Data visualization assignment

Michael Hornicek

November 10, 2022

Assignment

First load the data:

```
churn <- read.csv(file="./churn.txt", stringsAsFactors = FALSE)
```

33.

There are no missing values for any of the variables

```
na_count <- sapply(churn, function(value) sum(length(which(is.na(value)))))
na_count <- data.frame(na_count)
na_count
```

34.

There are only three unique values for area code (408, 415, 510), but there are 51 unique values for states. This is either completely impossible (if USA has unique area codes in the whole country) or at least incredibly unlikely (if USA has unique area codes per state).

```
table(churn$State)
length(unique(churn$State))

table(churn$Area.Code)
length(unique(churn$Area.Code))

barplot(table(churn$State), las=2, cex.names=0.5)
barplot(table(churn$Area.Code))
```

35.

To graphically identify outliers, a histogram can be used.

```
par(mfrow=c(1, 1))
hist(churn$CustServ.Calls,
     breaks=10,
     col="blue",
     border="black",
     xlab="Calls to customer service",
     ylab="counts",
     main="Histogram of customer service calls")
box(which="plot", lty="solid", col="black")
summary(churn$CustServ.Calls)
```

In the histogram, we can see that there are are two people who made 9 calls to customer service. This could be considered an outlier, since the distribution is very left-skewed and these data points are all the way on the right on the graph, and there is only a very small number of them, compared to the total number of records.

36.

IQR We can use the summary function to find the quartiles for the Customer calls variable:

```
summary(churn$CustServ.Calls)
```

This gives a Q1 of 1, and Q3 of 2. Therefore, the IQR is 1. Values are considered outliers if they are less than Q1 - 1.5IQR or more than Q3 + 1.5IQR. Because the number of customer calls cannot be negative, there are no outliers on the lower end. On the higher end, customers with 4 or more calls would be considered outliers.

Z-score

To find the range for outliers using the Z-score, we first need to get the mean and the standard deviation. The mean was given by the `summary` function above, it is 1.563. The standard deviation can be found using the `sd` function:

```
sd(churn$CustServ.calls)
```

The sd is 1.3155. Using the Z-score method, values are considered to be outliers if their Z-score is greater than 3 or less than -3. In other words, values which are further than 3 standard deviations away from the mean. For the lower end this would be 1.563 - 3(1.3155), which would give a negative value, just like the IQR method did. Therefore, there are again no outliers on the lower end. On the higher end, values greater than 1.563 + 3(1.3155) are considered to be outliers, so customers with 6 calls or more.

37.

```
zscore_day_mins <- ((churn$Day.Mins-mean(churn$Day.Mins))/sd(churn$Day.Mins))
zscore_day_mins
```

38.

a)

```
day_mins_skew <- ((3*(mean(churn$Day.Mins) - median(churn$Day.Mins)))/sd(churn$Day.Mins))
```

The skewness of *day minutes* is 0.2066

b)

```
zscore_day_mins_skew <- ((3*(mean(zscore_day_mins) - median(zscore_day_mins)))/sd(zscore_day_mins))
```

Where `zscore_day_mins` was obtained using the method in 37. The skewness of Z-score standardized *day minutes* is also 0.2066, the same as for regular *day minutes*. This is expected, Z-score standardization has no effect on skewness.

c)

I would expect *day minutes* to be nearly perfectly symmetric, because the skewness is very close to zero, so the mean and the median are very close together, which is the case for symmetric distributions.

39

```
par(mfrow = c(1, 1))
qqnorm(churn$Day.Mins,
       datax=TRUE,
       col="red",
```

```
       main="Normal Q-Q Plot of Day Minutes")
qqline(churn$Day.Mins,
      col="blue",
      datax=TRUE)
```

day minutes is almost perfectly normally distributed, as the majority of the points on the normal probability plot are very close to the straight line.

40

a)

```
par(mfrow = c(1, 1))
qqnorm(churn$Intl.Mins,
       datax=TRUE,
       col="red",
       main="Normal Q-Q Plot of International Minutes")
qqline(churn$Intl.Mins,
       col="blue",
       datax=TRUE)
```

b)

The distribution has a "fat left tail" - there are many more customers who have 0 international minutes than there would be if the data were normally distibuted. This can be confirmed using the `table()` function or a histogram of the variable.

c)

```
churn["has_international_minutes"] <- unlist(lapply(churn$Intl.Mins, function(x) as.numeric(x > 0)))
```

d)

```
churn["non_zero_international_minutes"] <-unlist(lapply(churn$Intl.Mins, function(x) if (x>0) x else NA)

par(mfrow = c(1, 1))
qqnorm(churn$non_zero_international_minutes,
       datax=TRUE,
       col="red",
       main="Normal Q-Q Plot of Non-zero International Minutes")
qqline(churn$Intl.Mins,
       col="blue",
       datax=TRUE)
```

The derived variable is much closer to a normal distribution than the original variable. By getting rid of the points near 0, there are now very few points which are not on the straight line of the normal probability plot.

41.

To transform using Z-score standardization:

```
zscore_night_mins <- ((churn$Night.Mins-mean(churn$Night.Mins))/sd(churn$Night.Mins))
zscore_night_mins
```

The range of standardized values can be seen from a histogram:

```
par(mfrow=c(1, 1))
hist(zscore_night_mins,
```

```
     breaks=10,
     col="blue",
     border="black",
     xlim = c(-6, 6),
     xlab="Z-standardized night call minutes",
     ylab="Customer count",
     main="Histogram of night call minutes")
box(which="plot", lty="solid", col="black")
```

All the values are within 4 standard deviations of the mean, as the Z-scores range from -4 to 4. If we use the summary function:

```
summary(zscore_night_mins)
```

We get that the actual maximum is 3.84 and the minimum is -3.51. This means that the variable includes data which can be considered outliers.

library(rmarkdown) render("assignment_hornicek.rmd")