ML_Assignment_1

Setup

```
#Calling packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.1
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.1
```

```
library(modeest)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.4.1

## Registered S3 method overwritten by 'psych':
##   method         from
##   plot.residuals rmutil

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.1
```

```
## Loading required package: lattice
```

Data Import

```r
#Importing Churn Data

mobileChurnDF = read_csv("C:\\Users\\Urza\\OneDrive\\BA 64060 Fundementals of Machine Learning\\Module 1
```

```
## Rows: 1000 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ContractType
## dbl (8): CallFailures, SubscriptionLength, DataUsage, VoiceMinutes, Customer...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 1

```r
#Dimensions
dim(mobileChurnDF)
```

```
## [1] 1000    9
```

```r
#First six rows by call failures
mobileChurnDF %>% arrange(desc(CallFailures)) %>% head()
```

```
## # A tibble: 6 x 9
##   CallFailures SubscriptionLength DataUsage VoiceMinutes CustomerSupportCalls
##          <dbl>              <dbl>     <dbl>        <dbl>                 <dbl>
## 1           20                 21    5.95          2089.                    4
## 2           20                 24    0.603         3451.                    1
## 3           20                  1    9.42          1353.                    5
## 4           20                  1    5.70           421.                    3
## 5           20                 10    2.01          3014.                    1
## 6           20                 13    0.0926        3925.                    4
## # i 4 more variables: ContractType <chr>, MonthlyCharges <dbl>,
## #   RoamingUsage <dbl>, Churn <dbl>
```

```r
#Last six rows by call failures
mobileChurnDF %>% arrange(desc(CallFailures)) %>% tail()
```

```
## # A tibble: 6 x 9
##   CallFailures SubscriptionLength DataUsage VoiceMinutes CustomerSupportCalls
##          <dbl>              <dbl>     <dbl>        <dbl>                <dbl>
## 1            0                 23     9.71          179.                    5
## 2            0                  9     3.49          604.                    0
## 3            0                  1     8.41         2077.                    5
## 4            0                 10     0.934         115.                    4
## 5            0                  2     7.01         3142.                    0
## 6            0                 15     0.690        3889.                    5
## # i 4 more variables: ContractType <chr>, MonthlyCharges <dbl>,
## #   RoamingUsage <dbl>, Churn <dbl>
```

```
#Looking at the structure of the data
str(mobileChurnDF)
```

```
## spc_tbl_ [1,000 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ CallFailures        : num [1:1000] 16 4 0 9 3 17 16 14 6 3 ...
##  $ SubscriptionLength  : num [1:1000] 11 9 8 9 8 3 8 10 10 2 ...
##  $ DataUsage           : num [1:1000] 4.194 8.409 0.654 8.833 7.246 ...
##  $ VoiceMinutes        : num [1:1000] 4836 1695 4384 2610 2890 ...
##  $ CustomerSupportCalls: num [1:1000] 2 5 3 0 3 1 1 4 3 1 ...
##  $ ContractType        : chr [1:1000] "Monthly" "Monthly" "Monthly" "Monthly" ...
##  $ MonthlyCharges      : num [1:1000] 24.3 82.5 52.9 32.3 58.2 ...
##  $ RoamingUsage        : num [1:1000] 2.6 5.28 3.17 3.03 8.91 ...
##  $ Churn               : num [1:1000] 0 1 0 0 0 1 1 1 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   CallFailures = col_double(),
##   ..   SubscriptionLength = col_double(),
##   ..   DataUsage = col_double(),
##   ..   VoiceMinutes = col_double(),
##   ..   CustomerSupportCalls = col_double(),
##   ..   ContractType = col_character(),
##   ..   MonthlyCharges = col_double(),
##   ..   RoamingUsage = col_double(),
##   ..   Churn = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#Checking for missing values
colMeans(is.na(mobileChurnDF))
```

```
##         CallFailures   SubscriptionLength            DataUsage
##                    0                    0                    0
##         VoiceMinutes CustomerSupportCalls         ContractType
##                    0                    0                    0
##       MonthlyCharges         RoamingUsage                Churn
##                    0                    0                    0
```

```
#Summary
summary(mobileChurnDF)
```

```
##   CallFailures    SubscriptionLength   DataUsage        VoiceMinutes
```

```
##  Min.   : 0.000   Min.   : 1.00      Min.   :0.04479   Min.   :   1.913
##  1st Qu.: 5.000   1st Qu.: 6.00      1st Qu.:2.45883   1st Qu.:1392.148
##  Median :10.000   Median :12.00      Median :5.07325   Median :2626.685
##  Mean   : 9.985   Mean   :12.09      Mean   :5.09635   Mean   :2564.964
##  3rd Qu.:16.000   3rd Qu.:18.00      3rd Qu.:7.82260   3rd Qu.:3712.721
##  Max.   :20.000   Max.   :24.00      Max.   :9.99831   Max.   :4998.703
##  CustomerSupportCalls ContractType       MonthlyCharges   RoamingUsage
##  Min.   :0.000        Length:1000        Min.   :20.07   Min.   :0.01345
##  1st Qu.:1.000        Class :character   1st Qu.:37.57   1st Qu.:2.32212
##  Median :2.000        Mode  :character   Median :56.91   Median :4.94221
##  Mean   :2.394                           Mean   :58.42   Mean   :4.95070
##  3rd Qu.:4.000                           3rd Qu.:77.45   3rd Qu.:7.44860
##  Max.   :5.000                           Max.   :99.96   Max.   :9.99680
##      Churn
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :1.000
##  Mean   :0.504
##  3rd Qu.:1.000
##  Max.   :1.000
```

```
#Mode CallFailures
mlv(mobileChurnDF$CallFailures, method = 'mfv')
```

```
## [1] 9
```

```
#Mode SubscriptionLength
mlv(mobileChurnDF$SubscriptionLength, method = 'mfv')
```

```
## [1] 12
```

```
#Mode DataUsage No Mode
head(mlv(mobileChurnDF$DataUsage, method = 'mfv'))
```

```
## [1] 0.04479119 0.04739877 0.04740768 0.07698255 0.08304643 0.09255755
```

```
#Mode VoiceMinutes No mode
head(mlv(mobileChurnDF$VoiceMinutes, method = 'mfv'))
```

```
## [1]  1.912871 11.692593 14.096979 20.760797 21.805209 32.058203
```

```
#Mode CustomerSupportCalls
mlv(mobileChurnDF$CustomerSupportCalls, method = 'mfv')
```

```
## [1] 0
```

```
#Mode MonthlyCharges No mode
head(mlv(mobileChurnDF$MonthlyCharges, method = 'mfv'))
```

```
## [1] 20.06781 20.07184 20.21478 20.40233 20.45981 20.46626
```

```r
#Mode RoamingUsage No mode
head(mlv(mobileChurnDF$RoamingUsage, method = 'mfv'))
```

```
## [1] 0.01345361 0.01458429 0.03109833 0.03279057 0.04421746 0.06000759
```

```r
#Mode Churn
mlv(mobileChurnDF$Churn, method = 'mfv')
```

```
## [1] 1
```

```r
#Mode or value repeated the most for character
mlv(mobileChurnDF$ContractType, method = 'mfv')
```

```
## [1] "Monthly"
```

```r
#Additional discriptions of data
describe(mobileChurnDF)
```

```
##                       vars    n    mean      sd  median trimmed     mad   min
## CallFailures             1 1000    9.98    6.12   10.00   10.00    7.41  0.00
## SubscriptionLength       2 1000   12.09    6.85   12.00   12.04    8.90  1.00
## DataUsage                3 1000    5.10    2.96    5.07    5.11    3.97  0.04
## VoiceMinutes             4 1000 2564.96 1411.60 2626.69 2577.77 1760.39  1.91
## CustomerSupportCalls     5 1000    2.39    1.72    2.00    2.37    2.97  0.00
## ContractType*            6 1000    1.51    0.50    2.00    1.51    0.00  1.00
## MonthlyCharges           7 1000   58.42   23.07   56.91   58.09   29.64 20.07
## RoamingUsage             8 1000    4.95    2.89    4.94    4.95    3.80  0.01
## Churn                    9 1000    0.50    0.50    1.00    0.50    0.00  0.00
##                         max   range  skew kurtosis    se
## CallFailures          20.00   20.00 -0.01    -1.22  0.19
## SubscriptionLength    24.00   23.00  0.03    -1.15  0.22
## DataUsage             10.00    9.95 -0.02    -1.29  0.09
## VoiceMinutes        4998.70 4996.79 -0.07    -1.14 44.64
## CustomerSupportCalls   5.00    5.00  0.06    -1.29  0.05
## ContractType*          2.00    1.00 -0.03    -2.00  0.02
## MonthlyCharges        99.96   79.89  0.08    -1.19  0.73
## RoamingUsage          10.00    9.98 -0.01    -1.22  0.09
## Churn                  1.00    1.00 -0.02    -2.00  0.02
```

Question 2

The first descriptive statistic looked at was the dimension of the data set. It has 1000 rows and 9 variables, the number of variables corresponds with the description given to us. Next I looked at the head (top six) and tail (bottom six) rows of the data set after it had been sorted in descending order based on the number of call failures. Looking over these tables the Call failures seems to correlate with the Churn and would warrant additional investigation. Next, was the structure of the data which showed that all but one variable is numeric. The one character variable is Contract type. This categorical variable would need to be transformed later into a factor or as dummy variables if we are interested in using it in a regression model. Next, missing values were checked using colMeans and is.na, the result of this showed that there was no NA values in the data set. Looking at the summary table we have the mean, median, min, max, 1st and 3rd

quartile. From the summary it shows that most variable means and medians are very close to each other, except for the Churn which has a mean of .504 which indicates that the number of customers that leave and the customers that stayed are just about 50/50 in this data set. Another observation is the large values for VoiceMinutes and the large range between min and max which may need to be investigated further. However, it is not uncommon for people to use not use their phone for calls and only text/data. The mode was determined for each variable which showed the most frequent value, an observation from the modes shows that DataUsage, VoiceMinutes, MonthlyCharges, and RoamingUsage do not have a mode because all values are unique. Note: For clarity in the document I used head to show only the first six values instead of all 1000 for each no mode variable. The last "mode" determined was for the categorical value which indicated that the most used type of contract is Monthly. The mode for Churn was 1 which idicates that more customers are leaving the mobile company then staying. Finally the description table was found for the data which provided additional information from summary, this information included standard deviation and skew. The VoiceMinutes has the largest standard deviation indicating data is spread out from the mean. All variables have minimal skew indicating normal distributions.

Question 3

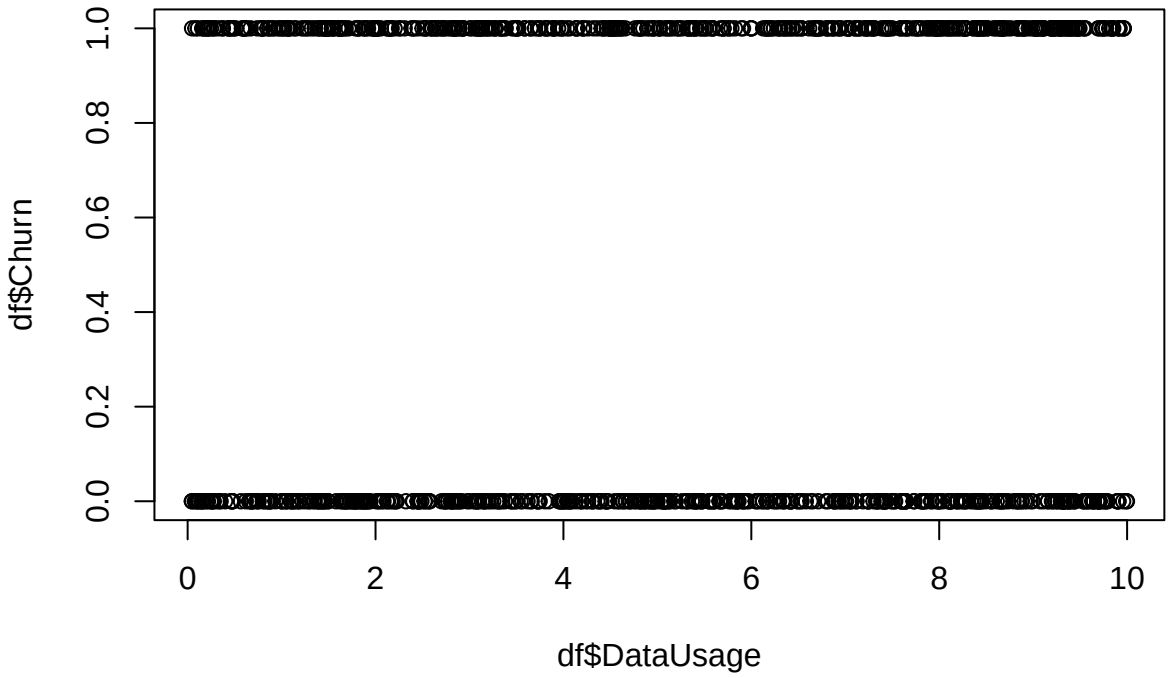
```
#Creation of a new table using select
df = mobileChurnDF %>% select(DataUsage, Churn)

#Description of the Data
describe(df)
```

```
##           vars    n mean   sd median trimmed  mad  min max range  skew kurtosis
## DataUsage    1 1000  5.1 2.96   5.07    5.11 3.97 0.04  10  9.95 -0.02    -1.29
## Churn        2 1000  0.5 0.50   1.00    0.50 0.00 0.00   1  1.00 -0.02    -2.00
##             se
## DataUsage 0.09
## Churn     0.02
```

```
#Frequency table showing how many of each data amount have remained a customer or left.
table(round(df$DataUsage, digits = 0), df$Churn, dnn = c("Data", "Churn"))
```

```
##     Churn
## Data  0  1
##    0  27 22
##    1  51 51
##    2  51 51
##    3  48 56
##    4  43 41
##    5  54 49
##    6  46 40
##    7  42 47
##    8  57 56
##    9  48 70
##    10 29 21
```

```
#Plot of Churn Vs DataUsage
plot(df$DataUsage, df$Churn)
```




```
#Correlation
cor(df$DataUsage, df$Churn)
```

```
## [1] 0.0190528
```

Question 3 Insights

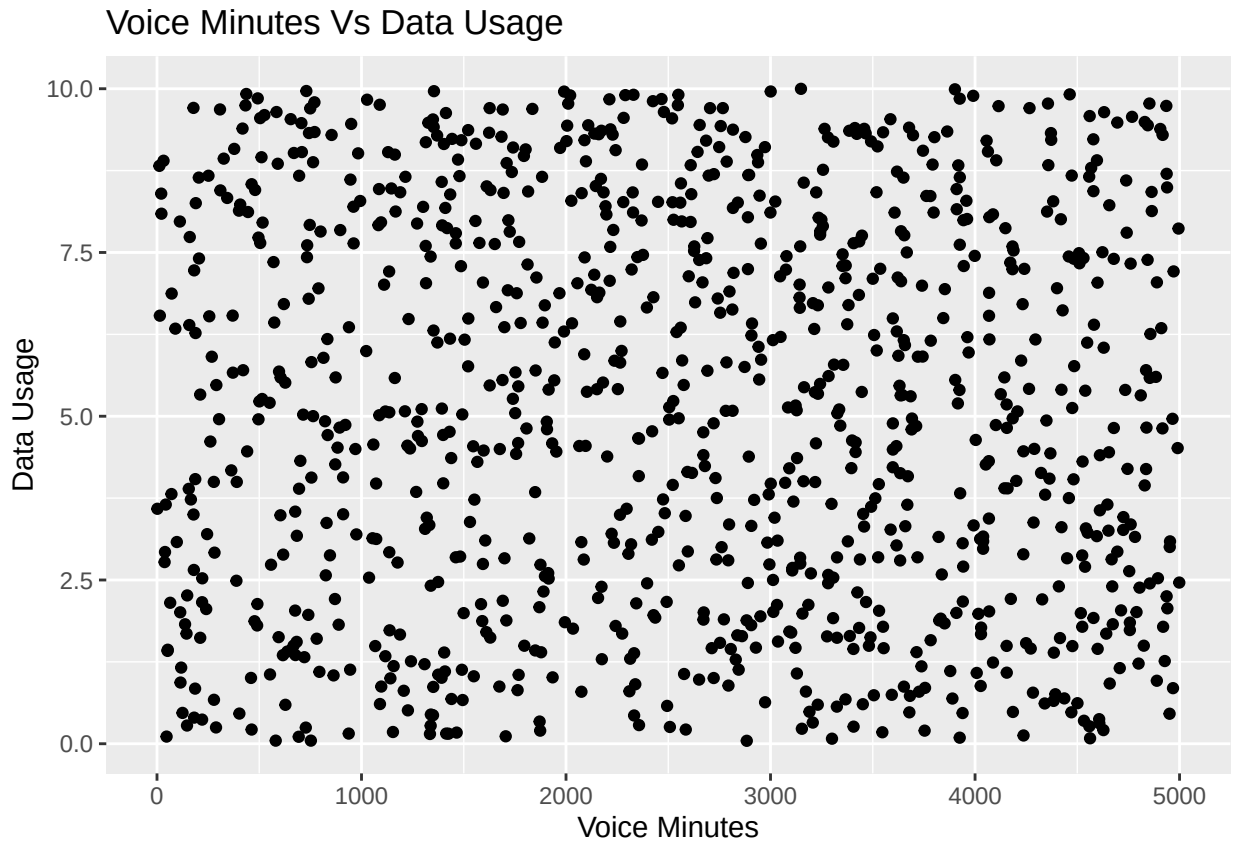
There does not appear to be any correlation between Churn and Data Usage. Looking at the frequency table which has the data usage amount rounded to the nearest integer, it appears that a similar amount of customers stay or leave the mobile provider no matter what the level of data usage is. The plot of Churn Vs DataUsage shows the un-rounded amounts of data usage and there does not appear to be any relationship. Finally, I calculated the correlation which came out to 0.01 which indicates very small to no correlation between the two variables.

Question 4

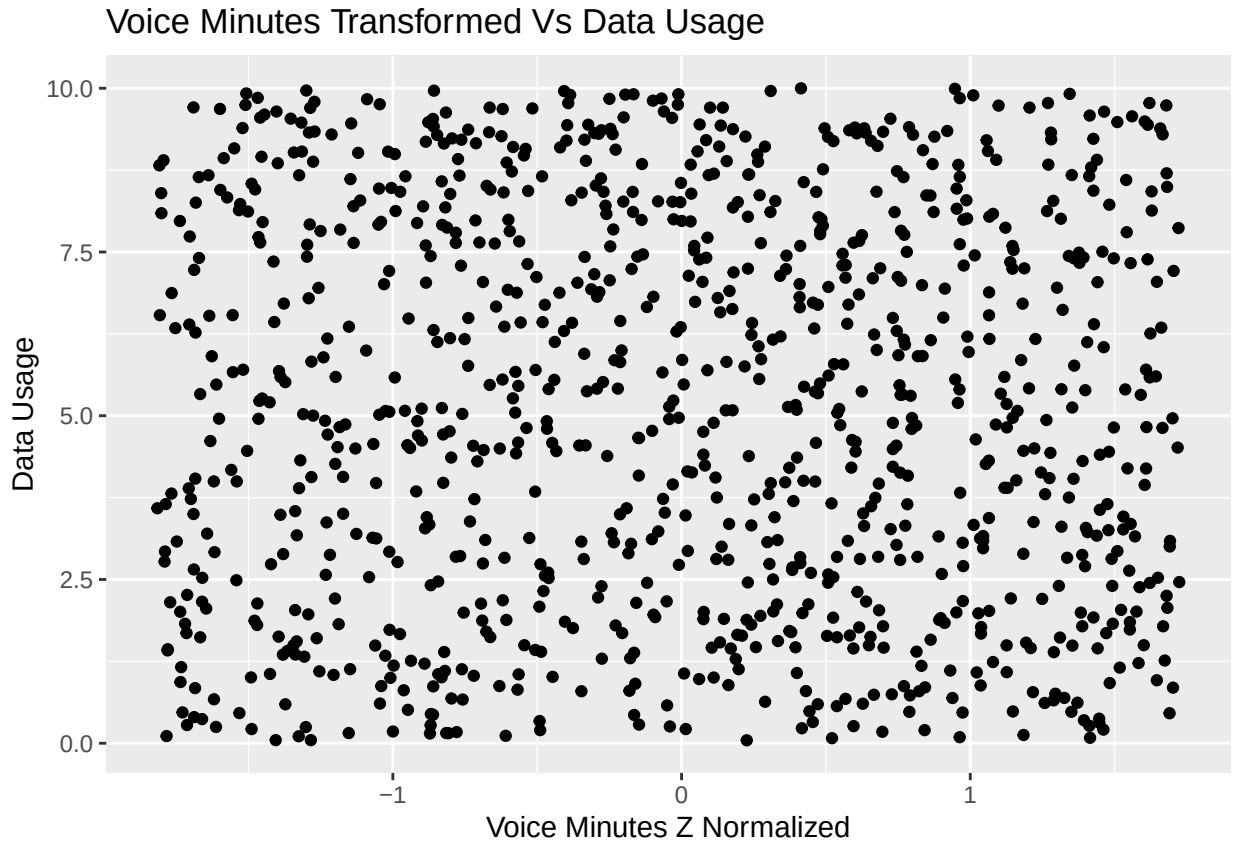
```
#Z-score transformation

mobileChurnDF$VoiceMinutes_Z = scale(mobileChurnDF$VoiceMinutes)
summary(mobileChurnDF$VoiceMinutes_Z)
```

```
##        V1
##  Min.   :-1.81571
##  1st Qu.:-0.83084
##  Median : 0.04372
##  Mean   : 0.00000
##  3rd Qu.: 0.81309
##  Max.   : 1.72410
```

```
ggplot(mobileChurnDF) + geom_point(aes(x = mobileChurnDF$VoiceMinutes, y = mobileChurnDF$DataUsage)) + g
```




```
ggplot(mobileChurnDF) + geom_point(aes(x = mobileChurnDF$VoiceMinutes_Z, y = mobileChurnDF$DataUsage)) +
```



Question 4 Insights

With out knowing what this data will be used for I decided to transform the VoiceMinutes Variable in order to decrease it's range. This happens by transforming the data to have a mean of 0 and a standard deviation of 1. Voice Minutes ranges from 1.91 Minutes to 4998.7 Minutes. This is a wide range as well as much larger numbers when compared to the rest of the variables. Changing the scale did not change the distribution of the data as can be seen in the two charts. The first is using the original VoiceMinutes while the second chart uses the transformed VoiceMinutes_Z.