# CIND 123 - Data Analytics: Basic Methods

Rui Zhang

Assignment 1 (10%)

Rui Zhang

DHA 500736315

---

## Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown http://rmarkdown.rstudio.com.

Use RStudio for this assignment. Complete the assignment by inserting your R code wherever you see the string "#INSERT YOUR ANSWER HERE".

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

Submit **both** the rmd and generated output files. Failing to submit both files will be subject to mark deduction.

## Sample Question and Solution

Use `seq()` to create the vector $(1, 2, 3, \ldots, 10)$.

```r
seq(1,10)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

## Question 1

a) Use the `seq()` function to create the vector $(1, 7, 13, \ldots, 61)$. Note that each term in this sequence is of the form $1 + 6n$ where $n = 0, \ldots, 10$.

```r
seq(1, 61, by=6)
```

```
##  [1]  1  7 13 19 25 31 37 43 49 55 61
```

b) Use `rep()` to create the vector $(2, 3, 4, \ldots, 2, 3, 4, \ldots, 2, 3, 4)$ in which the sequence $(2, 3, 4)$ is repeated 5 times.

```r
rep(2:4, 5)
```

```
##  [1] 2 3 4 2 3 4 2 3 4 2 3 4 2 3 4
```

c) To convert factor to number, would it be correct to use the following commands? Explain your answer.

```
factorVar <- factor(c(1, 6, 5.4, 3.2));as.numeric(factorVar)
```

```r
# after converting factorvar to numeric, it should assign back
factorVar <- factor(c(1, 6, 5.4, 3.2))
factorVar <- as.numeric(factorVar)
```

d) A comma-separated values file `dataset.csv` consists of missing values represented by question marks (`?`) and exclamation mark (`!`). How can you read this type of files in R?

```r
# replace "?" and "!" by "NA"
# read.csv("dataset.csv", na.strings=c("?","!","NA"))
```

## Question 2

a) Compute:

$$\sum_{n=10}^{100} n^3$$

```r
sum((10:100)^3)
```

```
## [1] 25500475
```

b) Compute:

$$\sum_{n=1}^{10} \left( \frac{2^n}{n^2} + \frac{n^4}{4^n} \right)$$

```r
sum(2^(1:10)/(1:10)^2+(1:10)^4/4^(1:10))
```

```
## [1] 35.80589
```

c) Compute:

$$\sum_{n=0}^{10} \frac{1}{(n+1)!}$$

```r
sum(1/factorial((0:10)+1))
```

```
## [1] 1.718282
```

d) Compute:

$$\prod_{n=3}^{33} \left( 3n + \frac{3}{\sqrt[3]{n}} \right)$$

```r
n <- (3:33)
sum(3*n + 3/n^(1/3))
```

```
## [1] 1712.463
```

e) Explain the output of this R-command: `c(0:5)[NA]`

```r
c(0:5)[NA]
```

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

```r
# Create 5 vector, each one has NA value
```

f) What is the difference between is.vector() and is.numeric() functions?

```r
# is.vector tests if input argument is vector
# is.vector tests if each element of input argument is vecor
```

g) List at least three advantages and three disadvantages of using `RShiny` package?

```r
# 1. easy to build interactive web apps straight from R
# 2. extend Shiny apps with CSS themes, htmlwidgets, and JavaScript actions
# 3. Shiny apps are easy to write. No web development skills are required
# 4. You can communicate results via interactive charts, visualizations, text, or tables
# 5. Built-in capabilities let you share your work easily with colleagues and friends.
```

## Question 3

`iris` dataset gives the measurements in centimeters of the variables sepal length, sepal width, petal length and petal width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Install the `iris` dataset on your computer using the command `install.packages("datasets")`. Then, load the `datasets` package into your session using the following command.

```
library(datasets)
```

a) Display the first six rows of the `iris` data set.

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

b) Compute the average of the first four variables (Sepal.Length, Sepal.Width, Petal.Length and Petal.Width) using `sapply()` function.

Hint: You might need to consider removing the `NA` values, otherwise the average will not be computed.

```
sapply(iris[c(1:4)], mean, na.rm = TRUE)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     5.843333     3.057333     3.758000     1.199333
```

c) Show how to use R to replace the missing values in this dataset with plausible ones.

```
iris$Sepal.Length[is.na(iris$Sepal.Length)] <- mean(iris$Sepal.Length, na.rm=T)
iris$Sepal.Width[is.na(iris$Sepal.Width)] <- mean(iris$Sepal.Width, na.rm=T)
iris$Petal.Length[is.na(iris$Petal.Length)] <- mean(iris$Petal.Length, na.rm=T)
iris$Petal.Width[is.na(iris$Petal.Width)] <- mean(iris$Petal.Width, na.rm=T)
```

d) Compute the standard deviation for only the first and the third variables (Sepal.Length and Petal.Length)

```
sd(iris$Sepal.Length, na.rm=T)
```

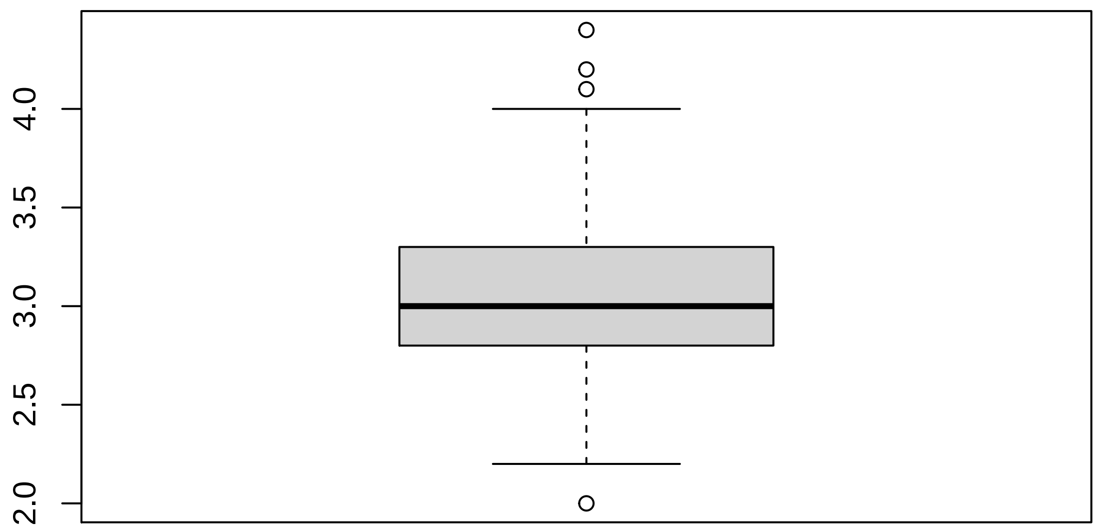
```
## [1] 0.8280661
```

```
sd(iris$Petal.Length, na.rm=T)
```

```
## [1] 1.765298
```

e) Construct a boxplot for `Sepal.Width` variable, then display the values of all the outliers. Explain how these outliers have been calculated.

```r
boxplot(iris$Sepal.Width)
```




```r
# outliers are points below lower fence Q1-1.5(IQR) or higer than Upper fence Q3+1.5(IQR)
```

f) Compute the upper quartile of the `Sepal.Width` variable with two different methods.

```r
quantile(iris$Sepal.Width, 0.75, na.rm=T)
```

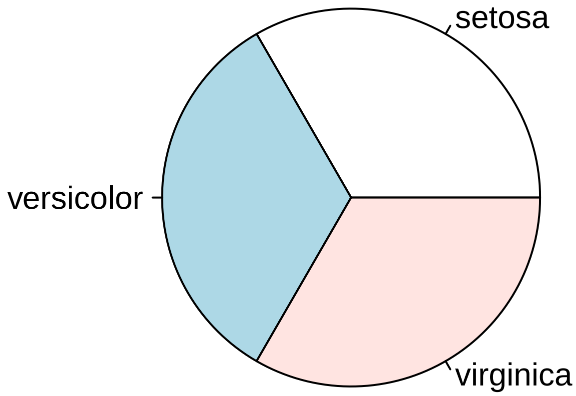
```
## 75%
## 3.3
```

```r
summary(iris$Sepal.Width)[5]
```

```
## 3rd Qu.
##     3.3
```

g) Construct a pie chart to describe the species with 'Sepal.Length' less than 7 centimeters.

```r
labels <- unique(iris$Species)
sepalLess7 = iris[iris$Sepal.Length<7,]
setosa = sepalLess7[sepalLess7$Species==labels[1],]
versicolor = sepalLess7[sepalLess7$Species==labels[2],]
virginica = sepalLess7[sepalLess7$Species==labels[3],]
pie(c(length(setosa),length(versicolor),length(versicolor)), labels)
```




END of Assignment #1.