DSC/CSC/TCS 462 Assignment 1

Due September 19, 2019 by 3:15pm

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?
- Who, if anyone, did you work with on this assignment?
- What questions do you have relating to any of the material we have covered so far in class?

For this assignment, you are expected to use the `ggplot2` library in `R` for completing all the graphics. The Chapter 2 and Chapter 3 R code files from class contain examples of using the `ggplot2` library for all our needs. To learn more about graphics using `ggplot2`, please read through the guide available here: http://www.cookbook-r.com/Graphs/. This is a wonderful open source textbook that walks through examples of many different graphics in `ggplot2`. If you have not done so already, start by installing the library. In the R console (i.e. NOT in your .RMD file), run the code `install.packages("ggplot2")`. Then, in your .RMD file, load the library as follows:

```
library(ggplot2)
```

1. Monitoring ocean conditions is an important task to ensure the safety of people in the sorrounding waters. Increased wave sizes can also be detrimental to beaches, causing for faster erosion of the shoreline. At the same time, large waves are perfect for surfing and increased tourism. Data collected on waves off the shores of Mooloolaba, Australia are contained in the dataset titled "waves.csv."
   a. Read the data into RStudio and summarize the data with the `summary()` function.
   b. Create a histogram of `Hmax`. Make sure to appropriately title the histogram and label the axes. Comment on the center, shape, and spread. Use a total of 13 bins.
   c. How many bins does Sturges' formula suggest we use for the histogram in part b? Show your work.
   d. Calculate the mean, median, and 20% trimmed mean of the maximum wave height. Report the mean, median, and 20% trimmed mean on the histogram. In particular, create a red vertical line on the histogram at the mean, and report the value of the mean in red next to the line using the form "$\bar{x} =$". Create a blue vertical line on the histogram at the median, and report the value of the median in blue next to the line using the form "$\tilde{x} =$". Create a green vertical line on the histogram at the 20% trimmed mean, and report the value of the 20% trimmed mean in green next to the line using the form "$\bar{x}_{20} =$" (to get $\bar{x}_{20}$ to print on the plot, use `bar(x)[20]` within the `paste()` function).
   e. Calculate and report the 25th and 75th percentiles.
   f. Calculate and report the interquartile range.

g. Calculate and report the standard span, the lower fence, and the upper fence.
h. Are there any outliers? Subset the outlying points. Use code based on the following:

```
waves[waves$Hmax >= upper_fence, ] #upper outliers
waves[waves$Hmax <= lower_fence, ] #lower outliers
# Use upper and lower fence values from part g.
```

i. Calculate and report the variance, standard deviation, and coefficient of variation of the maximum wave heigt.
j. We have seen from the histogram that the data are skewed. Calculate and report the skewness. Comment on this value and how it matches with what you visually see in the histogram.
k. Use a Box-Cox power transformation to appropriately transform the data. In particular, use the `boxcox()` function in the `MASS` library. Report the recommended transformation. Do not apply this transformation to the data yet. (Note: the `boxcox` function automatically produces a plot. You do NOT need to make this in `ggplot2`.)
l. Apply the exact Box-Cox recommended transformation (rounded to three decimal places) to the data (this transformation is hereon referred to as the Box-Cox transformed data). Use the `summary()` function to summarize the results of this transformation.
m. Create a histogram of the Box-Cox transformed data. On this histogram, report the mean, median, and 20% trimmed mean using the same formatting options as in part d above. Comment on the center shape and spread.
n. As an alternative to the Box-Cox transformation, let's also use a square root transformation. Apply the square root transformation to the original `Hmax` data (this transformation is hereon referred to as the square root transformed data). Use the `summary()` function to summarize the results of this transformation.
o. Create a histogram of the square root transformed data. On this histogram, report the mean, median, and 20% trimmed mean using the same formatting options as in part d above. Comment on the center shape and spread.
p. Create a qqplot for the original data, a qqplot for the Box-Cox transformed data, and a qqplot of the square root transformed data. Comment on the results.
q. Evaluate the empirical rule for the original data, the Box-Cox transformed data, and the square root trasnformed data. In particular, make a table similar to that on slide 94 of Chapter 2 notes. Comment on the results. Do either of the transformed data seem to be "better" to work with? Note, you can use code similar to the following to answer this question:

```
###Create a matrix named "mat" with 9 rows & 5 columns
mat <- matrix(NA, nrow=9, ncol=5)

### Set row names and column names
rownames(mat) <- c("Original","","","Box-Cox","","","Square Root","","")
colnames(mat) <- c("x","xbar-k*s", "xbar+k*s", "Theoretical %","Actual %")
```

```r
### Fill in known quantities
mat[,1] <- c(1,2,3)
mat[,4]<-c(68,95,99.7)

### Fill in calculated values
### (I only give a preview of this and leave the remaining calculations for you).
### I use "orig" as the original data, "bcdat" as the Box-Cox transformed data,
### and "srdat" as the square root-transformed data
### Name your variables anything you'd like.
mat[3,2] <- mean(orig)-3*sd(orig)
mat[2,3] <- mean(orig)+2*sd(orig)
mat[4,5] <- sum(bcdat >mean(bcdat)-1*sd(bcdat)
                & bcdat < mean(bcdat)+1*sd(bcdat))/length(bcdat)*100
mat[9,5] <- sum(srdat >mean(srdat)-3*sd(srdat)
                & srdat < mean(srdat)+3*sd(srdat))/length(srdat)*100

### Create a table
library(knitr)
kable(x=mat, digits=2,row.names=T, format="markdown")
```

2. We now will explore a dataset relating to the SAT. The dataset is available on Blackboard in the file named "SAT.csv"
   a. Read the data into RStudio and print the first 6 entries of the dataset.
   b. Plot total SAT score (y axis) against teacher salary (x axis). Make sure to label your axes. Comment on the form, strength, and direction of the plot.
   c. Calculate the correlation between SAT score and teacher salary. Why is this an interesting result?
   d. From `help(SAT)`, we can see that `frac` is the percentage of all elligible students taking the SAT. Create a categorical variable based on `frac`. In particular, categorize `frac` into "low", "medium", and "high" groups, based on whether the percentage was in the range "(0,22)", "(22,49)","(49,81)", respectively. Do this using the cut function as follows:

```r
SAT$new_var <- cut(SAT$frac,breaks=c(0,22,49,81),labels=c(1,2,3))
```

   e. Plot total SAT score (y axis) against teacher salary (x axis), but now color points based on which `frac` group then fall into. You can do this by specifying the `col=new_var` option in the `plot()` function. Comment on the results.
   f. Calculate the correlation between SAT score and teacher salary for each of the three groups specified by `new_var`. Comment on the results. You can use the following code to get started:

```r
cor(SAT$salary[SAT$new_var==1], SAT$sat[SAT$new_var==1])
```

You should see that the correlations for each group have a different sign from when all the data are considered together. This is an example of ***Simpson's paradox****, since by considering a third variable (in this case* ***frac****) the direction of association between two*

variables of interest has changed.

g. Create two new categorical variables. The first will be an indicator for whether `expend` is greater than 7. The second will be an indicator for whether `sat` is greater than 900. Use the following example code:

```
SAT$expendCAT <- ifelse(SAT$expend > 7, "high", "low")
```

h. Create a stacked barplot with a bar for each SAT score group. Each bar should be broken up into two pieces: one for high expenditures and one for low expenditures Make sure to label your axes and add a legend. Comment on the results. Why does this result seem counterintuitive?