# D208 – Predictive Modeling (Task 1)

Morrell J. Parrish

Western Governors University

Table of Contents

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*

Code Preview:

```
library(readxl)
churn_clean <- read_excel("~/Desktop/WGU/WGU 2022
/D208/d9rkejv84kd9rk30fi2l/churn_clean.xlsx")
View(churn_clean)
```

- check for missing/null values
  - sapply(churn_clean, function(x) sum(is.na(x)))

```
          CaseOrder         Customer_id         Interaction                 UID
                  0                   0                   0                   0
               City               State              County                 Zip
                  0                   0                   0                   0
                Lat                 Lng          Population                Area
                  0                   0                   0                   0
           TimeZone                 Job            Children                 Age
                  0                   0                   0                   0
             Income             Marital              Gender               Churn
                  0                   0                   0                   0
Outage_sec_perweek               Email           Contacts Yearly_equip_failure
                  0                   0                   0                   0
            Techie            Contract         Port_modem              Tablet
                  0                   0                   0                   0
   InternetService               Phone            Multiple      OnlineSecurity
                  0                   0                   0                   0
      OnlineBackup   DeviceProtection         TechSupport         StreamingTV
                  0                   0                   0                   0
   StreamingMovies     PaperlessBilling       PaymentMethod              Tenure
                  0                   0                   0                   0
     MonthlyCharge    Bandwidth_GB_Year               Item1               Item2
                  0                   0                   0                   0
             Item3               Item4               Item5               Item6
                  0                   0                   0                   0
             Item7               Item8
                  0                   0
```

- examine the data structure
  - str(churn_clean)

```
tibble [10,000 × 50] (S3: tbl_df/tbl/data.frame)
 $ CaseOrder          : num [1:10000] 1 2 3 4 5 6 7 8 9 10 ...
 $ Customer_id        : chr [1:10000] "K409198" "S120509" "K191035" "D90850" ...
 $ Interaction        : chr [1:10000] "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0
f7d4ac2524" "344d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...
 $ UID                : chr [1:10000] "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb
8a" "f1784cfa9f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...
 $ City               : chr [1:10000] "Point Baker" "West Branch" "Yamhill" "Del Mar" ...
 $ State              : chr [1:10000] "AK" "MI" "OR" "CA" ...
 $ County             : chr [1:10000] "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...
 $ Zip                : num [1:10000] 99927 48661 97148 92014 77461 ...
 $ Lat                : num [1:10000] 56.3 44.3 45.4 33 29.4 ...
 $ Lng                : num [1:10000] -133.4 -84.2 -123.2 -117.2 -95.8 ...
 $ Population         : num [1:10000] 38 10446 3735 13863 11352 ...
 $ Area               : chr [1:10000] "Urban" "Urban" "Urban" "Suburban" ...
 $ TimeZone           : chr [1:10000] "America/Sitka" "America/Detroit" "America/Los_Angeles" "America/L
os_Angeles" ...
 $ Job                : chr [1:10000] "Environmental health practitioner" "Programmer, multimedia" "Chie
f Financial Officer" "Solicitor" ...
 $ Children           : num [1:10000] 0 1 4 1 0 3 0 2 2 1 ...
 $ Age                : num [1:10000] 68 27 50 48 83 83 79 30 49 86 ...
 $ Income             : num [1:10000] 28562 21705 9610 18925 40074 ...
 $ Marital            : chr [1:10000] "Widowed" "Married" "Widowed" "Married" ...
 $ Gender             : chr [1:10000] "Male" "Female" "Female" "Male" ...
 $ Churn              : chr [1:10000] "No" "Yes" "No" "No" ...
 $ Outage_sec_perweek : num [1:10000] 7.98 11.7 10.75 14.91 8.15 ...
 $ Email              : num [1:10000] 10 12 9 15 16 15 10 16 20 18 ...
 $ Contacts           : num [1:10000] 0 0 0 2 2 3 0 0 2 1 ...
 $ Yearly_equip_failure: num [1:10000] 1 1 1 0 1 1 1 0 3 0 ...
 $ Techie             : chr [1:10000] "No" "Yes" "Yes" "Yes" ...
 $ Contract           : chr [1:10000] "One year" "Month-to-month" "Two Year" "Two Year" ...
 $ Port_modem         : chr [1:10000] "Yes" "No" "Yes" "No" ...
```

- 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

```
> 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$Job<-NULL
> churn_clean$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
>
```

- examine the data summary
    - summary(churn_clean)

```
> summary(churn_clean)
      Area          Children          Age             Income             Marital           Gender
 Urban   :3327   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Widowed      :2027   Female   :5025
 Suburban:3346   1st Qu.: 1.000   1st Qu.:18.00   1st Qu.:2500   Married      :1911   Male     :4744
 Rural   :3327   Median : 2.000   Median :36.00   Median :4996   Separated    :2014   Nonbinary: 231
                 Mean   : 3.088   Mean   :36.08   Mean   :4997   Never Married:1956
                 3rd Qu.: 4.000   3rd Qu.:54.00   3rd Qu.:7495   Divorced     :2092
                 Max.   :11.000   Max.   :72.00   Max.   :9993
 Churn      Outage_sec_perweek Yearly_equip_failure Techie             Contract    Port_modem Tablet
 No :7350   Min.   :   1       Min.   :1.000        No :8321   Month-to-month:5456   No :5166   No :7009
 Yes:2650   1st Qu.:2501       1st Qu.:1.000        Yes:1679   One year      :2102   Yes:4834   Yes:2991
            Median :5000       Median :1.000                   Two Year      :2442
            Mean   :4997       Mean   :1.398
            3rd Qu.:7492       3rd Qu.:2.000
            Max.   :9986       Max.   :6.000
   InternetService Phone     Multiple   OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV
 DSL        :3463   No : 933   No :5392   No :6424       No :5494     No :5614         No :6250    No :5071
 Fiber Optic:4408   Yes:9067   Yes:4608   Yes:3576       Yes:4506     Yes:4386         Yes:3750    Yes:4929
 None       :2129


 StreamingMovies PaperlessBilling                 PaymentMethod      Tenure       MonthlyCharge
 No :5110        No :4118         Bank Transfer(automatic):2229   Min.   : 1.000   Min.   : 79.98
 Yes:4890        Yes:5882         Credit Card (automatic) :2083   1st Qu.: 7.918   1st Qu.:139.98
                                  Electronic Check        :3398   Median :35.431   Median :167.48
                                  Mailed Check            :2290   Mean   :34.526   Mean   :172.62
                                                                  3rd Qu.:61.480   3rd Qu.:200.73
                                                                  Max.   :71.999   Max.   :290.16

 Bandwidth_GB_Year      Tenure_group    Age_group
 Min.   : 155.5    > 60 Month :2805   Min.   :1.00
 1st Qu.:1236.5    0-12 Month :3643   1st Qu.:1.00
 Median :3279.5    12-24 Month:1253   Median :2.00
 Mean   :3392.3    24-48 Month: 666   Mean   :2.48
 3rd Qu.:5586.1    48-60 Month:1633   3rd Qu.:3.00
 Max.   :7159.0                       Max.   :4.00
```

## 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')
```

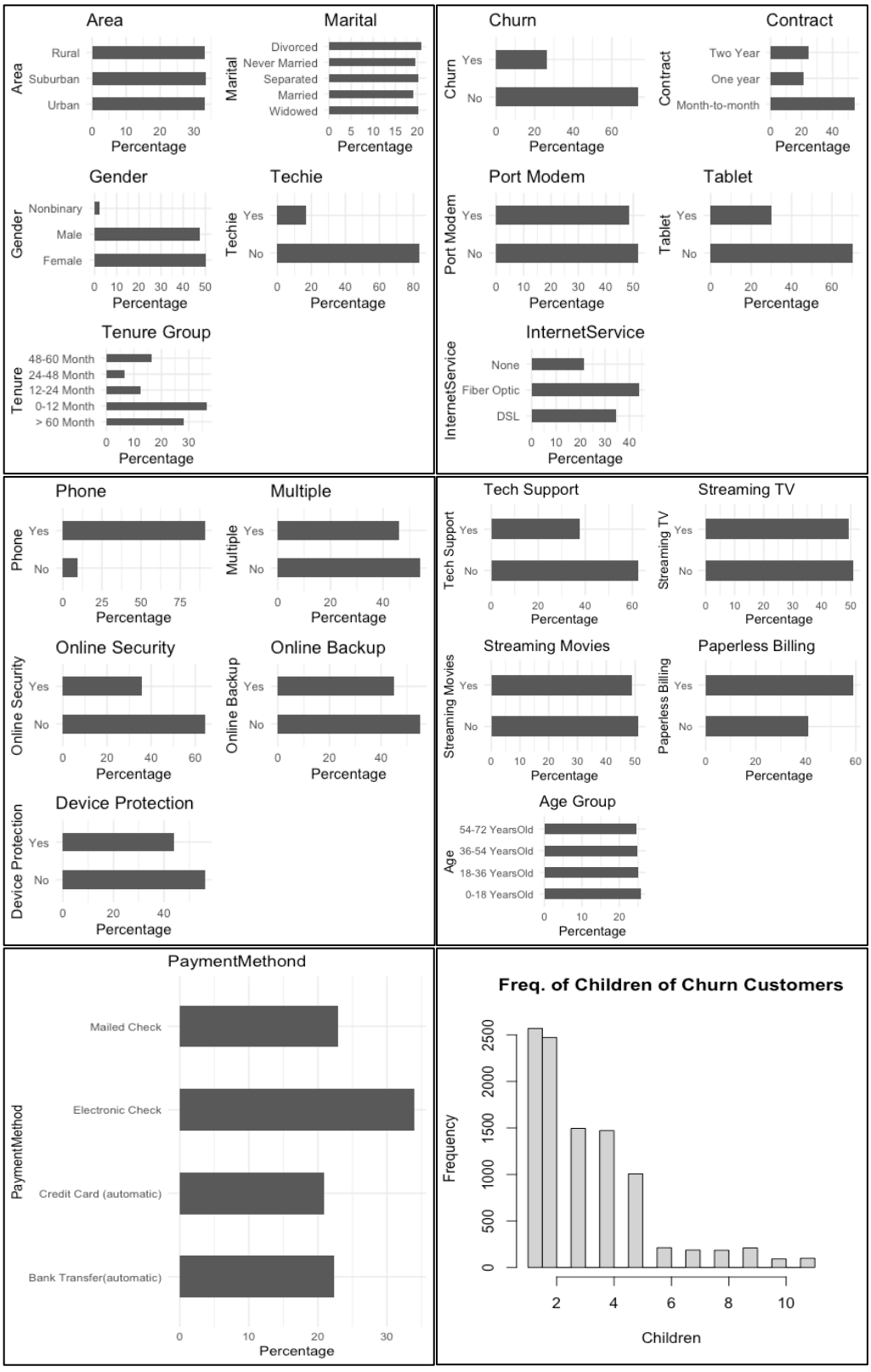
```
    }else if (Outage_sec_perweek >20 ){
       return('> 20')
    }
 }

 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)
```

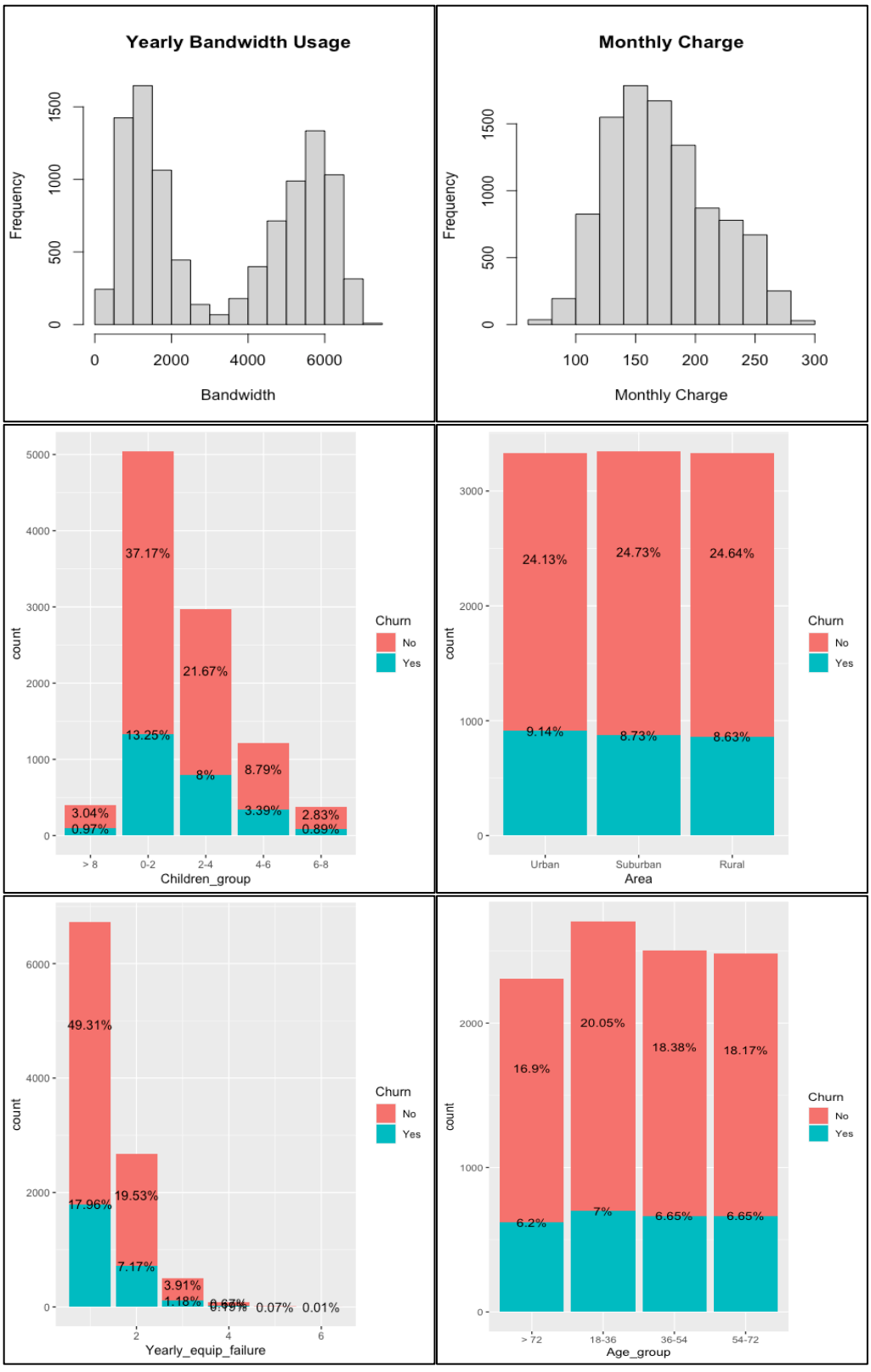
```
# Create univariate and bivariate visualizations

 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()
```

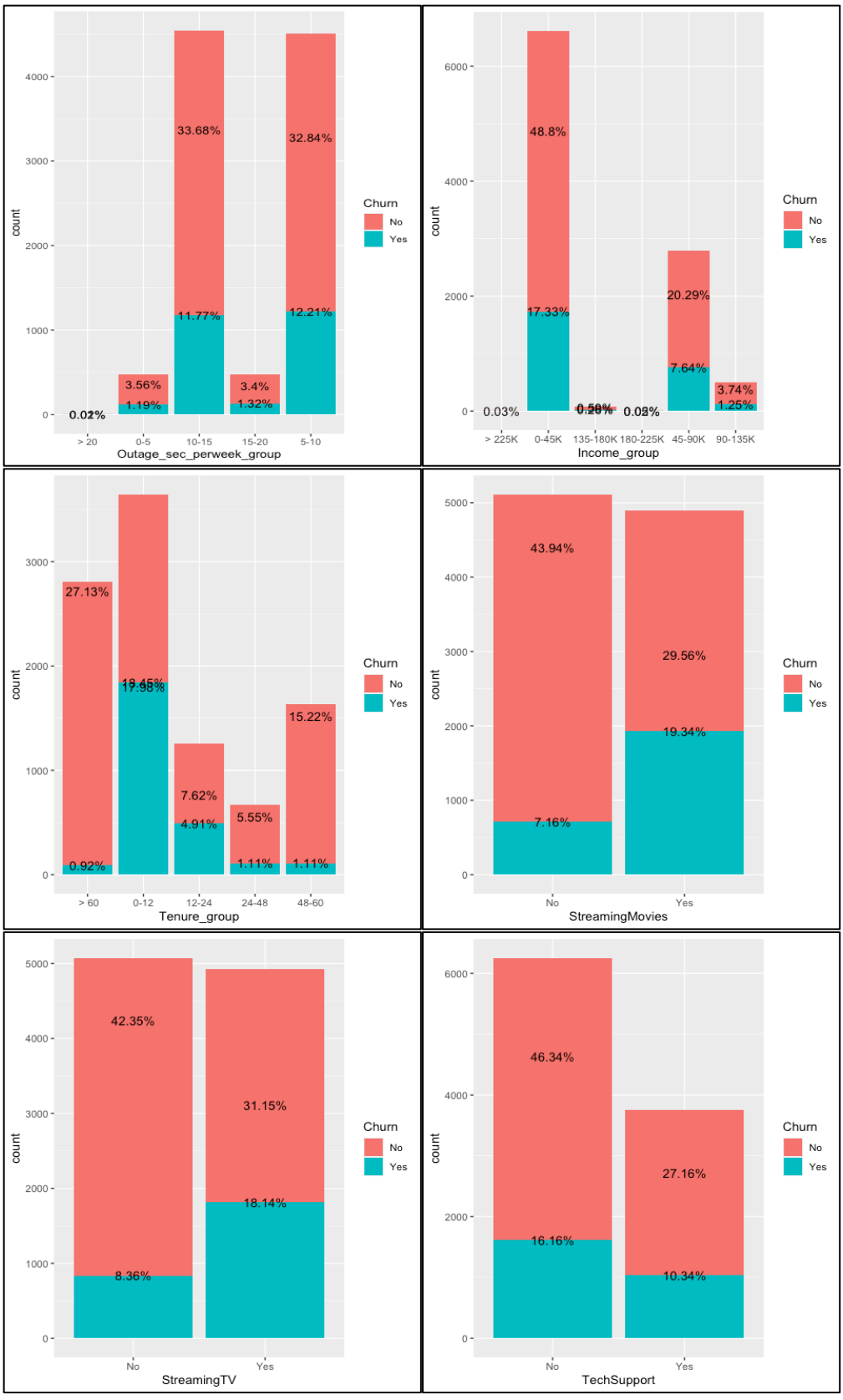
```
 p18<- ggplot(churn_clean, aes(x=StreamingMovies)) + ggtitle("Streaming Movies") + xlab("Streaming Movies") +geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_minimal()
 p19<- ggplot(churn_clean, aes(x=PaperlessBilling)) + ggtitle("Paperless 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')
```

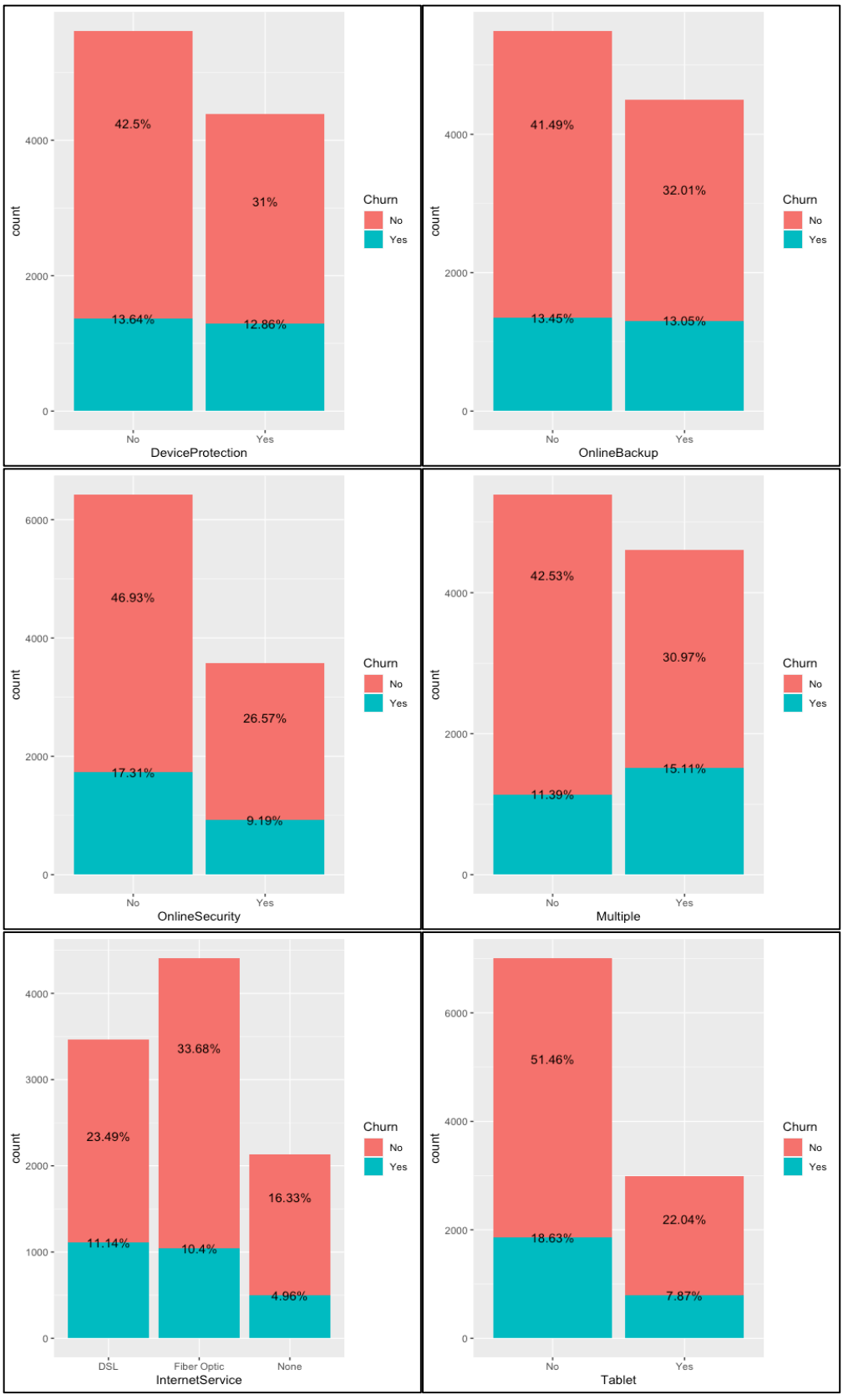
```
 tp5<- ggplot(churn_clean, aes(x=Income_group,, fill=Churn)) + geom_bar(position = 'stack',stat='count') +
geom_text(aes(label = paste0(round(prop.table(..count..) * 100, 2),'%')), stat = 'count')
 tp6<- ggplot(churn_clean, aes(x=Marital, fill=Churn)) + geom_bar(position = 'stack',stat='count') +
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)
```

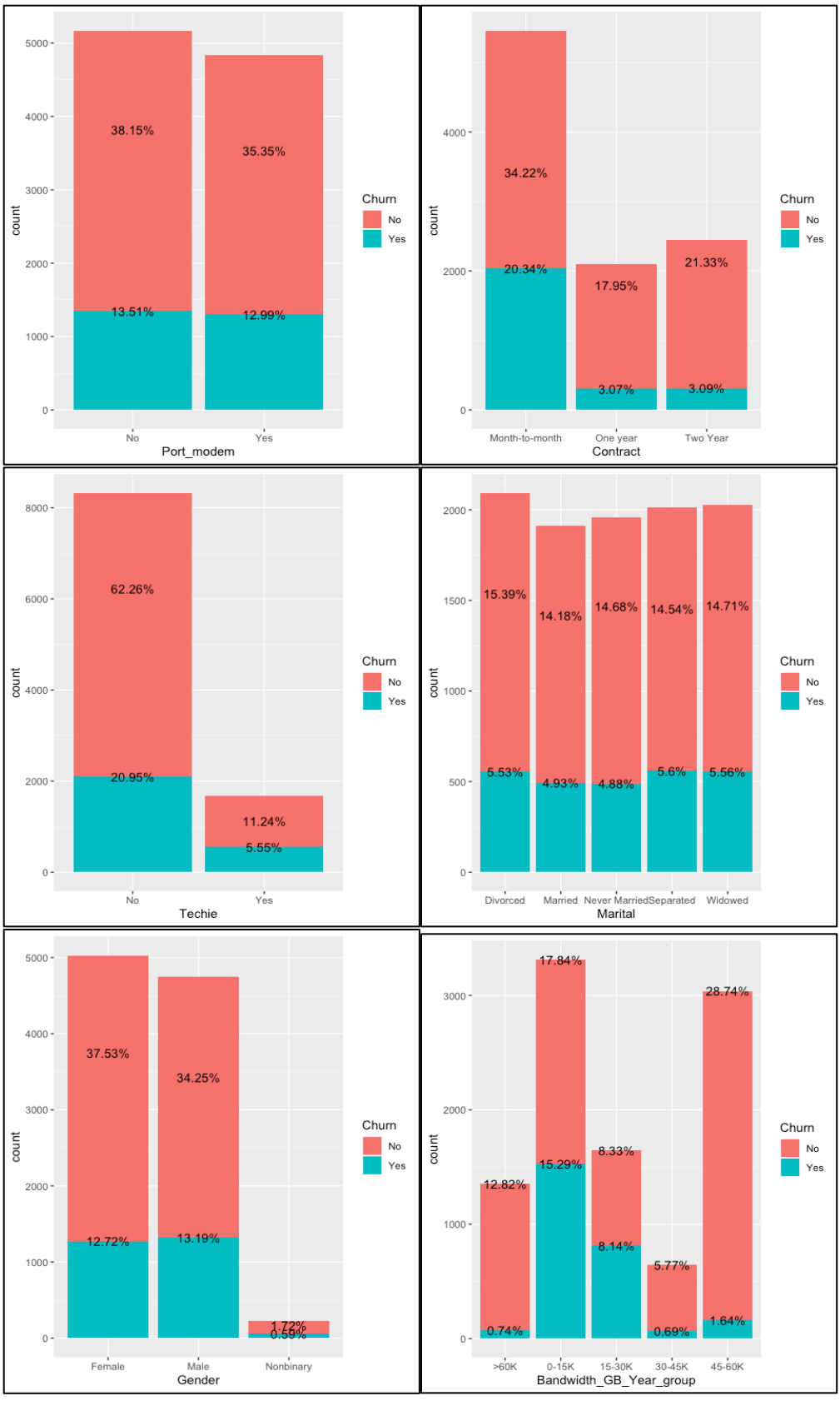
## C4. Univariate and Bivariate Visualizations




Yearly Bandwidth Usage
Frequency
0
500
1000
1500
0
2000
4000
6000
Bandwidth
Monthly Charge
Frequency
0
500
1000
1500
100
150
200
250
300
Monthly Charge
count
5000
4000
3000
2000
1000
0
37.17%
21.67%
13.25%
8%
8.79%
3.39%
3.04%
0.97%
2.83%
0.89%
> 8
0-2
2-4
4-6
6-8
Children_group
Churn
No
Yes
count
3000
2000
1000
0
24.13%
24.73%
24.64%
9.14%
8.73%
8.63%
Urban
Suburban
Rural
Area
Churn
No
Yes
count
6000
4000
2000
0
49.31%
17.96%
19.53%
7.17%
3.91%
1.18%
0.67%
0.19%
0.07%
0.01%
2
4
6
Yearly_equip_failure
Churn
No
Yes
count
2000
1000
0
16.9%
6.2%
20.05%
7%
18.38%
6.65%
18.17%
6.65%
> 72
18-36
36-54
54-72
Age_group
Churn
No
Yes


count
Churn
No
Yes
0.02%
3.56%
1.19%
33.68%
11.77%
3.4%
1.32%
32.84%
12.21%
> 20
0-5
10-15
15-20
5-10
Outage_sec_perweek_group
0.03%
48.8%
17.33%
20.29%
7.64%
3.74%
1.25%
> 225K
0-45K
135-180K
180-225K
45-90K
90-135K
Income_group
27.13%
0.92%
18.45%
17.98%
7.62%
4.91%
5.55%
1.11%
15.22%
1.11%
> 60
0-12
12-24
24-48
48-60
Tenure_group
43.94%
7.16%
29.56%
19.34%
No
Yes
StreamingMovies
42.35%
8.36%
31.15%
18.14%
No
Yes
StreamingTV
46.34%
16.16%
27.16%
10.34%
No
Yes
TechSupport


42.5%
31%
13.64%
12.86%
count
Churn
No
Yes
No
Yes
DeviceProtection
41.49%
32.01%
13.45%
13.05%
OnlineBackup
46.93%
26.57%
17.31%
9.19%
OnlineSecurity
42.53%
30.97%
11.39%
15.11%
Multiple
23.49%
33.68%
16.33%
11.14%
10.4%
4.96%
DSL
Fiber Optic
None
InternetService
51.46%
22.04%
18.63%
7.87%
Tablet

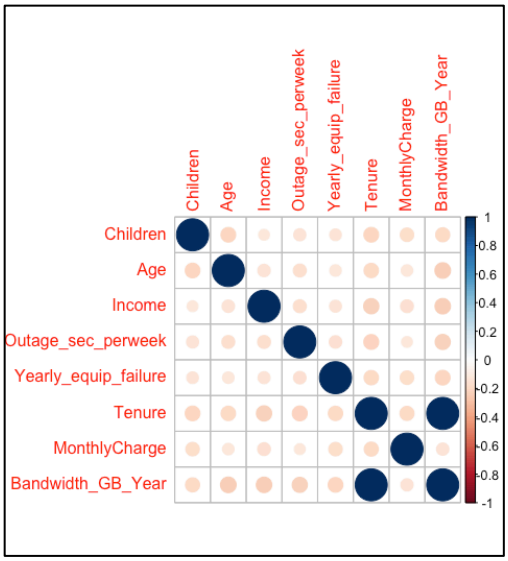

count
Churn
No
Yes
38.15%
35.35%
13.51%
12.99%
No
Yes
Port_modem
34.22%
20.34%
17.95%
3.07%
21.33%
3.09%
Month-to-month
One year
Two Year
Contract
62.26%
20.95%
11.24%
5.55%
Techie
15.39%
5.53%
14.18%
4.93%
14.68%
4.88%
14.54%
5.6%
14.71%
5.56%
Divorced
Married
Never Married
Separated
Widowed
Marital
37.53%
12.72%
34.25%
13.19%
1.72%
0.55%
Female
Male
Nonbinary
Gender
12.82%
0.74%
17.84%
15.29%
8.33%
8.14%
5.77%
0.69%
28.74%
1.64%
>60K
0-15K
15-30K
30-45K
45-60K
Bandwidth_GB_Year_group

## C5. Churn Data Set

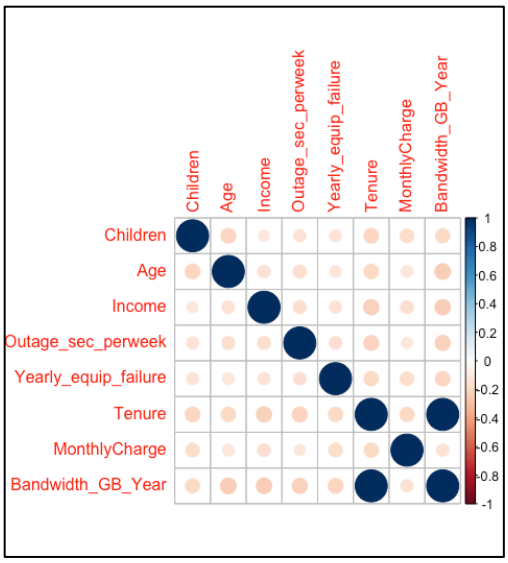
The prepared dataset used for this analysis has been uploaded with assessment file.

## D. Model Comparison and Analysis

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




```
> cor(churn_clean[, unlist(lapply(churn_clean, is.numeric))])
                        Children          Age       Income Outage_sec_perweek Yearly_equip_failure       Tenure
Children             1.000000000 -0.029731540  0.013353408       0.0026149827         0.0074829356 -0.005091318
Age                 -0.029731540  1.000000000 -0.002955682      -0.0077870124         0.0084594658  0.016979273
Income               0.013353408 -0.002955682  1.000000000      -0.0061159834         0.0048119913 -0.001023863
Outage_sec_perweek   0.002614983 -0.007787012 -0.006115983       1.0000000000         0.0009642827  0.004243537
Yearly_equip_failure 0.007482936  0.008459466  0.004811991       0.0009642827         1.0000000000  0.012311950
Tenure              -0.005091318  0.016979273 -0.001023863       0.0042435366         0.0123119498  1.000000000
MonthlyCharge       -0.009781399  0.010728512 -0.001233154       0.0184606529        -0.0070524907 -0.003336810
Bandwidth_GB_Year    0.025584816 -0.014723648  0.000873916       0.0055929965         0.0119099970  0.991495192
                    MonthlyCharge Bandwidth_GB_Year
Children            -0.009781399       0.025584816
Age                  0.010728512      -0.014723648
Income              -0.001233154       0.000873916
Outage_sec_perweek   0.018460653       0.005592997
Yearly_equip_failure -0.007052491       0.011909997
Tenure              -0.003336810       0.991495192
MonthlyCharge        1.000000000       0.060406431
Bandwidth_GB_Year    0.060406431       1.000000000
```




**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

# 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