Customer Churn Prediction

By Group One: Okeke Onyedikachukwu, Babatunde Odumuyiwa, Maduabughichi Achilefu, Eeshaan Ali Syed

Andy Chen 11/23/23 Probability and Statistics

Table of Contents

Introduction

About Dataset

With the rapid development of telecommunication industry, the service providers are inclined more towards expansion of the subscriber base. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. It is stated that the cost of acquiring a new customer is far more than that for retaining the existing one. Therefore, it is imperative for the telecom industries to use advanced analytics to understand consumer behavior and in-turn predict the association of the customers as whether or not they will leave the company.

Content

This data set contains customer level information for a telecom company. Various attributes related to the services used are recorded for each customer.

Problem Statement

Some possible insights could be -

What variables are contributing to customer churn?

What actions can be taken to stop them from leaving?

Each row represents a customer and each column contains attributes related to customer as described in the column description.

Dataset Description

Features	Description	Data Type
Churn	1 if customer cancelled service, 0 if not	Integer
AccountWeeks	Number of weeks customer has had active account	Integer
ContractRenewal	1 if customer recently renewed contract, 0 if not	Integer
DataPlan	1 if customer has data plan, 0 if not	Integer
DataUsage	Gigabytes of monthly data usage	Number
CustServCalls	Number of calls into customer service	Integer
DayMins	Average daytime minutes per month	Number
DayCalls	Average number of daytime calls	Integer
MonthlyCharge	Average monthly bill	Number
OverageFee	Largest overage fee in last 12 months	Number
RoamingMins	Roaming calls by customer	Number
DataUsagecat	Categorization of data users based on data consumed	Character

PHASE I

Introduction of Data set and Data Summary

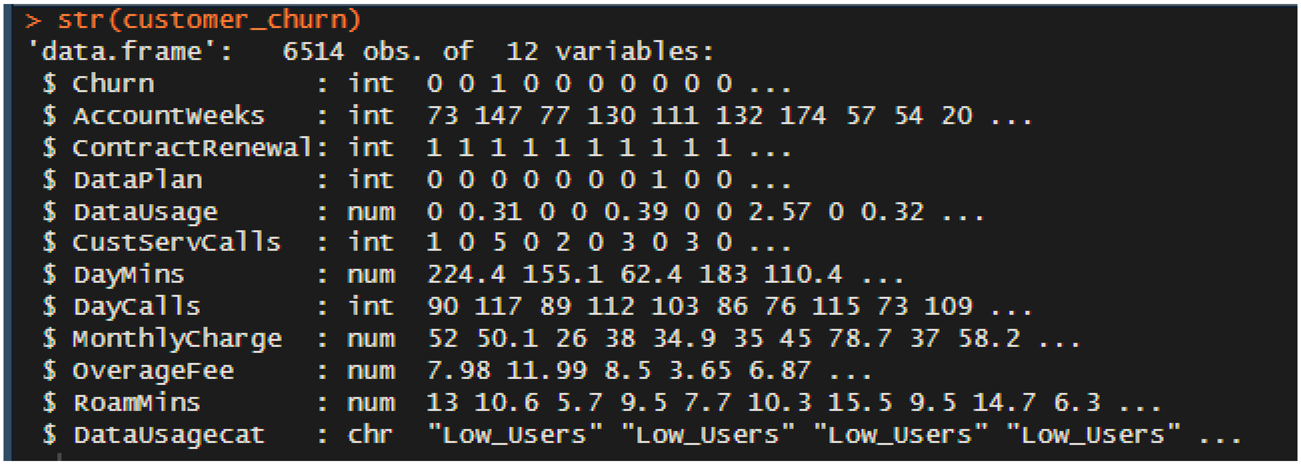
Reading the data in R, summarizing and checking for noise in the dataset.

The first phase is to read the into R using the code below:

```
Customer_churn <- read.csv("C:/Users/onyif/Downloads/customer_churn.csv")
Customer_churn
```

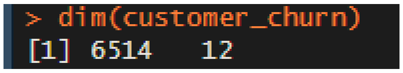
A sub-categorization was introduced to the data set were an extra column created was called DataUsagecat was created to further categorize data users into High, medium and low data users. The code below was used to achieve the sub-categorization.

```
Customer_churn$DataUsagecat[Customer_churn$DataUsage <1] <- "Low_Users"
Customer_churn$DataUsagecat[Customer_churn$DataUsage >=1 & Customer_churn$DataUsage <=2.7] <-"Avg
_Users"
Customer_churn$DataUsagecat[Customer_churn$DataUsage >2.7] <-"High_Users"
```



```
> str(customer_churn)
'data.frame':	6514 obs. of  12 variables:
 $ Churn          : int  0 0 1 0 0 0 0 0 0 0 ...
 $ AccountWeeks   : int  73 147 77 130 111 132 174 57 54 20 ...
 $ ContractRenewal: int  1 1 1 1 1 1 1 1 1 1 ...
 $ DataPlan       : int  0 0 0 0 0 0 0 1 0 0 ...
 $ DataUsage      : num  0 0.31 0 0 0.39 0 0 2.57 0 0.32 ...
 $ CustServCalls  : int  1 0 5 0 2 0 3 0 3 0 ...
 $ DayMins        : num  224.4 155.1 62.4 183 110.4 ...
 $ DayCalls       : int  90 117 89 112 103 86 76 115 73 109 ...
 $ MonthlyCharge  : num  52 50.1 26 38 34.9 35 45 78.7 37 58.2 ...
 $ OverageFee     : num  7.98 11.99 8.5 3.65 6.87 ...
 $ RoamMins       : num  13 10.6 5.7 9.5 7.7 10.3 15.5 9.5 14.7 6.3 ...
 $ DataUsagecat   : chr  "Low_Users" "Low_Users" "Low_Users" "Low_Users" ...
```

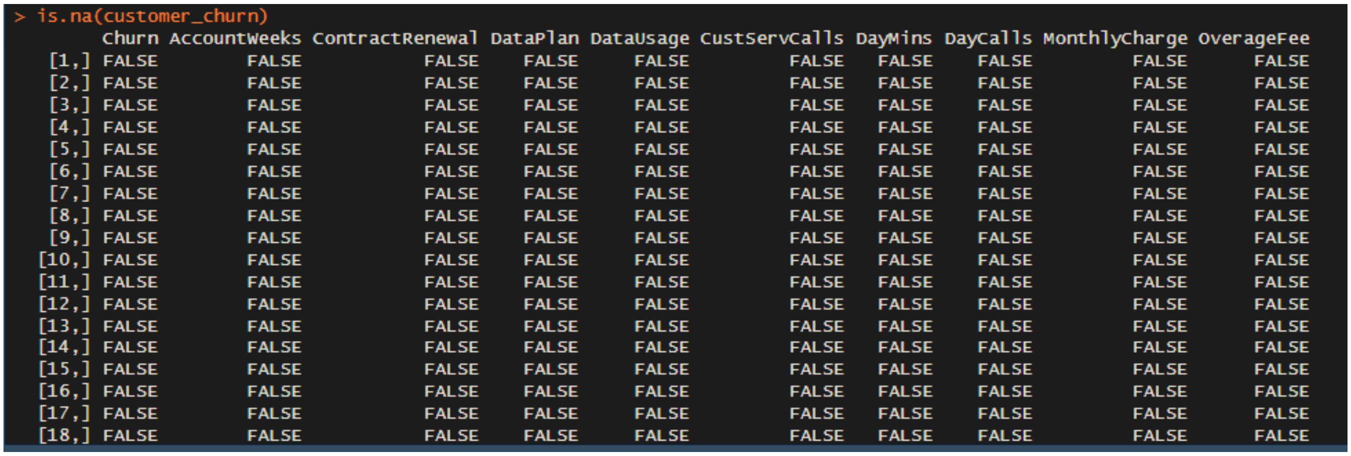
```
dim(Customer_churn)
```



```
> dim(customer_churn)
[1] 6514   12
```

The data set as seen above has 6,514 rows and 12 columns

Also, we have to check for NA or missing from the data set, this can be achieved using the code below is.na(Customer_churn)

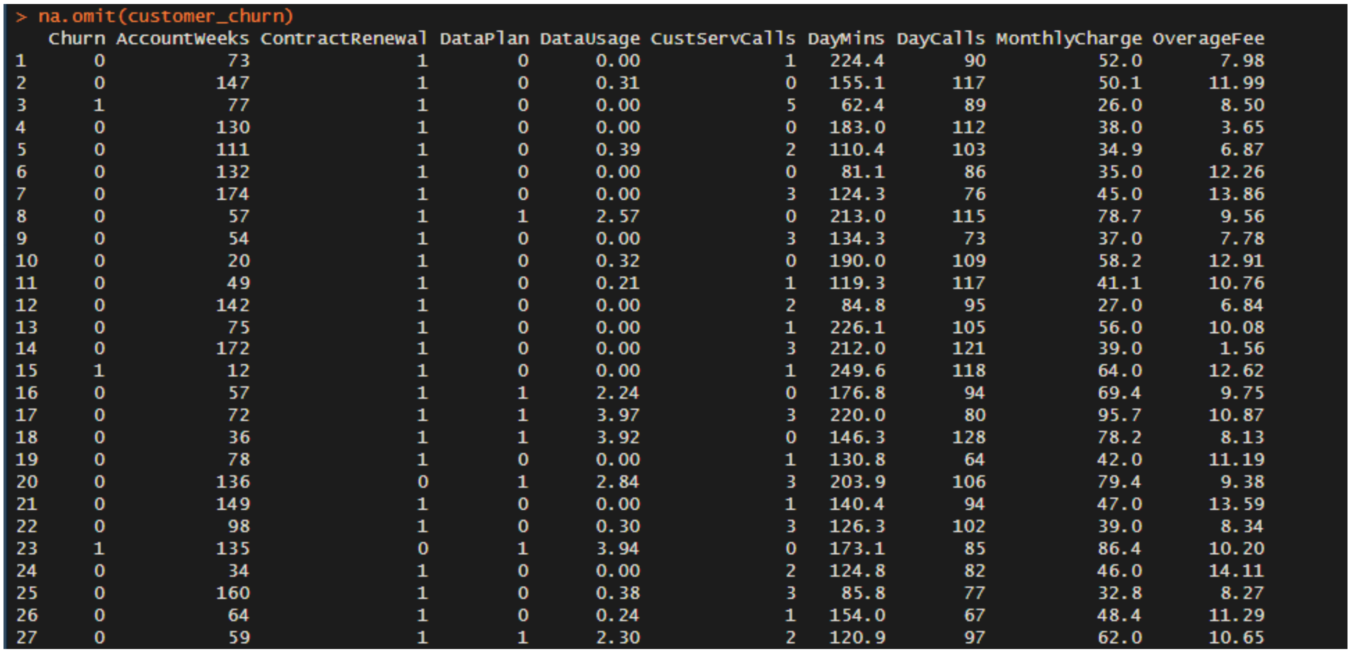


```
> is.na(customer_churn)
       Churn AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls DayMins DayCalls MonthlyCharge OverageFee
  [1,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [2,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [3,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [4,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [5,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [6,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [7,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [8,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
  [9,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [10,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [11,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [12,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [13,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [14,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [15,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [16,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [17,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
 [18,] FALSE        FALSE           FALSE    FALSE     FALSE         FALSE   FALSE    FALSE         FALSE      FALSE
```

The output shows no NA or missing values from our data set.

To double check we need to omit any NA/missing values from our data set, this can be achieved by application of the code below:

na.omit(Customer_churn)

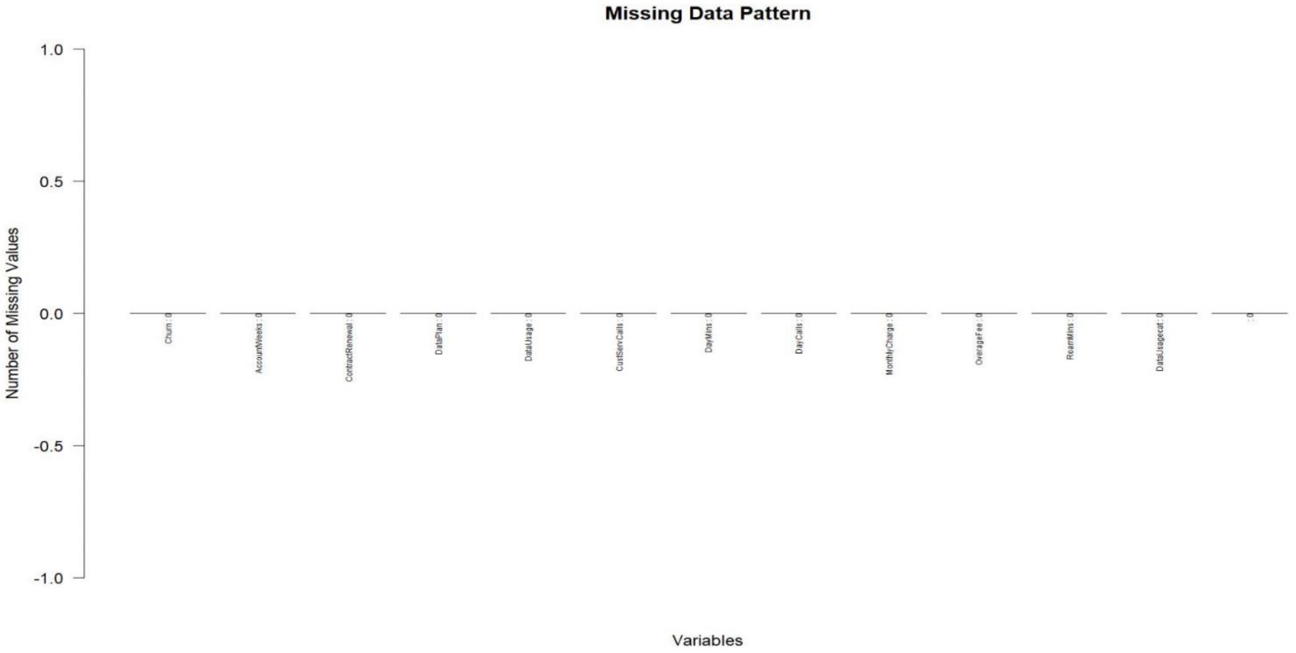


```
> na.omit(customer_churn)
   Churn AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls DayMins DayCalls MonthlyCharge OverageFee
1      0           73               1        0      0.00             1   224.4       90          52.0       7.98
2      0          147               1        0      0.31             0   155.1      117          50.1      11.99
3      1           77               1        0      0.00             5    62.4       89          26.0       8.50
4      0          130               1        0      0.00             0   183.0      112          38.0       3.65
5      0          111               1        0      0.39             2   110.4      103          34.9       6.87
6      0          132               1        0      0.00             0    81.1       86          35.0      12.26
7      0          174               1        0      0.00             3   124.3       76          45.0      13.86
8      0           57               1        1      2.57             0   213.0      115          78.7       9.56
9      0           54               1        0      0.00             3   134.3       73          37.0       7.78
10     0           20               1        0      0.32             0   190.0      109          58.2      12.91
11     0           49               1        0      0.21             1   119.3      117          41.1      10.76
12     0          142               1        0      0.00             2    84.8       95          27.0       6.84
13     0           75               1        0      0.00             1   226.1      105          56.0      10.08
14     0          172               1        0      0.00             3   212.0      121          39.0       1.56
15     1           12               1        0      0.00             1   249.6      118          64.0      12.62
16     0           57               1        1      2.24             0   176.8       94          69.4       9.75
17     0           72               1        1      3.97             3   220.0       80          95.7      10.87
18     0           36               1        1      3.92             0   146.3      128          78.2       8.13
19     0           78               1        0      0.00             1   130.8       64          42.0      11.19
20     0          136               0        1      2.84             3   203.9      106          79.4       9.38
21     0          149               1        0      0.00             1   140.4       94          47.0      13.59
22     0           98               1        0      0.30             3   126.3      102          39.0       8.34
23     1          135               0        1      3.94             0   173.1       85          86.4      10.20
24     0           34               1        0      0.00             2   124.8       82          46.0      14.11
25     0          160               1        0      0.38             3    85.8       77          32.8       8.27
26     0           64               1        0      0.24             1   154.0       67          48.4      11.29
27     0           59               1        1      2.30             2   120.9       97          62.0      10.65
```

na.omit package omits NA or missing values. The output shows NA or missing values has been omitted from the data set.

Seeing from our data set that there is no missing values or NA, it is important to check for noise in our data set and to achieve this we apply the mice function below:

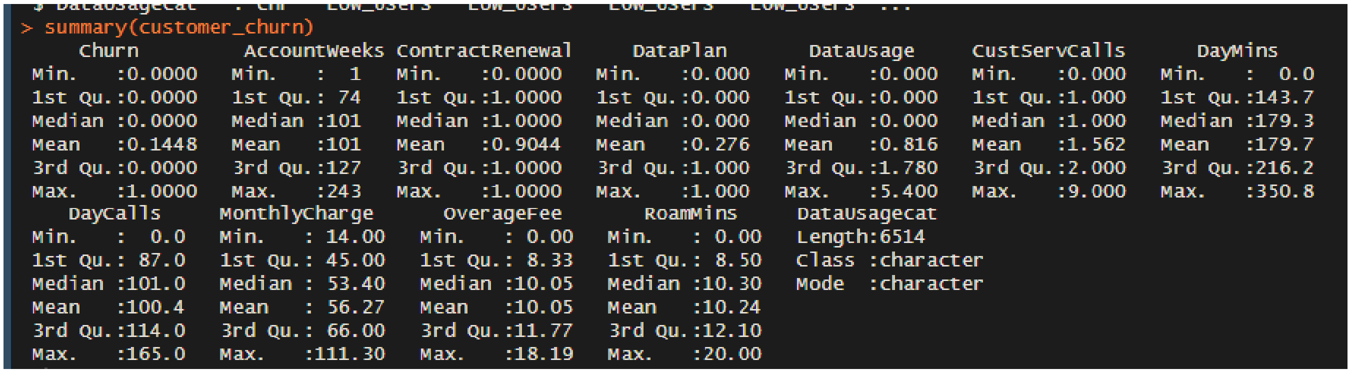
```
library(mice)
md.pattern(Customer_churn)
```



The output shows no NA or missing value pattern from our data set.

We need to have a brief summary of our data set and to achieve this we make use of the code below:

```
summary(Customer_churn)
```



```
 $ DataUsagecat   : chr  Low_users   Low_users   Low_users   Low_users  ...
> summary(customer_churn)
     Churn         AccountWeeks ContractRenewal     DataPlan       DataUsage      CustServCalls      DayMins    
 Min.   :0.0000   Min.   :  1   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :  0.0  
 1st Qu.:0.0000   1st Qu.: 74   1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:143.7  
 Median :0.0000   Median :101   Median :1.0000   Median :0.000   Median :0.000   Median :1.000   Median :179.3  
 Mean   :0.1448   Mean   :101   Mean   :0.9044   Mean   :0.276   Mean   :0.816   Mean   :1.562   Mean   :179.7  
 3rd Qu.:0.0000   3rd Qu.:127   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.780   3rd Qu.:2.000   3rd Qu.:216.2  
 Max.   :1.0000   Max.   :243   Max.   :1.0000   Max.   :1.000   Max.   :5.400   Max.   :9.000   Max.   :350.8  
    DayCalls      MonthlyCharge      OverageFee       RoamMins     DataUsagecat      
 Min.   :  0.0   Min.   : 14.00   Min.   : 0.00   Min.   : 0.00   Length:6514       
 1st Qu.: 87.0   1st Qu.: 45.00   1st Qu.: 8.33   1st Qu.: 8.50   Class :character  
 Median :101.0   Median : 53.40   Median :10.05   Median :10.30   Mode  :character  
 Mean   :100.4   Mean   : 56.27   Mean   :10.05   Mean   :10.24                     
 3rd Qu.:114.0   3rd Qu.: 66.00   3rd Qu.:11.77   3rd Qu.:12.10                     
 Max.   :165.0   Max.   :111.30   Max.   :18.19   Max.   :20.00                     
```

The summary package is applied to summarize our data set highlighting valuable insight in our data.

Phase II

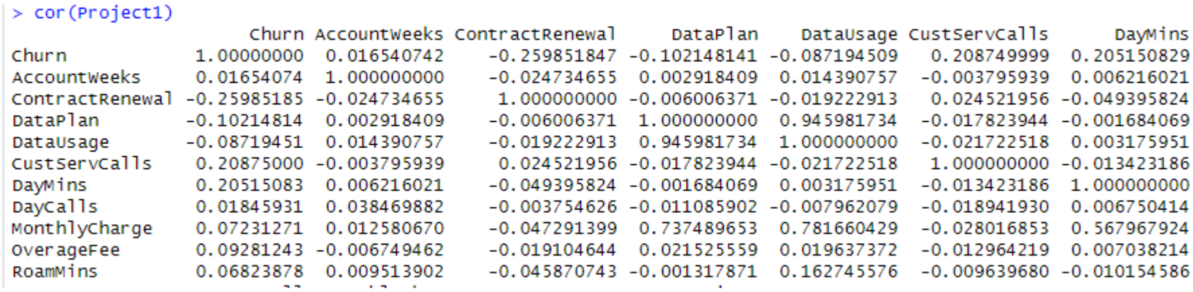
Data Analysis and Visualization

This phase carrying out analysis on the data set using different functions to help analyze and provide valuable insights to the problem statement

Find the correlation among the numeric feature in the data set

```
####Correlation plot of all features #####

cor(Customer_churn)
```

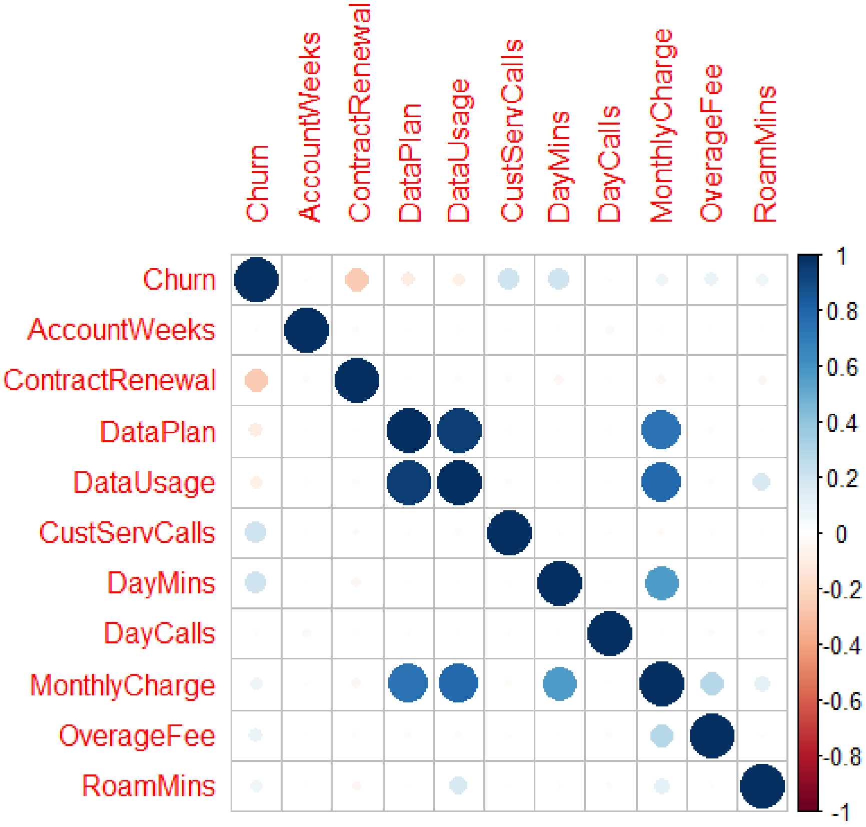


```
> cor(Project1)
                      Churn AccountWeeks ContractRenewal     DataPlan    DataUsage CustServCalls      DayMins
Churn            1.00000000  0.016540742    -0.259851847 -0.102148141 -0.087194509   0.208749999  0.205150829
AccountWeeks     0.01654074  1.000000000    -0.024734655  0.002918409  0.014390757  -0.003795939  0.006216021
ContractRenewal -0.25985185 -0.024734655     1.000000000 -0.006006371 -0.019222913   0.024521956 -0.049395824
DataPlan        -0.10214814  0.002918409    -0.006006371  1.000000000  0.945981734  -0.017823944 -0.001684069
DataUsage       -0.08719451  0.014390757    -0.019222913  0.945981734  1.000000000  -0.021722518  0.003175951
CustServCalls    0.20875000 -0.003795939     0.024521956 -0.017823944 -0.021722518   1.000000000 -0.013423186
DayMins          0.20515083  0.006216021    -0.049395824 -0.001684069  0.003175951  -0.013423186  1.000000000
DayCalls         0.01845931  0.038469882    -0.003754626 -0.011085902 -0.007962079  -0.018941930  0.006750414
MonthlyCharge    0.07231271  0.012580670    -0.047291399  0.737489653  0.781660429  -0.028016853  0.567967924
OverageFee       0.09281243 -0.006749462    -0.019104644  0.021525559  0.019637372  -0.012964219  0.007038214
RoamMins         0.06823878  0.009513902    -0.045870743 -0.001317871  0.162745576  -0.009639680 -0.010154586
```

To make more sense from the data set a corrplot function was used to create a correlation plot for easy inter pretation of the data.

```
##### produce one correlation plot #####
corrplot(Project_cor,method="circle",mar=c(1,1,0,1))
```

Correlation Plot



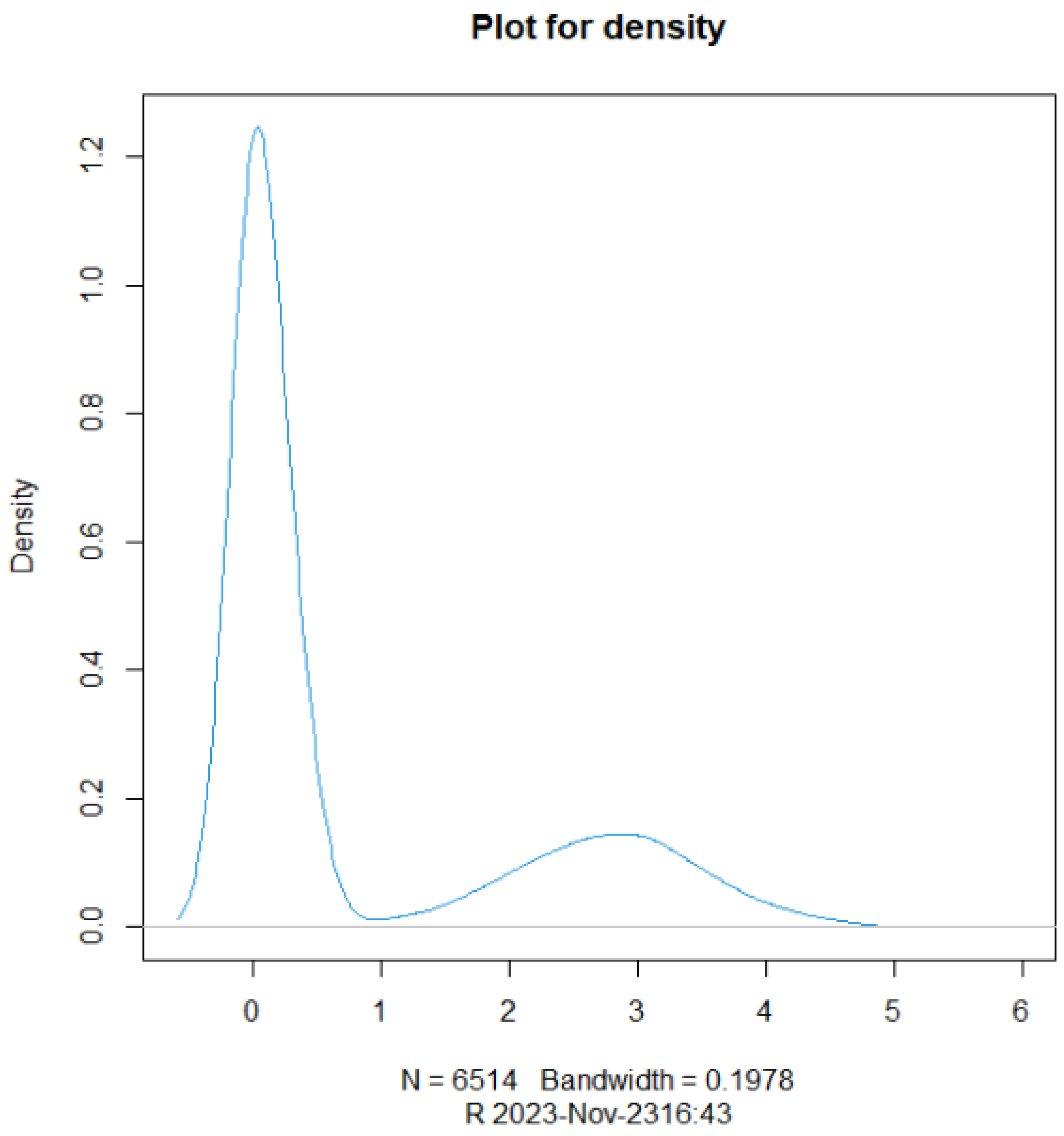
The data has shown some trends which is highlighted below:

- The data shows a strong correlation between **data plan and data usage** which meanings that as **data subscription grows data usage grows as well**.
- The data also shows a **weak correlation between churn and data usage which means that customers who frequently buy data plans are not likely to churn out of the network.**

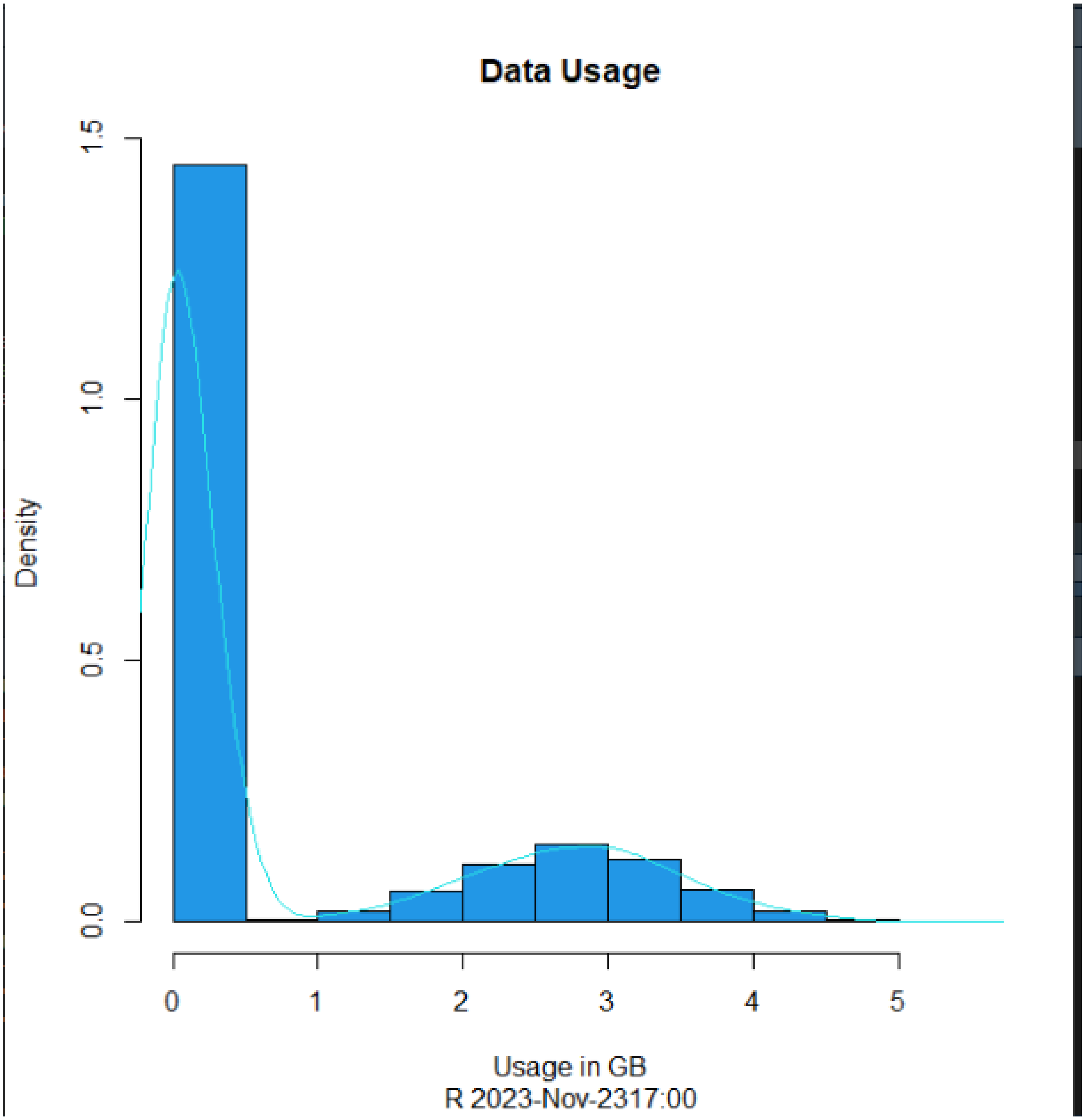
- The data also shows there is a weak correlation between Contract Renewal and churn, which indicates **that customers who renew their contract are not likely to churn out of the network**.
- The data shows a strong relationship between monthly charge and data usage/data plan indicating that customers who buy monthly plan and use data contributes to the revenue generation of the network

Analysis from the correlation plot has shown that customers who buy data plan are less likely to churn, so let's have an idea of data usage and active data plan distribution from the data set. This is achieved using various visualization graphs

Density and Histogram Plot

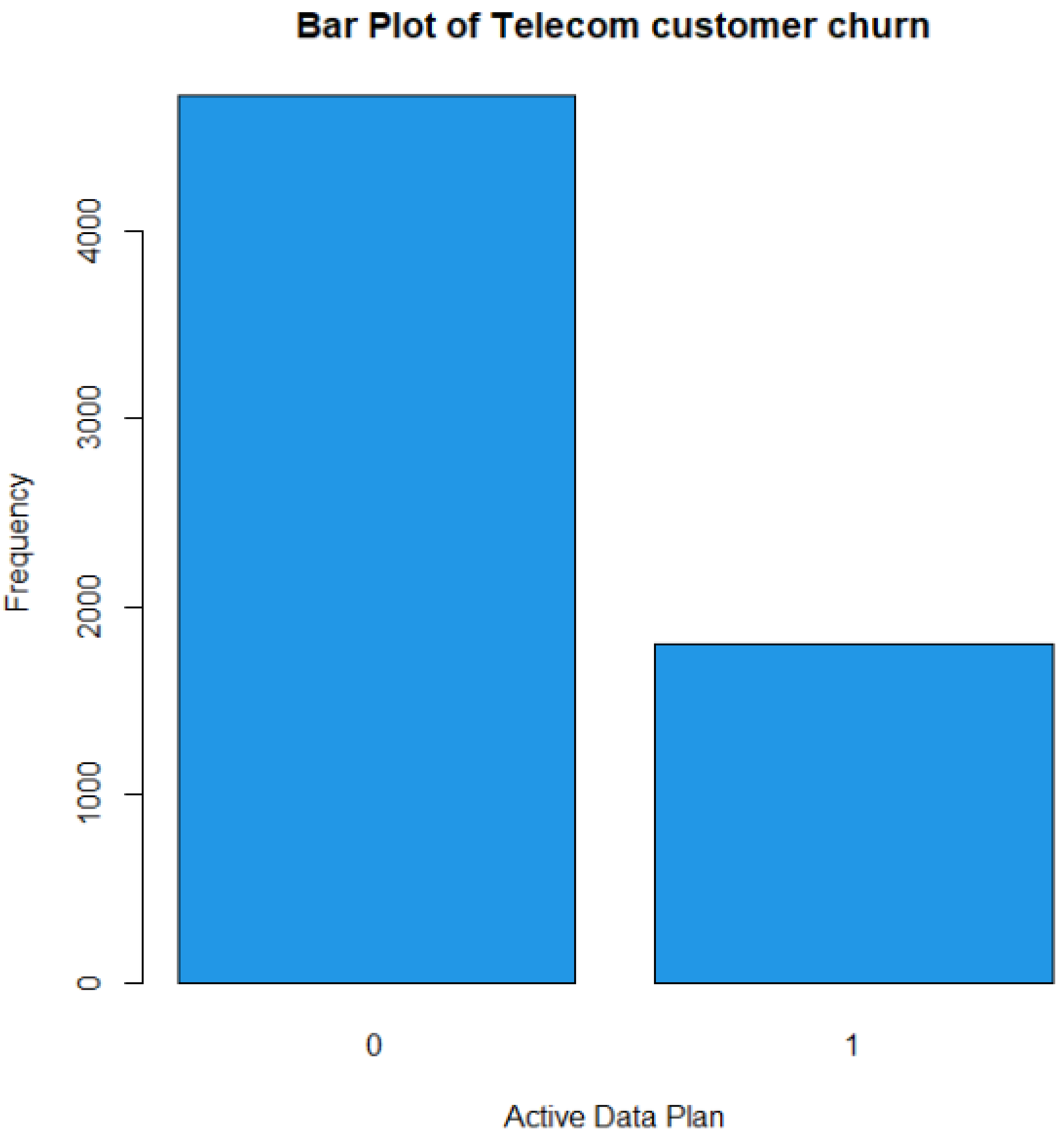


The density package was used to generate the output. The output shows a high frequency of customers below 1GB in data usage



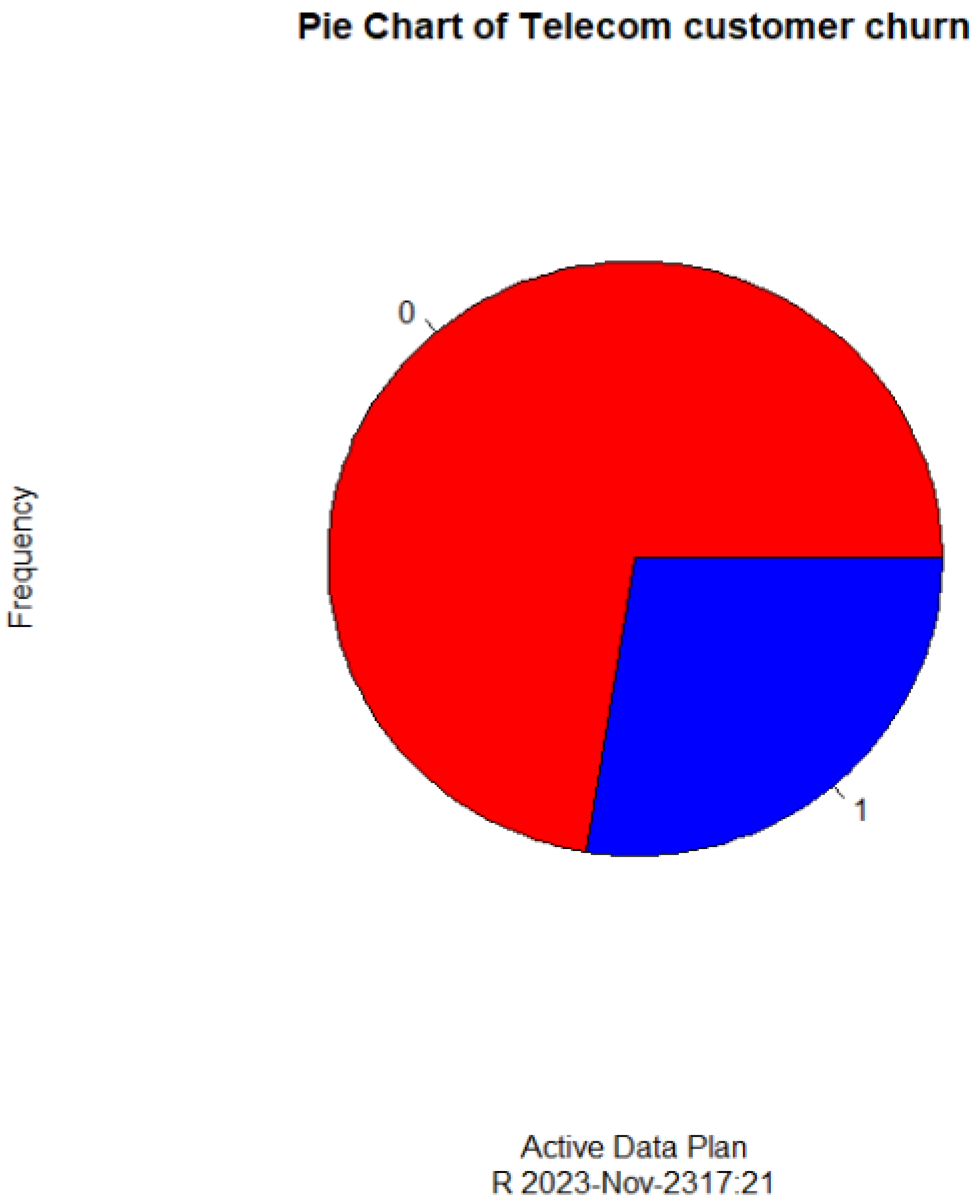
The output shows a normal distribution for Data usage.

Bar Plot



R 2023-Nov-2317:03

Pie Chart



The shows output large portion of customers are at the risk of churning out of the network.

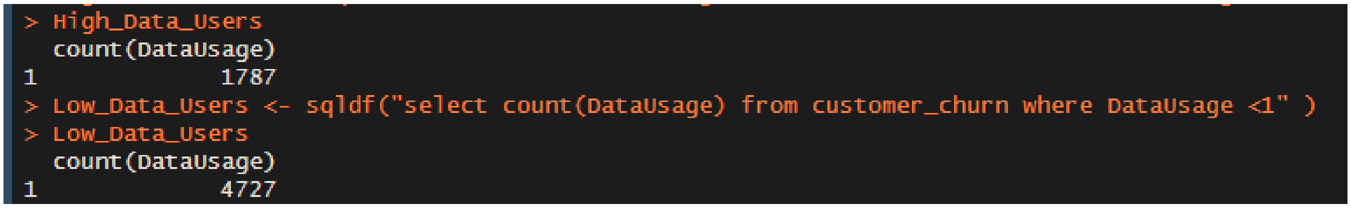
- The density and histogram chart reveals data consumption below **1GB** has the highest number and the histogram shows a normal distribution in data consumption with the mean around **2.7GB**.
- The data in the Bar plot shows the **distribution between active subscribers and inactive subscriptions**. **1 means customers with active data plans while 0 indicates customers without a data plan**. The data shows huge disparity between customers without data plan to customers with active data plan indicating

that more customer might churn out the network if measures were not put in place to reduce the churn rate.

To find the number of between high and low data users, we need to figure out the count of customers in each category and this achieved using some data manipulation techniques using the SQLDF package in R.

The code below was used to achieve this:

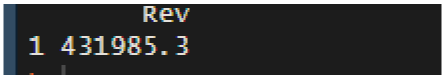
```
High_Data_Users <- sqldf("select count(DataUsage) from Customer_churn where DataUsa
ge >=1" )
Low_Data_Users <- sqldf("select count(DataUsage) from Customer_churn where DataUsag
e <1" )
```



```
> High_Data_Users
  count(DataUsage)
1             1787
> Low_Data_Users <- sqldf("select count(DataUsage) from customer_churn where DataUsage <1" )
> Low_Data_Users
  count(DataUsage)
1             4727
```

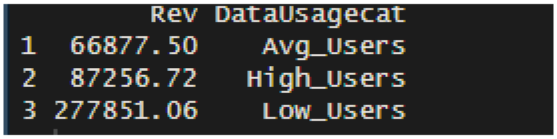
The output shows **72.6%** of customers with low data usage against **27.4%** with high data consumption, indicating a high probability of customers churning out of the network. Now let's take a look at the revenue categorization of customers in the data set.

```
##### Total Revenue ###
Rev1 <- sqldf("select sum(MonthlyCharge + OverageFee) As Rev from Customer_churn" )
```



```
       Rev
1 431985.3
```

```
##### Revenue breakdown by DataUsageCategory ###
Rev <- sqldf("select sum(MonthlyCharge + OverageFee) As Rev,DataUsagecat from Custo
mer_churn group by DataUsagecat")
```



```
        Rev DataUsagecat
1  66877.50    Avg_Users
2  87256.72   High_Users
3 277851.06    Low_Users
```

The output shows the Low data users generate **64.2%** of the revenue while High data users generate **20.1%** and Average data users generate **15.5%** of revenue for the company. From the output of this query there is a clear indication that a lot has to be done in improving the revenue generation of the company. When data usage grows monthly charge grows as well, meaning a lot of efforts for strategic sales conversion has to be made, cross-selling for most of the customers into buying higher data plans, there by increase the probability for customer loyalty rather than customer churn

Phase III

Data Insights and Recommendations

- The data shows a strong correlation between **data plan and data usage** which meanings that as **data subscription grows data usage grows as well**.
- The data also shows a **weak correlation between churn and data usage which means that customers who frequently buy data plans are not likely to churn out of the network.**
- The data also shows there is a weak correlation between Contract Renewal and churn, which indicates **that customers who renew their contract are not likely to churn out of the network**.
- Large revenue share from low data users which indicates a risk in as many customers are at the verge of churning out of the network.

Recommendations

Remember that the telecom industry is evolving rapidly, and staying competitive and meeting customer expectations is an ongoing challenge. By focusing on exceptional customer service, transparent pricing, and personalization, you can reduce churn and build long-lasting customer relationships. Below are some of the recommendations to help reduce customer churn and improve customer loyalty.

Retention Offers: Identify high-risk customers and provide targeted retention offers, such as discounts or additional services, to encourage them to stay.

Bundling Services: Offer bundles of telecom services (e.g., TV, internet, and mobile) to incentivize customers to stay with your company.

Loyalty Programs: Develop customer loyalty programs that reward long-term customers with exclusive benefits and discounts.

Quality Network and Service: Ensure the quality and reliability of your network and services. Invest in infrastructure to reduce dropped calls, slow data speeds, and service interruptions.

Competitive Analysis: Stay up-to-date with the offerings and pricing of your competitors. Ensure that your plans and services remain competitive.

Innovative Services: Stay at the forefront of technology by offering innovative services such as 5G, IoT solutions, and advanced security features.

Customer Retention Teams: Establish specialized customer retention teams to identify and address at-risk customers and help resolve their issues.

CODES

```
Customer_churn <- read.csv("C:/Users/onyif/Downloads/telecom_churn.csv")
Customer_churn

#####Correlation plot of all features #####
cor(Customer_churn)
is.na(Customer_churn)
na.omit(Customer_churn)

##### Sub Categorization of the Data set #####
Customer_churn$DataUsagecat[Customer_churn$DataUsage <1] <- "Low_Users"
Customer_churn$DataUsagecat[Customer_churn$DataUsage >=1 & Customer_churn$DataUsage <=2.7] <-"Avg
_Users"
Customer_churn$DataUsagecat[Customer_churn$DataUsage >2.7] <-"High_Users"


##### produce one correlation plot #####
Project_cor <- cor(Customer_churn,method="pearson")
corrplot(Project_cor,method="circle",mar=c(1,1,0,1))

#### produce one density and histogram plots together####
usage_den1 <- density(Customer_churn$DataUsage) plot(usage_den,col=4,main="Plot for
density",sub=paste("R",format(Sys.time(),"%Y-%b-%d%H:%S")))

hist(Customer_churn$DataUsage,main="Data Usage",prob=T,xlab="Usage in GB",ylab="Fre
quency",sub=paste("R",format(Sys.time(),"%Y-%b-%d%H:%S")),col=4)
lines(usage_den1,col=5)

#### produce pie plot/bar plot and a histogram based on the features of your choice
in your dataset ###
T=table(Customer_churn$DataPlan)
T
barplot(T,main="Bar Plot of Telecom customer churn",xlab="Active Data Plan",ylab="F
requency",sub=paste("R",format(Sys.time(),"%Y-%b-%d%H:%S")),col=4)

#### pie chart ###
T=table(Customer_churn$DataPlan)
T
colors<-c("red","blue")
```

```
pie(T,main="Pie Chart of Telecom customer churn",xlab="Active Data Plan",ylab="Freq
uency",sub=paste("R",format(Sys.time(),"%Y-%b-%d%H:%S")),col=colors)

#### Histogram Chart ####

hist(Customer_churn$DataUsage,main="Histogram of Data Usage",xlab="Data Usage in GB
",ylab="Frequency",sub=paste("R",format(Sys.time(),"%Y-%b-%d%H:%S")),col=4)

##### Total Revenue ###
Rev1 <- sqldf("select sum(MonthlyCharge + OverageFee) As Rev from Customer_churn" )

##### Revenue breakdown by DataUsageCategory ###
Rev <- sqldf("select sum(MonthlyCharge + OverageFee) As Rev,DataUsagecat from Custo
mer_churn group by DataUsagecat")
```

References

https://www.kaggle.com/datasets/barun2104/telecom-churn/data

R in Action, R. Kabacoff, 2nd edition, Manning, ISBN 978-1-617-29138-8, Chapter 4.

R in Action, R. Kabacoff, 2nd edition, Manning, ISBN 978-1-617-29138-8, Chapter 1 & 2.

R in Action, R. Kabacoff, 2nd edition, Manning, ISBN 978-1-617-29138-8, Chapter 4.