2023

Multiple Linear Regression Model

Task 1: LINEAR REGRESSION MODELING

Part I: Research Question (A)

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

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

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

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

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

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

Part II: Method Justification (B)

Describe multiple linear regression methods by doing the following:

1. **Summarize four assumptions of a multiple linear regression model.**

Multiple linear regression models (MLR) depend on assumptions of the dataset for the regression to be viable.

There are four main assumptions of a linear regression model:

1. Linear Relationship – A MLR depends on the dataset having a linear relationship between the independent variable (x) and the dependent variable (y). Often, a scatterplot is helpful in determining linear relationships (Statistics Solutions, 2023).
2. Multivariate Normality – MLR assume that residuals are normally distributed (Statistics Solutions, 2023). Residuals within a MLR are the difference between a data point and the regression line (Glen, 2023). Residuals are often referred to as an error term (Hayes, 2021).
3. No Multicollinearity – MLR draws the assumption that independent variables are not highly correlated with one another. When multicollinearity is present, it indicates independent variables are too highly correlated with each other (Statistics Solutions, 2023).

4. Homoscedasticity – This is a condition in which the variance of the error terms, or residuals, in a regression model is constant. In other words, the variances of the data points are generally the same for all data points in the specified dataset (Kenton, 2022).

It is important to note for the purposes of this assessment that the MLR model assumes that the dependent variable(s) being examined are continuous (Berry, 2005).

2. Describe two benefits of using Python or R in support of various phases of the analysis.

The two benefits of using R Programming for this assessment is:

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

3. Explain why multiple linear regression an appropriate technique is to use for analyzing the research question summarized in part I.

MLR is an appropriate technique because it is a model that is used to predict the value of a target variable based on the values of one or more independent variables. In this case, the target variable that is being examined will be the Tenure of customers based on the multiple independent variables within the dataset (Reilly, 2023). It is important to note for the purposes of this assessment that the MLR model assumes that the dependent variable(s) being examined are continuous (Berry, 2005).

Part III: Data Preparation (C)

Summarize the data preparation process for multiple linear regression analysis by doing the following:

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

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

My methodology to achieve the data goals are as follows:

1. Make a copy of the data
2. Import data into R programming.
3. Examine the structure of the data to better understand the dataset.
4. Examine and clean the data for potential missing data, renaming columns, duplications, data errors, anomalies, removal of unneeded variables, or anything else that might aid in the analysis.
5. Summarize data by discovering the distribution and potential outliers within the variables that might alter the statistical analysis of the dataset using both histograms and boxplots. Handle outliers as necessary.
6. Summarize usable data by utilizing measures of central tendency (mean, median, mode).

The steps used to clean the data and the code are found within the next step of the assessment.

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

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

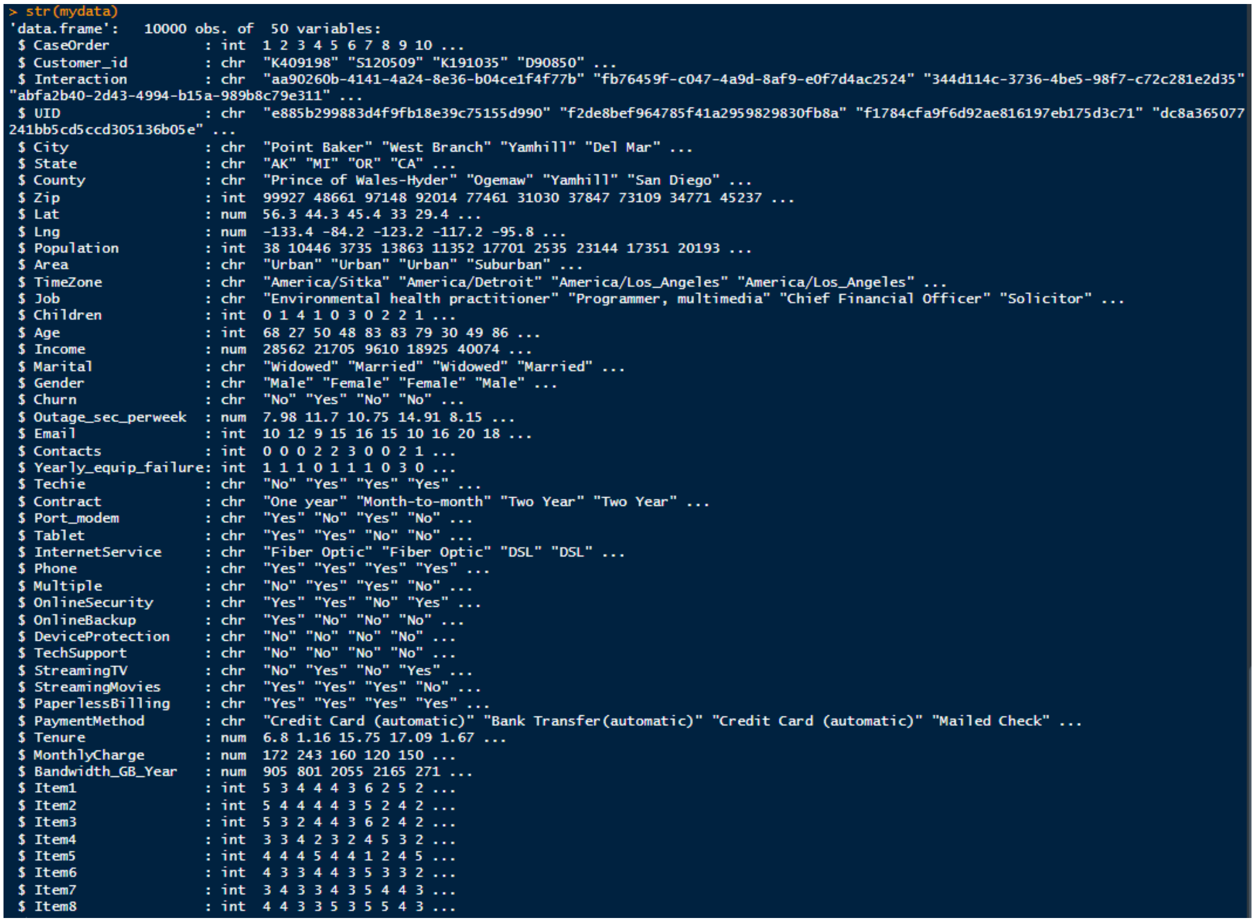
Using R, packages were imported to conduct analysis. Once the packages were imported, setwd() was used to create a working directory. Then, importing the .csv file was used using read.csv():

```
# Packages that will be used for regression:

        library(tidyverse)
        library(dplyr)
        library(plyr)
        library(readr)
        library(ggplot2)
        library(gridExtra)
        library(stats)
        library(gplots)

# Setting the working directory:

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

# Importing the dataset:

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

# Renaming the dataset:

        mydata <- churn_df
```

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

```
# Summary/Structure of Data

    str(mydata)
    summary(mydata)
```

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



```
> str(mydata)
'data.frame':   10000 obs. of  50 variables:
 $ CaseOrder          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Customer_id        : chr  "K409198" "S120509" "K191035" "D90850" ...
 $ Interaction        : chr  "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "344d114c-3736-4be5-98f7-c72c281e2d35"
"abfa2b40-2d43-4994-b15a-989b8c79e311" ...
 $ UID                : chr  "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a" "f1784cfa9f6d92ae816197eb175d3c71" "dc8a365077
241bb5cd5ccd305136b05e" ...
 $ City               : chr  "Point Baker" "West Branch" "Yamhill" "Del Mar" ...
 $ State              : chr  "AK" "MI" "OR" "CA" ...
 $ County             : chr  "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...
 $ Zip                : int  99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...
 $ Lat                : num  56.3 44.3 45.4 33 29.4 ...
 $ Lng                : num  -133.4 -84.2 -123.2 -117.2 -95.8 ...
 $ Population         : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
 $ Area               : chr  "Urban" "Urban" "Urban" "Suburban" ...
 $ TimeZone           : chr  "America/Sitka" "America/Detroit" "America/Los_Angeles" "America/Los_Angeles" ...
 $ Job                : chr  "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Officer" "Solicitor" ...
 $ Children           : int  0 1 4 1 0 3 0 2 2 1 ...
 $ Age                : int  68 27 50 48 83 83 79 30 49 86 ...
 $ Income             : num  28562 21705 9610 18925 40074 ...
 $ Marital            : chr  "Widowed" "Married" "Widowed" "Married" ...
 $ Gender             : chr  "Male" "Female" "Female" "Male" ...
 $ Churn              : chr  "No" "Yes" "No" "No" ...
 $ Outage_sec_perweek : num  7.98 11.7 10.75 14.91 8.15 ...
 $ Email              : int  10 12 9 15 16 15 10 16 20 18 ...
 $ Contacts           : int  0 0 0 2 2 3 0 0 2 1 ...
 $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
 $ Techie             : chr  "No" "Yes" "Yes" "Yes" ...
 $ Contract           : chr  "One year" "Month-to-month" "Two Year" "Two Year" ...
 $ Port_modem         : chr  "Yes" "No" "Yes" "No" ...
 $ Tablet             : chr  "Yes" "Yes" "No" "No" ...
 $ InternetService    : chr  "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...
 $ Phone              : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Multiple           : chr  "No" "Yes" "Yes" "No" ...
 $ OnlineSecurity     : chr  "Yes" "Yes" "No" "Yes" ...
 $ OnlineBackup       : chr  "Yes" "No" "No" "No" ...
 $ DeviceProtection   : chr  "No" "No" "No" "No" ...
 $ TechSupport        : chr  "No" "No" "No" "No" ...
 $ StreamingTV        : chr  "No" "Yes" "No" "Yes" ...
 $ StreamingMovies    : chr  "Yes" "Yes" "Yes" "No" ...
 $ PaperlessBilling   : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ PaymentMethod      : chr  "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)" "Mailed Check" ...
 $ Tenure             : num  6.8 1.16 15.75 17.09 1.67 ...
 $ MonthlyCharge      : num  172 243 160 120 150 ...
 $ Bandwidth_GB_Year  : num  905 801 2055 2165 271 ...
 $ Item1              : int  5 3 4 4 4 3 6 2 5 2 ...
 $ Item2              : int  5 4 4 4 4 3 5 2 4 2 ...
 $ Item3              : int  5 3 2 4 4 3 6 2 4 2 ...
 $ Item4              : int  3 3 4 2 3 2 4 5 3 2 ...
 $ Item5              : int  4 4 4 5 4 4 1 2 4 5 ...
 $ Item6              : int  4 3 3 4 4 3 5 3 3 2 ...
 $ Item7              : int  3 4 3 3 4 3 5 4 4 3 ...
 $ Item8              : int  4 4 3 3 5 3 5 5 4 3 ...
```

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

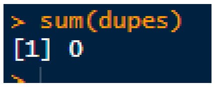
Lastly, there are 8 variables concerning survey information: Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8.

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

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

```
        # Searching for Duplicates

                dupes <- duplicated(mydata)

        # Summing to see if duplicates are present:

                sum(dupes)
```

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



```
> sum(dupes)
[1] 0
>
```

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

These can be removed:

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

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

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

The following is the code to convert to binary:

```
# Creating Dummy Variables for Categorical Data

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

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

Code:

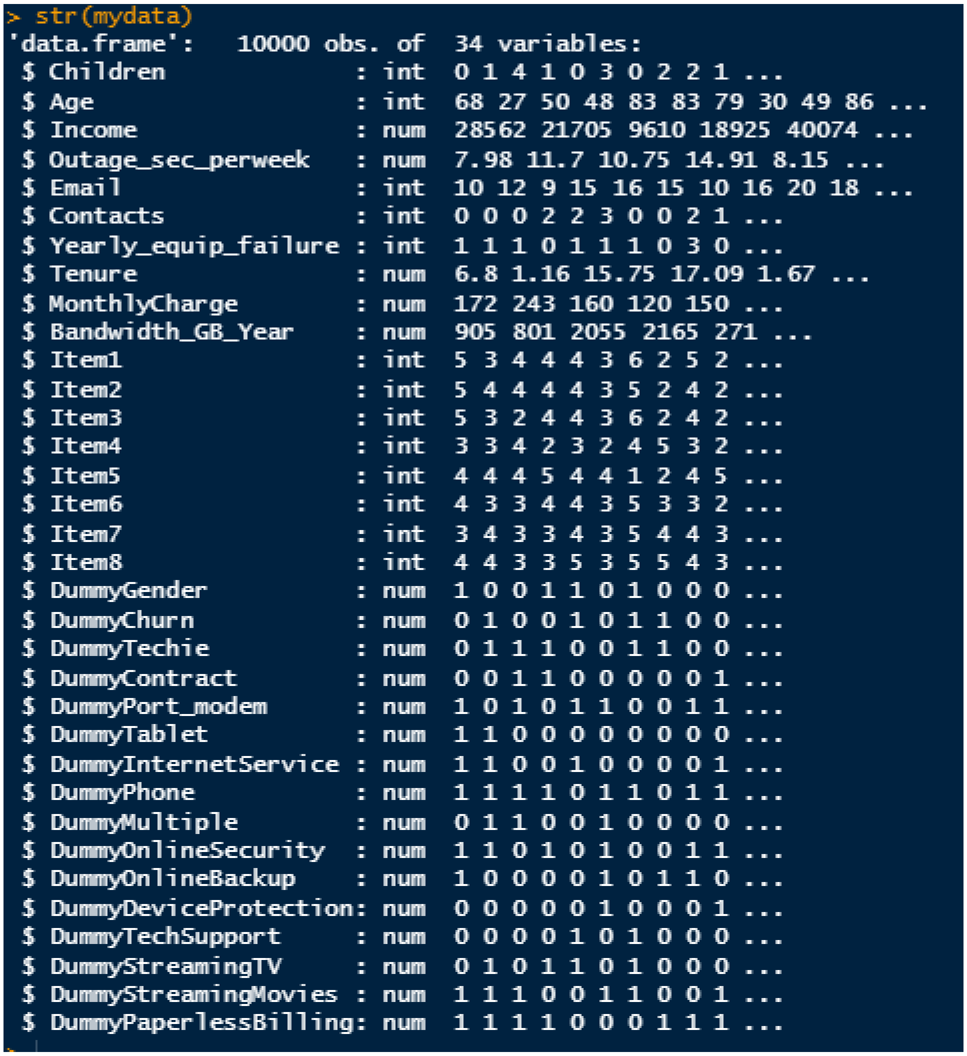
```
# Dropping all old categorical variables:

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

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

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

Code:

str(mydata)

Output:



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

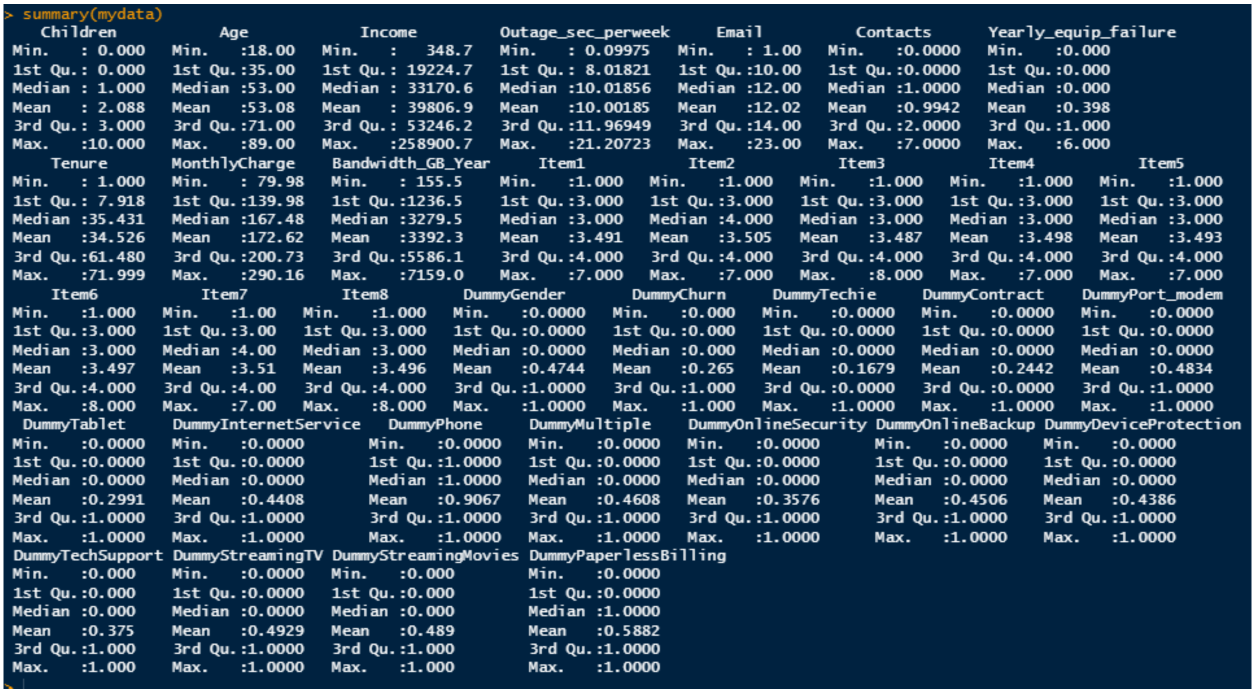
The columns have been removed, replaced, and categories are now binary.

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

# Look for missing data points via summary()

Summary(mydata)

The output of this command:



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

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

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

```
# Changing Column Names of Ordinal Data:

        colnames(mydata)[colnames(mydata) == 'Item1'] <- 'TimelyResponse'
        colnames(mydata)[colnames(mydata) == 'Item2'] <- 'TimelyFixes'
        colnames(mydata)[colnames(mydata) == 'Item3'] <- 'TimelyReplacements'
        colnames(mydata)[colnames(mydata) == 'Item4'] <- 'Reliability'
        colnames(mydata)[colnames(mydata) == 'Item5'] <- 'Options'
        colnames(mydata)[colnames(mydata) == 'Item6'] <- 'RespectfulResponse'
        colnames(mydata)[colnames(mydata) == 'Item7'] <- 'CourtExchange'
        colnames(mydata)[colnames(mydata) == 'Item8'] <- 'EvidenceActiveListening'
```

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

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

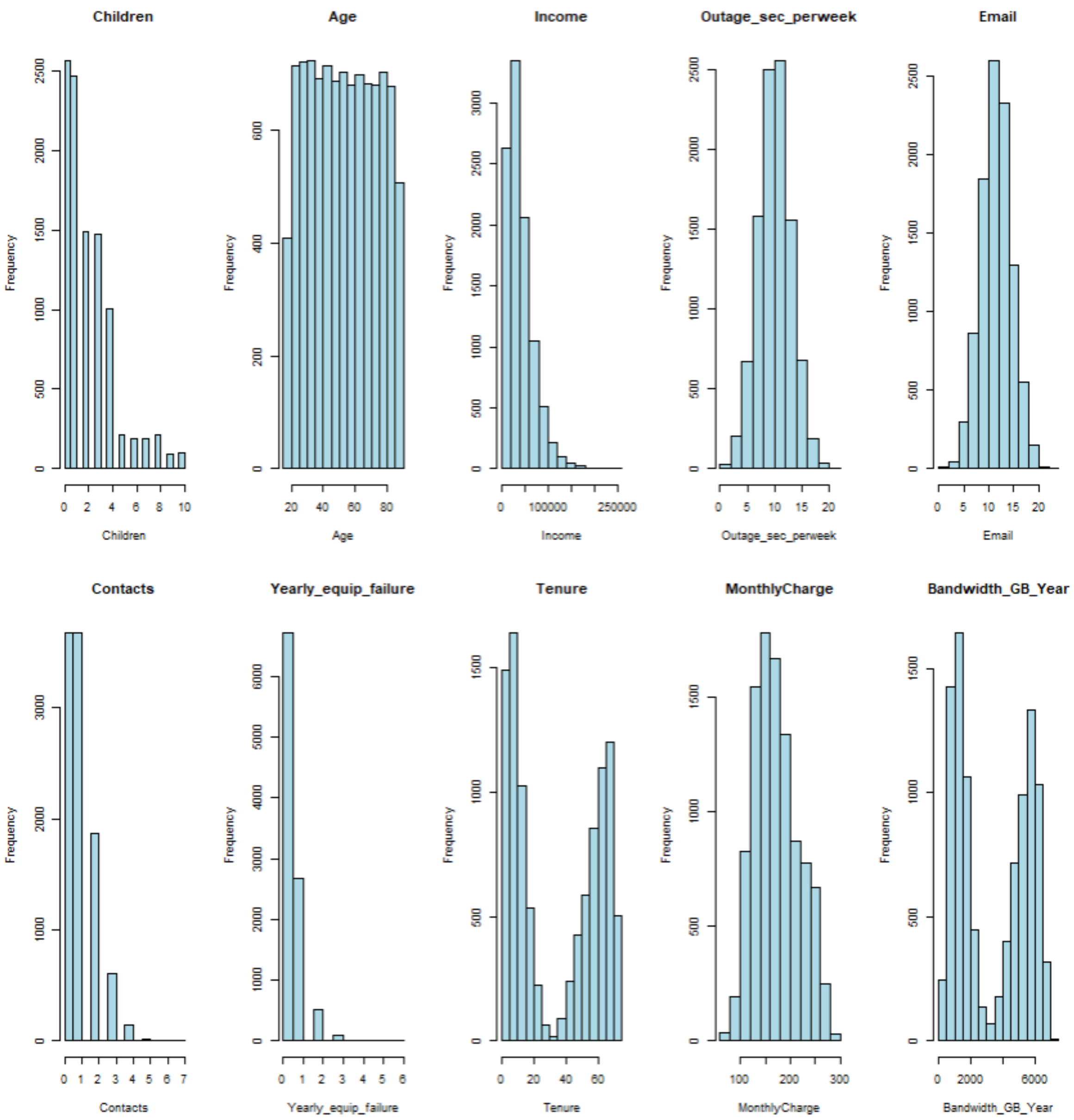
To begin, analyzing both continuous and categorical variables is required.

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

Histograms of Continuous Variables:

Code:

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

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


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

par(mfrow = c(2, 5))

# Creating the histograms:

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

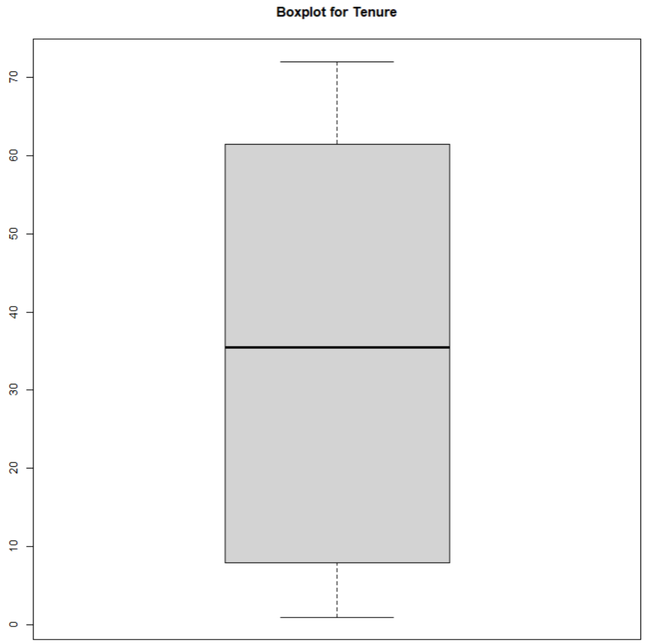


<u>Boxplot for each continuous variable:</u>

1. Tenure – No outliers present.

   Code:

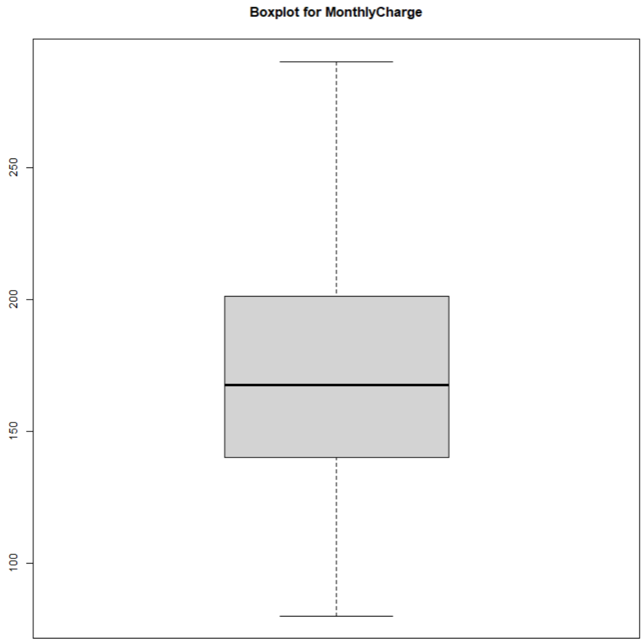
   boxplot(mydata$Tenure, main = 'Boxplot for Tenure')$out



2. MonthlyCharges – No outliers present.

   Code:

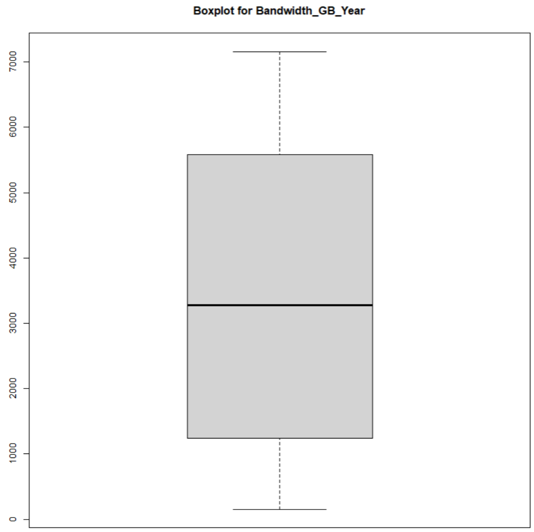
   ```
   boxplot(mydata$MonthlyCharge, main = 'Boxplot for MonthlyCharge')$out
   ```



3. Bandwidth_GB_Year – No outliers present.

   Code:

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



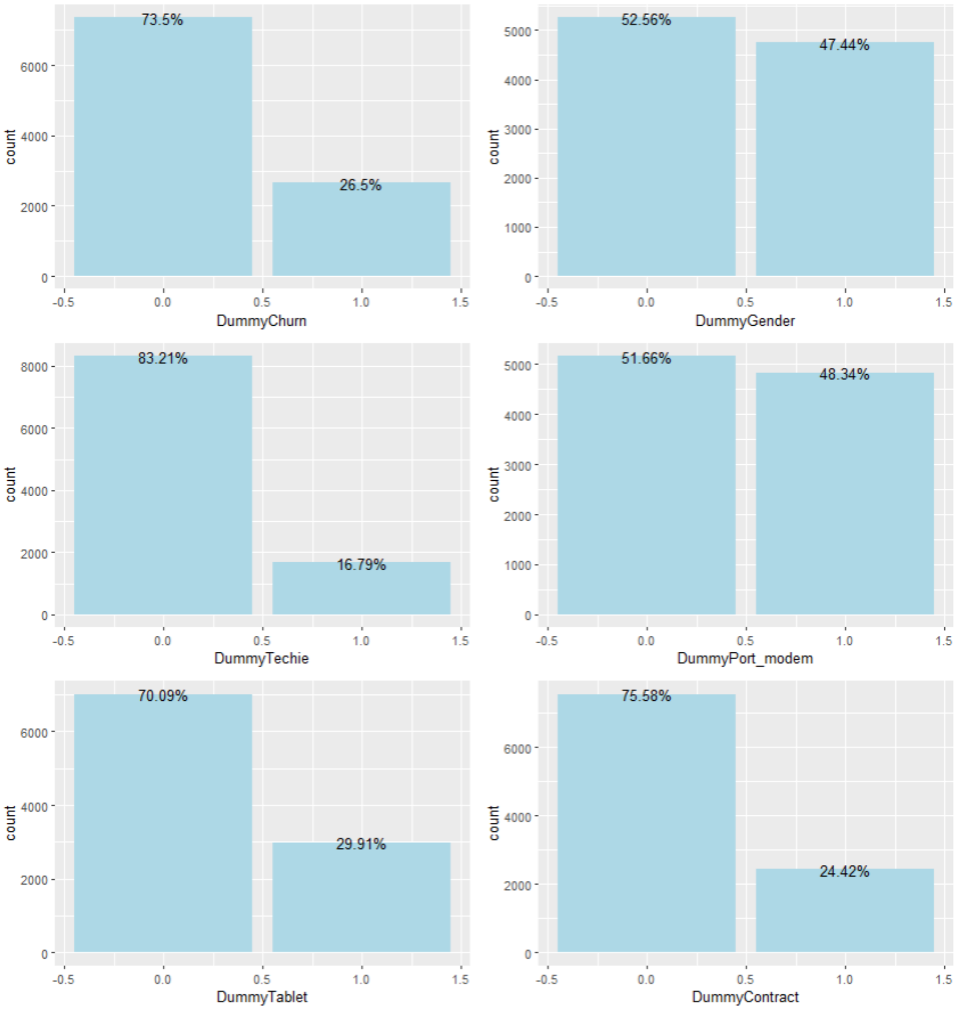
There does not appear to be any outliers in the modified dataset.

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

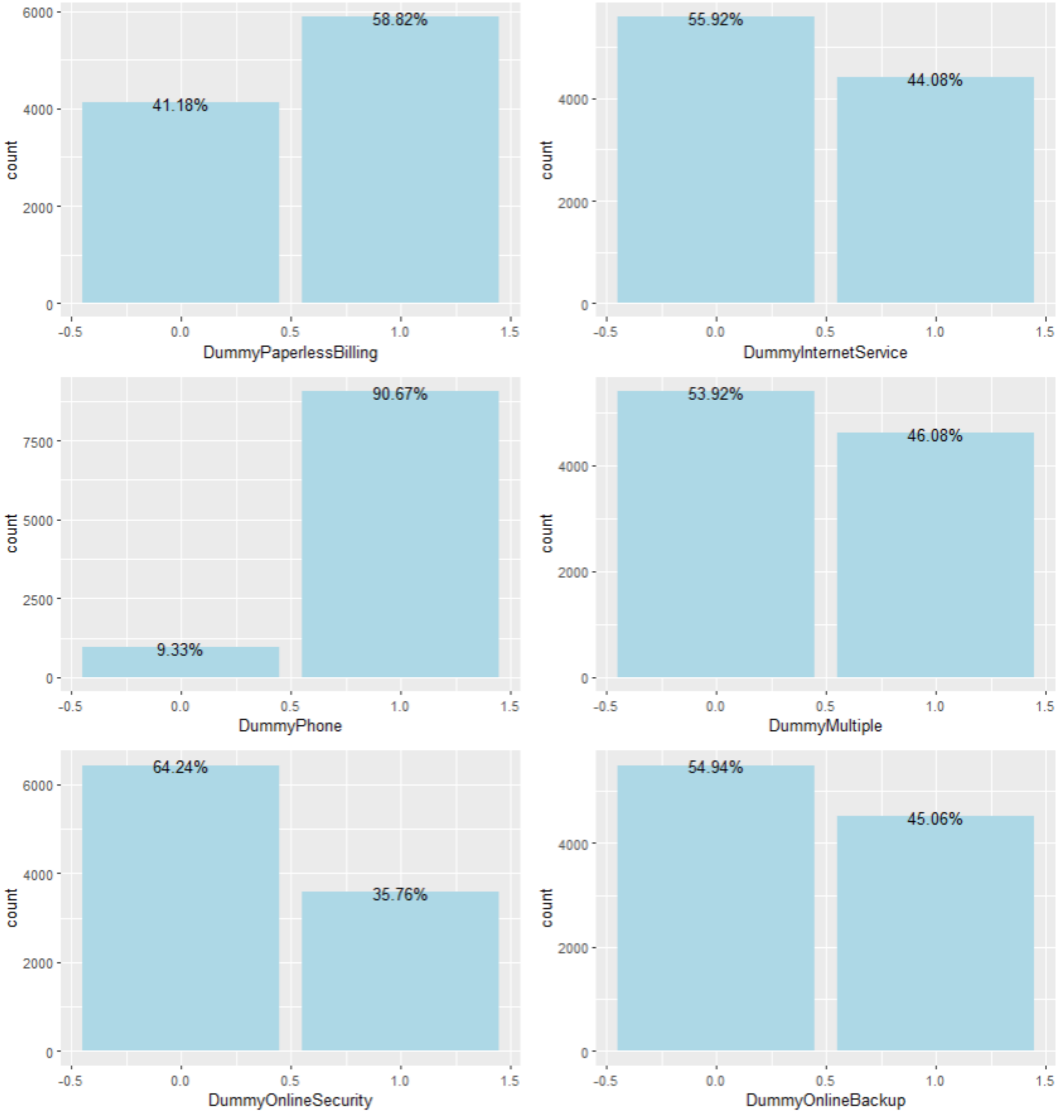
This is the following code to create the visualizations that allow summaries of the categorical variables:

```
# Summary of Independent Variables

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

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

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

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

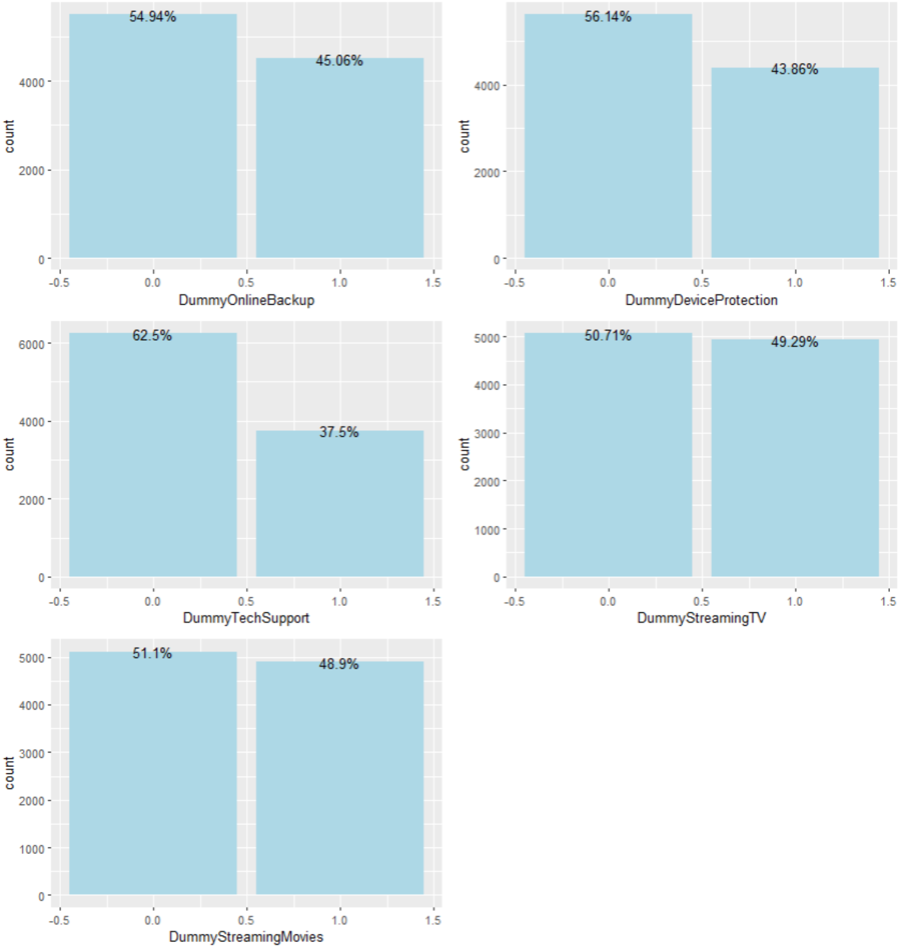
Once these were created, using grid.arrange() allowed the create the following visualizations (they were broken up into 3 grids to make it more readable):



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



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

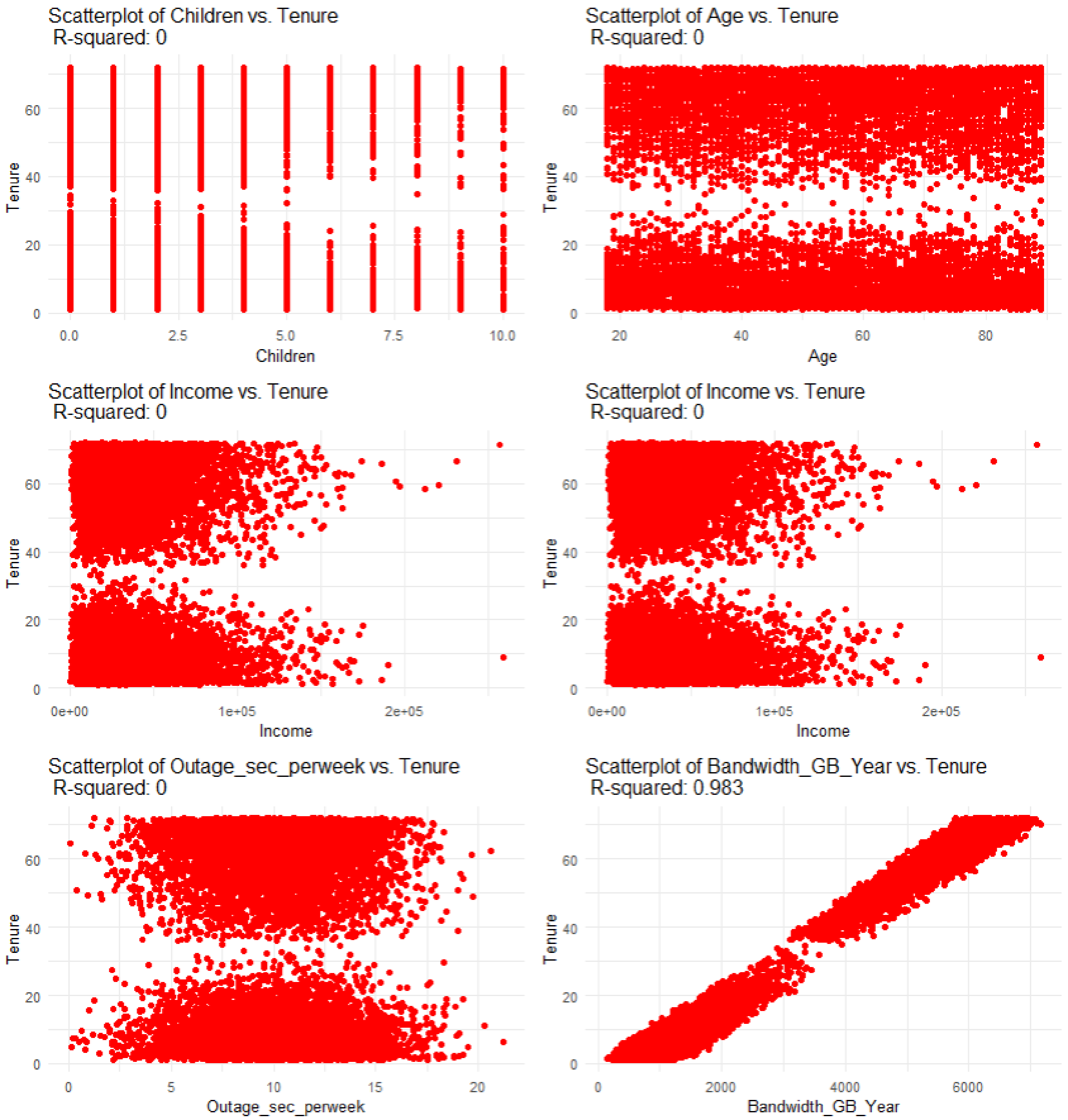


As seen in the above visualization, over 56% of their customers do not have online backups. More than half (> 56%) do not have device protection on their devices. Additionally, only 36% of customers have a technical support add-on. When it comes up Streaming TV, there is an almost 50/50 on customers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

Next, **bivariate statistics** are conducted. Running scatterplots is an appropriate way to see linear relationships within the dependent variable: Tenure. Also taking the r-squared value to determine there is a relationship between Tenure and the other variables.

All code for these scatterplots will be as follows but change the **X** variable for each plot:

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

    ggplot(mydata, aes(x = Children, y = Tenure)) +  geom_point(color =
    'red') + labs(title = paste('Scatterplot of Children vs. Tenure\n',
                     'R-squared:', round(cor(mydata$Children,
    mydata$Tenure)^2, 3)),
          x = 'Children',
    y = 'Tenure') +
theme_minimal()
```



Scatterplot of Children vs. Tenure
R-squared: 0

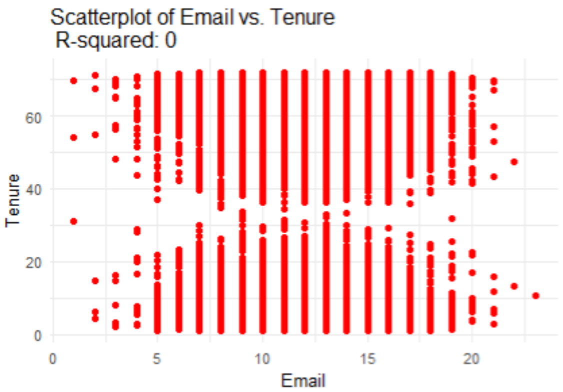
Scatterplot of Age vs. Tenure
R-squared: 0

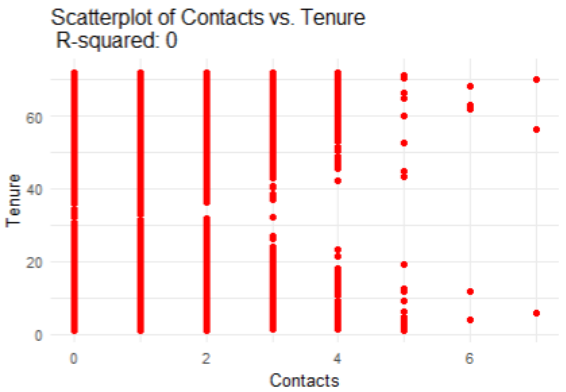
Scatterplot of Income vs. Tenure
R-squared: 0

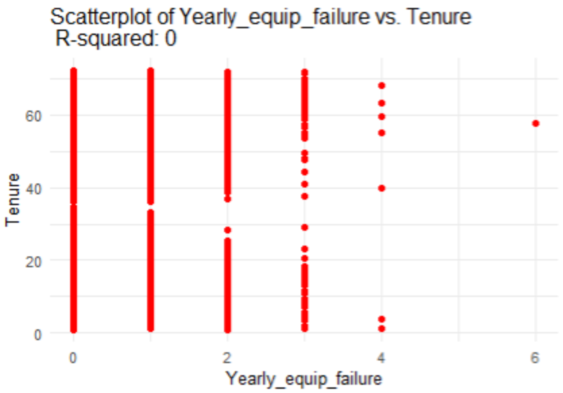
Scatterplot of Income vs. Tenure
R-squared: 0

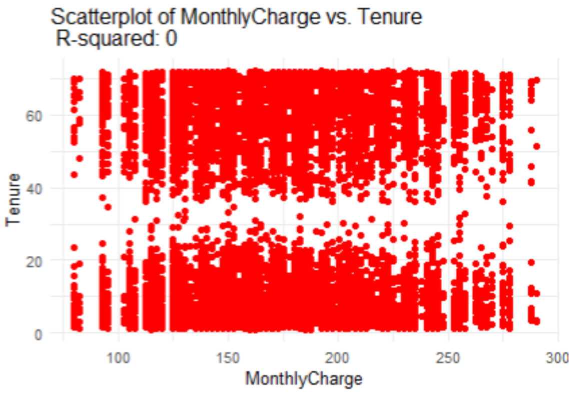
Scatterplot of Outage_sec_perweek vs. Tenure
R-squared: 0

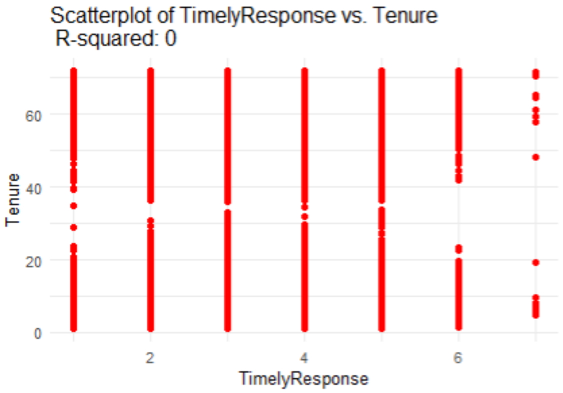
Scatterplot of Bandwidth_GB_Year vs. Tenure
R-squared: 0.983

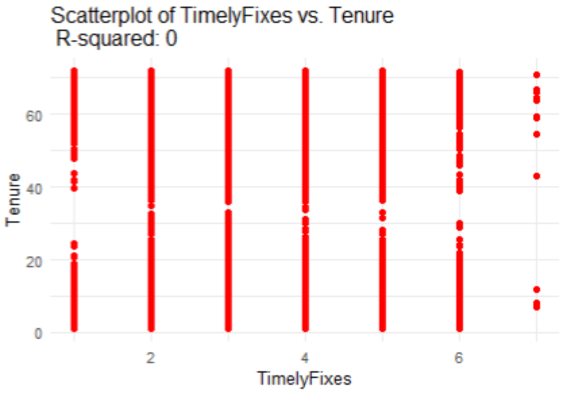

Scatterplot of Email vs. Tenure
R-squared: 0

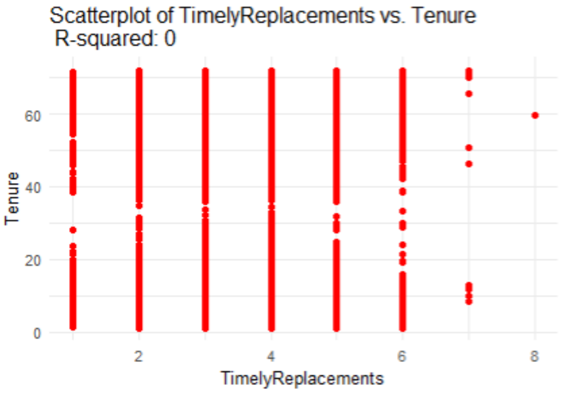

Scatterplot of Contacts vs. Tenure
R-squared: 0


Scatterplot of Yearly_equip_failure vs. Tenure
R-squared: 0

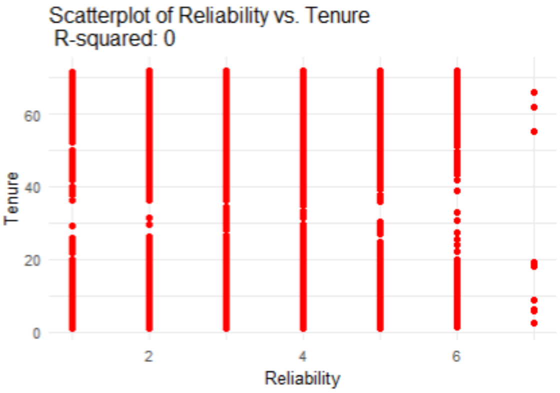

Scatterplot of MonthlyCharge vs. Tenure
R-squared: 0


Scatterplot of TimelyResponse vs. Tenure
R-squared: 0

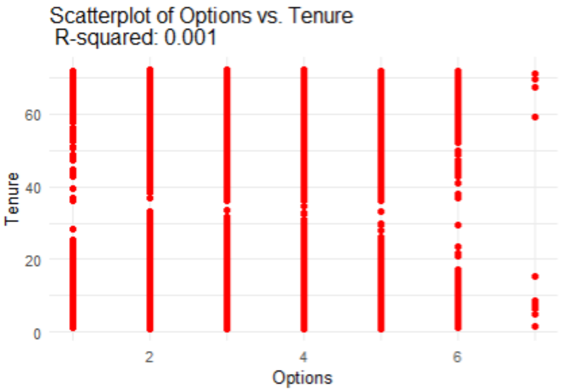

Scatterplot of TimelyFixes vs. Tenure
R-squared: 0



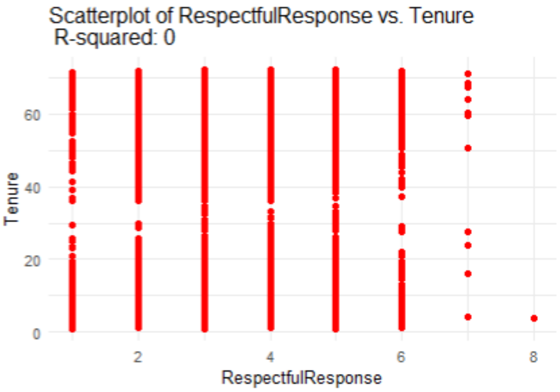
Scatterplot of TimelyReplacements vs. Tenure
R-squared: 0



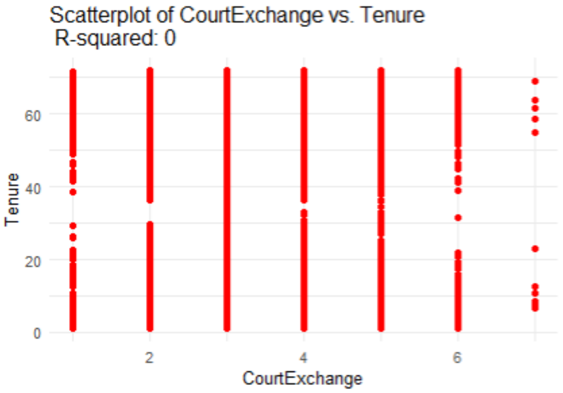
Scatterplot of Reliability vs. Tenure
R-squared: 0



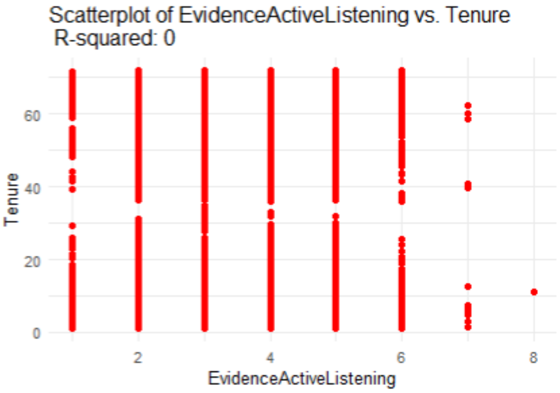
Scatterplot of Options vs. Tenure
R-squared: 0.001



Scatterplot of RespectfulResponse vs. Tenure
R-squared: 0



Scatterplot of CourtExchange vs. Tenure
R-squared: 0



Scatterplot of EvidenceActiveListening vs. Tenure
R-squared: 0

All scatterplots express a low R-Square except Tenure vs. Bandwidth_GB_Year. The R-Square Value for this scatterplot was 0.983. However, more analysis is needed to understand the relationship between Tenure and the other independent variables. A high R-square does not necessarily mean causation.

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

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

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

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

# .csv of data transformation

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

This will be uploaded with the assessment.

Part IV: Model Comparison and Analysis (D)

Compare an initial and a reduced linear regression model by doing the following:

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

The multiple linear regression model was performed to include all independent variables with Tenure being the dependent variable.

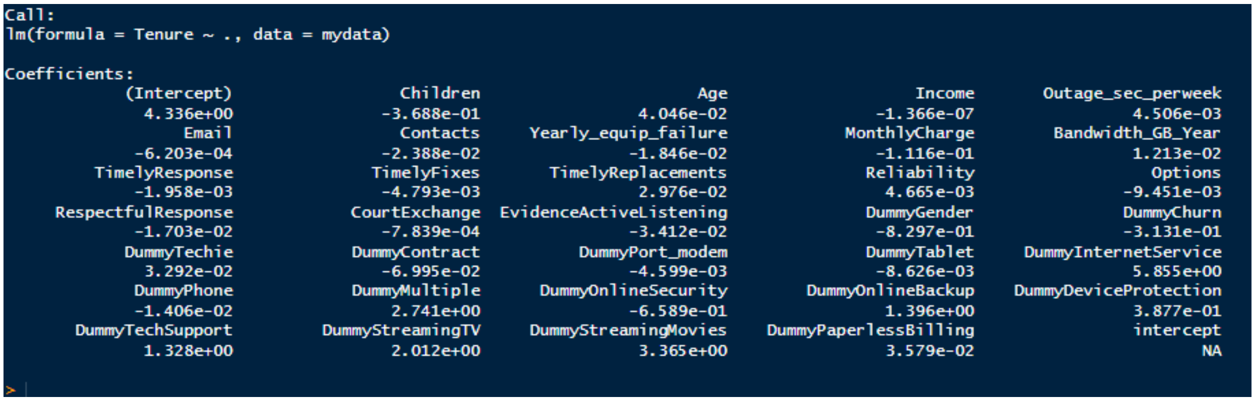
Code:

# Using all independent variable minus Tenure:

```
lm_all_variables <- lm(Tenure ~ ., data = mydata)
```

# Printing out the results:

```
print(lm_all_variables)
```



```
Call:
lm(formula = Tenure ~ ., data = mydata)

Coefficients:
            (Intercept)                Children                     Age                  Income      Outage_sec_perweek
              4.336e+00              -3.688e-01               4.046e-02              -1.366e-07               4.506e-03
                  Email                Contacts    Yearly_equip_failure           MonthlyCharge       Bandwidth_GB_Year
             -6.203e-04              -2.388e-02              -1.846e-02              -1.116e-01               1.213e-02
         TimelyResponse             TimelyFixes      TimelyReplacements             Reliability                 Options
             -1.958e-03              -4.793e-03               2.976e-02               4.665e-03              -9.451e-03
     RespectfulResponse           CourtExchange EvidenceActiveListening             DummyGender              DummyChurn
             -1.703e-02              -7.839e-04              -3.412e-02              -8.297e-01              -3.131e-01
            DummyTechie           DummyContract         DummyPort_modem             DummyTablet    DummyInternetService
              3.292e-02              -6.995e-02              -4.599e-03              -8.626e-03               5.855e+00
             DummyPhone           DummyMultiple     DummyOnlineSecurity       DummyOnlineBackup   DummyDeviceProtection
             -1.406e-02               2.741e+00              -6.589e-01               1.396e+00               3.877e-01
       DummyTechSupport        DummyStreamingTV    DummyStreamingMovies   DummyPaperlessBilling               intercept
              1.328e+00               2.012e+00               3.365e+00               3.579e-02                      NA

> 
```

A total of 34 variables (including Tenure): Tenure = 4.336e+00 (intercept) - 3.688e-01 (Children) + 4.046e-02 (Age) – 1.366e-07 (Income) + 4.506e-03 (Outage_sec_perweek) – 6.203e-04 (Email) – 2.388e-02 (Contacts) – 1.846e-02 (Yearly_equip_failure) – 1.116e-01 (MonthlyCharge) + 2.213e-02 (Bandwidth_GB_Year) - -1.958e-03 (TimelyResponse) – 4.793e-03 (TimelyFixes) + 2.976e-02 (TimelyReplacements) +

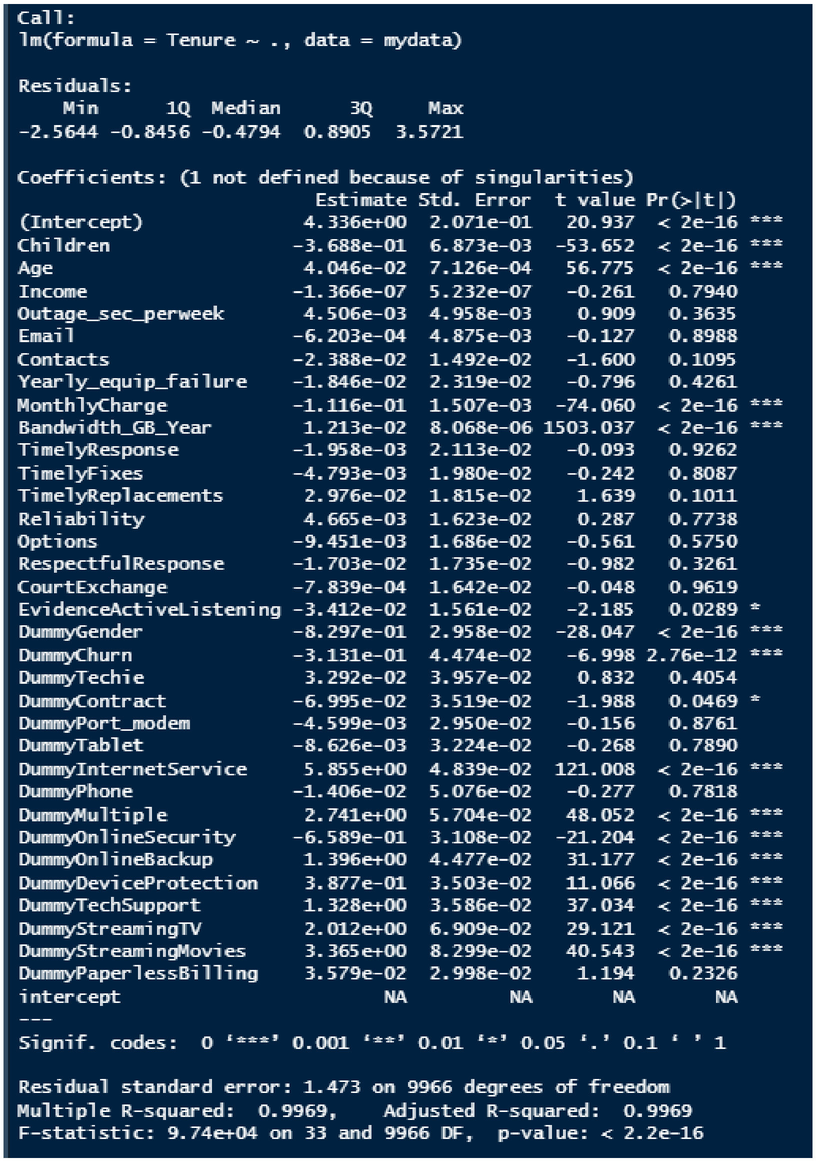
4.665e-03 (Reliability) – 9.451e-03 (Options) – 1.703e-0 (RepectfulResponse) – 7.829e-04 (CourtExchange) – 3.412e-02 (EvidenceActiveListening) – 8.297e-01 (DummyGender) – 3.131e-01 (DummyChurn) + 3.292e-0-2 (DummyTechie) – 6.995e-02 (DummyContract) – 4.599e-03 (DummyPort_modem) – 8.626e-03 (DummyTablet) + 5.855 (DummyInternetService) – 1.406e-02 (DummyPhone) + 2.741 (DummyMultiple) – 0.6589 (DummyOnlineSecurity) + 1.396 (DummyOnlineBackup) + 3.877e-01 (DummyDeviceProtection) + 1.328 (DummyTechSupport) + 2.012 (DummyStreamingTV) + 3.365 (DummyStreamingMovies) + 3.579e-02 (DummyPaperlessBilling)

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

Code:

```
summary(lm_all_variables)
```

Output:



```
Call:
lm(formula = Tenure ~ ., data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5644 -0.8456 -0.4794  0.8905  3.5721 

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error  t value Pr(>|t|)    
(Intercept)              4.336e+00  2.071e-01   20.937  < 2e-16 ***
Children                -3.688e-01  6.873e-03  -53.652  < 2e-16 ***
Age                      4.046e-02  7.126e-04   56.775  < 2e-16 ***
Income                  -1.366e-07  5.232e-07   -0.261   0.7940    
Outage_sec_perweek       4.506e-03  4.958e-03    0.909   0.3635    
Email                   -6.203e-04  4.875e-03   -0.127   0.8988    
Contacts                -2.388e-02  1.492e-02   -1.600   0.1095    
Yearly_equip_failure    -1.846e-02  2.319e-02   -0.796   0.4261    
MonthlyCharge           -1.116e-01  1.507e-03  -74.060  < 2e-16 ***
Bandwidth_GB_Year        1.213e-02  8.068e-06 1503.037  < 2e-16 ***
TimelyResponse          -1.958e-03  2.113e-02   -0.093   0.9262    
TimelyFixes             -4.793e-03  1.980e-02   -0.242   0.8087    
TimelyReplacements       2.976e-02  1.815e-02    1.639   0.1011    
Reliability              4.665e-03  1.623e-02    0.287   0.7738    
Options                 -9.451e-03  1.686e-02   -0.561   0.5750    
RespectfulResponse      -1.703e-02  1.735e-02   -0.982   0.3261    
CourtExchange           -7.839e-04  1.642e-02   -0.048   0.9619    
EvidenceActiveListening -3.412e-02  1.561e-02   -2.185   0.0289 *  
DummyGender             -8.297e-01  2.958e-02  -28.047  < 2e-16 ***
DummyChurn              -3.131e-01  4.474e-02   -6.998 2.76e-12 ***
DummyTechie              3.292e-02  3.957e-02    0.832   0.4054    
DummyContract           -6.995e-02  3.519e-02   -1.988   0.0469 *  
DummyPort_modem         -4.599e-03  2.950e-02   -0.156   0.8761    
DummyTablet             -8.626e-03  3.224e-02   -0.268   0.7890    
DummyInternetService     5.855e+00  4.839e-02  121.008  < 2e-16 ***
DummyPhone              -1.406e-02  5.076e-02   -0.277   0.7818    
DummyMultiple            2.741e+00  5.704e-02   48.052  < 2e-16 ***
DummyOnlineSecurity     -6.589e-01  3.108e-02  -21.204  < 2e-16 ***
DummyOnlineBackup        1.396e+00  4.477e-02   31.177  < 2e-16 ***
DummyDeviceProtection    3.877e-01  3.503e-02   11.066  < 2e-16 ***
DummyTechSupport         1.328e+00  3.586e-02   37.034  < 2e-16 ***
DummyStreamingTV         2.012e+00  6.909e-02   29.121  < 2e-16 ***
DummyStreamingMovies     3.365e+00  8.299e-02   40.543  < 2e-16 ***
DummyPaperlessBilling    3.579e-02  2.998e-02    1.194   0.2326    
intercept                       NA         NA       NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.473 on 9966 degrees of freedom
Multiple R-squared:  0.9969,	Adjusted R-squared:  0.9969 
F-statistic: 9.74e+04 on 33 and 9966 DF,  p-value: < 2.2e-16
```

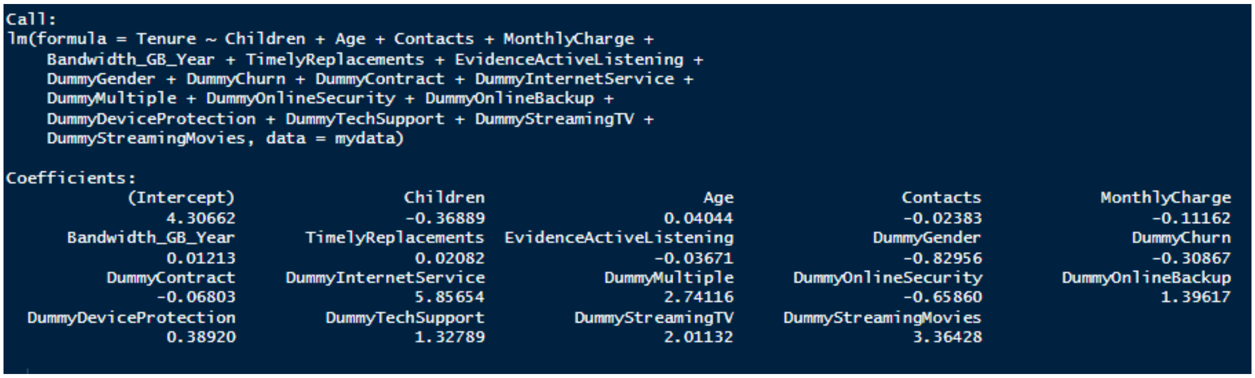
There is an R-Squared value of 0.9969 (multiple and adjusted). This shows that 99.69% of the variation within the data frame is within the model. This is extremely high and may be explained by suggesting the model has strong multicollinearity. If this is true, a reduction of the variable may help determine the exact cause of this.

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

R provides a function that allows for a stepwise regression:

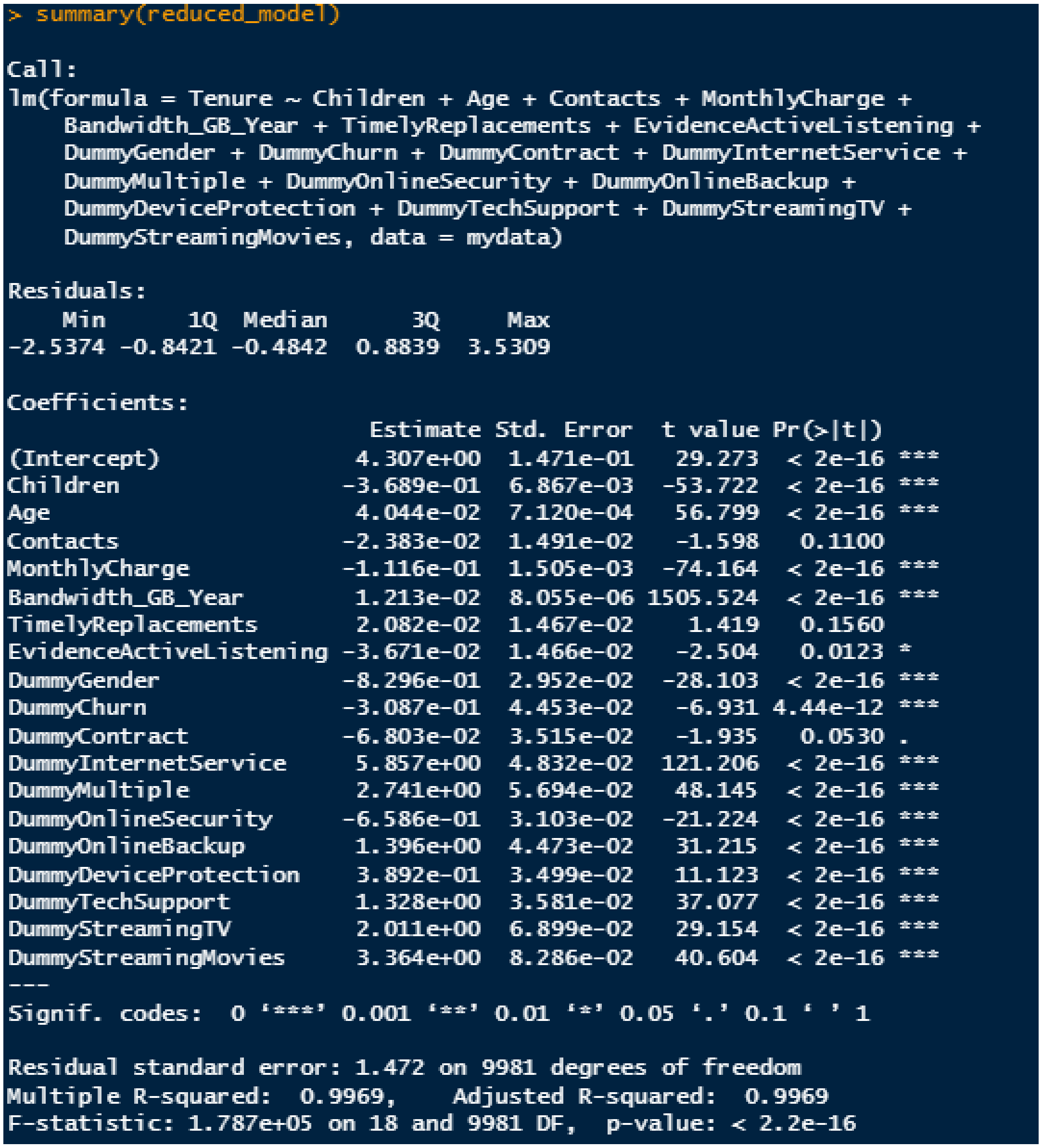
```
# Reducing the model backwards

        reduced_model <- step(lm_all_variables, direction = "backward")
```

The results of this reduced the model from 34 variables to 19 variables. Of the 19, Tenure (the dependent variable) is included with 18 independent variables: Children, Age, Contact, MonthlyCharge, Bandwidth_GB_Year, TimelyReplacements, EvidenceActiveListening, DummyGender, DummyChurn, DummyContract, DummyInternetService, DummyMultiple, DummyOnlineSecruity, DummyOnlineBackup, DummyDeviceProtect, DummyTechSupport, DummyStreamingTV, and DummyStreamingMovies.



```
Call:
lm(formula = Tenure ~ Children + Age + Contacts + MonthlyCharge +
    Bandwidth_GB_Year + TimelyReplacements + EvidenceActiveListening +
    DummyGender + DummyChurn + DummyContract + DummyInternetService +
    DummyMultiple + DummyOnlineSecurity + DummyOnlineBackup +
    DummyDeviceProtection + DummyTechSupport + DummyStreamingTV +
    DummyStreamingMovies, data = mydata)

Coefficients:
          (Intercept)                Children                      Age                 Contacts            MonthlyCharge
              4.30662                -0.36889                  0.04044                 -0.02383                 -0.11162
    Bandwidth_GB_Year      TimelyReplacements  EvidenceActiveListening              DummyGender               DummyChurn
              0.01213                 0.02082                 -0.03671                 -0.82956                 -0.30867
        DummyContract    DummyInternetService            DummyMultiple      DummyOnlineSecurity        DummyOnlineBackup
             -0.06803                 5.85654                  2.74116                 -0.65860                  1.39617
DummyDeviceProtection        DummyTechSupport         DummyStreamingTV     DummyStreamingMovies
              0.38920                 1.32789                  2.01132                  3.36428
```

The following is the summary:



```
> summary(reduced_model)

Call:
lm(formula = Tenure ~ Children + Age + Contacts + MonthlyCharge + 
    Bandwidth_GB_Year + TimelyReplacements + EvidenceActiveListening + 
    DummyGender + DummyChurn + DummyContract + DummyInternetService + 
    DummyMultiple + DummyOnlineSecurity + DummyOnlineBackup + 
    DummyDeviceProtection + DummyTechSupport + DummyStreamingTV + 
    DummyStreamingMovies, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5374 -0.8421 -0.4842  0.8839  3.5309 

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)    
(Intercept)              4.307e+00  1.471e-01   29.273  < 2e-16 ***
Children                -3.689e-01  6.867e-03  -53.722  < 2e-16 ***
Age                      4.044e-02  7.120e-04   56.799  < 2e-16 ***
Contacts                -2.383e-02  1.491e-02   -1.598   0.1100    
MonthlyCharge           -1.116e-01  1.505e-03  -74.164  < 2e-16 ***
Bandwidth_GB_Year        1.213e-02  8.055e-06 1505.524  < 2e-16 ***
TimelyReplacements       2.082e-02  1.467e-02    1.419   0.1560    
EvidenceActiveListening -3.671e-02  1.466e-02   -2.504   0.0123 *  
DummyGender             -8.296e-01  2.952e-02  -28.103  < 2e-16 ***
DummyChurn              -3.087e-01  4.453e-02   -6.931 4.44e-12 ***
DummyContract           -6.803e-02  3.515e-02   -1.935   0.0530 .  
DummyInternetService     5.857e+00  4.832e-02  121.206  < 2e-16 ***
DummyMultiple            2.741e+00  5.694e-02   48.145  < 2e-16 ***
DummyOnlineSecurity     -6.586e-01  3.103e-02  -21.224  < 2e-16 ***
DummyOnlineBackup        1.396e+00  4.473e-02   31.215  < 2e-16 ***
DummyDeviceProtection    3.892e-01  3.499e-02   11.123  < 2e-16 ***
DummyTechSupport         1.328e+00  3.581e-02   37.077  < 2e-16 ***
DummyStreamingTV         2.011e+00  6.899e-02   29.154  < 2e-16 ***
DummyStreamingMovies     3.364e+00  8.286e-02   40.604  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.472 on 9981 degrees of freedom
Multiple R-squared:  0.9969,	Adjusted R-squared:  0.9969 
F-statistic: 1.787e+05 on 18 and 9981 DF,  p-value: < 2.2e-16
```

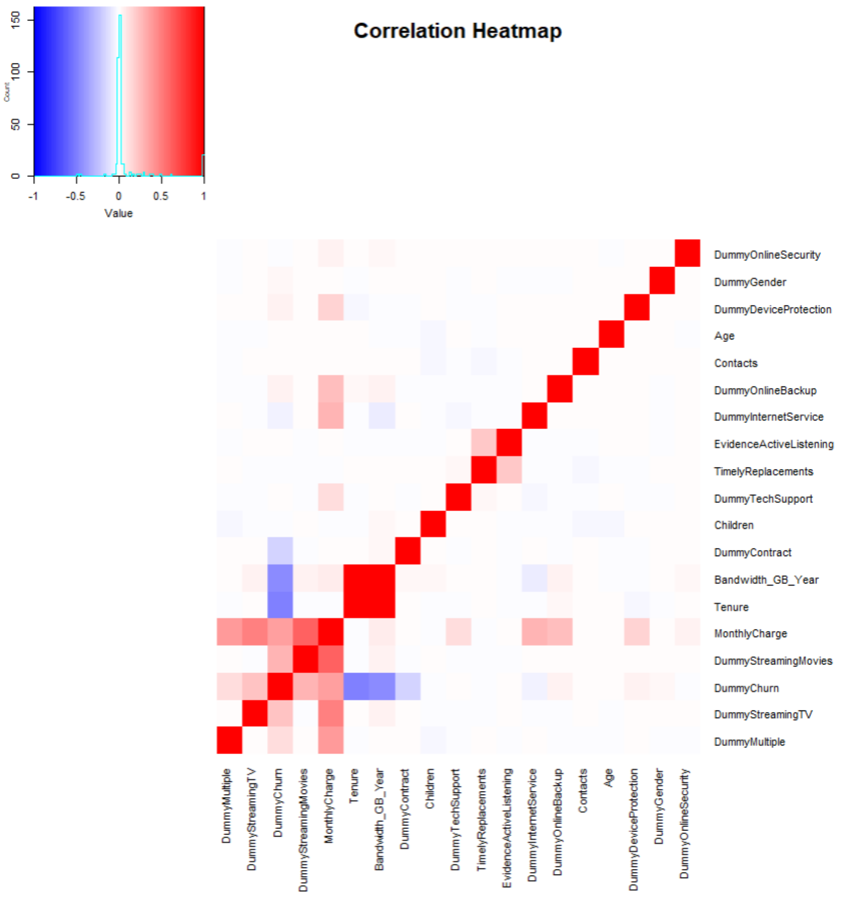
A correlation heatmap can provide further insights:

Code:

# Listing variables to go into heatmap

heatmap_data <- mydata[, c('Tenure', 'Children', 'Age', 'Contacts', 'MonthlyCharge', 'Bandwidth_GB_Year', 'TimelyReplacements', 'EvidenceActiveListening', DummyGender', 'DummyChurn', 'DummyContract', 'DummyInternetService', 'DummyMultiple', 'DummyOnlineSecurity', 'DummyOnlineBackup', DummyDeviceProtection', 'DummyTechSupport', 'DummyStreamingTV', 'DummyStreamingMovies')]

```
# Creating the heatmap

heatmap.2(cor_matrix,
        trace = "none",
        col = colorRampPalette(c("blue", "white", "red"))(100),
        main = "Correlation Heatmap",
        key.title = NA,
        cexRow = 0.9, cexCol = 0.9,
        margins = c(10, 10),
        dendrogram = "none")
```



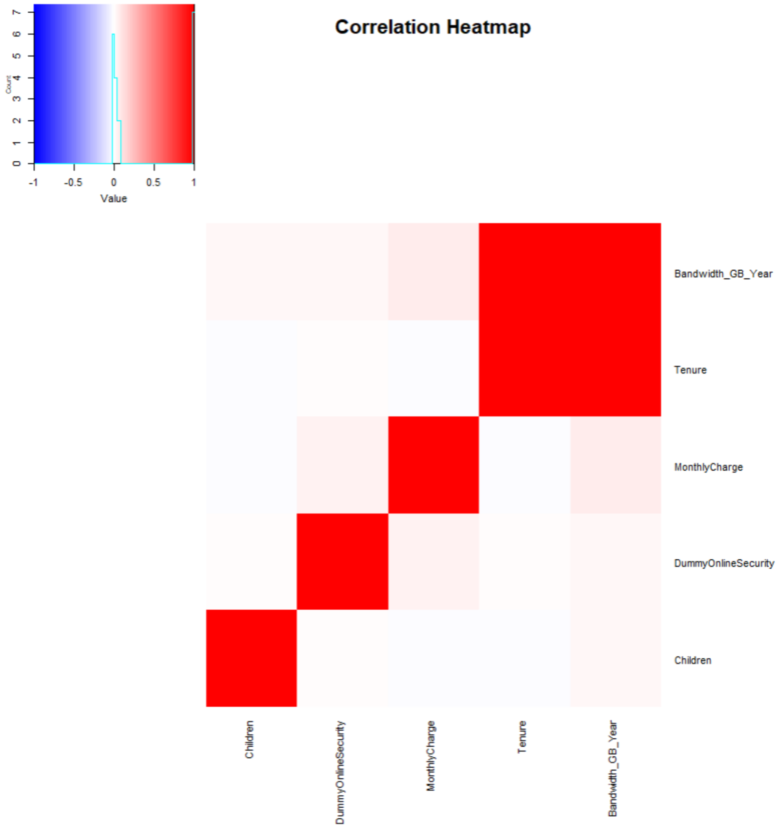
As seen in the above heatmap, Tenure and Bandwidth_GB_Year have a strong positive correlation.

To reduce it further and remove those variables that do not seem to have any correlation:

Code:

```
# Reduced heatmap
```

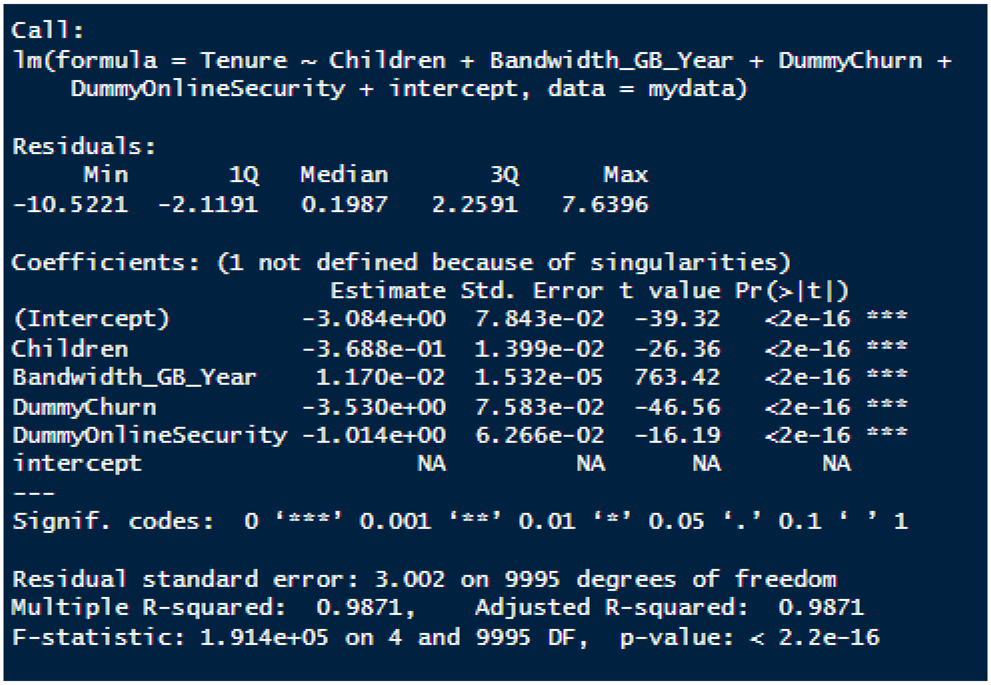
```
heatmap_data_2 <- mydata[, c('Tenure', 'Children', 'MonthlyCharge',
'Bandwidth_GB_Year', 'DummyOnlineSecurity')]

cor_matrix_2 <- cor(heatmap_data_2)

heatmap.2(cor_matrix_2,
          trace = "none",  # Remove dendrogram traces
          col = colorRampPalette(c("blue", "white", "red"))(100),  # Color palette
          main = "Correlation Heatmap",
          key.title = NA,  # Remove color key title
          cexRow = 0.9, cexCol = 0.9,  # Adjust label size
          margins = c(10, 10),  # Add some margin space
          dendrogram = "none"  # Remove both row and column dendrograms
)
```



It is evident that Tenure and Bandwidth have a strong positive correlation.

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

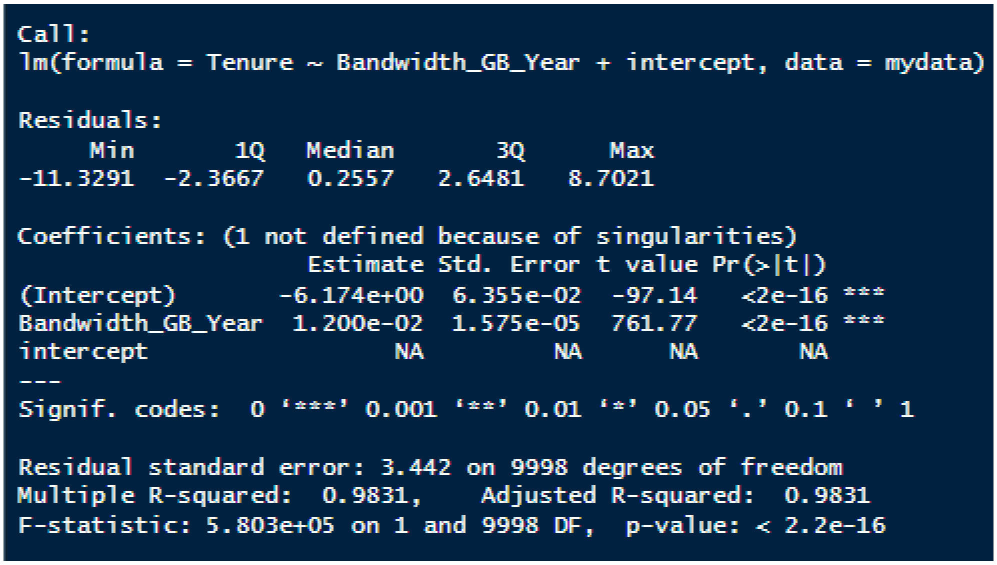
   After evaluating the models, it is clear that Tenure and Bandwidth have a strong correlation. Reducing the model to show the dependent variable Tenure and 4 independent variables (MonthlyCharge, Bandwidth_GB_Year, Children, and DummyOnlineSecurity) produced the following results:



```
Call:
lm(formula = Tenure ~ Children + Bandwidth_GB_Year + DummyChurn +
    DummyOnlineSecurity + intercept, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5221  -2.1191   0.1987   2.2591   7.6396

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.084e+00  7.843e-02  -39.32   <2e-16 ***
Children            -3.688e-01  1.399e-02  -26.36   <2e-16 ***
Bandwidth_GB_Year    1.170e-02  1.532e-05  763.42   <2e-16 ***
DummyChurn          -3.530e+00  7.583e-02  -46.56   <2e-16 ***
DummyOnlineSecurity -1.014e+00  6.266e-02  -16.19   <2e-16 ***
intercept                   NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.002 on 9995 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9871
F-statistic: 1.914e+05 on 4 and 9995 DF,  p-value: < 2.2e-16
```

These independent variables were chosen because they seemed to show some correlation in the heatmap and p-values.

However, reducing it even further to only Tenure and Bandwidth produced the following:



```
Call:
lm(formula = Tenure ~ Bandwidth_GB_Year + intercept, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3291  -2.3667   0.2557   2.6481   8.7021

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6.174e+00  6.355e-02  -97.14   <2e-16 ***
Bandwidth_GB_Year  1.200e-02  1.575e-05  761.77   <2e-16 ***
intercept                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.442 on 9998 degrees of freedom
Multiple R-squared:  0.9831,    Adjusted R-squared:  0.9831
F-statistic: 5.803e+05 on 1 and 9998 DF,  p-value: < 2.2e-16
```

As seen, by reducing the model to its most simplistic form of Tenure ~ Bandwidth, the R-Squared value is 0.9831 or 98.31%. However, with including a few other independent variables (Bandwidth, Children, DummyChurn and MonthlyCharge) the R-Squared value is 0.9871 or 98.71%. In addition, the reduction that includes more independent variables has a lower Residual Standard Error. While this is probably not that significant in terms of the models, it is interesting to note.

Nonetheless, it is evident that Bandwidth and Tenure have a direct linear relationship in this analysis.

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

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

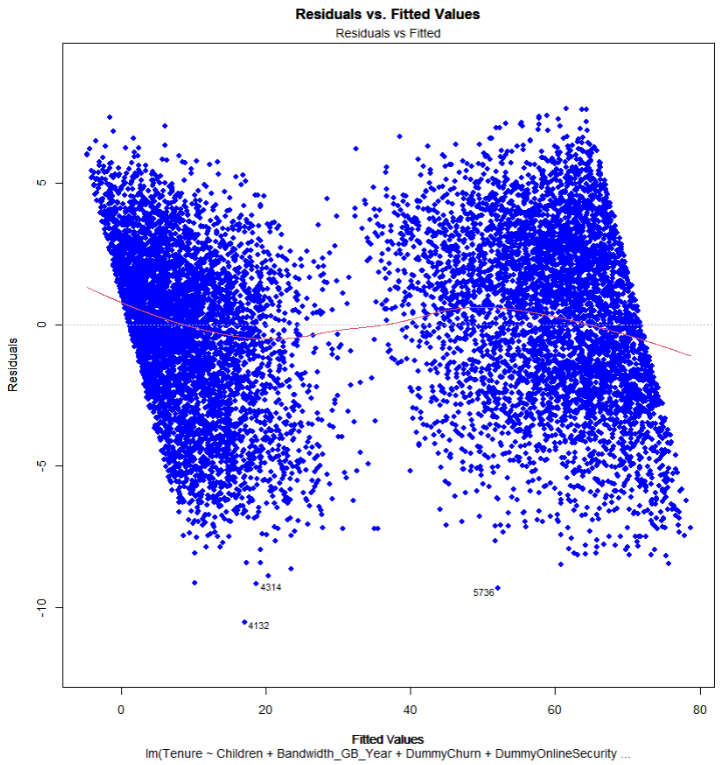
Model Evaluation Metric

The reduced model as seen above shows a high R-squared value of greater than 98%. Interestingly, the initial model had an R-squared value greater than 99%. While both models exhibited high correlation, the reduced model did suggest that both Tenure and Bandwidth have a linear relationship.

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

Residual Plot:

```
# Residual Plot:

par(mfrow = c(1, 1))  # Set up a single plot

# Create the residuals vs. fitted values plot

plot(lm_tenure_reduced, which = 1, col = "blue", pch = 16,
     main = "Residuals vs. Fitted Values")

# Add labels for readability

title(main = "Residuals vs. Fitted Values", xlab = "Fitted Values", ylab =
"Residuals")
```



The model’s residual standard error is Residual standard error: 3.002 on 9995 degrees of freedom.

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

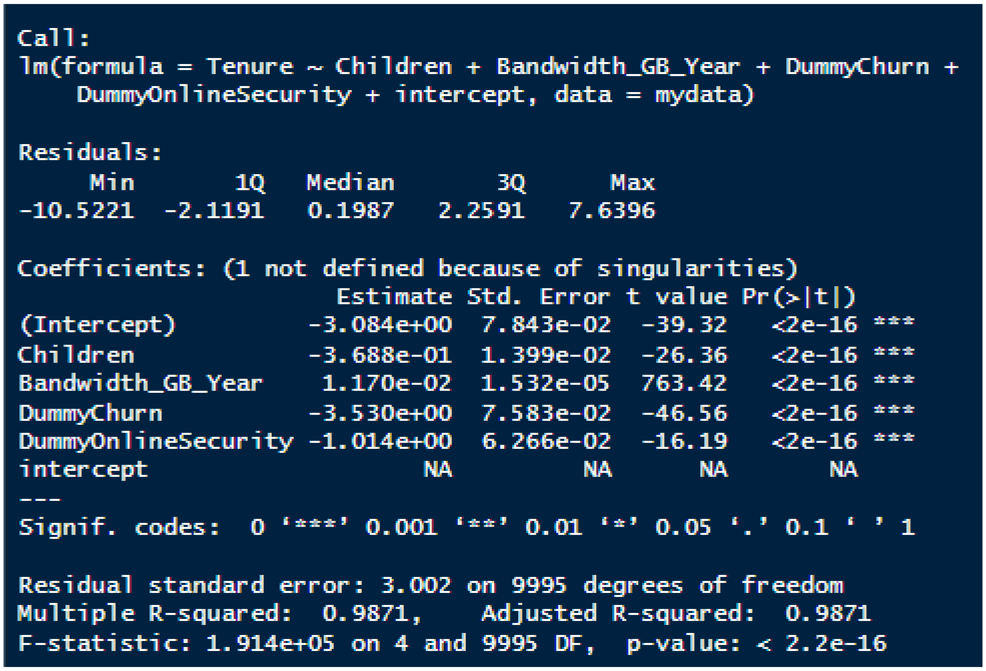
   Code will be provided in an R Source file and a .txt file attached to this assessment: task_1_code_R.txt. In addition, coding has been provided above.

Part V: Data Summary and Implications

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

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

   Regression equation for the reduced model:



```
Call:
lm(formula = Tenure ~ Children + Bandwidth_GB_Year + DummyChurn +
    DummyOnlineSecurity + intercept, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5221  -2.1191   0.1987   2.2591   7.6396

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3.084e+00  7.843e-02  -39.32   <2e-16 ***
Children            -3.688e-01  1.399e-02  -26.36   <2e-16 ***
Bandwidth_GB_Year    1.170e-02  1.532e-05  763.42   <2e-16 ***
DummyChurn          -3.530e+00  7.583e-02  -46.56   <2e-16 ***
DummyOnlineSecurity -1.014e+00  6.266e-02  -16.19   <2e-16 ***
intercept                   NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.002 on 9995 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9871
F-statistic: 1.914e+05 on 4 and 9995 DF,  p-value: < 2.2e-16
```

The final equation for the model is:

Tenure = -3.084 (intercept) – 0.3688 (Children) + 0.0117 (Bandwidth) – 3.530 (DummyChurn) – 1.014 (DummyOnlineSecurity)

Interpretation of the coefficients of the reduced model:

The coefficients suggest that for each additional increase in the units of Tenure a customer continues services:

- Customers will have a decrease in the number of children by -0.368.
- Customers will likely have increased average bandwidth by 0.0117 GBs
- Customers are likely to churn based on the predicted value of Tenure be decreased by 3.530 units. In other words, if a customer churn they likely have less Tenure credits.
- If customers have Online Security, they are expected to have 1.014 less Tenure units.

All coefficients are making the assumption that all other variables within the dataset are being held constant.

The statistical and practical significance of the reduced model

The reduced model suggests a few statistical and practical significance. First, the p-values (< 2.2e-16) is very small for the F-Statistic. This indicates at least one of the predictor variables is significantly correlated to the dependent variable (Tenure). Based on the heatmap and reduced model, it may be assumed Bandwidth_GB_Year has a significant impact. This is even more evident with the high t-value Bandwidth_GB_Year has and its associated small p-value. Suffice it to say, there is significance within the model.

The limitations of the data analysis

The main limitation of the analysis is the potential multicollinearity occurring between Tenure and Bandwidth_GB_Year. The high correlation can lead to the instability of the coefficients as they can deviate immensely based on which independent variables are within the model. Independent variables should be independent of one another and with such high correlation, it can cause problems with both the interpretation of the results and the fit of the model, which may have occurred with this analysis (Frost, 2023).

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

   The recommendation to the Telecommunication company is to focus their attention on customer service and maintaining their services with limited downtime to ensure customers do not churn. Customers use more bandwidth the longer they are tenured within the company. Ensure those customers do not have a reason to leave and maintain their business.

Part VI: Demonstration

G. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements:

Included in the upload of assessment.

- an identification of the version of the programming environment
- a comparison of the initial multiple linear regression model you used and the reduced linear regression model you used in your analysis
- an interpretation of the coefficients of the reduced model

Works Cited

Berry, W. B. (2005). *Probit/Logit and Other Binary Models*. Retrieved from Science Direct: https://www.sciencedirect.com/topics/social-sciences/multiple-regression#:~:text=Multiple%20regression%20is%20the%20most,the%20dependent%20variable%20is%20continuous.

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

Frost, J. (2023). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Retrieved from Statistics by Jim: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Glen, S. (2023). *Residual Values (Residuals) in Regression Analysis*. Retrieved from Statistics How To: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/residual/

Hayes, A. (2021, October 19). *Error Term: Definition, Example, and How to Calculate With Formula*. Retrieved December 1, 2023, from Investopia: https://www.investopedia.com/terms/e/errorterm.asp

JMP. (2023). *JMP*. Retrieved from Box Plot: https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis/box-plot.html#:~:text=Box%20plots%20highlight%20outliers,than%201.5%20times%20the%20IQR.

Kenton, W. (2022, December 31). *Homoskedastic: What It Means in Regression Modeling, With Example*. Retrieved December 1, 2023, from Investopia: https://www.investopedia.com/terms/h/homoskedastic.asp#:~:text=Homoskedastic%20(also%20spelled%20%22homoscedastic%22,of%20the%20predictor%20variable%20changes.

Reilly, J. (2023, January 26). *Data Science for Churn Prediction*. Retrieved December 1, 2023, from Akkio: https://www.akkio.com/post/churn-prediction

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