­

**Linear Regression Modeling D208**

Nicole Reiswig

College of Information Technology, Western Governors University

Keiona Middleton

November 01, 2023

**Linear Regression Modeling D208**

**Table of Contents**

A1 Research Question 3

A2 Objectives and Goals for Analysis 3

B1 Assumptions 3

B2 Programming Language and Benefit 3

B3 Justification of using Regression 3

C1 Data Cleaning 3

C2 Data Exploration 8

C3 Visualizations 10

C4 Data Transformation (Data Wrangling) 65

C5 Prepared Dataset 70

D1 Initial Model 70

D2 Model Reduction Method and Justification 72

D3 Reduced Model 72

E1 Model Comparison 73

E2 Multiple Linear Regression 74

E3 Code 100

F1 Regression Equation, Coefficients 100

F2 Recommendations 100

G Panopto Recording 100

H Sources 101

**Part I: Research Question**

A.  The purpose of this data analysis is the following:

1.  The research question that will be answered utilizing multiple linear regression modeling is, “What factors affect length of tenure?”

2.  The goals of the analysis are to determine which variables have an impact on the length of tenure. In determining which factors have the greatest impact, this information can be used to increase tenure which will result in positive cash flow.

**Part II: Method Justification**

B.

1.  The four assumptions of a multiple linear regression model according to (Zach, 2020) are that there is a linear relationship between the x and y variables, the residuals are independent, there is homoscedasticity and normality. A linear relationship between x and y mean that there is a relationship in the form of a straight line and is checked through visualizations. In order for the residuals to be independent this means there must not be correlation among the variables and that they independent and random this is checked through visualizations. Homoscedasticity is when the error does not change across the values of of the independent variables (Middleton, (2023) and this is verified through VIF the variance inflation factor. Lastly normality is when the residuals are normally distributed and this is checked through visualizations.

2.  R is the programming language used for thimeanss multiple linear 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, and olsrr.

3.  Multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I because (Zach. 2020) linear regression is the most commonly used technique in statistics and it is used to quantify the relationship between predictor variables and response variables. In this specific analysis we are quantifying the relationship between the predictor variable Tenure which is a continuous variable and all the other variables in the churn data set which are the independent variables. The multiple linear regression analysis provides us insight as to how strong a relationship is between the dependent and independent variables. Therefore the multiple linear regression analysis is a great technique for answering business questions with large data sets and finding the most significant factors affecting the business model. We utilized the kitchen sink approach in this analysis and utilized all the variables and reduced the model to fit. This narrowed down the results of which variables have a significant relationship with how long a customers length of tenure is with the company.

**Part III: Data Preparation**

C.  The data preparation process for multiple linear regression analysis is the following:

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

#checking working directory

getwd()

#data profiling

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

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

dim(churn\_clean208)

library(tidyverse)

glimpse(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

b <- boxplot(churn\_clean208$CaseOrder)

b <- boxplot(churn\_clean208$Zip)

b <- boxplot(churn\_clean208$Lat)

b <- boxplot(churn\_clean208$Lng)

b <- boxplot(churn\_clean208$Population)

b <- boxplot(churn\_clean208$Children)

b <- boxplot(churn\_clean208$Age)

b <- boxplot(churn\_clean208$Income)

b <- boxplot(churn\_clean208$Outage\_sec\_perweek)

b <- boxplot(churn\_clean208$Email)

b <- boxplot(churn\_clean208$Contacts)

b <- boxplot(churn\_clean208$Yearly\_equip\_failure)

b <- boxplot(churn\_clean208$Tenure)

b <- boxplot(churn\_clean208$MonthlyCharge)

b <- boxplot(churn\_clean208$Bandwidth\_GB\_Year)

b <- boxplot(churn\_clean208$Item1)

b <- boxplot(churn\_clean208$Item2)

b <- boxplot(churn\_clean208$Item3)

b <- boxplot(churn\_clean208$Item4)

b <- boxplot(churn\_clean208$Item5)

b <- boxplot(churn\_clean208$Item6)

b <- boxplot(churn\_clean208$Item7)

b <- boxplot(churn\_clean208$Item8)

#count and range of variables with outliers

children\_query <- churn\_clean208[which(churn\_clean208$Children > 7), ]

str(children\_query)

income\_query <- churn\_clean208[which(churn\_clean208$Income > 100000), ]

str(income\_query)

osw\_query <- churn\_clean208[which(churn\_clean208$Outage\_sec\_perweek > 20), ]

str(osw\_query)

osw2\_query <- churn\_clean208[which(churn\_clean208$Outage\_sec\_perweek < 0), ]

str(osw2\_query)

email\_query <- churn\_clean208[which(churn\_clean208$Email > 20), ]

str(email\_query)

email2\_query <- churn\_clean208[which(churn\_clean208$Email < 4), ]

str(email2\_query)

contacts\_query <- churn\_clean208[which(churn\_clean208$Contacts > 5), ]

str(contacts\_query)

yef\_query <- churn\_clean208[which(churn\_clean208$Yearly\_equip\_failure > 2), ]

str(yef\_query)

mc\_query <- churn\_clean208[which(churn\_clean208$MonthlyCharge > 300), ]

str(mc\_query)

item1\_query <- churn\_clean208[which(churn\_clean208$item1 > 5), ]

str(item1\_query)

item1.1\_query <- churn\_clean208[which(churn\_clean208$item1 < 2), ]

str(item1.1\_query)

item2\_query <- churn\_clean208[which(churn\_clean208$item2 > 5), ]

str(item1\_query)

item2.1\_query <- churn\_clean208[which(churn\_clean208$item2 < 2), ]

str(item2.1\_query)

#count and range of item3

item3\_query <- churn\_clean208[which(churn\_clean208$item3 > 5), ]

str(item3\_query)

item3.1\_query <- churn\_clean208[which(churn\_clean208$item3 < 2), ]

str(item3.1\_query)

item4\_query <- churn\_clean208[which(churn\_clean208$item4 > 5), ]

str(item4\_query)

item4.1\_query <- churn\_clean208[which(churn\_clean208$item4 < 2), ]

str(item4.1\_query)

item5\_query <- churn\_clean208[which(churn\_clean208$item5 > 5), ]

str(item5\_query)

item5.1\_query <- churn\_clean208[which(churn\_clean208$item5 < 2), ]

str(item5.1\_query)

item6\_query <- churn\_clean208[which(churn\_clean208$item6 > 5), ]

str(item6\_query)

item6.1\_query <- churn\_clean208[which(churn\_clean208$item6 < 2), ]

str(item6.1\_query)

item7\_query <- churn\_clean208[which(churn\_clean208$item7 > 5), ]

str(item7\_query)

item7.1\_query <- churn\_clean208[which(churn\_clean208$item7 < 2), ]

str(item7.1\_query)

item8\_query <- churn\_clean208[which(churn\_clean208$item8 > 5), ]

str(item8\_query)

item8.1\_query <- churn\_clean208[which(churn\_clean208$item8 < 2), ]

str(item8.1\_query)

2.  The dependent variable is the Tenure variable and all independent variables in the churn datatset are included in the summary statistics and are required to answer the research question.

> #summary statistics

> summary(churn\_clean208)

CaseOrder Customer\_id Interaction UID City

Min. : 1 Length:10000 Length:10000 Length:10000 Length:10000

1st Qu.: 2501 Class :character Class :character Class :character Class :character

Median : 5000 Mode :character Mode :character Mode :character Mode :character

Mean : 5000

3rd Qu.: 7500

Max. :10000

State County Zip Lat Lng

Length:10000 Length:10000 Min. : 601 Min. :17.97 Min. :-171.69

Class :character Class :character 1st Qu.:26293 1st Qu.:35.34 1st Qu.: -97.08

Mode :character Mode :character Median :48870 Median :39.40 Median : -87.92

Mean :49153 Mean :38.76 Mean : -90.78

3rd Qu.:71867 3rd Qu.:42.11 3rd Qu.: -80.09

Max. :99929 Max. :70.64 Max. : -65.67

Population Area TimeZone Job Children

Min. : 0 Length:10000 Length:10000 Length:10000 Min. : 0.000

1st Qu.: 738 Class :character Class :character Class :character 1st Qu.: 0.000

Median : 2910 Mode :character Mode :character Mode :character Median : 1.000

Mean : 9757 Mean : 2.088

3rd Qu.: 13168 3rd Qu.: 3.000

Max. :111850 Max. :10.000

Age Income Marital Gender Churn

Min. :18.00 Min. : 348.7 Length:10000 Length:10000 Min. :0.000

1st Qu.:35.00 1st Qu.: 19224.7 Class :character Class :character 1st Qu.:0.000

Median :53.00 Median : 33170.6 Mode :character Mode :character Median :0.000

Mean :53.08 Mean : 39806.9 Mean :0.265

3rd Qu.:71.00 3rd Qu.: 53246.2 3rd Qu.:1.000

Max. :89.00 Max. :258900.7 Max. :1.000

Outage\_sec\_perweek Email Contacts Yearly\_equip\_failure Techie

Min. : 0.09975 Min. : 1.00 Min. :0.0000 Min. :0.000 Min. :0.0000

1st Qu.: 8.01821 1st Qu.:10.00 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000

Median :10.01856 Median :12.00 Median :1.0000 Median :0.000 Median :0.0000

Mean :10.00185 Mean :12.02 Mean :0.9942 Mean :0.398 Mean :0.1679

3rd Qu.:11.96949 3rd Qu.:14.00 3rd Qu.:2.0000 3rd Qu.:1.000 3rd Qu.:0.0000

Max. :21.20723 Max. :23.00 Max. :7.0000 Max. :6.000 Max. :1.0000

Contract Port\_modem Tablet InternetService Phone

Length:10000 Min. :0.0000 Min. :0.0000 Length:10000 Min. :0.0000

Class :character 1st Qu.:0.0000 1st Qu.:0.0000 Class :character 1st Qu.:1.0000

Mode :character Median :0.0000 Median :0.0000 Mode :character Median :1.0000

Mean :0.4834 Mean :0.2991 Mean :0.9067

3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000

Max. :1.0000 Max. :1.0000 Max. :1.0000

Multiple OnlineSecurity OnlineBackup DeviceProtection TechSupport

Min. :0.0000 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.:0.0000 1st Qu.:0.000

Median :0.0000 Median :0.0000 Median :0.0000 Median :0.0000 Median :0.000

Mean :0.4608 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.:1.0000 3rd Qu.:1.000

Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.000

StreamingTV StreamingMovies PaperlessBilling PaymentMethod Tenure

Min. :0.0000 Min. :0.000 Min. :0.0000 Length:10000 Min. : 1.000

1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 Class :character 1st Qu.: 7.918

Median :0.0000 Median :0.000 Median :1.0000 Mode :character 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 Timely\_Replacements

Min. : 79.98 Min. : 155.5 Min. :1.000 Min. :1.000 Min. :1.000

1st Qu.:139.98 1st Qu.:1236.5 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000

Median :167.48 Median :3279.5 Median :3.000 Median :4.000 Median :3.000

Mean :172.62 Mean :3392.3 Mean :3.491 Mean :3.505 Mean :3.487

3rd Qu.:200.73 3rd Qu.:5586.1 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000

Max. :290.16 Max. :7159.0 Max. :7.000 Max. :7.000 Max. :8.000

Reliability Options Respectful\_Response Courteous\_Exchange Active\_Listening

Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000

1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:3.00 1st Qu.:3.000

Median :3.000 Median :3.000 Median :3.000 Median :4.00 Median :3.000

Mean :3.498 Mean :3.493 Mean :3.497 Mean :3.51 Mean :3.496

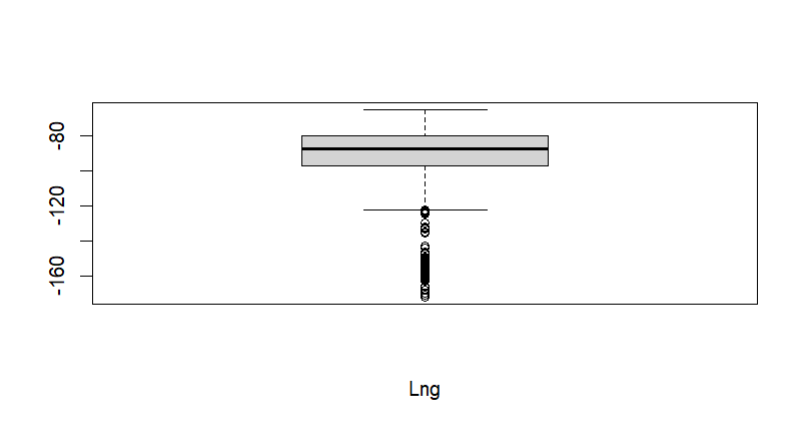
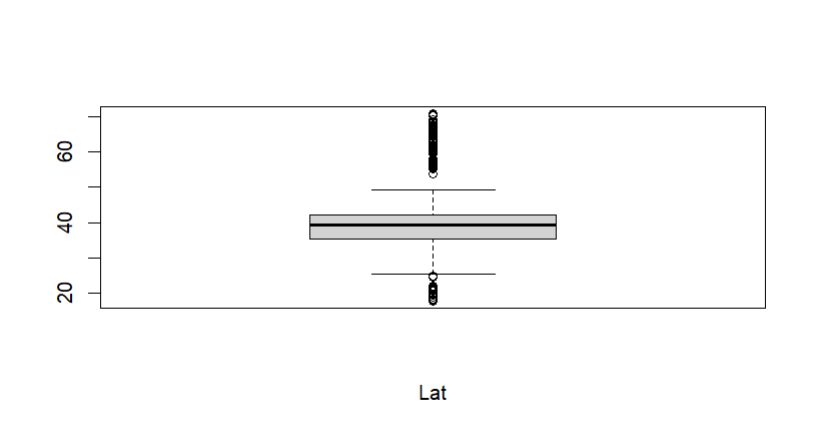
3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.00 3rd Qu.:4.000

Max. :7.000 Max. :7.000 Max. :8.000 Max. :7.00 Max. :8.000

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

A diagram of a case order

Description automatically generatedA diagram of a zip code

Description automatically generatedA diagram of a graph

Description automatically generatedA diagram of a diagram

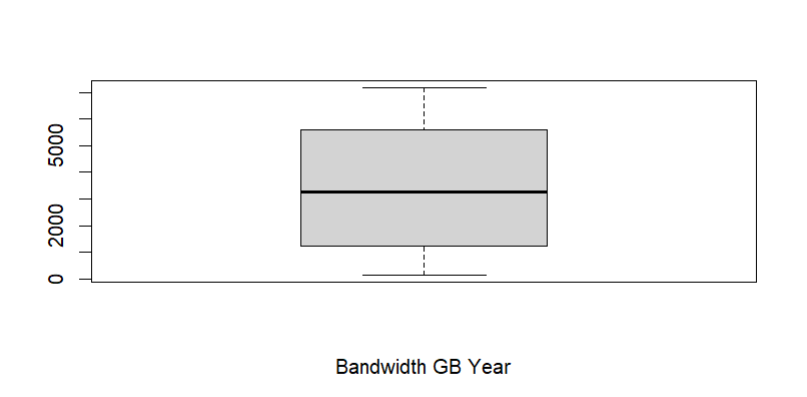
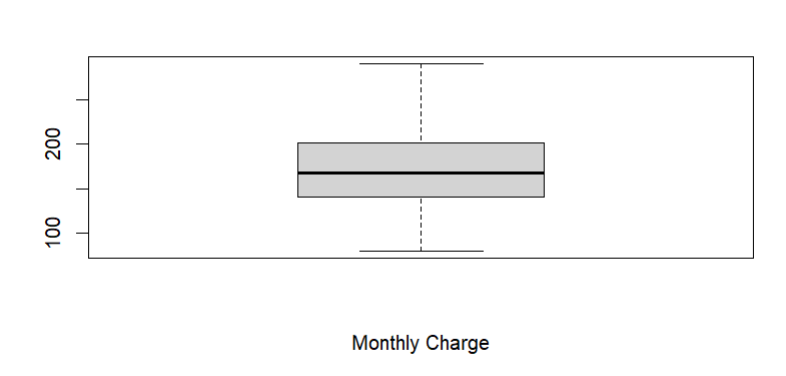
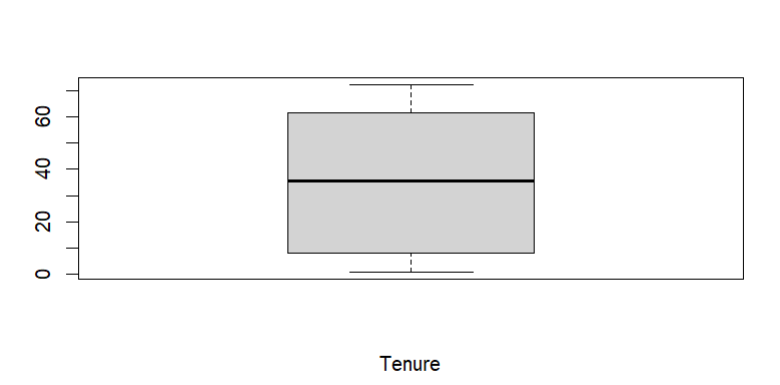
Description automatically generatedA diagram of income

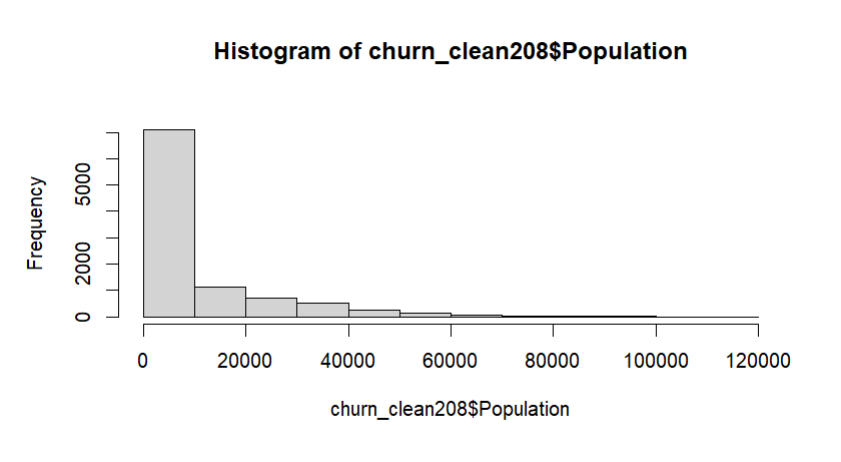
Description automatically generatedA diagram of a line

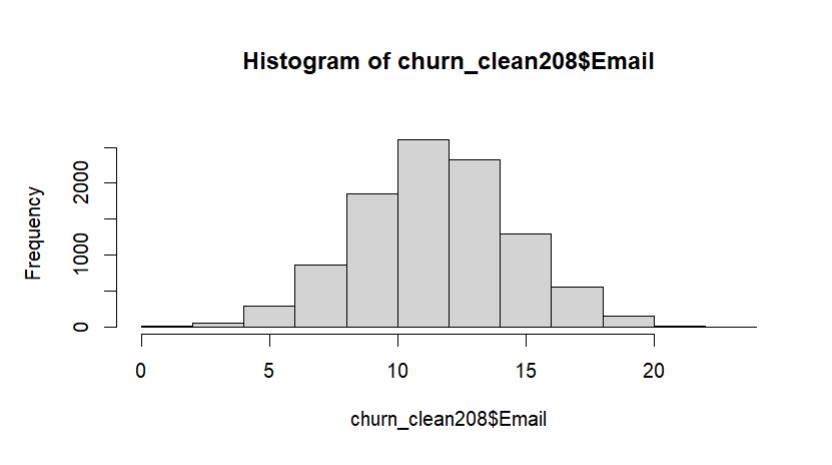
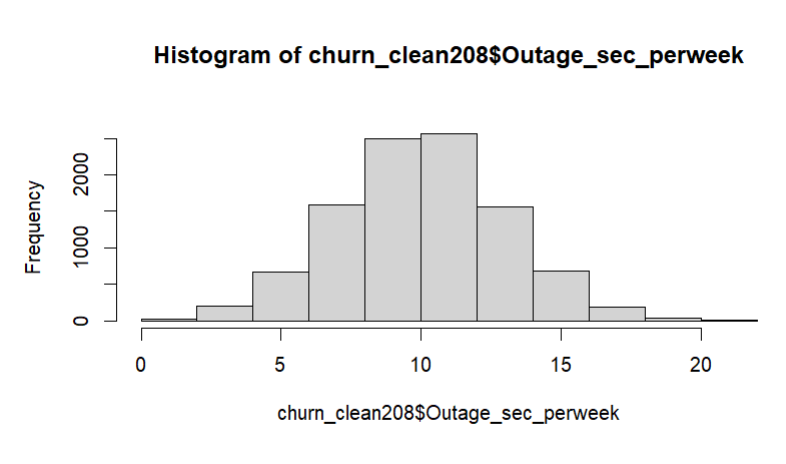
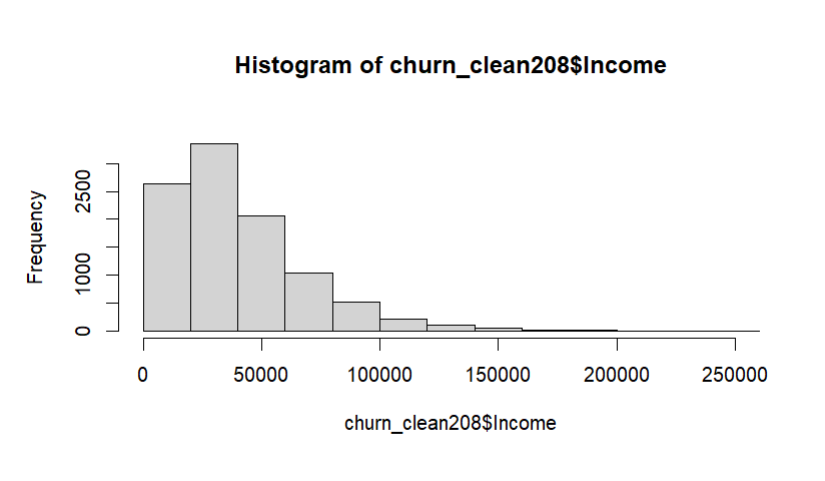
Description automatically generatedA line drawing of a line

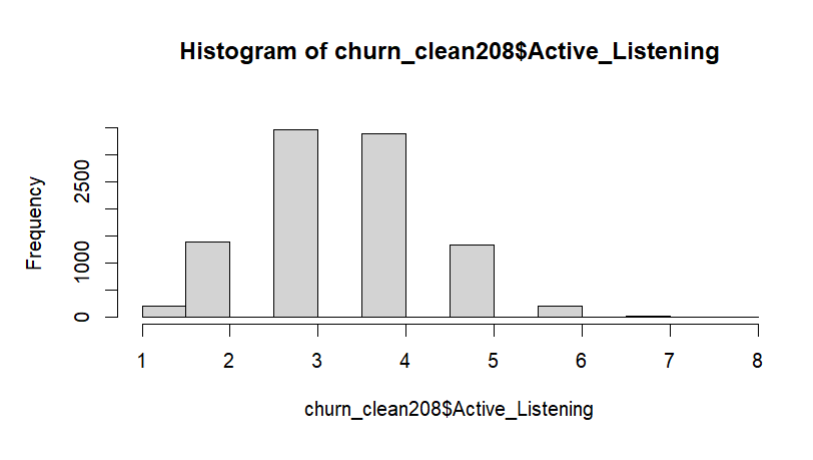
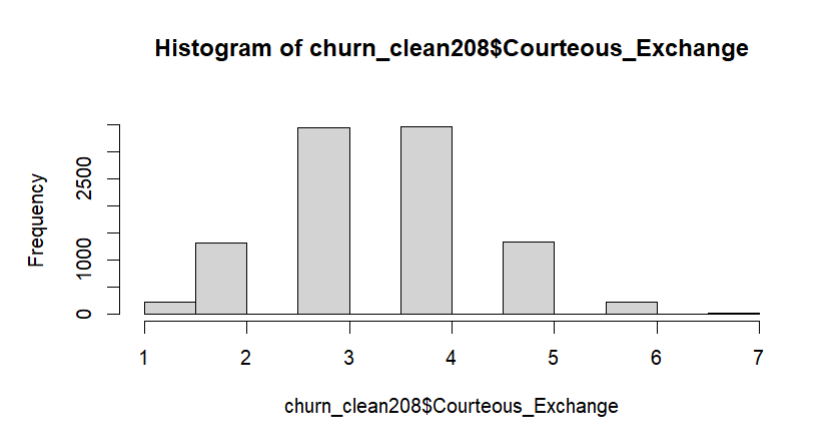
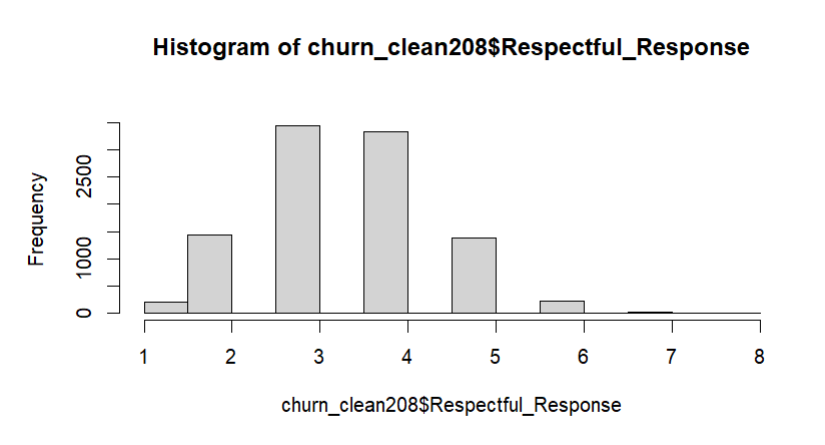
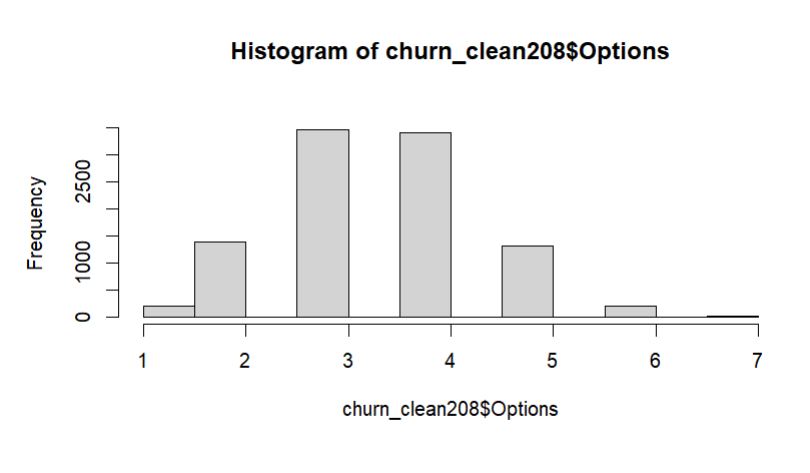
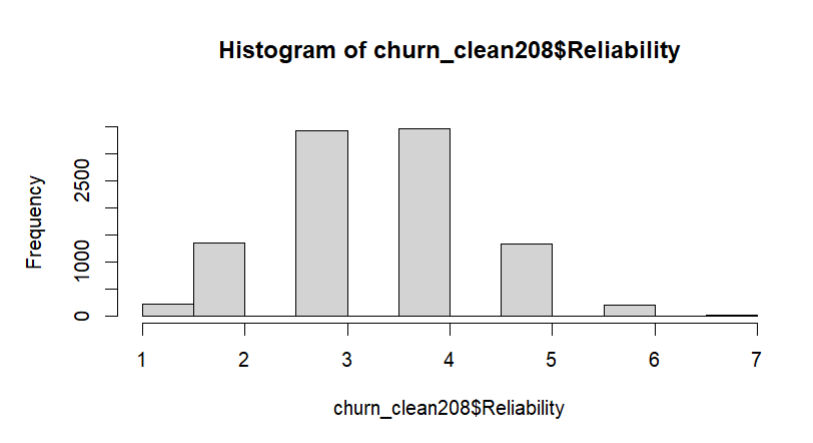
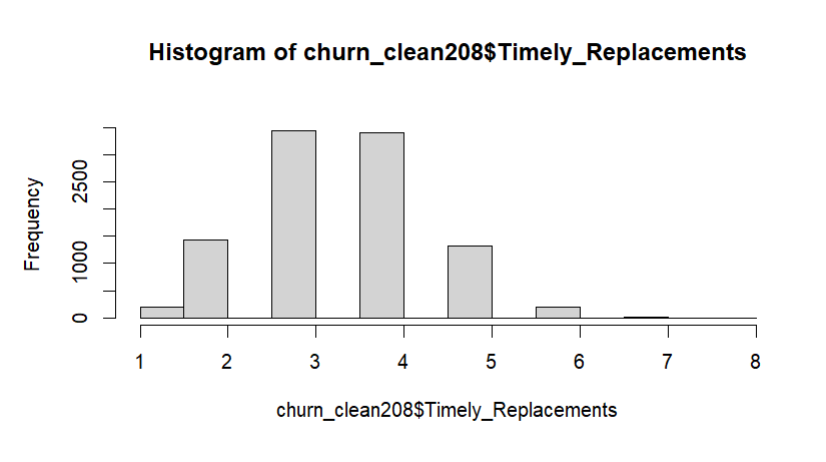
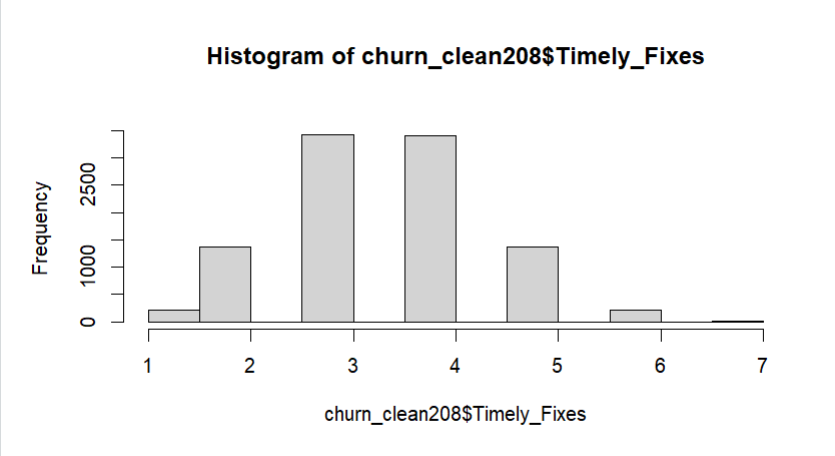
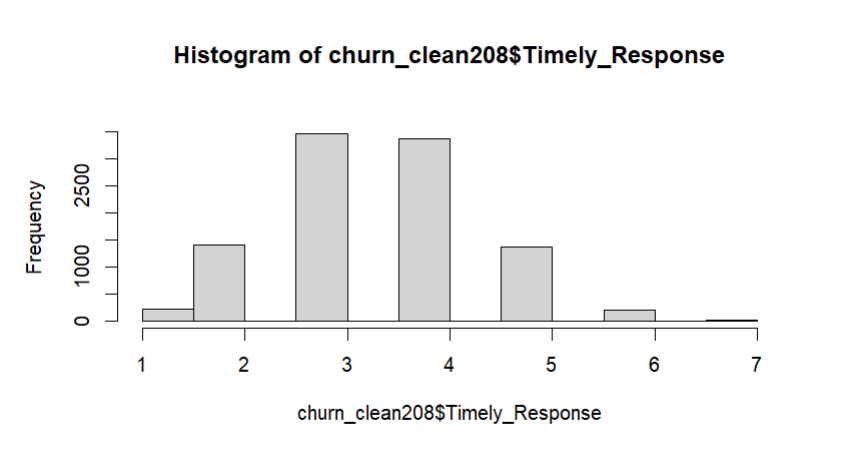
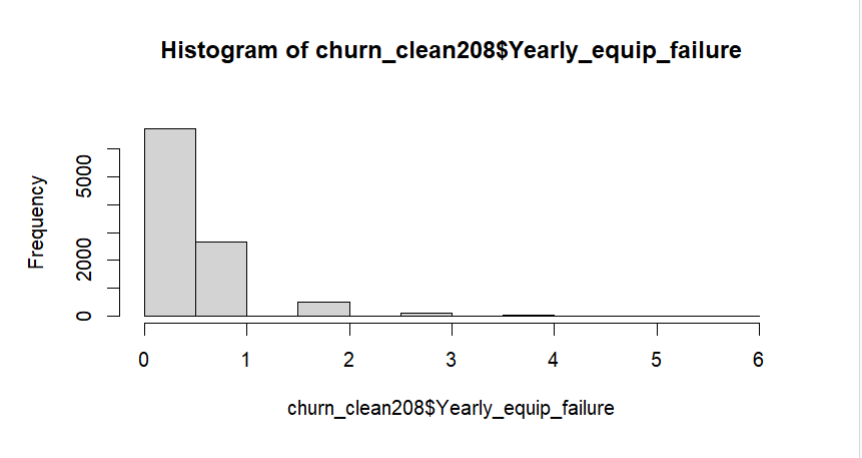
Description automatically generated with medium confidenceA diagram of a graph

Description automatically generated with medium confidenceA diagram of a graph

Description automatically generated

A graph of a number of children

Description automatically generated with medium confidenceA graph of a graph

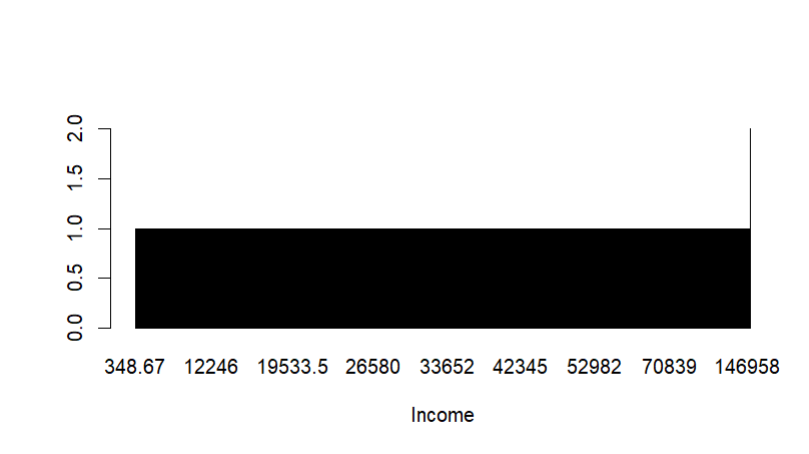
Description automatically generated with medium confidence

A row of squares with text

Description automatically generatedA graph of a number of people

Description automatically generated with medium confidenceA graph of a number of children

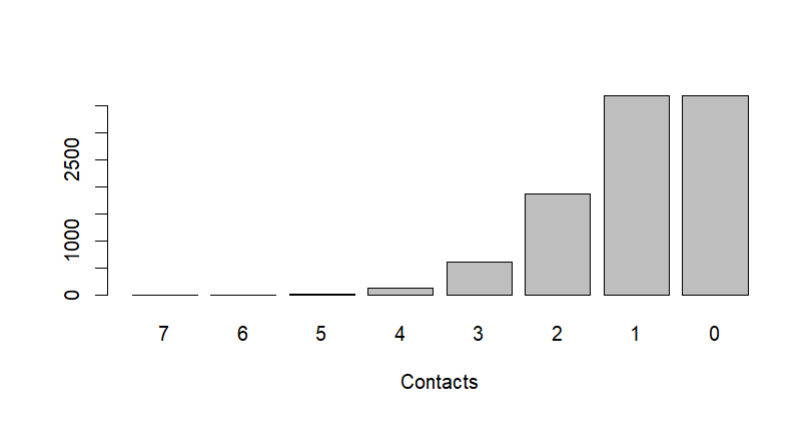
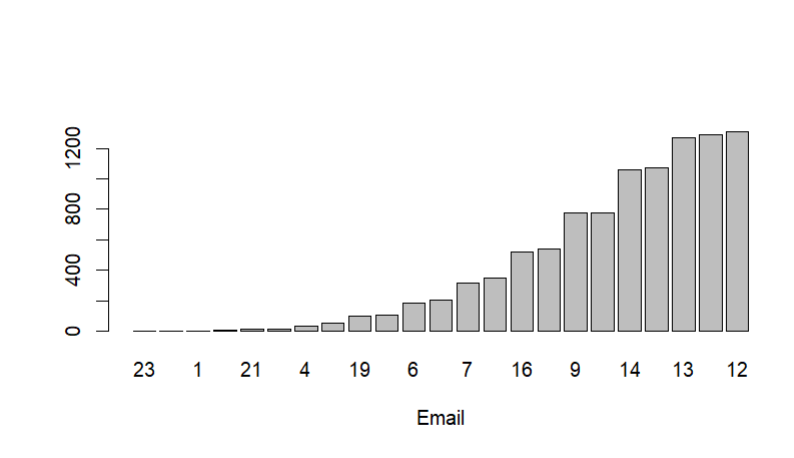
Description automatically generatedA graph of age and age

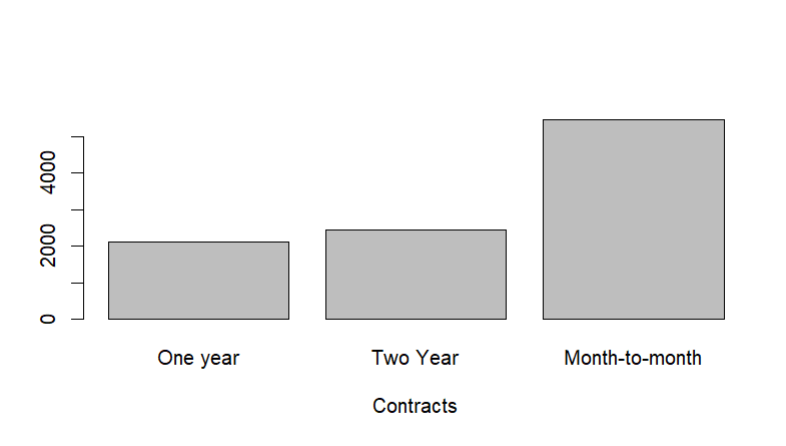
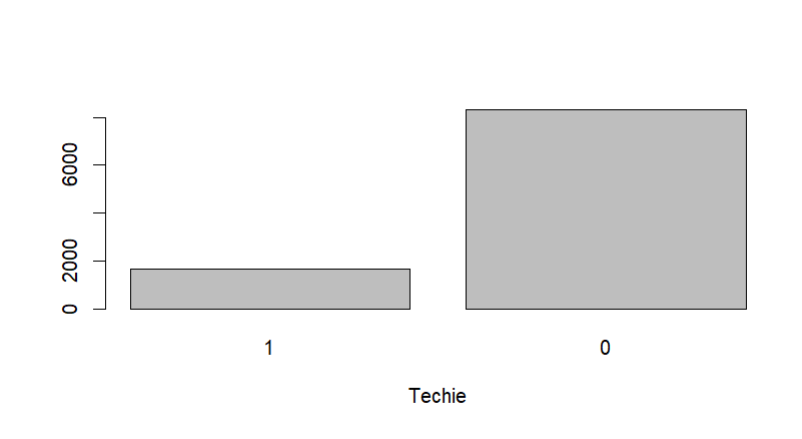
Description automatically generatedA row of rectangular objects

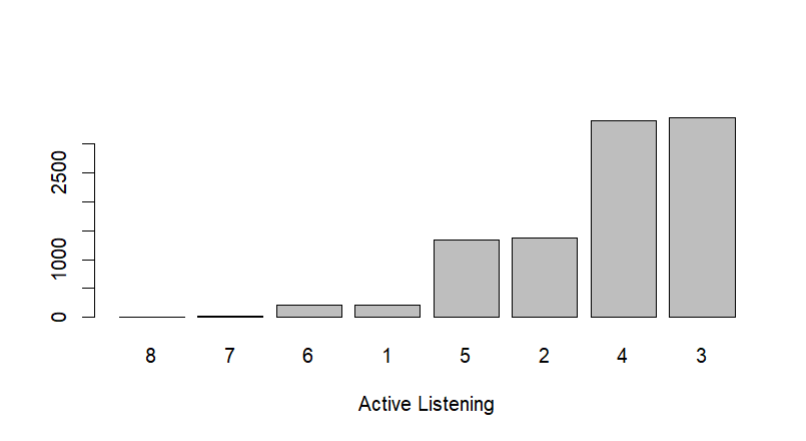
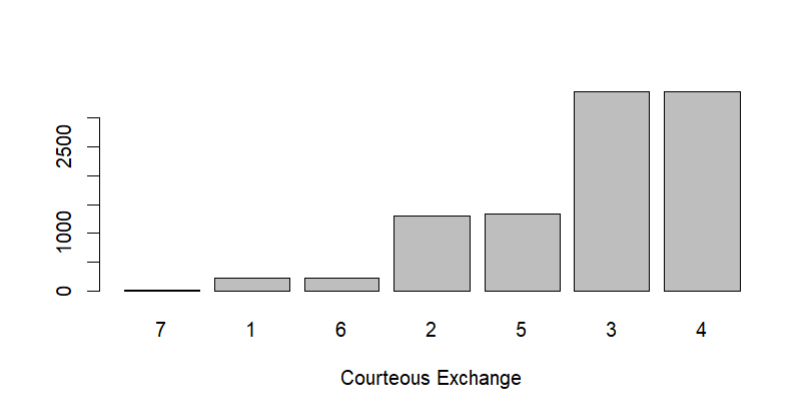
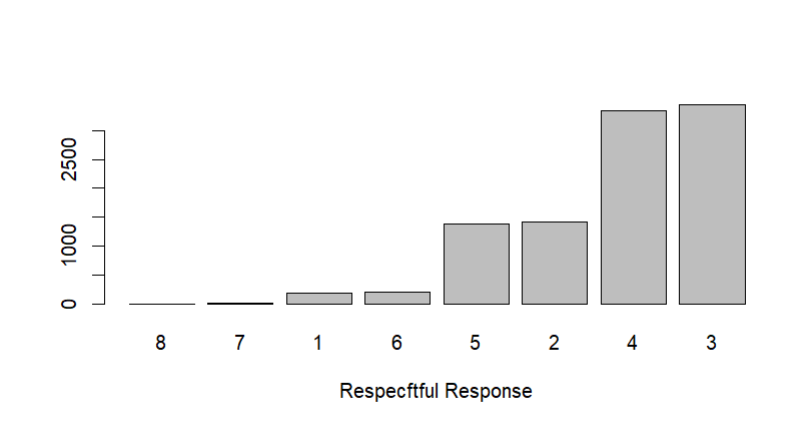
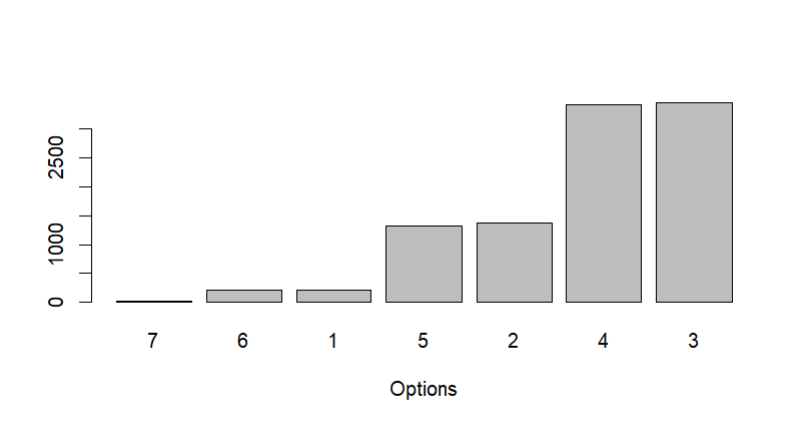
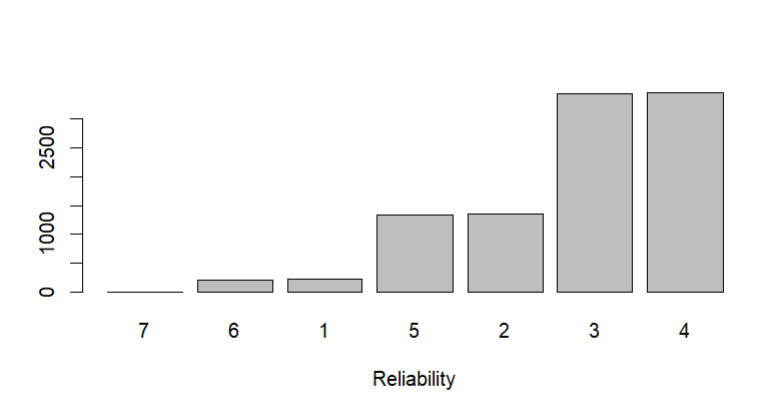
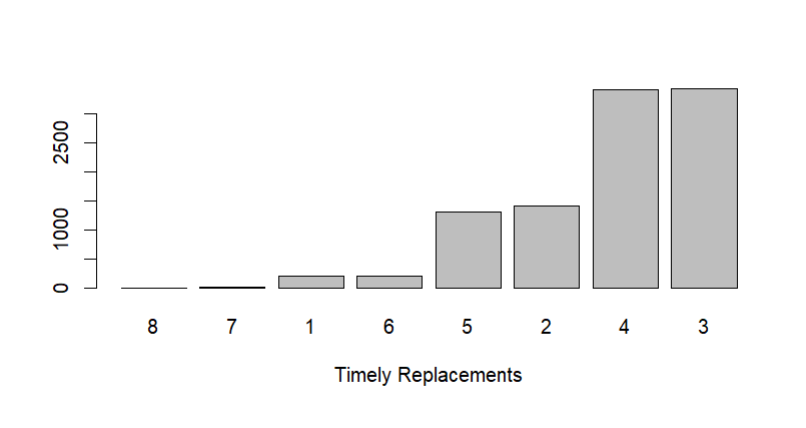
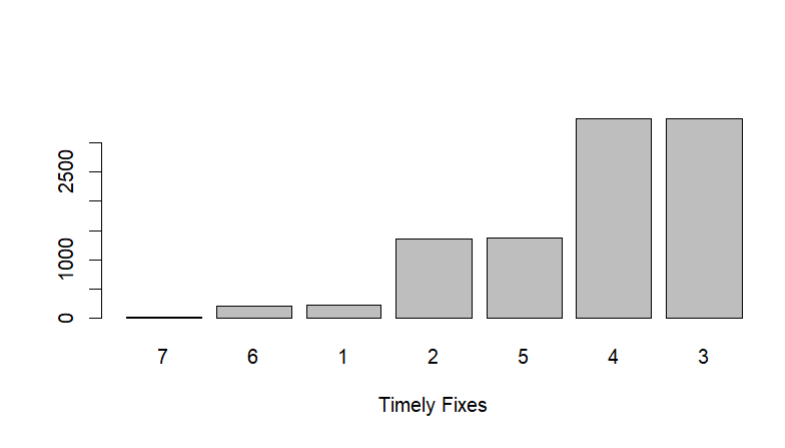
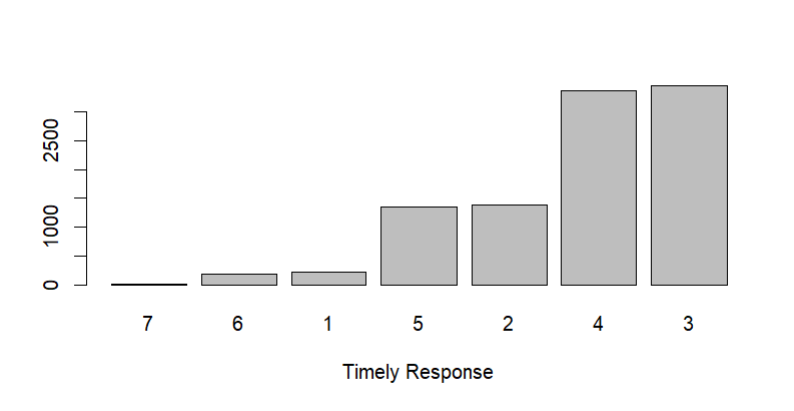
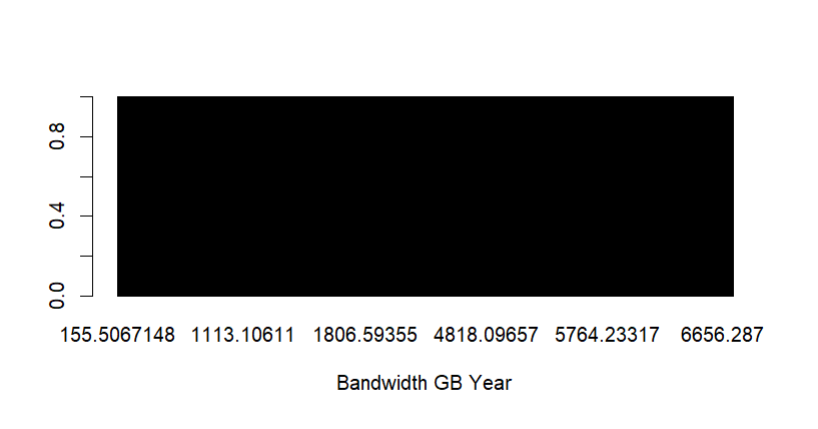
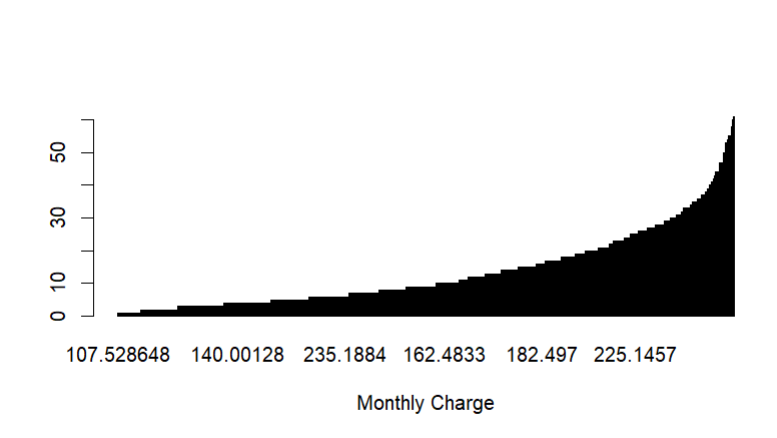
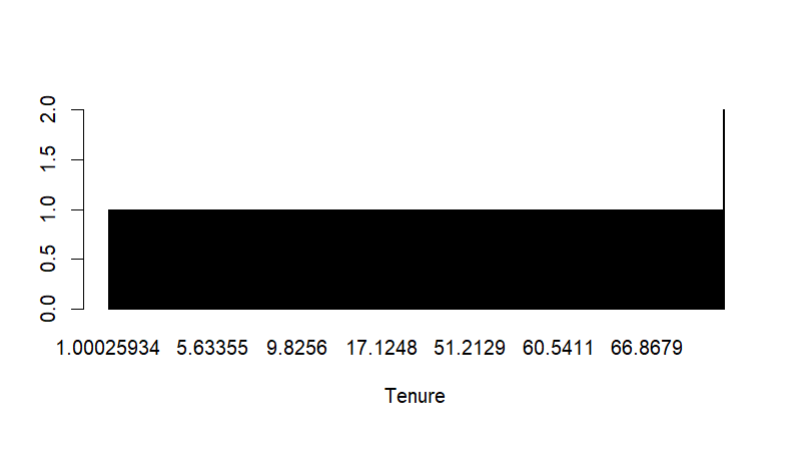
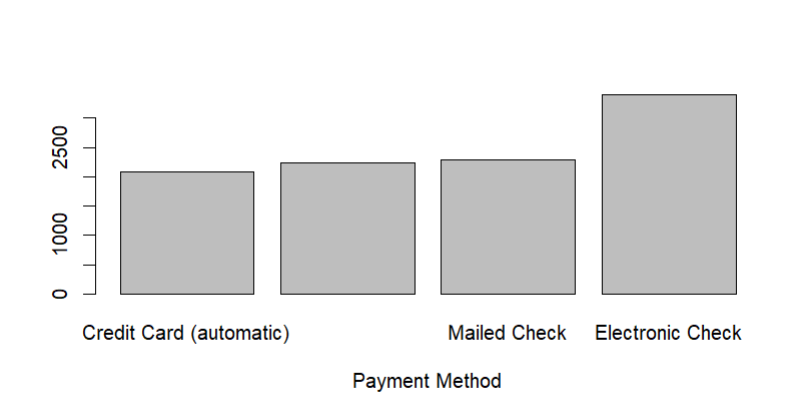
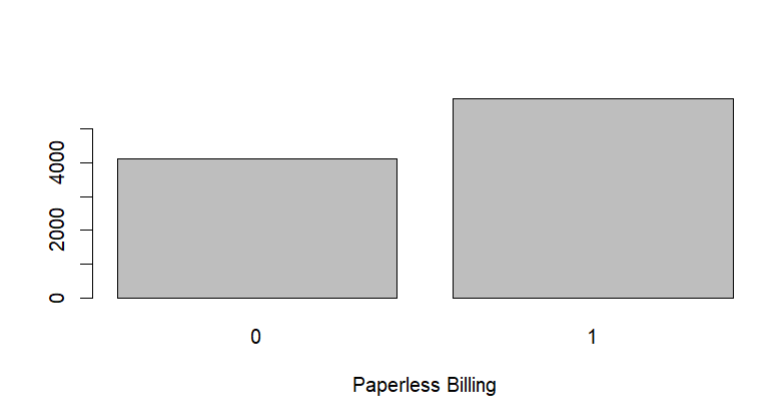
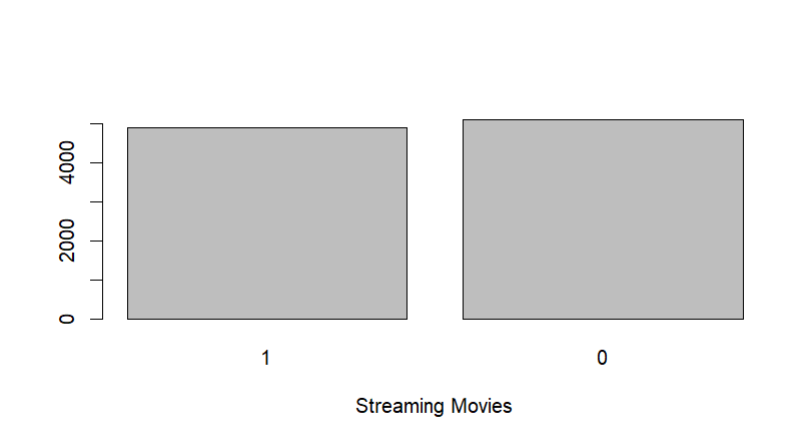
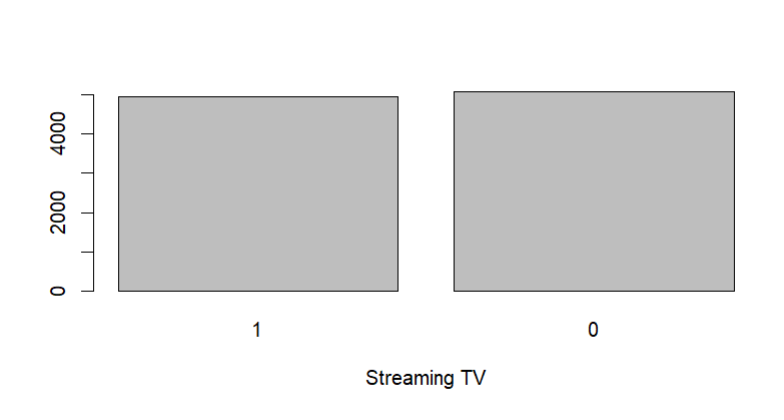
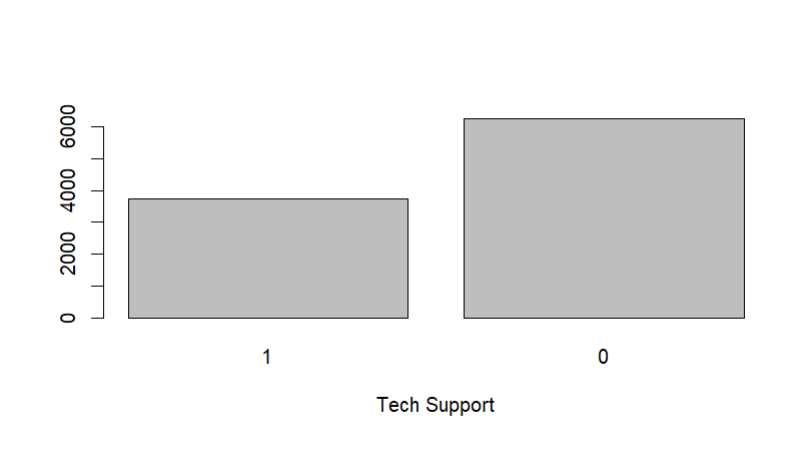
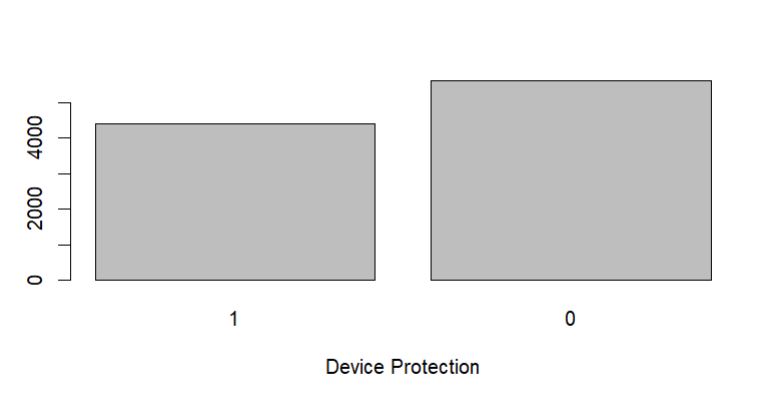
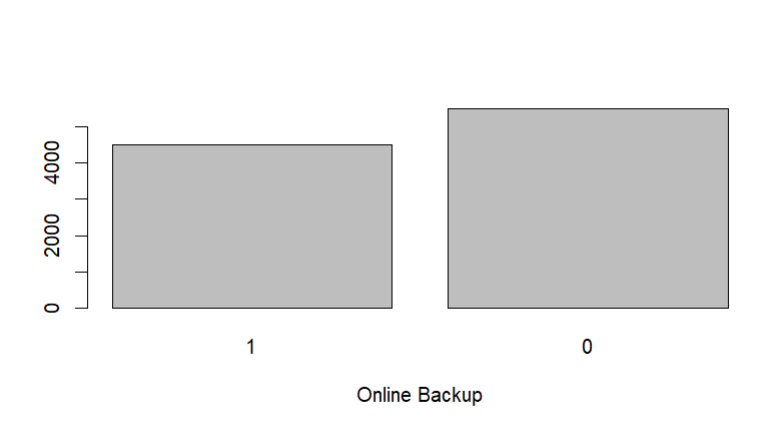
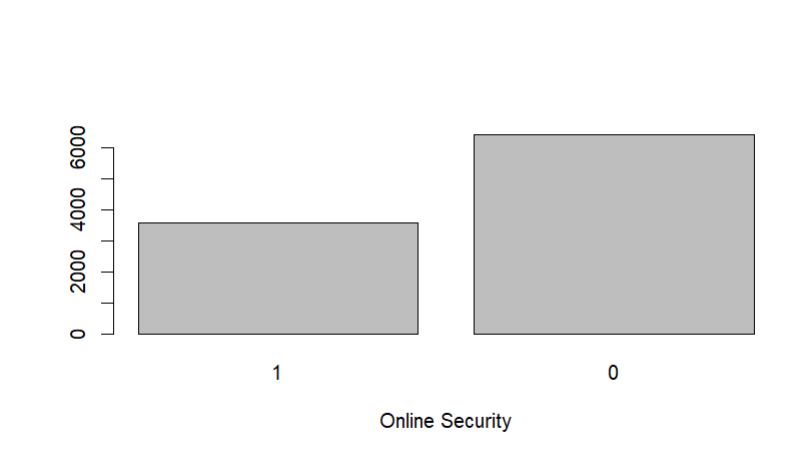
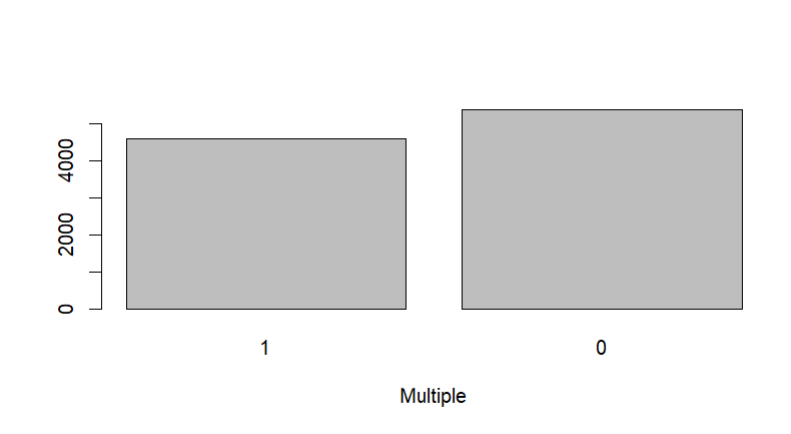
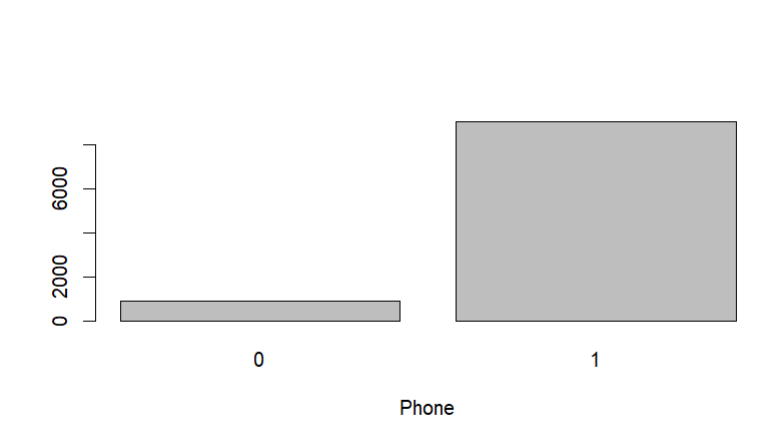
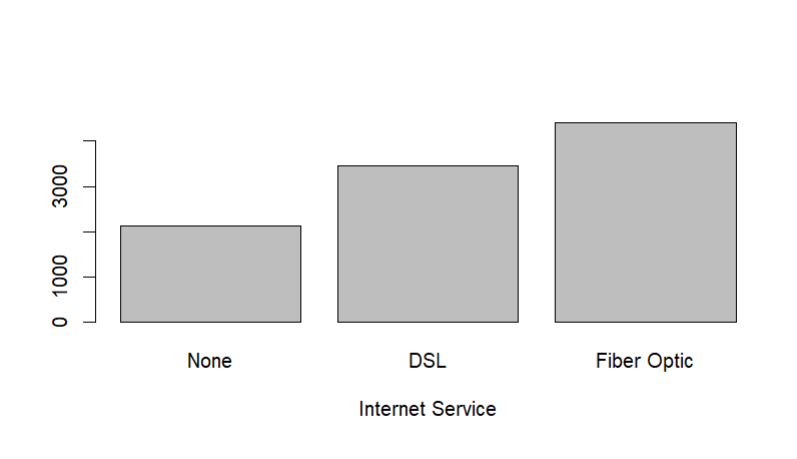
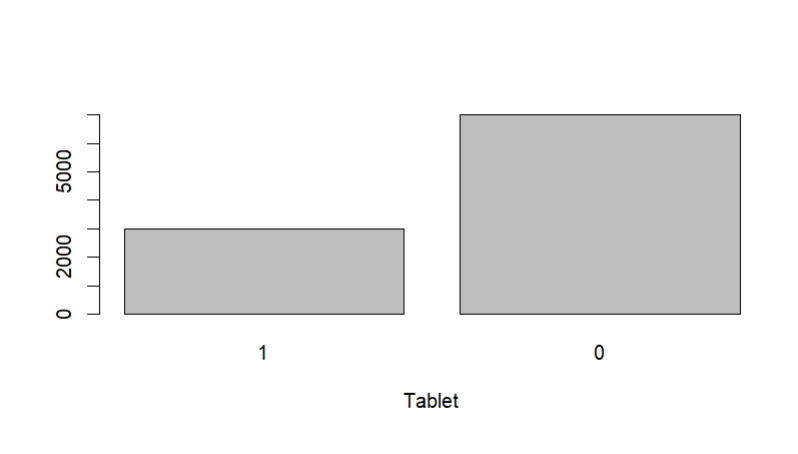
Description automatically generatedA row of gray squares with black text

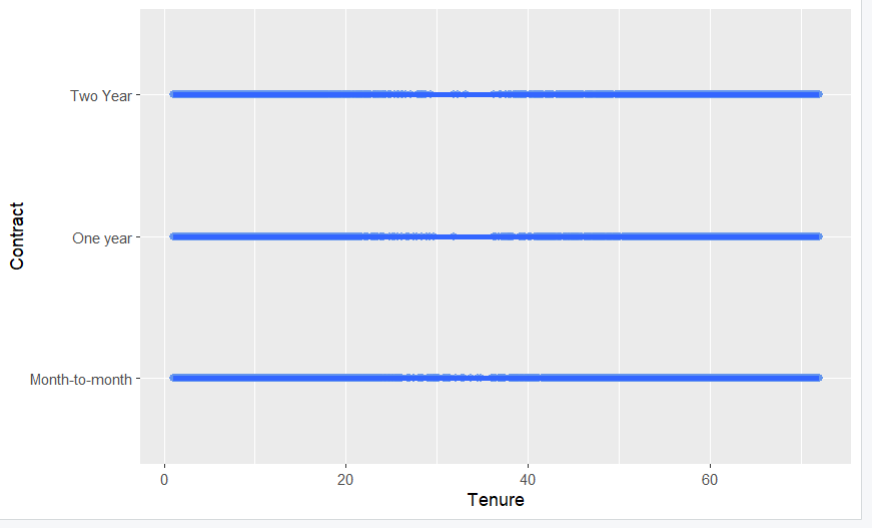
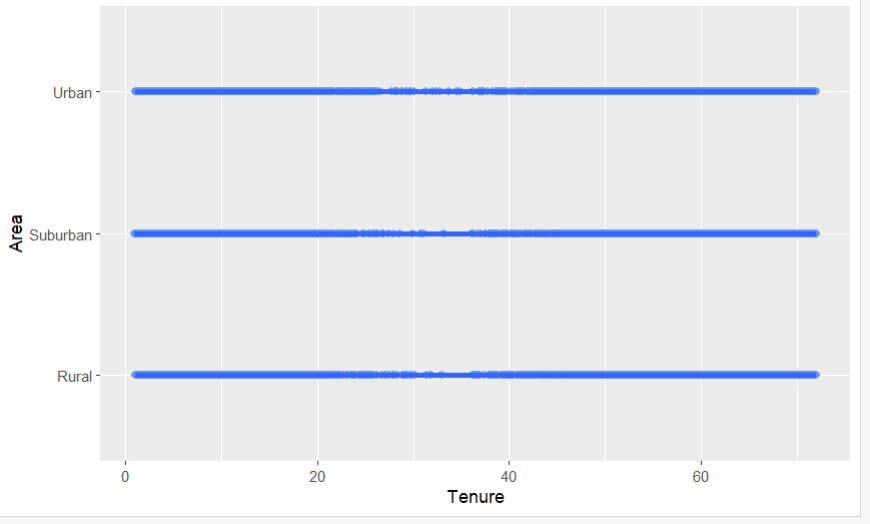
Description automatically generatedA graph of a number of squares

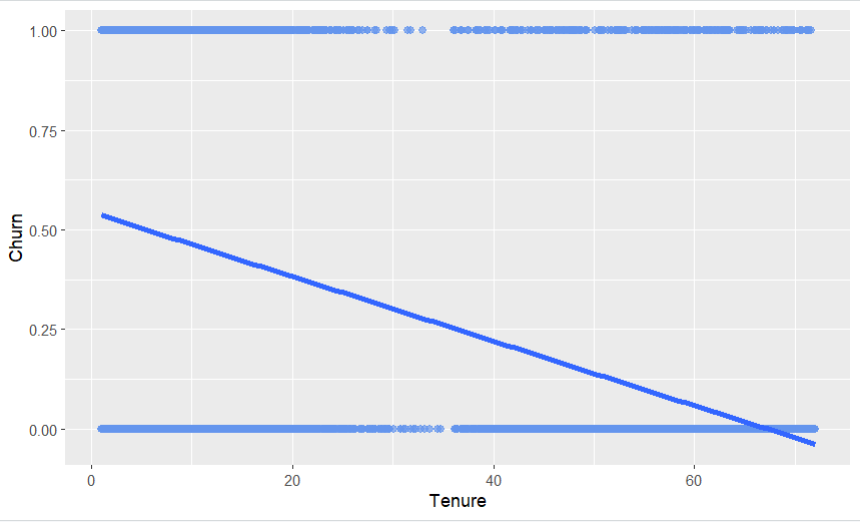
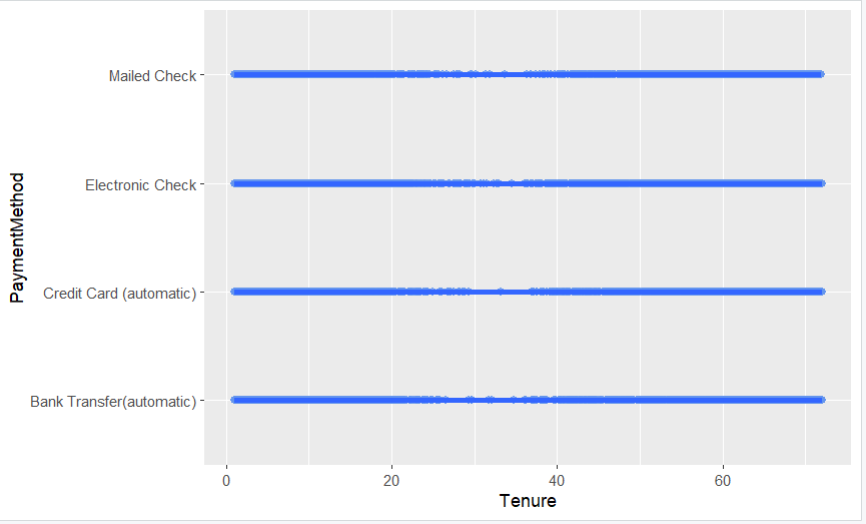
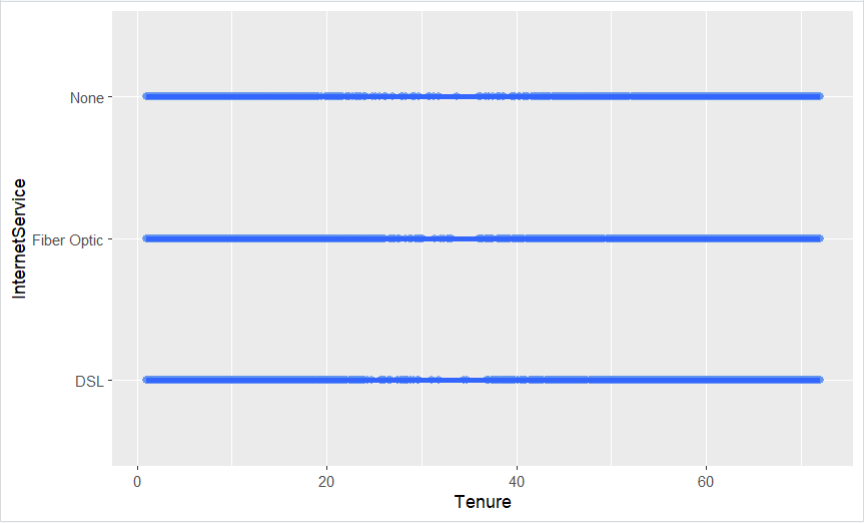
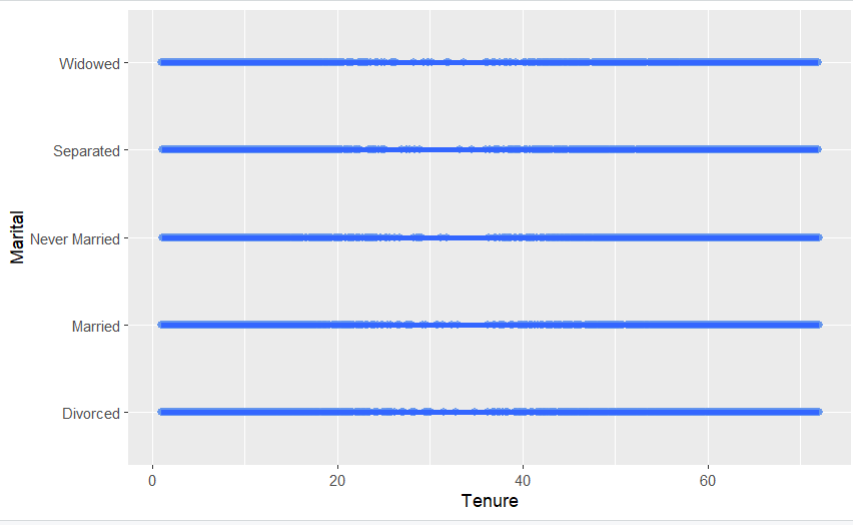
Description automatically generated with medium confidenceA black line with numbers

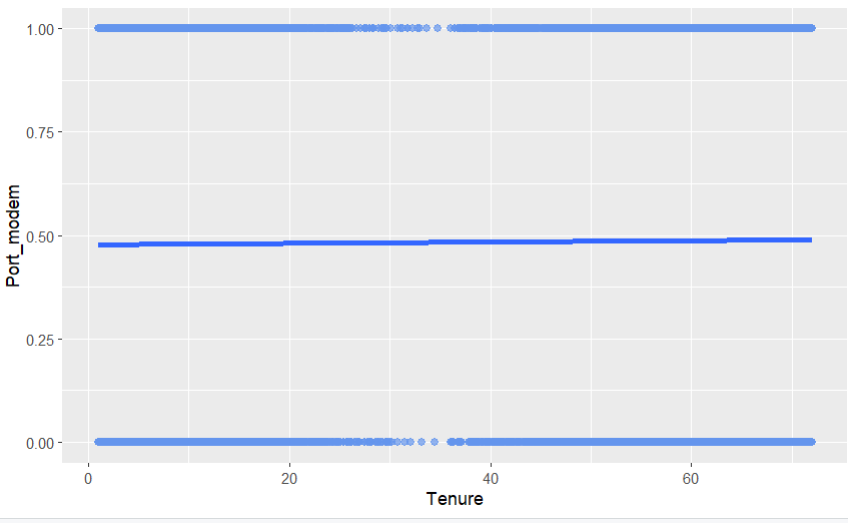
Description automatically generatedA graph of a graph showing a number of objects

Description automatically generated with medium confidenceA couple of squares with text

Description automatically generated with medium confidence

A graph with blue lines

Description automatically generatedA graph with blue lines

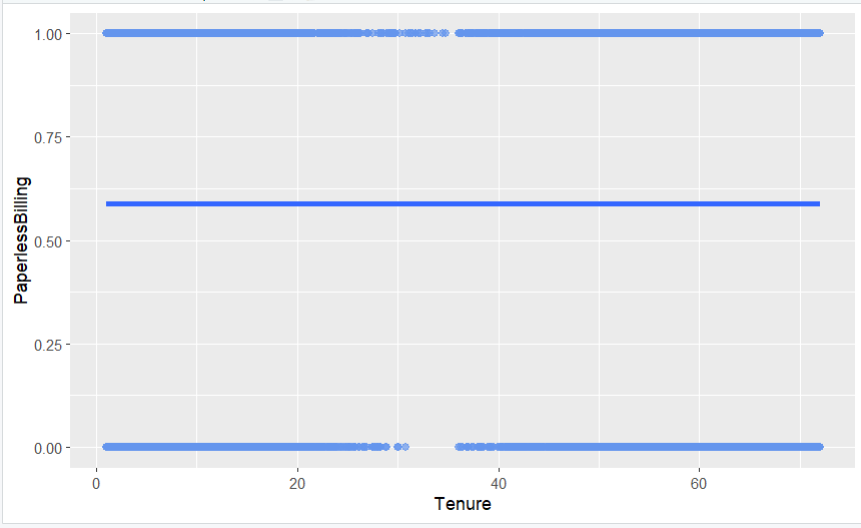
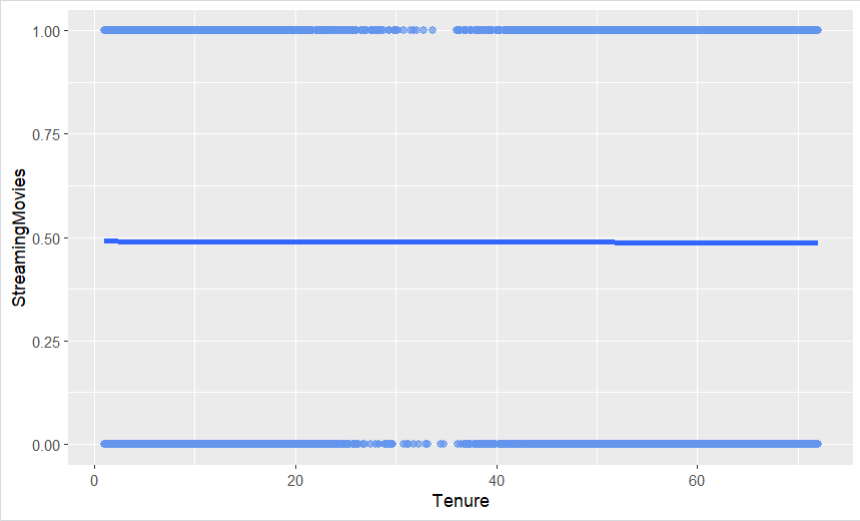
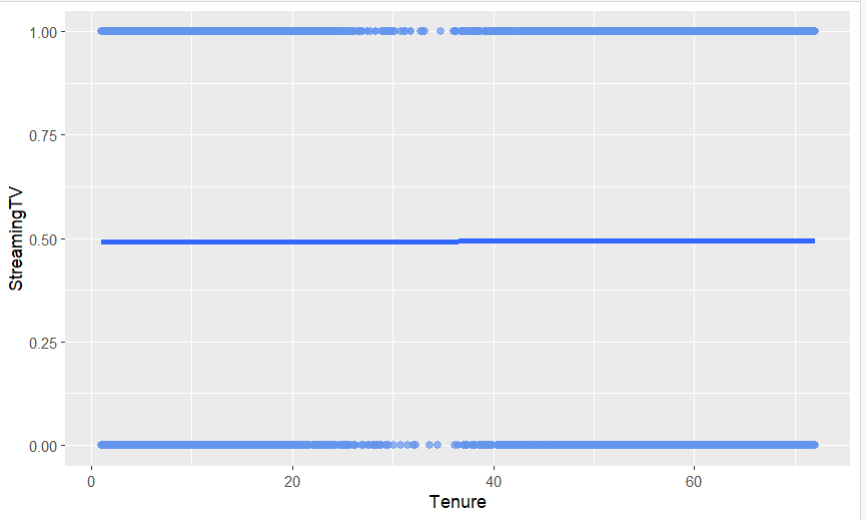
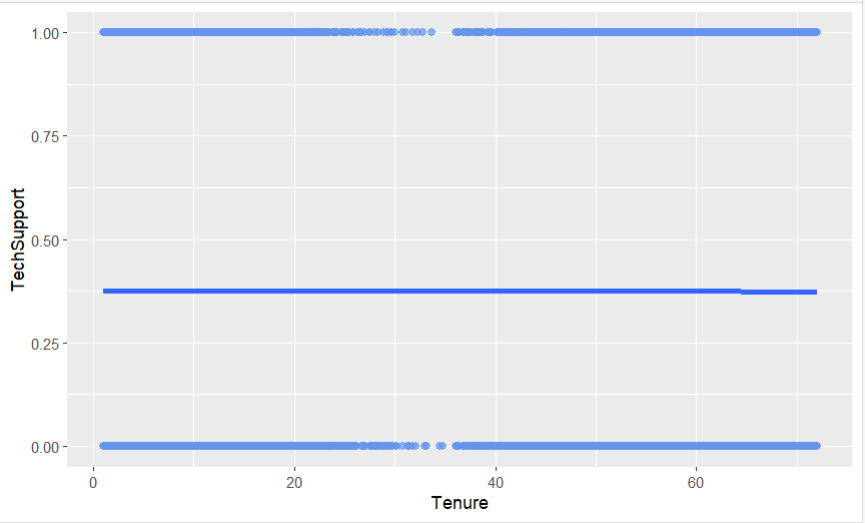
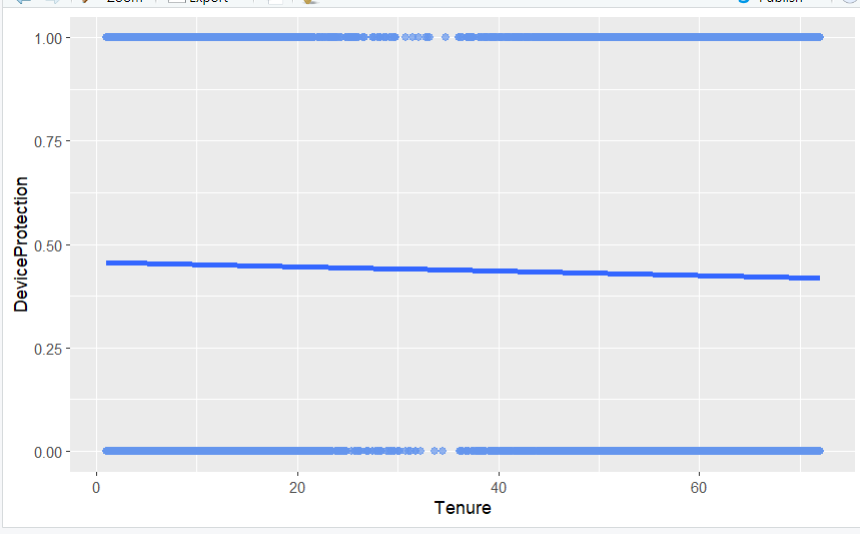
Description automatically generatedA graph with blue lines

Description automatically generatedA graph with blue lines

Description automatically generatedA graph with blue lines

Description automatically generatedA graph with blue lines

Description automatically generatedA graph with blue lines

Description automatically generated

#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 [in-text citation: (R programming 101, n.d.)]

barplot(sort(table(churn\_clean208$Area)))

barplot(sort(table(churn\_clean208$TimeZone)))

barplot(sort(table(churn\_clean208$Children)))

barplot(sort(table(churn\_clean208$Age)))

barplot(sort(table(churn\_clean208$Income)))

barplot(sort(table(churn\_clean208$Marital)))

barplot(sort(table(churn\_clean208$Gender)))

barplot(sort(table(churn\_clean208$Churn)))

barplot(sort(table(churn\_clean208$Outage\_sec\_perweek)))

barplot(sort(table(churn\_clean208$Email)))

barplot(sort(table(churn\_clean208$Contacts)))

barplot(sort(table(churn\_clean208$Yearly\_equip\_failure)))

barplot(sort(table(churn\_clean208$Techie)))

barplot(sort(table(churn\_clean208$Contract)))

barplot(sort(table(churn\_clean208$Port\_modem)))

barplot(sort(table(churn\_clean208$Tablet)))

barplot(sort(table(churn\_clean208$InternetService)))

barplot(sort(table(churn\_clean208$Phone)))

barplot(sort(table(churn\_clean208$Multiple)))

barplot(sort(table(churn\_clean208$OnlineSecurity)))

barplot(sort(table(churn\_clean208$OnlineBackup)))

barplot(sort(table(churn\_clean208$DeviceProtection)))

barplot(sort(table(churn\_clean208$TechSupport)))

barplot(sort(table(churn\_clean208$StreamingTV)))

barplot(sort(table(churn\_clean208$StreamingMovies)))

barplot(sort(table(churn\_clean208$PaperlessBilling)))

barplot(sort(table(churn\_clean208$PaymentMethod)))

barplot(sort(table(churn\_clean208$Tenure)))

barplot(sort(table(churn\_clean208$MonthlyCharge)))

barplot(sort(table(churn\_clean208$Bandwidth\_GB\_Year)))

barplot(sort(table(churn\_clean208$Timely\_Response)))

barplot(sort(table(churn\_clean208$Timely\_Fixes)))

barplot(sort(table(churn\_clean208$Timely\_Replacements)))

barplot(sort(table(churn\_clean208$Reliability)))

barplot(sort(table(churn\_clean208$Options)))

barplot(sort(table(churn\_clean208$Respectful\_Response)))

barplot(sort(table(churn\_clean208$Courteous\_Exchange)))

barplot(sort(table(churn\_clean208$Active\_Listening)))

#BiVariate Graph of each variable

library(ggplot2)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

ggplot(data = A\_treat,

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

geom\_point(color = "cornflowerblue",

alpha = .7,

size = 2) +

geom\_smooth(method = "lm",

se = FALSE,

linewidth = 1.5)

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 linear 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

install.packages("fastDummies")

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

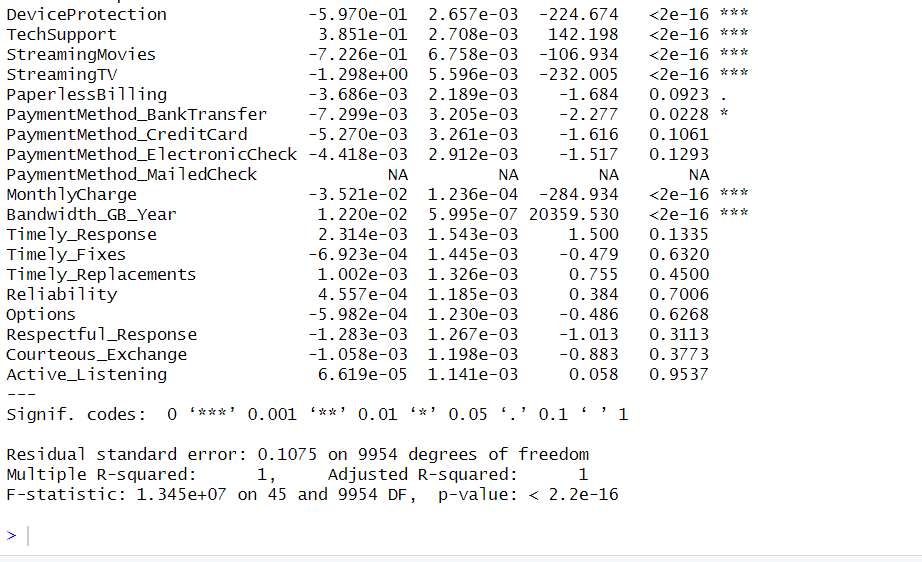
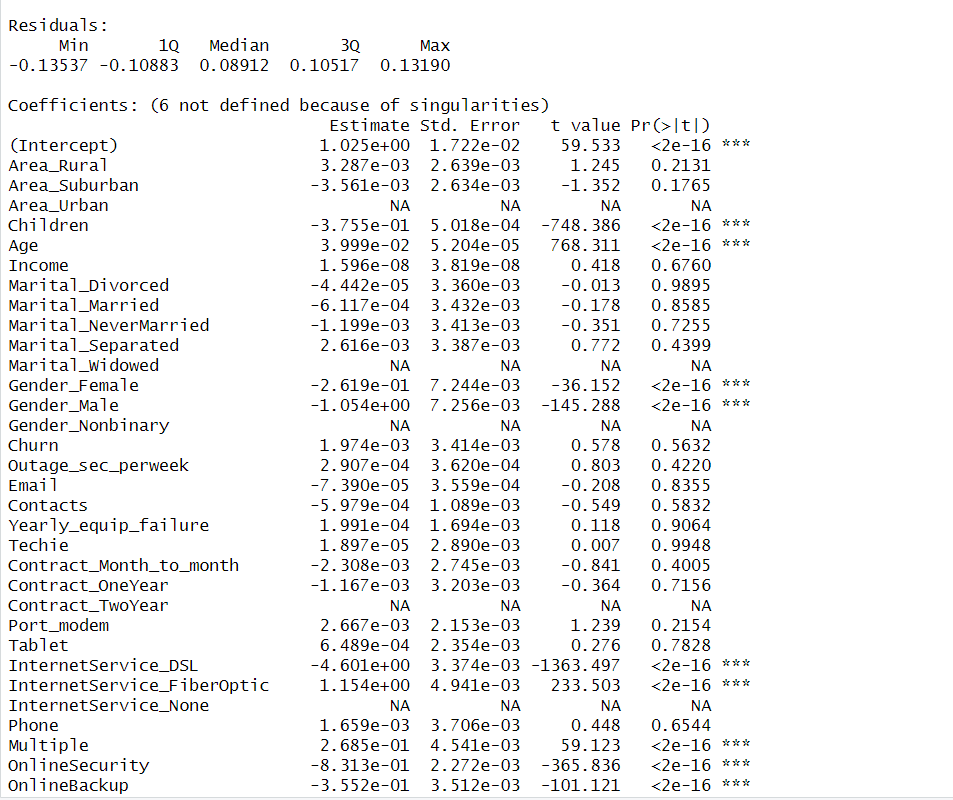
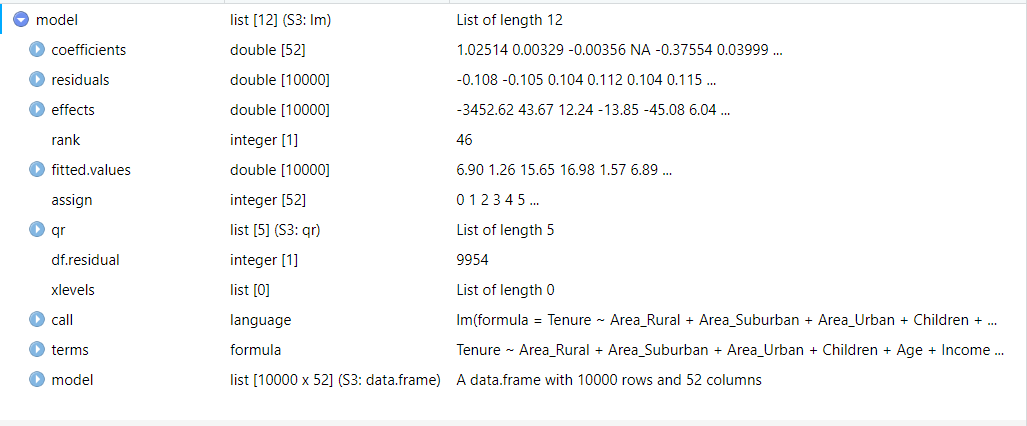
P\_treat <- dummy\_cols(churn\_clean208, select\_columns = "PaymentMethod")

5.  See attached code.

**Part IV: Model Comparison and Analysis**

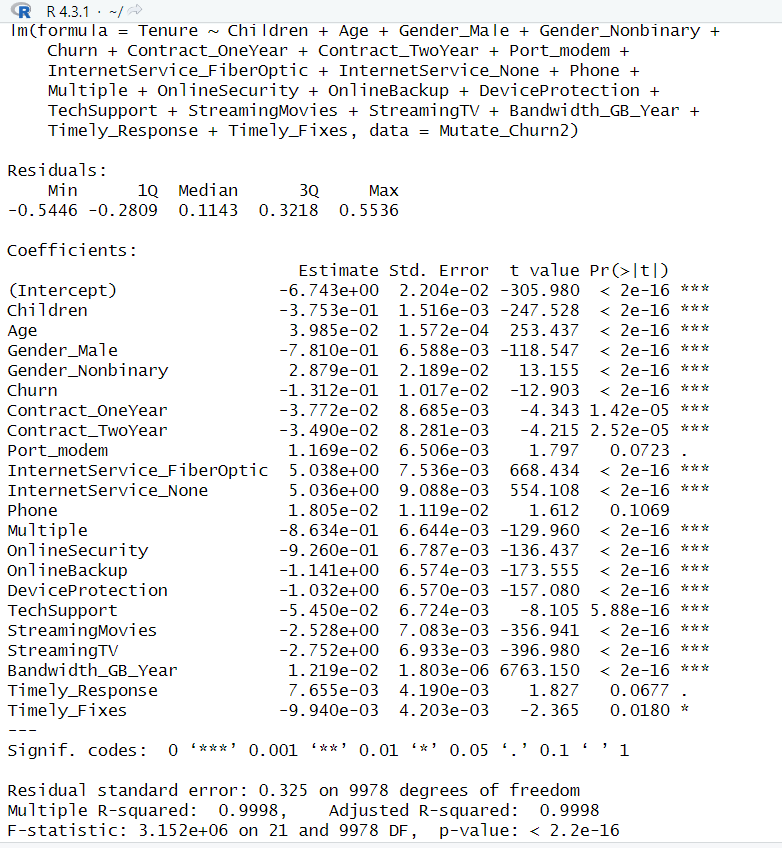
D.  Comparison of the initial and a reduced linear regression model:

1. The initial multiple linear regression model residuals are acceptable as the median is close to zero at .08 with a min of -.135 and a max of .136. This tells us that the model is close to evenly distributed and symmetrical. 1Q and 3Q are also similar values, one being positive and one being negative. The initial models std. error is .1075, the lower the number the better. The Pr(>|t|) is <2.2e-16 which is significantly less than .05 meaning that there is significance. The Rsquared and adjusted Rsquared were equal to 1 in the initial model.



2.  Feature selection reduction was done utilizing a few 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. Second, we reduced MonthlyCharge because the VIF score was greater than 10 which means there’s multicollinearity. We must remove features with multicollinearity to perform multiple linear regression analysis. Lastly, we utilized the wrapper method which is a feature selection reduction method. We utilized the backward stepwise reduction method before finalizing our linear regression model.

3. The reduced linear regression model residuals show a similar min and max with the min at -.5446 and the max at .5536. Q1 is -.2809 and Q3 at .3218 which is slightly not as close in number as the initial model was. The median is .1143 which is slightly higher than the initial model. The coefficiants in the reduced model has an intercept of -6.743e+00, std. error of 2.204e-02 which has improved slightly from the initial model as the smaller the number the better because there is less room for error. The Pr(>|t|) came in at the same value of <2e-16 which shows it is very significant as it is much less than .05. The residual standard of error is .325, both Rsquared and adjusted Rsquared are .9998 which is just below the initial model, and the F stat is 3.152e+06.



E.

1.  The data analysis process includes comparing the initial multiple linear regression model and reduced linear regression model, utilizing the residual standard error 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 multiple linear regression model which are explained in section B1. Ordinary least squares method was performed, the model was reduced based on these findings, the correlation was determined utilizing the Pearsons method and the model was reduced again based on those findings, ols\_vif\_tol was checked and the model was reduced again. Next, the backward stepwise wrapper feature selection reduction method was performed four times to complete the model feature reduction. The residual standard error on the initial model was .1075 while the reduced model has a residual standard error of .325. This meaning the initial model predicts the length of tenure with an average error of .1075, while the reduced model predicts tenure with an average error of .325. On the initial model, the Rsquared and adjusted Rsquared were 1 while on the reduced model they were .9998, this is the percentage variation in dependent that is explained by the independent variable. This measn there is a 99% variation in Y that is explained by the predicting variables. The F stat on the initial model was 1.345e+07 and the reduced model was 3.152e+06, neither of these numbers signify significance in the overall regression as it is much greater than .05. In the initial model there were 71 variables while the reduced model had 39.

2.  The output for all calculations of the analysis that were performed are included below:

 #Create initial model- linear regression modeling [in-text citation: (Zach, 2020)]

model <- lm(formula = Tenure ~ 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 + Churn + 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 = Mutate\_Churn1)

summary(model)

#Reduce model

Mutate\_Churn2 <- subset(Mutate\_Churn1, select = -c(Job, TimeZone, Population, Lat, Lng, Zip, County, State, City, UID, Interaction, Customer\_id))

#Change name of column

colnames(Mutate\_Churn2)[colnames(Mutate\_Churn2) == 'Contract\_Month-to-month'] <- 'Contract\_Month\_To\_Month'

model2 <- lm(formula = Tenure ~ 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 + Churn + 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 +

Bandwidth\_GB\_Year + Timely\_Response +

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

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = Mutate\_Churn2)

View(model2)

summary(model2)

#OLS

library(olsrr)

OLSMutateChurn2 <- ols\_regress(Tenure ~ 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 + Churn + 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 +

Bandwidth\_GB\_Year + Timely\_Response +

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

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = Mutate\_Churn2)

summary(OLSMutateChurn2)

#View summary of variables

summary(Mutate\_Churn2)

#Check variable type

str(Mutate\_Churn2)

#Check correlation

library(corrr)

Churncor <- cor(Mutate\_Churn2)

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

Mutate\_Churn2 <- subset(Mutate\_Churn2, select = -c(PaymentMethod\_BankTransfer, PaymentMethod\_CreditCard, PaymentMethod\_ElectronicCheck, PaymentMethod\_MailedCheck, Marital\_Divorced, Marital\_Married, Marital\_NeverMarried, Marital\_Widowed))

Mutate\_Churn2 <- subset(Mutate\_Churn2, select = -c(Marital\_Separated ))

#K-1 method

Mutate\_Churn2 <- subset(Mutate\_Churn2, select = -c(Area\_Rural, Contract\_Month\_To\_Month, Gender\_Female, InternetService\_DSL))

#Check correlation using Pearson method

library(corrr)

Churncor <- cor(Mutate\_Churn2)

#Export

write.csv(Churncor, "C:/Users/ntrei/OneDrive/Documents/MSDA/Churncor208.csv")

#Correlation -1 perfect negative correlation, 0 is no correlation, 1 is perfect positive correlation

#Bandwidth/Tenure .99 correlation, Timely replacements/fixes and responses ~.5 correlation

library(olsrr)

ols\_vif\_tol(model)

ols\_vif\_tol(model2)

#Drop MonthlyCharge high VIF of 24.383121

Mutate\_Churn2 <- subset(Mutate\_Churn2, select = -c(MonthlyCharge))

#lm() model

modelreduced <- lm(formula = Tenure ~ Area\_Suburban + Area\_Urban + Children + Age +

Income + Gender\_Male + Gender\_Nonbinary + Churn + 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 +

Bandwidth\_GB\_Year + Timely\_Response + Timely\_Fixes + Timely\_Replacements + Reliability + Options +

Respectful\_Response + Courteous\_Exchange + Active\_Listening, data = Mutate\_Churn2)

summary(modelreduced)

library(olsrr)

ols\_vif\_tol(modelreduced)

#Rsquared and adjust Rsquared identical, port modem, phone, timely response great than .05 Pvalue

Mutate\_Churn2 %>%

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

correlate() %>%

shave() %>%

rplot(print\_cor = TRUE) +

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

#Call proper packages

library(tidyverse)

library(caret)

library(leaps)

library(MASS)

# Stepwise regression model [in text citation: (Pickard, M. (N.D.)) & (Zach. (2019)]

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

trace = FALSE)

summary(step.model)

models <- regsubsets(Tenure~., data = Mutate\_Churn2, nvmax = 50,

method = "backward")

Summarymodels <- summary(models)

#Stepwise regression 2

step.model2 <- stepAIC(modelreduced, direction = "backward",

trace = FALSE)

summary(step.model2)

models2 <- regsubsets(Tenure~., data = Mutate\_Churn2, nvmax = 50,

method = "backward")

Summarymodels2 <- summary(models)

models$coefficients

models$anova

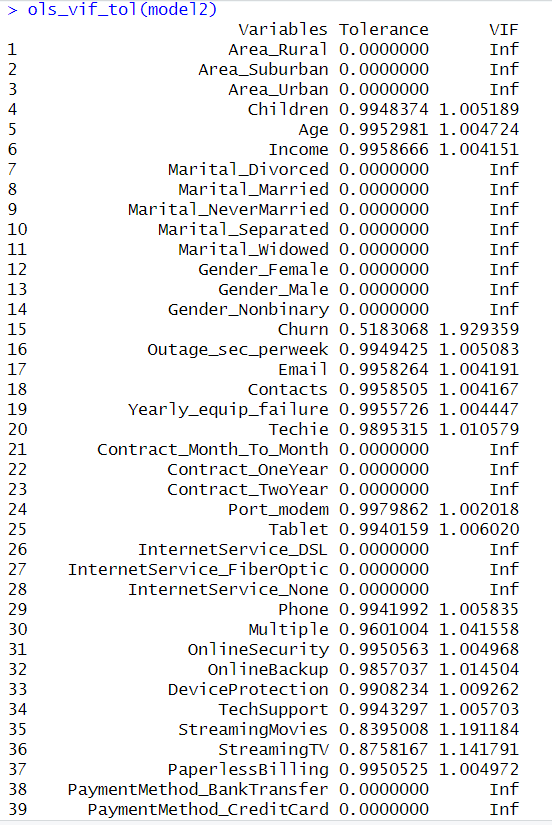
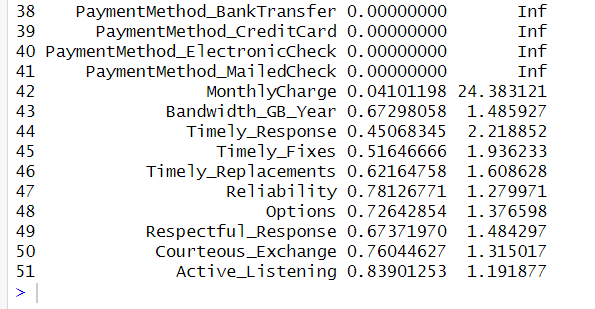
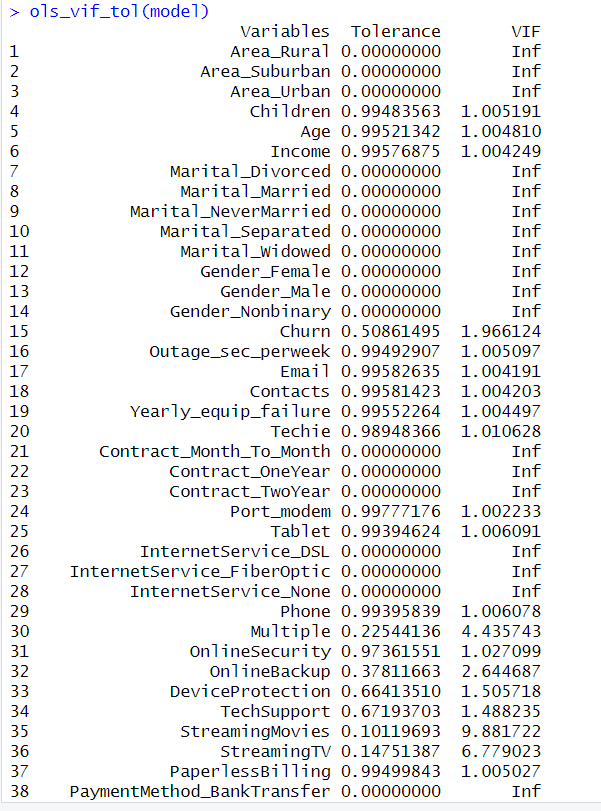
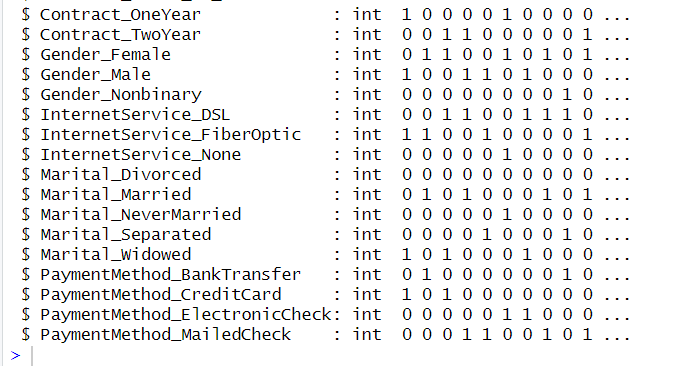
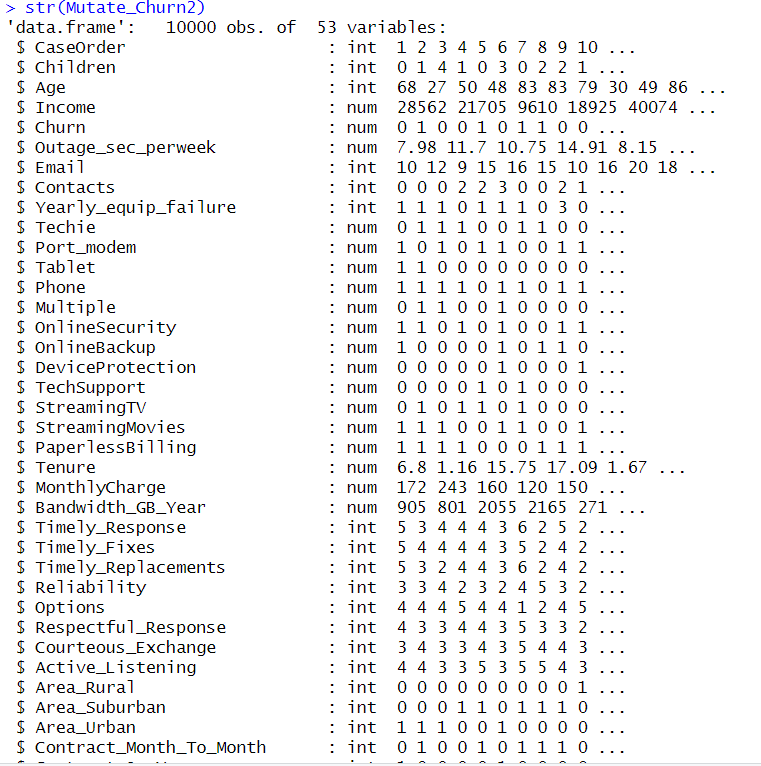
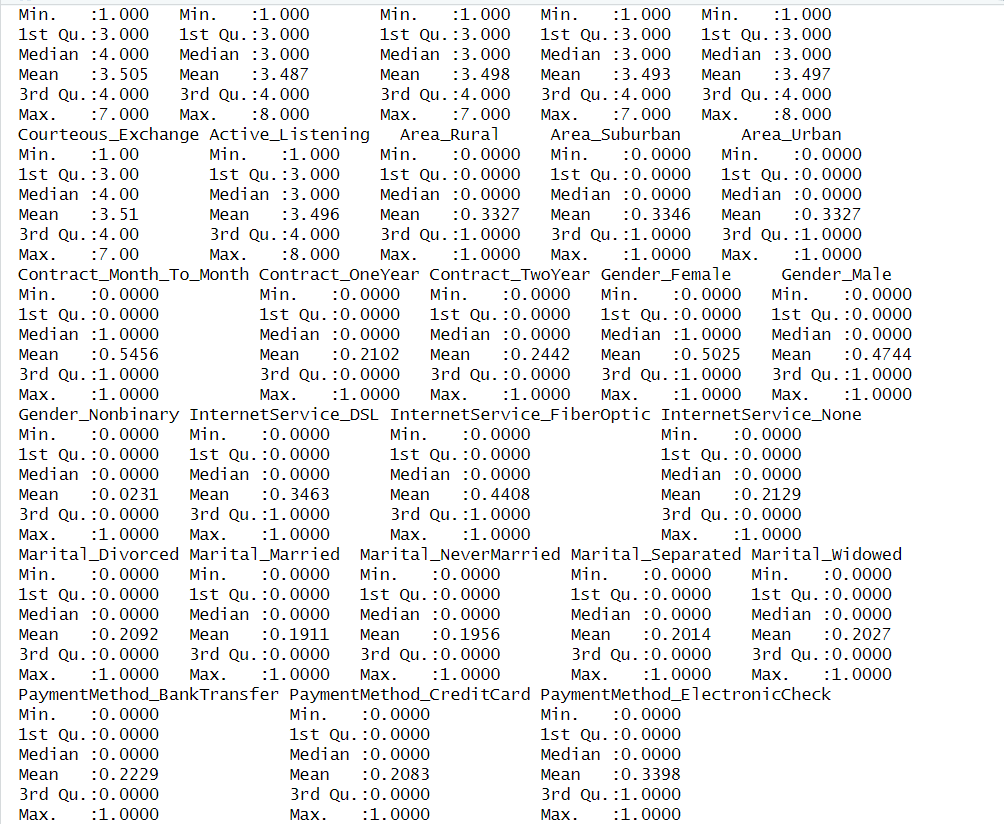
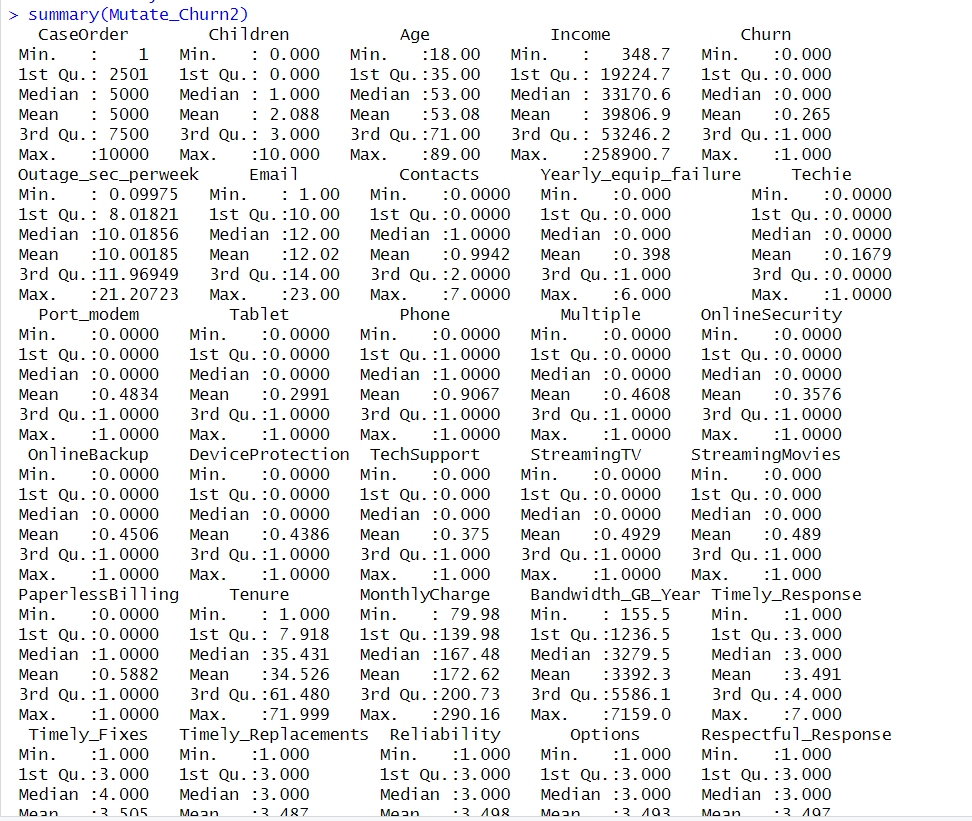
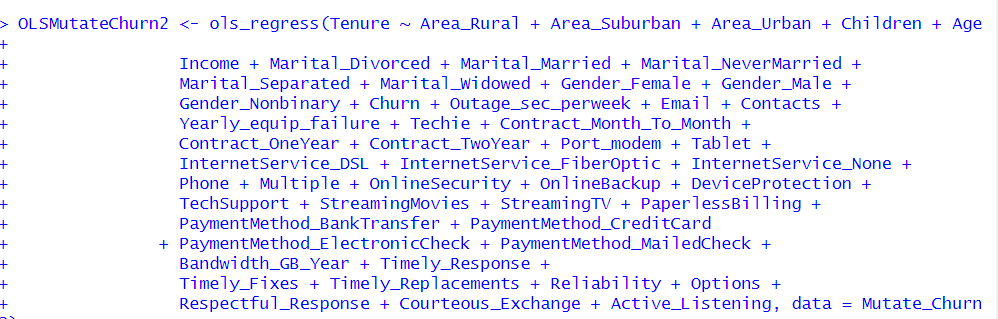
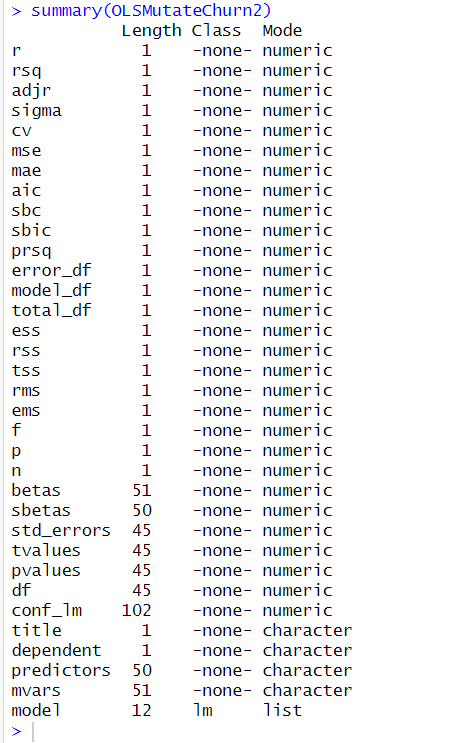
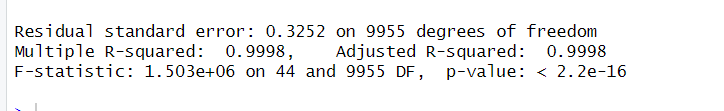
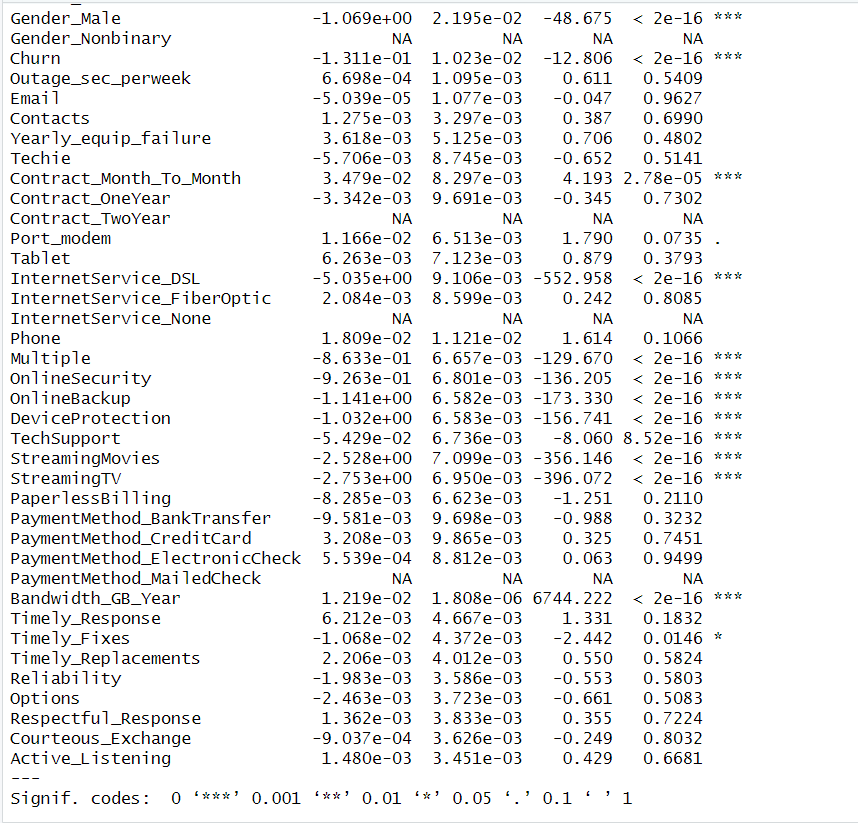
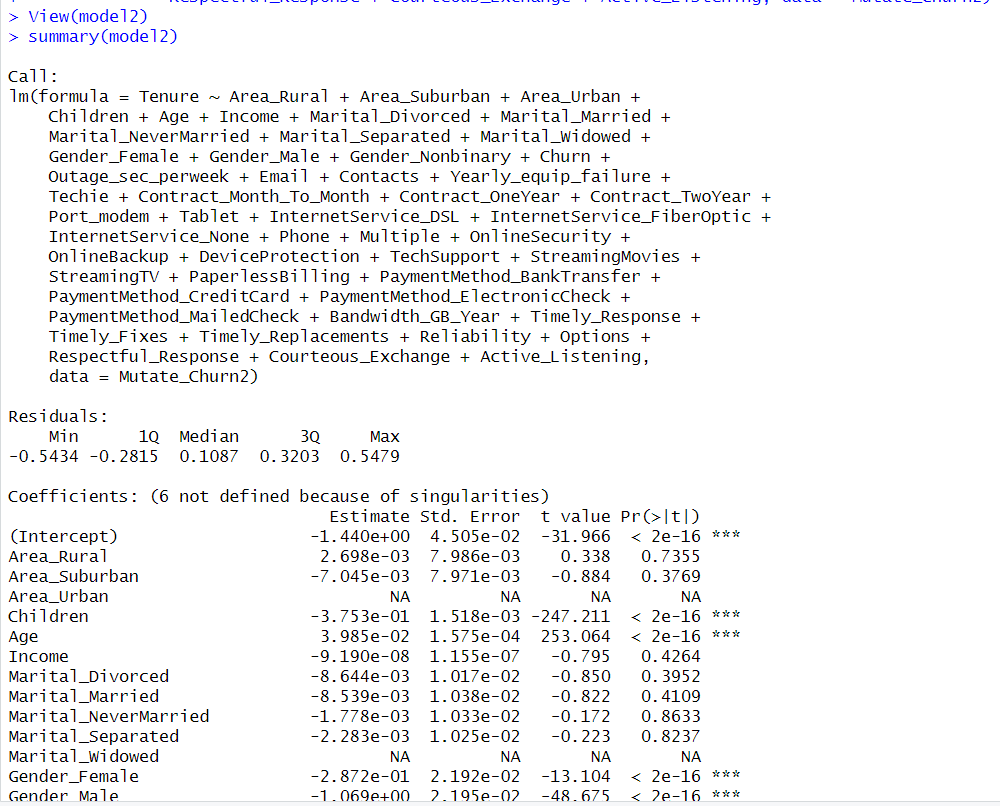
#Visualizations

plot(models)

plot(models2)

#Export

write.csv(Mutate\_Churn2, "C:/Users/ntrei/OneDrive/Documents/MSDA/MutateChurn2.208.csv")



Call:

lm(formula = Tenure ~ Area\_Suburban + Area\_Urban + Children +

Age + Income + Gender\_Male + Gender\_Nonbinary + Churn + 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 + Bandwidth\_GB\_Year + Timely\_Response +

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

Respectful\_Response + Courteous\_Exchange + Active\_Listening,

data = Mutate\_Churn2)

Residuals:

Min 1Q Median 3Q Max

-0.5493 -0.2820 0.1098 0.3198 0.5444

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.731e+00 3.882e-02 -173.386 < 2e-16 \*\*\*

Area\_Suburban -9.745e-03 7.966e-03 -1.223 0.2212

Area\_Urban -2.732e-03 7.981e-03 -0.342 0.7321

Children -3.753e-01 1.518e-03 -247.265 < 2e-16 \*\*\*

Age 3.984e-02 1.574e-04 253.188 < 2e-16 \*\*\*

Income -9.098e-08 1.155e-07 -0.788 0.4309

Gender\_Male -7.816e-01 6.601e-03 -118.409 < 2e-16 \*\*\*

Gender\_Nonbinary 2.871e-01 2.191e-02 13.101 < 2e-16 \*\*\*

Churn -1.304e-01 1.022e-02 -12.762 < 2e-16 \*\*\*

Outage\_sec\_perweek 6.567e-04 1.095e-03 0.600 0.5486

Email -3.142e-05 1.076e-03 -0.029 0.9767

Contacts 1.192e-03 3.294e-03 0.362 0.7174

Yearly\_equip\_failure 3.728e-03 5.121e-03 0.728 0.4666

Techie -5.651e-03 8.739e-03 -0.647 0.5179

Contract\_OneYear -3.794e-02 8.697e-03 -4.362 1.30e-05 \*\*\*

Contract\_TwoYear -3.450e-02 8.293e-03 -4.161 3.20e-05 \*\*\*

Port\_modem 1.158e-02 6.511e-03 1.779 0.0753 .

Tablet 6.402e-03 7.118e-03 0.899 0.3684

InternetService\_FiberOptic 5.038e+00 7.546e-03 667.562 < 2e-16 \*\*\*

InternetService\_None 5.036e+00 9.099e-03 553.431 < 2e-16 \*\*\*

Phone 1.791e-02 1.121e-02 1.598 0.1101

Multiple -8.635e-01 6.653e-03 -129.788 < 2e-16 \*\*\*

OnlineSecurity -9.261e-01 6.797e-03 -136.249 < 2e-16 \*\*\*

OnlineBackup -1.141e+00 6.580e-03 -173.375 < 2e-16 \*\*\*

DeviceProtection -1.032e+00 6.580e-03 -156.828 < 2e-16 \*\*\*

TechSupport -5.439e-02 6.733e-03 -8.078 7.34e-16 \*\*\*

StreamingMovies -2.529e+00 7.095e-03 -356.389 < 2e-16 \*\*\*

StreamingTV -2.753e+00 6.946e-03 -396.289 < 2e-16 \*\*\*

PaperlessBilling -8.200e-03 6.619e-03 -1.239 0.2155

Bandwidth\_GB\_Year 1.219e-02 1.806e-06 6750.416 < 2e-16 \*\*\*

Timely\_Response 6.100e-03 4.665e-03 1.308 0.1910

Timely\_Fixes -1.071e-02 4.370e-03 -2.451 0.0143 \*

Timely\_Replacements 2.368e-03 4.010e-03 0.591 0.5548

Reliability -2.027e-03 3.584e-03 -0.565 0.5718

Options -2.463e-03 3.721e-03 -0.662 0.5080

Respectful\_Response 1.343e-03 3.830e-03 0.351 0.7258

Courteous\_Exchange -8.565e-04 3.624e-03 -0.236 0.8132

Active\_Listening 1.524e-03 3.448e-03 0.442 0.6586

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3251 on 9962 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 1.788e+06 on 37 and 9962 DF, p-value: < 2.2e-16

> ols\_vif\_tol(modelreduced)

Variables Tolerance VIF

1 Area\_Suburban 0.7481267 1.336672

2 Area\_Urban 0.7473242 1.338107

3 Children 0.9949164 1.005110

4 Age 0.9961308 1.003884

5 Income 0.9963566 1.003657

6 Gender\_Male 0.9728755 1.027881

7 Gender\_Nonbinary 0.9755786 1.025033

8 Churn 0.5199298 1.923337

9 Outage\_sec\_perweek 0.9960135 1.004002

10 Email 0.9963453 1.003668

11 Contacts 0.9967070 1.003304

12 Yearly\_equip\_failure 0.9965075 1.003505

13 Techie 0.9905118 1.009579

14 Contract\_OneYear 0.8416004 1.188212

15 Contract\_TwoYear 0.8327005 1.200912

16 Port\_modem 0.9981729 1.001830

17 Tablet 0.9951566 1.004867

18 InternetService\_FiberOptic 0.7528763 1.328240

19 InternetService\_None 0.7617673 1.312737

20 Phone 0.9949805 1.005045

21 Multiple 0.9610368 1.040543

22 OnlineSecurity 0.9956889 1.004330

23 OnlineBackup 0.9858901 1.014312

24 DeviceProtection 0.9913839 1.008691

25 TechSupport 0.9946386 1.005390

26 StreamingMovies 0.8402390 1.190138

27 StreamingTV 0.8763560 1.141089

28 PaperlessBilling 0.9958483 1.004169

29 Bandwidth\_GB\_Year 0.6786434 1.473528

30 Timely\_Response 0.4510482 2.217058

31 Timely\_Fixes 0.5169864 1.934287

32 Timely\_Replacements 0.6221377 1.607361

33 Reliability 0.7819367 1.278876

34 Options 0.7267652 1.375960

35 Respectful\_Response 0.6743127 1.482991

36 Courteous\_Exchange 0.7606624 1.314644

37 Active\_Listening 0.8401095 1.190321

Call:

lm(formula = Tenure ~ Children + Age + Gender\_Male + Gender\_Nonbinary +

Churn + Contract\_OneYear + Contract\_TwoYear + Port\_modem +

InternetService\_FiberOptic + InternetService\_None + Phone +

Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +

TechSupport + StreamingMovies + StreamingTV + Bandwidth\_GB\_Year +

Timely\_Response + Timely\_Fixes, data = Mutate\_Churn2)

Residuals:

Min 1Q Median 3Q Max

-0.5446 -0.2809 0.1143 0.3218 0.5536

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.743e+00 2.204e-02 -305.980 < 2e-16 \*\*\*

Children -3.753e-01 1.516e-03 -247.528 < 2e-16 \*\*\*

Age 3.985e-02 1.572e-04 253.437 < 2e-16 \*\*\*

Gender\_Male -7.810e-01 6.588e-03 -118.547 < 2e-16 \*\*\*

Gender\_Nonbinary 2.879e-01 2.189e-02 13.155 < 2e-16 \*\*\*

Churn -1.312e-01 1.017e-02 -12.903 < 2e-16 \*\*\*

Contract\_OneYear -3.772e-02 8.685e-03 -4.343 1.42e-05 \*\*\*

Contract\_TwoYear -3.490e-02 8.281e-03 -4.215 2.52e-05 \*\*\*

Port\_modem 1.169e-02 6.506e-03 1.797 0.0723 .

InternetService\_FiberOptic 5.038e+00 7.536e-03 668.434 < 2e-16 \*\*\*

InternetService\_None 5.036e+00 9.088e-03 554.108 < 2e-16 \*\*\*

Phone 1.805e-02 1.119e-02 1.612 0.1069

Multiple -8.634e-01 6.644e-03 -129.960 < 2e-16 \*\*\*

OnlineSecurity -9.260e-01 6.787e-03 -136.437 < 2e-16 \*\*\*

OnlineBackup -1.141e+00 6.574e-03 -173.555 < 2e-16 \*\*\*

DeviceProtection -1.032e+00 6.570e-03 -157.080 < 2e-16 \*\*\*

TechSupport -5.450e-02 6.724e-03 -8.105 5.88e-16 \*\*\*

StreamingMovies -2.528e+00 7.083e-03 -356.941 < 2e-16 \*\*\*

StreamingTV -2.752e+00 6.933e-03 -396.980 < 2e-16 \*\*\*

Bandwidth\_GB\_Year 1.219e-02 1.803e-06 6763.150 < 2e-16 \*\*\*

Timely\_Response 7.655e-03 4.190e-03 1.827 0.0677 .

Timely\_Fixes -9.940e-03 4.203e-03 -2.365 0.0180 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.325 on 9978 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 3.152e+06 on 21 and 9978 DF, p-value: < 2.2e-16

Call:

lm(formula = Tenure ~ Children + Age + Gender\_Male + Gender\_Nonbinary +

Churn + Contract\_OneYear + Contract\_TwoYear + Port\_modem +

InternetService\_FiberOptic + InternetService\_None + Phone +

Multiple + OnlineSecurity + OnlineBackup + DeviceProtection +

TechSupport + StreamingMovies + StreamingTV + Bandwidth\_GB\_Year +

Timely\_Response + Timely\_Fixes, data = Mutate\_Churn2)

Residuals:

Min 1Q Median 3Q Max

-0.5446 -0.2809 0.1143 0.3218 0.5536

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.743e+00 2.204e-02 -305.980 < 2e-16 \*\*\*

Children -3.753e-01 1.516e-03 -247.528 < 2e-16 \*\*\*

Age 3.985e-02 1.572e-04 253.437 < 2e-16 \*\*\*

Gender\_Male -7.810e-01 6.588e-03 -118.547 < 2e-16 \*\*\*

Gender\_Nonbinary 2.879e-01 2.189e-02 13.155 < 2e-16 \*\*\*

Churn -1.312e-01 1.017e-02 -12.903 < 2e-16 \*\*\*

Contract\_OneYear -3.772e-02 8.685e-03 -4.343 1.42e-05 \*\*\*

Contract\_TwoYear -3.490e-02 8.281e-03 -4.215 2.52e-05 \*\*\*

Port\_modem 1.169e-02 6.506e-03 1.797 0.0723 .

InternetService\_FiberOptic 5.038e+00 7.536e-03 668.434 < 2e-16 \*\*\*

InternetService\_None 5.036e+00 9.088e-03 554.108 < 2e-16 \*\*\*

Phone 1.805e-02 1.119e-02 1.612 0.1069

Multiple -8.634e-01 6.644e-03 -129.960 < 2e-16 \*\*\*

OnlineSecurity -9.260e-01 6.787e-03 -136.437 < 2e-16 \*\*\*

OnlineBackup -1.141e+00 6.574e-03 -173.555 < 2e-16 \*\*\*

DeviceProtection -1.032e+00 6.570e-03 -157.080 < 2e-16 \*\*\*

TechSupport -5.450e-02 6.724e-03 -8.105 5.88e-16 \*\*\*

StreamingMovies -2.528e+00 7.083e-03 -356.941 < 2e-16 \*\*\*

StreamingTV -2.752e+00 6.933e-03 -396.980 < 2e-16 \*\*\*

Bandwidth\_GB\_Year 1.219e-02 1.803e-06 6763.150 < 2e-16 \*\*\*

Timely\_Response 7.655e-03 4.190e-03 1.827 0.0677 .

Timely\_Fixes -9.940e-03 4.203e-03 -2.365 0.0180 \*

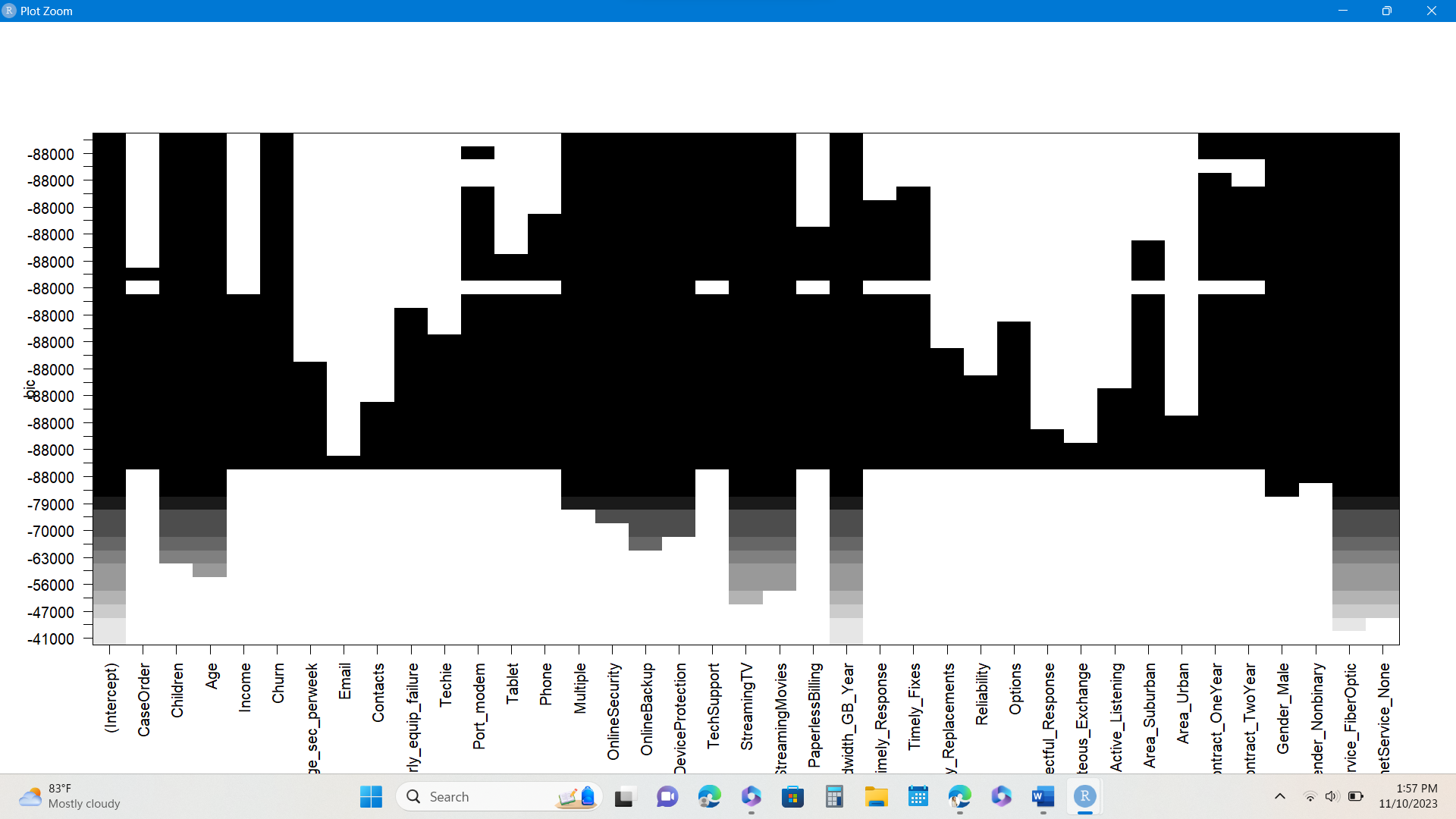
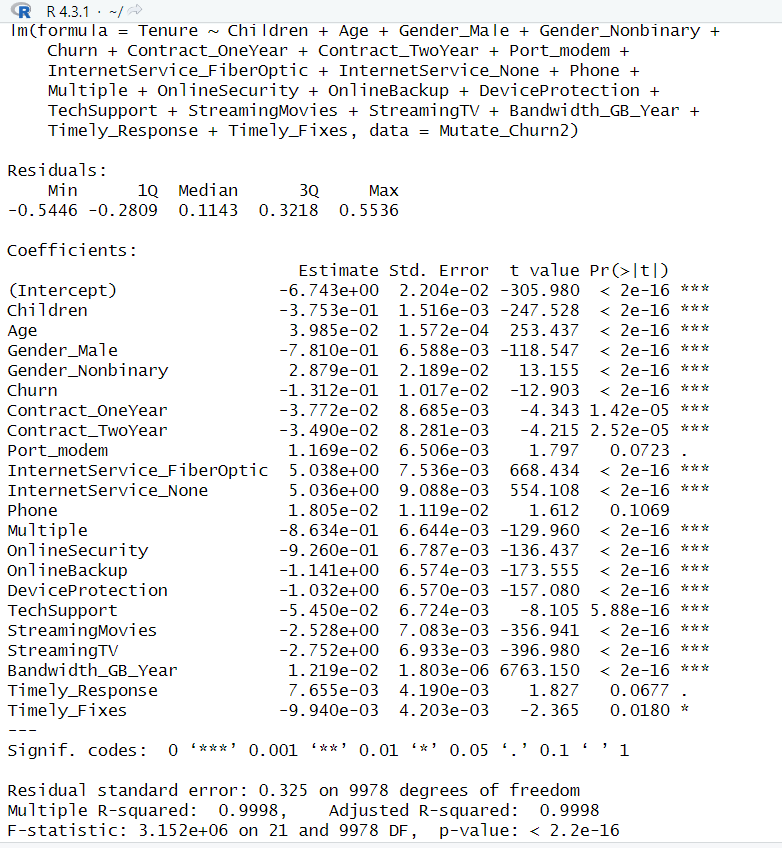
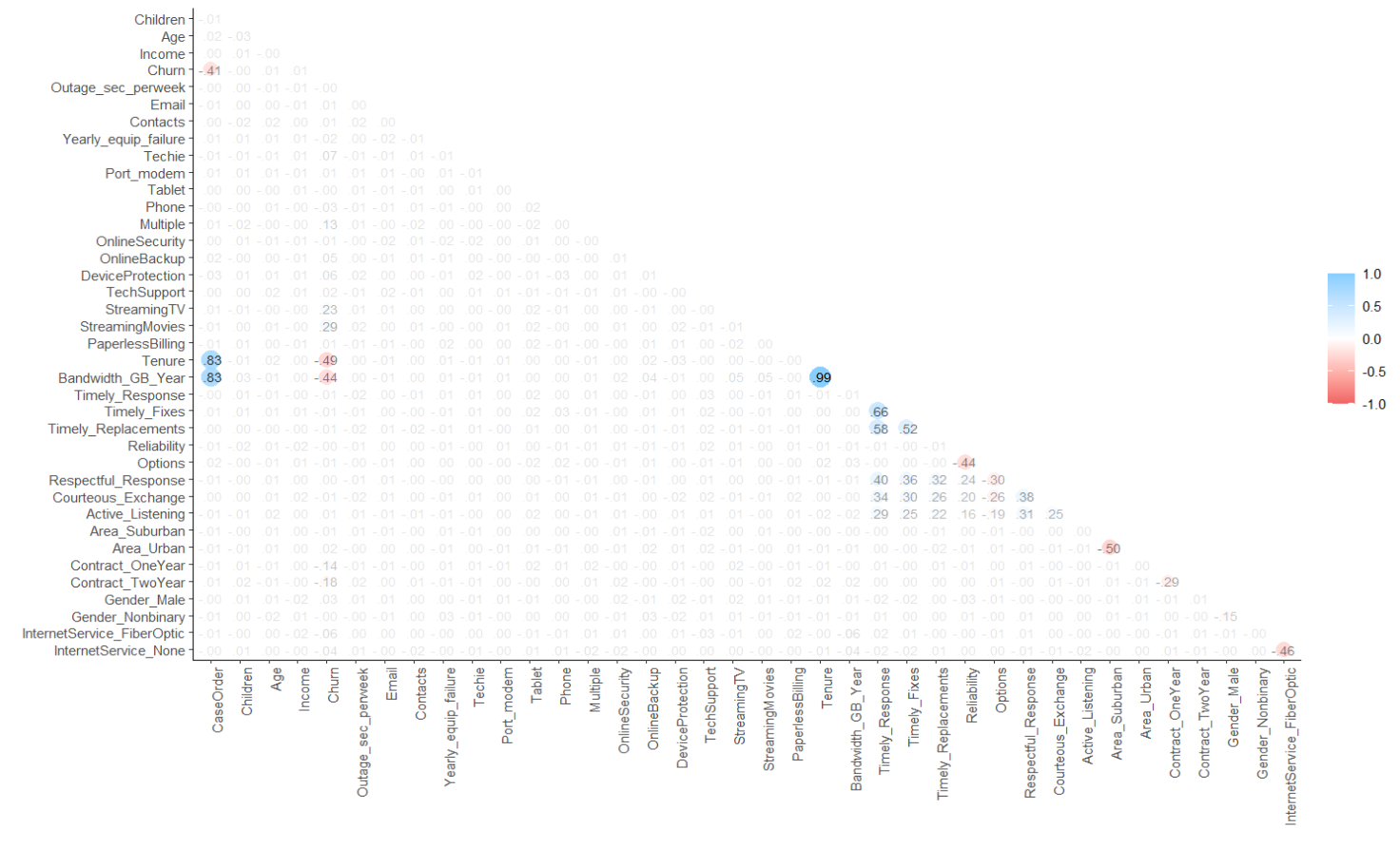
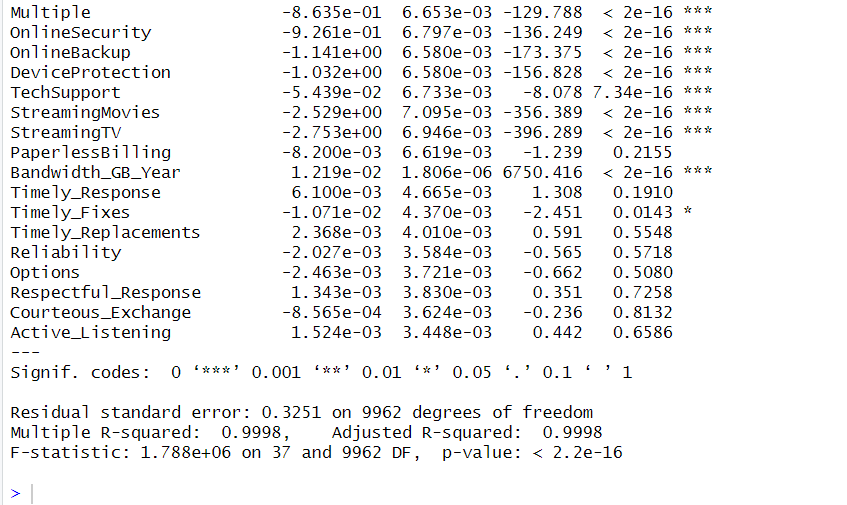
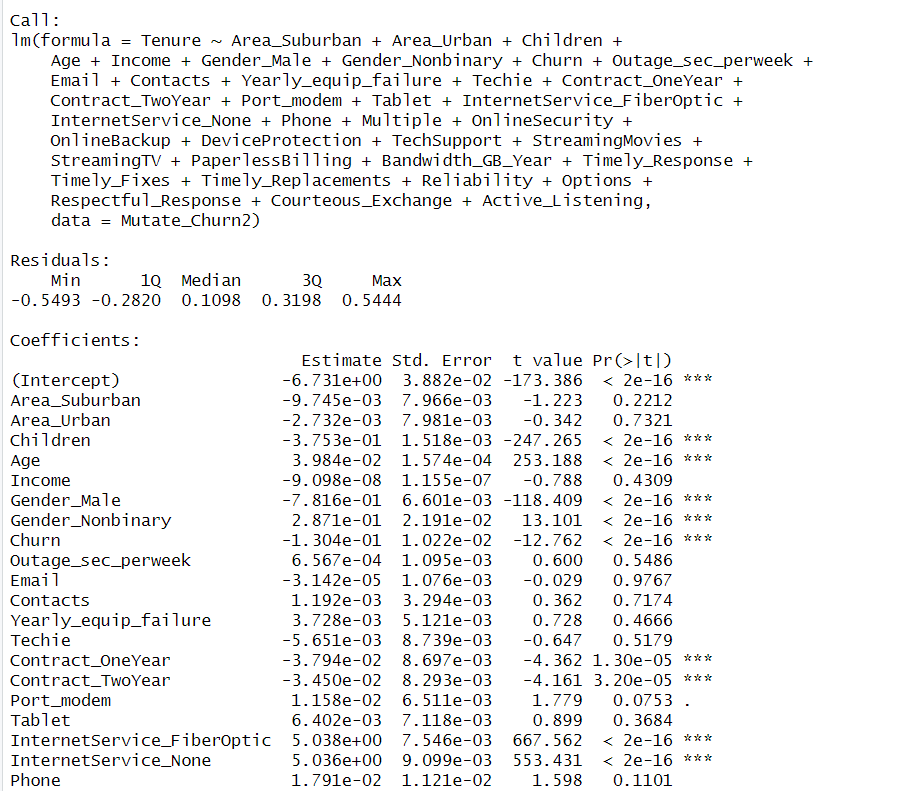
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.325 on 9978 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 3.152e+06 on 21 and 9978 DF, p-value: < 2.2e-16



3.  See attached code.

**Part V: Data Summary and Implications**

F.  Summary of the findings and assumptions:

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

•   The regression formula is as follows:

(Pricejza, 2023), where Y is the dependent variable, B0 is Y intercept at the time of 0 and B1 is:

[(Σx22)(Σx1y)  – (Σx1x2)(Σx2y)]  / [(Σx12) (Σx22) – (Σx1x2)2]

X1 is the independent variable and E is the random error, residuals (Zach, 2020).

•   The coefficients of the reduced model indicate that the reduced model is statistically significant because the Pr(>|t|) value is less then .05 and the residuals demonstrate even distrubtion.

•   The statistical and practical significance of the reduced model is that there variables with little to no significance on the length of tenure have been reduced out of the model. Each of the remaining variables has significance to the length of tenure but the linear regression model as a whole is not as significant.

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

2.  The recommended course of action based on these results is to utilize this information to address more narrow business questions. We now know that certain variables have a direct impact on length of tenure.

**Part VI: Demonstration**

G.  See the attached Panopto recording demonstration.

**References**

Zach. (2020). *The four assumptions of linear regression.*

[The Four Assumptions of Linear Regression - Statology](https://www.statology.org/linear-regression-assumptions/)

Western Governors University. (2023). *Python or R.*

<https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.htm>

Zach. (2020). *Four examples of using multiple linear regression.*

[4 Examples of Using Linear Regression in Real Life - Statology](https://www.statology.org/linear-regression-real-life-examples/)

Zach. (2021) *How to merge multiple data frames in R.*

[How to Merge Multiple Data Frames in R (With Examples) - Statology](https://www.statology.org/merge-multiple-data-frames-in-r/)

Eloriaga, J. (2021). *Testing for multicollinearity in R.*

[Testing for Multicollinearity in R - YouTube](https://www.youtube.com/watch?v=2HQxVmrX57M)

Zach. (2019). *A complete guide to stepwise regression in R.*

[A Complete Guide to Stepwise Regression in R - Statology](https://www.statology.org/stepwise-regression-r/#Example%202:%20Backward%20Stepwise%20Selection)

Pickard, M. (N.D.). *Information and feature importance.*

[Information and feature importance | R (datacamp.com)](https://campus.datacamp.com/courses/dimensionality-reduction-in-r/foundations-of-dimensionality-reduction?ex=4)

Pricejza. (2023). *Simple linear regression equation example.*

[Simple linear regression equation example - pricejza (weebly.com)](https://pricejza.weebly.com/blog/simple-linear-regression-equation-example)

Zach. (2020). *Multiple linear regression by hand.*

[Multiple Linear Regression by Hand (Step-by-Step) - Statology](https://www.statology.org/multiple-linear-regression-by-hand/)