Tucker Atwood
WGU MSDA
D208 Task 1
3/31/24

A1. Research Question

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

A2. Analysis Objectives

The goal of this analysis is to understand the relationship between various factors and the amount of time a customer stays with the service. These relationships may then be used to inform the areas of focus that will ensure customers stay longer with the service.

B1. Regression Assumptions

In order to perform this multiple linear 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 dependent variable.
- **Multicollinearity**: independent variables are not highly correlated with each other.
- **Residual Normality**: the residuals of the final reduced model follow a Normal distribution.
- **Homoscedasticity**: the residuals of the final reduced model have a constant variance.

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 linear regression. Within R, the following packages will be utilized: naniar, plyr, dplyr, ggplot2, fastDummies, and car. These packages were chosen to provide several functions that will make the predictive modeling process more effective.

B3. Regression Justification

Multiple linear regression will be used to answer this research question because the response variable to be studied, customer tenure, is a quantitative continuous variable, and the question aims to find explanatory variables that can be used to predict customer tenure.

C1. Data Cleaning

To ensure the results of the multiple linear 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_Task1_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 linear 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.

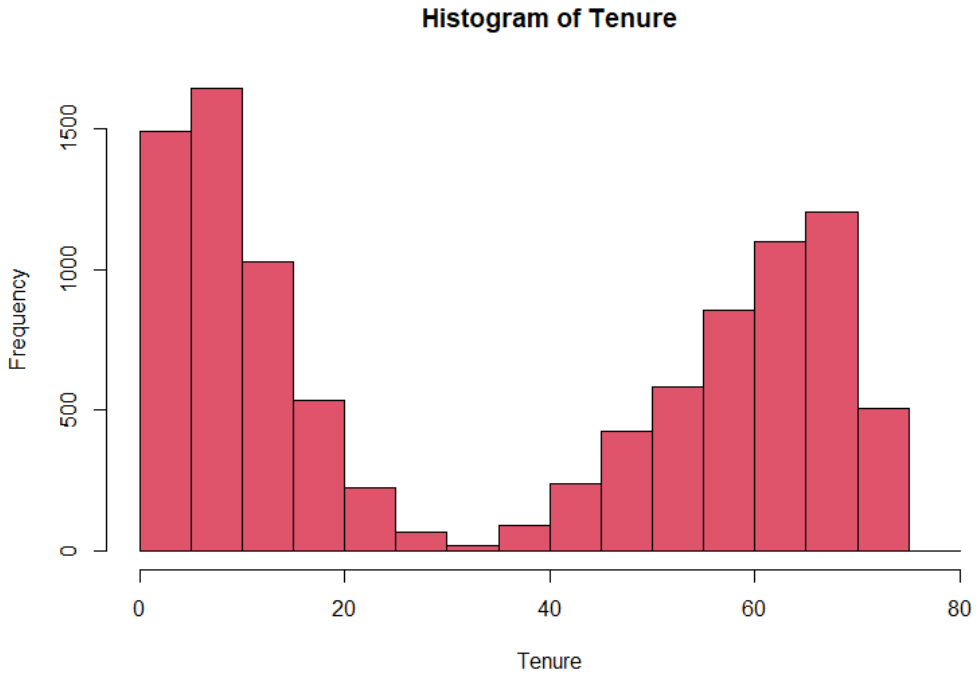
```
 DSL Fiber Optic        None
3463        4408        2129
```

  - 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.

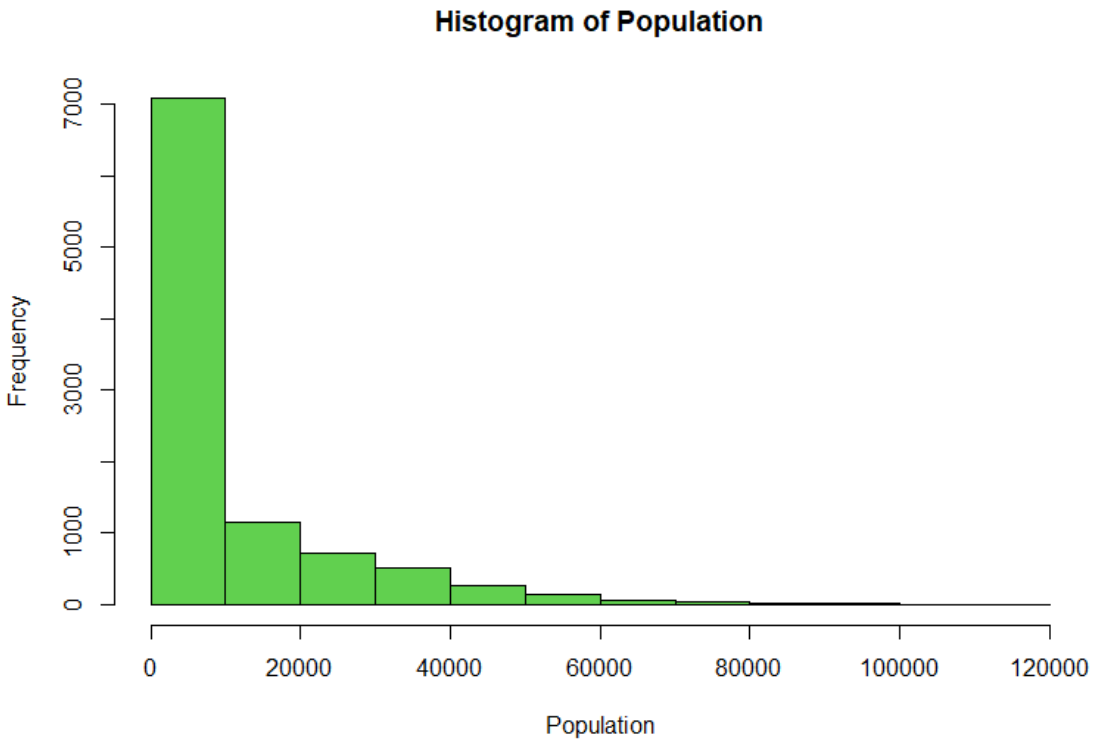
C3. Visualizations

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.

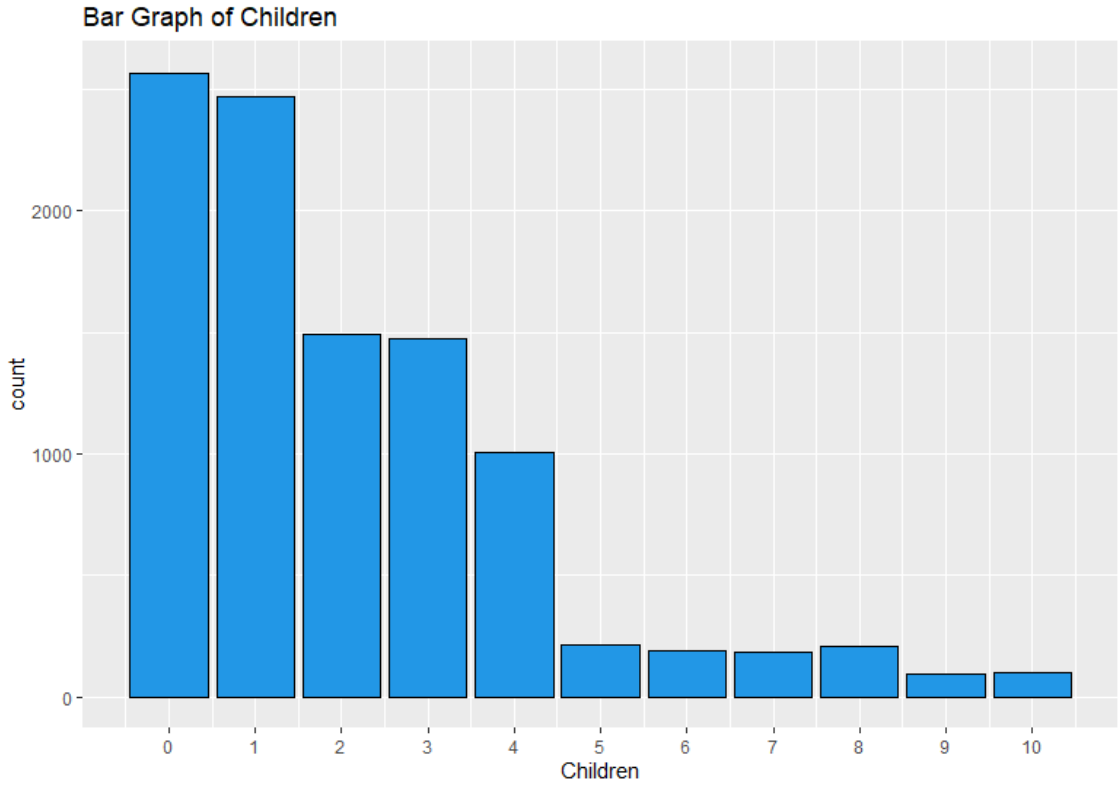
- **Tenure:**



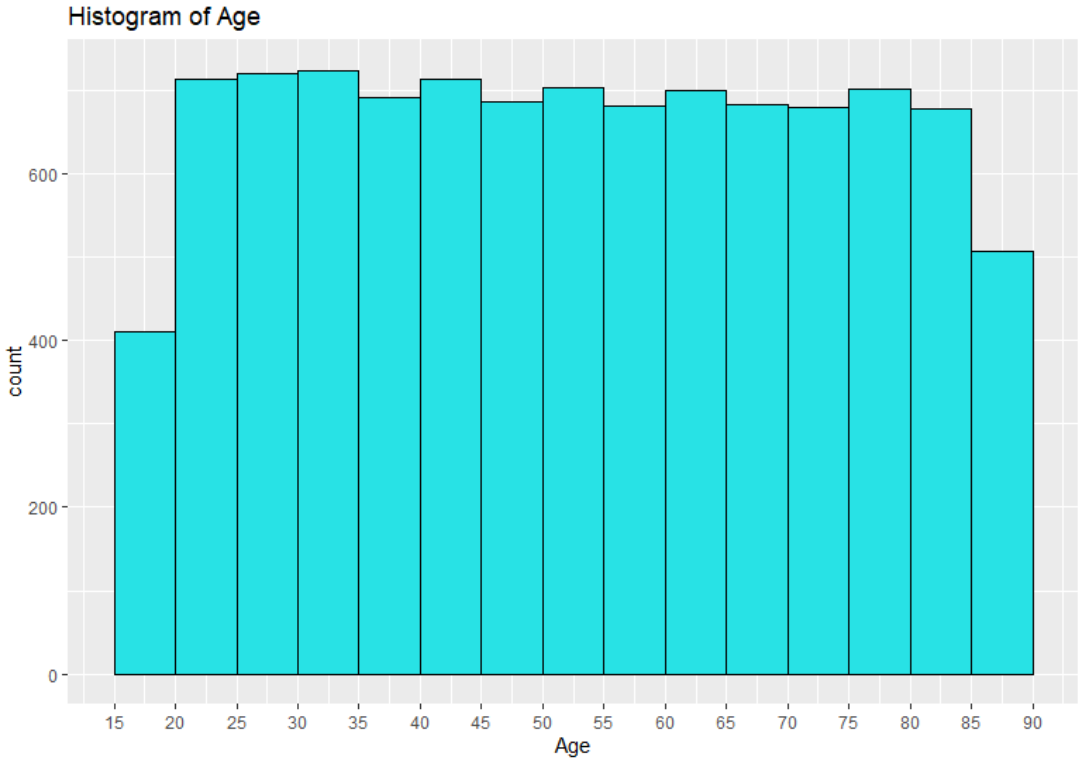
- **Population:**



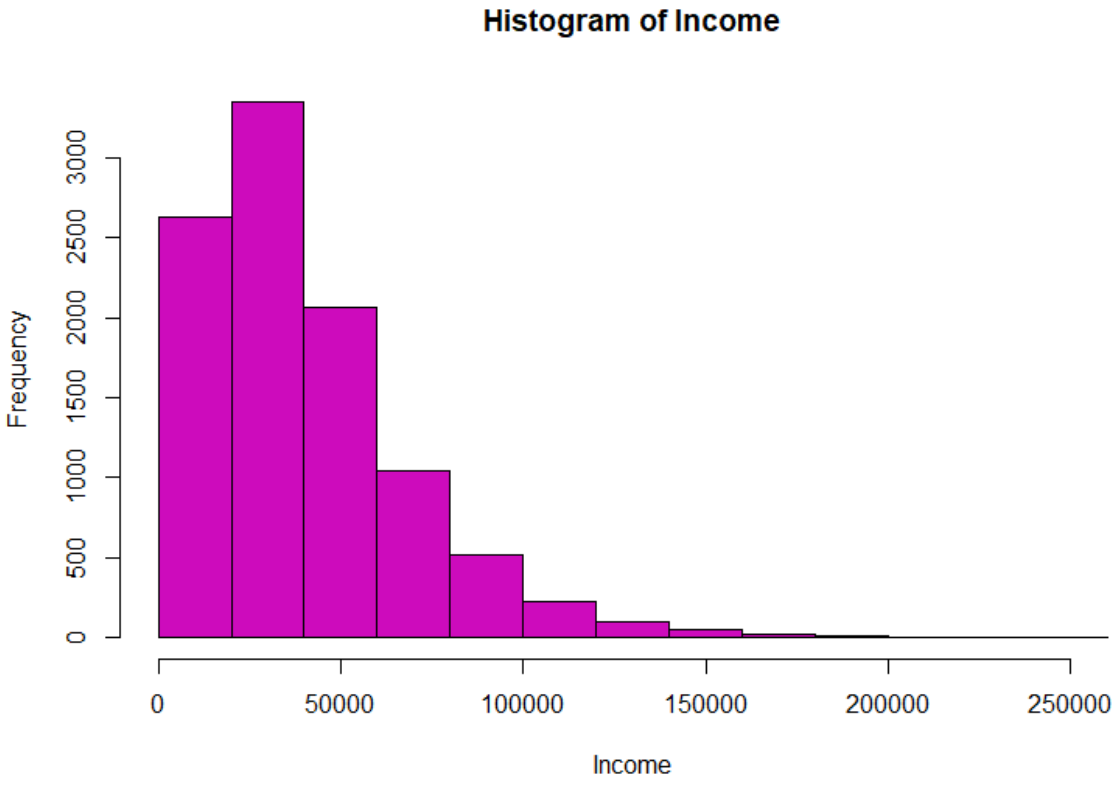
- **Children:**



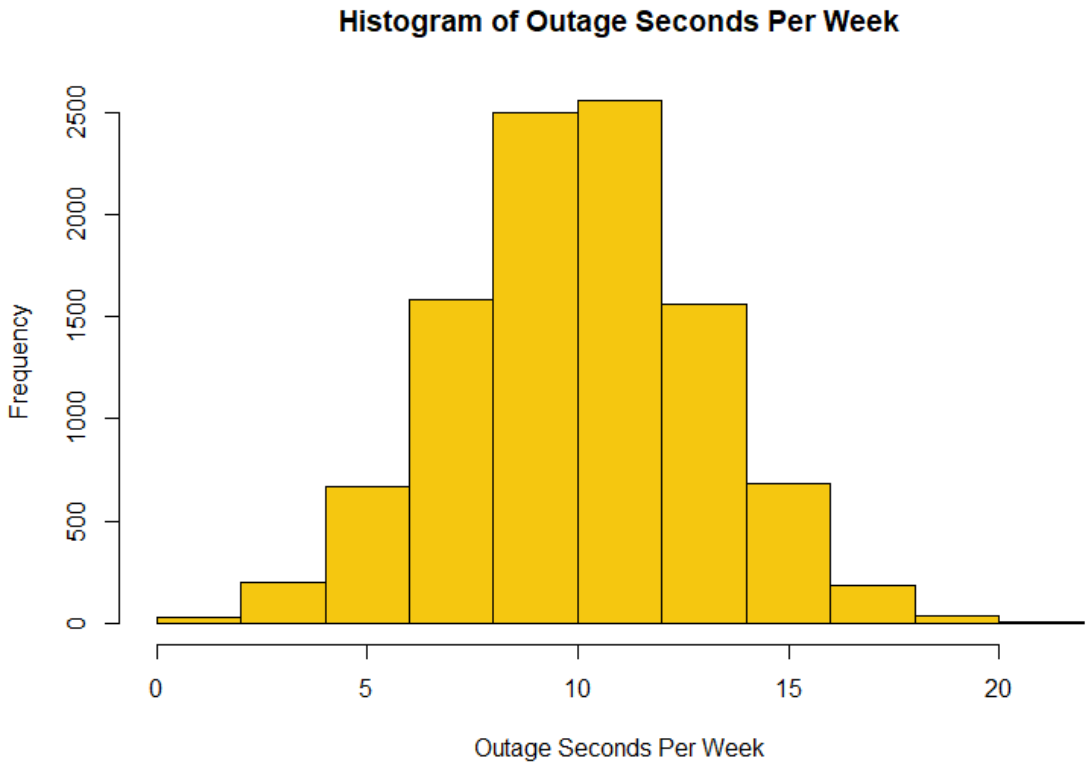
- **Age:**


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

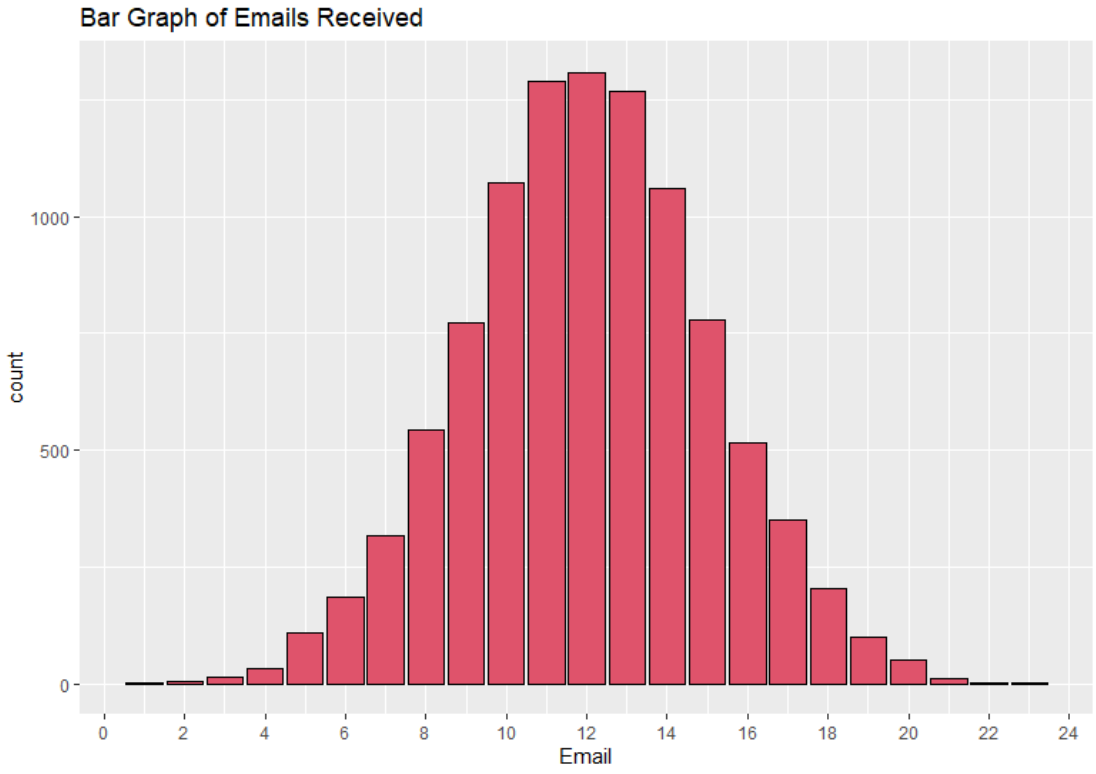
- **Income:**


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

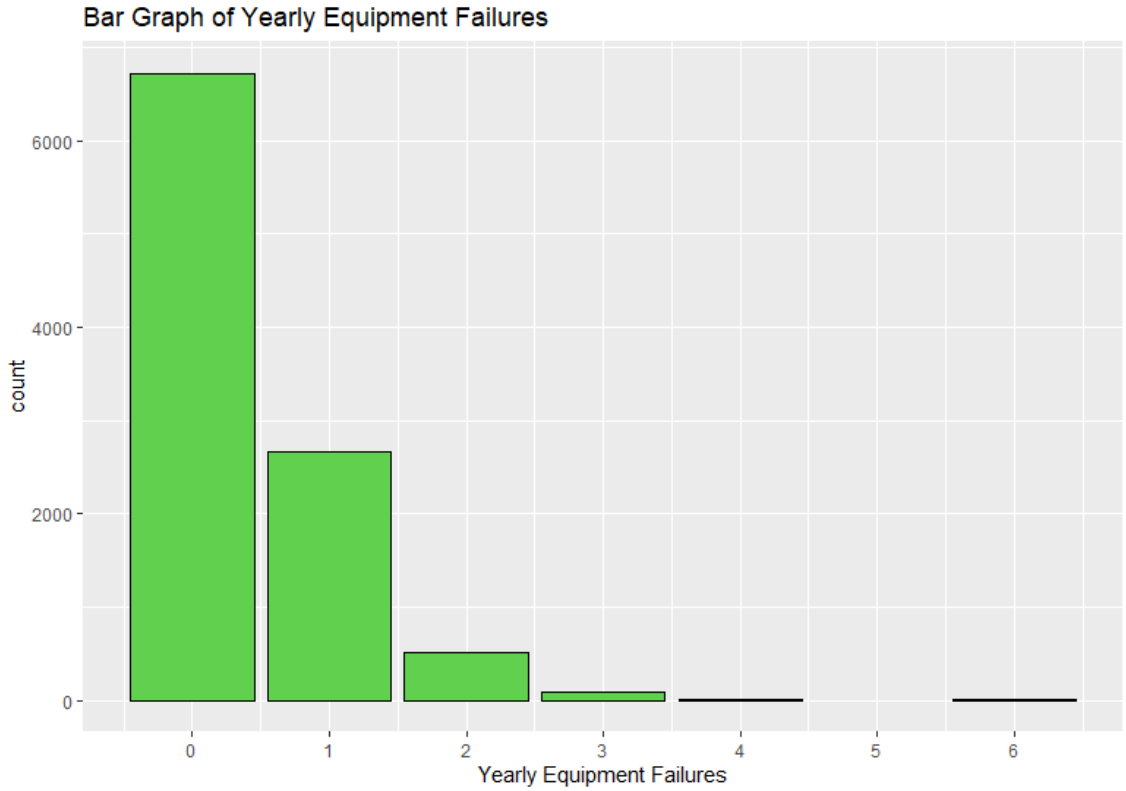
- **Outage_sec_perweek:**



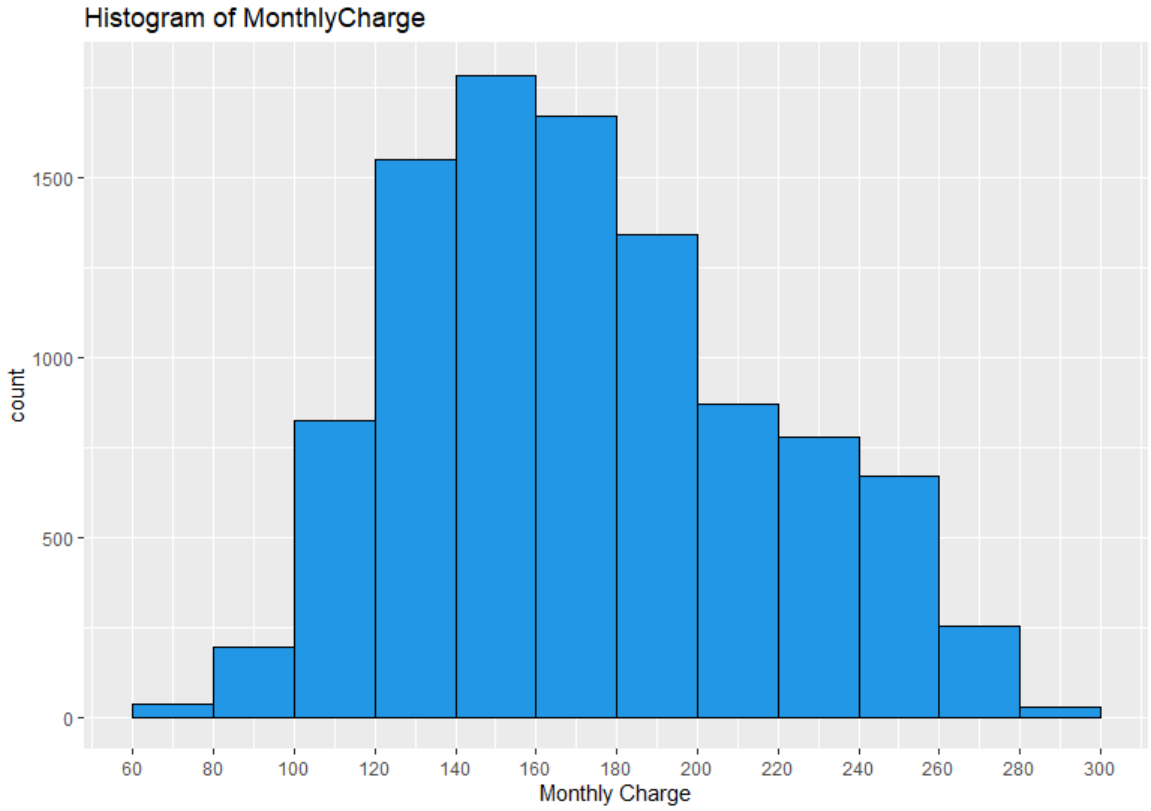
- **Email:**



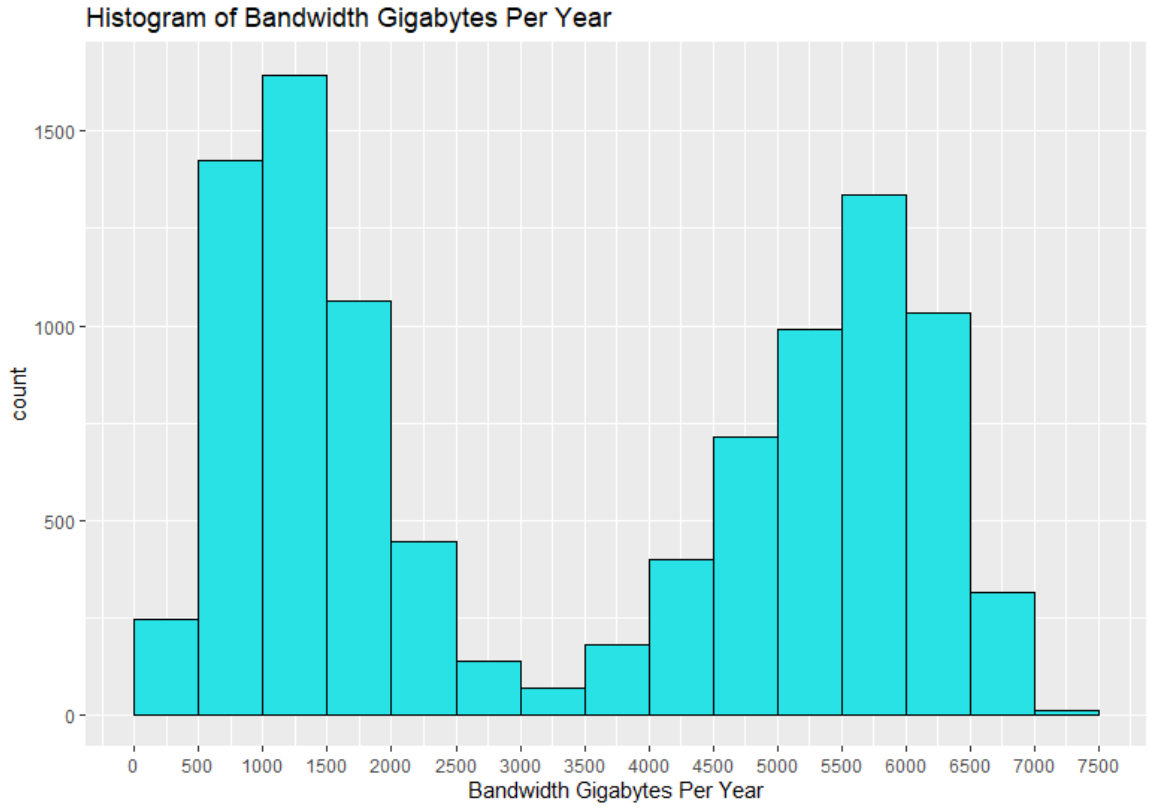
- **Yearly_equip_failure:**



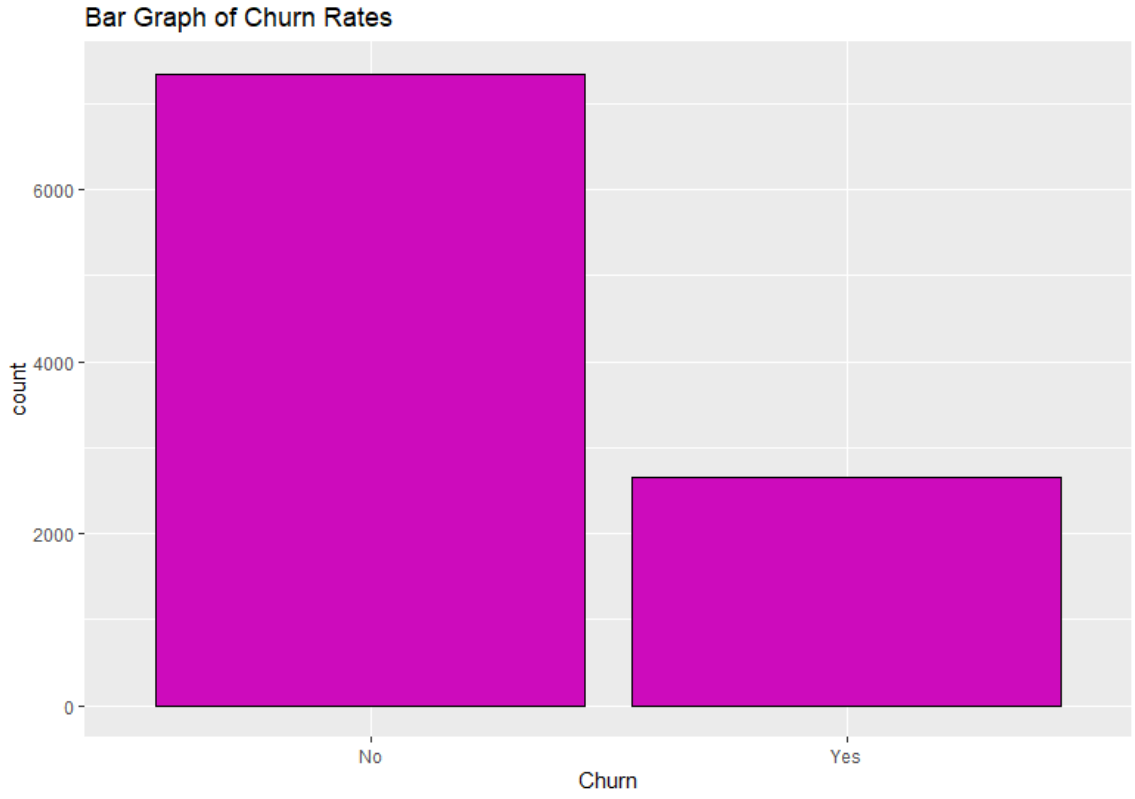
- **MonthlyCharge:**



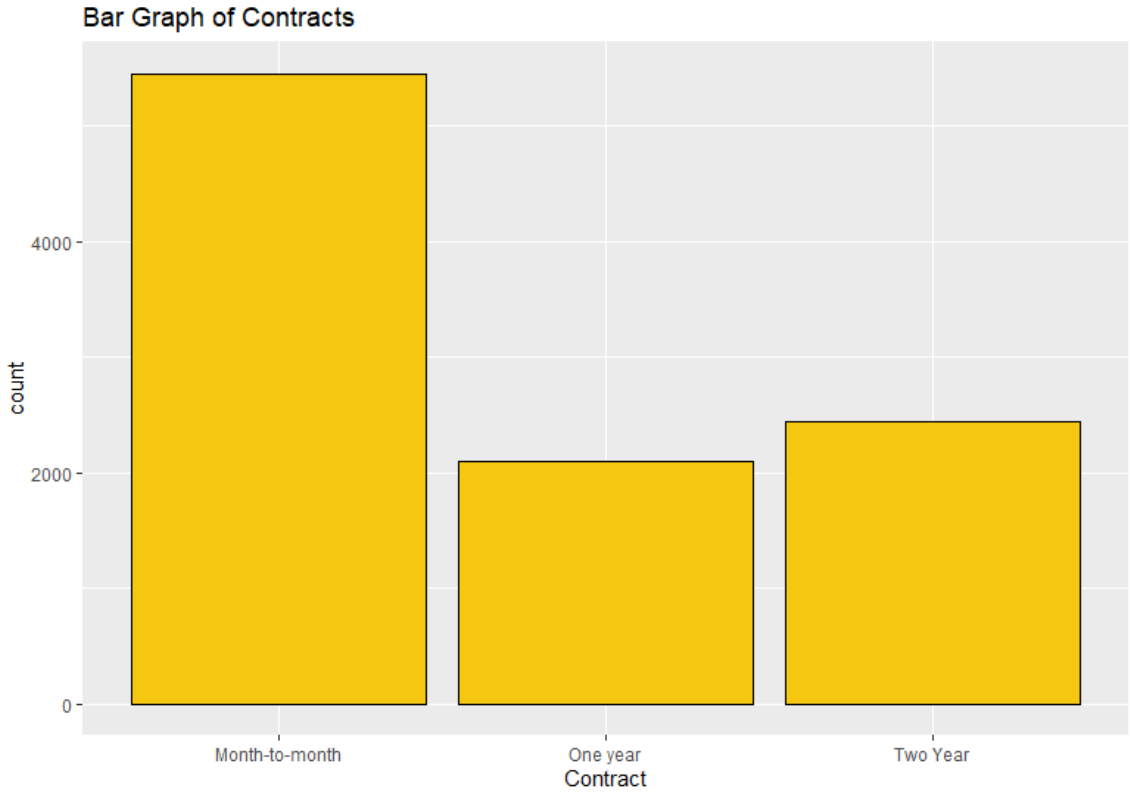
- **Bandwidth_GB_Year:**



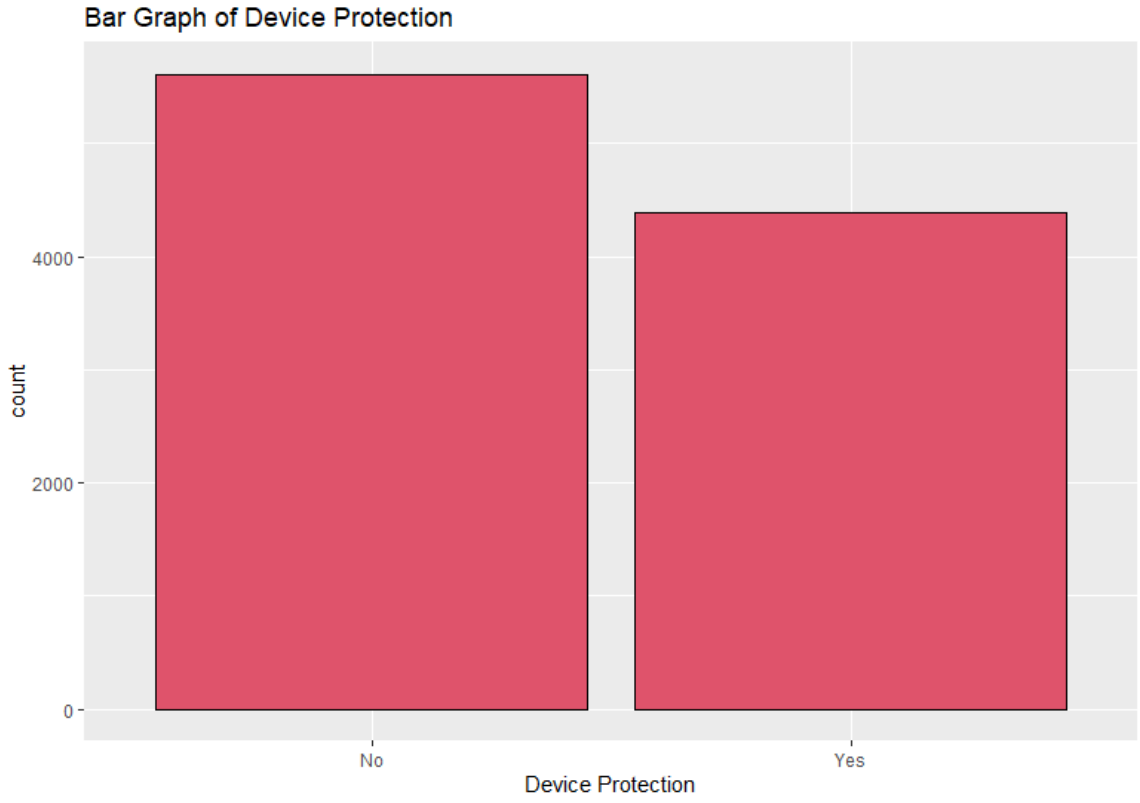
- **Churn:**



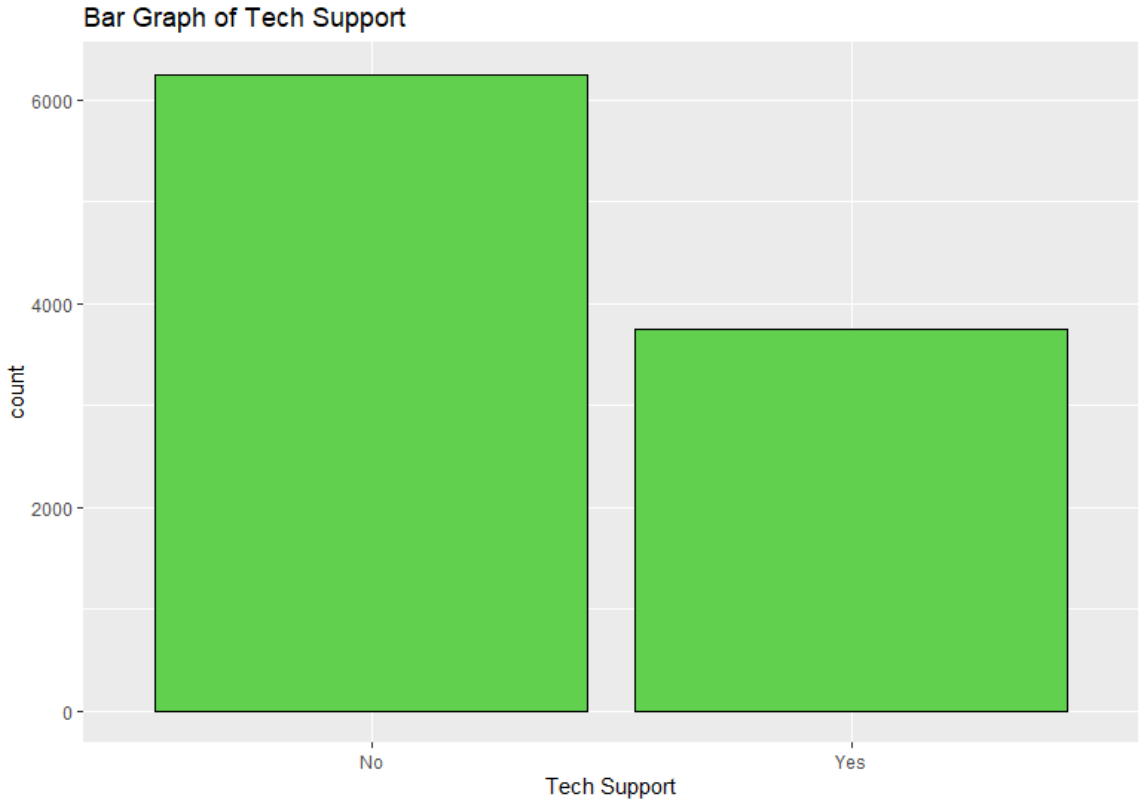
- **Contract:**



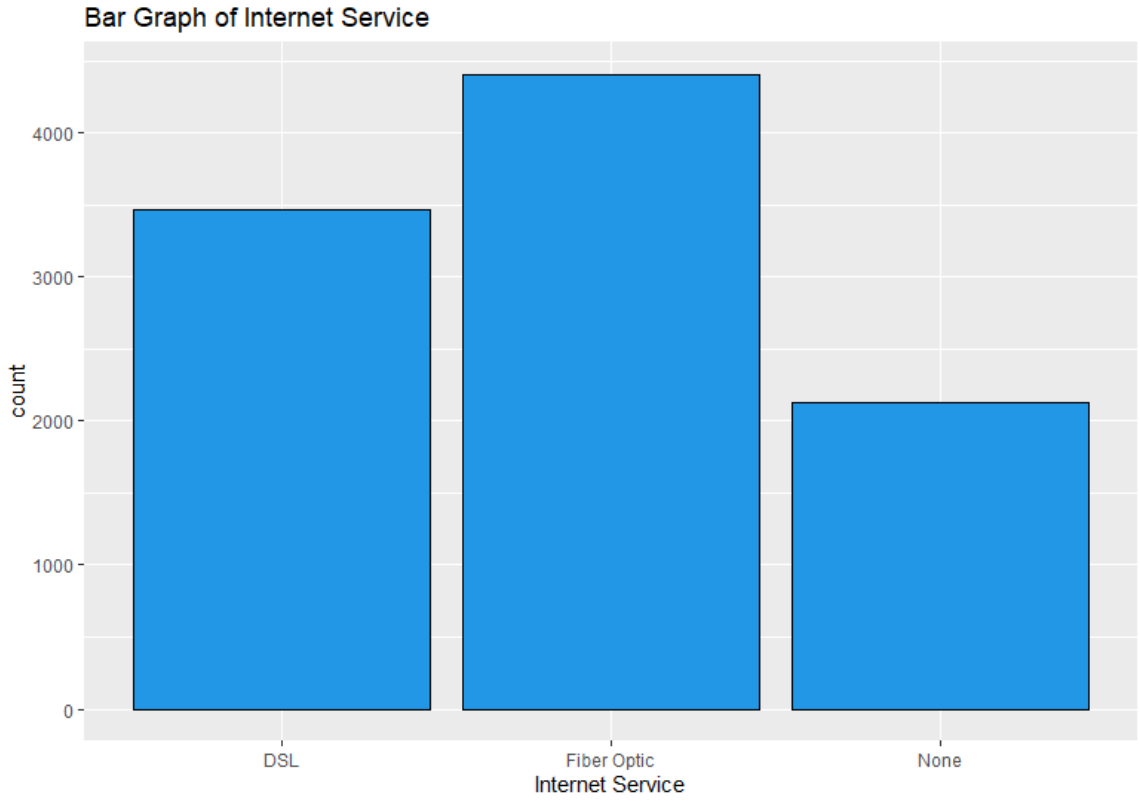
- **DeviceProtection:**



- **TechSupport:**

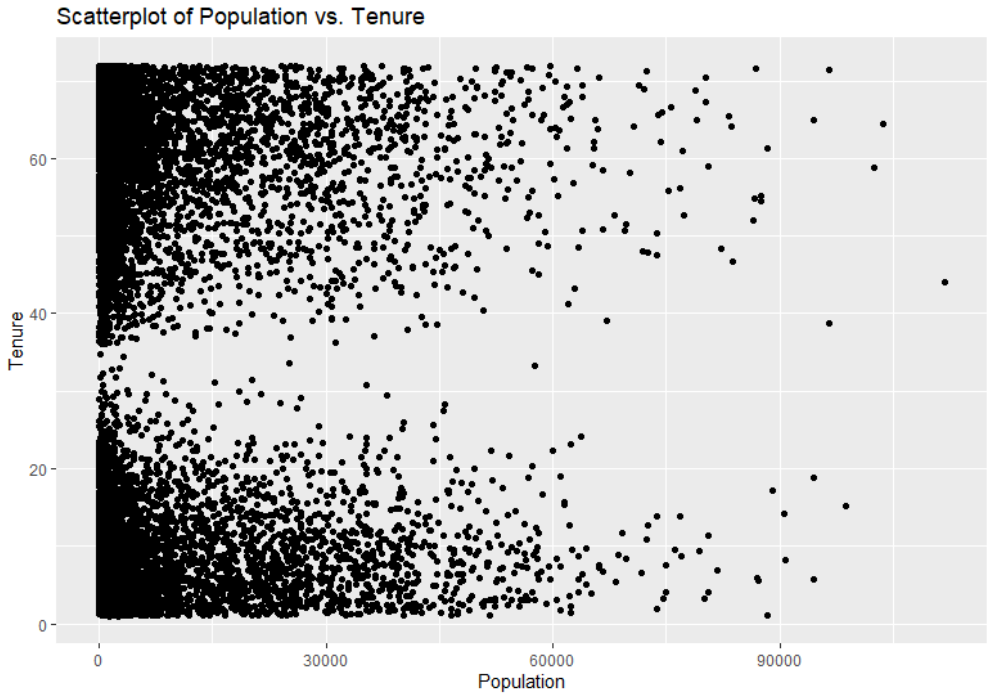


- **InternetService:**

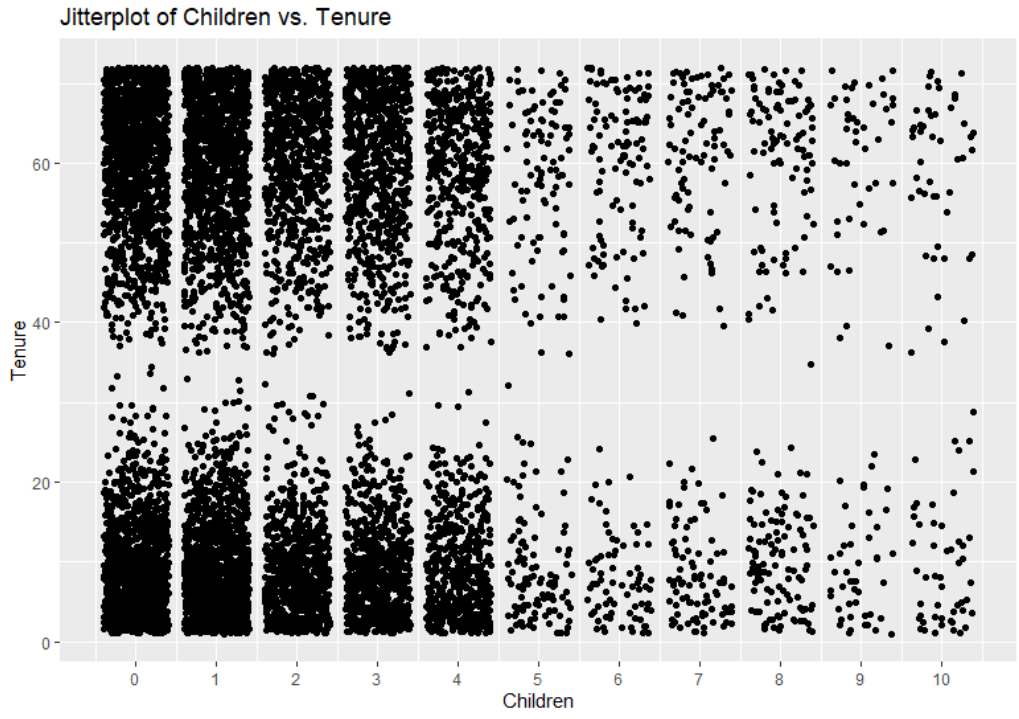


The following graphs provide further insight into the **bivariate distributions** of the relationships between each independent variable relevant to the research question and the dependent variable, Tenure.

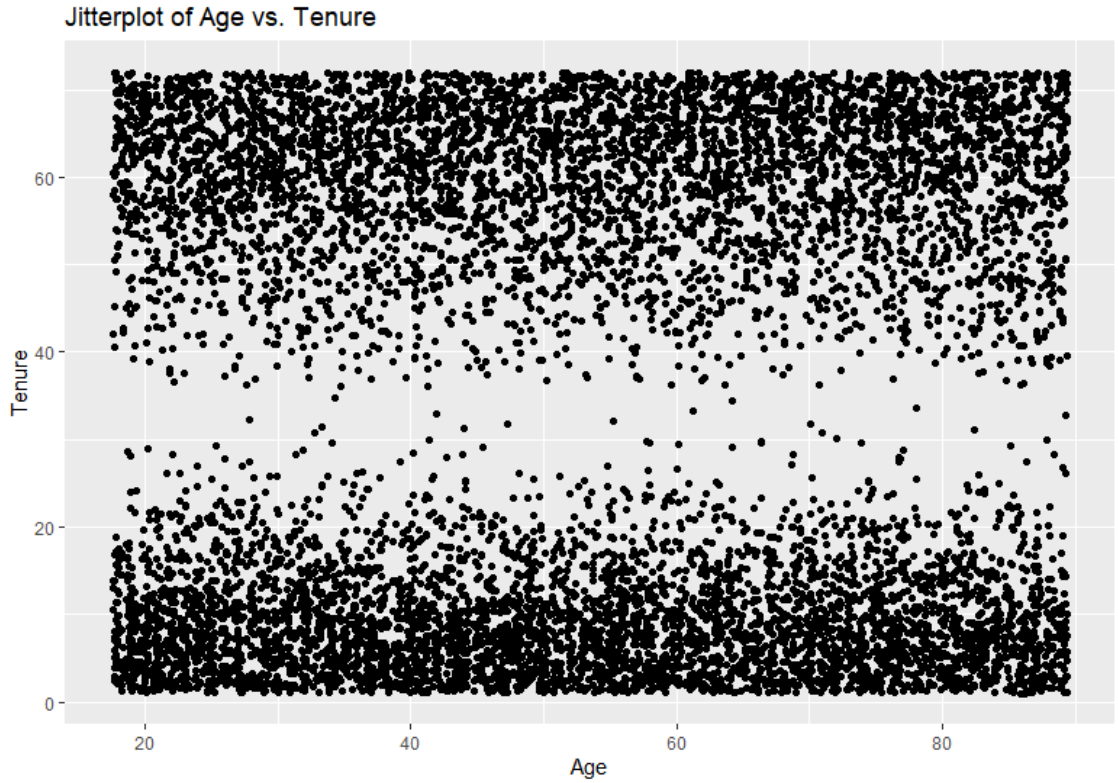
- **Population and Tenure:**



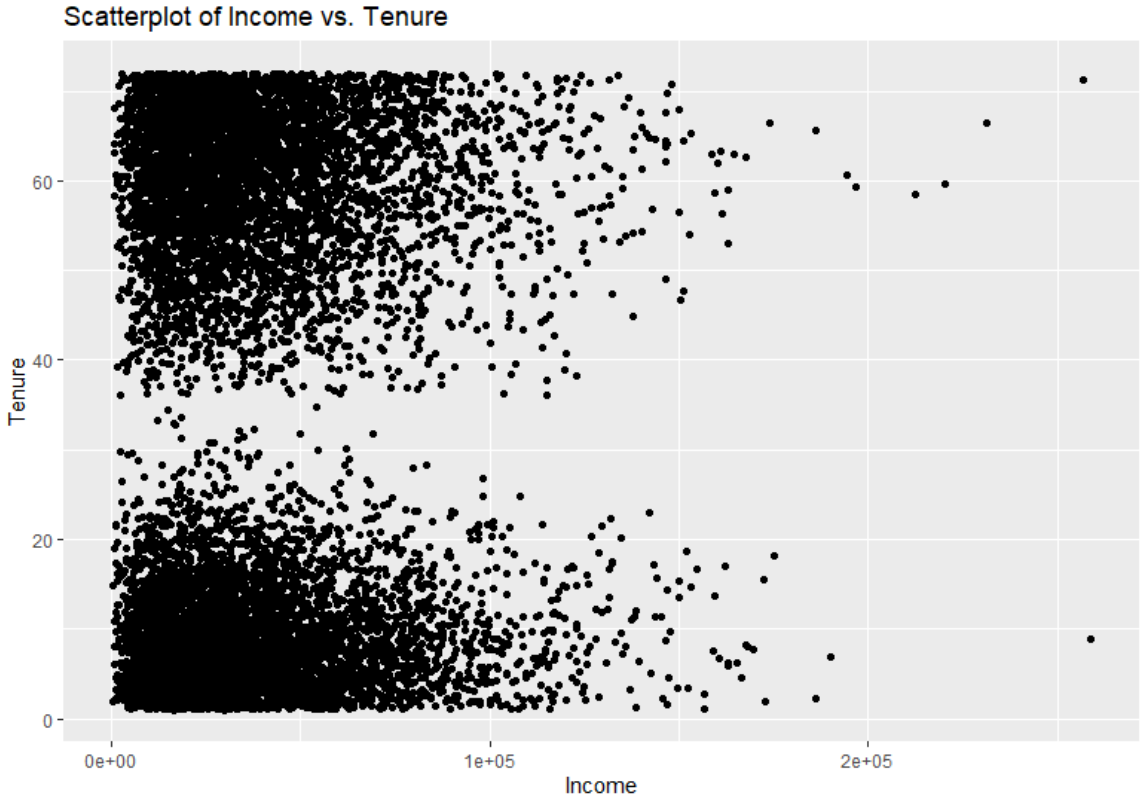
- **Children and Tenure:**



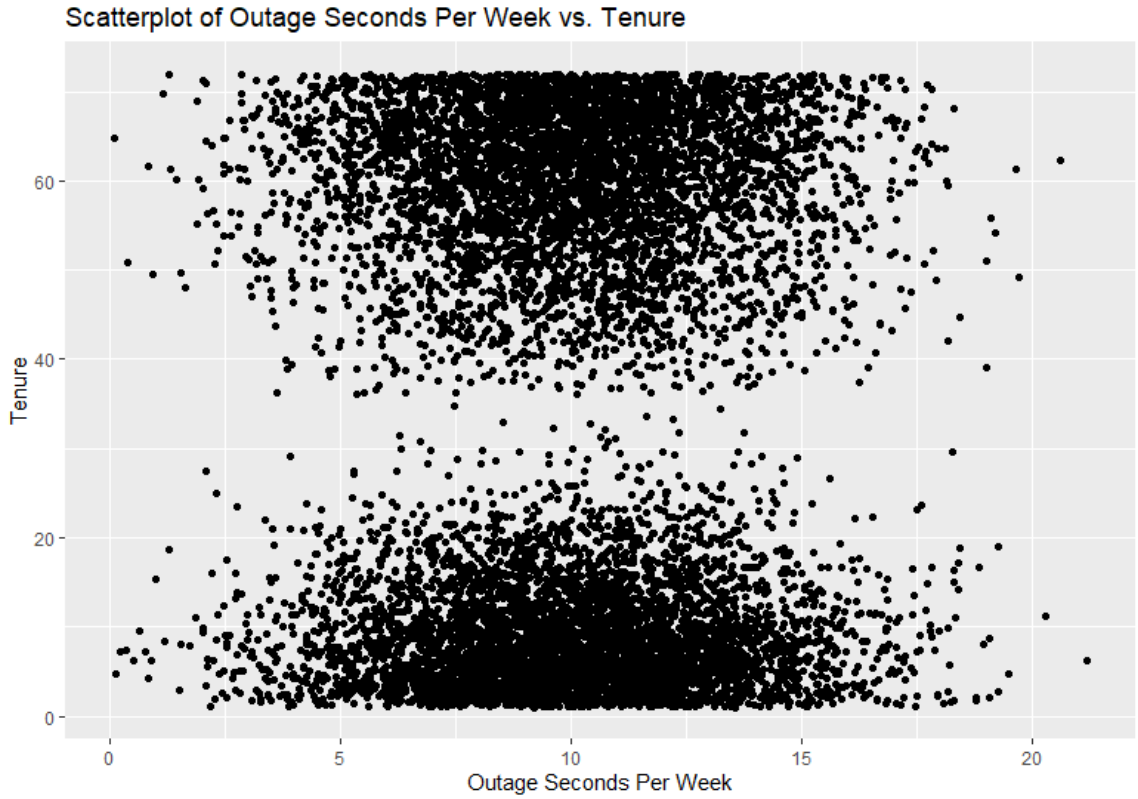
- **Age and Tenure:**



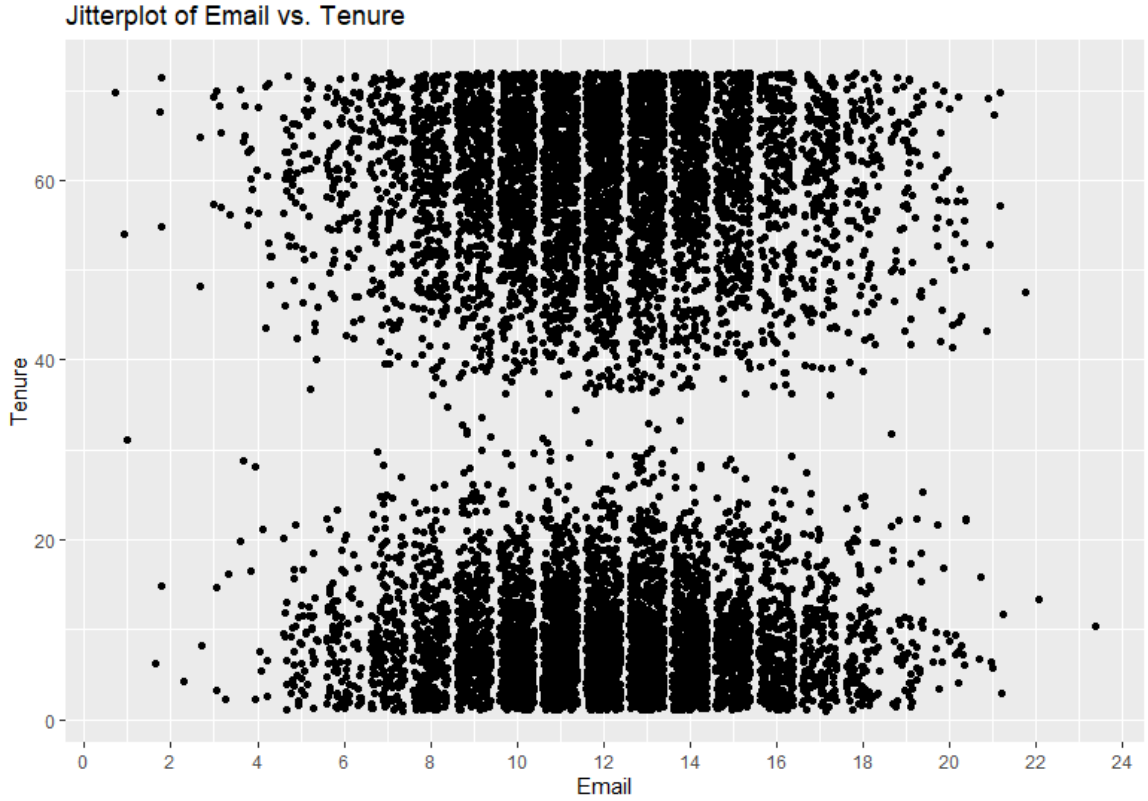
- **Income and Tenure:**



- **Outage_sec_perweek and Tenure:**

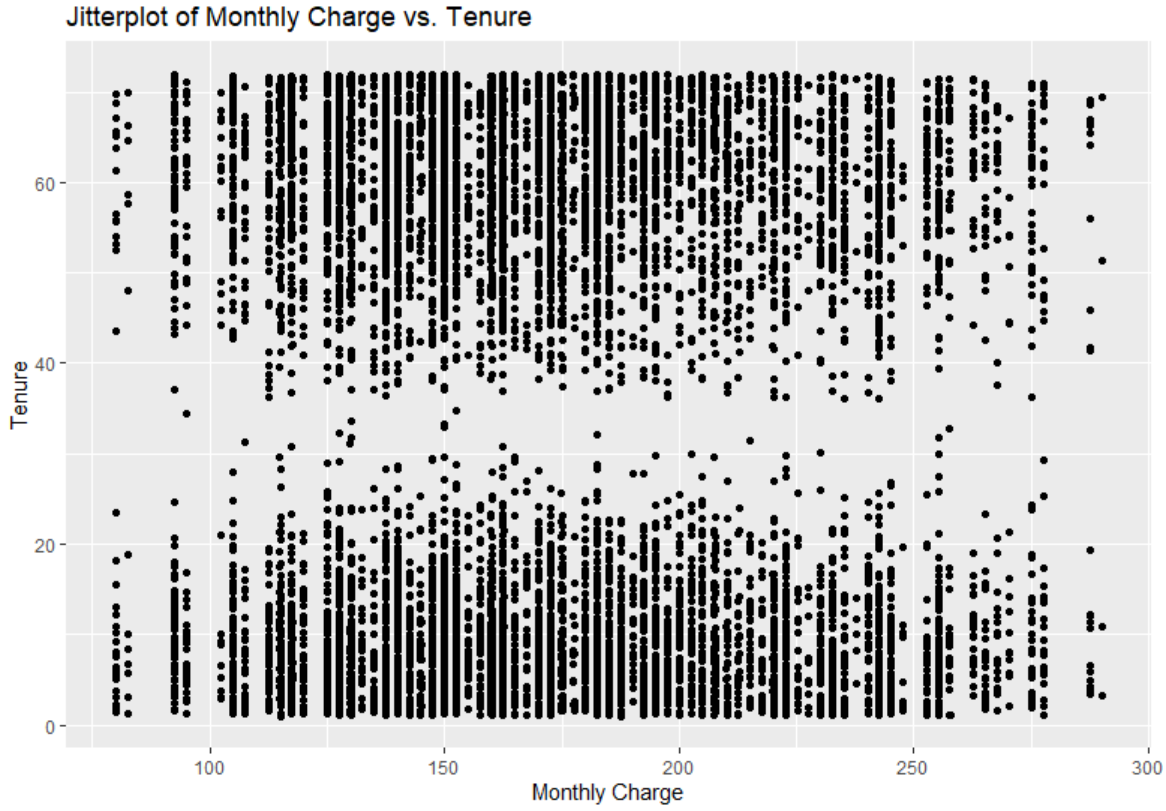


- **Email and Tenure:**

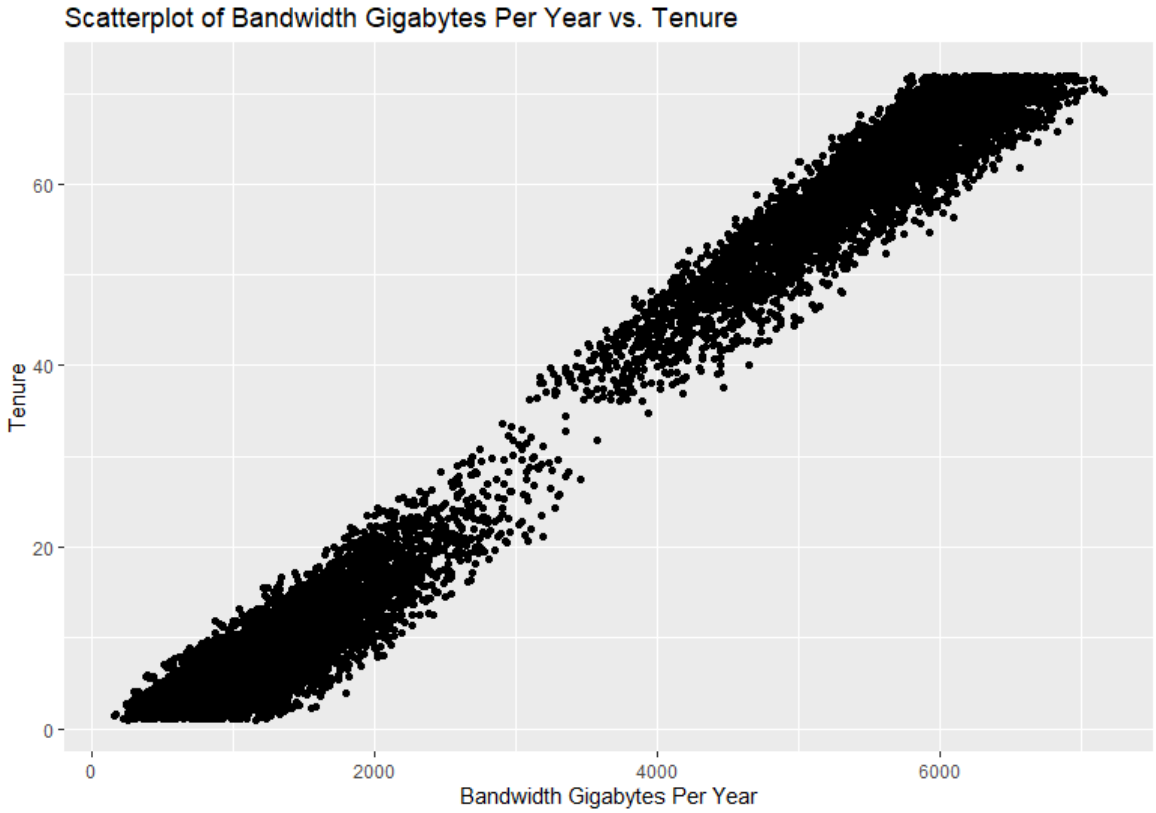


- **Yearly_equip_failure and Tenure:**

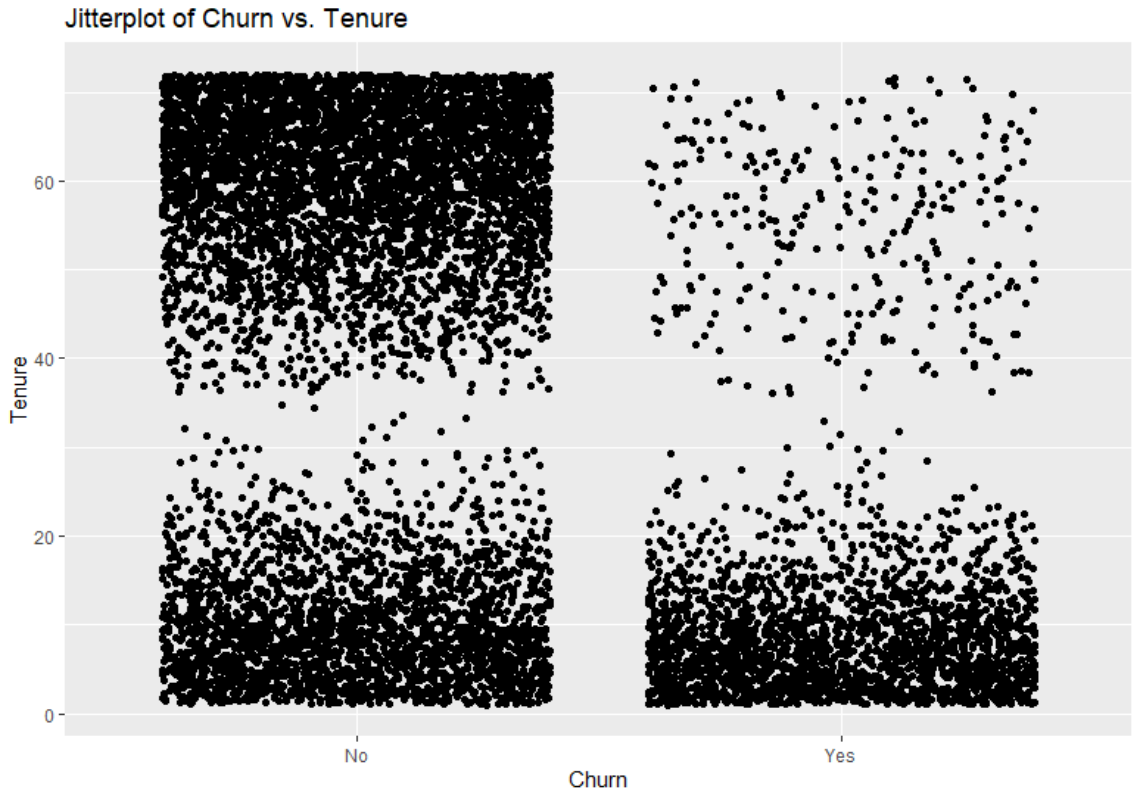
- **MonthlyCharge and Tenure:**



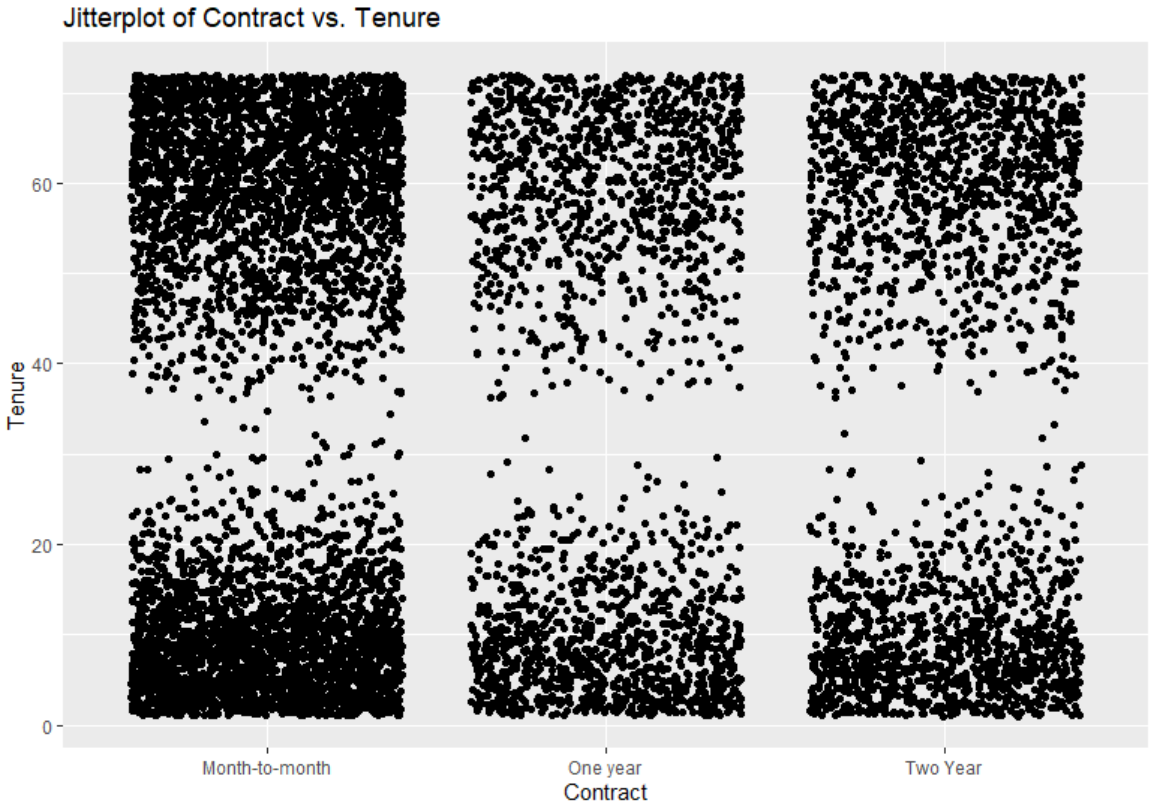
- **Bandwidth_GB_Year and Tenure:**



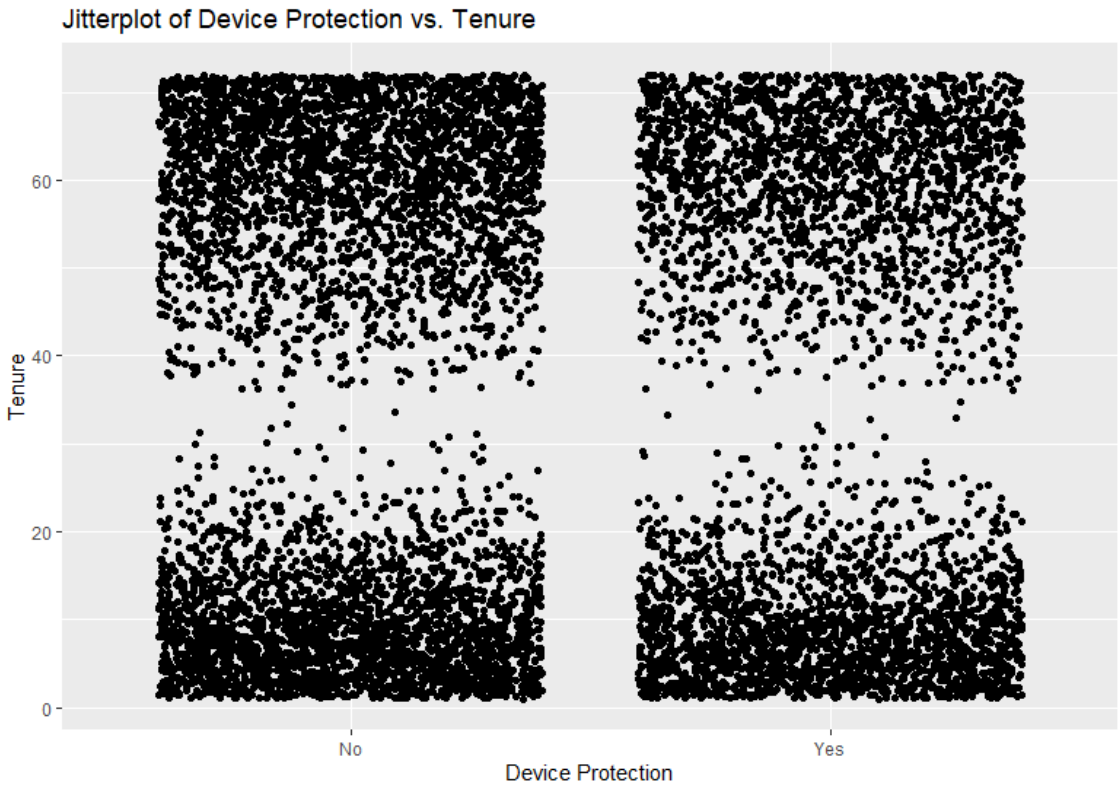
- **Churn and Tenure:**



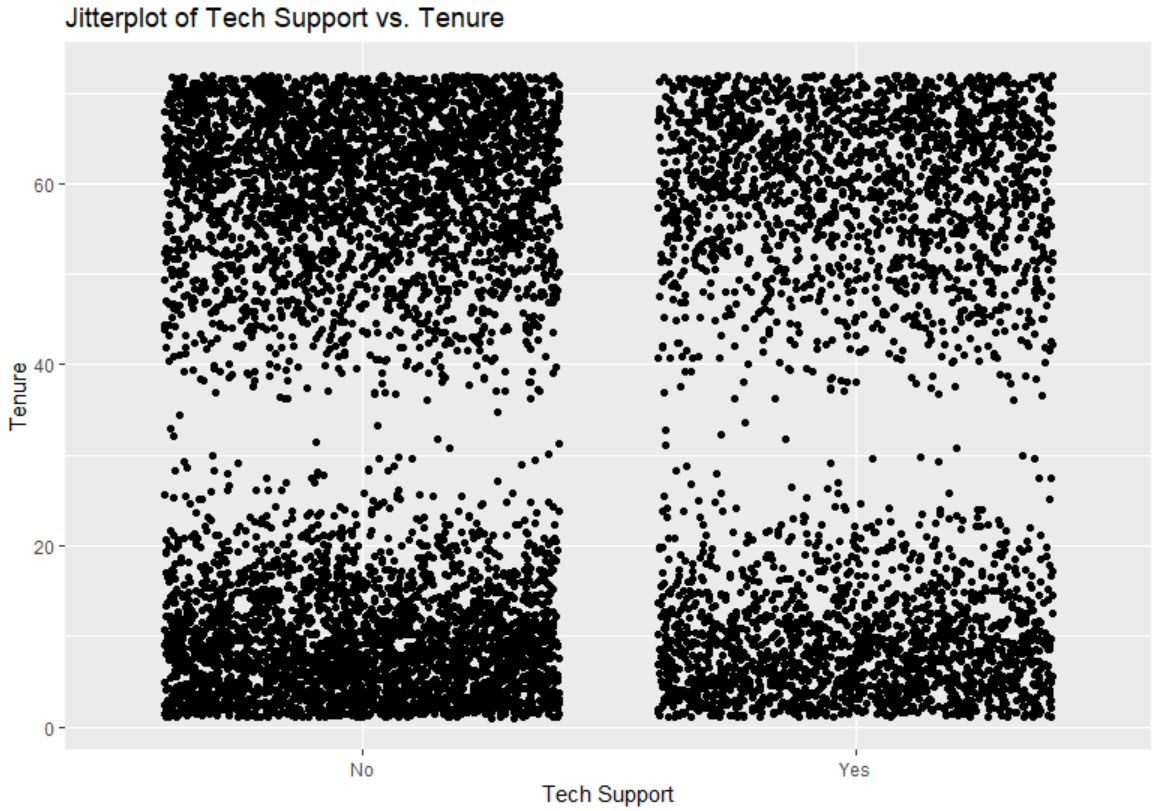
- **Contract and Tenure:**



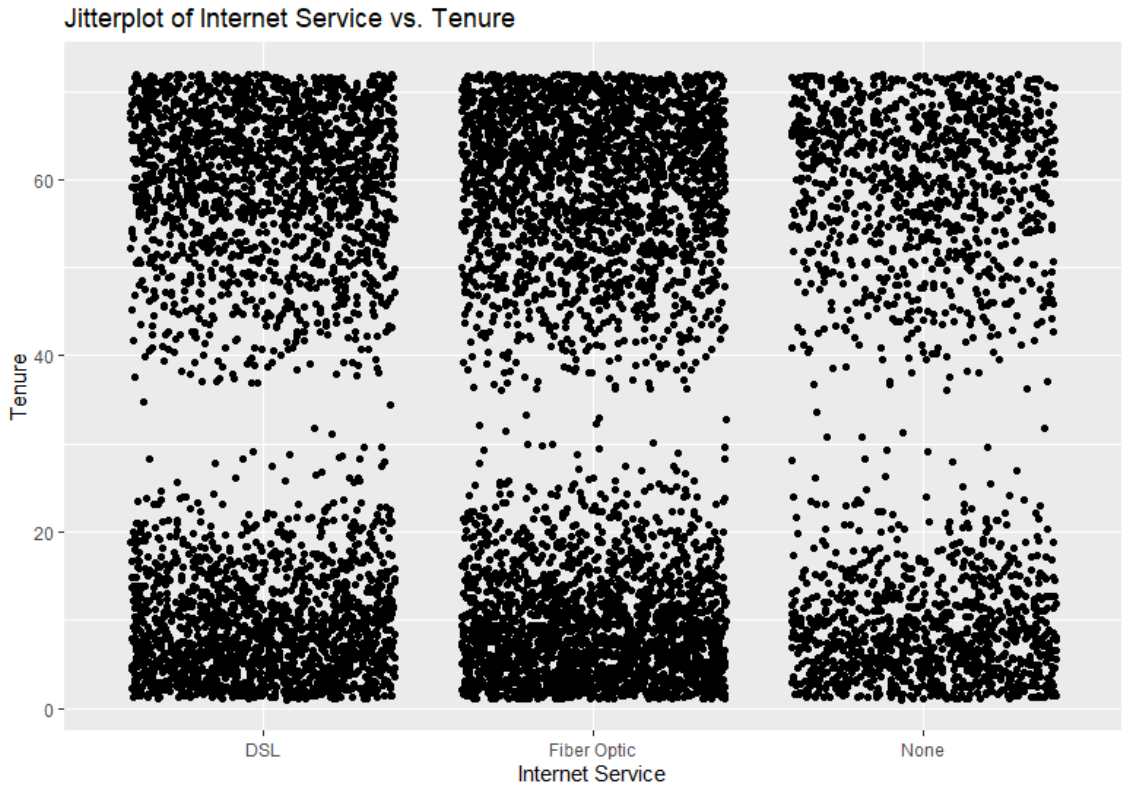
- **DeviceProtection and Tenure:**



- **TechSupport and Tenure:**



- **InternetService and Tenure:**



C4. Data Transformation

In order to perform the required multiple linear regression with inclusion of categorical 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_Task1_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 dataset has been written into a CSV file and attached with the following name: churn_new.csv.

D1. Initial Model

An initial linear regression model was constructed, with Tenure 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:
lm(formula = Tenure ~ Population + Children + Age + Income +
    Churn + Outage_sec_perweek + Email + Yearly_equip_failure +
    DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
    InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year +
    Contract_Two_Year, data = churn)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80425 -0.52481 -0.07797  0.59937  1.71806

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 1.199e+00  5.903e-02   20.310  < 2e-16 ***
Population                 -1.218e-07  5.362e-07   -0.227   0.8203
Children                   -3.794e-01  3.608e-03 -105.153  < 2e-16 ***
Age                         3.998e-02  3.741e-04  106.892  < 2e-16 ***
Income                      6.088e-07  2.744e-07    2.219   0.0265 *
Churn                      -1.208e-01  2.426e-02   -4.981 6.44e-07 ***
Outage_sec_perweek         -2.843e-04  2.601e-03   -0.109   0.9130
Email                       3.482e-04  2.558e-03    0.136   0.8917
Yearly_equip_failure        8.655e-03  1.217e-02    0.711   0.4769
DeviceProtection           -4.409e-01  1.584e-02  -27.837  < 2e-16 ***
TechSupport                 5.599e-01  1.617e-02   34.618  < 2e-16 ***
MonthlyCharge              -4.933e-02  2.277e-04 -216.683  < 2e-16 ***
Bandwidth_GB_Year           1.218e-02  4.292e-06 2838.166  < 2e-16 ***
InternetService_Fiber_Optic 1.623e+00  2.179e-02   74.489  < 2e-16 ***
InternetService_DSL        -4.396e+00  2.161e-02 -203.488  < 2e-16 ***
Contract_One_Year          -2.279e-02  2.068e-02   -1.102   0.2704
Contract_Two_Year          -4.530e-02  1.972e-02   -2.298   0.0216 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7734 on 9983 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 7.3e+05 on 16 and 9983 DF,  p-value: < 2.2e-16
```

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. Any results that were less than 1% likely to be calculated by 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_perweek** (p-value 0.9130):

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.80449 -0.52500 -0.07786  0.59913  1.71882 

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                1.196e+00  5.323e-02   22.467  < 2e-16 ***
Population                -1.221e-07  5.362e-07   -0.228   0.8199    
Children                  -3.794e-01  3.608e-03 -105.159  < 2e-16 ***
Age                        3.998e-02  3.740e-04  106.902  < 2e-16 ***
Income                     6.091e-07  2.744e-07    2.220   0.0264 *  
Churn                     -1.208e-01  2.426e-02   -4.980 6.45e-07 ***
Email                      3.470e-04  2.558e-03    0.136   0.8921    
Yearly_equip_failure       8.651e-03  1.217e-02    0.711   0.4771    
DeviceProtection          -4.410e-01  1.584e-02  -27.845  < 2e-16 ***
TechSupport                5.599e-01  1.617e-02   34.625  < 2e-16 ***
MonthlyCharge             -4.933e-02  2.276e-04 -216.739  < 2e-16 ***
Bandwidth_GB_Year          1.218e-02  4.292e-06 2838.306  < 2e-16 ***
InternetService_Fiber_Optic  1.623e+00  2.178e-02   74.499  < 2e-16 ***
InternetService_DSL       -4.396e+00  2.160e-02 -203.515  < 2e-16 ***
Contract_One_Year         -2.278e-02  2.068e-02   -1.102   0.2705    
Contract_Two_Year         -4.533e-02  1.971e-02   -2.299   0.0215 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7733 on 9984 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991 
F-statistic: 7.787e+05 on 15 and 9984 DF,  p-value: < 2.2e-16
```

- **Removal of Email** (p-value 0.8921):

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.80574 -0.52420 -0.07858  0.59896  1.71847 

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                1.200e+00  4.339e-02   27.663  < 2e-16 ***
Population                -1.208e-07  5.361e-07   -0.225   0.8217    
Children                  -3.794e-01  3.607e-03 -105.165  < 2e-16 ***
Age                        3.998e-02  3.740e-04  106.908  < 2e-16 ***
Income                     6.088e-07  2.744e-07    2.219   0.0265 *  
Churn                     -1.208e-01  2.426e-02   -4.980 6.47e-07 ***
Yearly_equip_failure       8.624e-03  1.216e-02    0.709   0.4784    
DeviceProtection          -4.410e-01  1.584e-02  -27.846  < 2e-16 ***
TechSupport                5.600e-01  1.617e-02   34.639  < 2e-16 ***
MonthlyCharge             -4.933e-02  2.276e-04 -216.751  < 2e-16 ***
Bandwidth_GB_Year          1.218e-02  4.291e-06 2838.554  < 2e-16 ***
InternetService_Fiber_Optic  1.623e+00  2.178e-02   74.503  < 2e-16 ***
InternetService_DSL       -4.396e+00  2.160e-02 -203.525  < 2e-16 ***
Contract_One_Year         -2.281e-02  2.067e-02   -1.104   0.2698    
Contract_Two_Year         -4.532e-02  1.971e-02   -2.299   0.0215 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7733 on 9985 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991 
F-statistic: 8.344e+05 on 14 and 9985 DF,  p-value: < 2.2e-16
```

- **Removal of Population** (p-value 0.8217):

```
Call:
lm(formula = Tenure ~ Children + Age + Income + Churn + Yearly_equip_failure +
    DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
    InternetService_Fiber_Optic + InternetService_DSL + Contract_One_Year +
    Contract_Two_Year, data = churn)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8047 -0.5243 -0.0779  0.5992  1.7189

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 1.199e+00  4.307e-02   27.842  < 2e-16 ***
Children                   -3.794e-01  3.607e-03 -105.170  < 2e-16 ***
Age                         3.998e-02  3.740e-04  106.916  < 2e-16 ***
Income                      6.093e-07  2.743e-07    2.221   0.0264 *
Churn                      -1.207e-01  2.426e-02   -4.978 6.52e-07 ***
Yearly_equip_failure        8.636e-03  1.216e-02    0.710   0.4777
DeviceProtection           -4.409e-01  1.583e-02  -27.847  < 2e-16 ***
TechSupport                 5.600e-01  1.616e-02   34.643  < 2e-16 ***
MonthlyCharge              -4.933e-02  2.276e-04 -216.762  < 2e-16 ***
Bandwidth_GB_Year           1.218e-02  4.291e-06 2838.801  < 2e-16 ***
InternetService_Fiber_Optic 1.623e+00  2.178e-02   74.506  < 2e-16 ***
InternetService_DSL        -4.396e+00  2.160e-02 -203.543  < 2e-16 ***
Contract_One_Year          -2.278e-02  2.067e-02   -1.102   0.2706
Contract_Two_Year          -4.538e-02  1.971e-02   -2.302   0.0213 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7733 on 9986 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 8.987e+05 on 13 and 9986 DF,  p-value: < 2.2e-16
```

- **Removal of Yearly_equip_failure** (p-value 0.4777):

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.79622 -0.52428 -0.07845  0.60006  1.71551

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 1.202e+00  4.282e-02   28.076  < 2e-16 ***
Children                   -3.794e-01  3.607e-03 -105.170  < 2e-16 ***
Age                         3.999e-02  3.739e-04  106.930  < 2e-16 ***
Income                      6.103e-07  2.743e-07    2.225   0.0261 *
Churn                      -1.209e-01  2.425e-02   -4.986 6.28e-07 ***
DeviceProtection           -4.410e-01  1.583e-02  -27.851  < 2e-16 ***
TechSupport                 5.600e-01  1.616e-02   34.647  < 2e-16 ***
MonthlyCharge              -4.933e-02  2.276e-04 -216.768  < 2e-16 ***
Bandwidth_GB_Year           1.218e-02  4.291e-06 2838.903  < 2e-16 ***
InternetService_Fiber_Optic 1.623e+00  2.178e-02   74.508  < 2e-16 ***
InternetService_DSL        -4.396e+00  2.160e-02 -203.547  < 2e-16 ***
Contract_One_Year          -2.267e-02  2.067e-02   -1.097   0.2727
Contract_Two_Year          -4.548e-02  1.971e-02   -2.307   0.0211 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7732 on 9987 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 9.736e+05 on 12 and 9987 DF,  p-value: < 2.2e-16
```

- **Removal of Contract_One_Year** (p-value 0.2727)

```
Call:
lm(formula = Tenure ~ Children + Age + Income + Churn + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    InternetService_DSL + Contract_Two_Year, data = churn)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79533 -0.52501 -0.08001  0.60181  1.71369

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 1.197e+00  4.259e-02   28.116  < 2e-16 ***
Children                   -3.794e-01  3.606e-03 -105.225  < 2e-16 ***
Age                         3.999e-02  3.739e-04  106.944  < 2e-16 ***
Income                      6.096e-07  2.743e-07    2.222   0.0263 *
Churn                      -1.135e-01  2.330e-02   -4.873 1.12e-06 ***
DeviceProtection           -4.409e-01  1.583e-02  -27.845  < 2e-16 ***
TechSupport                 5.604e-01  1.616e-02   34.678  < 2e-16 ***
MonthlyCharge              -4.937e-02  2.247e-04 -219.750  < 2e-16 ***
Bandwidth_GB_Year           1.218e-02  4.240e-06 2873.284  < 2e-16 ***
InternetService_Fiber_Optic 1.624e+00  2.173e-02   74.746  < 2e-16 ***
InternetService_DSL        -4.397e+00  2.160e-02 -203.581  < 2e-16 ***
Contract_Two_Year          -3.789e-02  1.846e-02   -2.053   0.0401 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7733 on 9988 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 1.062e+06 on 11 and 9988 DF,  p-value: < 2.2e-16
```

- **Removal of Contract_Two_Year** (p-value 0.0401)

```
Call:
lm(formula = Tenure ~ Children + Age + Income + Churn + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    InternetService_DSL, data = churn)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7828 -0.5272 -0.0822  0.6035  1.7018

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 1.190e+00  4.246e-02   28.037  < 2e-16 ***
Children                   -3.796e-01  3.606e-03 -105.275  < 2e-16 ***
Age                         4.000e-02  3.740e-04  106.947  < 2e-16 ***
Income                      6.110e-07  2.744e-07    2.227    0.026 *
Churn                      -1.031e-01  2.274e-02   -4.532 5.91e-06 ***
DeviceProtection           -4.405e-01  1.584e-02  -27.818  < 2e-16 ***
TechSupport                 5.609e-01  1.616e-02   34.710  < 2e-16 ***
MonthlyCharge              -4.942e-02  2.234e-04 -221.250  < 2e-16 ***
Bandwidth_GB_Year           1.218e-02  4.218e-06 2888.076  < 2e-16 ***
InternetService_Fiber_Optic 1.626e+00  2.173e-02   74.809  < 2e-16 ***
InternetService_DSL        -4.398e+00  2.159e-02 -203.663  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7734 on 9989 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 1.168e+06 on 10 and 9989 DF,  p-value: < 2.2e-16
```

- **Removal of Income** (p-value 0.026):

```
Call:
lm(formula = Tenure ~ Children + Age + Churn + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    InternetService_DSL, data = churn)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76177 -0.52988 -0.08519  0.60411  1.67983

Coefficients:
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 1.215e+00  4.106e-02   29.582  < 2e-16 ***
Children                   -3.795e-01  3.606e-03 -105.237  < 2e-16 ***
Age                         3.999e-02  3.740e-04  106.917  < 2e-16 ***
Churn                      -1.027e-01  2.274e-02   -4.516 6.37e-06 ***
DeviceProtection           -4.400e-01  1.584e-02  -27.783  < 2e-16 ***
TechSupport                 5.613e-01  1.616e-02   34.725  < 2e-16 ***
MonthlyCharge              -4.942e-02  2.234e-04 -221.224  < 2e-16 ***
Bandwidth_GB_Year           1.218e-02  4.219e-06 2887.555  < 2e-16 ***
InternetService_Fiber_Optic 1.625e+00  2.173e-02   74.786  < 2e-16 ***
InternetService_DSL        -4.397e+00  2.160e-02 -203.610  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7735 on 9990 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 1.297e+06 on 9 and 9990 DF,  p-value: < 2.2e-16
```

To ensure variables within this model did not exhibit multicollinearity, a **variance inflation factor (VIF)** was calculated for each variable, with the following results:

```
          Children                Age              Churn      DeviceProtection
          1.002037           1.001701           1.683815           1.032103
       TechSupport      MonthlyCharge Bandwidth_GB_Year InternetService_Fiber_Optic
          1.023357           1.538065           1.420574           1.945899
InternetService_DSL
          1.764597
```

VIF measures the ratio of a factor's variance in the full model (with all other factors) to the variance by itself. If there existed 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. All VIF values calculated for this model were below this cutoff value. Therefore, there is no indication of multicollinearity in this model.

D3. Reduced Model

The following explanatory variables were deemed statistically insignificant and removed from the initial linear regression model in a backward stepwise elimination method: Outage_sec_perweek, Email, Population, Yearly_equip_failure, Contract_One_Year, Contract_Two_Year, and Income.

The following explanatory variables were deemed statistically significant to the linear regression and were retained in the reduced model: Children, Age, Churn, DeviceProtection, TechSupport, MonthlyCharge, Bandwidth_GB_Year, InternetService_Fiber_Optic, and InternetService_DSL.

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

```
Call:
lm(formula = Tenure ~ Children + Age + Churn + DeviceProtection +
    TechSupport + MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
    InternetService_DSL, data = churn)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76177 -0.52988 -0.08519  0.60411  1.67983

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                1.215e+00  4.106e-02   29.582  < 2e-16 ***
Children                  -3.795e-01  3.606e-03 -105.237  < 2e-16 ***
Age                        3.999e-02  3.740e-04  106.917  < 2e-16 ***
Churn                     -1.027e-01  2.274e-02   -4.516 6.37e-06 ***
DeviceProtection          -4.400e-01  1.584e-02  -27.783  < 2e-16 ***
TechSupport                5.613e-01  1.616e-02   34.725  < 2e-16 ***
MonthlyCharge             -4.942e-02  2.234e-04 -221.224  < 2e-16 ***
Bandwidth_GB_Year          1.218e-02  4.219e-06 2887.555  < 2e-16 ***
InternetService_Fiber_Optic  1.625e+00  2.173e-02   74.786  < 2e-16 ***
InternetService_DSL       -4.397e+00  2.160e-02 -203.610  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7735 on 9990 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 1.297e+06 on 9 and 9990 DF,  p-value: < 2.2e-16
```

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: residual standard error, adjusted R-squared, and F-statistic probability.

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

```
Residual standard error: 0.7734 on 9983 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 7.3e+05 on 16 and 9983 DF,  p-value: < 2.2e-16
```

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

```
Residual standard error: 0.7735 on 9990 degrees of freedom
Multiple R-squared:  0.9991,    Adjusted R-squared:  0.9991
F-statistic: 1.297e+06 on 9 and 9990 DF,  p-value: < 2.2e-16
```

There is **a minor difference** between the model metrics for the initial and reduced models. Both show statistically significant results, and many metrics are the same or 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:

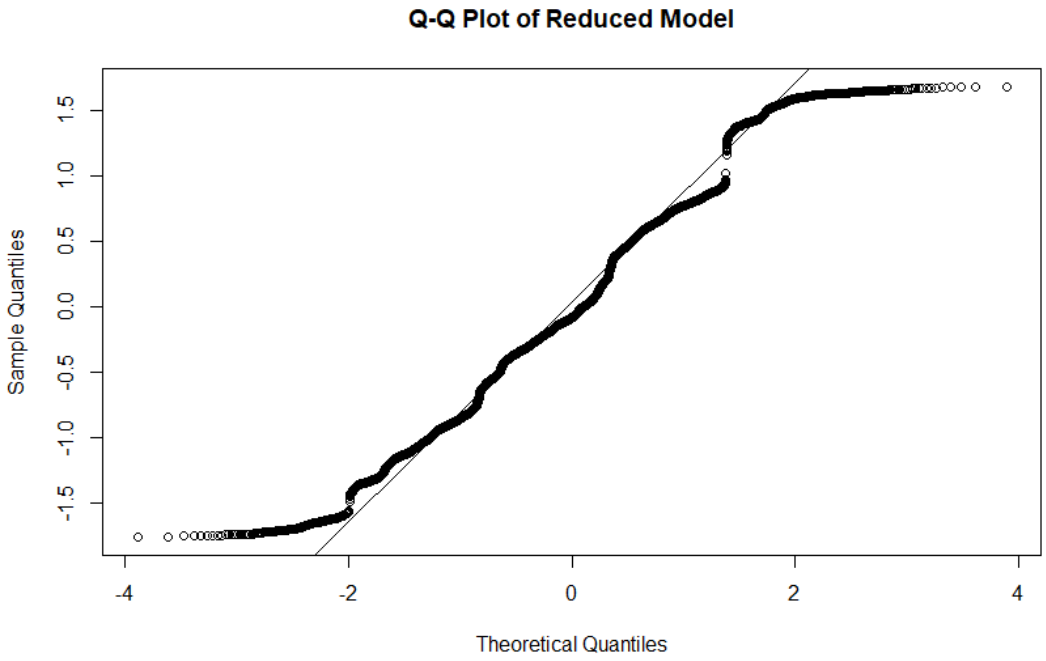
- **Residual standard error:**
    - The residual standard error represents the expected differences between predicted values and observed values for the response variable, Tenure.
    - The residual standard errors for the initial and reduced model are within 0.0001 of each other; the reduced model has a very slightly higher residual standard error.
    - This indicates that the reduced model can predict Tenure with nearly the same accuracy as the initial model, despite having less overall information. The removed information, then, was likely unnecessary as a predictor.
- **Adjusted R-squared:**
    - Adjusted R-squared indicates the percentage of variance in the response variable, Tenure, which can be explained by the explanatory variables. In contrast to multiple R-squared, it does not increase due to a higher number of predictor variables.
    - The adjusted R-squared values for both the initial and reduced model are equivalent down to the ten-thousandth decimal place: 0.9991.
    - This indicates that, for both models, 99.91% of variance in Tenure can be explained by the explanatory variables. Since the variables in both models can explain the same amount of variance, and since adjusted R-squared only increases if a predictor variable adds to the explanatory power of the model, the same amount of information can be communicated with fewer variables in the reduced model.
- **F-statistic probability:**
    - The F-statistic and its associated probability represent the overall significance of the group of variables in the model. If the probability, also known as the p-value, is very low, this means the results were unlikely to be due to chance or luck, and there is a significant relationship between the group of explanatory variables and the response variable.
    - The F-statistic p-values for both the initial and reduced models are extremely low; the values are less than 0.0000000000000002.
    - This indicates that both models are statistically significant and provide meaningful information about the response variable, Tenure. However, since the reduced model uses fewer variables to produce results with similar significance, it would be preferred over the initial model due to clarity and efficiency of results.

E2. Calculations

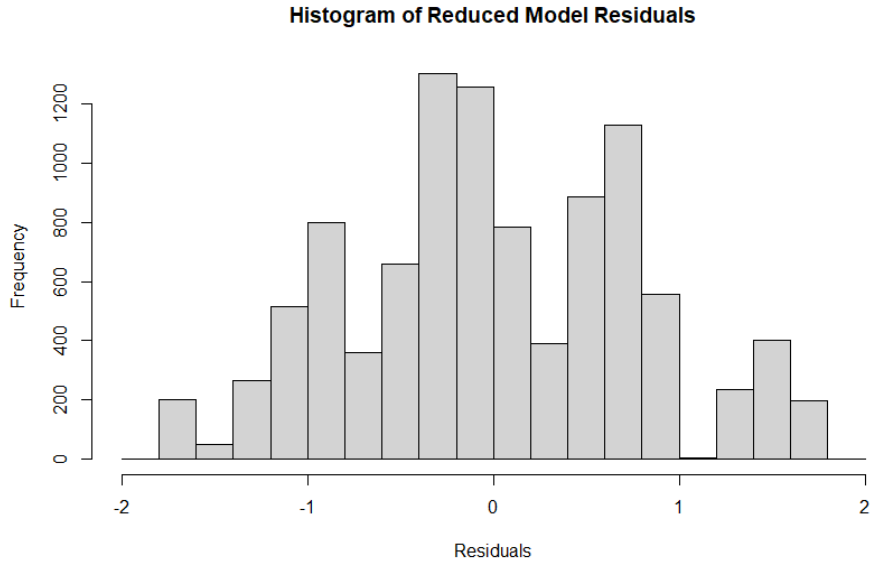
The following summarizes all calculations performed in this analysis, including relevant visualizations of the residuals in the reduced model.

- **Residual Standard Error**: 0.7735 on 9990 degrees of freedom.

- **Q-Q Plot of reduced model**:



- **Histogram of residuals in reduced model**:



E3. Linear 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_Task1_Code.R.

```
# CREATING INITIAL MODEL:

in_model <- lm(Tenure ~ Population + Children + Age + Income + Churn + 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)

summary(in_model)

# REDUCING MODEL:

re_model <- lm(Tenure ~ Population + Children + Age + Income + Churn + Email +
Yearly_equip_failure +
        DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
InternetService_Fiber_Optic +
        InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn)
# remove Outage_sec_perweek

summary(re_model)

re_model <- lm(Tenure ~ Population + Children + Age + Income + Churn + Yearly_equip_failure
+
        DeviceProtection + TechSupport + MonthlyCharge + Bandwidth_GB_Year +
InternetService_Fiber_Optic +
        InternetService_DSL + Contract_One_Year + Contract_Two_Year, churn)
# remove Email

summary(re_model)

re_model <- lm(Tenure ~ Children + Age + Income + Churn + Yearly_equip_failure +
DeviceProtection + TechSupport +
        MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
InternetService_DSL +
        Contract_One_Year + Contract_Two_Year, churn)
# remove Population

summary(re_model)
```

```
re_model <- lm(Tenure ~ Children + Age + Income + Churn + DeviceProtection + TechSupport +
               MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
InternetService_DSL +
               Contract_One_Year + Contract_Two_Year, churn)
# remove Yearly_equip_failure

summary(re_model)

re_model <- lm(Tenure ~ Children + Age + Income + Churn + DeviceProtection + TechSupport +
               MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
InternetService_DSL +
               Contract_Two_Year, churn)
# remove Contract_One_Year

summary(re_model)

re_model <- lm(Tenure ~ Children + Age + Income + Churn + DeviceProtection + TechSupport +
               MonthlyCharge + Bandwidth_GB_Year + InternetService_Fiber_Optic +
InternetService_DSL, churn)
# remove Contract_Two_Year

summary(re_model)

re_model <- lm(Tenure ~ Children + Age + Churn + DeviceProtection + TechSupport +
MonthlyCharge +
               Bandwidth_GB_Year + InternetService_Fiber_Optic + InternetService_DSL, churn)
# remove Income

summary(re_model)

library(car)
vif(re_model)

re_res <- resid(re_model)

qqnorm(re_res, main = "Q-Q Plot of Reduced Model")
qqline(re_res)

hist(re_res, main = "Histogram of Reduced Model Residuals", xlab = "Residuals", breaks = seq(-
2,2,0.2))
```

F1. Regression Equation

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

```
Coefficients:
                            Estimate
(Intercept)                1.215e+00
Children                  -3.795e-01
Age                        3.999e-02
Churn                     -1.027e-01
DeviceProtection          -4.400e-01
TechSupport                5.613e-01
MonthlyCharge             -4.942e-02
Bandwidth_GB_Year          1.218e-02
InternetService_Fiber_Optic 1.625e+00
InternetService_DSL       -4.397e+00
```

These produce the following final equation of the reduced linear model, with coefficients rounded to the nearest hundredth value:

$$\begin{aligned} Tenure = \ & 1.22 - 0.38(Children) + 0.04(Age) - 0.10(Churn) - 0.44(DeviceProtection) \\ & + 0.56(TechSupport) - 0.05(MonthlyCharge) \\ & + 0.01(Bandwidth\_GB\_Year) \ + \ 1.63(InternetService\_Fiber\_Optic) \\ & - 4.40(InternetService\_DSL) \end{aligned}$$

In this equation, each coefficient can be **interpreted** as such:

- The **y-intercept, 1.22**, is the expected value for the response variable, Tenure, when all explanatory variables have a value of 0. Thus, a customer with a value of 0 in all listed variables in the reduced model would be expected to have been with the service for approximately 1.22 months.
- The **Children coefficient, -0.38**, indicates that an increase of 1 in the Children variable is associated with an approximate decrease of 0.38 in the Tenure variable. Thus, a customer with 1 more child than another is expected to have been with the service for 0.38 fewer months than the other customer.
- The **Age coefficient, 0.04**, indicates that an increase of 1 in the Age variable is associated with an approximate increase of 0.04 in the Tenure variable. Thus, a customer 1 year older than another is expected to have been with the service for 0.04 more months than the other customer.
- The **Churn coefficient, -0.10**, indicates that an increase of 1 in the Churn variable is associated with an approximate decrease of 0.1 in the Tenure variable. Since Churn is a binary value, this means that a customer who has left the service (represented by "1" in the transformed numeric variable) is expected to have been with the service for 0.1 fewer months than a customer who has continued with the service (represented by "0" in the transformed numeric variable).
- The **DeviceProtection coefficient, -0.44**, indicates that an increase of 1 in the DeviceProtection variable is associated with an approximate decrease of 0.44 in the Tenure variable. Since DeviceProtection is a binary value, this means that a customer who has a device protection add-on (represented by "1" in the transformed numeric variable) is

expected to have been with the service for 0.44 fewer months than a customer who does not have a device protection add-on (represented by "0" in the transformed numeric variable).

- The **TechSupport coefficient, 0.56**, indicates that an increase of 1 in the TechSupport variable is associated with an approximate increase of 0.56 in the Tenure variable. Since TechSupport is a binary value, this means that a customer who has a technical support add-on (represented by "1" in the transformed numeric variable) is expected to have been with the service for 0.56 more months than a customer who does not have a technical support add-on (represented by "0" in the transformed numeric variable).
- The **MonthlyCharge coefficient, -0.05**, indicates that an increase of 1 in the MonthlyCharge variable is associated with an approximate decrease of 0.05 in the Tenure variable. Thus, a customer being charged $1 more per month than another is expected to have been with the service for 0.05 fewer months than the other customer.
- The **Bandwidth_GB_Year coefficient, 0.01**, indicates that an increase of 1 in the Bandwidth_GB_Year variable is associated with an approximate increase of 0.01 in the Tenure variable. Thus, a customer with 1 more gigabyte of data used per year than another is expected to have been with the service for 0.01 more months than the other customer.
- The **InternetService_Fiber_Optic coefficient, 1.63**, indicates that an increase of 1 in the InternetService_Fiber_Optic variable is associated with an approximate increase of 1.63 in the Tenure variable. Since it is a binary value, this means that a customer who has Fiber Optic for an internet provider (represented by "1" in the transformed numeric variable) is expected to have been with the service for 1.63 more months than a customer who does not have Fiber Optic (represented by "0" in the transformed numeric variable).
- The **InternetService_DSL coefficient, -4.4**, indicates that an increase of 1 in the InternetService_DSL variable is associated with an approximate decrease of 4.4 in the Tenure variable. Since it is a binary value, this means that a customer who has DSL for an internet provider (represented by "1" in the transformed numeric variable) is expected to have been with the service for 4.4 fewer months than a customer who does not have DSL (represented by "0" in the transformed numeric variable).

The significance of these results is summarized as follows:

- This final reduced model has been determined to be **statistically significant**.
    - All factors have p-values much lower than the given alpha level of 0.01.
    - The overall F-statistic probability is also extremely low.
    - This means the probability that the calculated coefficient estimates are the result of 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 customer factors and how long a customer continues with the service (avoiding customer churn).

    - These results can be used to analyze what makes customers more or less likely to stay longer with the service, which can inform further research and implementation of strategies focused on maximizing Tenure.
    - For example, since Bandwidth Gigabytes Per Year was found to have a statistically significant relationship with Tenure, efforts can be made to target customers with low data usage and attempt to increase their data usage, which may encourage them to stay longer with the service.

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

- Only a subset of variables was selected in the initial data exploration and linear regression. These variables were selected due to their perceived practical connection to customer Tenure; however, it is possible for one or more unselected variables in the initial dataset to have significant relationships with Tenure. 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 Tenure, it is important to remember that correlation does not imply causation. These connections could be the result of confounding variables or simply coincidences, so conclusions that assume direct causations between correlated variables should be treated with caution.

F2. Recommendations

Based on the results of the multiple linear regression, strong relationships have been identified between the response variable, Tenure, and 8 explanatory variables. These relationships should be explored further, as they may uncover ways to increase the time customers spend with the service, which would inherently reduce churn rate.

- There is a negative relationship between Children and Tenure. This information should be used to inform future research on whether customers with more children stay with the service for less time. If so, perhaps family deals can be promoted and directed specifically at customers with children.
- There is a positive relationship between Age and Tenure. This may be simply due to the nature of time (the more time a customer is with the service, the older they get), or it may indicate that older customers are more likely to stay with the service. If so, it may be beneficial to create deals aimed specifically at young customers, in hopes of getting them to maintain their status with the service.
- There is a negative relationship between Churn and Tenure. This may imply that customers who have been with the service for less time are more likely to discontinue. If so, more incentives could be created to ensure newer customers continue with the service.

- There is a negative relationship between DeviceProtection and Tenure. This may indicate something wrong with the device protection add-on program. Further research could be conducted to determine if the program is not working as intended; perhaps surveys could be sent out to customers with the device protection to measure their satisfaction with it.
- There is a positive relationship between TechSupport and Tenure. This may indicate 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 negative relationship between MonthlyCharge and Tenure. This may indicate that customers who are being charged less are more likely to stay longer with the service. Of course, lowering monthly charge values 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 positive relationship between Bandwidth_GB_Year and Tenure. This may imply that customers who use more data are more likely to stay longer with the service. Efforts could be made to encourage customers to use more data, potentially increasing their reliance on the service and their likelihood of continuing with it.
- There is a positive relationship between InternetService_Fiber_Optic and Tenure and a negative relationship between InternetService_DSL and Tenure. This may indicate a tendency for customers with Fiber Optic as an internet provider to stay with the service for longer, and for customers with DSL to stay with the service for less time. Further research may include sending surveys to Fiber Optic customers to determine what may be keeping them with the service. Any insight gained from collecting more information could be applied to DSL customers to encourage them to stay in the same manner.

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.

I. Content References

There is no content in this analysis that has been quoted, paraphrased, summarized, or otherwise requires direct citation.