BSDS 100: Intro to Data Science with R
Assignment 3

by James D. Wilson (University of San Francisco)

Before we begin, make sure that you have R and RStudio properly installed. Also make sure that you understand how to use the knitr package and RMarkdownfrom the Lecture 3 on the course website https://github.com/jdwilson4/Intro-Data-Science-2017.

Directions: For all questions in this assignment, write complete sentences and fully answer any question that is asked. Provide all R code and solutions by *knitting* your final RStudio file into a single file named [your_name]_CA3.pdf. This assignment is due next Tuesday at the beginning of class. Late assignments will automatically have 10 points deducted.

1. *Loading pre-stored data and viewing the data:* We first consider analyzing the *iris* dataset, which was collected by the famous statistician Sir Ronald Fisher (in the early 1900s). This dataset is already available in R, so we only need to call the dataset from the console. Load and view, and read a description about the data using the below commands.

```
#Load the data
data(iris)

#Look at the data
iris

#Read a description of the data
help(iris)
```

When data is first loaded to your R console, it is stored as a **data frame** structure. Data frames are a tabular representation of data and can contain any mix of characters, numerical quantities, or factors. An important aspect of the data frame is that it contains both row and column names. Often, we want to extract a subset of the data to take a closer look. This can be done directly using the row or column name *or* by calling the number of the row/column you want to extract. Follow the example below for an illustration.

```
#Display the column and row names of the data
colnames(iris)
rownames(iris)

#Look at the 10th row
iris[10, ]

#Look at the 3rd column
iris[, 3]

#Alternatively, just look at the variable "Petal.Length"
iris$Petal.Length

#Store columns 1-2 and rows 10-20 for later use
```

```
subset.data <- iris[10:20, 1:2]

#Store the species names for later use
Species.names <- iris$Species
```

As we see above, the $ symbol can be used to call a column as long as it is followed by the *exact* column name. **Note:** Remember that `R` is case sensitive! In the command line (not in your *knit* file), run the command *iris$petal.Length* to see what happens in your console.

Questions

(a) What are the different variables in this dataset? What are the measurements of each of these variables for the 10th sample?

(b) How many samples are in this dataset? What are the different species of flower that have been measured?

2. *Summary Statistics*: With any dataset, one of the first forms of exploratory analysis involves calculating summary statistics from the data like the five number summary: the minimum, median, third quartile, maximum and mean. We are also interested in the variation of a variable as measured by its standard deviation.

   - One can use the *summary()* function to calculate a five number summary. The *apply()* function can be used to find the standard deviation of each row or column of a data frame or matrix structure. We note that the *apply()* function is very general as it can apply any other `R` command across rows or columns of a dataset. Here, we apply the *sd()* function across the columns (and then rows) of the *iris* dataset. Calculate these summary statistics of the *iris* dataset using the code below.

     ```
     #calculate the 5-number summary of the {\it iris} dataset
     five.num.summary <- summary(iris)

     #calculate the standard deviation across the first four columns of the dataset
     st.devs.cols <- apply(iris[,1:4], 2, sd)

     #calculate the standard deviation across the samples of the dataset
     st.devs.rows <- apply(iris[, 1:4], 1, sd)
     ```

   - Now, we can vizualize the quantitative summary of the 4 variables by building a boxplot using the *boxplot()* command. Plot a boxplot using the following command:

     ```
     boxplot(iris[, 1:4], main = "Boxplot of Iris Variables")
     ```

Questions

(a) How many of each species are there in the dataset?

(b) Which two variables have a median that is smaller than their corresponding mean?

(c) What is the standard deviation of the sepal length measurements?

(d) Calculate the 5-number summary and standard deviations of the subset you extracted earlier (*subset.data*). What is the standard deviation of the sepal length measurements for this subset?