Name:\_Rahul Avadhoot\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Last 4 Digits of Student ID Number:\_1787\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Introduction to R for Data Analysis in the Health Sciences**

**BIOST 509**

**Homework Assignment #1**

**Due: 9:00 AM on October 4, 2017**

Instructions

Enter your numerical and/or written answers to the questions into this Word file or create your own Word document (.doc or .docx) or pdf file with your answers. Your work will be graded as 1 (complete) if at least 70% of the questions are answered correctly (and/or an obvious “good faith” effort), or 0 (incomplete) otherwise. For this homework assignment, also include the R commands that you used to obtain your answers.

All questions for this homework are related to an airquality dataset with daily air quality measurements from New York between May and September 1973. The following variables are included:

* Ozone: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
* Solar.R: Solar radiation in Langleys in the frequency band 4000–7700 Angstroms from 0800 to 1200 hours at Central Park
* Wind: Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
* Temp: Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.
* Month: Number of month
* Day: Number of day during month

The dataset is available on the course canvas site (airquality.csv within the datasets file folder), and on the web at <http://faculty.washington.edu/tathornt/Biost509/DataSets/airquality.csv>.

First, download the file from canvas to your computer and import the dataset into RStudio, or import the dataset directly from the web using the “Import Dataset” tab within the RStudio Environment. Then, answer questions 1-6 below using this dataset.

1. What are the dimensions of the data set?

Number of rows = 153

Number of variables (columns) = 6

R Command: dim(airquality)

2. What is the largest daily temperature? (Don’t find it by looking through the data - use an R function!)

97 F

R Command: max(airquality$Temp)

3. On which month and day did the highest temperature occur? (Don’t find it by looking through the data - use an R function!)

Month: 8

Day: 28

R Command: airquality[which.max(airquality$Temp),]

4. What is the R command you used to identify the month and day of the maximum temperature?

R Command: airquality[which.max(airquality$Temp),]

5. Compute the following descriptive statistics for temperature.

Mean temperature: 77.88 F

Standard deviation of temperature: 9.46527 F

Median temperature: 79.00 F

R Command: summary(airquality$Temp)

sd(airquality$Temp)

6. The ozone variable has missing values. What value do you get when you try to compute the mean of the ozone variable using R’s default mean function? Why did R return this value for the mean? (We will see how to fix this problem later.)

mean(airquality$Ozone) returns NA.

R returns NA whenever there are missing values present in the data

We can fix this problem by using mean(airquality$Ozone,na.rm = TRUE) command