Churn

Manish Grewal

29/07/2020

## Load the libraries

library(ggplot2)  
library(dplyr)  
library(tidyr)

## Set working directory

setwd("C:/Users/manish.grewal/emdp/R/churn")

## Load the dataset

churn <- read.csv("churn.csv")

# Use the churn data set, and complete the following exercises:

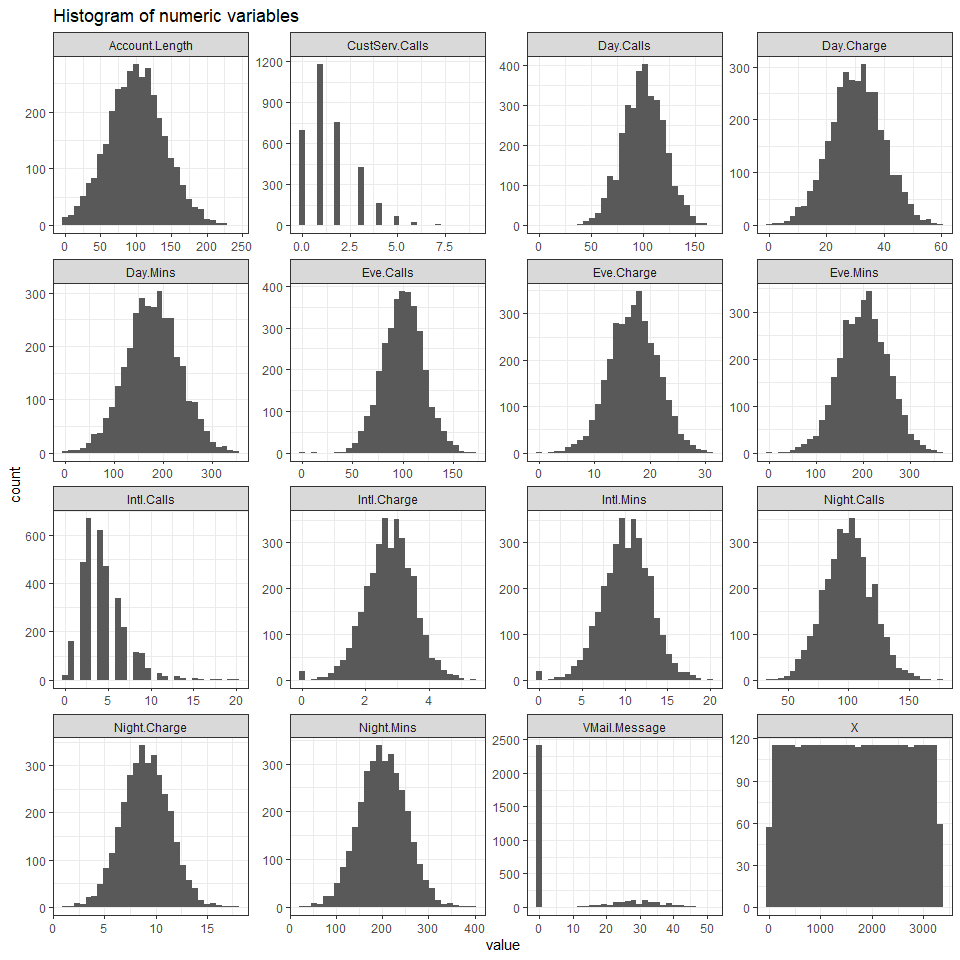
## 1. Explore whether there are missing values for any of the variables.

any( is.na( churn ) )

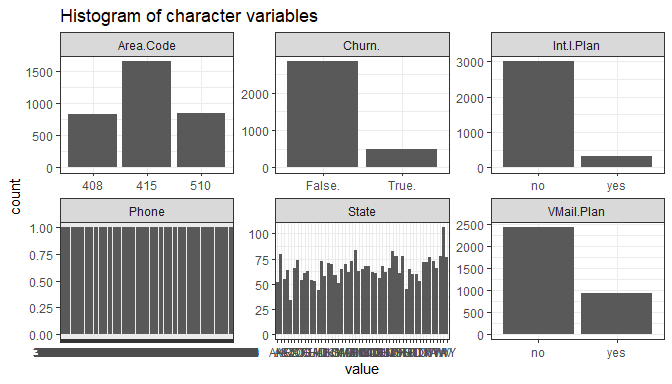
## [1] FALSE

There are no missing values. summary(churn) also does not show any NA values.

churn$Area.Code = as.character(churn$Area.Code)  
churn %>%   
 select\_if(is.numeric) %>%   
 gather() %>%   
 ggplot(aes(value)) +  
 facet\_wrap(~ key, scales = "free") +  
 geom\_histogram() +  
 labs(title = "Histogram of numeric variables") +  
 theme\_bw()



churn %>%   
 select\_if(is.character) %>%   
 gather() %>%   
 ggplot(aes(value)) +  
 facet\_wrap(~ key, scales = "free") +   
 geom\_histogram(stat = "count") +  
 labs(title = "Histogram of character variables") +  
 theme\_bw()

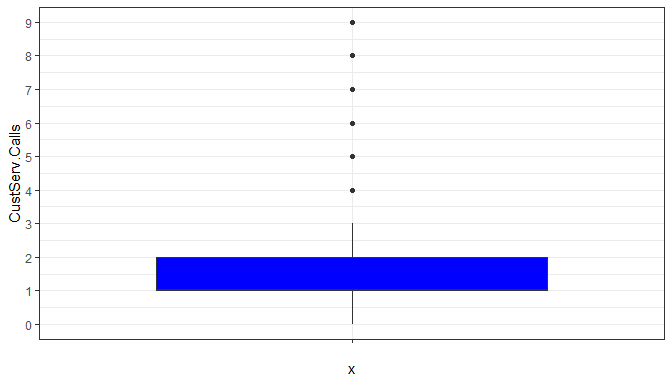


Histogram of Area.Code shows only 3 different values, which may require further exploration as the same 3 area codes are appearing for different states.

VMail.Message also shows a high number of records with 0 value.

## 2. Use a graph to visually determine whether there are any outliers among the number of calls to customer service.

ggplot(churn) +  
 aes(x = "", y = CustServ.Calls) +  
 geom\_boxplot(fill = "blue") +  
 scale\_y\_continuous(n.breaks = 10) +  
 theme\_bw()



The box plot shows the values 4 to 9 are outliers among the number of calls to customer service.

## 3. Identify the range of customer service calls that should be considered outliers, using:

### a. the Z-score method;

churn$CustServ.Calls\_zscore <- (churn$CustServ.Calls - mean(churn$CustServ.Calls)) / sd(churn$CustServ.Calls)   
  
outliers <- churn %>%  
 filter(CustServ.Calls\_zscore < -3 |  
 CustServ.Calls\_zscore > 3) %>%   
 select(CustServ.Calls, CustServ.Calls\_zscore)  
  
sort(unique(outliers$CustServ.Calls))

## [1] 6 7 8 9

Outliers using the z-score method are the customers who made the following number of calls:  
**6, 7, 8, 9**

### b. the IQR method

summary(churn$CustServ.Calls)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 1.000 1.563 2.000 9.000

Q1 <- as.numeric(quantile(churn$CustServ.Calls, 0.25)); Q1

## [1] 1

Q3 <- as.numeric(quantile(churn$CustServ.Calls, 0.75)); Q3

## [1] 2

IQR <- Q3 - Q1; IQR

## [1] 1

outliers <- churn %>%  
 filter(CustServ.Calls < Q1 - 1.5\*IQR | CustServ.Calls > Q3 + 1.5\*IQR) %>%  
 select(CustServ.Calls)  
sort(unique(outliers$CustServ.Calls))

## [1] 4 5 6 7 8 9

Outliers using the IQR method are the customers who made the following number of calls:  
**4, 5, 6, 7, 8, 9**

## 4. Transform the day minutes attribute using Z-score standardization.

churn$Day.Mins\_zscore <- (churn$Day.Mins - mean(churn$Day.Mins)) / sd(churn$Day.Mins)   
head(churn$Day.Mins\_zscore)

## [1] 1.5665319 -0.3336877 1.1681284 2.1962665 -0.2400537 0.8009362

## 5. Work with skewness as follows:

### a. Calculate the skewness of day minutes.

skew <- 3 \* (mean(churn$Day.Mins) - median(churn$Day.Mins)) / sd(churn$Day.Mins)  
skew

## [1] 0.02065993

### b. Then calculate the skewness of the Z-score standardized day minutes. Comment.

skew\_z <- 3 \* (mean(churn$Day.Mins\_zscore) - median(churn$Day.Mins\_zscore)) / sd(churn$Day.Mins\_zscore)  
skew\_z

## [1] 0.02065993

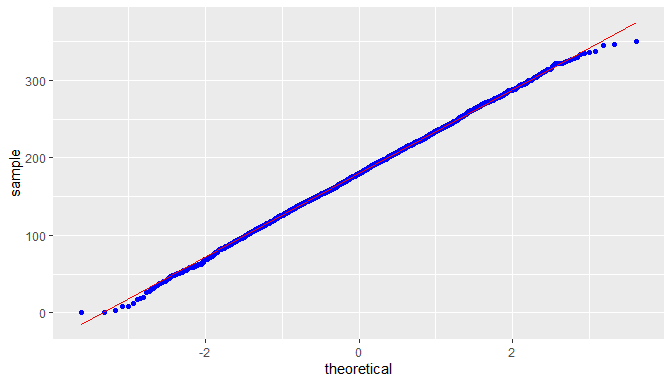
The skewness remains the same after z-score standardization.

### c. Based on the skewness value, would you consider day minutes to be skewed or nearly perfectly symmetric?

The skewness of Day.Mins (0.0206599) is close to zero but positive. Therefore, day minutes is nearly perfectly symmetric with a very slight right skewness.

## 6. Construct a normal probability plot of day minutes. Comment on the normality of the data.

ggplot(churn, aes(sample = Day.Mins)) +  
 stat\_qq(col = "blue") +  
 stat\_qq\_line(col = "red")

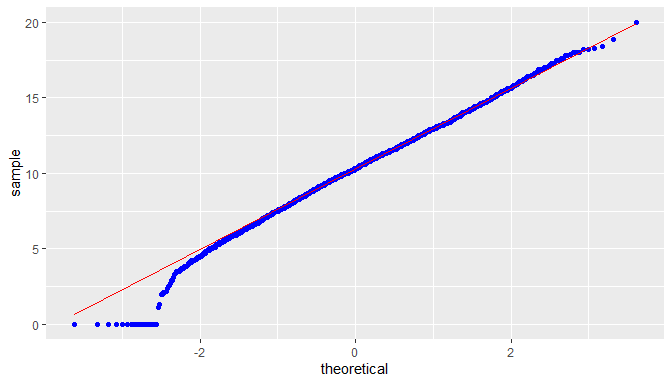


As the sample matches the normal line very closely, the data is nearly perfectly normal.

## 7. Work with international minutes as follows:

### a. Construct a normal probability plot of international minutes.

ggplot(churn, aes(sample = Intl.Mins)) +  
 stat\_qq(col = "blue") +  
 stat\_qq\_line(col = "red")



The normal probability curve indicates left skewness. This is confirmed from the summary of Intl.Mins where the mean (10.24) is a less than the median (10.3). The skewness calculation returns a negative value (-0.0673817) and confirms the interpretation of left skewness.

summary(churn$Intl.Mins)

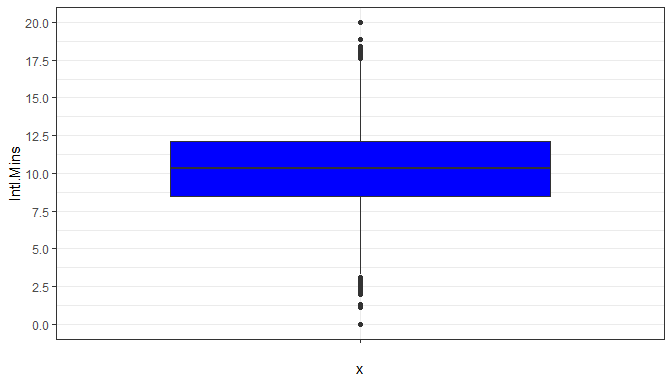
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 8.50 10.30 10.24 12.10 20.00

skew <- 3 \* (mean(churn$Intl.Mins) - median(churn$Intl.Mins)) / sd(churn$Intl.Mins)  
skew

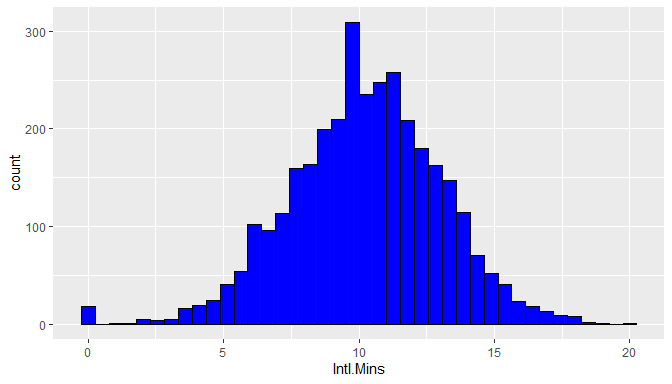
## [1] -0.06738167

### b. What is preventing this variable from being normally distributed.

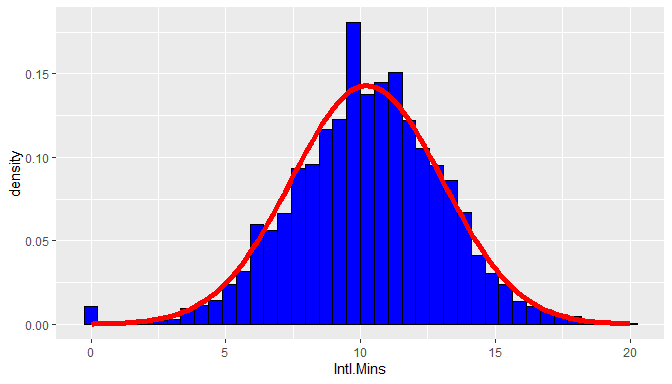
ggplot(churn) +  
 aes(x = "", y = Intl.Mins) +  
 geom\_boxplot(fill = "blue") +  
 scale\_y\_continuous(n.breaks = 10) +  
 theme\_bw()



ggplot(churn, aes(Intl.Mins)) +  
 geom\_histogram(bins = 40, color = "black", fill = "blue")



ggplot(churn, aes(Intl.Mins)) +  
 geom\_histogram(aes(y = ..density..), bins = 40, color = "black", fill = "blue") +  
 stat\_function(fun = dnorm, args = list(mean(churn$Intl.Mins), sd(churn$Intl.Mins)), color = "red", size = 2)



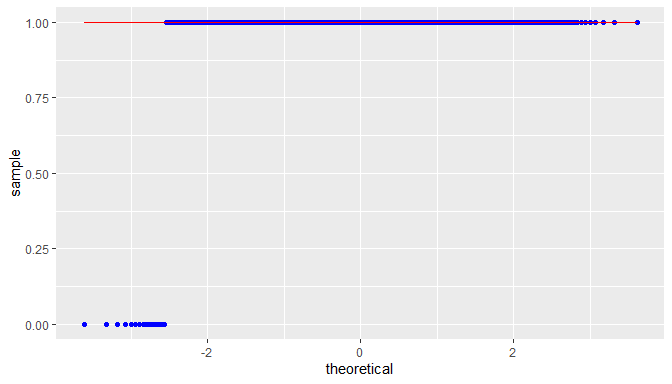
The boxplot shows outlier below 3 and above 17.5. However, the histogram shows around 20 outliers with value 0. This concentration of outliers with 0 value is preventing the data from being normally distributed.

### c. Construct a flag variable to deal with the situation in (b).

churn <- mutate(churn, Intl.Mins.NZ = ifelse(Intl.Mins == 0, 0, 1))

### d. Construct a normal probability plot of the derived variable nonzero international minutes.

ggplot(churn, aes(sample = Intl.Mins.NZ)) +  
 stat\_qq(col = "blue") +  
 stat\_qq\_line(col = "red")



### e. Comment on the normality of the derived variable.

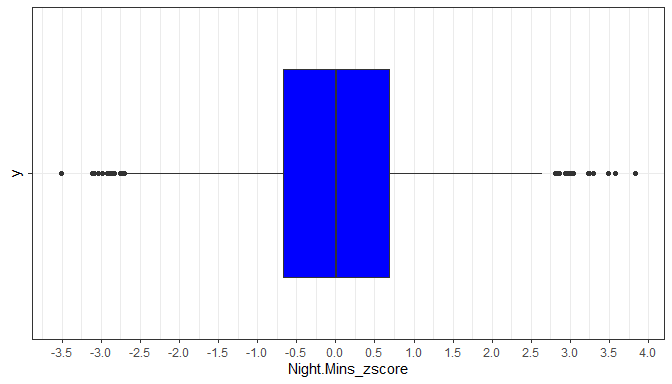
The derived variable shows the normal probability where its value is 1 and diverges where its value is zero. The zero values are making the distribution left skewed.

churn$Intl.Mins.NZ\_z <- scale(churn$Intl.Mins.NZ)  
skew <- 3 \* (mean(churn$Intl.Mins.NZ\_z) - median(churn$Intl.Mins.NZ\_z)) / sd(churn$Intl.Mins.NZ\_z)  
skew

## [1] -0.2210297

## 8. Transform the night minutes attribute using Z-score standardization. Using a graph, describe the range of the standardized values.

churn$Night.Mins\_zscore <- scale(churn$Night.Mins)  
  
ggplot(churn) +  
 aes(y = "", x = Night.Mins\_zscore) +  
 geom\_boxplot(fill = "blue") +  
 scale\_x\_continuous(n.breaks = 20) +  
 theme\_bw()



From the boxplot, we can see that

* the z-score has a median of 0
* Ist quartile = -0.7
* 3rd quartile = 0.7
* the IQR is Q3 - Q1 = 1.4
* the range is from -2.6 to 2.6 (excluding outliers)
* the positive outliers range from 2.8 to 3.8
* the negative outliers range from -3.5 to -2.6