D208 Predictive Modeling

Adam Eyerman

2022-07-25

# Part I: Research Question

**A1)** **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 regression.**

How long will a customer’s tenure be with the company? Can any combination of variables predict this value accurately?

**A2) Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.**

The objective of the data analysis is to be able to predict the tenure of a customer using a variety of explanatory variables. If a linear regression equation can predict tenure within a reasonable certainty, then the stakeholders will benefit from understanding the variables impacting customer retention over time. Adjustments can then be made to any positive or negative correlations for the variables within the regression model to maximize tenure of customers.

# Part II: Method Justification

**B1)** **Summarize the assumptions of a multiple regression model.**

According to (“The Five Assumptions of Multiple Linear Regression”, 2021) we must make sure that five assumptions are met when performing a multiple linear regression.

* A linear relationship exists between each independent variable and the dependent variable.
* None of the independent variables are highly correlated
* Each observation is independent
* Residuals have the same variance at every point in model
* Residuals of the model have a normal distribution

If one assumption is violated, then the model may be unreliable.

**B2) Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.**

I will be using R and RStudio for my analysis using multiple linear regression. According to (“R: What Is R?”, n.d.) R provides a wide variety of statistical and graphical techniques and is highly extensible. This includes linear and non-linear models and the ability to utilize extensions through packages to further its capabilities. R offers an open source and easy to navigate environment for data wrangling, manipulating, cleaning, and analyzing. My analysis will involve all these tasks and R serves as great tool to gather data, prepare it for analysis, and run predictive models. Also, the R-Markdown feature (RMD) is an efficient way to organize all aspects of the data analysis process and creates either an HTML, PDF, or Word document to present the analysis.

**B3)** **Explain why multiple regression is an appropriate technique to analyze the research question summarized in Part I.**

Multiple linear regression is a predictive model used for estimating the relationship between two or more independent variables and a singular dependent variable (Bevans, 2022). The dependent variable is continuous and therefore is appropriate for this technique. If the dependent variable was binary, then a logistic regression would be utilized. The question involves more than one independent variable, a continuous dependent variable, and a need for a predictive model to be successfully analyzed. In my analysis I am attempting to find a model that can predict the tenure of a particular customer. The assumption will be that there is more than one explanatory variable needed to answer this question accurately. By formulating this equation, I can determine which variables positively and negatively impact the length of tenure. Multiple linear regression fits this description and can provide an equation to answer the research question.

# Part III: Data Preparation

**C1) Describe your data preparation goals and the data manipulations that will be used to achieve the goals.**

The main goal of the data preparation process will be to create a clean and reliable data set on all necessary predictor variables and the singular response variable. The first step will be to upload the churn data set and examine the columns for any NULL or missing values that may skew the data. If any NULL or missing values are found then a imputation process will be done to normalize the data accordingly using central tendencies of mean, median or mode. Next would be to examine for any outliers within the continuous variables. If found these values could be removed if they are clearly beyond the scope of the variable parameters. Then negative values found in variables that logically can only be positive to be valid will be switched to their absolute value. Also, the survey response categories (Item1-8) will be modified to their specific names for clarity. Since the variables being using for the model will include numeric, integer, binary and categorical values I will convert the binary and categorical variables into numeric values to utilize them for the multiple linear regression equation. For example, the Gender variable consists of “Male”, “Female” and “Nonbinary” for possible values. I would replace each value type with a number (0,1,2) which makes it possible to use within the regression model. Once the needed variables are cleaned and prepared, I will create a new data set (churn\_prepared) that will be used for developing the regression model.

**C2) Discuss the summary statistics, including the target variable and *all* predictor variables that you will need to gather from the data set to answer the research question.**

The churn\_clean data set consists of 50 variables and 10,000 observations. For the research question, the data set will be reduced to 34 variables to create a more targeted approach to the analysis. The removal of CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital, PaymentMethod is needed to consolidate the data set. The 34 remaining variables will comprise of numeric, integer, and categorical data. The categorical variables were converted to integer to include them in the linear regression model. These include Gender, Churn, Techie, Contract, Port\_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling. The Gender variable has 3 distinct values and will be converted to 0, 1, 2 respectively while the remaining variables will be converted to 0 or 1 based on the Yes or No response value. The numeric variables will be Tenure, Outage\_sec\_perweek, MonthlyCharge, Bandwidth\_GB\_Year, and Income. The integer variables will be Children, Age, Yearly\_equip\_failure, Email, Contacts, Timely\_response, Timely\_fixes, Timely\_replacements, Reliability, Options, Respectful\_response, Courteous\_exchange, and Active\_listening.

Preparation: Using this data set I will gather the summary statistics. Running the summary() function I will acquire the Minimum, 1st Quartile, Median, Mean, 3rd Quartile, and Maximum for each variable. Also both univariate and bivariate visualizations will be analyzed. This will provide information about the distribution of the variables within the data set. For the variables to be effective in the linear regression model the distribution must be normal.

Assessment: After assessing the summary statistics of the predictor variables the data showed no issues in regard to distribution. Details shown in C3 and C4.

**C3)** **Explain the steps used to prepare the data for the analysis, including the annotated code.**

The data set was prepared using the following steps:

* Load the data set “churn\_clean.csv”
* Install package “tidyverse” and load library accordingly
* Check for NULL or missing values
* Rename columns (Item1-8) to respective names for more clarity
* Change categorical variables to integer
* Replace categorical variables with integer version
* Remove unnecessary variables that will not be included in linear regression model
* View summary statistics using
* Display univariate visualizations for distribution analysis
* Display bivariate visualization for distribution analysis with target variable

#Install package “tidyverse” install.packages(“tidyverse”)

#Load library  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --  
## v ggplot2 3.3.6 v purrr 0.3.4  
## v tibble 3.1.7 v dplyr 1.0.9  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1  
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

#Load the data set "churn\_clean.csv"  
churn\_clean <- read.csv('churn\_clean.csv')  
  
#Check for NULL or missing values  
colSums(is.na(churn\_clean))

## CaseOrder Customer\_id Interaction   
## 0 0 0   
## UID City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## TimeZone Job Children   
## 0 0 0   
## Age Income Marital   
## 0 0 0   
## Gender Churn Outage\_sec\_perweek   
## 0 0 0   
## Email Contacts Yearly\_equip\_failure   
## 0 0 0   
## Techie Contract Port\_modem   
## 0 0 0   
## Tablet InternetService Phone   
## 0 0 0   
## Multiple OnlineSecurity OnlineBackup   
## 0 0 0   
## DeviceProtection TechSupport StreamingTV   
## 0 0 0   
## StreamingMovies PaperlessBilling PaymentMethod   
## 0 0 0   
## Tenure MonthlyCharge Bandwidth\_GB\_Year   
## 0 0 0   
## Item1 Item2 Item3   
## 0 0 0   
## Item4 Item5 Item6   
## 0 0 0   
## Item7 Item8   
## 0 0

#Rename columns (Item1-8) to respective names for more clarity  
churn\_updated <- churn\_clean %>% rename(Timely\_response = Item1, Timely\_fixes = Item2, Timely\_replacements = Item3, Reliability = Item4, Options = Item5, Respectful\_response = Item6, Courteous\_exchange = Item7, Active\_listening = Item8)   
  
#Change categorical variables to integer  
churn\_new <- churn\_clean$Churn # Replicating vector  
churn\_new <- as.character(churn\_new) # Converting factor to character  
churn\_new[churn\_new == "Yes"] <- 1 # Replacing Yes by 1  
churn\_new[churn\_new == "No"] <- 0 # Replacing No by 0  
churn\_new <- as.integer(churn\_new) # Converting character to numeric  
  
gender\_new <- churn\_clean$Gender # Replicating vector  
gender\_new <- as.character(gender\_new) # Converting factor to character  
gender\_new[gender\_new == "Male"] <- 2 # Replacing Yes by 2  
gender\_new[gender\_new == "Female"] <- 1 # Replacing No by 1  
gender\_new[gender\_new == "Nonbinary"] <- 0 # Replacing No by 0  
gender\_new <- as.integer(gender\_new) # Converting character to numeric  
  
  
techie\_new <- churn\_clean$Techie # Replicating vector  
techie\_new <- as.character(techie\_new) # Converting factor to character  
techie\_new[techie\_new == "Yes"] <- 1 # Replacing Yes by 1  
techie\_new[techie\_new == "No"] <- 0 # Replacing No by 0  
techie\_new <- as.integer(techie\_new) # Converting character to numeric  
   
  
paperless\_billing\_new <- churn\_clean$PaperlessBilling # Replicating vector  
paperless\_billing\_new <- as.character(paperless\_billing\_new)# Converting factor to character  
paperless\_billing\_new[paperless\_billing\_new == "Yes"] <- 1 # Replacing Yes by 1  
paperless\_billing\_new[paperless\_billing\_new == "No"] <- 0 # Replacing No by 0  
paperless\_billing\_new <- as.integer(paperless\_billing\_new)# Converting character to numeric  
   
  
streaming\_movies\_new <- churn\_clean$StreamingMovies # Replicating vector  
streaming\_movies\_new <- as.character(streaming\_movies\_new) # Converting factor to character  
streaming\_movies\_new[streaming\_movies\_new == "Yes"] <- 1 # Replacing Yes by 1  
streaming\_movies\_new[streaming\_movies\_new == "No"] <- 0 # Replacing No by 0  
streaming\_movies\_new <- as.integer(streaming\_movies\_new) # Converting character to numeric  
  
  
streaming\_tv\_new <- churn\_clean$StreamingTV # Replicating vector  
streaming\_tv\_new <- as.character(streaming\_tv\_new) # Converting factor to character  
streaming\_tv\_new[streaming\_tv\_new == "Yes"] <- 1 # Replacing Yes by 1  
streaming\_tv\_new[streaming\_tv\_new == "No"] <- 0 # Replacing No by 0  
streaming\_tv\_new <- as.integer(streaming\_tv\_new) # Converting character to numeric  
   
  
techsupport\_new <- churn\_clean$TechSupport # Replicating vector  
techsupport\_new <- as.character(techsupport\_new) # Converting factor to character  
techsupport\_new[techsupport\_new == "Yes"] <- 1 # Replacing Yes by 1  
techsupport\_new[techsupport\_new == "No"] <- 0 # Replacing No by 0  
techsupport\_new <- as.integer(techsupport\_new) # Converting character to numeric  
  
  
device\_protection\_new <- churn\_clean$DeviceProtection # Replicating vector  
device\_protection\_new <- as.character(device\_protection\_new) # Converting factor to character  
device\_protection\_new[device\_protection\_new == "Yes"] <- 1# Replacing Yes by 1  
device\_protection\_new[device\_protection\_new == "No"] <- 0 # Replacing No by 0  
device\_protection\_new <- as.integer(device\_protection\_new) # Converting character to numeric  
   
  
online\_backup\_new <- churn\_clean$OnlineBackup # Replicating vector  
online\_backup\_new <- as.character(online\_backup\_new) # Converting factor to character  
online\_backup\_new[online\_backup\_new == "Yes"] <- 1 # Replacing Yes by 1  
online\_backup\_new[online\_backup\_new == "No"] <- 0 # Replacing No by 0  
online\_backup\_new <- as.integer(online\_backup\_new) # Converting character to numeric  
   
  
online\_security\_new <- churn\_clean$OnlineSecurity # Replicating vector  
online\_security\_new <- as.character(online\_security\_new) # Converting factor to character  
online\_security\_new[online\_security\_new == "Yes"] <- 1 # Replacing Yes by 1  
online\_security\_new[online\_security\_new == "No"] <- 0 # Replacing No by 0  
online\_security\_new <- as.integer(online\_security\_new) # Converting character to numeric  
   
  
multiple\_new <- churn\_clean$Multiple # Replicating vector  
multiple\_new <- as.character(multiple\_new) # Converting factor to character  
multiple\_new[multiple\_new == "Yes"] <- 1 # Replacing Yes by 1  
multiple\_new[multiple\_new == "No"] <- 0 # Replacing No by 0  
multiple\_new <- as.integer(multiple\_new) # Converting character to numeric  
  
  
phone\_new <- churn\_clean$Phone # Replicating vector  
phone\_new <- as.character(phone\_new) # Converting factor to character  
phone\_new[phone\_new == "Yes"] <- 1 # Replacing Yes by 1  
phone\_new[phone\_new == "No"] <- 0 # Replacing No by 0  
phone\_new <- as.integer(phone\_new) # Converting character to numeric  
  
  
tablet\_new <- churn\_clean$Tablet # Replicating vector  
tablet\_new <- as.character(tablet\_new) # Converting factor to character  
tablet\_new[tablet\_new == "Yes"] <- 1 # Replacing Yes by 1  
tablet\_new[tablet\_new == "No"] <- 0 # Replacing No by 0  
tablet\_new <- as.integer(tablet\_new) # Converting character to numeric  
  
  
port\_modem\_new <- churn\_clean$Port\_modem # Replicating vector  
port\_modem\_new <- as.character(port\_modem\_new) # Converting factor to character  
port\_modem\_new[port\_modem\_new == "Yes"] <- 1 # Replacing Yes by 1  
port\_modem\_new[port\_modem\_new == "No"] <- 0 # Replacing No by 0  
port\_modem\_new <- as.integer(port\_modem\_new) # Converting character to numeric  
   
  
contract\_new <- churn\_clean$Contract # Replicating vector  
contract\_new <- as.character(contract\_new) # Converting factor to character  
contract\_new[contract\_new == "Two Year"] <- 2 # Replacing Yes by 2  
contract\_new[contract\_new == "One year"] <- 1 # Replacing No by 1  
contract\_new[contract\_new == "Month-to-month"] <- 0 # Replacing No by 0  
contract\_new <- as.integer(contract\_new) # Converting character to numeric  
   
  
internet\_new <- churn\_clean$InternetService # Replicating vector  
internet\_new <- as.character(internet\_new) # Converting factor to character  
internet\_new[internet\_new == "Fiber Optic"] <- 2 # Replacing Yes by 2  
internet\_new[internet\_new == "DSL"] <- 1 # Replacing No by 1  
internet\_new[internet\_new == "None"] <- 0 # Replacing No by 0  
internet\_new <- as.integer(internet\_new) # Converting character to numeric  
   
  
#Replace categorical variables with integer version  
churn\_updated$Churn <- churn\_new  
churn\_updated$Contract <- contract\_new  
churn\_updated$DeviceProtection <- device\_protection\_new  
churn\_updated$Gender <- gender\_new  
churn\_updated$InternetService <- internet\_new  
churn\_updated$Multiple <- multiple\_new  
churn\_updated$OnlineSecurity <- online\_security\_new  
churn\_updated$OnlineBackup <- online\_backup\_new  
churn\_updated$PaperlessBilling <- paperless\_billing\_new  
churn\_updated$Phone <- phone\_new  
churn\_updated$Port\_modem <- port\_modem\_new  
churn\_updated$StreamingTV <- streaming\_tv\_new  
churn\_updated$StreamingMovies <- streaming\_movies\_new  
churn\_updated$Tablet <- tablet\_new  
churn\_updated$Techie <- techie\_new  
churn\_updated$TechSupport <- techsupport\_new  
  
#Remove unnecessary variables that will not be included in linear regression model  
churn\_prepared <- subset(churn\_updated, select = -c(CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital, PaymentMethod))  
   
#View summary statistics using  
summary(churn\_prepared)

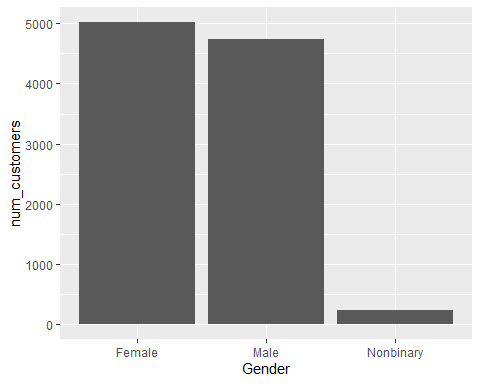
## Children Age Income Gender   
## Min. : 0.000 Min. :18.00 Min. : 348.7 Min. :0.000   
## 1st Qu.: 0.000 1st Qu.:35.00 1st Qu.: 19224.7 1st Qu.:1.000   
## Median : 1.000 Median :53.00 Median : 33170.6 Median :1.000   
## Mean : 2.088 Mean :53.08 Mean : 39806.9 Mean :1.451   
## 3rd Qu.: 3.000 3rd Qu.:71.00 3rd Qu.: 53246.2 3rd Qu.:2.000   
## Max. :10.000 Max. :89.00 Max. :258900.7 Max. :2.000   
## Churn Outage\_sec\_perweek Email Contacts   
## Min. :0.000 Min. : 0.09975 Min. : 1.00 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.: 8.01821 1st Qu.:10.00 1st Qu.:0.0000   
## Median :0.000 Median :10.01856 Median :12.00 Median :1.0000   
## Mean :0.265 Mean :10.00185 Mean :12.02 Mean :0.9942   
## 3rd Qu.:1.000 3rd Qu.:11.96949 3rd Qu.:14.00 3rd Qu.:2.0000   
## Max. :1.000 Max. :21.20723 Max. :23.00 Max. :7.0000   
## Yearly\_equip\_failure Techie Contract Port\_modem   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.398 Mean :0.1679 Mean :0.6986 Mean :0.4834   
## 3rd Qu.:1.000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :6.000 Max. :1.0000 Max. :2.0000 Max. :1.0000   
## Tablet InternetService Phone Multiple   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.0000 Median :1.000 Median :1.0000 Median :0.0000   
## Mean :0.2991 Mean :1.228 Mean :0.9067 Mean :0.4608   
## 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :2.000 Max. :1.0000 Max. :1.0000   
## OnlineSecurity OnlineBackup DeviceProtection TechSupport   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.000   
## Mean :0.3576 Mean :0.4506 Mean :0.4386 Mean :0.375   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.000   
## StreamingTV StreamingMovies PaperlessBilling Tenure   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. : 1.000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.: 7.918   
## Median :0.0000 Median :0.000 Median :1.0000 Median :35.431   
## Mean :0.4929 Mean :0.489 Mean :0.5882 Mean :34.526   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:61.480   
## Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :71.999   
## MonthlyCharge Bandwidth\_GB\_Year Timely\_response Timely\_fixes   
## Min. : 79.98 Min. : 155.5 Min. :1.000 Min. :1.000   
## 1st Qu.:139.98 1st Qu.:1236.5 1st Qu.:3.000 1st Qu.:3.000   
## Median :167.48 Median :3279.5 Median :3.000 Median :4.000   
## Mean :172.62 Mean :3392.3 Mean :3.491 Mean :3.505   
## 3rd Qu.:200.73 3rd Qu.:5586.1 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :290.16 Max. :7159.0 Max. :7.000 Max. :7.000   
## Timely\_replacements Reliability Options Respectful\_response  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :3.000 Median :3.000 Median :3.000 Median :3.000   
## Mean :3.487 Mean :3.498 Mean :3.493 Mean :3.497   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :8.000 Max. :7.000 Max. :7.000 Max. :8.000   
## Courteous\_exchange Active\_listening  
## Min. :1.00 Min. :1.000   
## 1st Qu.:3.00 1st Qu.:3.000   
## Median :4.00 Median :3.000   
## Mean :3.51 Mean :3.496   
## 3rd Qu.:4.00 3rd Qu.:4.000   
## Max. :7.00 Max. :8.000

**C4) Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.**

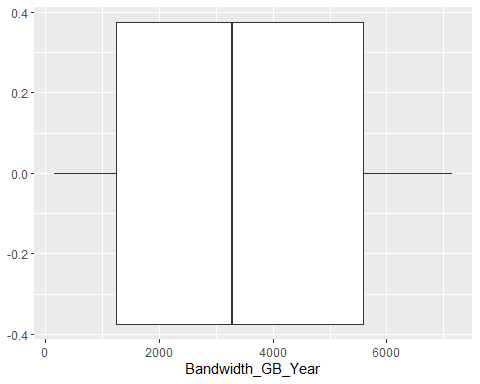
#Display univariate visualizations for distribution analysis  
num\_customers = 1  
  
#Churn  
churn\_plot <- ggplot(churn\_clean,  
 aes(Churn,num\_customers)) +  
 geom\_bar(stat = "identity")  
churn\_plot



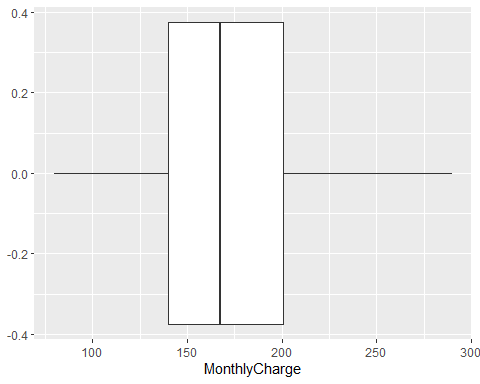
#Gender  
Gender\_plot <- ggplot(churn\_clean,  
 aes(Gender,num\_customers)) +  
 geom\_bar(stat = "identity")  
Gender\_plot



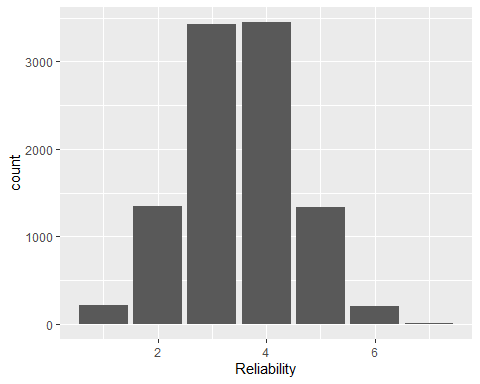
#Bandwidth  
Bandwidth\_plot <- ggplot(churn\_clean,  
 aes(Bandwidth\_GB\_Year)) +  
 geom\_boxplot()  
Bandwidth\_plot



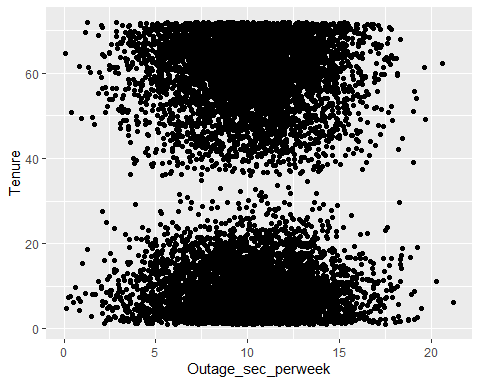
#MonthlyCharge  
MonthlyCharge\_plot <- ggplot(churn\_clean,  
 aes(MonthlyCharge)) +  
 geom\_boxplot()  
MonthlyCharge\_plot



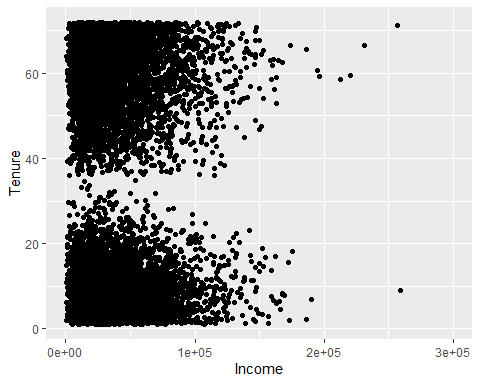
#Reliability  
Reliability\_plot <- ggplot(churn\_prepared,  
 aes(Reliability)) +  
 geom\_bar()  
Reliability\_plot



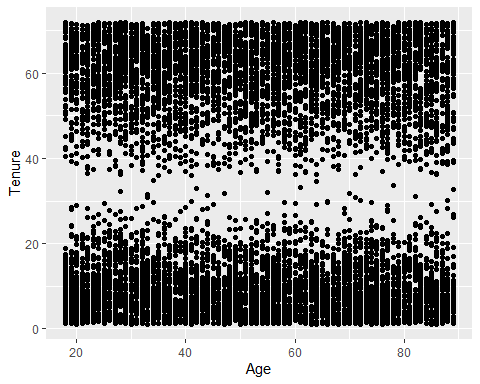
#Display bivariate visualization for distribution analysis with target variable  
  
#Outage(Bi)  
Outage\_biplot <- ggplot(churn\_clean,  
 aes(Outage\_sec\_perweek, Tenure)) +  
 geom\_point()   
Outage\_biplot



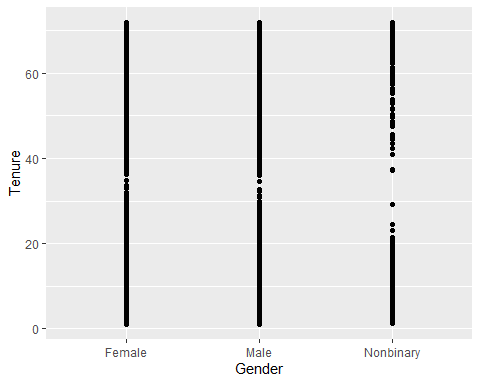
#Income(Bi)  
Income\_biplot <- ggplot(churn\_clean,  
 aes(Income, Tenure)) +  
 geom\_point() +  
 xlim(0,300000)  
Income\_biplot



#Age(Bi)  
Age\_biplot <- ggplot(churn\_clean,  
 aes(Age, Tenure)) +  
 geom\_point()   
Age\_biplot



#Gender(Bi)  
Gender\_biplot <- ggplot(churn\_clean,  
 aes(Gender, Tenure)) +  
 geom\_point()   
Gender\_biplot



**C5) Provide a copy of the prepared data set.**

File exported to csv file and converted to .xlsx and is attached as another file in submission.

#Code of Prepared Data Set  
library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

fwrite(churn\_prepared, "C:\\Users\\eyerm\\Dropbox\\PC\\Downloads\\churn\_prepared.csv")

"C:\Users\eyerm\Dropbox\PC\Downloads\churn\_prepared.csv"

*#View summary statistics using*  
summary(churn\_prepared)

## Children Age Income Gender   
## Min. : 0.000 Min. :18.00 Min. : 348.7 Min. :0.000   
## 1st Qu.: 0.000 1st Qu.:35.00 1st Qu.: 19224.7 1st Qu.:1.000   
## Median : 1.000 Median :53.00 Median : 33170.6 Median :1.000   
## Mean : 2.088 Mean :53.08 Mean : 39806.9 Mean :1.451   
## 3rd Qu.: 3.000 3rd Qu.:71.00 3rd Qu.: 53246.2 3rd Qu.:2.000   
## Max. :10.000 Max. :89.00 Max. :258900.7 Max. :2.000   
## Churn Outage\_sec\_perweek Email Contacts   
## Min. :0.000 Min. : 0.09975 Min. : 1.00 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.: 8.01821 1st Qu.:10.00 1st Qu.:0.0000   
## Median :0.000 Median :10.01856 Median :12.00 Median :1.0000   
## Mean :0.265 Mean :10.00185 Mean :12.02 Mean :0.9942   
## 3rd Qu.:1.000 3rd Qu.:11.96949 3rd Qu.:14.00 3rd Qu.:2.0000   
## Max. :1.000 Max. :21.20723 Max. :23.00 Max. :7.0000   
## Yearly\_equip\_failure Techie Contract Port\_modem   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.398 Mean :0.1679 Mean :0.6986 Mean :0.4834   
## 3rd Qu.:1.000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :6.000 Max. :1.0000 Max. :2.0000 Max. :1.0000   
## Tablet InternetService Phone Multiple   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :0.0000 Median :1.000 Median :1.0000 Median :0.0000   
## Mean :0.2991 Mean :1.228 Mean :0.9067 Mean :0.4608   
## 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :2.000 Max. :1.0000 Max. :1.0000   
## OnlineSecurity OnlineBackup DeviceProtection TechSupport   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.000   
## Mean :0.3576 Mean :0.4506 Mean :0.4386 Mean :0.375   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.000   
## StreamingTV StreamingMovies PaperlessBilling Tenure   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. : 1.000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.: 7.918   
## Median :0.0000 Median :0.000 Median :1.0000 Median :35.431   
## Mean :0.4929 Mean :0.489 Mean :0.5882 Mean :34.526   
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:61.480   
## Max. :1.0000 Max. :1.000 Max. :1.0000 Max. :71.999   
## MonthlyCharge Bandwidth\_GB\_Year Timely\_response Timely\_fixes   
## Min. : 79.98 Min. : 155.5 Min. :1.000 Min. :1.000   
## 1st Qu.:139.98 1st Qu.:1236.5 1st Qu.:3.000 1st Qu.:3.000   
## Median :167.48 Median :3279.5 Median :3.000 Median :4.000   
## Mean :172.62 Mean :3392.3 Mean :3.491 Mean :3.505   
## 3rd Qu.:200.73 3rd Qu.:5586.1 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :290.16 Max. :7159.0 Max. :7.000 Max. :7.000   
## Timely\_replacements Reliability Options Respectful\_response  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :3.000 Median :3.000 Median :3.000 Median :3.000   
## Mean :3.487 Mean :3.498 Mean :3.493 Mean :3.497   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :8.000 Max. :7.000 Max. :7.000 Max. :8.000   
## Courteous\_exchange Active\_listening  
## Min. :1.00 Min. :1.000   
## 1st Qu.:3.00 1st Qu.:3.000   
## Median :4.00 Median :3.000   
## Mean :3.51 Mean :3.496   
## 3rd Qu.:4.00 3rd Qu.:4.000   
## Max. :7.00 Max. :8.000

str(churn\_prepared)

## '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 ...  
## $ Gender : int 2 1 1 2 2 1 2 1 0 1 ...  
## $ Churn : int 0 1 0 0 1 0 1 1 0 0 ...  
## $ 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 : int 0 1 1 1 0 0 1 1 0 0 ...  
## $ Contract : int 1 0 2 2 0 1 0 0 0 2 ...  
## $ Port\_modem : int 1 0 1 0 1 1 0 0 1 1 ...  
## $ Tablet : int 1 1 0 0 0 0 0 0 0 0 ...  
## $ InternetService : int 2 2 1 1 2 0 1 1 1 2 ...  
## $ Phone : int 1 1 1 1 0 1 1 0 1 1 ...  
## $ Multiple : int 0 1 1 0 0 1 0 0 0 0 ...  
## $ OnlineSecurity : int 1 1 0 1 0 1 0 0 1 1 ...  
## $ OnlineBackup : int 1 0 0 0 0 1 0 1 1 0 ...  
## $ DeviceProtection : int 0 0 0 0 0 1 0 0 0 1 ...  
## $ TechSupport : int 0 0 0 0 1 0 1 0 0 0 ...  
## $ StreamingTV : int 0 1 0 1 1 0 1 0 0 0 ...  
## $ StreamingMovies : int 1 1 1 0 0 1 1 0 0 1 ...  
## $ PaperlessBilling : int 1 1 1 1 0 0 0 1 1 1 ...  
## $ 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 ...  
## $ Timely\_response : int 5 3 4 4 4 3 6 2 5 2 ...  
## $ Timely\_fixes : int 5 4 4 4 4 3 5 2 4 2 ...  
## $ Timely\_replacements : int 5 3 2 4 4 3 6 2 4 2 ...  
## $ Reliability : int 3 3 4 2 3 2 4 5 3 2 ...  
## $ Options : int 4 4 4 5 4 4 1 2 4 5 ...  
## $ Respectful\_response : int 4 3 3 4 4 3 5 3 3 2 ...  
## $ Courteous\_exchange : int 3 4 3 3 4 3 5 4 4 3 ...  
## $ Active\_listening : int 4 4 3 3 5 3 5 5 4 3 ...

# 

# Part IV: Model Comparison and Analysis

**D1)** **Construct an initial multiple regression model from *all* predictors that were identified in Part C2.**

***Initial Multiple Regression Model:***

Initial\_Model <- lm(Tenure ~ Children + Age + Income + gender\_new + churn\_new + Outage\_sec\_perweek + Email +  
 Contacts + Yearly\_equip\_failure + techie\_new + port\_modem\_new + tablet\_new +   
 internet\_new + phone\_new + multiple\_new + online\_backup\_new + online\_security\_new +   
 streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year + MonthlyCharge +   
 paperless\_billing\_new + contract\_new + device\_protection\_new + techsupport\_new +   
 Timely\_response + Timely\_fixes + Timely\_replacements + Reliability + Options +   
 Respectful\_response + Active\_listening + Courteous\_exchange, data = churn\_prepared)  
   
  
summary(Initial\_Model)

##   
## Call:  
## lm(formula = Tenure ~ Children + Age + Income + gender\_new +   
## churn\_new + Outage\_sec\_perweek + Email + Contacts + Yearly\_equip\_failure +   
## techie\_new + port\_modem\_new + tablet\_new + internet\_new +   
## phone\_new + multiple\_new + online\_backup\_new + online\_security\_new +   
## streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year +   
## MonthlyCharge + paperless\_billing\_new + contract\_new + device\_protection\_new +   
## techsupport\_new + Timely\_response + Timely\_fixes + Timely\_replacements +   
## Reliability + Options + Respectful\_response + Active\_listening +   
## Courteous\_exchange, data = churn\_prepared)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8347 -2.4516 0.8095 1.8636 4.2794   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.798e+00 3.307e-01 -11.486 < 2e-16 \*\*\*  
## Children -3.640e-01 1.077e-02 -33.794 < 2e-16 \*\*\*  
## Age 4.014e-02 1.117e-03 35.951 < 2e-16 \*\*\*  
## Income -1.284e-06 8.195e-07 -1.567 0.1171   
## gender\_new -7.324e-01 4.270e-02 -17.154 < 2e-16 \*\*\*  
## churn\_new -1.502e+00 7.080e-02 -21.215 < 2e-16 \*\*\*  
## Outage\_sec\_perweek 9.859e-03 7.767e-03 1.269 0.2043   
## Email 7.702e-04 7.638e-03 0.101 0.9197   
## Contacts -2.182e-02 2.338e-02 -0.934 0.3506   
## Yearly\_equip\_failure -2.534e-02 3.633e-02 -0.698 0.4855   
## techie\_new 5.937e-02 6.200e-02 0.958 0.3383   
## port\_modem\_new 4.544e-02 4.621e-02 0.983 0.3254   
## tablet\_new -4.015e-03 5.050e-02 -0.080 0.9366   
## internet\_new 3.222e-01 5.293e-02 6.088 1.18e-09 \*\*\*  
## phone\_new 1.664e-02 7.951e-02 0.209 0.8342   
## multiple\_new -1.373e+00 9.606e-02 -14.290 < 2e-16 \*\*\*  
## online\_backup\_new -1.461e+00 7.453e-02 -19.601 < 2e-16 \*\*\*  
## online\_security\_new -9.935e-01 4.873e-02 -20.387 < 2e-16 \*\*\*  
## streaming\_movies\_new -3.113e+00 1.431e-01 -21.761 < 2e-16 \*\*\*  
## streaming\_tv\_new -3.206e+00 1.186e-01 -27.036 < 2e-16 \*\*\*  
## Bandwidth\_GB\_Year 1.195e-02 1.255e-05 952.171 < 2e-16 \*\*\*  
## MonthlyCharge 1.865e-02 2.596e-03 7.185 7.18e-13 \*\*\*  
## paperless\_billing\_new 7.979e-02 4.695e-02 1.699 0.0893 .   
## contract\_new -2.080e-01 2.912e-02 -7.141 9.88e-13 \*\*\*  
## device\_protection\_new -1.172e+00 5.674e-02 -20.658 < 2e-16 \*\*\*  
## techsupport\_new -3.535e-01 5.764e-02 -6.134 8.90e-10 \*\*\*  
## Timely\_response 5.525e-02 3.310e-02 1.669 0.0951 .   
## Timely\_fixes -5.692e-02 3.101e-02 -1.836 0.0664 .   
## Timely\_replacements 2.829e-02 2.844e-02 0.995 0.3199   
## Reliability -1.218e-02 2.543e-02 -0.479 0.6319   
## Options -3.682e-02 2.640e-02 -1.394 0.1632   
## Respectful\_response -1.128e-02 2.718e-02 -0.415 0.6781   
## Active\_listening -5.552e-02 2.446e-02 -2.269 0.0233 \*   
## Courteous\_exchange -3.640e-03 2.572e-02 -0.142 0.8874   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.307 on 9966 degrees of freedom  
## Multiple R-squared: 0.9924, Adjusted R-squared: 0.9924   
## F-statistic: 3.951e+04 on 33 and 9966 DF, p-value: < 2.2e-16

***Initial Equation:***

**D2) Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.**

The procedure I used to reduce the model was p-value significance. The initial model consisted of 33 variables and was reduced to 15 variables. The 17 removed variables did not have p-values under the required 0.05 and therefore did not provide enough statistical relevance to be included in the reduction model. P-value is assessing the correlation between the predictor variable and the target variable. A value over 0.05 suggest the predictor variable has an independent relationship to the target variable and is not needed in the model for analyzing the research question.

**D3)** **Provide a reduced multiple regression model that includes *both* categorical and continuous variables.**

***Reduced Multiple Regression Model:***

Reduced\_Model <- lm(Tenure ~ Children + Age + gender\_new + internet\_new +   
 multiple\_new + online\_backup\_new + online\_security\_new +   
 streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year + MonthlyCharge +   
 contract\_new + device\_protection\_new + techsupport\_new +   
 Active\_listening, data = churn\_prepared)  
  
summary(Reduced\_Model)

##   
## Call:  
## lm(formula = Tenure ~ Children + Age + gender\_new + internet\_new +   
## multiple\_new + online\_backup\_new + online\_security\_new +   
## streaming\_movies\_new + streaming\_tv\_new + Bandwidth\_GB\_Year +   
## MonthlyCharge + contract\_new + device\_protection\_new + techsupport\_new +   
## Active\_listening, data = churn\_prepared)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4679 -2.7155 0.9777 1.8343 3.5316   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.945e+00 2.337e-01 -16.885 < 2e-16 \*\*\*  
## Children -3.687e-01 1.100e-02 -33.522 < 2e-16 \*\*\*  
## Age 4.033e-02 1.141e-03 35.349 < 2e-16 \*\*\*  
## gender\_new -7.631e-01 4.355e-02 -17.523 < 2e-16 \*\*\*  
## internet\_new 4.365e-01 5.381e-02 8.112 5.58e-16 \*\*\*  
## multiple\_new -1.376e+00 9.809e-02 -14.026 < 2e-16 \*\*\*  
## online\_backup\_new -1.435e+00 7.615e-02 -18.844 < 2e-16 \*\*\*  
## online\_security\_new -9.723e-01 4.976e-02 -19.541 < 2e-16 \*\*\*  
## streaming\_movies\_new -3.241e+00 1.460e-01 -22.206 < 2e-16 \*\*\*  
## streaming\_tv\_new -3.326e+00 1.210e-01 -27.492 < 2e-16 \*\*\*  
## Bandwidth\_GB\_Year 1.209e-02 1.084e-05 1115.387 < 2e-16 \*\*\*  
## MonthlyCharge 1.325e-02 2.639e-03 5.022 5.21e-07 \*\*\*  
## contract\_new -1.225e-02 2.824e-02 -0.434 0.6645   
## device\_protection\_new -1.162e+00 5.797e-02 -20.052 < 2e-16 \*\*\*  
## techsupport\_new -3.178e-01 5.885e-02 -5.399 6.85e-08 \*\*\*  
## Active\_listening -4.382e-02 2.294e-02 -1.910 0.0562 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.358 on 9984 degrees of freedom  
## Multiple R-squared: 0.9921, Adjusted R-squared: 0.992   
## F-statistic: 8.314e+04 on 15 and 9984 DF, p-value: < 2.2e-16

***Reduced Equation:***

**E1) Explain your data analysis process by comparing the initial and reduced multiple regression models, including the following elements:**

**•  the logic of the variable selection technique**

**•  the model evaluation metric**

**•  a residual plot**

Both models were constructed of predictor variables that included demographic, survey, and continuous data. However, the reduced model removed some of the redundancies within the initial model through the assessment of the P-value or significance of each variable on the regression equation. The initial model was reduced from 33 to 15 predictor variables by eliminating the variables with a P-value over 0.05. This consolidated the model down and maintained the R-squared value as well as the adjusted R-squared value, which accounts for variation explained using selected variables. Calling a summary function to the model displayed the necessary information to determine which variables would be removed using the column (Pr(>|t|) as the evaluation metric. As stated above, variables with values above 0.05 were removed for the reduced model.

The residual plots show a slightly more concise reading for the reduced model than the initial model, which is expected. However, both models have an odd distribution in terms of the density plot, Q-Q plot, and residual plot. It appears neither model is demonstrating a result that would suggest the linear model is appropriate.

These plots show a pattern and lack the randomness needed to provide confidence in the model. The density plot does not have a bell shape normal distribution rather a double peak which also suggests the models are not necessarily ideal. Lasty, the Q-Q plot shows wide deviation from the fitted line which aligns with the density plot and residual plot conclusions.

*#Initial Model Residual Plot*  
initial\_residual <- resid(Initial\_Model)  
plot(fitted(Initial\_Model), initial\_residual)

Chart

Description automatically generated

plot(density(initial\_residual))

Chart, histogram

Description automatically generated

qqnorm(initial\_residual)  
qqline(initial\_residual)

Chart, histogram

Description automatically generated

*#Reduced Model Residual Plot*  
reduced\_residual <- resid(Reduced\_Model)  
plot(fitted(Reduced\_Model), reduced\_residual)

A picture containing diagram

Description automatically generated

plot(density(reduced\_residual))

Chart, diagram, histogram

Description automatically generated

qqnorm(reduced\_residual)  
qqline(reduced\_residual)

Chart, diagram

Description automatically generated

**E2) Provide the output and *any* calculations of the analysis you performed, including the model’s residual error.**

***Note: The output should include the predictions from the refined model you used to perform the analysis.***

Code provided above in Part IV: D1-D3

**E3) Provide the code used to support the implementation of the multiple regression models.**

Code provided above in Part IV: D1-D3

**F1) Discuss the results of your data analysis, including the following elements:**

**•  a regression equation for the reduced model**

**•  an interpretation of coefficients of the statistically significant variables of the model**

**•  the statistical and practical significance of the model**

**•  the limitations of the data analysis**

***Reduced Equation:***

The coefficients of the statistically significant variables of the reduced regression model represent the positive or negative relationship to the target variable. For example, the more children a customer has the lower the Tenure appears to be based on the equation as the -0.37 value or intercept for the Children variable is negative. In opposition to that the Age variable has an intercept of 0.04 meaning that the higher the Age of a customer the higher the Tenure would be due to the positive value of 0.04. Evaluating the model, you can see the predictor variables show both negative and positive correlations to the target variable.

According to (“How to Interpret Adjusted R-Squared (With Examples)”,2022) the R-squared value is the proportion of the variance in the target variable that can be explained by the predictor variables in the model. The adjusted r-squared value considers the added predictor variables and if they increase or decrease the amount of variance accounted for. The model produced an adjusted r-squared value of 0.992 which demonstrates that the predictor variables account for nearly all the variance within the regression model. This value was slightly different in the initial model which was 0.9224. Maintaining the adjusted r-squared value after the reduction shows the reduced model maintains the relevance of the predictor variables while removing unnecessary or redundant variables. The linear regression model shows a statistical significance, but the residual plots suggest the analysis may not be ideal due to the lack of a true linear relationship between predictor variables and the target variable.

This analysis is limited by the amount of data collected and the ability to update the data set over time. With a larger set of data and a recurring update the analysis may be more accurate and relevant.

**F2) Recommend a course of action based on your results.**

My recommended course of action would be to evaluate the predictor variables in the reduced regression model and their relationship with the target variable. Age, internetnew, Bandwidth\_GB\_Year, and MonthlyCharge all have a positive relationship with the target variable (Tenure). The remaining predictor variables have a negative relationship with target variable. A further investigation would be needed to fully understand why these relationships exist. However, my analysis would suggest that more bandwidth, a higher monthly charge, the use of fiber optic internet and older individuals tend to have a higher tenure with the company. I would focus on marketing fiber optic internet as my main action moving forward. Fiber optic internet is a higher quality product, and the assumption would be made that customers with this service have higher bandwidth usage due to better quality. Also, fiber optic is assumed to cost more than DSL internet service leading to a higher monthly charge. Though the model residuals suggest the regression model is not ideal for predicting an exact tenure prediction, a few key predictor variables show a relationship with the target variable that can facilitate an actionable step that can lead to creating a higher tenured customer on average.

**G.  Provide a Panopto video recording that includes *all* of the following elements:**

**•  a demonstration of the functionality of the code used for the analysis**

**•  an identification of the version of the programming environment**

**•  a comparison of the two multiple regression models you used in your analysis**

**•  an interpretation of the coefficients.**

[Mon Jul 25 2022 3:05:42 PM (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=8a78640e-b30e-4ce5-a880-aedd013ac8ea)

**H.  List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.**

Z. (2021, November 16). *The Five Assumptions of Multiple Linear Regression*. Statology. <https://www.statology.org/multiple-linear-regression-assumptions/#:%7E:text=However%2C%20before%20we%20perform%20multiple%20linear%20regression%2C%20we,variables%20are%20highly%20correlated%20with%20each%20other.%203>.

*R: What is R?* (n.d.). R-Project. <https://www.r-project.org/about.html>

Bevans, R. (2022, June 1). *Multiple Linear Regression | A Quick Guide (Examples)*. Scribbr. <https://www.scribbr.com/statistics/multiple-linear-regression/#:%7E:text=%20The%20formula%20for%20a%20multiple%20linear%20regression,many%20independent%20variables%20you%20are%20testing%20More%20>

Z. (2022, March 24). *How to Interpret Adjusted R-Squared (With Examples)*. Statology. <https://www.statology.org/adjusted-r-squared-interpretation/>

**I.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**