D208 – Predictive Modeling (Task 1)

Morrell J. Parrish

Western Governors University

Table of Contents

[A. Research Question 3](#_Toc103158799)

[A2. Objective or Goals 3](#_Toc103158800)

[B. Assumptions Summary 3](#_Toc103158801)

[B2. Benefits Of Chosen Analytical Tool(s) 3](#_Toc103158802)

[B3. Chosen Technique Explanation 3](#_Toc103158803)

[C. Data Preparation Description 3](#_Toc103158804)

[C2. Summary of Statistics 3](#_Toc103158805)

[C3. Data Preparation Steps 4](#_Toc103158806)

[C4. Univariate and Bivariate Visualizations 4](#_Toc103158807)

[C5. Churn Data Set 4](#_Toc103158808)

[D. Model Comparison and Analysis 4](#_Toc103158809)

[D2. Justification of Based Variable Selection Procedure and Model Evaluation Metric 4](#_Toc103158810)

[D3. Multiple Regression Model (Categorical and Continuous Variables) 4](#_Toc103158811)

[E. Data Set Analyzation 4](#_Toc103158812)

[E2. Data Set Analyzation 5](#_Toc103158813)

[E3. Regression Model Code 5](#_Toc103158814)

[F. Summary 5](#_Toc103158815)

[F2. Recommended Course of Action 5](#_Toc103158816)

[G. Panopto video recording 5](#_Toc103158817)

[References 7](#_Toc103158818)

D208 – Predictive Modeling (Task 1)

# A. Research Question

During this course of research, we will explore and identify which variable(s) within our dataset affects the churn rate?

## A2. Objective or Goals

The objective of this analysis is to use exploratory data methods to determine which variables within our dataset are indicators for churn. “The churn rate, also known as the rate of attrition or customer churn; is the frequency in which consumers discontinue doing business with a company. It is commonly represented as the percentage of service subscribers who cancel their memberships within a specified time frame” (Frankenfield, 2022).

This analysis will provide clarity on how well a business retains its customers, which in essence may be a reflection on the quality of service the business is providing; however, there can be some limitations within this analysis because this analysis does not take into consideration the types of customers leaving; for example - maybe the customers leaving are the ones who signed up during a promotional period and now the promotional period is over they no longer need your services, so they cancel their subscription(s).

# B. Assumptions Summary

Regression analysis is a set of statistical processes used in statistical modeling to estimate the relationships between a dependent variable (often referred to as the 'outcome' or 'response' variable) and one or more independent variables (often referred to as 'predictors,' 'covariates,' 'explanatory variables,' or 'features'). Linear regression is the most common type of regression analysis, in which the line (or a more complex linear combination) that best fits the data according to a specific mathematical criterion is found (Wikimedia Foundation, 2022).

A linear regression model makes the following assumptions:

* The dependent and independent variables show a linear relationship between the slope and the intercept
* The independent variable is not random
* The value of the residual (error) is zero
* The value of the residual (error) is constant across all observations
* The value of the residual (error) is not correlated across all observations
* The residual (error) values follow the normal distribution.

## B2. Benefits Of Chosen Analytical Tool(s)

The chosen analytical tool for this analysis will be *R*. Both *Python* and *R* have strength and weaknesses; however, *R* is capable of handling very large datasets; the dataset used in this analysis contains 10000 observations and 50 variables. Both *R* and *Python* have packages/libraries which allow you to cleanse, manage, transform, and perform analysis and statistics. Another reason we will be using *R* is because some of its primary purposes are to evaluate statistical relations and create linear regression models.

## B3. Chosen Technique Explanation

Since the variable(s) used to analyze the research question are continuous integers, a multiple regression model is an appropriate technique; also a multiple regression model will allow us to add or remove independent variables, this will help determine if they have an impact on "Churn," the target variable; ultimately, this will influence the company’s decisions.

# C. Data Preparation Description

To use the churn dataset in our analysis we will first need to prepare the data.

The following steps were taken to prepare the dataset for analysis:

* import the dataset into *R*
* evaluate the dataset, remove null or missing values
* remove demographics, and personal identification
  + caseorder, customer\_id, interaction, UID, city, state, county, zip, lat, lng, population, area, timezone, job, email, contacts
* remove any outliners

## C2. Summary of Statistics.

There are 9 continuous variables and 17 categorical variables; there are 10,000 observations and 25 predictor variables and 1 targeted variable (churn). See chart below statistics summary.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Value | Data Type | Statistical Summary |
| children | numerical | continuous | median = 1, mean = 2 |
| age | numerical | continuous | median = 53, mean 53, max = 89, min = 18 |
| income | numerical | continuous | median = 33170.60, mean = 39806.90, max = 258900.70, min = 348.70 |
| marital | partnered, widow, married | categorical, qualitative |  |
| gender | female, male, nonbinary | categorical, qualitative |  |
| churn (targeted variable) | yes or no | categorical, qualitative |  |
| outage\_sec\_perweek | numerical | continuous | median = 10.01856 , mean = 10.00185, min = 0.09975, max = 21.20723 |
| yearly\_equip\_failure | numerical | continuous | median = 0, mean = 0.398, min = 1, max = 6 |
| techie | yes or no | categorical, qualitative |  |
| contract | yes or no | categorical, qualitative |  |
| port\_modem | yes or no | categorical, qualitative |  |
| tablet | yes or no | categorical, qualitative |  |
| internetservice | yes or no | categorical, qualitative |  |
| phone | yes or no | categorical, qualitative |  |
| multiple | yes or no | categorical, qualitative |  |
| onlinesecurity | yes or no | categorical, qualitative |  |
| onlinebackup | yes or no | categorical, qualitative |  |
| deviceprotection | yes or no | categorical, qualitative |  |
| techsupport | yes or no | categorical, qualitative |  |
| streamingtv | yes or no | categorical, qualitative |  |
| streamingmovings | yes or no | categorical, qualitative |  |
| paperlessbilling | yes or no | categorical, qualitative |  |
| paymentmethod | bank transfer (automatic), credit card(automatic), electronic check, mailed | categorical, qualitative |  |
| tenure | numerical | continuous | median = 35.431, mean = 34.526, min = 1, max =71.999 |
| monthlycharge | numerical | continuous | median = 167.48, mean = 167.48, min = 79.98, max = 290.16 |
| bandwidth | numerical | continuous | median = 3279.5, mean = 3392.3, min = 155.5, max = 7159.0 |

## C3. Data Preparation Steps

The following steps were taken to prepare the data for analysis

* import the dataset into *R*

Graphical user interface, text, application

Description automatically generated

* check for missing/null values
  + sapply(churn\_clean, function(x) sum(is.na(x)))

A picture containing graphical user interface

Description automatically generated

* examine the data structure
  + str(churn\_clean)

Text

Description automatically generated

* remove independent variables, demographics, and personal identification not being used in the analysis
  + caseorder, customer\_id, interaction, UID, city, state, county, zip, lat, lng, population, area, timezone, job, marital, email, item1, item2, item3, item4, item4, item6, item7, item8

Text

Description automatically generated

* examine the data summary
  + summary(churn\_clean)

A screenshot of a computer

Description automatically generated with low confidence

**Code:**

*# Install the following libraries*

library(plyr)

library(corrplot)

library(ggplot2)

library(gridExtra)

library(ggthemes)

library(caret)

library(MASS)

library(randomForest)

library(party)

library(readxl)

churn\_clean <- read\_excel("~/Desktop/WGU/WGU2022/D208/d9rkejv84kd9rk30fi2l/churn\_clean.xlsx")

View(churn\_clean)

str(churn\_clean)

summary(churn\_clean)

sapply(churn\_clean, function(x) sum(is.na(x)))

*# Removing non relevant columns from the analysis*

churn\_clean$CaseOrder <- NULL

churn\_clean$Customer\_id <- NULL

churn\_clean$Interaction <- NULL

churn\_clean$UID <- NULL

churn\_clean$City <- NULL

churn\_clean$State <- NULL

churn\_clean$County <- NULL

churn\_clean$Zip <-NULL

churn\_clean$Lat <-NULL

churn\_clean$Lng <-NULL

churn\_clean$Population <-NULL

churn\_clean$TimeZone <-NULL

churn\_clean$Job <-NULL

churn\_clean$Contacts<-NULL

churn\_clean$Email<-NULL

churn\_clean$Item1<-NULL

churn\_clean$Item2 <-NULL

churn\_clean$Item3 <-NULL

churn\_clean$Item4 <-NULL

churn\_clean$Item5 <-NULL

churn\_clean$Item6 <-NULL

churn\_clean$Item7 <-NULL

churn\_clean$Item8 <-NULL

*# View summary data with the removed columns*

summary(churn\_clean)

*# Change Churn Categories to 1 and 2*

churn\_clean$Churn <- as.factor(mapvalues(churn\_clean$Churn, from=c("Yes","No"), to=c("1", "2")))

*# Revalue variable data*

unique(churn\_clean$Area)

data<-churn\_clean$Area

int\_dict<-c(1="Urban" , 2= "Suburban" , 3="Rural" )

int\_val<-revalue (x=data, replace = int\_dict)

view(int\_dict)

churn\_clean$Area<-as.factor(int\_val)

View(churn\_clean)

unique(churn\_clean$Marital)

data<-churn\_clean$Marital

int\_dict<-c("Widowed" = 1, "Married" = 2, "Separated" = 3, "Never Married" = 4, "Divorced" = 5)

int\_val<-revalue (x=data, replace = int\_dict)

view(int\_dict)

churn\_clean$Marital<-as.numeric(int\_val)

View(churn\_clean)

*# Creating Groups for the category Tenure*

group\_Tenure <- function(Tenure){

if (Tenure >= 0 & Tenure <= 12){

return('0-12')

}else if(Tenure >12 & Tenure <= 24){

return('12-24')

}else if (Tenure > 24 & Tenure <= 48){

return('24-48')

}else if (Tenure > 48 & Tenure <=60){

return('48-60')

}else if (Tenure > 60){

return(' > 60')

}

}

churn\_clean$Tenure\_group <- sapply(churn\_clean$Tenure,group\_Tenure)

churn\_clean$Tenure\_group <- as.factor(churn\_clean$Tenure\_group)

*# Creating Groups for the category Age*

group\_Age <- function(Age){

if (Age >= 18 & Age <= 36){

return('18-36')

}else if(Age > 36 & Age <= 54){

return('36-54')

}else if (Age > 54 & Age <= 72){

return('54-72')

}else if (Age > 72){

return('> 72')

}

}

churn\_clean$Age\_group <- sapply(churn\_clean$Age,group\_Age)

churn\_clean$Age\_group <- as.factor(churn\_clean$Age\_group)

*# Creating Groups for the category Children*

group\_Children <- function(Children){

if (Children >= 0 & Children <= 2){

return('0-2')

}else if(Children > 2 & Children <= 4){

return('2-4')

}else if (Children > 4 & Children <= 6){

return('4-6')

}else if (Children > 6 & Children <=8){

return('6-8')

}else if (Children > 8){

return('> 8')

}

}

churn\_clean$Children\_group <- sapply(churn\_clean$Children,group\_Children)

churn\_clean$Children\_group <- as.factor(churn\_clean$Children\_group)

*# Creating Groups for the category Income*

group\_Income <- function(Income){

if (Income >= 0 & Income <= 45000){

return('0-45')

}else if(Income > 45000 & Income <= 90000){

return('45-90')

}else if (Income > 90000 & Income <= 135000){

return('90-135')

}else if (Income > 135000 & Income <=180000){

return('135-180')

}else if (Income >180000 & Income <=225000){

return('180-225')

}else if (Income >225000 ){

return('>225')

}

}

churn\_clean$Income\_group <- sapply(churn\_clean$Income,group\_Income)

churn\_clean$Income\_group <- as.factor(churn\_clean$Income\_group)

*# Creating Groups for the category Outage Secs Per Week*

group\_Outage\_sec\_perweek<- function(Outage\_sec\_perweek){

if (Outage\_sec\_perweek >= 0 & Outage\_sec\_perweek <= 5){

return('0-5')

}else if(Outage\_sec\_perweek >5 & Outage\_sec\_perweek <= 10){

return('5-10')

}else if (Outage\_sec\_perweek >10 & Outage\_sec\_perweek <= 15){

return('10-15')

}else if (Outage\_sec\_perweek >15 & Outage\_sec\_perweek <=20){

return('>15-20')

}else if (Outage\_sec\_perweek >20 ){

return('> 20')

}

}

churn\_clean$ Outage\_sec\_perweek\_group <- sapply(churn\_clean$Outage\_sec\_perweek,group\_Outage\_sec\_perweek)

churn\_clean$ Outage\_sec\_perweek\_group <- as.factor(churn\_clean$Outage\_sec\_perweek\_group)

*# Creating Groups for the category Bandwidth*

group\_Bandwidth\_GB\_Year<- function(Bandwidth){

if (Bandwidth\_GB\_Year >= 0 & Bandwidth\_GB\_Year <= 1500){

return('0-15K')

}else if(Bandwidth\_GB\_Year >5 & Bandwidth\_GB\_Year <= 10){

return('15-30K')

}else if (Bandwidth\_GB\_Year >10 & Bandwidth\_GB\_Year <= 15){

return('30-45K')

}else if (Bandwidth\_GB\_Year >15 & Bandwidth\_GB\_Year <=20){

return('>45-60K')

}else if (Bandwidth\_GB\_Year >20 ){

return('> 60')

}

}

churn\_clean$Bandwidth\_GB\_Year\_group<-sapply(churn\_clean$Bandwidth\_GB\_Year,group\_Bandwidth\_GB\_year)

churn\_clean$Bandwidth\_GB\_Year\_group<-as.factor(churn\_clean$Bandwidth\_GB\_Year\_group)

*# Convert categorical variables to factors*

churn\_clean$Area<-as.factor(churn\_clean$Area)

churn\_clean$Marital<-as.factor(churn\_clean$Marital)

churn\_clean$Gender<-as.factor(churn\_clean$Gender)

churn\_clean$Churn<-as.factor(churn\_clean$Churn)

churn\_clean$Techie<-as.factor(churn\_clean$Techie)

churn\_clean$Contract<-as.factor(churn\_clean$Contract)

churn\_clean$Port\_modem<-as.factor(churn\_clean$Port\_modem)

churn\_clean$Tablet<-as.factor(churn\_clean$Tablet)

churn\_clean$InternetService<-as.factor(churn\_clean$InternetService)

churn\_clean$Phone<-as.factor(churn\_clean$Phone)

churn\_clean$Multiple<-as.factor(churn\_clean$Multiple)

churn\_clean$OnlineSecurity<-as.factor(churn\_clean$OnlineSecurity)

churn\_clean$OnlineBackup<-as.factor(churn\_clean$OnlineBackup)

churn\_clean$DeviceProtection<-as.factor(churn\_clean$DeviceProtection)

churn\_clean$TechSupport<-as.factor(churn\_clean$TechSupport)

churn\_clean$StreamingTV<-as.factor(churn\_clean$StreamingTV)

churn\_clean$StreamingMovies<-as.factor(churn\_clean$StreamingMovies)

churn\_clean$PaperlessBilling<-as.factor(churn\_clean$PaperlessBilling)

churn\_clean$PaymentMethod<-as.factor(churn\_clean$PaymentMethod)

*# Install ggplot2 library and load ggplot2*

packages("ggplot2")

library(ggplot2)

*# Create univariate and bivariate visualizations*

p1 <- ggplot(churn\_clean, aes(x=Area)) + ggtitle("Area") + xlab("Area") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p2 <- ggplot(churn\_clean, aes(x=Marital)) + ggtitle("Marital") + xlab("Marital") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p3 <- ggplot(churn\_clean, aes(x=Gender)) + ggtitle("Gender") + xlab("Gender") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p4 <- ggplot(churn\_clean, aes(x=Techie)) + ggtitle("Techie") + xlab("Techie") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p5<- ggplot(churn\_clean, aes(x=Tenure\_group)) + ggtitle("Tenure Group") + xlab("Tenure") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p6<- ggplot(churn\_clean, aes(x=Churn)) + ggtitle("Churn") + xlab("Churn") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p7<- ggplot(churn\_clean, aes(x=Contract)) + ggtitle("Contract") + xlab("Contract") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p8<- ggplot(churn\_clean, aes(x=Port\_modem)) + ggtitle("Port Modem") + xlab("Port Modem") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p9<- ggplot(churn\_clean, aes(x=Tablet)) + ggtitle("Tablet") + xlab("Tablet") +geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p10<- ggplot(churn\_clean, aes(x=InternetService)) + ggtitle("InternetService") + xlab("InternetService") +geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p11<- ggplot(churn\_clean, aes(x=Phone)) + ggtitle("Phone") + xlab("Phone") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p12<- ggplot(churn\_clean, aes(x=Multiple)) + ggtitle("Multiple") + xlab("Multiple") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p13<- ggplot(churn\_clean, aes(x=OnlineSecurity)) + ggtitle("Online Security") + xlab("Online Security") +geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p14<- ggplot(churn\_clean, aes(x=OnlineBackup)) + ggtitle("Online Backup") + xlab("Online Backup") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p15<- ggplot(churn\_clean, aes(x=DeviceProtection)) + ggtitle("Device Protection") + xlab("Device Protection") + geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p16<- ggplot(churn\_clean, aes(x=TechSupport)) + ggtitle("Tech Support") + xlab("Tech Support") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p17<- ggplot(churn\_clean, aes(x=StreamingTV)) + ggtitle("Streaming TV") + xlab("Streaming TV") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p18<- ggplot(churn\_clean, aes(x=StreamingMovies)) + ggtitle("Streaming Movies") + xlab("Streaming Movies") +geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p19<- ggplot(churn\_clean, aes(x=PaperlessBilling)) + ggtitle("Paperless Billing") + xlab("Paperless Billing") + geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p20<- ggplot(churn\_clean, aes(x=PaymentMethod)) + ggtitle("PaymentMethond") + xlab("PaymentMethod") + geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

p21<- ggplot(churn\_clean, aes(x=Age\_group)) + ggtitle("Age Group") + xlab("Age") +

geom\_bar(aes(y = 100\*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord\_flip() + theme\_minimal()

grid.arrange(p1, p2, p3, p4, p5, ncol=2)

grid.arrange(p6, p7, p8, p9, p10, ncol=2)

grid.arrange(p11, p12, p13, p14, p15, ncol=2)

grid.arrange(p16, p17, p18, p19, p21, ncol=2)

grid.arrange(p20, ncol=1)

*# Create Histograms*

hist(churn\_clean$Age, main="Customers Age", xlab = "Age")

hist(churn\_clean$Children, main="Freq. of Children of Churn Customers", xlab = "Children")

hist(churn\_clean$Income, main="Customer Income", xlab = "Income")

hist(churn\_clean$MonthlyCharge, main="Monthly Charge", xlab = "Monthly Charge")

hist(churn\_clean$Bandwidth\_GB\_Year, main="Yearly Bandwidth Usage", xlab = "Bandwidth")

hist(churn\_clean$MonthlyCharge, main="Monthly Charge", xlab = "Monthly Charge")

*# install.packages("ggcorrplot") and load library*

install.packages("ggcorrplot")

library(ggcorrplot)

*# Create Correlation Matrix*

cor(churn\_clean[, unlist(lapply(churn\_clean, is.numeric))])

*# Create a dataframe of the correlation matirx*

df<-cor(churn\_clean[, unlist(lapply(churn\_clean, is.numeric))])

*# Create Correlation Plot*

corrplot(cor(df))

*# Plot independent variables against targeted variable (churn)*

tp1 <- ggplot(churn\_clean, aes(x=Age\_group,, fill=Churn)) + geom\_bar(position = 'stack',

stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),

'%')), stat = 'count')

tp2 <- ggplot(churn\_clean, aes(x=Children\_group,, fill=Churn)) + geom\_bar(position = 'stack',

stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),

'%')), stat = 'count')

tp3 <- ggplot(churn\_clean, aes(x=Area, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp4 <- ggplot(churn\_clean, aes(x=Outage\_sec\_perweek\_group, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp5<- ggplot(churn\_clean, aes(x=Income\_group,, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp6<- ggplot(churn\_clean, aes(x=Marital, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp7<- ggplot(churn\_clean, aes(x=Gender, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp8<- ggplot(churn\_clean, aes(x=Marital, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp10<- ggplot(churn\_clean, aes(x=Techie, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp11<- ggplot(churn\_clean, aes(x=Contract, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp12<- ggplot(churn\_clean, aes(x=Port\_modem, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp13<- ggplot(churn\_clean, aes(x=Tablet, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp14<- ggplot(churn\_clean, aes(x=InternetService, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp15<- ggplot(churn\_clean, aes(x=Phone, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp16<- ggplot(churn\_clean, aes(x=Multiple, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp17<- ggplot(churn\_clean, aes(x=OnlineSecurity, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp18<- ggplot(churn\_clean, aes(x=OnlineBackup, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp19<- ggplot(churn\_clean, aes(x=DeviceProtection, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp20<- ggplot(churn\_clean, aes(x=TechSupport, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp21<- ggplot(churn\_clean, aes(x=StreamingTV, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp22<- ggplot(churn\_clean, aes(x=StreamingMovies, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp23<- ggplot(churn\_clean, aes(x=Tenure\_group, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), stat = 'count')

tp24<- ggplot(churn\_clean, aes(x=Income\_group, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), position=PositionStack, stat = 'count')

tp25<- ggplot(churn\_clean, aes(x=Bandwidth\_GB\_Year\_group, fill=Churn)) + geom\_bar(position = 'stack',stat='count') + geom\_text(aes(label = paste0(round(prop.table(..count..) \* 100, 2),'%')), position=PositionStack, stat = 'count')

grid.arrange(tp1,tp2, tp3, tp4, tp5, ncol=2)

grid.arrange(tp6, tp7, tp8, tp10, ncol=2)

grid.arrange(tp11, tp12, tp13, tp14, tp15, ncol=2)

grid.arrange(tp16, tp17, tp18, tp19, tp21, ncol=2)

grid.arrange(tp22, tp23, tp24, tp25, ncol=2)

## C4. Univariate and Bivariate Visualizations

A picture containing text, screenshot, receipt

Description automatically generatedA picture containing chart

Description automatically generated

Diagram

Description automatically generated with low confidenceDiagram

Description automatically generated with medium confidence

Chart

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

## C5. Churn Data Set

The prepared dataset used for this analysis has been uploaded with assessment file.

# D. Model Comparison and Analysis

In order to continue our analysis, we need to create our initial multiple regression model

Chart, bubble chart

Description automatically generated

Table

Description automatically generated

Chart, bubble chart

Description automatically generated

## D2. Justification of Based Variable Selection Procedure and Model Evaluation Metric

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

## D3. Multiple Regression Model (Categorical and Continuous Variables)

3. Provide a reduced multiple regression model that includes both categorical and continuous variables.

# E. Data Set Analyzation

1. Explain your data analysis process by comparing the initial and reduced multiple regression models, including the following elements:

• the logic of the variable selection technique

• the model evaluation metric

• a residual plot

## E2. Data Set Analyzation

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

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

## E3. Regression Model Code

*# Create linear regression model*

*# Load the following libraries*

library(linearModel)

require(caTools)

set.seed(123)

sample <- sample.split(churn\_clean$Churn, SplitRatio = .75)

train <-subset(churn\_clean, sample == TRUE)

testing<-subset(churn\_clean, sample == FALSE)

*# Verify the test and training datasets*

dim(train)

[1] 7500 33

dim(testing)

[1] 2500 33

*# View the test and training datasets*

View(train)

View(testing)

*# Build the logical model*

# F. Summary

Summarize your findings and assumptions by doing the following:

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

• a regression equation for the reduced model

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

• the statistical and practical significance of the model

• the limitations of the data analysis

## F2. Recommended Course of Action

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

# G. Panopto video recording

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

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

• an identification of the version of the programming environment

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

• an interpretation of the coefficients.

[Video Link](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e36d2ccf-c985-49d9-b148-ae8a0129ad3d)

References

Frankenfield, J. (2022, February 8). Churn rate. Investopedia. Retrieved May 13, 2022, from https://www.investopedia.com/terms/c/churnrate.asp

Ray, S. (2020, June 26). Questions on multiple regression in R: Python. Analytics Vidhya. Retrieved May 13, 2022, from https://www.analyticsvidhya.com/blog/2015/10/regression-python-beginners/

Wikimedia Foundation. (2022, May 10). Regression analysis. Wikipedia. Retrieved May 13, 2022, from https://en.wikipedia.org/wiki/Regression\_analysis