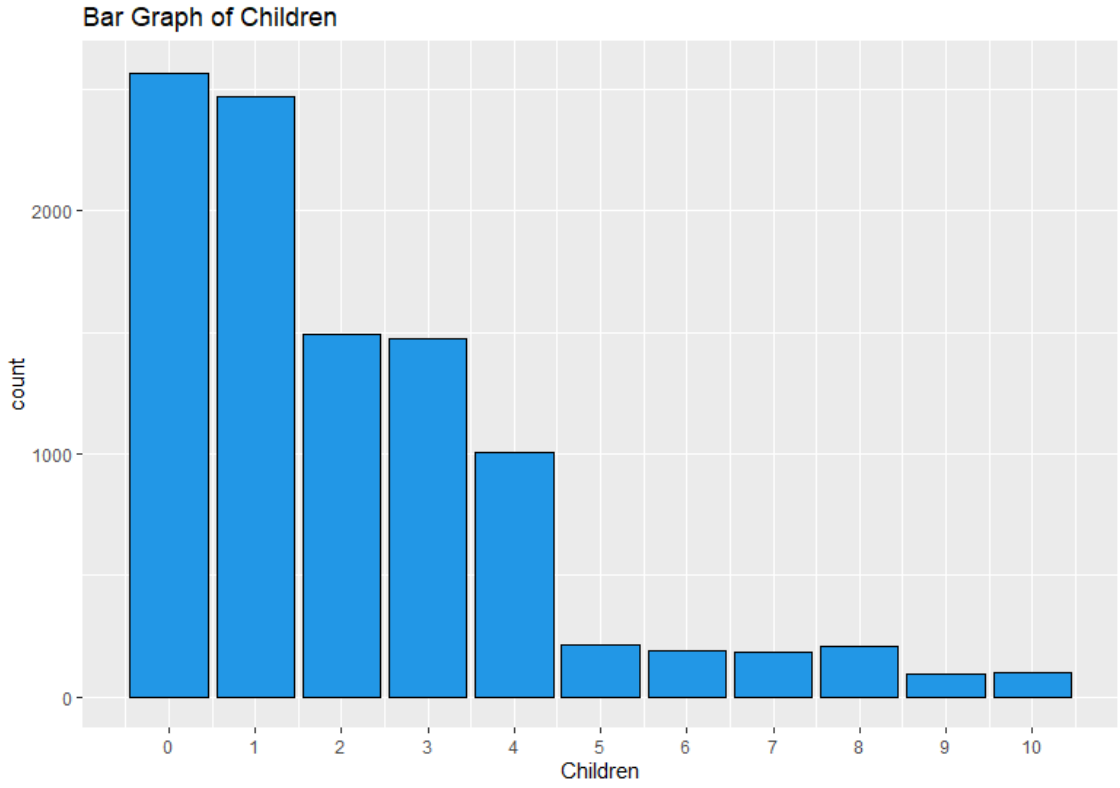
- **Tenure:**



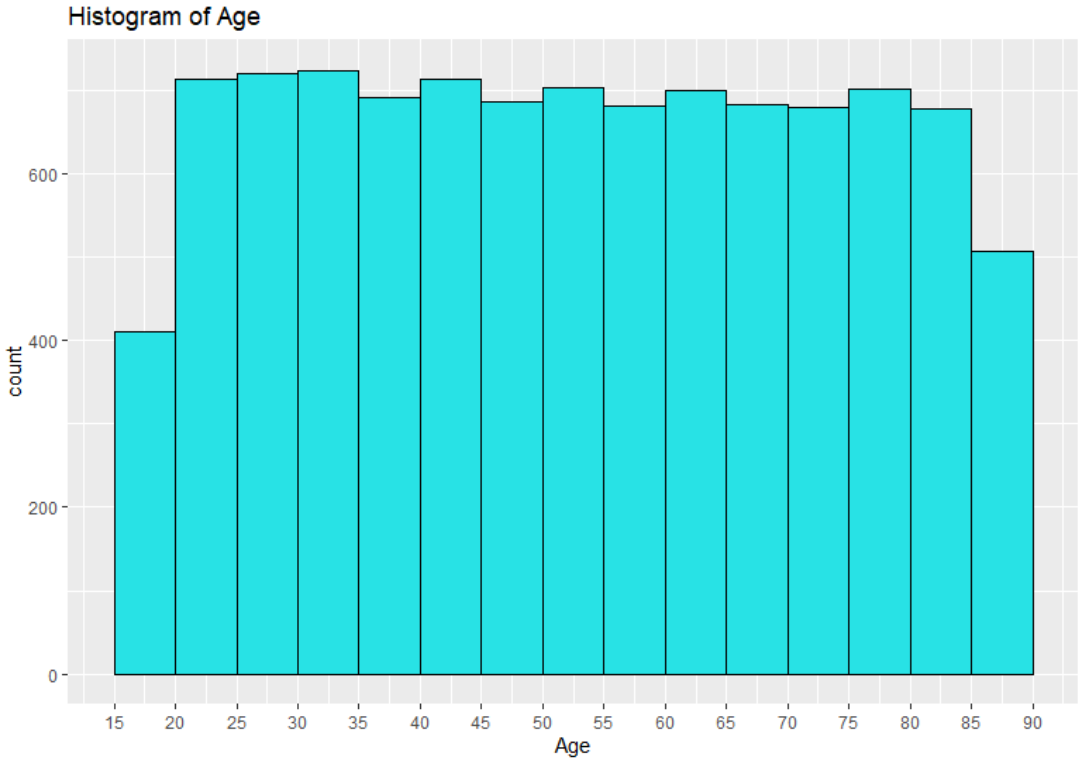
- **Population:**



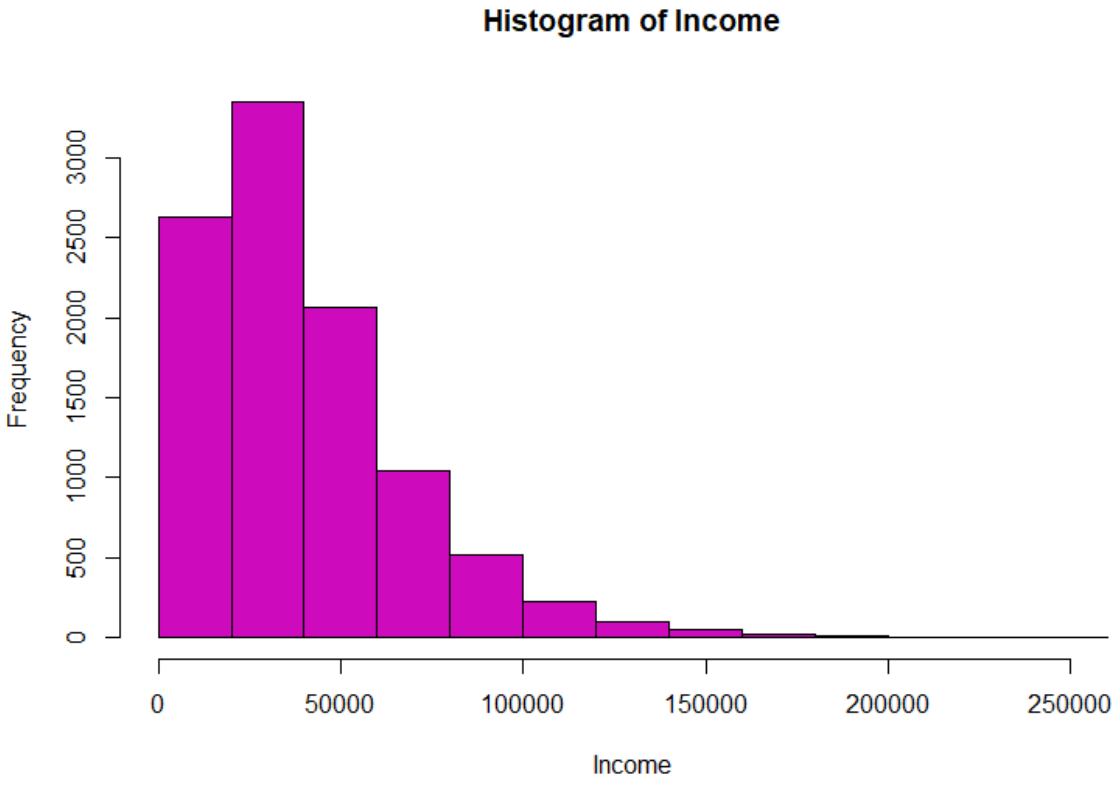
- **Children:**



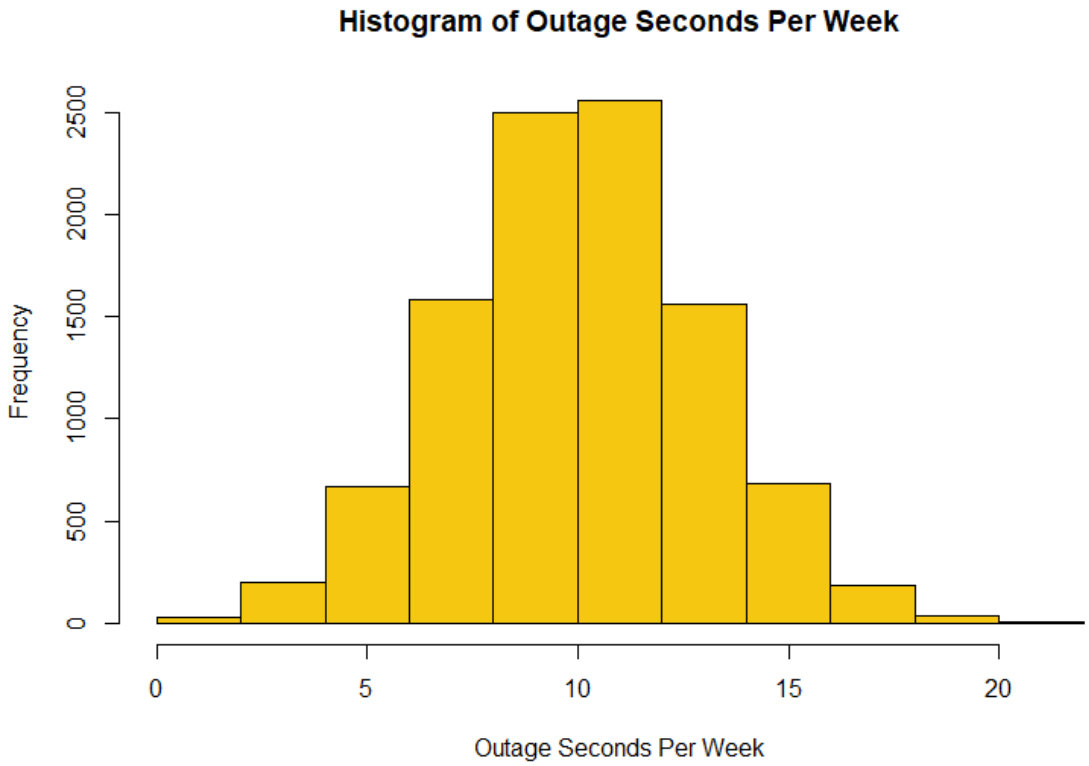
- **Age:**


Histogram of Age
count
0
200
400
600
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
Age

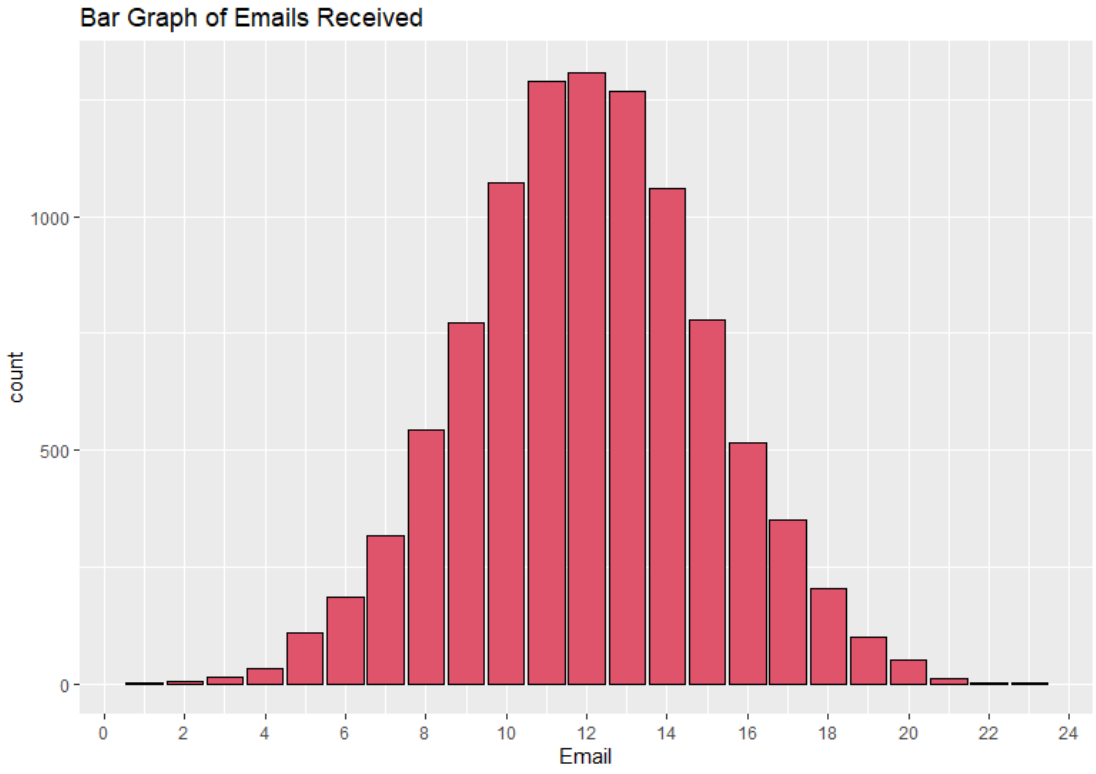
- **Income:**


Histogram of Income
Frequency
0
500
1000
1500
2000
2500
3000
0
50000
100000
150000
200000
250000
Income

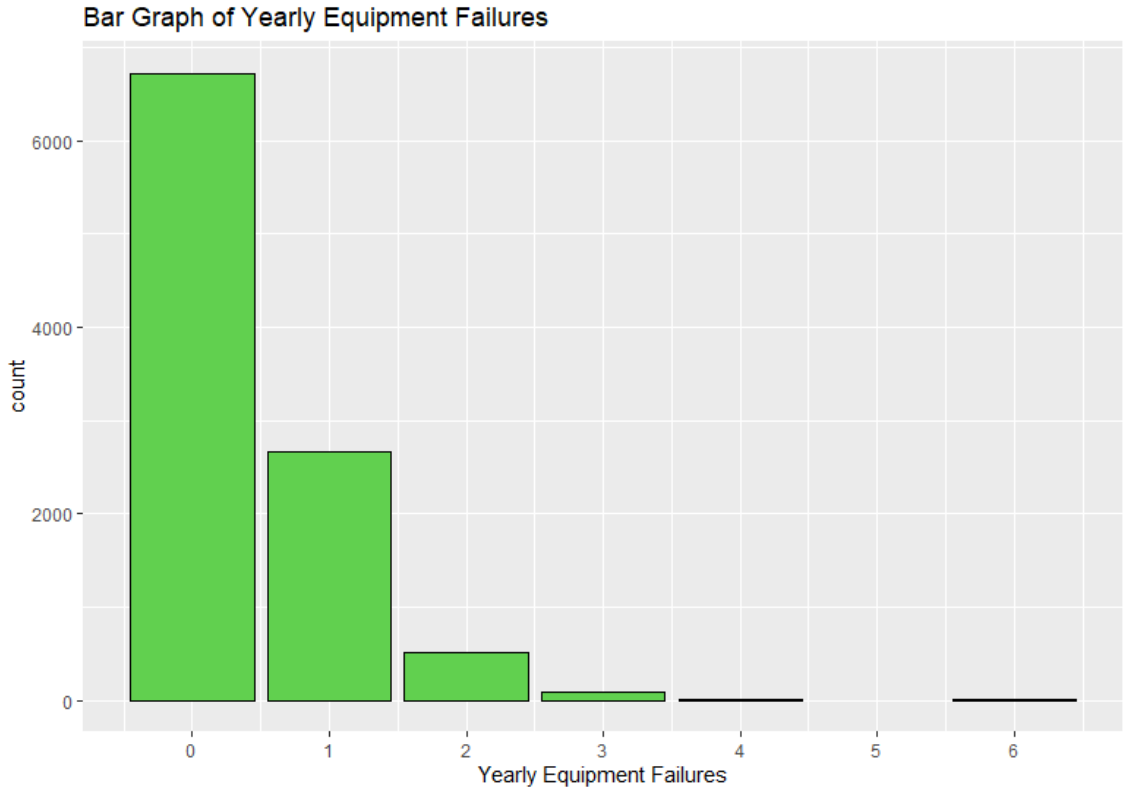
- **Outage_sec_perweek:**



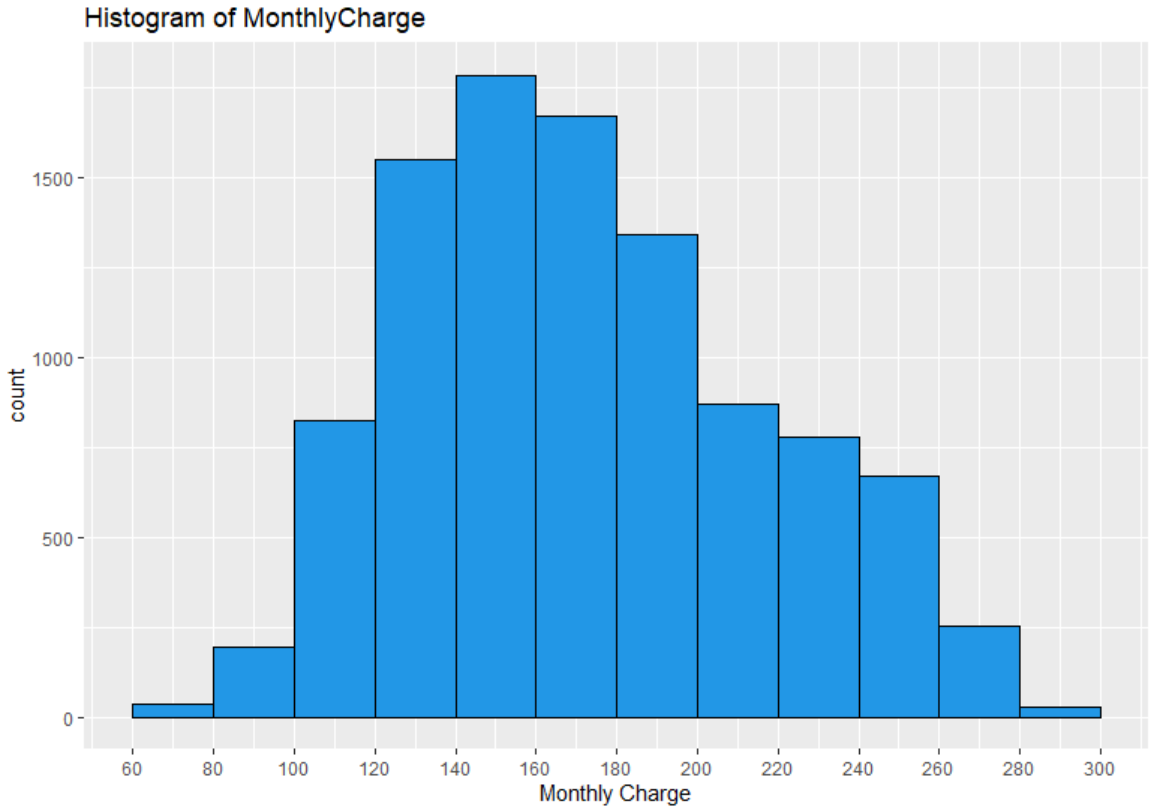
- **Email:**



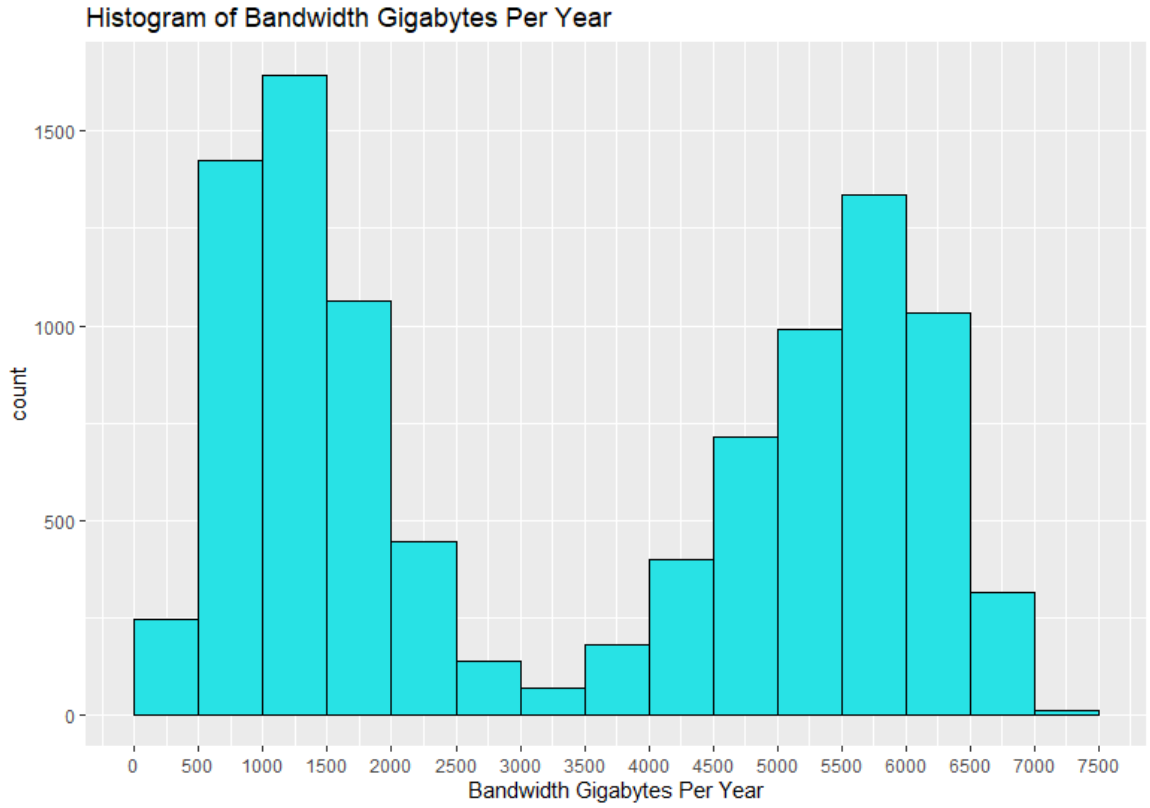
- **Yearly_equip_failure:**



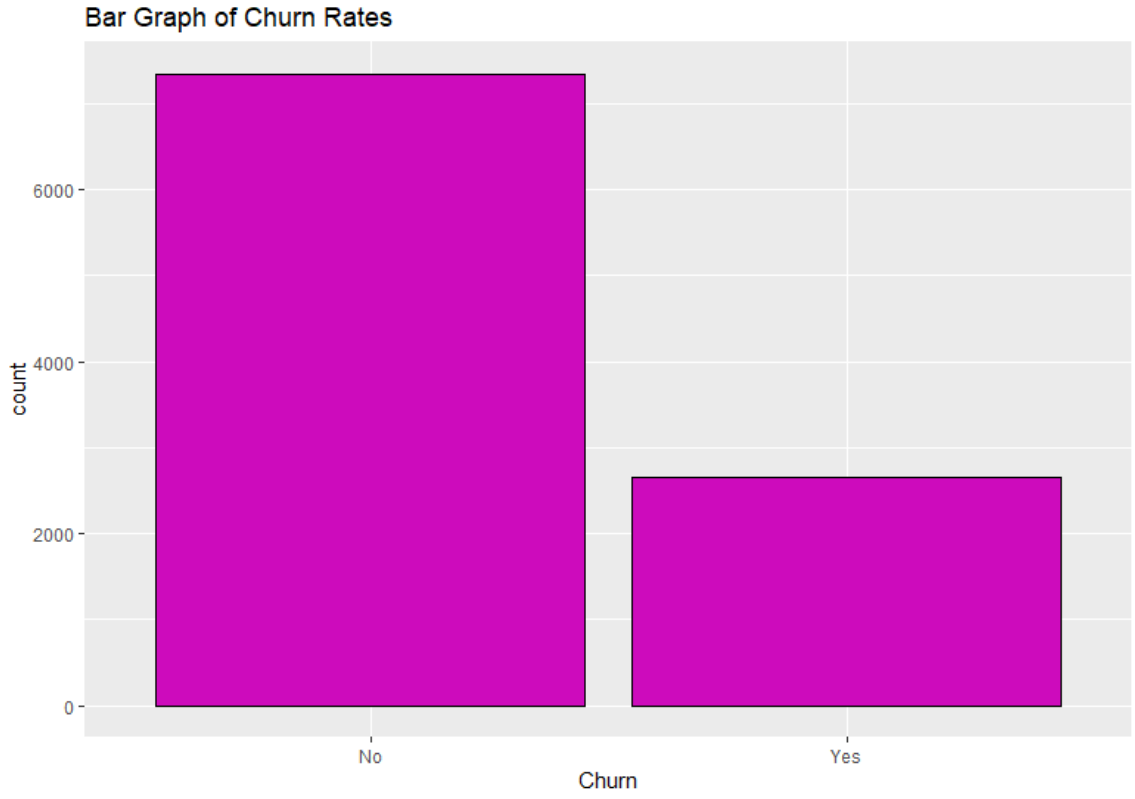
- **MonthlyCharge:**



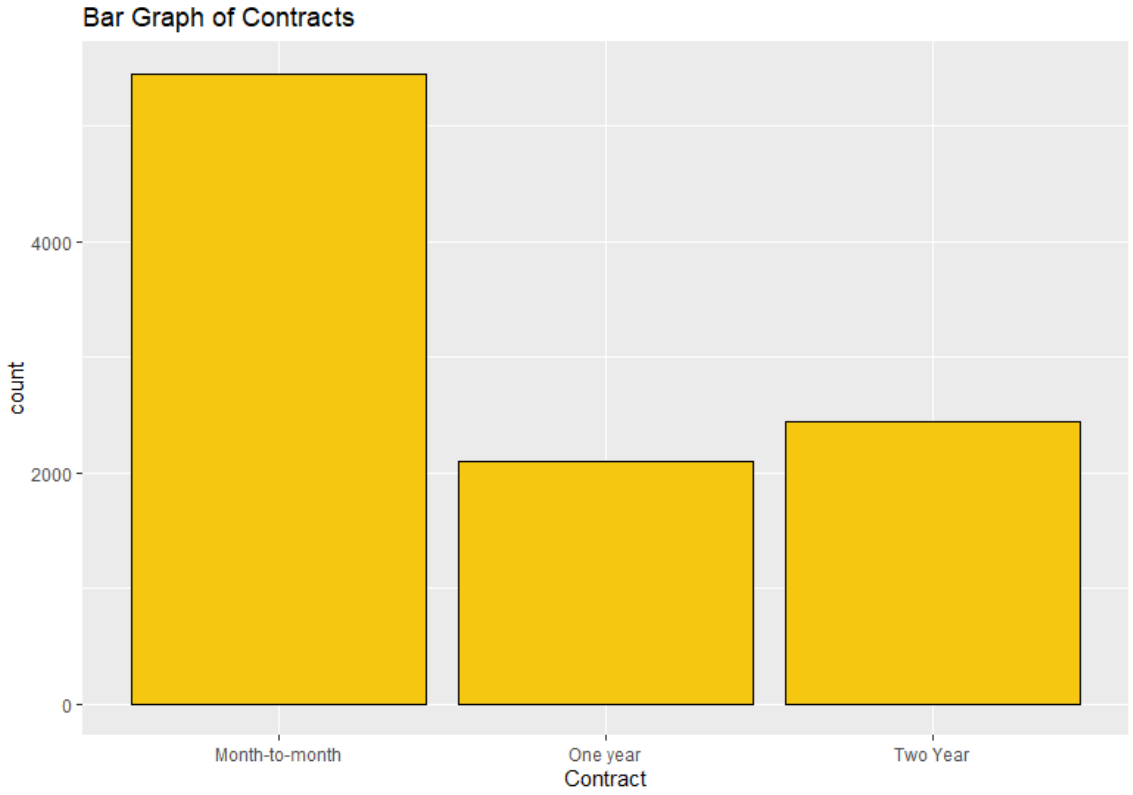
- **Bandwidth_GB_Year:**



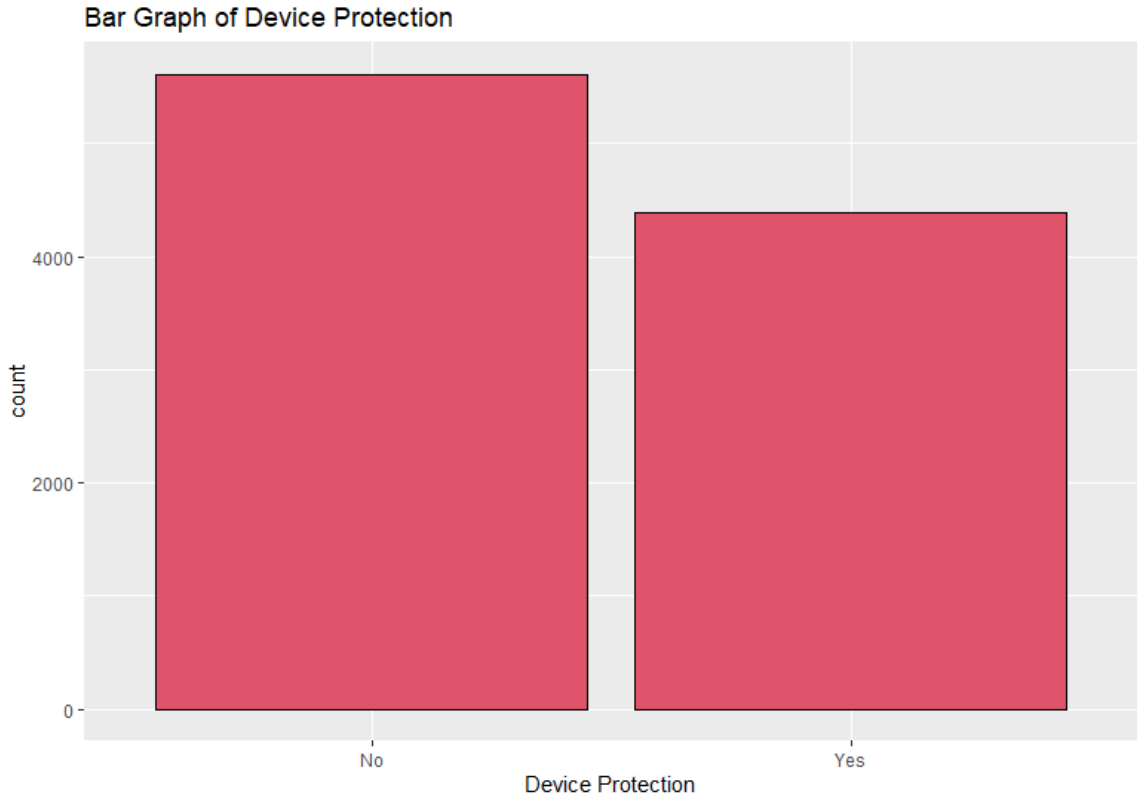
- **Churn:**



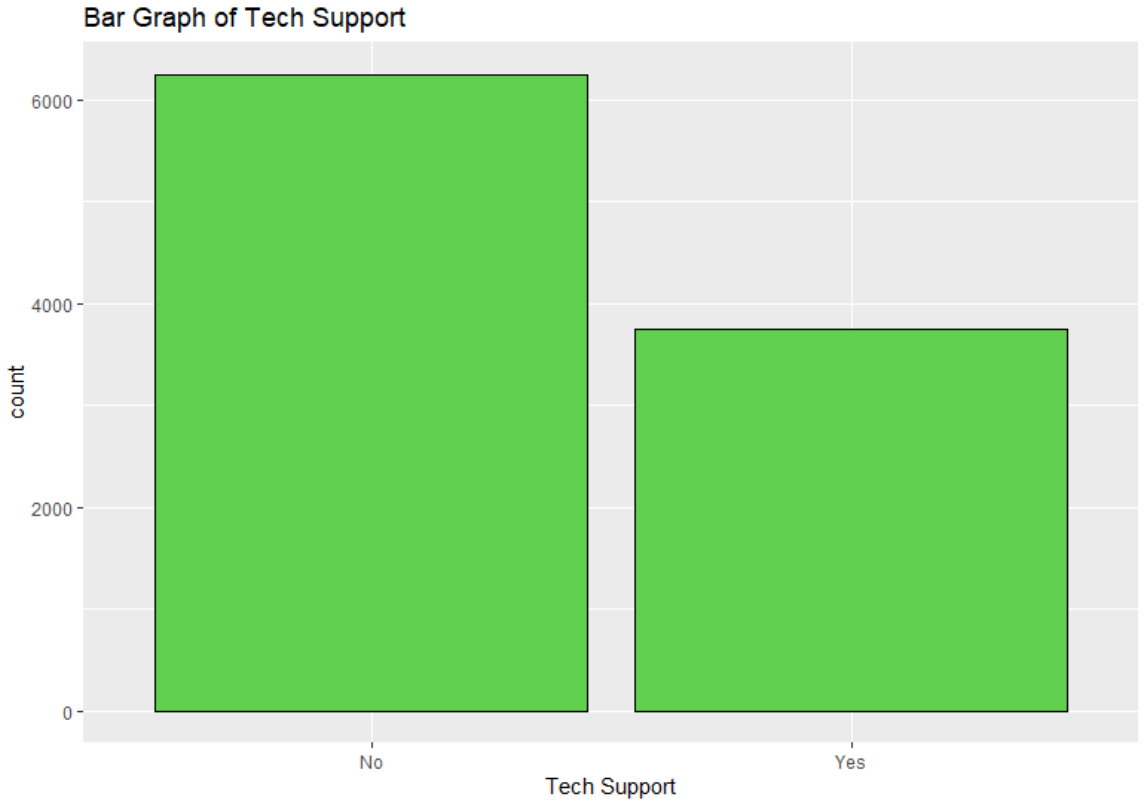
- **Contract:**



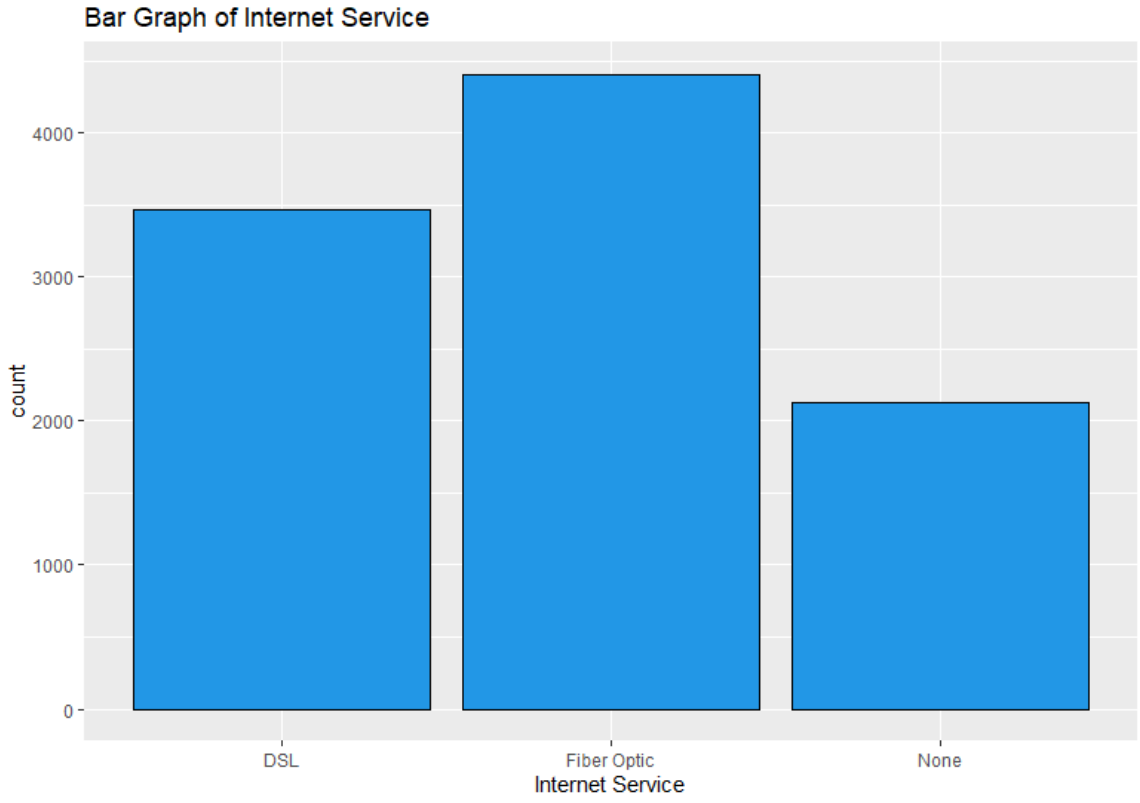
- **DeviceProtection:**



- **TechSupport:**

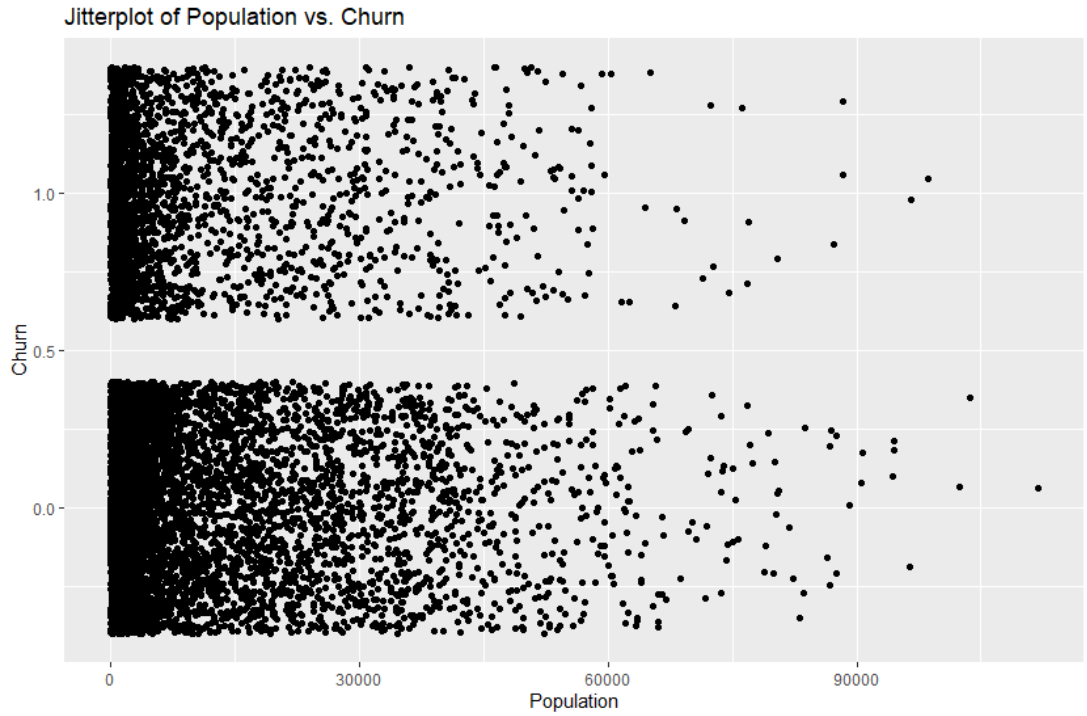


- **InternetService:**

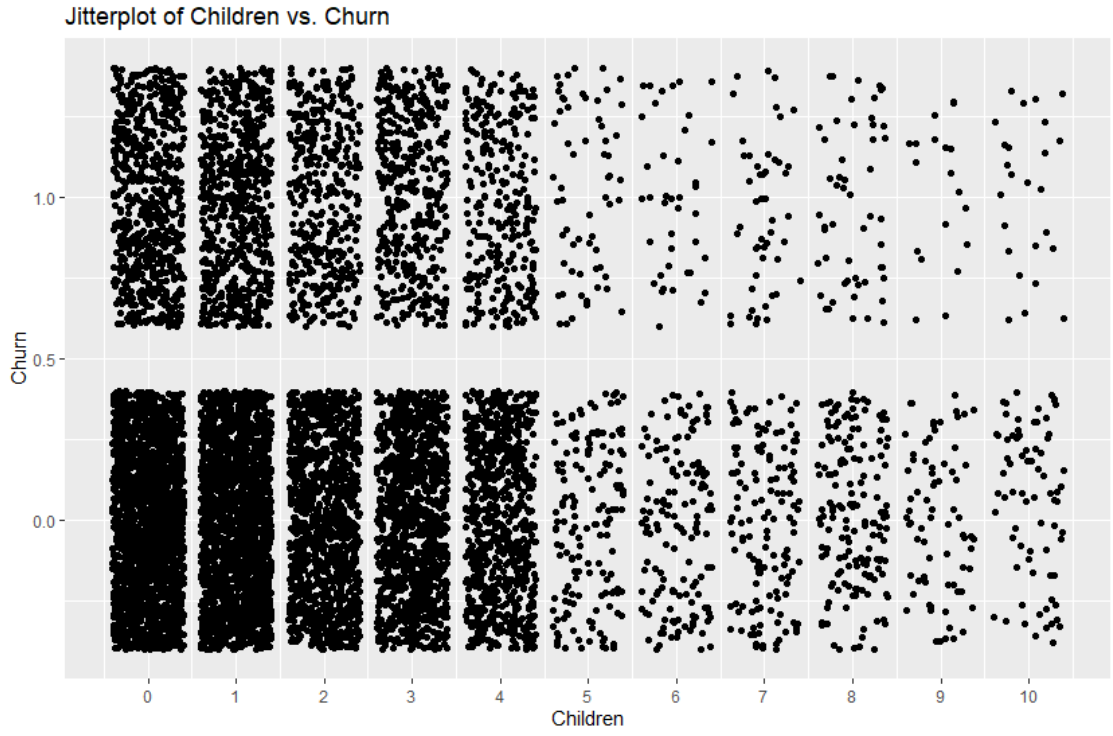


The following graphs provide further insight into the **bivariate distributions** of the relationships between each independent variable relevant to the research question and the dependent variable, Churn.

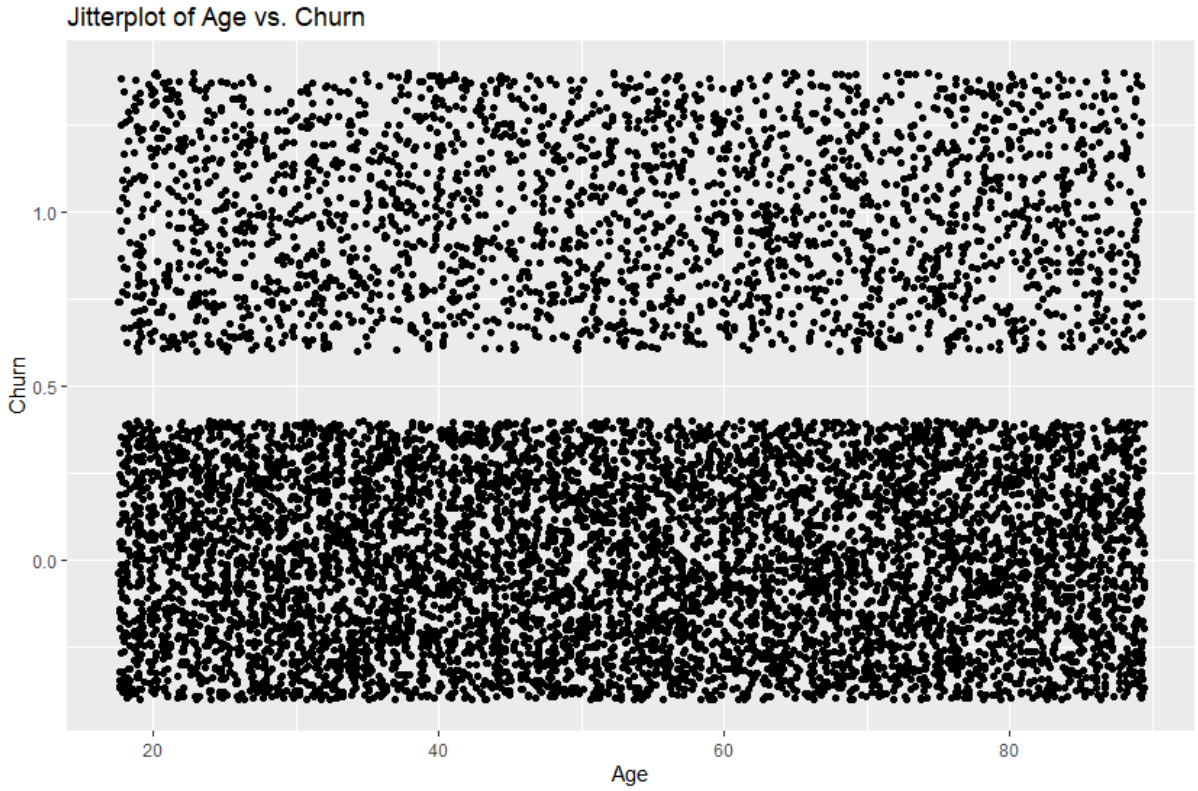
- **Population and Churn:**



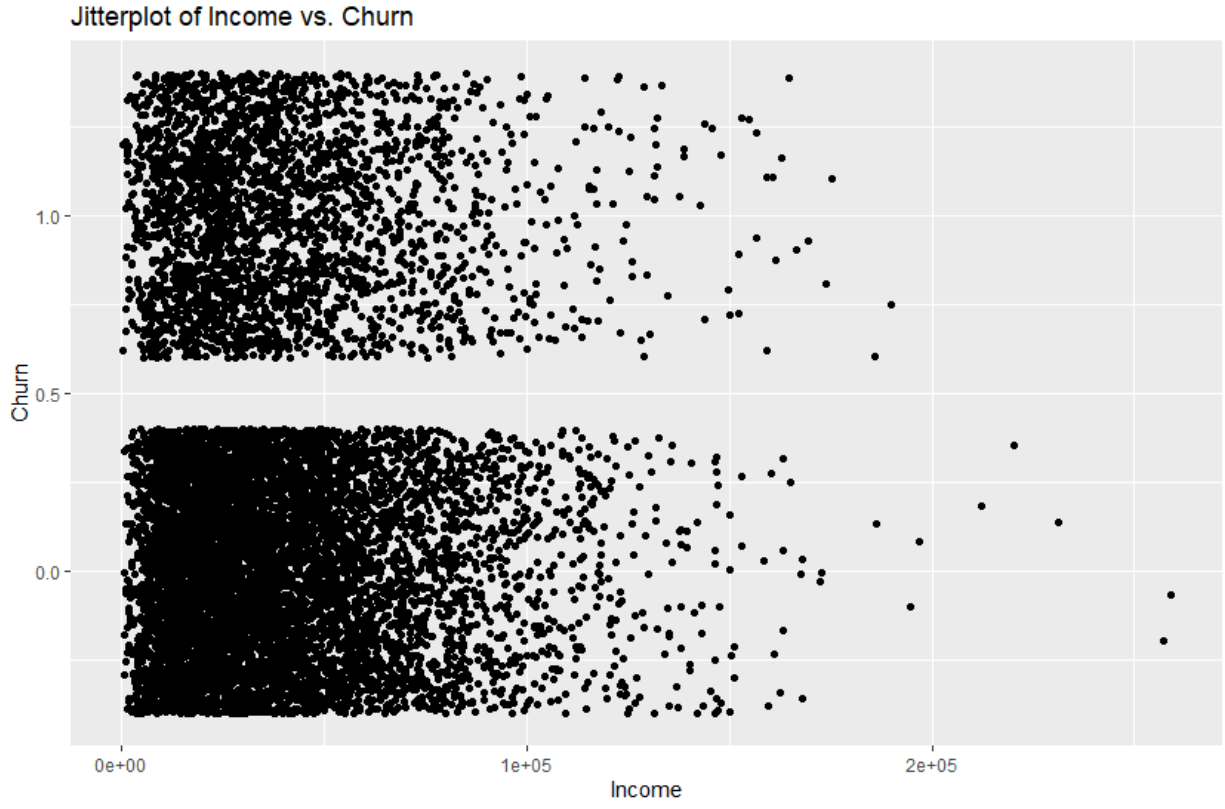
- **Children and Churn:**



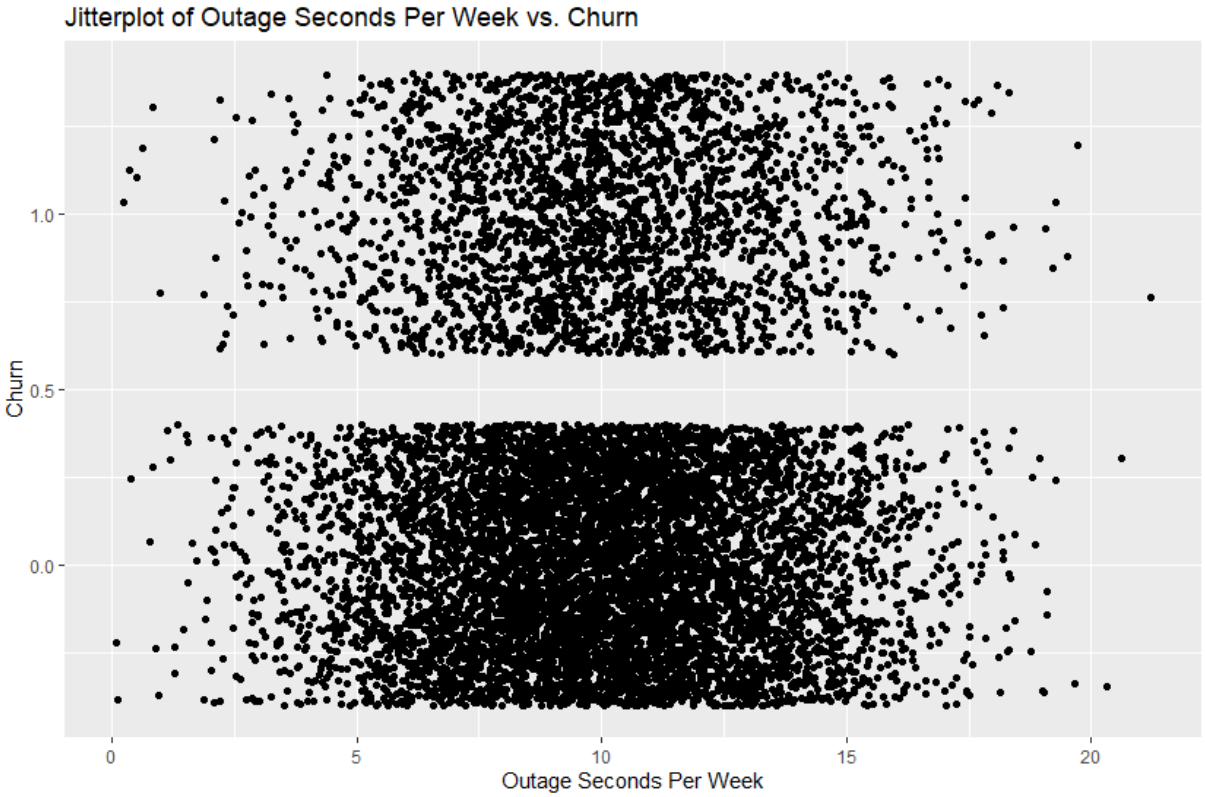
- **Age and Churn:**



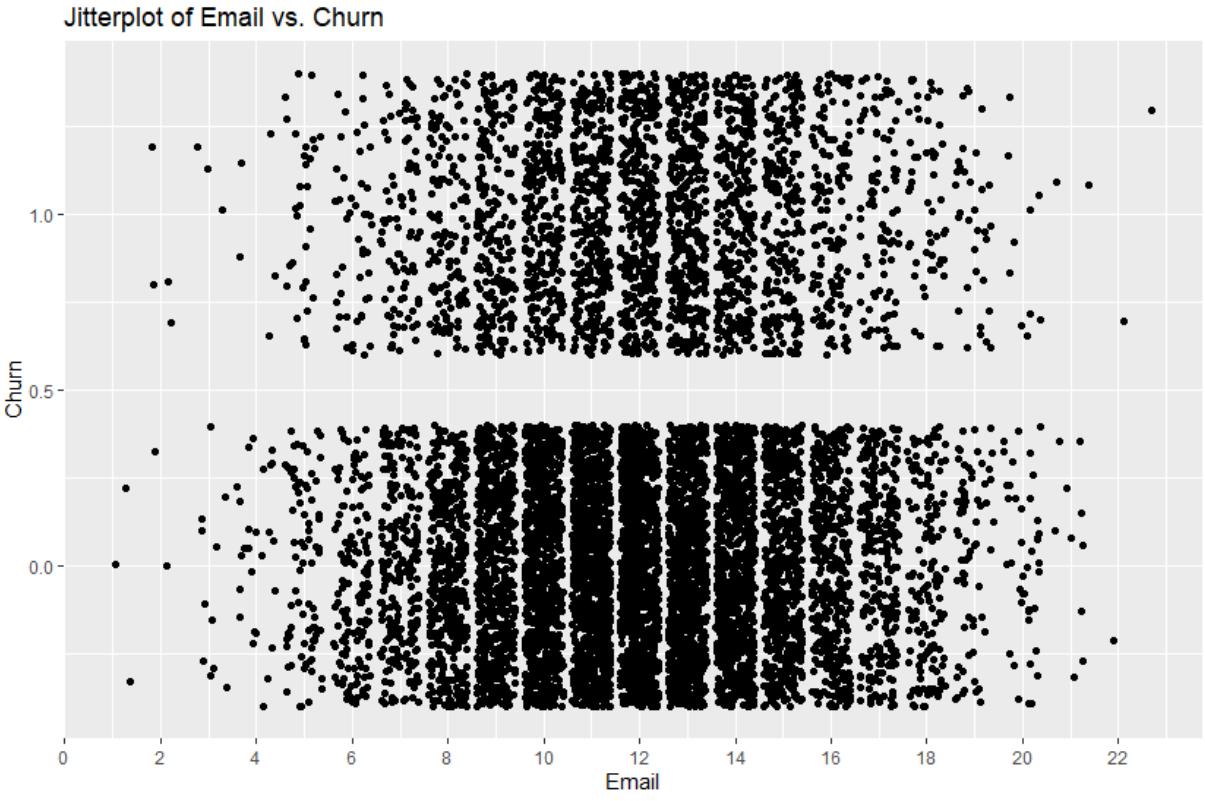
- **Income and Churn:**



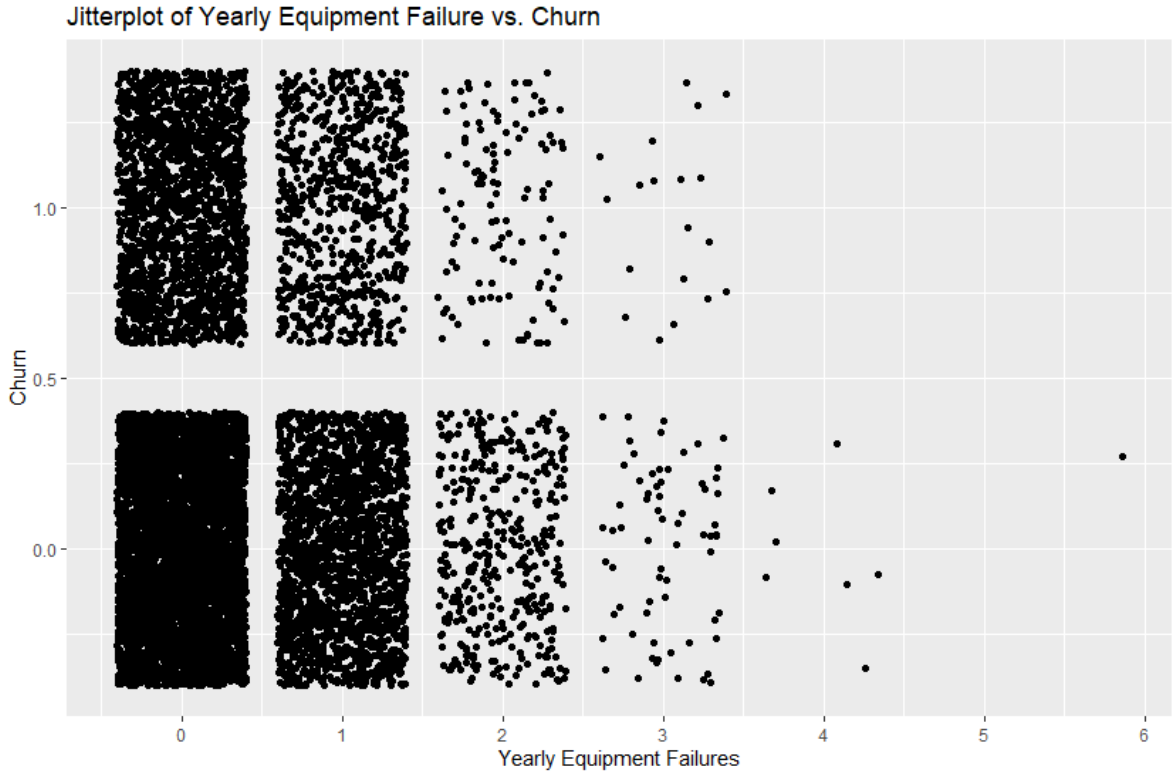
- **Outage_sec_perweek and Churn:**



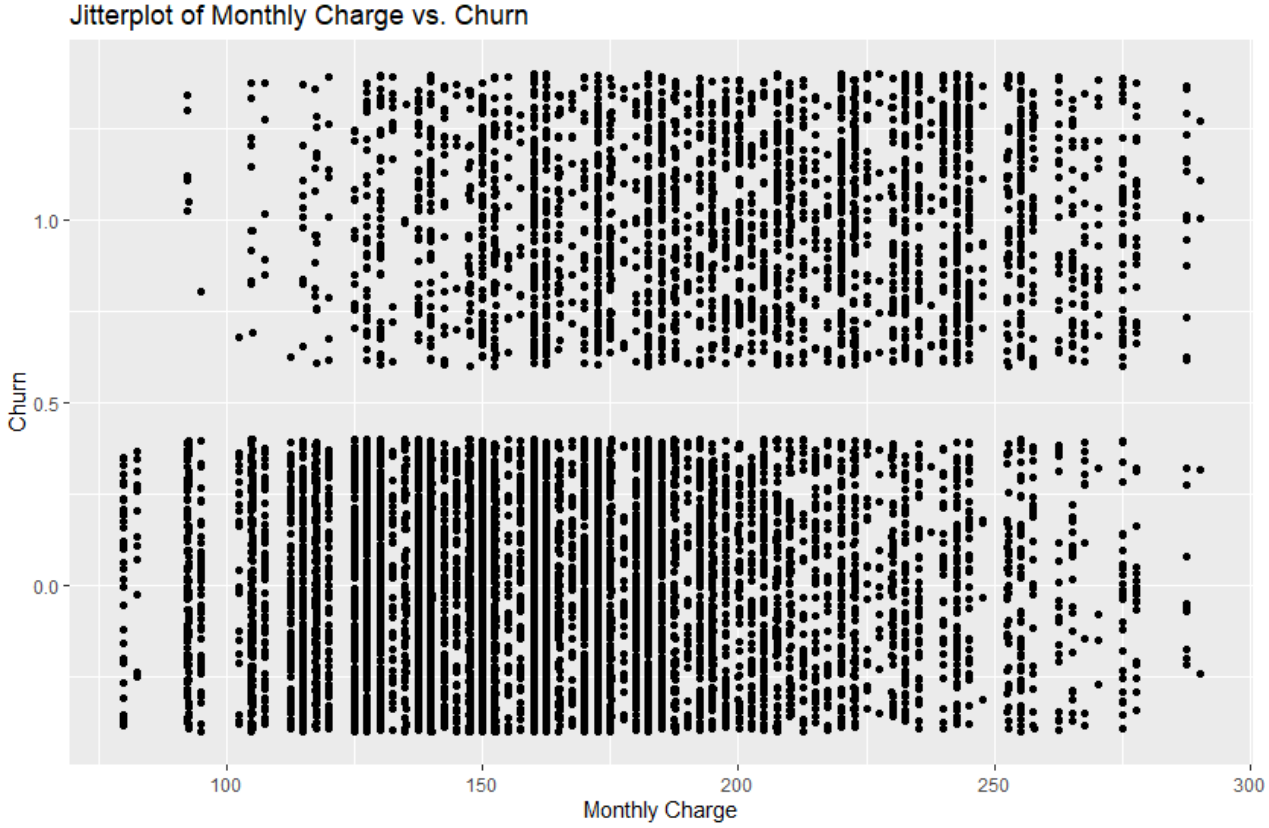
- **Email and Churn:**



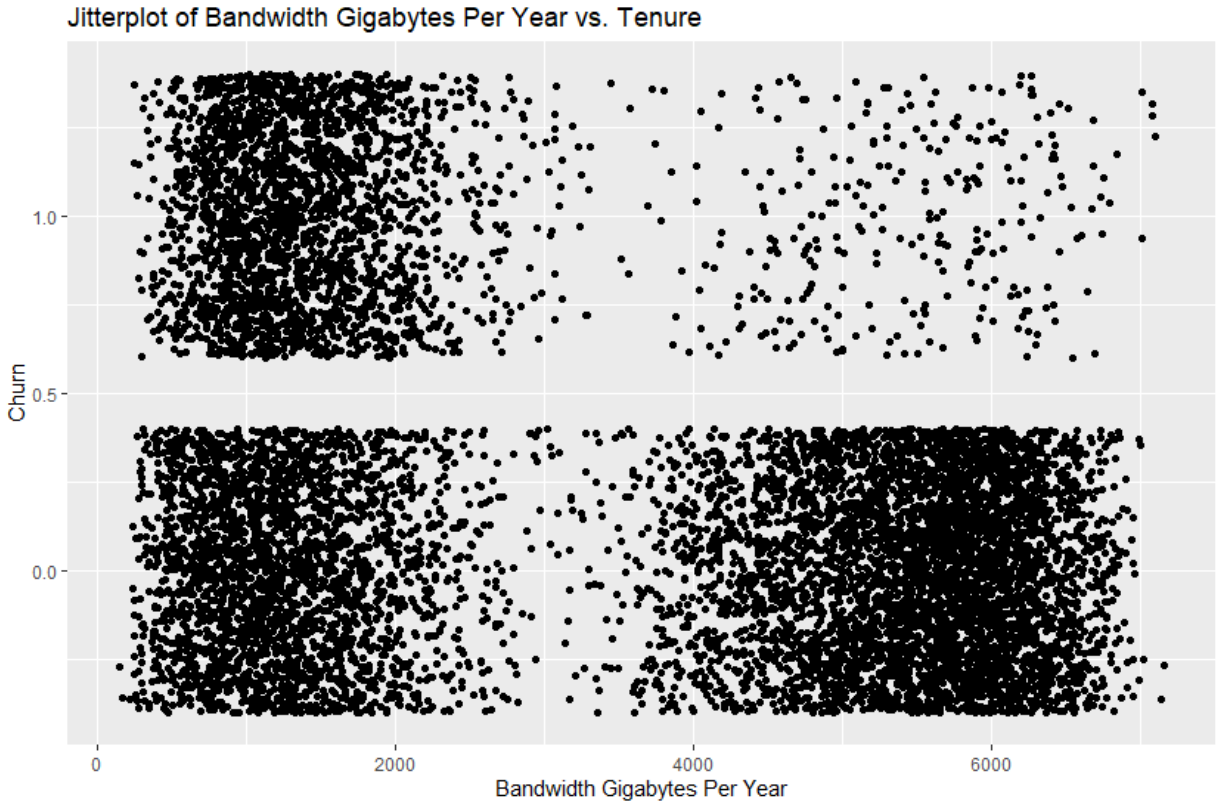
- **Yearly_equip_failure and Churn:**



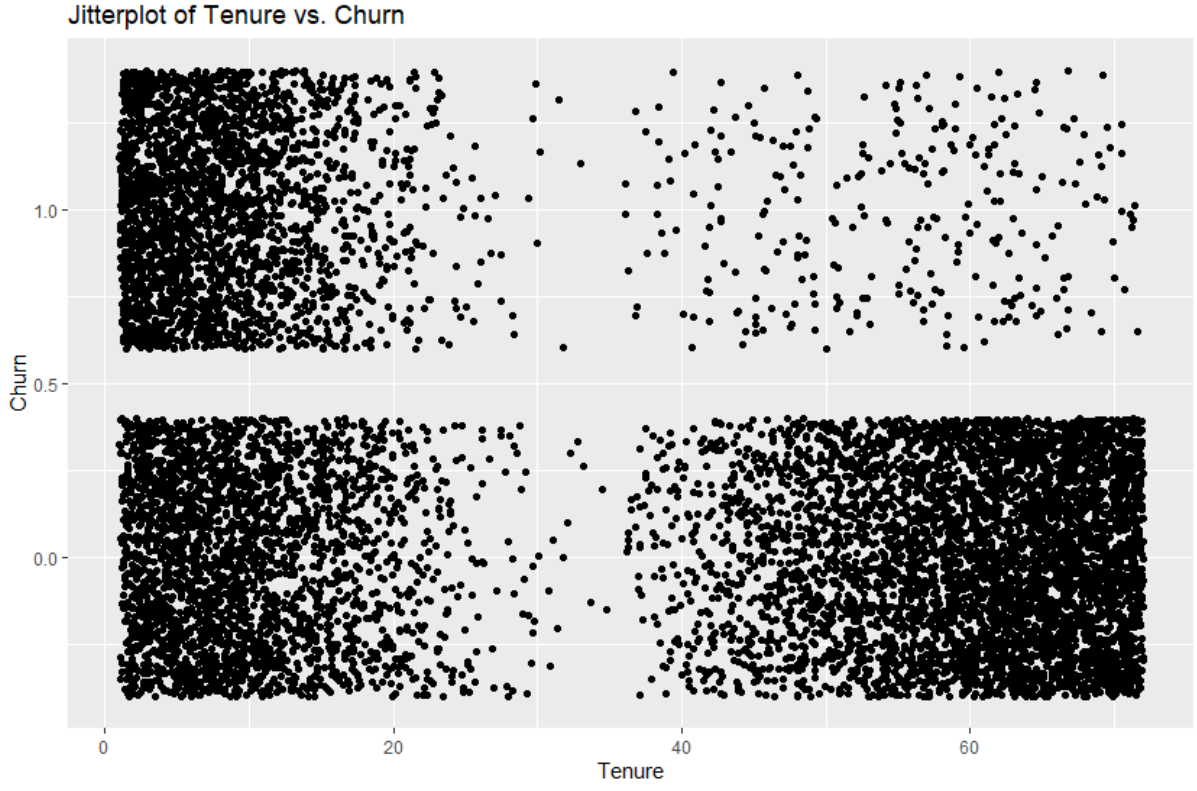
- **MonthlyCharge and Churn:**



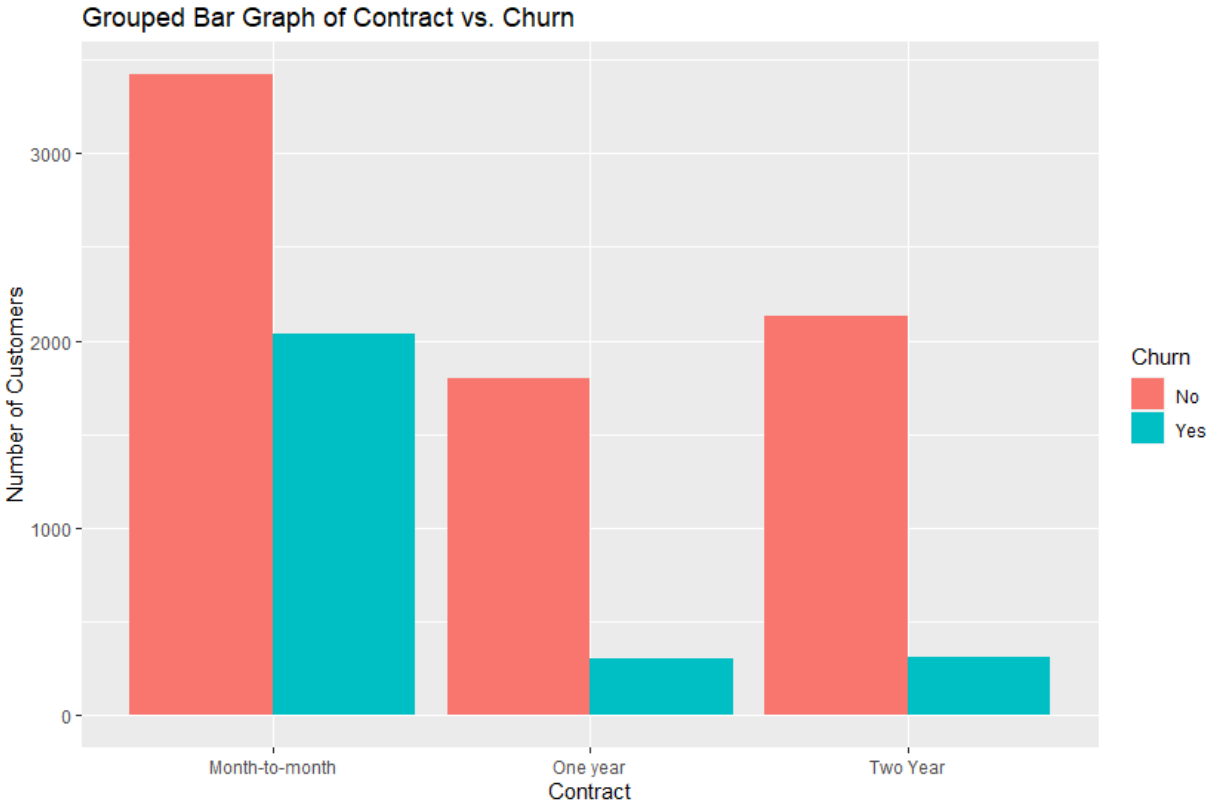
- **Bandwidth_GB_Year and Churn:**



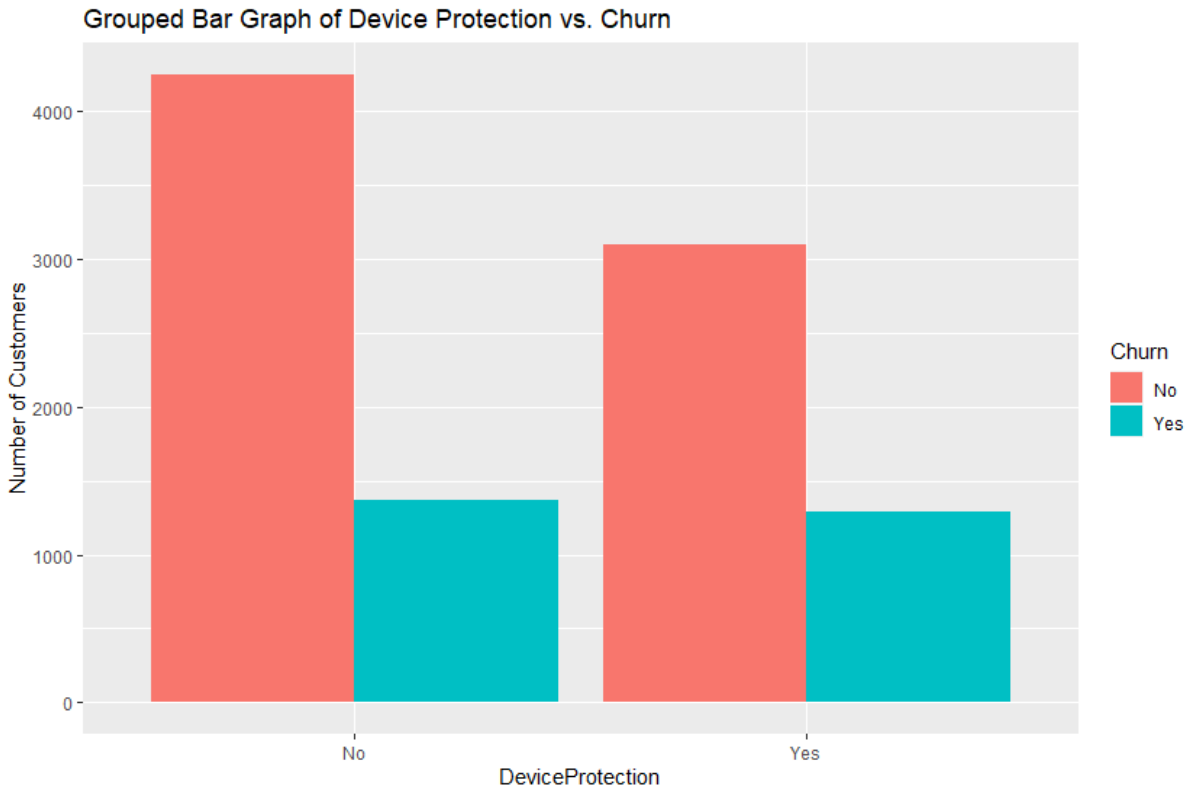
- **Tenure and Churn:**



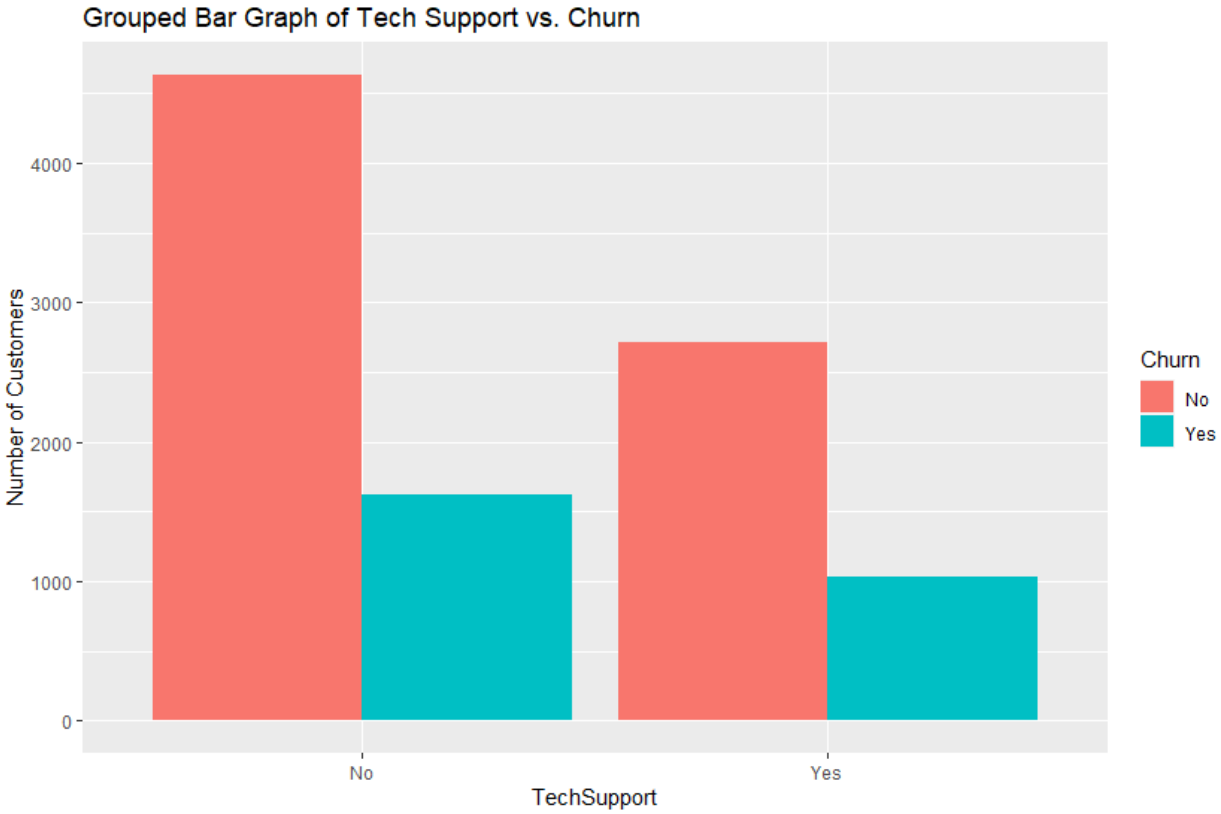
- **Contract and Churn:**



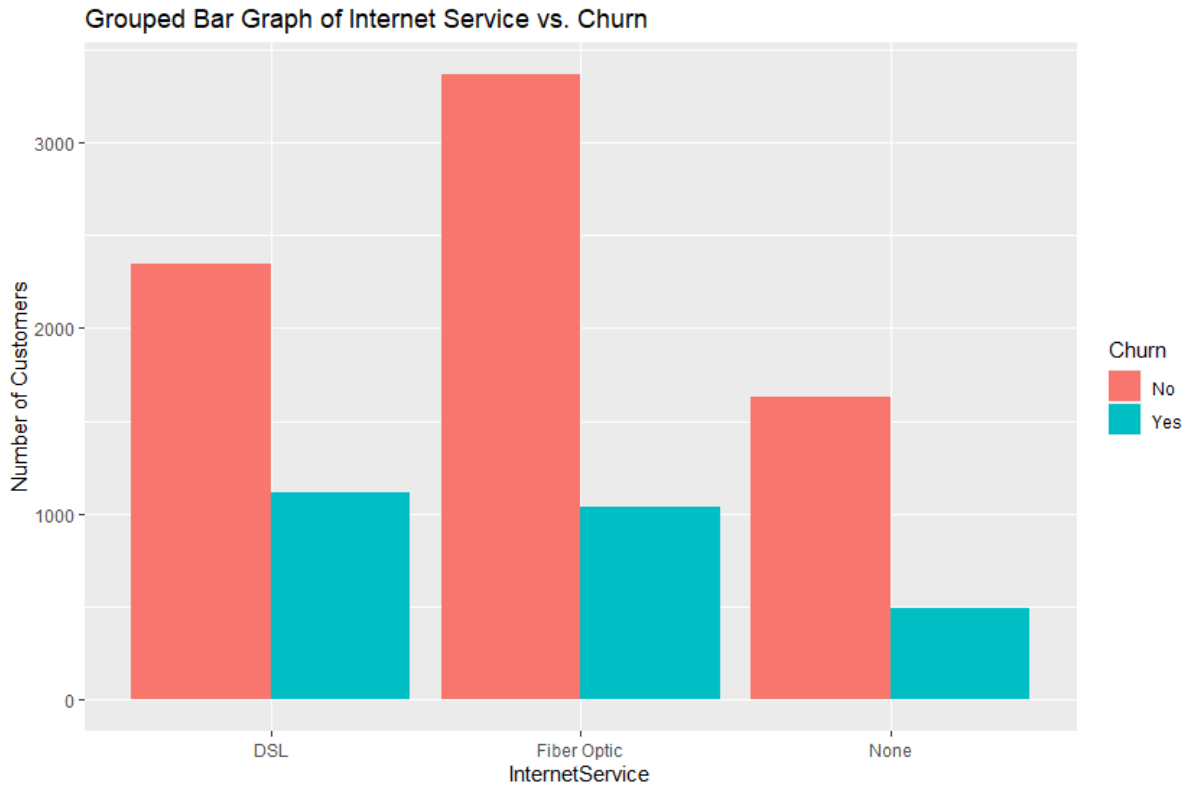
- **DeviceProtection and Churn:**



- **TechSupport and Churn:**



- **InternetService and Churn:**



C4. Data Transformation

In order to perform the multiple logistic regression with the inclusion of categorical response and explanatory variables, the necessary factors must be transformed into numeric variables. The following steps were taken to convert each variable from categorical to numeric.

- The **Churn** factor was converted by replacing all "No" responses with "0" and all "Yes" responses with "1."
- The **TechSupport** factor was converted by replacing all "No" responses with "0" and all "Yes" responses with "1."
- The **DeviceProtection** factor was converted by replacing all "No" responses with "0" and all "Yes" responses with "1."
- The **InternetService** factor was split into two new "dummy" variables using a one-hot encoding process:
    - **InternetService_DSL** contains a value of "0" for any customer who does not have DSL as an internet service provider, and a "1" for any customer who does have DSL as an internet service provider.
    - **InternetService_Fiber_Optic** contains a value of "0" for any customer who does not have Fiber Optic as an internet service provider, and a "1" for any customer who does have Fiber Optic as an internet service provider.
    - Note: an **InternetService_None** dummy variable **was not created**; a k-1 groups method was implemented to avoid potential multicollinearity. Any customers with a value of "0" listed for both InternetService_DSL and InternetService_Fiber_Optic were originally listed as "None" for InternetService.
- The **Contract** factor was split into two new "dummy" variables using a one-hot encoding process:
    - **Contract_One_Year** contains a value of "0" for any customer who does not have a one-year contract and a "1" for any customer who does have a one-year contract.
    - **Contract_Two_Year** contains a value of "0" for any customer who does not have a two-year contract and a "1" for any customer who does have a two-year contract.
    - Note: a **Contract_Month-to-month** dummy variable **was not created**; a k-1 groups method was implemented to avoid potential multicollinearity. Any customers with a value of "0" listed for both Contract_One_Year and Contract_Two_Year were originally listed as "Month-to-month" for Contract.

The following code executes the data transformations as described. An executable version of this code can be found in the attached file: Atwood_D208_Task2_Code.R.

```
churn$Churn <- as.numeric(revalue(churn$Churn, replace = c("No" = 0, "Yes" = 1)))
# converts Churn to numeric: 0 for No, 1 for Yes
```

```
churn$TechSupport <- as.numeric(revalue(churn$TechSupport, replace = c("No" = 0, "Yes" = 1)))
# converts TechSupport to numeric: 0 for No, 1 for Yes

churn$DeviceProtection <- as.numeric(revalue(churn$DeviceProtection, replace = c("No" = 0, "Yes" = 1)))
# converts DeviceProtection to numeric: 0 for No, 1 for Yes

library(fastDummies) # using fastDummies package
churn <- churn %>%
  dummy_cols("InternetService") %>%
  rename(InternetService_Fiber_Optic = 'InternetService_Fiber Optic') %>%
  select(-InternetService_None)
# re-expresses InternetService as numeric using one-hot encoding

churn <- churn %>%
  dummy_cols("Contract") %>%
  rename(Contract_One_Year = 'Contract_One year') %>%
  rename(Contract_Two_Year = 'Contract_Two Year') %>%
  select(-'Contract_Month-to-month')
# re-expresses Contract as numeric using one-hot encoding

churn <- churn %>%
  select(Tenure, Population, Children, Age, Income, Outage_sec_perweek, Email, Yearly_equip_failure, MonthlyCharge,
         Bandwidth_GB_Year, Churn, DeviceProtection, TechSupport, InternetService_DSL, InternetService_Fiber_Optic,
         Contract_One_Year, Contract_Two_Year)
# selecting only variables that will be used in the analysis
```

C5. Prepared Dataset

The resulting cleaned and prepared dataset has been written into a CSV file and attached with the following name: churn_new_task2.csv.

D1. Initial Model

An initial logistic regression model was constructed, with Churn as the response variable and all factors listed in section C2 as explanatory variables (including InternetService as two dummy variables, as described in section C4). The following is a summary of the **initial model statistics**:

```
Call:
glm(formula = Churn ~ Population + Children + Age + Income +
    Tenure + Outage_sec_perweek + Email + Yearly_equip_failure +
    DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
    InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year +
    Contract_Two_Year, family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8638  -0.2903  -0.0654   0.0899   3.5186

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.436e+00  3.108e-01 -20.705  < 2e-16 ***
Population                  -1.973e-06  2.610e-06  -0.756 0.449713
Children                    -6.253e-02  2.568e-02  -2.435 0.014888 *
Age                          8.321e-03  2.683e-03   3.101 0.001927 **
Income                       6.937e-07  1.318e-06   0.526 0.598593
Tenure                      -2.984e-01  4.893e-02  -6.099 1.07e-09 ***
Outage_sec_perweek           5.469e-04  1.254e-02   0.044 0.965213
Email                       -5.366e-03  1.229e-02  -0.437 0.662450
Yearly_equip_failure        -1.927e-02  5.863e-02  -0.329 0.742381
DeviceProtection            -3.539e-01  7.905e-02  -4.477 7.59e-06 ***
TechSupport                 -3.595e-01  8.198e-02  -4.385 1.16e-05 ***
MonthlyCharge                4.745e-02  2.741e-03  17.313  < 2e-16 ***
Bandwidth_GB_Year            2.301e-03  5.936e-04   3.876 0.000106 ***
InternetService_Fiber_Optic -1.489e+00  1.348e-01 -11.045  < 2e-16 ***
InternetService_DSL         -2.209e-01  2.357e-01  -0.937 0.348838
Contract_One_Year           -3.182e+00  1.214e-01 -26.198  < 2e-16 ***
Contract_Two_Year           -3.291e+00  1.194e-01 -27.559  < 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:  4638.6  on 9983  degrees of freedom
AIC: 4672.6

Number of Fisher Scoring iterations: 7
```

D2. Model Reduction

To refine the initial model, a **backward stepwise elimination** method was used to remove variables one at a time until only statistically significant variables remained in the model. The factors' p-values were used to determine statistical significance, with the following justification:

- The "t" statistic represents a standardized coefficient estimate, which measures the relative distance that a variable's coefficient estimate is from 0.
- Since a coefficient estimate of or near 0 would imply that the model likely does not require the relevant factor, variables with t statistics near 0 can be deemed statistically insignificant.
- The p-value, presented as "Pr(>|t|)" in the model summary, measures the probability of finding a t statistic as extreme or more extreme than the one found. If this probability is high (which results from a t statistic near 0), it is likely that the calculated coefficient estimate's value was simply random error. If this probability is low (which results from a t statistic far away from 0), it is unlikely the specific results were calculated by chance; thus, the results are statistically significant.

- An alpha level of 0.01 will be used to determine p-value significance. This means that any results found that were less than 1% likely to be calculated by random chance will be deemed statistically significant.

Variables were removed from the model one at a time, starting with the highest p-value and continuing until no variables remained with a p-value higher than the alpha level of 0.01. A new model summary was calculated after each individual removal in case the removal of one variable significantly changed the p-value of another.

- **Removal of Outage_sec_per week (p-value 0.9652):**

```
Call:
glm(formula = Churn ~ Population + Children + Age + Income +
    Tenure + Email + Yearly_equip_failure + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    InternetService_DSL + Contract_One_Year + Contract_Two_Year,
    family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8638  -0.2902  -0.0653   0.0900   3.5186

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.430e+00  2.839e-01 -22.651  < 2e-16 ***
Population                 -1.971e-06  2.610e-06  -0.755 0.450123
Children                   -6.255e-02  2.568e-02  -2.436 0.014841 *
Age                         8.321e-03  2.683e-03   3.102 0.001925 **
Income                      6.934e-07  1.318e-06   0.526 0.598706
Tenure                     -2.985e-01  4.891e-02  -6.102 1.05e-09 ***
Email                      -5.367e-03  1.229e-02  -0.437 0.662386
Yearly_equip_failure       -1.925e-02  5.863e-02  -0.328 0.742666
DeviceProtection           -3.539e-01  7.905e-02  -4.477 7.56e-06 ***
TechSupport                -3.594e-01  8.197e-02  -4.385 1.16e-05 ***
MonthlyCharge               4.745e-02  2.740e-03  17.315  < 2e-16 ***
Bandwidth_GB_Year           2.301e-03  5.934e-04   3.879 0.000105 ***
InternetService_Fiber_Optic -1.489e+00  1.348e-01 -11.048  < 2e-16 ***
InternetService_DSL        -2.211e-01  2.357e-01  -0.938 0.348237
Contract_One_Year          -3.182e+00  1.214e-01 -26.201  < 2e-16 ***
Contract_Two_Year          -3.291e+00  1.194e-01 -27.559  < 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:  4638.6  on 9984  degrees of freedom
AIC: 4670.6

Number of Fisher Scoring iterations: 7
```

- **Removal of Yearly_equip_failure (p-value 0.7427):**

```
Call:
glm(formula = Churn ~ Population + Children + Age + Income +
    Tenure + Email + DeviceProtection + TechSupport + MonthlyCharge +
    Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL +
    Contract_One_Year + Contract_Two_Year, family = "binomial",
    data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8613  -0.2903  -0.0652   0.0902   3.5155

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.438e+00  2.830e-01 -22.749  < 2e-16 ***
Population                 -1.966e-06  2.610e-06  -0.753 0.451192
Children                   -6.261e-02  2.568e-02  -2.438 0.014750 *
Age                         8.319e-03  2.683e-03   3.101 0.001931 **
Income                      6.885e-07  1.318e-06   0.523 0.601294
Tenure                     -2.985e-01  4.891e-02  -6.104 1.04e-09 ***
Email                      -5.308e-03  1.229e-02  -0.432 0.665841
DeviceProtection           -3.538e-01  7.904e-02  -4.476 7.60e-06 ***
TechSupport                -3.599e-01  8.196e-02  -4.391 1.13e-05 ***
MonthlyCharge               4.744e-02  2.740e-03  17.314  < 2e-16 ***
Bandwidth_GB_Year           2.302e-03  5.934e-04   3.880 0.000104 ***
InternetService_Fiber_Optic -1.489e+00  1.348e-01 -11.048  < 2e-16 ***
InternetService_DSL        -2.212e-01  2.357e-01  -0.939 0.347879
Contract_One_Year          -3.182e+00  1.214e-01 -26.203  < 2e-16 ***
Contract_Two_Year          -3.291e+00  1.194e-01 -27.561  < 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:  4638.7  on 9985  degrees of freedom
AIC: 4668.7

Number of Fisher Scoring iterations: 7
```

- **Removal of Email (p-value 0.6658):**

```
Call:
glm(formula = Churn ~ Population + Children + Age + Income +
    Tenure + DeviceProtection + TechSupport + MonthlyCharge +
    Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL +
    Contract_One_Year + Contract_Two_Year, family = "binomial",
    data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8720  -0.2906  -0.0653   0.0909   3.5149

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.501e+00  2.426e-01 -26.792  < 2e-16 ***
Population                 -1.982e-06  2.609e-06  -0.760 0.447451
Children                   -6.261e-02  2.568e-02  -2.438 0.014755 *
Age                         8.322e-03  2.683e-03   3.102 0.001923 **
Income                      6.969e-07  1.317e-06   0.529 0.596792
Tenure                     -2.985e-01  4.891e-02  -6.104 1.04e-09 ***
DeviceProtection           -3.541e-01  7.904e-02  -4.480 7.48e-06 ***
TechSupport                -3.603e-01  8.196e-02  -4.397 1.10e-05 ***
MonthlyCharge               4.743e-02  2.740e-03  17.312  < 2e-16 ***
Bandwidth_GB_Year           2.303e-03  5.934e-04   3.881 0.000104 ***
InternetService_Fiber_Optic -1.489e+00  1.347e-01 -11.048  < 2e-16 ***
InternetService_DSL        -2.218e-01  2.357e-01  -0.941 0.346708
Contract_One_Year          -3.181e+00  1.214e-01 -26.200  < 2e-16 ***
Contract_Two_Year          -3.291e+00  1.194e-01 -27.556  < 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:  4638.9  on 9986  degrees of freedom
AIC: 4666.9

Number of Fisher Scoring iterations: 7
```

- **Removal of Income (p-value 0.5968):**

```
Call:
glm(formula = Churn ~ Population + Children + Age + Tenure +
    DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
    InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year +
    Contract_Two_Year, family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8544  -0.2912  -0.0652   0.0904   3.5094

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.474e+00  2.369e-01 -27.325  < 2e-16 ***
Population                  -1.997e-06  2.609e-06  -0.765 0.444083
Children                    -6.203e-02  2.565e-02  -2.418 0.015609 *
Age                          8.263e-03  2.680e-03   3.083 0.002051 **
Tenure                      -2.977e-01  4.888e-02  -6.090 1.13e-09 ***
DeviceProtection            -3.526e-01  7.899e-02  -4.464 8.06e-06 ***
TechSupport                 -3.609e-01  8.195e-02  -4.403 1.07e-05 ***
MonthlyCharge                4.747e-02  2.739e-03  17.332  < 2e-16 ***
Bandwidth_GB_Year            2.292e-03  5.930e-04   3.865 0.000111 ***
InternetService_Fiber_Optic -1.490e+00  1.347e-01 -11.057  < 2e-16 ***
InternetService_DSL         -2.169e-01  2.355e-01  -0.921 0.357025
Contract_One_Year           -3.180e+00  1.214e-01 -26.199  < 2e-16 ***
Contract_Two_Year           -3.291e+00  1.194e-01 -27.560  < 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:  4639.1  on 9987  degrees of freedom
AIC: 4665.1

Number of Fisher Scoring iterations: 7
```

- **Removal of Population (p-value 0.4441):**

```
Call:
glm(formula = Churn ~ Children + Age + Tenure + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    InternetService_DSL + Contract_One_Year + Contract_Two_Year,
    family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8680  -0.2903  -0.0654   0.0899   3.5129

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.491779   0.235809 -27.530  < 2e-16 ***
Children                   -0.061946   0.025652  -2.415 0.015742 *
Age                         0.008279   0.002680   3.089 0.002010 **
Tenure                     -0.297847   0.048877  -6.094 1.10e-09 ***
DeviceProtection           -0.351224   0.078959  -4.448 8.66e-06 ***
TechSupport                -0.359409   0.081920  -4.387 1.15e-05 ***
MonthlyCharge               0.047438   0.002738  17.326  < 2e-16 ***
Bandwidth_GB_Year           0.002295   0.000593   3.870 0.000109 ***
InternetService_Fiber_Optic -1.488136   0.134708 -11.047  < 2e-16 ***
InternetService_DSL        -0.217165   0.235514  -0.922 0.356482
Contract_One_Year          -3.180399   0.121396 -26.199  < 2e-16 ***
Contract_Two_Year          -3.291564   0.119408 -27.566  < 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:  4639.7  on 9988  degrees of freedom
AIC: 4663.7

Number of Fisher Scoring iterations: 7
```

- **Removal of InternetService_DSL (p-value 0.3565):**

```
Call:
glm(formula = Churn ~ Children + Age + Tenure + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    Contract_One_Year + Contract_Two_Year, family = "binomial",
    data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8702  -0.2923  -0.0652   0.0905   3.5160

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.5556899  0.2257722 -29.037  < 2e-16 ***
Children                   -0.0463778  0.0192969  -2.403   0.0162 *
Age                         0.0066603  0.0020240   3.291   0.0010 ***
Tenure                     -0.2576309  0.0219497 -11.737  < 2e-16 ***
DeviceProtection           -0.3329128  0.0764023  -4.357 1.32e-05 ***
TechSupport                -0.3819086  0.0782130  -4.883 1.05e-06 ***
MonthlyCharge               0.0493824  0.0017571  28.104  < 2e-16 ***
Bandwidth_GB_Year           0.0018043  0.0002612   6.909 4.88e-12 ***
InternetService_Fiber_Optic -1.5261972  0.1283847 -11.888  < 2e-16 ***
Contract_One_Year          -3.1815142  0.1213893 -26.209  < 2e-16 ***
Contract_Two_Year          -3.2885266  0.1192256 -27.582  < 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:  4640.6  on 9989  degrees of freedom
AIC: 4662.6

Number of Fisher Scoring iterations: 7
```

- **Removal of Children (p-value 0.0162):**

```
Call:
glm(formula = Churn ~ Age + Tenure + DeviceProtection + TechSupport +
    MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    Contract_One_Year + Contract_Two_Year, family = "binomial",
    data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8649  -0.2921  -0.0653   0.0894   3.5152

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.6111966  0.2247714 -29.413  < 2e-16 ***
Age                         0.0058624  0.0019951   2.938   0.0033 **
Tenure                     -0.2357461  0.0198919 -11.851  < 2e-16 ***
DeviceProtection           -0.3247034  0.0762794  -4.257 2.07e-05 ***
TechSupport                -0.3916254  0.0780819  -5.016 5.29e-07 ***
MonthlyCharge               0.0504754  0.0017032  29.636  < 2e-16 ***
Bandwidth_GB_Year           0.0015401  0.0002362   6.519 7.08e-11 ***
InternetService_Fiber_Optic -1.6214744  0.1221960 -13.269  < 2e-16 ***
Contract_One_Year          -3.1789099  0.1212900 -26.209  < 2e-16 ***
Contract_Two_Year          -3.2866802  0.1191948 -27.574  < 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:  4646.4  on 9990  degrees of freedom
AIC: 4666.4

Number of Fisher Scoring iterations: 7
```

To check for variables within this model that exhibit multicollinearity, a **variance inflation factor (VIF)** was calculated for each variable, with the following results:

```
                Age                  Tenure           DeviceProtection                TechSupport
           1.198703              117.457807                 1.047684                  1.051320
     MonthlyCharge       Bandwidth_GB_Year InternetService_Fiber_Optic         Contract_One_Year
           3.621904              130.423691                 2.642317                  1.366439
Contract_Two_Year
           1.397725
```

VIF measures the ratio of a factor's variance in the full model (with all other factors) to the variance by itself. If there exists a significant relationship between two or more explanatory variables, this could be inaccurately reflected in their relationships with the response variable. The standard cutoff for VIF significance is 10. There are two variables with a VIF much greater than 10: Tenure and Bandwidth_GB_Year. The variable with the highest VIF, Bandwidth_GB_Year, was therefore removed from the model.

- **Removal of Bandwidth_GB_Year (VIF 130.4237):**

```
Call:
glm(formula = Churn ~ Age + Tenure + DeviceProtection + TechSupport +
    MonthlyCharge + InternetService_Fiber_Optic + Contract_One_Year +
    Contract_Two_Year, family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8965  -0.2976  -0.0677   0.0951   3.5123

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -6.3188465  0.2166703 -29.163  < 2e-16 ***
Age                          0.0005879  0.0018144   0.324 0.745933
Tenure                      -0.1083351  0.0027172 -39.870  < 2e-16 ***
DeviceProtection            -0.2749748  0.0755549  -3.639 0.000273 ***
TechSupport                 -0.4645898  0.0770054  -6.033 1.61e-09 ***
MonthlyCharge                0.0567068  0.0014563  38.940  < 2e-16 ***
InternetService_Fiber_Optic -2.1764325  0.0897875 -24.240  < 2e-16 ***
Contract_One_Year           -3.1378432  0.1204027 -26.061  < 2e-16 ***
Contract_Two_Year           -3.2236309  0.1174942 -27.437  < 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:  4689.5  on 9991  degrees of freedom
AIC: 4707.5

Number of Fisher Scoring iterations: 7
```

This variable reduction due to multicollinearity led to an increase in the p-value for Age above the determined alpha level of 0.01, so the Age variable was removed as well.

- **Removal of Age:**

```
Call:
glm(formula = Churn ~ Tenure + DeviceProtection + TechSupport +
    MonthlyCharge + InternetService_Fiber_Optic + Contract_One_Year +
    Contract_Two_Year, family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8906  -0.2989  -0.0677   0.0953   3.5165

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -6.288474   0.195118 -32.229  < 2e-16 ***
Tenure                     -0.108335   0.002717 -39.868  < 2e-16 ***
DeviceProtection           -0.274564   0.075541  -3.635 0.000278 ***
TechSupport                -0.464288   0.076997  -6.030 1.64e-09 ***
MonthlyCharge               0.056710   0.001456  38.942  < 2e-16 ***
InternetService_Fiber_Optic -2.176217   0.089781 -24.239  < 2e-16 ***
Contract_One_Year          -3.138966   0.120353 -26.081  < 2e-16 ***
Contract_Two_Year          -3.224337   0.117484 -27.445  < 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:  4689.6  on 9992  degrees of freedom
AIC: 4705.6

Number of Fisher Scoring iterations: 7
```

Once again, a VIF was calculated for each remaining variable to ensure no multicollinearity, with the following results:

```
                    Tenure           DeviceProtection                TechSupport             MonthlyCharge
                  2.223951                  1.036896                   1.033017                  2.688015
InternetService_Fiber_Optic          Contract_One_Year          Contract_Two_Year
                  1.438843                  1.352020                   1.372286
```

The remaining variables all have VIF values less than the cutoff of 10. Also, the p-values for all remaining variables in the reduced model are less than the alpha level of 0.01. Therefore, no further reduction is required as specified by the factor elimination process.

D3. Reduced Model

The following explanatory variables were deemed statistically insignificant or exhibited multicollinearity and were removed from the logistic regression model in a backward stepwise elimination method: Outage_sec_perweek, Yearly_equip_failure, Email, Income, Population, InternetService_DSL, Children, Bandwidth_GB_Year, and Age.

The following explanatory variables were deemed statistically significant to the logistic regression model and were retained in the reduced model: Tenure, DeviceProtection, TechSupport, MonthlyCharge, InternetService_Fiber_Optic, Contract_One_Year, and Contract_Two_Year.

The following is a summary of the **reduced model statistics**:

```
Call:
glm(formula = Churn ~ Tenure + DeviceProtection + TechSupport +
    MonthlyCharge + InternetService_Fiber_Optic + Contract_One_Year +
    Contract_Two_Year, family = "binomial", data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8906  -0.2989  -0.0677   0.0953   3.5165

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -6.288474   0.195118 -32.229  < 2e-16 ***
Tenure                       -0.108335   0.002717 -39.868  < 2e-16 ***
DeviceProtection             -0.274564   0.075541  -3.635 0.000278 ***
TechSupport                  -0.464288   0.076997  -6.030 1.64e-09 ***
MonthlyCharge                 0.056710   0.001456  38.942  < 2e-16 ***
InternetService_Fiber_Optic -2.176217   0.089781 -24.239  < 2e-16 ***
Contract_One_Year            -3.138966   0.120353 -26.081  < 2e-16 ***
Contract_Two_Year            -3.224337   0.117484 -27.445  < 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:  4689.6  on 9992  degrees of freedom
AIC: 4705.6

Number of Fisher Scoring iterations: 7
```

E1. Model Comparison

To measure the effectiveness and statistical efficiency of both the initial and reduced regression models, the following model metrics will be compared and evaluated: confusion matrix accuracy, McFadden’s R-squared, and residual deviance.

The following metrics represent the **initial model**:

- Confusion matrix accuracy: 0.896.
- McFadden’s R-squared: 0.599.
- Residual deviance: 4638.6.

The following metrics represent the **reduced model**:

- Confusion matrix accuracy: 0.898.
- McFadden’s R-squared: 0.594.
- Residual deviance: 4689.6.

There is **a minor difference** between the model metrics for the initial and reduced models. Both show statistically significant results, and many metrics are very close. This may imply that the removal of several variables did not make noticeable changes to the significance of the results; however, it could also show that the removed variables did not contribute heavily to the initial model. Evidence of this can be found in each metric:

- **Confusion Matrix Accuracy:**
  - The confusion matrix accuracy represents the proportion of customers whose Churn status was correctly predicted by the logistic regression model.
  - The accuracy metrics for the initial and reduced models are 0.896 and 0.898, respectively. This indicates that both models are useful predictors of Churn status, predicting at high rates within 0.2% of each other.
  - This indicates that the reduced model can predict Churn with slightly better accuracy than the initial model, despite having less overall information. The removed information, then, was likely unnecessary as a predictor.
- **McFadden's R-squared:**
  - McFadden's R-squared measures a model's goodness of fit for predicting Churn on a scale from 0 to 1, with values closer to 1 indicating a better fit for the data.
  - The initial model has a slightly higher McFadden's R-squared value than the reduced model: 0.599 to 0.594. These values indicate that both models are good fits for the data and are quite similar in their effectiveness.
  - This indicates that the initial model is only a marginally better fit for predicting Churn than the reduced model, despite having several more variables' worth of information. The reduced model fits the predictions more efficiently.
- **Residual deviance:**
  - The residual deviance is a value designed to be compared to a dataset's null deviance to determine how well the response variable can be predicted by a model. The null deviance measures the predictive power of a model with only an intercept value, while the residual deviance measures the predictive power of a model with all of its predictor variables. A lower residual deviance indicates a better model.
  - The null deviance for both the initial and reduced model was 11,564.4. The residual deviance for the initial model was 4638.6 and for the reduced model was 4689.6. Both models have much lower residual deviance than null deviance, with the initial model having a slightly lower residual deviance.
  - This indicates that both models are good predictors of Churn status, and though the initial model produces slightly better results, the reduced model uses fewer variables with similar significance. Therefore, the reduced model would be preferred over the initial model due to the clarity of the results.

E2. Calculations

The following summarizes all calculations performed in the analysis, including confusion matrices for the initial and reduced models and relevant metrics for each confusion matrix.

- **Initial Model Confusion Matrix:**

predicted_responses_initial_model	actual_responses 0	1
0	6881	567
1	469	2083

- **Initial Model Confusion Matrix Metrics:**

	.metric	.estimator	.estimate
	<chr>	<chr>	<dbl>
1	accuracy	binary	0.896
2	kap	binary	0.731
3	sens	binary	0.936
4	spec	binary	0.786
5	ppv	binary	0.924
6	npv	binary	0.816
7	mcc	binary	0.731
8	j_index	binary	0.722
9	bal_accuracy	binary	0.861
10	detection_prevalence	binary	0.745
11	precision	binary	0.924
12	recall	binary	0.936
13	f_meas	binary	0.930

- **Reduced Model Confusion Matrix:**

	actual_responses	
predicted_responses_reduced_model	0	1
0	6892	563
1	458	2087

- **Reduced Model Confusion Matrix Metrics:**

	.metric	.estimator	.estimate
	<chr>	<chr>	<dbl>
1	accuracy	binary	0.898
2	kap	binary	0.735
3	sens	binary	0.938
4	spec	binary	0.788
5	ppv	binary	0.924
6	npv	binary	0.820
7	mcc	binary	0.735
8	j_index	binary	0.725
9	bal_accuracy	binary	0.863
10	detection_prevalence	binary	0.746
11	precision	binary	0.924
12	recall	binary	0.938
13	f_meas	binary	0.931

E3. Logistic Regression Code

The following code executes the creation of the initial model and the backward stepwise elimination method used to create the reduced model as described. An executable version of this code can be found in the attached file: Atwood_D208_Task2_Code.R.

```
# CREATING INITIAL MODEL:

in_model <- glm(Churn ~ Population + Children + Age + Income + Tenure + Outage_sec_perweek
+ Email + Yearly_equip_failure +
    DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year + Contract_Two_Year,
churn, family = "binomial")

summary(in_model)
```

```
# REDUCING MODEL:

re_model <- glm(Churn ~ Population + Children + Age + Income + Tenure + Email + Yearly_equip_failure +
                DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Outage_sec_perweek

summary(re_model)

re_model <- glm(Churn ~ Population + Children + Age + Income + Tenure + Email + DeviceProtection + TechSupport +
                MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Yearly_equip_failure

summary(re_model)

re_model <- glm(Churn ~ Population + Children + Age + Income + Tenure + DeviceProtection + TechSupport +
                MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Email

summary(re_model)

re_model <- glm(Churn ~ Population + Children + Age + Tenure + DeviceProtection + TechSupport +
                MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Income

summary(re_model)

re_model <- glm(Churn ~ Children + Age + Tenure + DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
                InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
```

```
# removing Population

summary(re_model)

re_model <- glm(Churn ~ Children + Age + Tenure + DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
                InternetService_Fiber_Optic + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing InternetService_DSL

summary(re_model)

re_model <- glm(Churn ~ Age + Tenure + DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
                InternetService_Fiber_Optic + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Children

summary(re_model)

library(car) # using car package
vif(re_model) # checking Variance Inflation Factor

re_model <- glm(Churn ~ Age + Tenure + DeviceProtection + TechSupport + MonthlyCharge +
                InternetService_Fiber_Optic + Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Bandwidth_GB_Year

summary(re_model)

re_model <- glm(Churn ~ Tenure + DeviceProtection + TechSupport + MonthlyCharge + InternetService_Fiber_Optic +
                Contract_One_Year + Contract_Two_Year, churn, family = "binomial")
# removing Age

summary(re_model)

vif(re_model) # checking Variance Inflation Factor
```

F1. Regression Equation

The coefficients of each factor in the final reduced model are as follows:

```
Coefficients:
                              Estimate
(Intercept)                  -6.288474
Tenure                       -0.108335
DeviceProtection             -0.274564
TechSupport                  -0.464288
MonthlyCharge                 0.056710
InternetService_Fiber_Optic -2.176217
Contract_One_Year            -3.138966
Contract_Two_Year            -3.224337
```

These produce the following final equation of the reduced logistic model, with coefficients rounded to the nearest hundredth value. Note: $\hat{p}$ represents the probability of a customer leaving the service ("Yes" in the Churn category).

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -6.29 - 0.11(Tenure) - 0.27(DeviceProtection) - 0.46(TechSupport) + 0.06(MonthlyCharge) - 2.18(InternetService\_Fiber\_Optic) - 3.14(Contract\_One\_Year) - 3.22(Contract\_Two\_Year)$$

For interpretation purposes, all coefficients other than the intercept have been converted into **odds ratios** using the formula e^{β_n}, where β_n represents the coefficient estimate for the nth predictor variable. The resulting values have been calculated and rounded to the nearest hundredth:

- **Tenure**: odds ratio of 0.90.
- **DeviceProtection**: odds ratio of 0.76.
- **TechSupport**: odds ratio of 0.63.
- **MonthlyCharge**: odds ratio of 1.06.
- **InternetService_Fiber_Optic**: odds ratio of 0.11.
- **Contract_One_Year**: odds ratio of 0.04.
- **Contract_Two_Year**: odds ratio of 0.04.

An odds ratio represents the percent change in the probability of a customer churning when a predictor variable increases by 1. A ratio of exactly 1 indicates no change in this probability, a ratio less than 1 represents a decreased probability, and a ratio greater than 1 represents an increased probability. The odds ratios for this model can be **interpreted** as such:

- The **Tenure odds ratio**, 0.90, indicates that for every increase of 1 month in the Tenure variable, the probability of a customer churning decreases by 10% (0.90 – 1 = -0.10). A customer who is with the service for 1 more month than another is predicted to be 10% less likely to leave the service than the other customer.
- The **DeviceProtection odds ratio**, 0.76, indicates that for every increase of 1 in the DeviceProtection variable, the probability of a customer churning decreases by 24% (0.76 – 1 = -0.24). Since this is a binary variable, this indicates that a customer who does have a

device protection add-on is predicted to be 24% less likely to leave the service than a customer who does not have a device protection add-on.

- The **TechSupport odds ratio**, 0.63, indicates that for every increase of 1 in the TechSupport variable, the probability of a customer churning decreases by 37% (0.63 – 1 = -0.37). Since this is a binary variable, this indicates that a customer who does have a technical support add-on is predicted to be 37% less likely to leave the service than a customer who does not have a technical support add-on.
- The **MonthlyCharge odds ratio**, 1.06, indicates that for every increase of $1 in the MonthlyCharge variable, the probability of a customer churning increases by 6% (1.06 – 1 = 0.06). A customer who has a monthly charge value of $1 more than another customer is predicted to be 6% more likely to leave the service than the other customer.
- The **InternetService_Fiber_Optic odds ratio**, 0.11, indicates that for every increase of 1 in the InternetService_Fiber_Optic variable, the probability of a customer churning decreases by 89% (0.11 – 1 = -0.89). Since this is a binary variable, this indicates that a customer who does have Fiber Optic for an internet service provider is predicted to be 89% less likely to leave the service than a customer who does not have Fiber Optic.
- The **Contract_One_Year odds ratio**, 0.04, indicates that for every increase of 1 in the Contract_One_Year variable, the probability of a customer churning decreases by 96% (0.04 – 1 = -0.96). Since this is a binary variable, this indicates that a customer who does have a one-year contract is predicted to be 96% less likely to leave the service than a customer who does not have a one-year contract.
- The **Contract_Two_Year odds ratio**, 0.04, indicates that for every increase of 1 in the Contract_Two_Year variable, the probability of a customer churning decreases by 96% (0.04 – 1 = -0.96). Since this is a binary variable, this indicates that a customer who does have a two-year contract is predicted to be 96% less likely to leave the service than a customer who does not have a two-year contract.
- These interpretations and the relevant formula used were obtained with assistance from the following resource: https://www.statology.org/interpret-logistic-regression-coefficients/.

A better interpretation of the y-intercept can be made using the probability of churn, as calculated using the formula $\frac{e^{\beta_0}}{1+e^{\beta_0}}$, where β_0 represents the intercept coefficient. The resulting value has been calculated and **interpreted** as follows:

- **Y-intercept**: probability of 0.002.
- This is the probability of a customer with values of 0 in all predictor variables leaving the service (churning). Though it may not be possible to have a value of 0 in each predictor variable, a hypothetical customer with a 0 in each relevant variable is predicted to churn at a 0.2% probability.
- This interpretation and the relevant formula used were obtained with assistance from the following resource: https://www.statology.org/interpret-logistic-regression-intercept/.

The significance of these results is summarized as follows:

- This final reduced model has been determined to be **statistically significant**.
    - All factors in the reduced model have extremely low p-values, much lower than the given alpha level of 0.01.
    - The confusion matrix accuracy, McFadden's R-squared, and residual deviance metrics all indicate statistical significance in the reduced model.
    - This means the probability that the model results were calculated by chance or luck is very low. There is a real, measurable relationship between the explanatory variables and the response variable.
- The model has been determined to be **practically significant** in understanding the relationship between various factors and whether a customer continues with the service.
    - These results can be used to analyze what makes customers more or less likely to stay with the service, which can inform further research and implementation of strategies focused on minimizing customer churn.
    - For example, since the device protection add-on program was found to have a statistically significant relationship with customer churn, efforts can be made to target customers without these add-ons and convince them to sign up for one, which may encourage them to stay with the service.

There are potential **limitations** to the processes and results of this multiple logistic regression:

- Only a subset of variables was selected in the initial data exploration and regression. These variables were selected due to their perceived practical connection to customer Churn; however, it is possible for one or more unselected variables in the initial dataset to have significant relationships with Churn. Their exclusion may affect the final results.
- All outliers found were determined to be reasonable within the context of the variables and were retained. These outliers may have been the result of data entry errors, mistranslated units, or other data quality issues which, if unchanged, could lead to incorrect results and assumptions made based on proceeding calculations.
- Though it was determined that several factors had meaningful relationships with Churn, it is important to remember that correlation does not imply causation. These connections could be the result of confounding variables or simply coincidences.

F2. Recommendations

Based on the results of the multiple logistic regression, strong relationships have been identified between the response variable, Churn, and 7 explanatory variables. These relationships should be explored further, as they may uncover ways to decrease customer churn rate.

- There is a negative relationship between Tenure and Churn. This indicates that customers who have been with the service for more time are less likely to leave the service. More incentives should be created to ensure newer customers continue with the service.

- There is a negative relationship between DeviceProtection and Churn. This indicates that the device protection add-on program is working well and should be advertised to more customers when they sign up to encourage them to stay longer.
- There is a negative relationship between TechSupport and Churn. This indicates that the technical support add-on program is working well and should be advertised to more customers when they sign up to encourage them to stay longer.
- There is a positive relationship between MonthlyCharge and Churn. This indicates that customers who are charged less are less likely to leave the service. Of course, lowering monthly charges would likely not be profitable for the organization, but this information may still be used to target specific customers who are at a higher risk of leaving the service.
- There is a negative relationship between InternetService_Fiber_Optic and Churn. This indicates that customers with Fiber Optic as an internet provider are less likely to leave the service. Surveys should be sent to Fiber Optic customers to determine what may be keeping them with the service. Any insight gained from collecting more information should be applied to customers with other internet providers to encourage them to stay.
- There is a negative relationship between Contract_One_Year and Churn. This indicates that customers with a one-year contract are less likely to leave the service. Surveys should be sent to these customers to determine what may be keeping them with the service. Any insight gained from collecting more information should be applied to customers with other contract lengths to encourage them to stay.
- There is a negative relationship between Contract_Two_Year and Churn. This indicates that customers with a two-year contract are less likely to leave the service. Surveys should be sent to these customers to determine what may be keeping them with the service. Any insight gained from collecting more information should be applied to customers with other contract lengths to encourage them to stay.

G. Panopto Video

Please refer to the link attached with a Panopto video recording. The video includes an explanation, execution, and output results of the referred code used to perform this analysis.

H. Third-Party Code References

WGU Courseware was used as a resource to learn the methods, concepts, and functions used to create the codes in this project, including DataCamp course tracks (datacamp.com), Dr. Keiona Middleton's D208 webinar videos, and the book *Data Science Using Python and R* by Chantal D. Larose and Daniel T. Larose.

The code used to calculate McFadden's R-squared was obtained from the following resource: https://www.statology.org/glm-r-squared/

This is referenced in-text in the code used to perform this analysis. This can be found as part of the attached file: Atwood_D208_Task2_Code.R.

I. Content References

The interpretations and the relevant formula used for the coefficient estimates and odds ratios of the predictor variables were obtained with assistance from the following resource:
https://www.statology.org/interpret-logistic-regression-coefficients/

The interpretation and the relevant formula used for the y-intercept of the logistic model were obtained with assistance from the following resource:
https://www.statology.org/interpret-logistic-regression-intercept/