­

**D208 Logistic Regression**

Nicole Reiswig

Western Governors University

Keiona Middleton

November 14, 2023

**Table of Contents**

**Part I: Research Question**

A.  1.  Research Question  3

A. 2.  Goals 3

**Part II: Method Justification**

B.  Logistic regression method 3

1.  Summary of assumptions 3

2.  Benefit of utilizing R 3

3.  Technique explanation 4

**Part III: Data Preparation**

C.  Summary of data preparation 4

1.  Data cleaning goals  4

2.  Summary statistics  18

3.  Univariate and bivariate visualizations  22

4.  Data transformation goals  39

5.  CSV file

**Part IV: Model Comparison and Analysis**

D.  Compare an initial and a reduced logistic regression model 46

1.  Initial logistic model  46

2.  Feature selection  49

3.  Reduced logistic regression 51

E.  Analysis of reduced logistic regression model

1.  Model evaluation metric 60

2.  Confusion matrix 61

3.  Code  65

**Part V: Data Summary and Implications**

F.  Summary of findings 65

1.  Discussion of results 65

2.  Recommended course of action 68

**Part VI: Demonstration**

G.  Panopto Recording 69

H.  Sources  70

# Logistic Regression Analysis

**Part I: Research Question**

A.

1.  The research question we are answering is, “What factors affect churn?” This is an important research question for the telecommunications company because we need to understand what factors influence customers to leave to retain their business.

2.  The logistic regression analysis is curvilinear and a machine learning classification algorithm.  The goal of this logistic regression analysis is to estimate if there is a relationship between the dependent and independent variables.

**Part II: Method Justification**

B.  A description of logistic regression methods provided by the following:

1.  Four assumptions of a logistic regression model are that the outcome is categorical such as yes/no or 1/0, the observations are independent of each other, there is a linear relationship between the logit and the outcome of the variables, and the errors do not need to be normally distributed, according to Kassambara, 2018. Logistic regression requires a large sample size.

2.  R is the programming language used for the logistic regression analysis. Two benefits of using R for this analysis are that (Western Governors University, 2023) it is very efficient because statistical models can be written with a few lines of code and R is great for statistical analysis and is often cited in academic journals. In this analysis, we utilize different packages in R such as dplyr, corrr, tidyverse, visdat, ggplot2, fastdummies, caret, leaps, mass, yardstick, and pscl. R was specifically created for statistics so many statistical functions are included in the base package of R which simplifies the code.

3.  Logistic regression is an appropriate technique to analyze the research question summarized in part I because it is a statistical tool to understand the relationship between categorical dependent variables and independent variables by estimating probabilities of outcomes, and our dependent variable is churn which is categorical yes/no and we will be measuring it against all other independent variables in the churn dataset to determine the probability that there is a relationship between the dependent and independent variables.

**Part III: Data Preparation**

C.  Summary of the data preparation methods:

1.  The data cleaning goal is to ensure no missing or null values, duplicates, outliers, or erroneous values. To achieve this goal, the steps used to clean the data are first to view and examine the data. We will detect duplicates and missing values. We will then detect outliers utilizing boxplot univariate visualizations. To clean the data we then retain, remove, or impute those values. See attached code.

#checking working directory

getwd()

#data profiling

churn\_clean208 <- read.csv("~/MSDA/churn\_clean208.csv")

str("~/MSDA/churn\_clean208")

#dimension of churn\_clean [in-text citation: (R programming 101, n.d.)]

dim(churn\_clean208)

library(tidyverse)

glimpse(churn\_clean208)

head(churn\_clean208)

summary(churn\_clean208)

#detect duplicates

duplicated("~/MSDA/churn\_clean208")

#sum of duplicated rows

sum(duplicated("~/MSDA/churn\_clean208"))

#detect missing values

colSums(is.na(churn\_clean208))

#visualize missing data

library(visdat)

vis\_miss(churn\_clean208)

#View data

churn\_clean208

#Boxplot of each variable - detect outliers - Univariate Graphs of each variable

boxplot(churn\_clean208$CaseOrder, xlab = "Case Order")

boxplot(churn\_clean208$Zip, xlab = "Zip Code")

boxplot(churn\_clean208$Lat, xlab = "Lat")

boxplot(churn\_clean208$Lng, xlab = "Lng")

boxplot(churn\_clean208$Population, xlab = "Population")

boxplot(churn\_clean208$Children, xlab = "Children")

boxplot(churn\_clean208$Age, xlab = "Age")

boxplot(churn\_clean208$Income, xlab = "Income")

boxplot(churn\_clean208$Outage\_sec\_perweek, xlab = "Outage Sec Per Week")

boxplot(churn\_clean208$Email, xlab = "Email")

boxplot(churn\_clean208$Contacts, xlab = "Contacts")

boxplot(churn\_clean208$Yearly\_equip\_failure, xlab = "Yearly Equip Failures")

boxplot(churn\_clean208$Tenure, xlab = "Tenure")

boxplot(churn\_clean208$MonthlyCharge, xlab = "Monthly Charge")

boxplot(churn\_clean208$Bandwidth\_GB\_Year, xlab = "Bandwidth GB Year")

#Retain Lat, Lng outliers expected

#Retain population to preserve sample size

#rename column names item 1- 8 [in-text citation: (Zach, 2022)]

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item1'] <- 'Timely\_Response'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item2'] <- 'Timely\_Fixes'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item3'] <- 'Timely\_Replacements'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item4'] <- 'Reliability'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item5'] <- 'Options'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item6'] <- 'Respectful\_Response'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item7'] <- 'Courteous\_Exchange'

colnames(churn\_clean208)[colnames(churn\_clean208) == 'Item8'] <- 'Active\_Listening'

#Verify columns were re-named successfully

glimpse(churn\_clean208)

#Histograms of variables with outliers- Univariate graphs

hist(churn\_clean208$Population)

hist(churn\_clean208$Children)

hist(churn\_clean208$Income)

hist(churn\_clean208$Outage\_sec\_perweek)

hist(churn\_clean208$Email)

hist(churn\_clean208$Contacts)

hist(churn\_clean208$Yearly\_equip\_failure)

hist(churn\_clean208$Timely\_Response)

hist(churn\_clean208$Timely\_Fixes)

hist(churn\_clean208$Timely\_Replacements)

hist(churn\_clean208$Reliability)

hist(churn\_clean208$Options)

hist(churn\_clean208$Respectful\_Response)

hist(churn\_clean208$Courteous\_Exchange)

hist(churn\_clean208$Active\_Listening)

#explore data variables- univariate graphs- summary statistics of categorical data [in-text citation: (R programming 101, n.d.)]

barplot(sort(table(churn\_clean208$Area)), xlab = "Area")

barplot(sort(table(churn\_clean208$TimeZone)), xlab = "Timezone")

barplot(sort(table(churn\_clean208$Children)), xlab = "Number of Children")

barplot(sort(table(churn\_clean208$Age)), xlab = "Age")

barplot(sort(table(churn\_clean208$Income)), xlab = "Income")

barplot(sort(table(churn\_clean208$Marital)), xlab = "Marital")

barplot(sort(table(churn\_clean208$Gender)), xlab = "Gender")

barplot(sort(table(churn\_clean208$Churn)), xlab = "Churn")

barplot(sort(table(churn\_clean208$Outage\_sec\_perweek)), xlab = "Outage Sec Per Week")

barplot(sort(table(churn\_clean208$Email)), xlab = "Email")

barplot(sort(table(churn\_clean208$Contacts)), xlab = "Contacts")

barplot(sort(table(churn\_clean208$Yearly\_equip\_failure)), xlab = "Yearly Equipment Failure")

barplot(sort(table(churn\_clean208$Techie)), xlab = "Techie")

barplot(sort(table(churn\_clean208$Contract)), xlab = "Contracts")

barplot(sort(table(churn\_clean208$Port\_modem)), xlab = "Port Modem")

barplot(sort(table(churn\_clean208$Tablet)), xlab = "Tablet")

barplot(sort(table(churn\_clean208$InternetService)), xlab = "Internet Service")

barplot(sort(table(churn\_clean208$Phone)), xlab = "Phone")

barplot(sort(table(churn\_clean208$Multiple)), xlab = "Multiple")

barplot(sort(table(churn\_clean208$OnlineSecurity)), xlab = "Online Security")

barplot(sort(table(churn\_clean208$OnlineBackup)), xlab = "Online Backup")

barplot(sort(table(churn\_clean208$DeviceProtection)), xlab = "Device Protection")

barplot(sort(table(churn\_clean208$TechSupport)), xlab = "Tech Support")

barplot(sort(table(churn\_clean208$StreamingTV)), xlab = "Streaming TV")

barplot(sort(table(churn\_clean208$StreamingMovies)), xlab = "Streaming Movies")

barplot(sort(table(churn\_clean208$PaperlessBilling)), xlab = "Paperless Billing")

barplot(sort(table(churn\_clean208$PaymentMethod)), xlab = "Payment Method")

barplot(sort(table(churn\_clean208$Tenure)), xlab = "Tenure")

barplot(sort(table(churn\_clean208$MonthlyCharge)), xlab = "Monthly Charge")

barplot(sort(table(churn\_clean208$Bandwidth\_GB\_Year)), xlab = "Bandwidth GB Year")

barplot(sort(table(churn\_clean208$Timely\_Response)), xlab = "Timely Response")

barplot(sort(table(churn\_clean208$Timely\_Fixes)), xlab = "Timely Fixes")

barplot(sort(table(churn\_clean208$Timely\_Replacements)), xlab = "Timely Replacements")

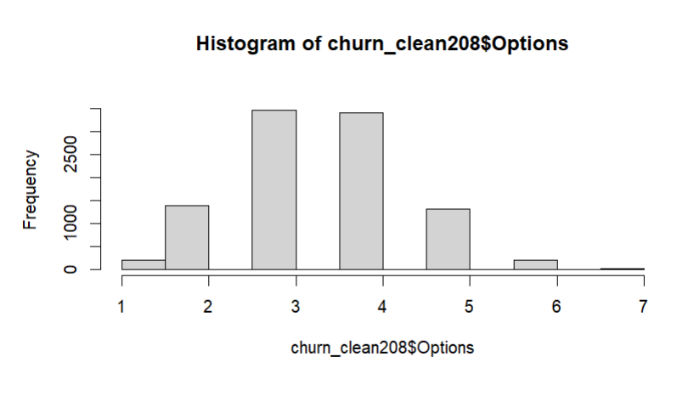
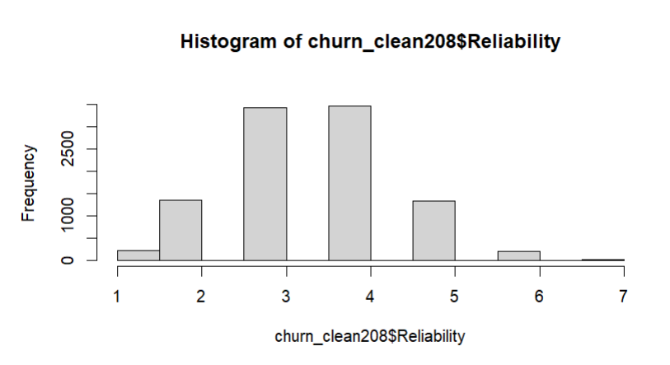
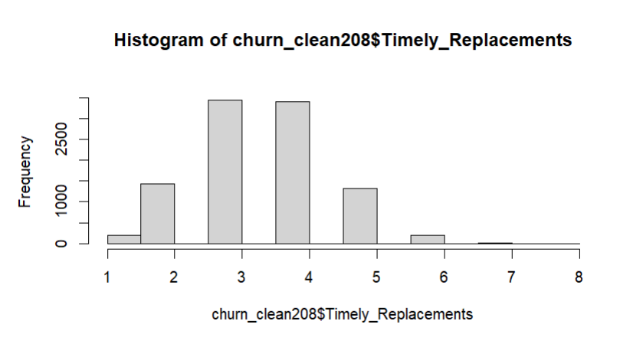
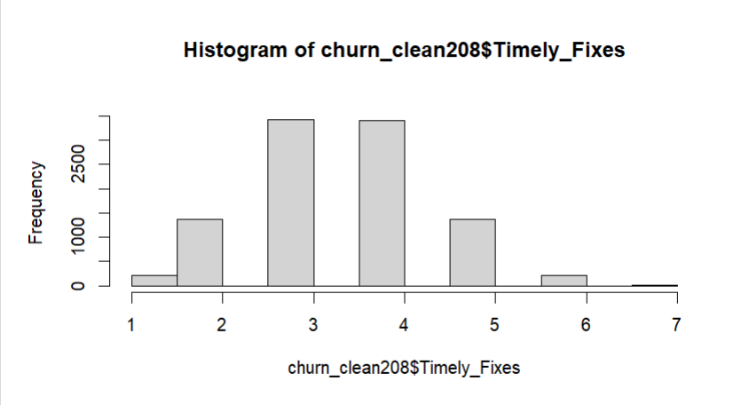
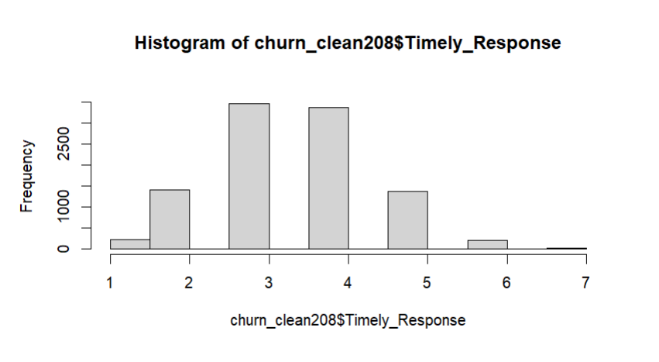
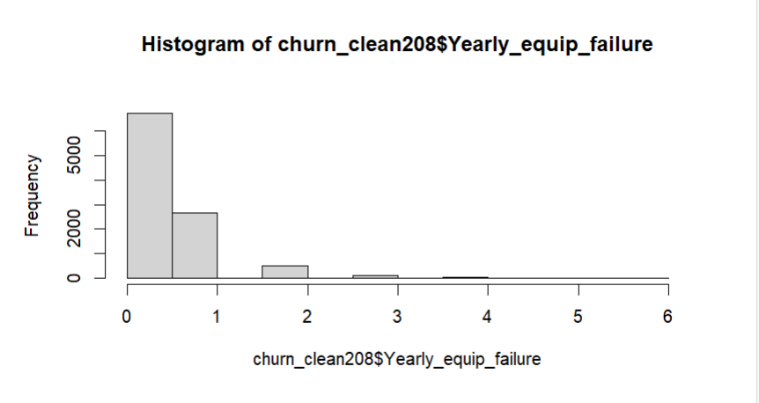
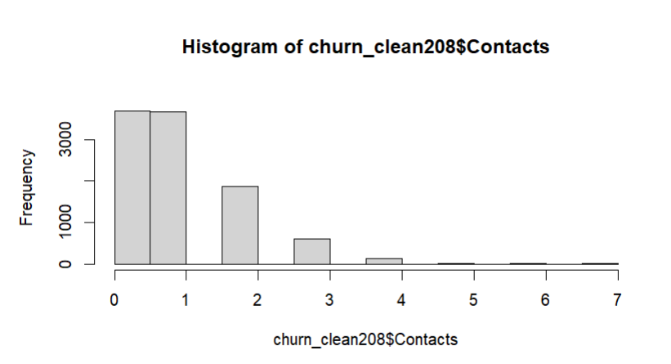
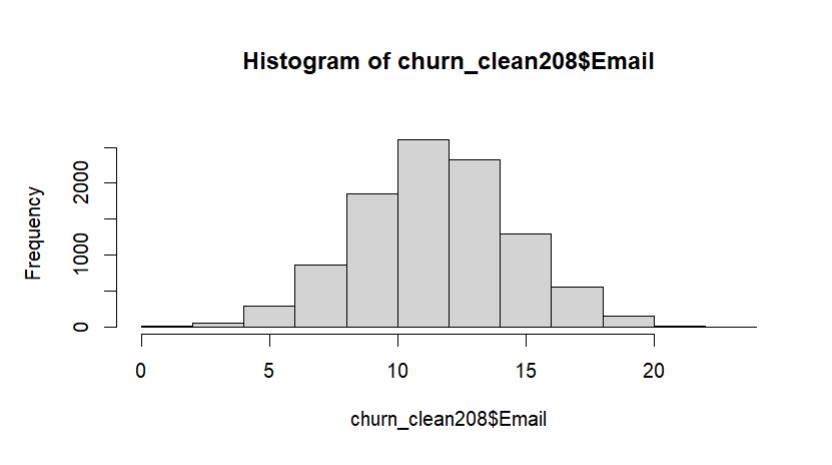
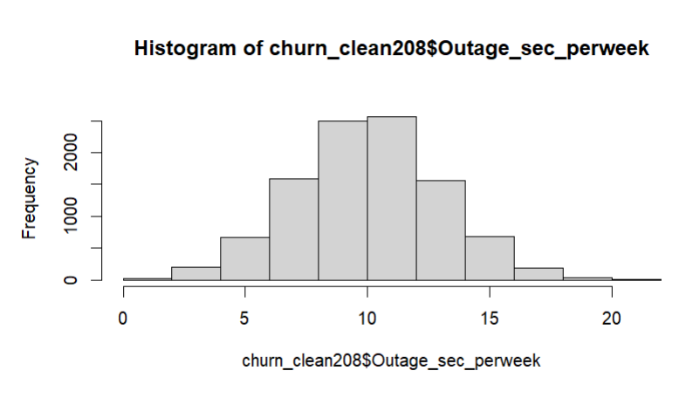
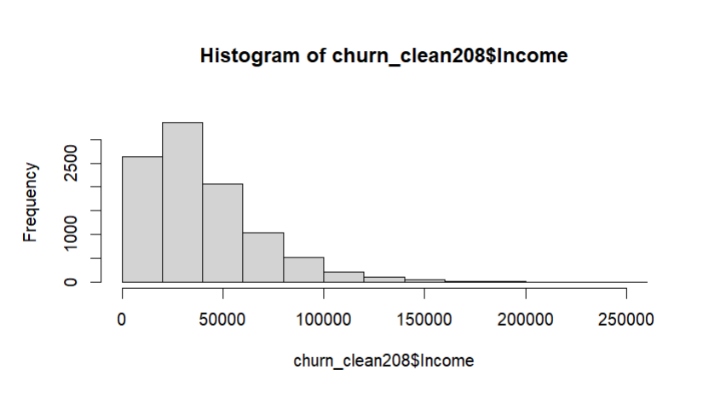
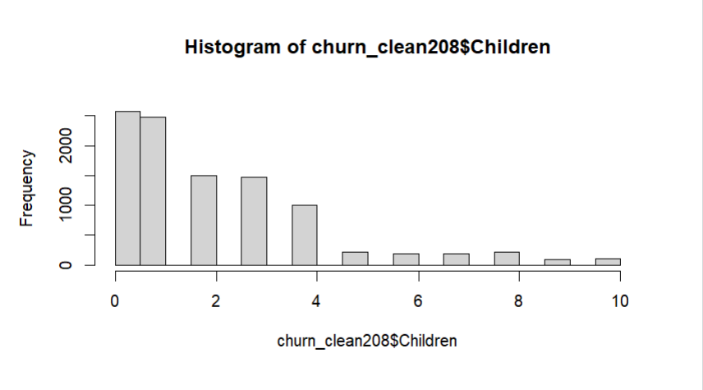
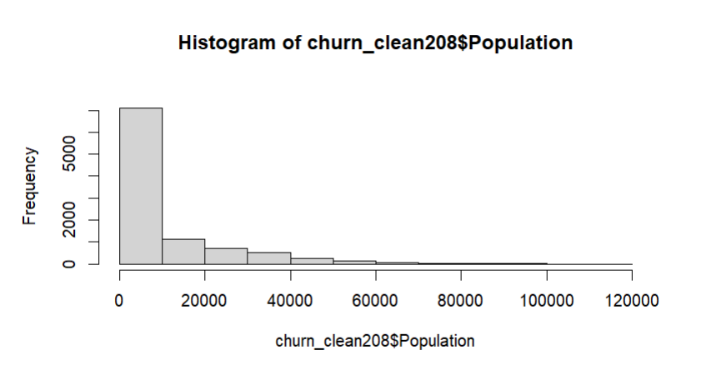
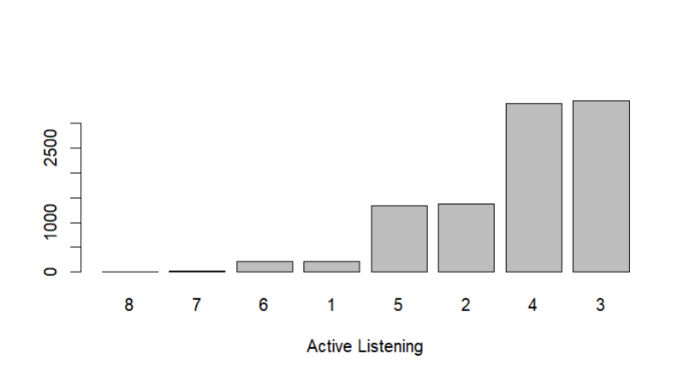
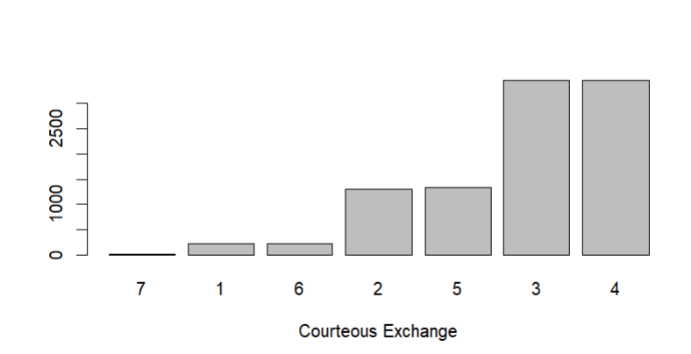
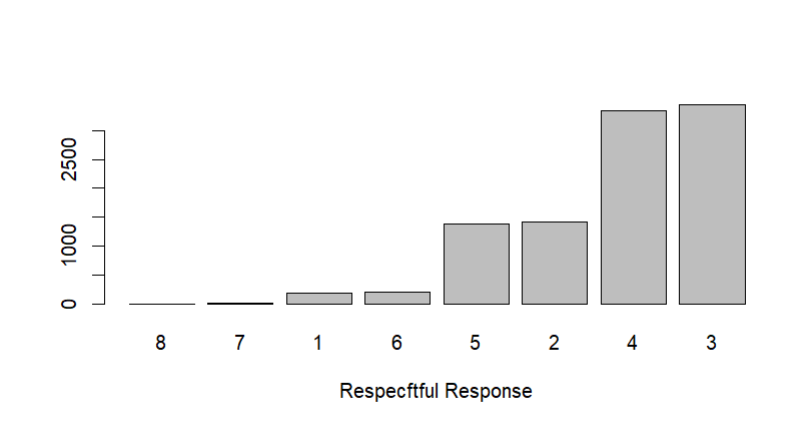
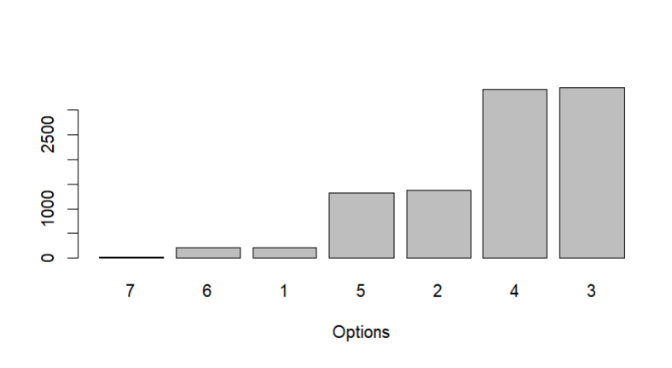
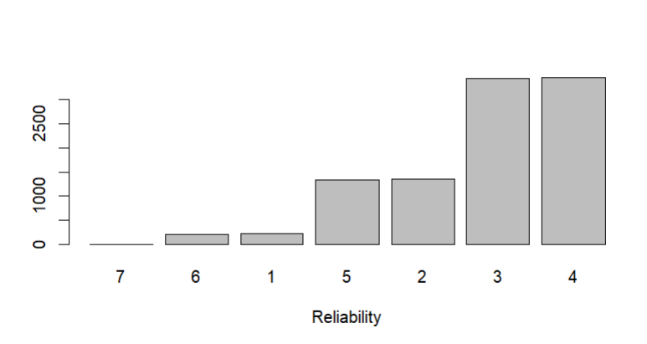
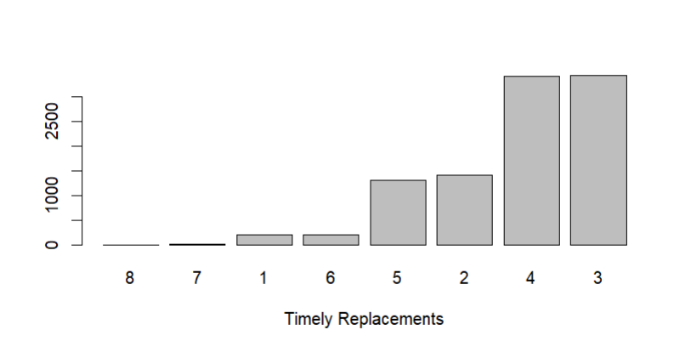
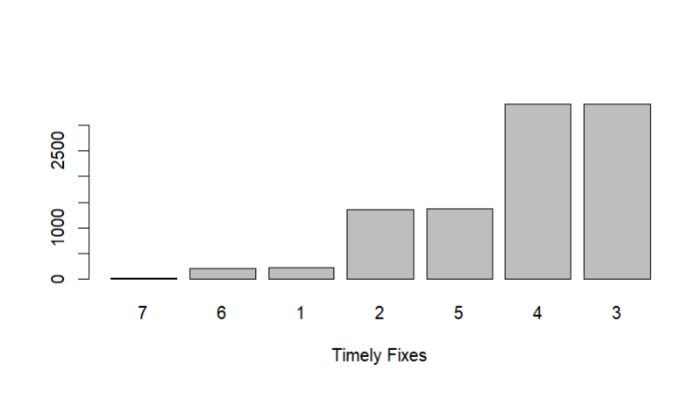
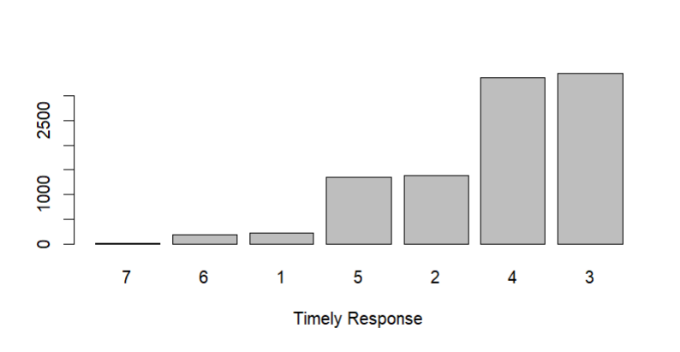
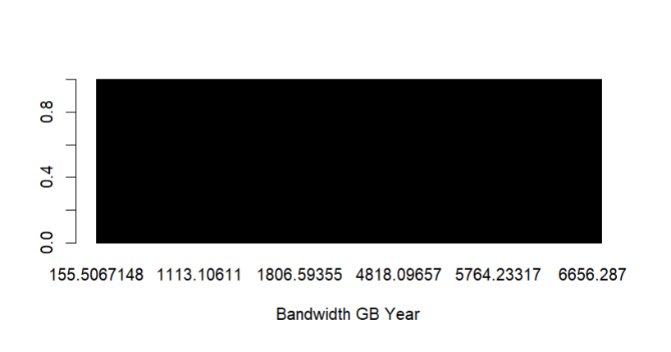
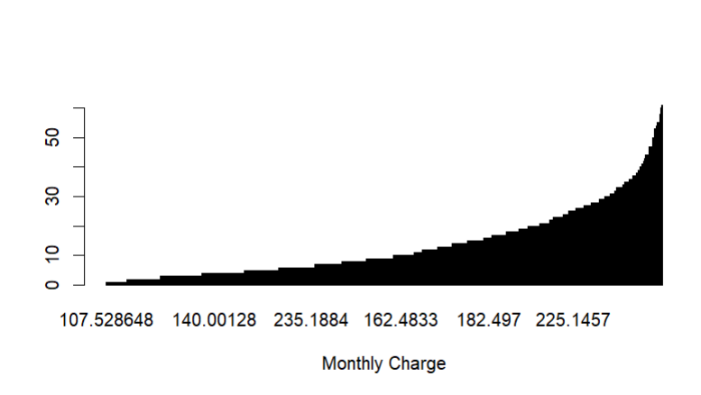
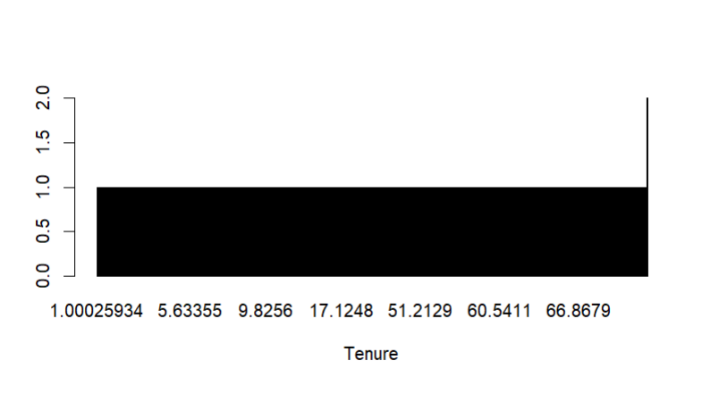
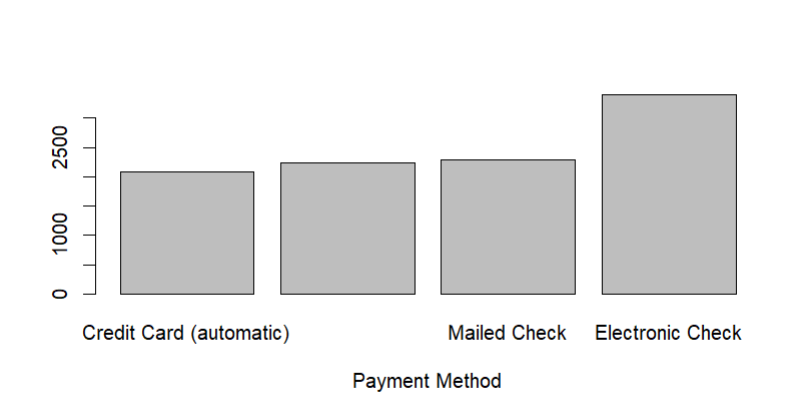
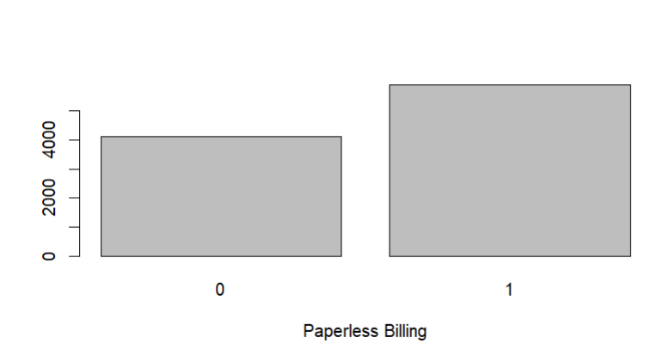
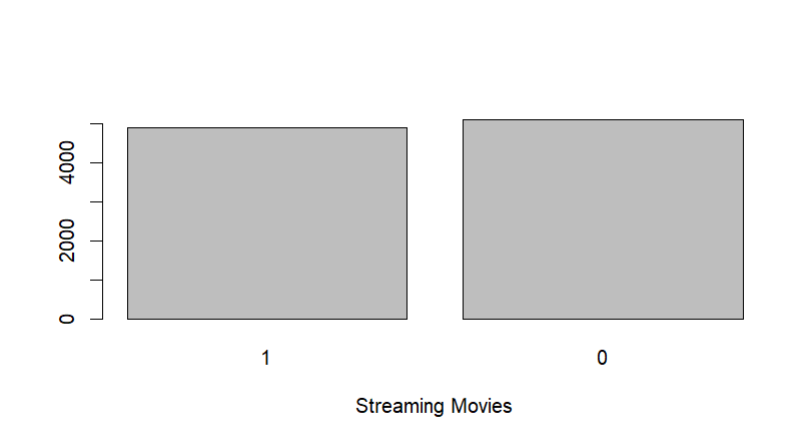
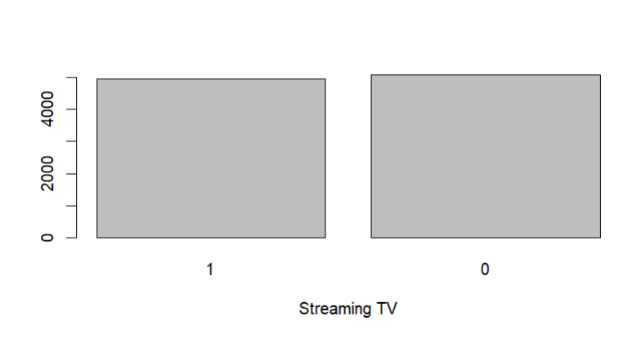
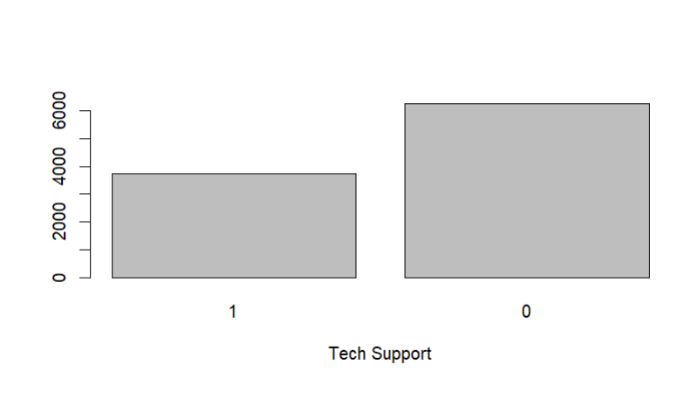
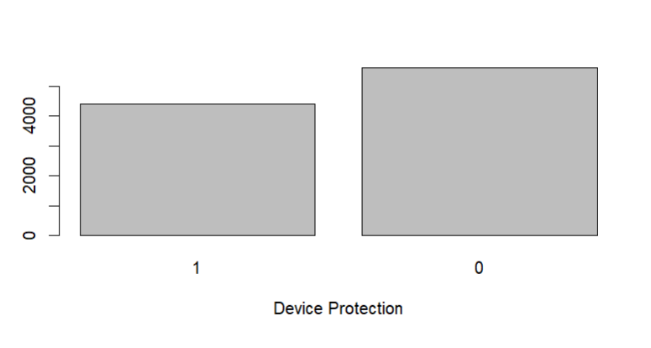
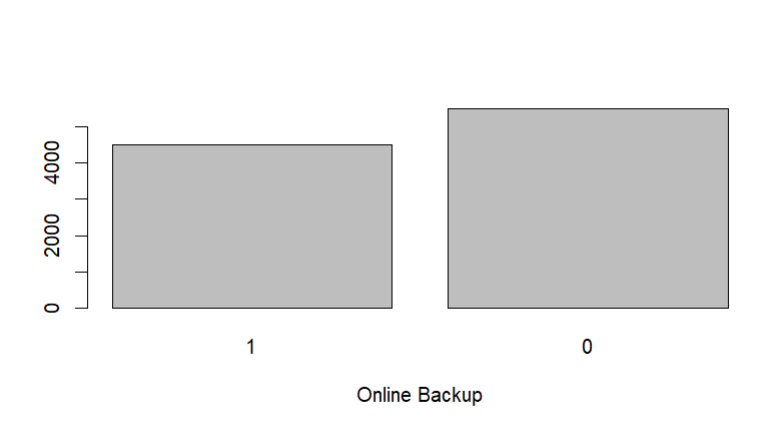
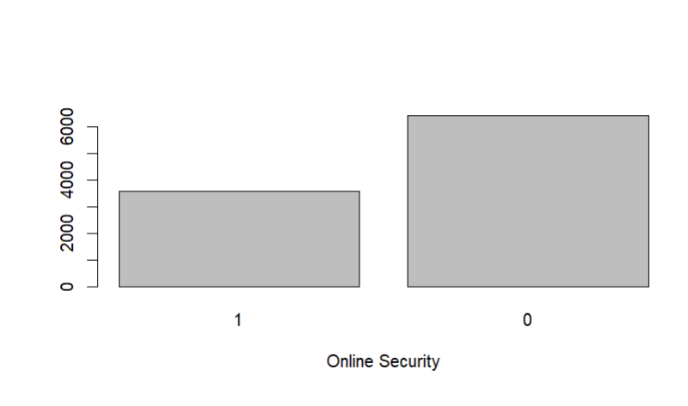
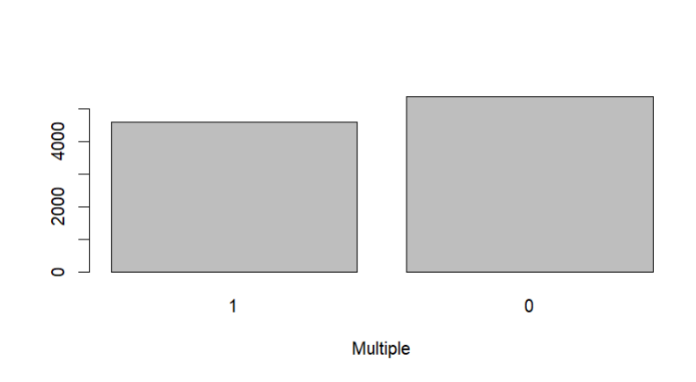
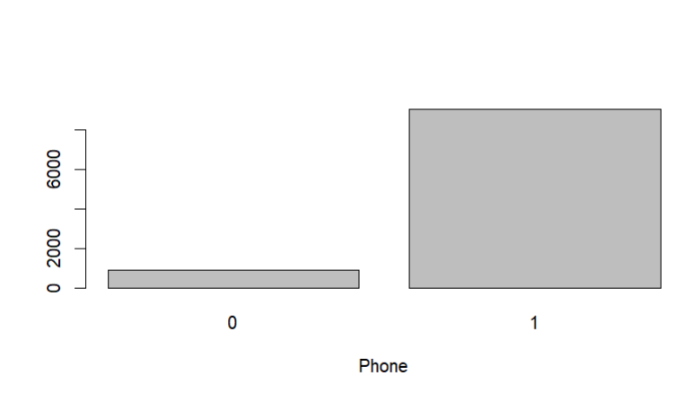
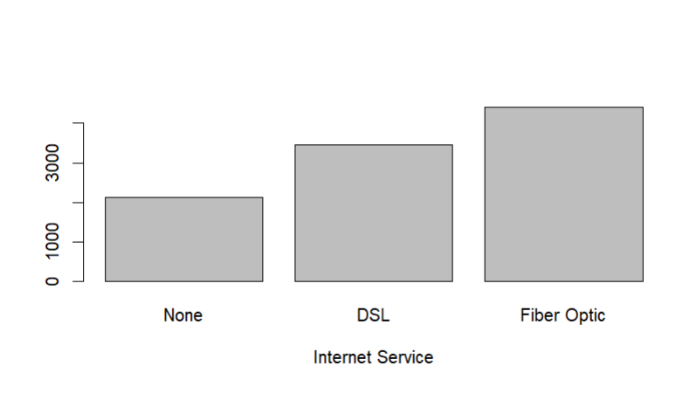
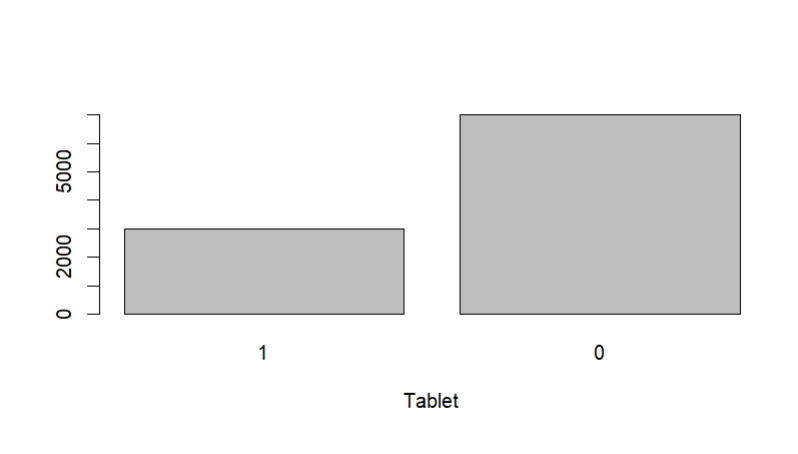
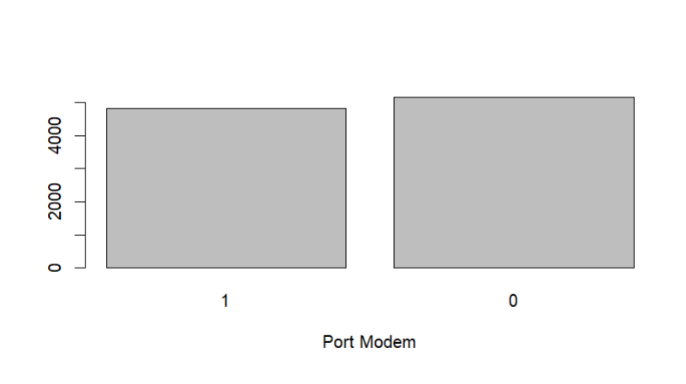
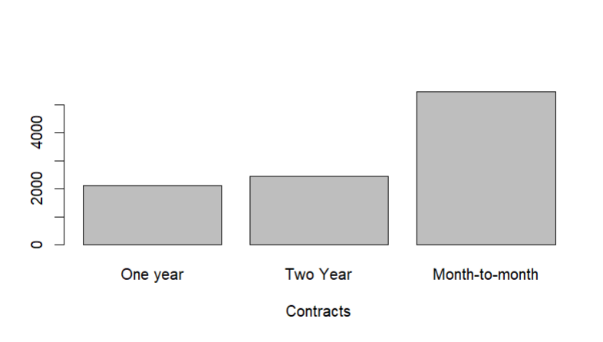
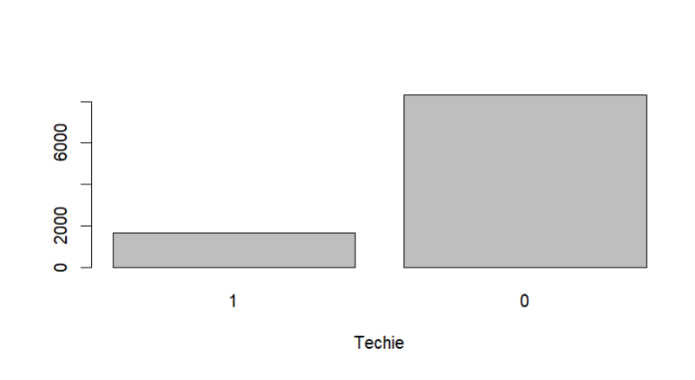
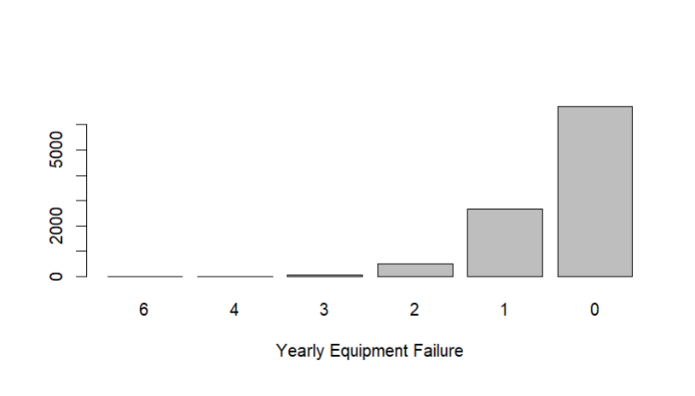
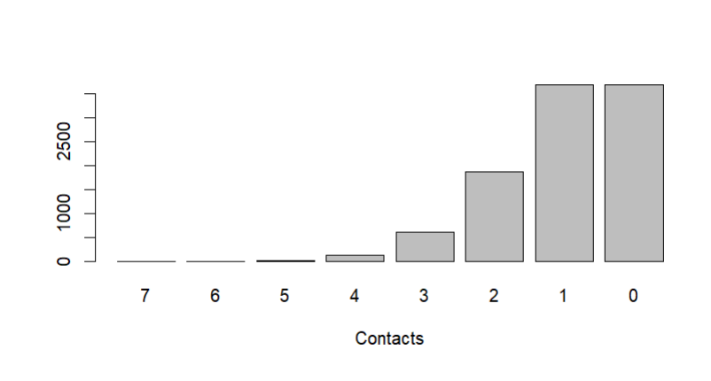
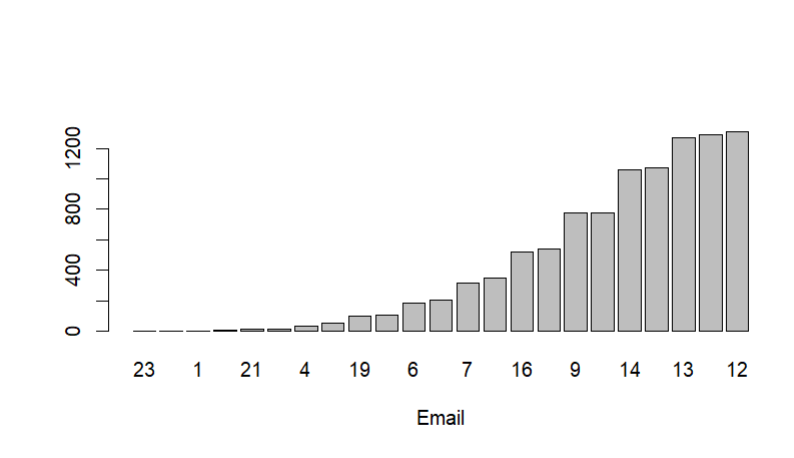
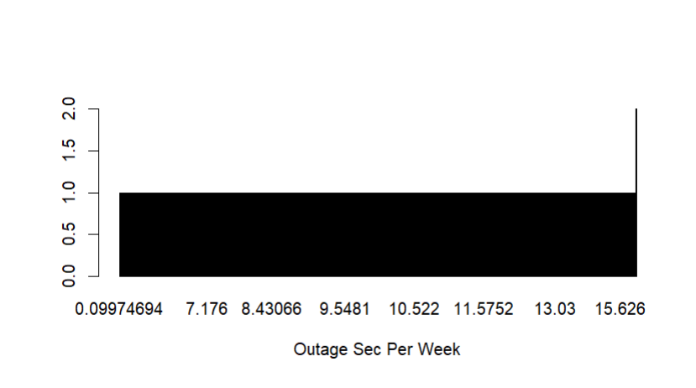
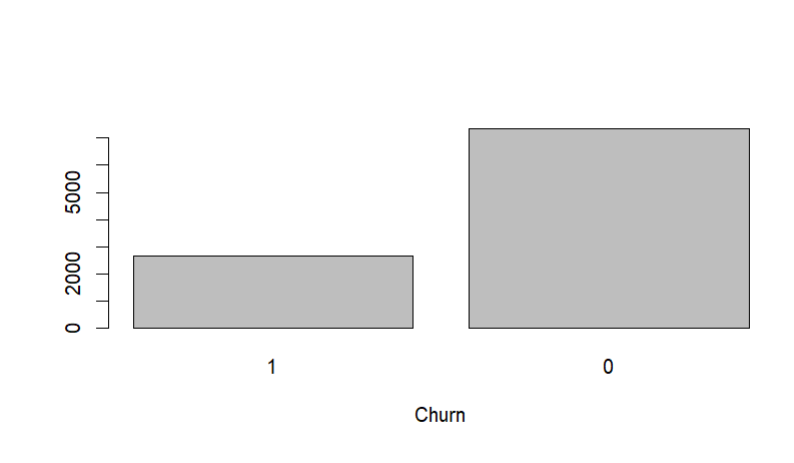
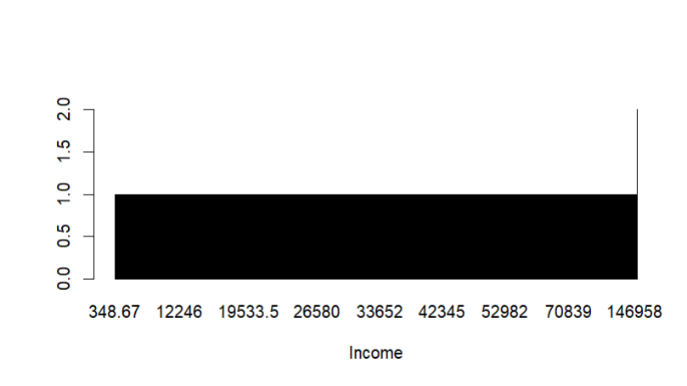
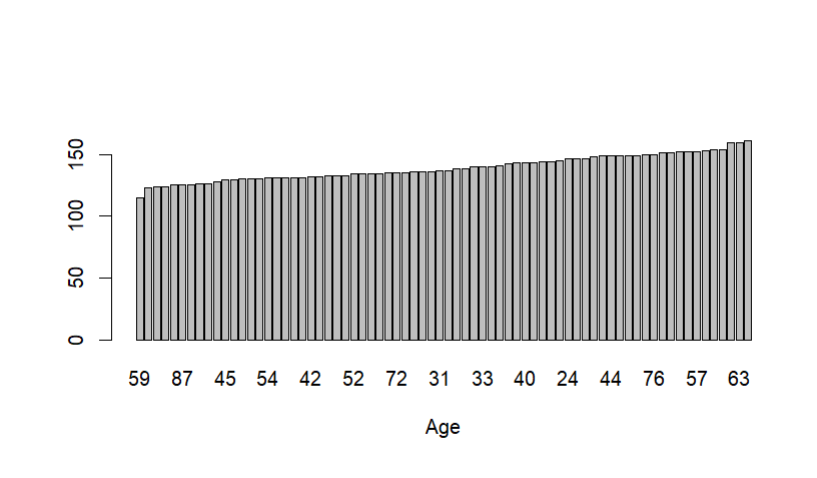
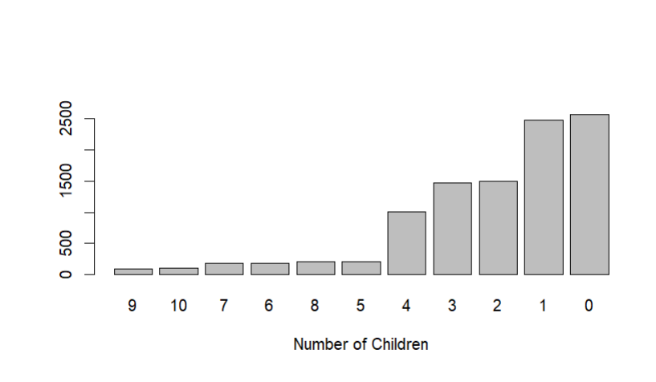
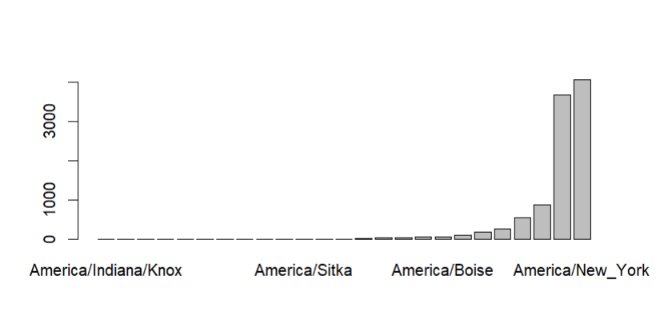
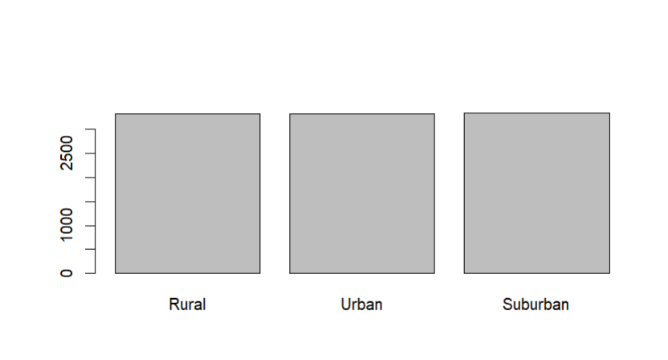
barplot(sort(table(churn\_clean208$Reliability)), xlab = "Reliability")

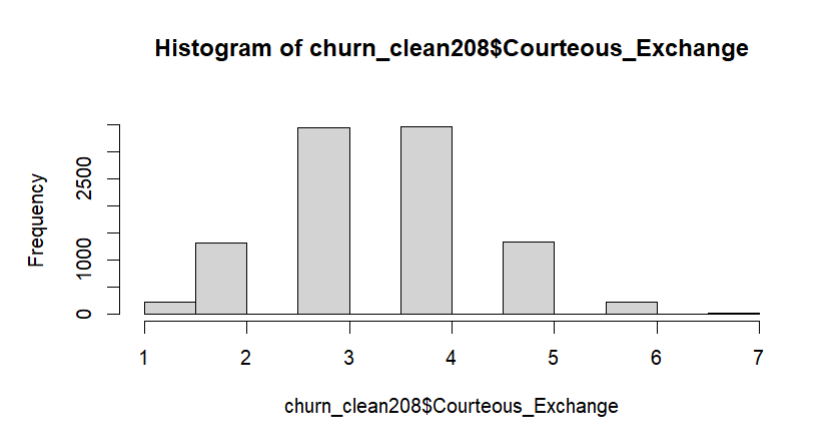
barplot(sort(table(churn\_clean208$Options)), xlab = "Options")

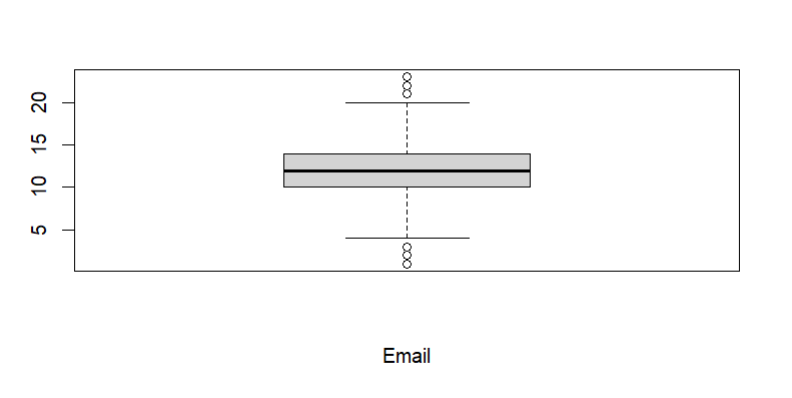
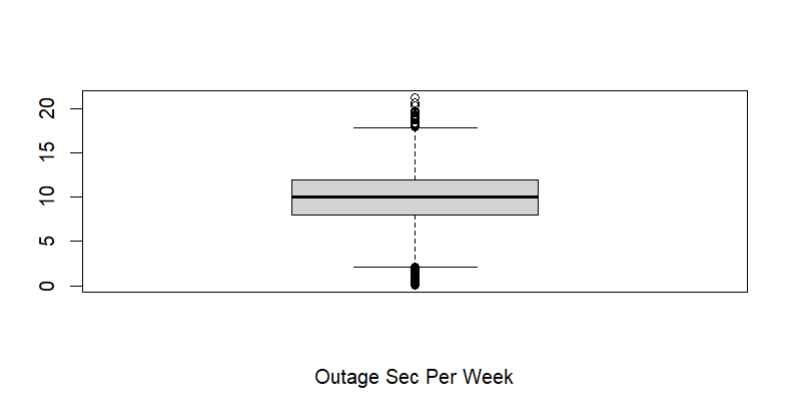
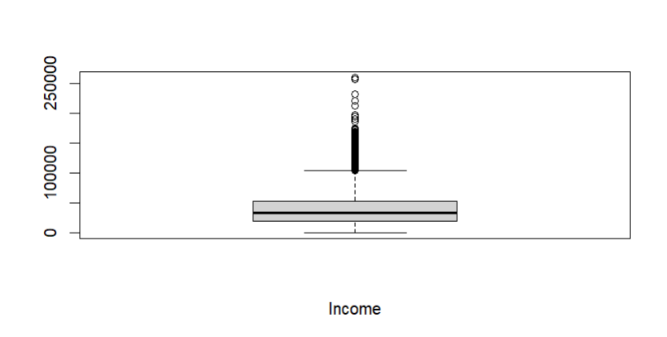
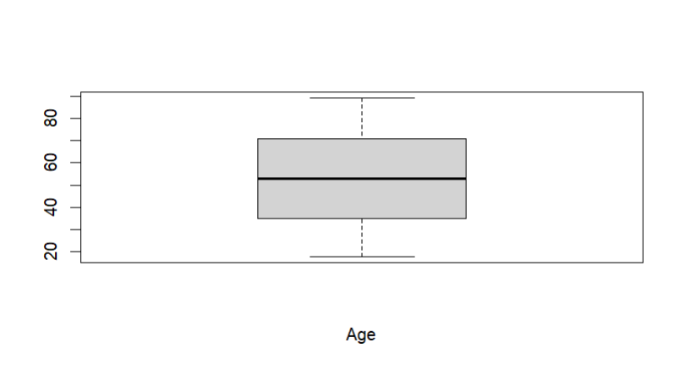
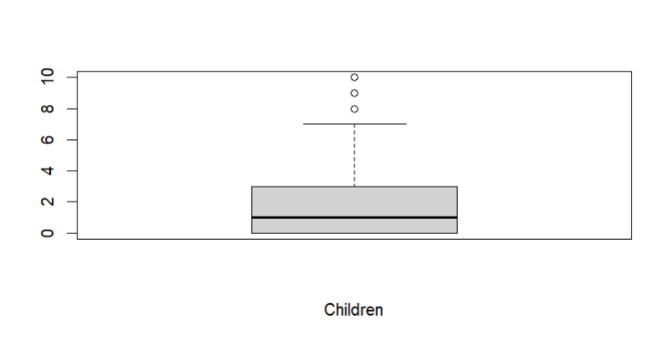
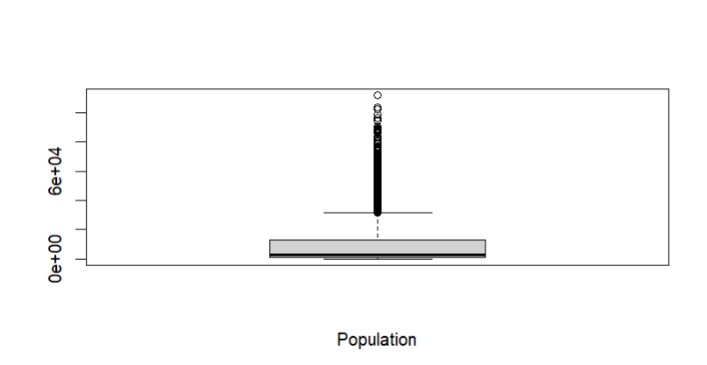
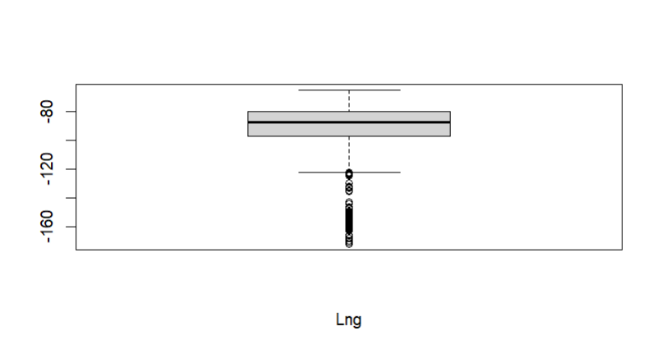
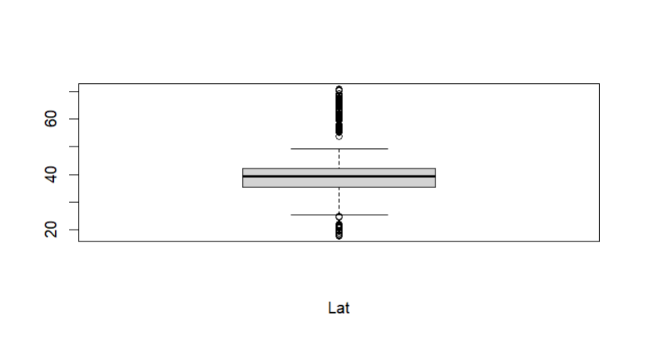
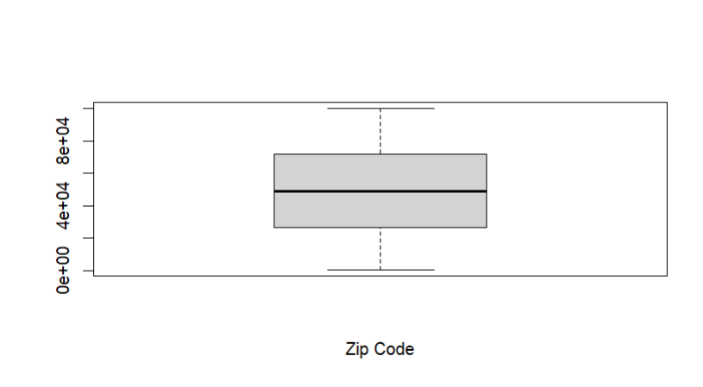
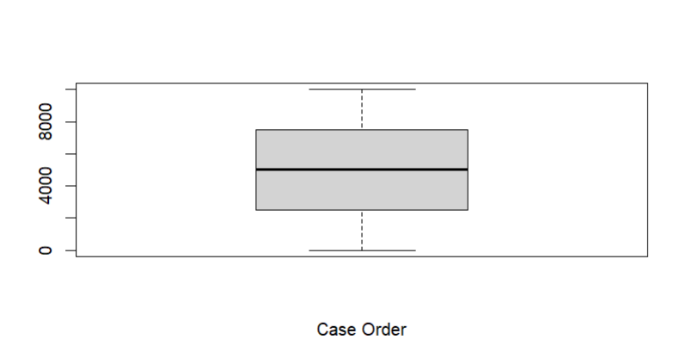
barplot(sort(table(churn\_clean208$Respectful\_Response)), xlab = "Respecftful Response")

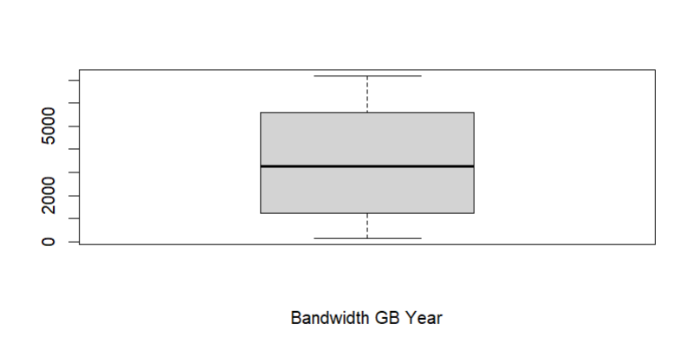
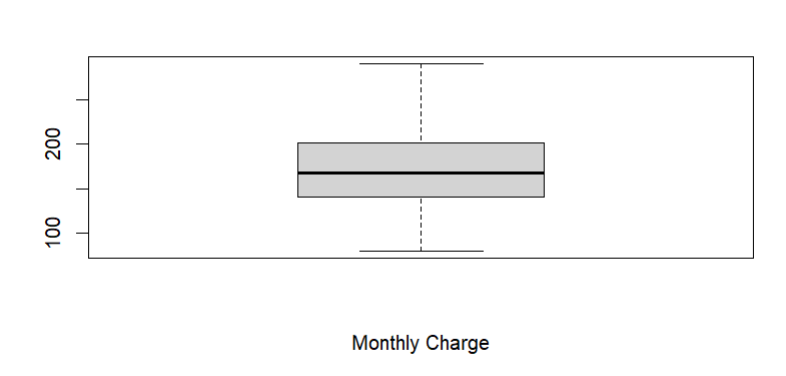
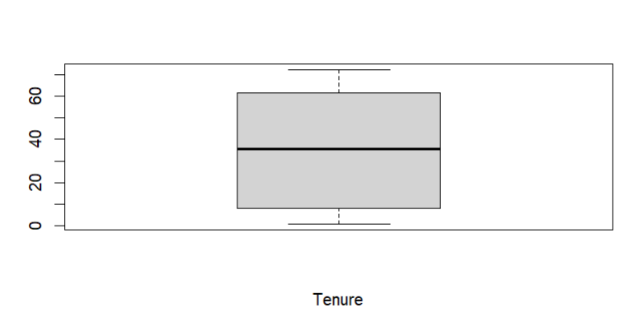
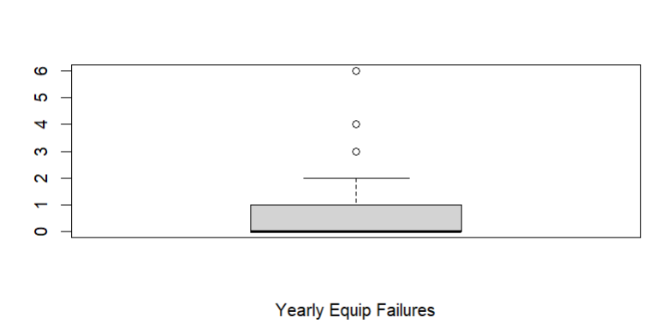
barplot(sort(table(churn\_clean208$Courteous\_Exchange)), xlab = "Courteous Exchange")

barplot(sort(table(churn\_clean208$Active\_Listening)), xlab = "Active Listening")

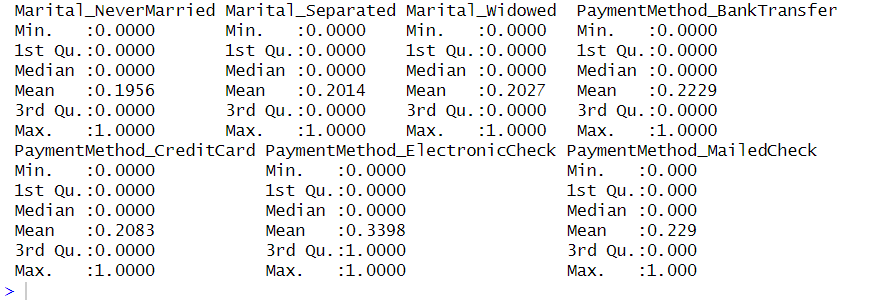
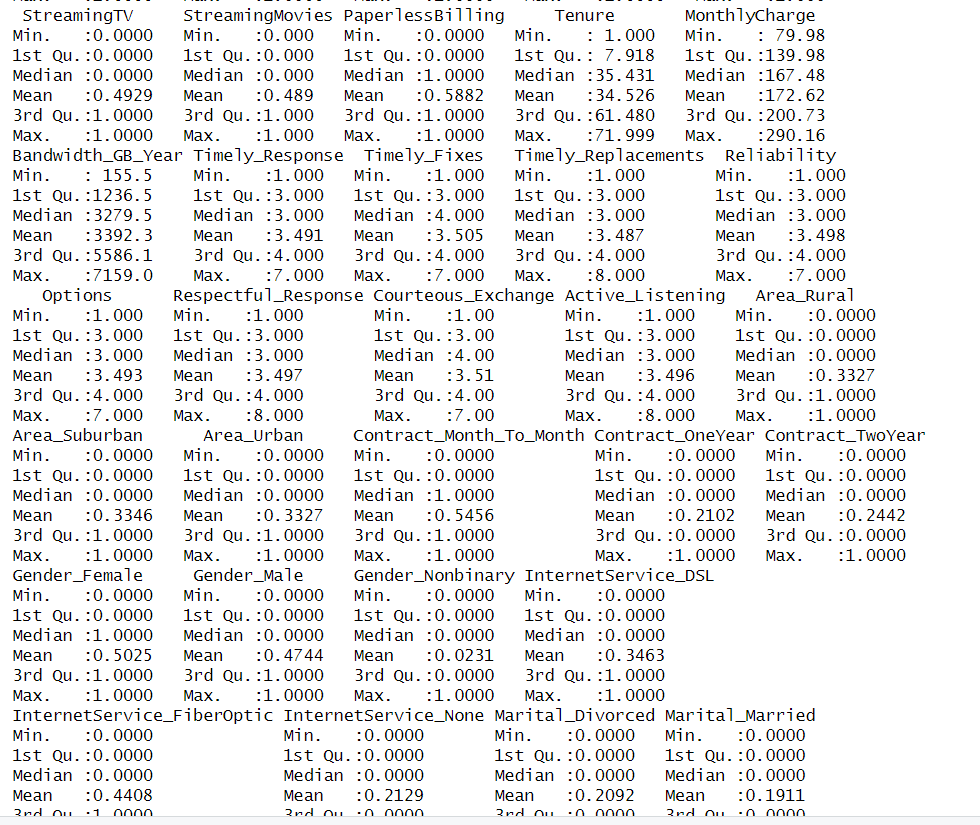
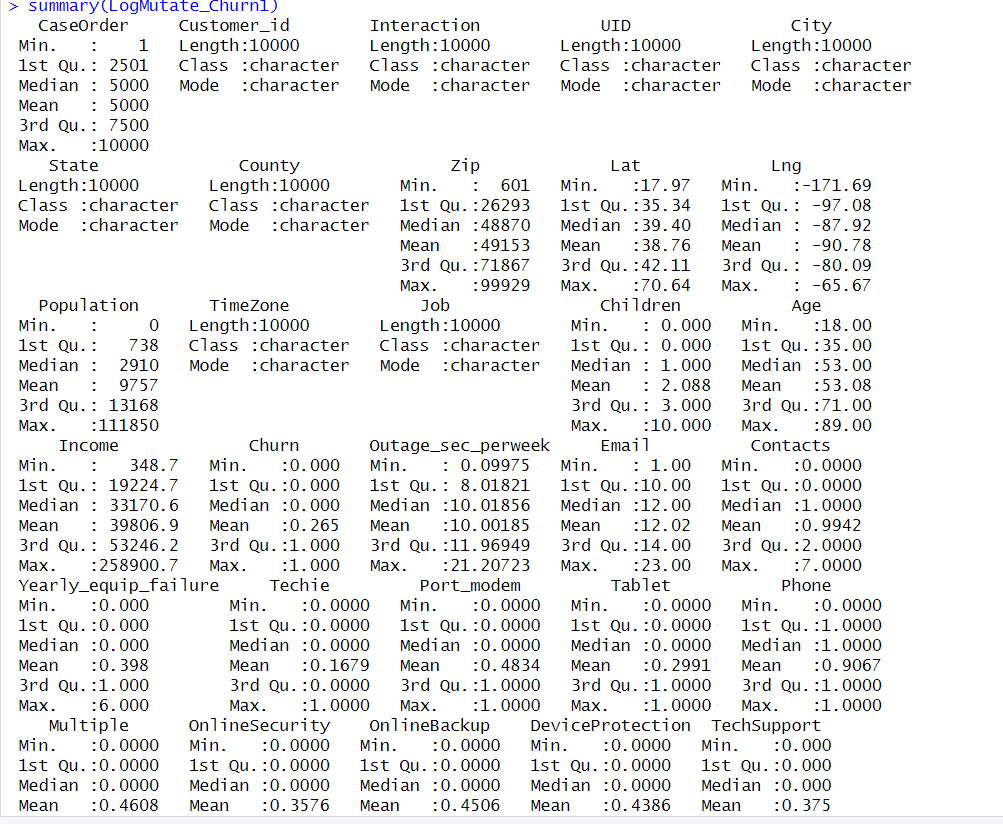
A graph with text overlay

Description automatically generatedA graph with text on it

Description automatically generatedA diagram of a graph

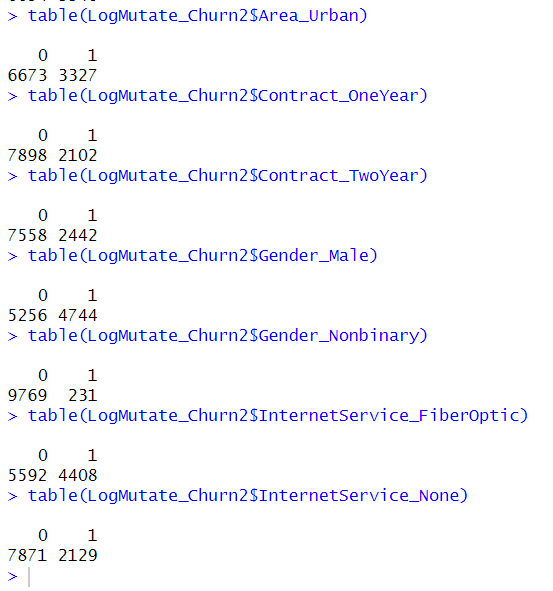
Description automatically generated with medium confidence

2.  The dependent variable is the Churn variable and all independent variables in the churn dataset are included in the summary statistics and are required to answer the research question. The categorical variables will be excluded aside from count as they are qualitative data and cannot be measured on a numerical scale. The categorical data is used for the count to determine the frequency or proportion of the category.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

3. Univariate visualizations were created for every variable in the analysis. Bivariate visualizations that include the dependent variable were created for every independent variable in the analysis. See attached code.

#BiVariate Graph of each variable

library(ggplot2)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Area)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Contract)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Gender)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Marital)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = InternetService)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = PaymentMethod)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Tenure)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Techie)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Port\_modem)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Tablet)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Phone)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = Multiple)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = OnlineSecurity)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = OnlineBackup)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = DeviceProtection)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = TechSupport)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = StreamingTV)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = StreamingMovies)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = churn\_clean208,

mapping = aes(x = Churn, y = PaperlessBilling)) +

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

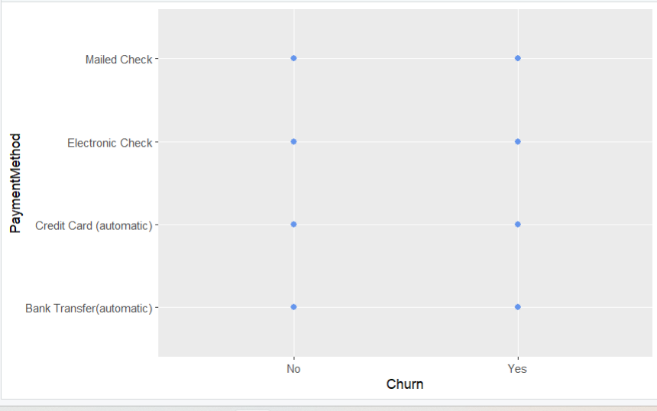
A screenshot of a graph

Description automatically generatedA grid of white squares with blue dots

Description automatically generatedA grid of white squares with blue dots

Description automatically generatedA graph with blue dots

Description automatically generatedA screenshot of a graph

Description automatically generatedA graph with blue lines

Description automatically generatedA screenshot of a graph

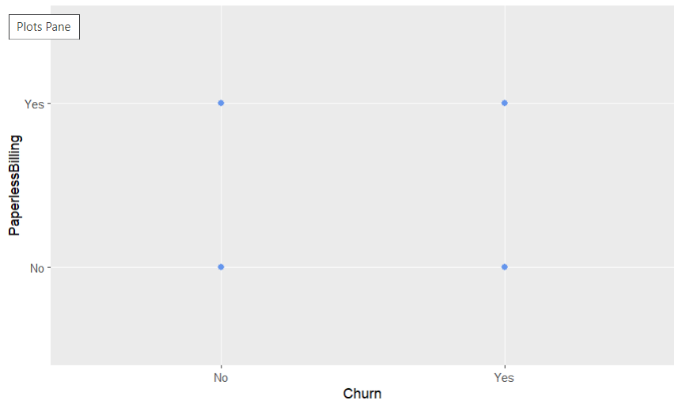
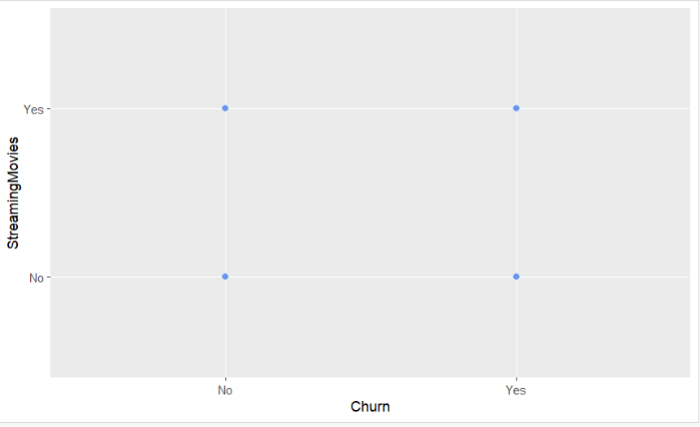
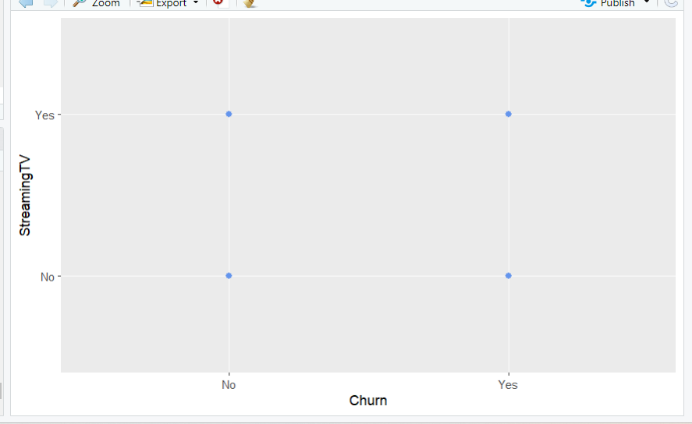
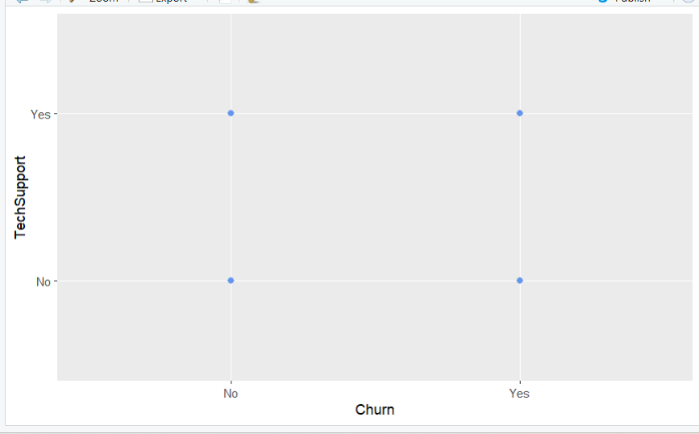
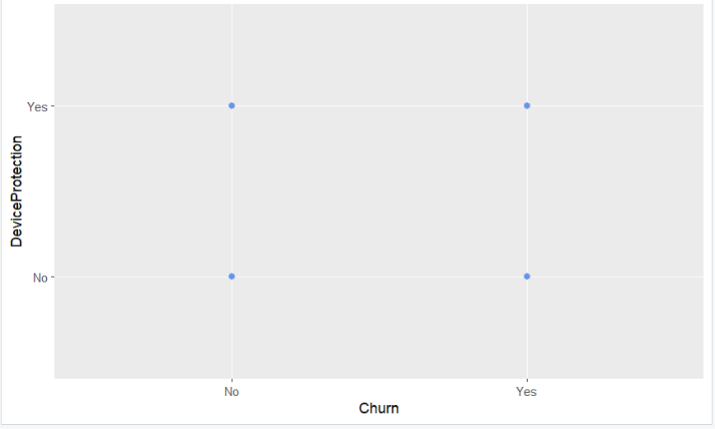
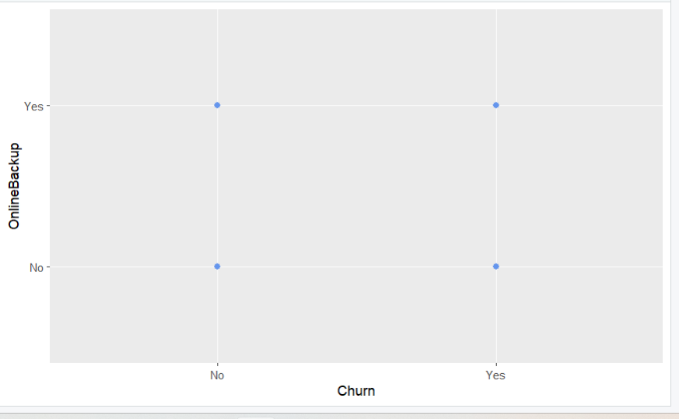
Description automatically generatedA screenshot of a computer

Description automatically generatedA graph with blue dots

Description automatically generatedA screenshot of a graph

Description automatically generatedA graph with many points

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated

4.  The data transformation goal is to transform the data to be easily readable and in a format that is clean and able to be utilized for linear regression while preserving the quality and size of the data. To transform the data, first, the variables named item 1- item 8 were re-named for the survey question that the recipients were answering. Next variables that were characters were encoded to numeric values so that they may be utilized in logistic regression. Lastly, one hot encoding and K-1 method was utilized to change categorical variables to numeric variables. See attached code.

#Encode categorical variables [in-text citation: (Statistics Globe, n.d.)]

#Encode Churn

churn\_clean208$Churn <- as.character(churn\_clean208$Churn)

churn\_clean208$Churn[churn\_clean208$Churn == "Yes"] <- 1

churn\_clean208$Churn[churn\_clean208$Churn == "No"] <- 0

churn\_clean208$Churn <- as.numeric(churn\_clean208$Churn)

#Encode Techie

churn\_clean208$Techie <- as.character(churn\_clean208$Techie)

churn\_clean208$Techie[churn\_clean208$Techie == "Yes"] <- 1

churn\_clean208$Techie[churn\_clean208$Techie == "No"] <- 0

churn\_clean208$Techie <- as.numeric(churn\_clean208$Techie)

#Encode Port Modem

churn\_clean208$Port\_modem <- as.character(churn\_clean208$Port\_modem)

churn\_clean208$Port\_modem[churn\_clean208$Port\_modem == "Yes"] <- 1

churn\_clean208$Port\_modem[churn\_clean208$Port\_modem == "No"] <- 0

churn\_clean208$Port\_modem <- as.numeric(churn\_clean208$Port\_modem)

#Encode Port Tablet

churn\_clean208$Tablet <- as.character(churn\_clean208$Tablet)

churn\_clean208$Tablet[churn\_clean208$Tablet == "Yes"] <- 1

churn\_clean208$Tablet[churn\_clean208$Tablet == "No"] <- 0

churn\_clean208$Tablet <- as.numeric(churn\_clean208$Tablet)

#Encode Phone

churn\_clean208$Phone <- as.character(churn\_clean208$Phone)

churn\_clean208$Phone[churn\_clean208$Phone == "Yes"] <- 1

churn\_clean208$Phone[churn\_clean208$Phone == "No"] <- 0

churn\_clean208$Phone <- as.numeric(churn\_clean208$Phone)

#Encode Multiple

churn\_clean208$Multiple <- as.character(churn\_clean208$Multiple)

churn\_clean208$Multiple[churn\_clean208$Multiple == "Yes"] <- 1

churn\_clean208$Multiple[churn\_clean208$Multiple == "No"] <- 0

churn\_clean208$Multiple <- as.numeric(churn\_clean208$Multiple)

#Encode Online Security

churn\_clean208$OnlineSecurity <- as.character(churn\_clean208$OnlineSecurity)

churn\_clean208$OnlineSecurity[churn\_clean208$OnlineSecurity == "Yes"] <- 1

churn\_clean208$OnlineSecurity[churn\_clean208$OnlineSecurity == "No"] <- 0

churn\_clean208$OnlineSecurity <- as.numeric(churn\_clean208$OnlineSecurity)

#Encode Online Backup

churn\_clean208$OnlineBackup <- as.character(churn\_clean208$OnlineBackup)

churn\_clean208$OnlineBackup[churn\_clean208$OnlineBackup == "Yes"] <- 1

churn\_clean208$OnlineBackup[churn\_clean208$OnlineBackup == "No"] <- 0

churn\_clean208$OnlineBackup <- as.numeric(churn\_clean208$OnlineBackup)

#Encode Device Protection

churn\_clean208$DeviceProtection <- as.character(churn\_clean208$DeviceProtection)

churn\_clean208$DeviceProtection[churn\_clean208$DeviceProtection == "Yes"] <- 1

churn\_clean208$DeviceProtection[churn\_clean208$DeviceProtection == "No"] <- 0

churn\_clean208$DeviceProtection <- as.numeric(churn\_clean208$DeviceProtection)

#Encode Tech Support

churn\_clean208$TechSupport <- as.character(churn\_clean208$TechSupport)

churn\_clean208$TechSupport[churn\_clean208$TechSupport == "Yes"] <- 1

churn\_clean208$TechSupport[churn\_clean208$TechSupport == "No"] <- 0

churn\_clean208$TechSupport <- as.numeric(churn\_clean208$TechSupport)

#Encode Streaming TV

churn\_clean208$StreamingTV <- as.character(churn\_clean208$StreamingTV)

churn\_clean208$StreamingTV[churn\_clean208$StreamingTV == "Yes"] <- 1

churn\_clean208$StreamingTV[churn\_clean208$StreamingTV == "No"] <- 0

churn\_clean208$StreamingTV <- as.numeric(churn\_clean208$StreamingTV)

#Encode Streaming Movies

churn\_clean208$StreamingMovies <- as.character(churn\_clean208$StreamingMovies)

churn\_clean208$StreamingMovies[churn\_clean208$StreamingMovies == "Yes"] <- 1

churn\_clean208$StreamingMovies[churn\_clean208$StreamingMovies == "No"] <- 0

churn\_clean208$StreamingMovies <- as.numeric(churn\_clean208$StreamingMovies)

#Encode Paperless Billing

churn\_clean208$PaperlessBilling <- as.character(churn\_clean208$PaperlessBilling)

churn\_clean208$PaperlessBilling[churn\_clean208$PaperlessBilling == "Yes"] <- 1

churn\_clean208$PaperlessBilling[churn\_clean208$PaperlessBilling == "No"] <- 0

churn\_clean208$PaperlessBilling <- as.numeric(churn\_clean208$PaperlessBilling)

#One-Hot Encoding Area

library(fastDummies)

A\_treat <- dummy\_cols(churn\_clean208, select\_columns = "Area")

#One-Hot Encoding Marital

M\_treat <- dummy\_cols(churn\_clean208, select\_columns = "Marital")

#One-Hot Encoding Gender

G\_treat <- dummy\_cols(churn\_clean208, select\_columns = "Gender")

#One-Hot Encoding Contract

C\_treat <- dummy\_cols(churn\_clean208, select\_columns = "Contract")

#One-Hot Encoding InternetService

I\_treat <- dummy\_cols(churn\_clean208, select\_columns = "InternetService")

#One-Hot Encoding PaymentMethod

P\_treat <- dummy\_cols(churn\_clean208, select\_columns = "PaymentMethod")

library(tidyverse)

#put all df in a list [in-text citation: (Zach, 2021)]

logdf\_list <- list(churn\_clean208, A\_treat, M\_treat, G\_treat, C\_treat, I\_treat, P\_treat)

#Merge all data frames together using mutate[in-text citation: (Zach, 2021)]

library (dplyr)

logMutate\_Churn <- mutate(churn\_clean208, A\_treat, C\_treat, G\_treat, I\_treat, M\_treat, P\_treat)

#Drop columns by name that were duplicates to the variables one-hot encoded

LogMutate\_Churn1 <- subset(logMutate\_Churn, select = -c(Area, Contract, Gender, InternetService, Marital, PaymentMethod))

#Rename columns with unexpected symbol

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'Marital\_Never Married'] <- 'Marital\_NeverMarried'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'Contract\_One year'] <- 'Contract\_OneYear'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'Contract\_Two Year'] <- 'Contract\_TwoYear'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'Contract\_Month-to-month'] <- 'Contract\_Month\_To\_Month'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'InternetService\_Fiber Optic'] <- 'InternetService\_FiberOptic'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'PaymentMethod\_Bank Transfer(automatic)'] <- 'PaymentMethod\_BankTransfer'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'PaymentMethod\_Credit Card (automatic)'] <- 'PaymentMethod\_CreditCard'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'PaymentMethod\_Electronic Check'] <- 'PaymentMethod\_ElectronicCheck'

colnames(LogMutate\_Churn1)[colnames(LogMutate\_Churn1) == 'PaymentMethod\_Mailed Check'] <- 'PaymentMethod\_MailedCheck'

5.  See attached code.

**Part IV: Model Comparison and Analysis**

D.  The initial and reduced logistic regression model are compared by doing the following:

1.  #Create initial model- logistic regression

logmodel <- glm(Churn ~ Area\_Rural + Area\_Suburban + Area\_Urban + Children + Age +

Income + Marital\_Divorced + Marital\_Married + Marital\_NeverMarried +

Marital\_Separated + Marital\_Widowed + Gender\_Female + Gender\_Male +

Gender\_Nonbinary + Tenure + Outage\_sec\_perweek + Email + Contacts +

Yearly\_equip\_failure + Techie + Contract\_Month\_To\_Month +

Contract\_OneYear + Contract\_TwoYear + Port\_modem + Tablet +

InternetService\_DSL + InternetService\_FiberOptic + InternetService\_None +

Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +

TechSupport + StreamingMovies + StreamingTV + PaperlessBilling +

PaymentMethod\_BankTransfer + PaymentMethod\_CreditCard +

PaymentMethod\_ElectronicCheck + PaymentMethod\_MailedCheck +

MonthlyCharge + Bandwidth\_GB\_Year + Timely\_Response +

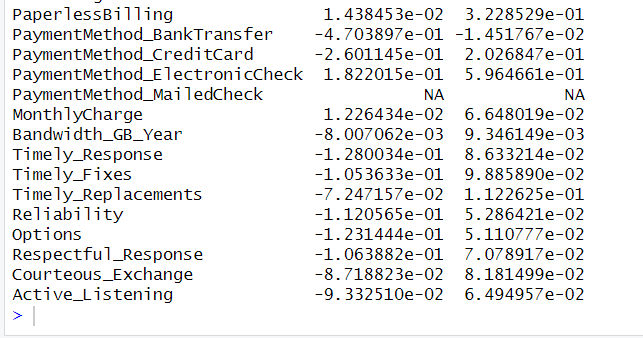
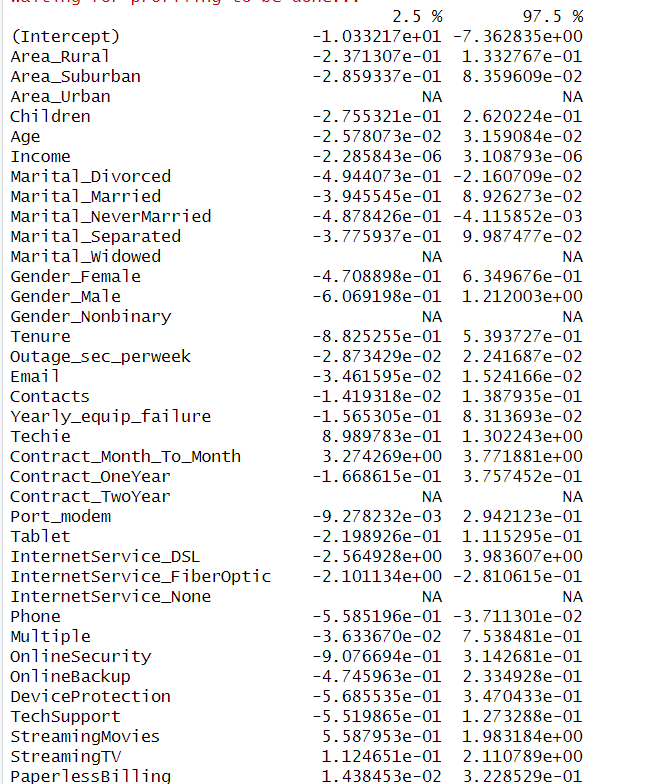
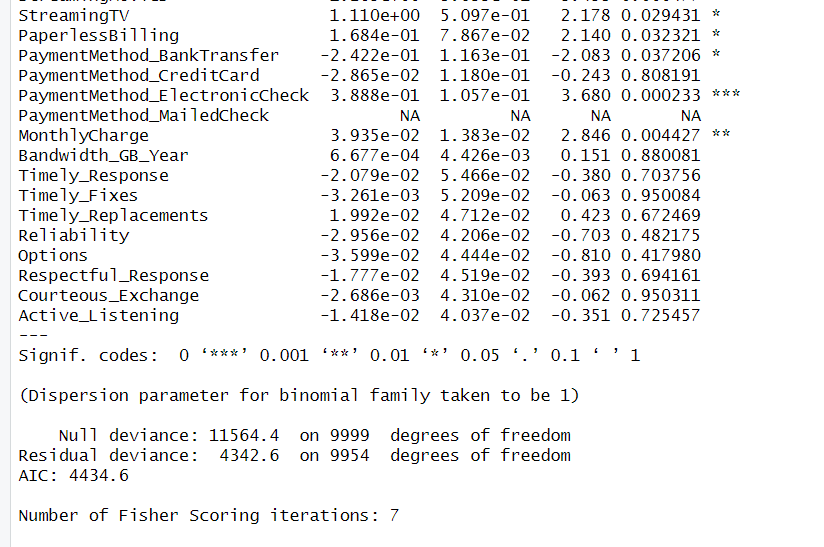
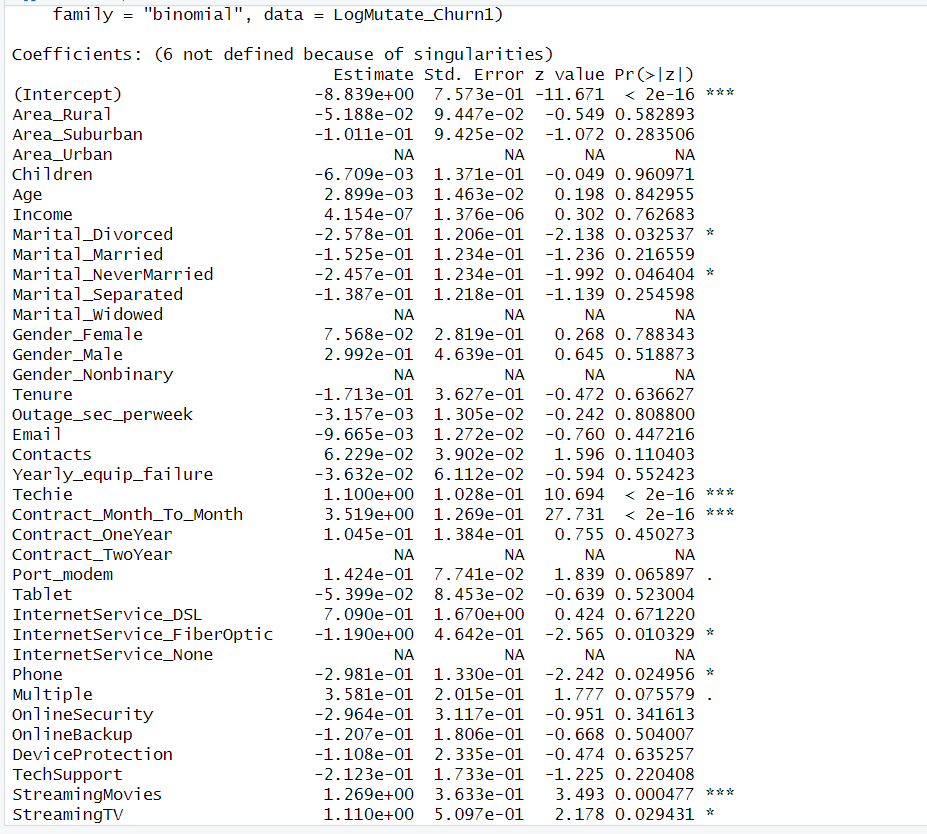
Timely\_Fixes + Timely\_Replacements + Reliability + Options +

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = LogMutate\_Churn1,

family = "binomial")

summary(logmodel)

confint(logmodel)



2.  Feature selection reduction was done utilizing two different feature selection methods. First categorical variables with more than three levels were reduced due to the cardinality being too great and we do not want to proliferate. Next, we utilized the wrapper method which is a feature selection reduction method. We utilized the backward stepwise reduction method before finalizing our logistic regression model.

#Reduce model

#Reduce Model Payment Method, Marital as cardinality is too great and we do not want to proliferate

#Reduce Model Payment Method, Marital as cardinality is too great and we do not want to proliferate

#K-1 method

LogMutate\_Churn2 <- subset(LogMutate\_Churn1, select = -c(Job, TimeZone, Population,

Lat, Lng, Zip, County, State,

City, UID, Interaction, Customer\_id,

PaymentMethod\_BankTransfer,

PaymentMethod\_CreditCard,

PaymentMethod\_ElectronicCheck,

PaymentMethod\_MailedCheck,

Marital\_Divorced, Marital\_Married,

Marital\_NeverMarried, Marital\_Widowed,

Marital\_Separated, Area\_Rural, Contract\_Month\_To\_Month,

Gender\_Female, InternetService\_DSL))

#Check correlation using Pearson method

library(corrr)

Churncor <- cor(LogMutate\_Churn2)

3.

#Create reduced model- logistic regression

logmodelreduced <- glm(Churn ~ Area\_Suburban + Area\_Urban + Children + Age +

Income + Gender\_Male + Gender\_Nonbinary + Tenure +

Outage\_sec\_perweek + Email + Contacts +

Yearly\_equip\_failure + Techie + Contract\_OneYear +

Contract\_TwoYear + Port\_modem + Tablet +

InternetService\_FiberOptic + InternetService\_None +

Phone + Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +

TechSupport + StreamingMovies + StreamingTV + PaperlessBilling +

MonthlyCharge + Bandwidth\_GB\_Year + Timely\_Response +

Timely\_Fixes + Timely\_Replacements + Reliability + Options +

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = LogMutate\_Churn1,

family = "binomial")

summary(logmodelreduced)

confint(logmodelreduced)

#Visualize

LogMutate\_Churn2 %>%

dplyr::select(where(is.numeric)) %>%

correlate() %>%

shave() %>%

rplot(print\_cor = TRUE) +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

plot(logmodelreduced)

#McFadden Rsquared

library(pscl)

pR2(logmodelreduced)

#Reduce model using feature selection method

#Call proper packages

library(tidyverse)

library(caret)

library(leaps)

library(MASS)

# Stepwise regression

log.step.model <- stepAIC(logmodelreduced, direction = "backward",

trace = FALSE)

summary(log.step.model)

logmodels <- regsubsets(Churn~., data = LogMutate\_Churn2, nvmax = 50,

method = "backward")

LogSummaryModels <- summary(logmodels)

log.step.model1 <- stepAIC(logmodelreduced, direction = "backward",

trace = FALSE)

summary(log.step.model1)

logmodels1 <- regsubsets(Churn~., data = LogMutate\_Churn2, nvmax = 50,

method = "backward")

LogSummaryModels1 <- summary(logmodels1)

#Visualizations

plot(logmodels1)

#Create confusion matrix

#Try a prediction

actualresponse <- LogMutate\_Churn1$Churn

logmodelprediction <- round(fitted(logmodelreduced))

confusionmatrix <- table(logmodelprediction, actualresponse)

confusionmatrix

library(ggplot2)

install.packages("yardstick")

library(yardstick)

confusionmatrixviz <- conf\_mat(confusionmatrix)

autoplot(confusionmatrixviz)

summary(confusionmatrixviz, event\_level = "second")

#Summary of categorical data

table(LogMutate\_Churn2$Techie)

prop.table(table(LogMutate\_Churn2$Techie))

table(LogMutate\_Churn2$Port\_modem)

prop.table(table(LogMutate\_Churn2$Port\_modem))

table(LogMutate\_Churn2$Tablet)

prop.table(table(LogMutate\_Churn2$Tablet))

table(LogMutate\_Churn2$Phone)

prop.table(table(LogMutate\_Churn2$Phone))

table(LogMutate\_Churn2$Multiple)

prop.table(table(LogMutate\_Churn2$Multiple))

table(LogMutate\_Churn2$OnlineSecurity)

table(LogMutate\_Churn2$OnlineBackup)

table(LogMutate\_Churn2$DeviceProtection)

table(LogMutate\_Churn2$TechSupport)

table(LogMutate\_Churn2$StreamingTV)

table(LogMutate\_Churn2$StreamingMovies)

table(LogMutate\_Churn2$PaperlessBilling)

table(LogMutate\_Churn2$Area\_Suburban)

table(LogMutate\_Churn2$Area\_Urban)

table(LogMutate\_Churn2$Contract\_OneYear)

table(LogMutate\_Churn2$Contract\_TwoYear)

table(LogMutate\_Churn2$Gender\_Male)

table(LogMutate\_Churn2$Gender\_Nonbinary)

table(LogMutate\_Churn2$InternetService\_FiberOptic)

table(LogMutate\_Churn2$InternetService\_None)

#Reduced model excluded categorical variables as they only provide 0/1

logmodelreduced <- glm(Churn ~ Children + Age + Income + Tenure +

Outage\_sec\_perweek + Email + Contacts +

Yearly\_equip\_failure +

MonthlyCharge + Bandwidth\_GB\_Year + Timely\_Response +

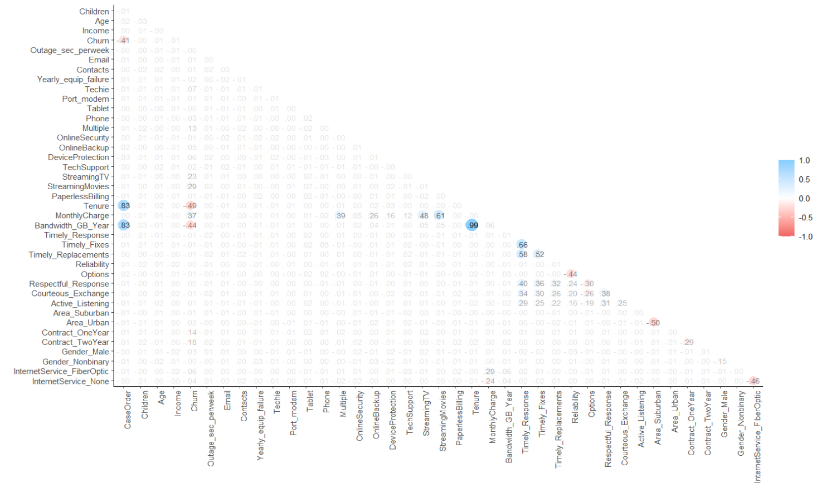
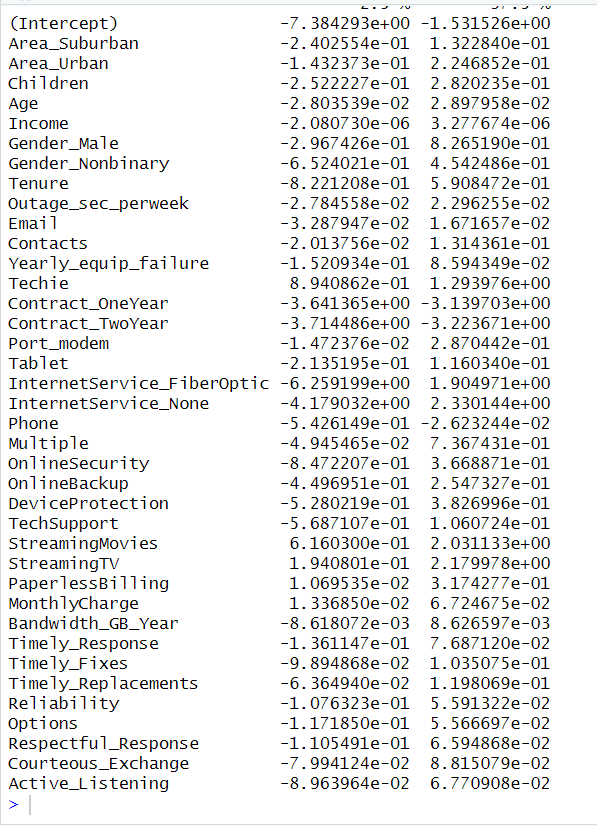
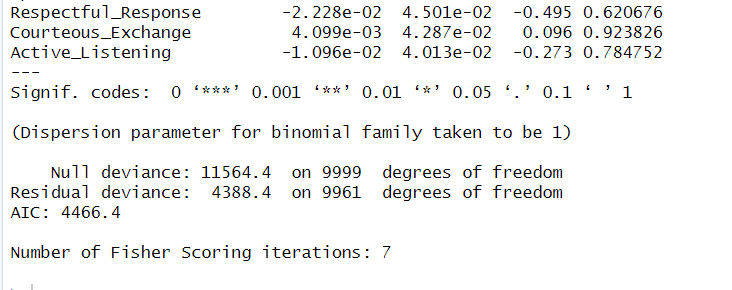
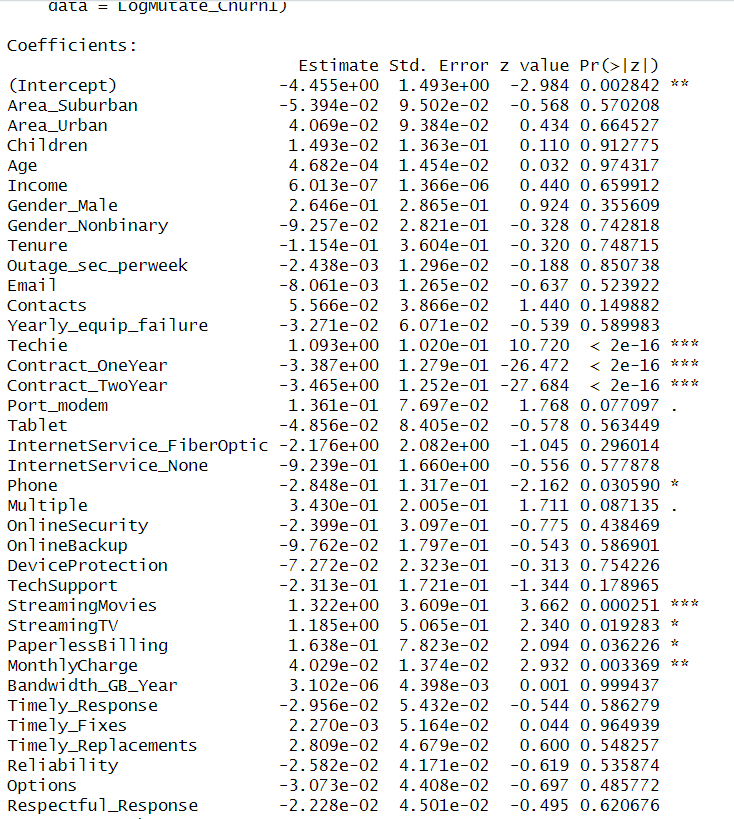
Timely\_Fixes + Timely\_Replacements + Reliability + Options +

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = LogMutate\_Churn1,

family = "binomial")

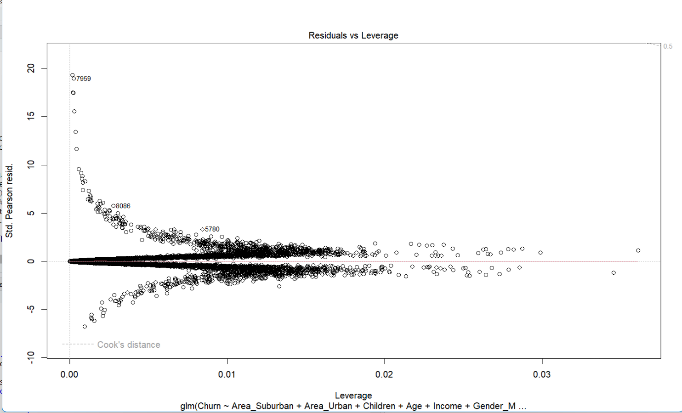
summary(logmodelreduced)

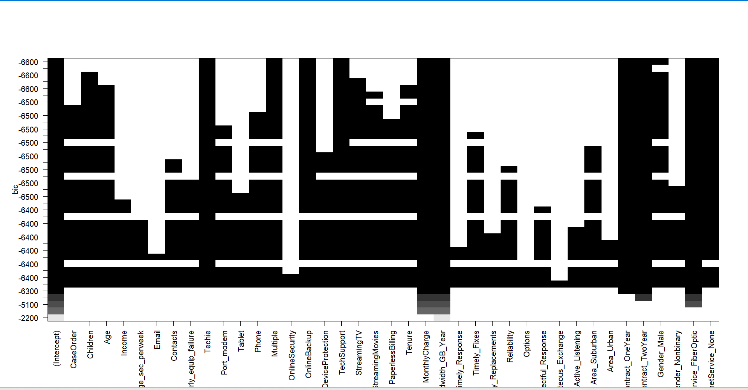
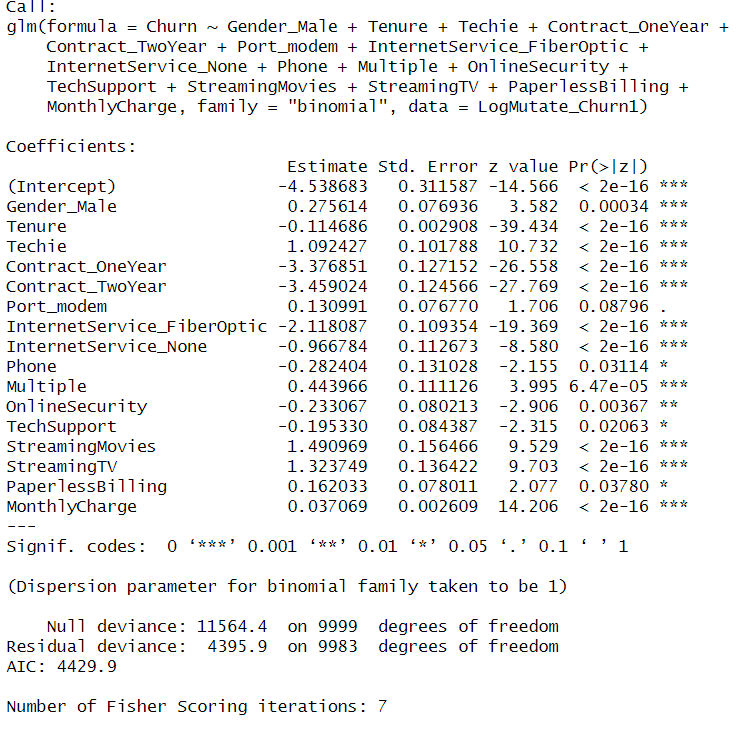
confint(logmodelreduced)

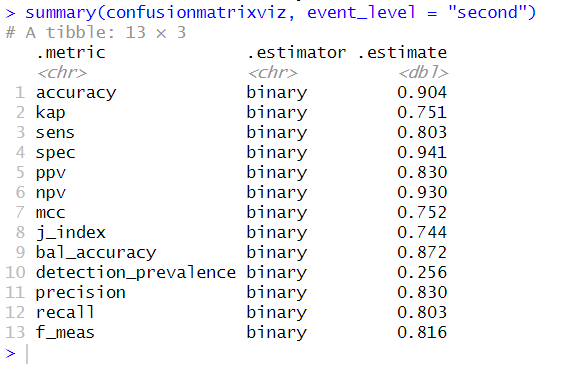
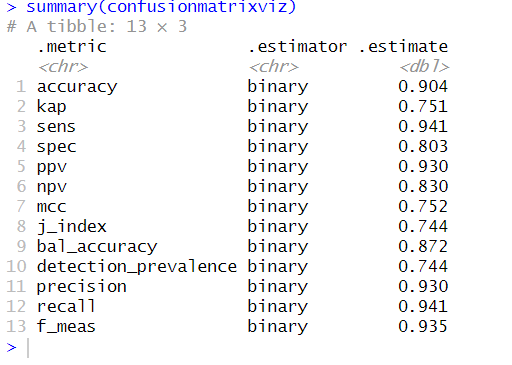
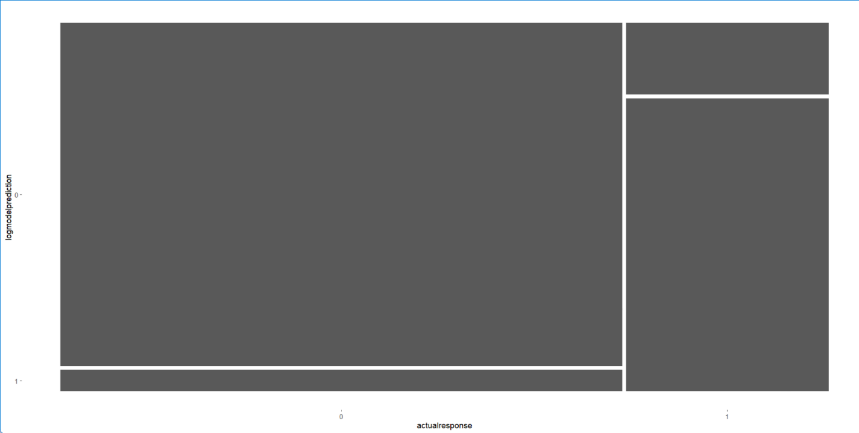
A graph of a function

Description automatically generated with medium confidenceA graph on a white background

Description automatically generatedA graph of a function

Description automatically generated with medium confidenceA close up of numbers

Description automatically generatedA screenshot of a computer code

Description automatically generated

E.  The data set was analyzed using the reduced logistic regression model by doing the following:

1.  The data analysis process includes comparing the initial logistic regression model and reduced logistic regression model, utilizing the AIC model evaluation metric. To begin the analysis process first the data needed to be cleaned and accurately prepared. Once that was complete it was reduced to meet the assumptions of a logistic regression model which are explained in section B1. Next, the backward stepwise wrapper feature selection reduction method was performed two times to complete the model feature reduction. The AIC on the initial model was 4434.6 while the reduced model has an AIC of 4429.9. This means the reduced model is a better fit than the initial model as it has a lower AIC. On the initial model, the null deviance was 11564.4 and the residual deviance was 4342.6 while on the reduced model the null deviance was 11,564.4 the same as the initial and the residual deviance on the reduced model was 4395.9 which is slightly higher than the initial model. The null deviance tells us if the response variable can be predicted by a model with only the intercept and the residual deviance tells us if the response variable can be predicted by a model with p predictor variables (Middleton. 2023). We will use the null deviance and residual deviance to calculate Chi Square and determine if the model is useful. To calculate Chi-Square, we subtract residual deviance from the null deviance. 7168.5 = 11564.4-4395.9 (Zach. 2020). We then use the Chi-Square to P-Value calculator with a Chi-Square of 7168.5 and 9983 degrees of freedom to calculate the P-Value of 1 (Zach. 2020). This is more than .05 which means this model is not useful (Middleton, 2023).

2.  The output and *all* calculations of the analysis performed are included below:

#Create confusion matrix

#Try a prediction

actualresponse <- LogMutate\_Churn1$Churn

logmodelprediction <- round(fitted(logmodelreduced))

confusionmatrix <- table(logmodelprediction, actualresponse)

confusionmatrix

library(ggplot2)

install.packages("yardstick")

library(yardstick)

confusionmatrixviz <- conf\_mat(confusionmatrix)

autoplot(confusionmatrixviz)

summary(confusionmatrixviz, event\_level = "second")

#Create reduced model no categorical variables

logmodelreduced <- glm(Churn ~ Children + Age + Income + Tenure +

Outage\_sec\_perweek + Email + Contacts +

Yearly\_equip\_failure +

MonthlyCharge + Bandwidth\_GB\_Year + Timely\_Response +

Timely\_Fixes + Timely\_Replacements + Reliability + Options +

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = LogMutate\_Churn1,

family = "binomial")

summary(logmodelreduced)

confint(logmodelreduced)

#Export

write.csv(LogMutate\_Churn2, "C:/Users/ntrei/OneDrive/Documents/MSDA/LogMutateChurn2.208.csv")

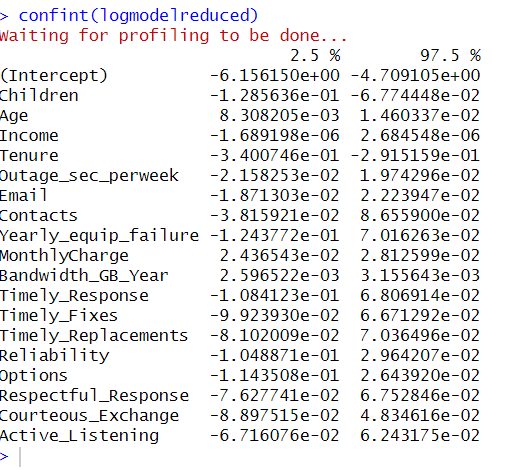
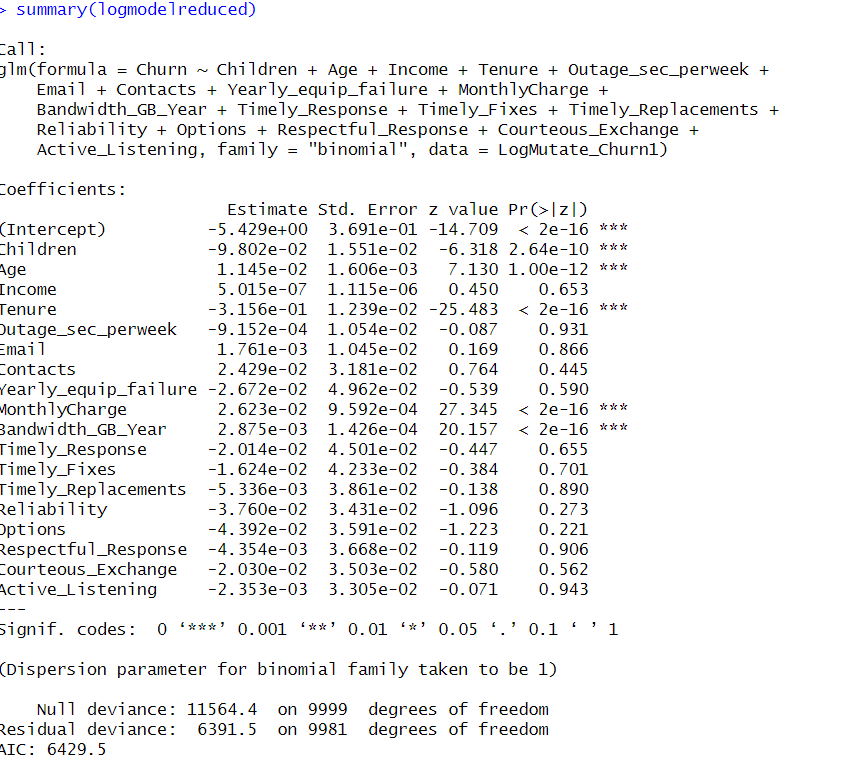
A screenshot of a computer code

Description automatically generatedA grey rectangular object with white lines

Description automatically generatedA screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated



3.  See the attached code.

**Part V: Data Summary and Implications**

F.  Summary of findings and assumptions:

1.  The results of the data analysis, include the following elements:

•   **log[p(X) / (1-p(X))]  =  β0 + β1X1 + β2X2 + … + βpXp (Zach. 2020)**

**log[p(X) / (1-p(X))]  = -5.429e+00 (y-intercept) + -9.082e-02 (Children) + 1.145e-02 (Age) + 5.015e-07 (Income) + -3.156e-04 (Tenure) + -9.152e-04 (Outage\_sec\_perweek) + 1.761e-03 (Email) + 2.429e-02 (Contacts) + -2.672e-02 (Yearly\_equip\_failure) + 2.623e-02 (Monthly charge) + 2.875e-03 (Bandwidth) + -2.014e-02 (Timely response) + -1.624e-02 (Timely fix) + -5.336e-03 (Timely replacements) + -3.760e-02 (Reliability) + -4.392e-02 (Options) + -4.354e-03 (Respectful response) + -2.030e-02 (Courteous exchange) + -2.353e-03 (Active listening).**

•   Logistic regression is a classification algorithm. The coefficients of the reduced model list the y-intercept as **-5.429e+00** and then go on to list the coefficient for the independent variables such as the Tenure coefficient at -3.156e-01 which is a negative meaning it is less likely that there is a relationship between Tenure and Churn while all other variables remain constant. The z value of Tenure is -25.483 which is significantly less than the desired 2 or greater which would be a significant value. The p-value of Tenure is significant at the .001 level with a p-value of <2e-16. Children's coefficient at **-9.082e-02 being a negative means it is less likely there is a relationship between Children and Churn while all other variables remain constant**. The z value of Children is -6.318 which is significantly less than the desired 2 or greater which would be significant. The p-value of Children is significant at the .001 level with a p-value of <2.64e-10. The age coefficient at **1.145e-02 being positive means it is more likely that there is a relationship between Age and Churn while all other variables remain constant**. The z value of Age is 7.130 which is significantly more than the desired 2 or greater which would be a significant value. The p-value of Age is significant at the .001 level with a p-value of 1.00e-12. The income coefficient at **5.015e-07 is more likely to have a relationship between Income and Churn, due to being positive, while all other variables remain constant**. The z value of Income is .450, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Income is insignificant due to being greater than .05 with a p-value of .653. Outage\_sec\_perweek coefficient at **-9.152e-04 is less likely to have a relationship between Outage\_sec\_perweek and Churn because it is negative, as long as all other variables remain constant**. The z value of Outage\_sec\_perweek is -.087 which is significantly less than the desired 2 or greater which would be a significant value. The p-value of Outage\_sec\_perweek is insignificant with a p-value of .931. The Email coefficient at **1.761e-03 is more likely to have a relationship between Email and Churn as long as all other variables remain constant**. The z value of Email is .169 which is significantly less than the desired 2 or greater which would be a significant value. The p-value of Email is insignificant with a p-value of .866. The Contacts coefficient at **2.429e-02 is more likely to have a relationship between Contacts and Churn as long as all other variables remain constant**. The z value of Contacts is .769, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Contacts is insignificant with a p-value of .445. Yearly\_equip\_failure coefficient at **-2.672e-02 is less likely as a negative to have a relationship between Yearly\_equip\_failures and Churn as long as all other variables remain constant**. The z value of Yearly\_equip\_failure is -.539, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Yearly\_equip\_failure is insignificant with a p-value of .590. Monthlycharge coefficient at 2.623e-02 is more likely to have a relationship between Monthlycharge and Churn as long as all other variables remain constant. The z value of Monthlycharge is 27.345, significantly more than the desired 2 or greater, which would be significant. The p-value of Monthlycharge is significant at the .001 level with a p-value of <2e-16. The bandwidth coefficient is 2.875e-03 which is more likely to have a relationship between Bandwidth and Churn as long as all other variables remain constant. The z value of Bandwidth is 20.157 which is significantly more than the desired 2 or greater which would be a significant value. The p-value of Bandwidth is significant at the .001 level with a p-value of <2e-16. The timely response coefficient at -2.014e-02 is negative and is less likely to have a relationship between Timely Response and Churn. The z value of Timely response is .447, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Timely response is insignificant with a p-value of .655. Timely fixes coefficient at -1.624e-02 is less likely to have a relationship between Timely Fixes and Churn because it is negative. The z value of Timely fixes is -.384 which is significantly less than the desired 2 or greater which would be a significant value. The p-value of Timely fixes is insignificant with a p-value of .701. Timely replacements coefficient at -5.336e-03 is less likely to have a relationship between Timely Replacements and Churn because it is negative. The z value of Timely replacements is -.138 which is significantly less than the desired 2 or greater which would be a significant value. The p-value of Timely replacements is insignificant with a p-value of .890. The reliability coefficient at -3.760e-02 is less likely to have a relationship between Reliability and Churn because it is negative, as long as all other variables remain constant. The z value of Reliability is -1.096, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Reliability is insignificant with a p-value of .273. The options coefficient at -4.392e-02 is less likely to have a relationship between Options and Churn because it is negative, as long as all other variables remain constant. The z value of Options is -1.223, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Options is insignificant with a p-value of .221. The respectful response coefficient at -4.354e-03 is less likely to have a relationship between Respectful response and Churn because it is negative, as long as all other variables remain constant. The z value of Respectful response is -.119 which is significantly less than the desired 2 or greater which would be a significant value. The p-value of Respectful response is insignificant with a p-value of .906. Courteous exchange coefficient at -2.00e-02 is less likely to have a relationship between Courteous Exchange and Churn as it is negative, as long as all other variables remain constant. The z value of Courteous exchange is -.580, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Courteous exchange is insignificant with a p-value of .562. The active listening coefficient at -2.353e-03 is less likely to have a relationship between Active Listening and Churn because it is negative, as long as all other variables remain constant. The z value of Active listening is -.071, significantly less than the desired 2 or greater, which would be a significant value. The p-value of Active listening is insignificant with a p-value of .943.

•   This logistic regression model demonstrates that there is statistical significance, and correlation between Churn and some of the independent variables, meaning Churn is not based on chance. This logistic regression model is practically significant because the information found to be statistically significant would be utilized to make business decisions.

•   The limitation of the data analysis is that correlation does not necessarily mean causation.

2.  Based on these results of the research question, “What factors affect churn?”, the recommended course of action is to utilize this information to address more narrow business questions and to reduce the model further. We now know that certain variables have a direct relationship to Churn. The recommended course of action would be to address these findings to reduce Churn.

**Part VI: Demonstration**

G.  See attached Panopto recording. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b9f9feeb-5a39-4c22-a9ce-b0e400eecd34>

**References**

Zach. (2020). *Introduction to logistic regression.*

[Introduction to Logistic Regression - Statology](https://www.statology.org/logistic-regression/)

Kassambara. (2018). *Logistic regression assumptions in R.*

[Logistic Regression Assumptions and Diagnostics in R - Articles - STHDA](http://sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/)

Zach. (2020). *Chi-square to p value calculator.*

[Chi-Square Score to P Value Calculator - Statology](https://www.statology.org/chi-square-p-value-calculator/)

Zach. (2020). *How to interpret null & residual deviance.*

[How to Interpret Null & Residual Deviance (With Examples) - Statology](https://www.statology.org/null-residual-deviance/)

Middleton, K. (2023). *Getting started with d208 part 2.*

[D208 - Webinar: Getting Started with D208 Part II (November) (panopto.com)](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=39bbe2db-de7d-4bf5-913b-af5c0003da9d)