Answers to the assignments

Chapter 1: Descriptive statistics

1.1 a) Mean: 6.6   Range: 8
Mode: 8   Lower quartile: 4
Median: 7   Upper quartile: 9
                                                     Interquartile range: 5

1.1 b) Mean: 5.91   Range: 9
Mode: 7   Lower quartile: 3.5
Median: 6.5   Upper quartile: 8.5
                                                     Interquartile range: 6

1.1 c) Assignment 1.1b was probably harder to do as you had to take the middle of two numbers to find the median, and the lower and upper quartiles.

1.1 d) These data sets are negatively skewed.

Explanation: The mean is lower than the median and mode, and so more values are concentrated on the right side (tail) of the distribution graph while the left tail of the distribution graph is longer.

1.2 a) The **View()** command opens a window in which you can inspect the data.

```
dataset1 <- c(2, 7, 4, 5, 8, 10, 10, 7, 9, 2, 8, 8, 9, 4, 6)
View(dataset1)
```

1.2 b) Mean: 6.6   Median: 7
Mode: 8   Range: 8

```
mean(dataset1) # Mean: 6.6
table(dataset1) # Mode: 8 is the most occurring number (3 times)
median(dataset1) # Median: 7
range(dataset1) # Range: 2 to 10 = 8
```

1.2 c) The **quantile()** command returns the minimum (0%), lower quartile (25%), median (50%), upper quartile (75%), and maximum (100%).

```
quantile(dataset1, type = 6)
```

1.2 d) Mean: 5.91   Range: 9
        Mode: 7   Lower quartile: 3.5
        Median: 6.5   Upper quartile: 8.5

```
dataset2 <- c(7, 7, 6, 5, 2, 1, 3, 7, 5, 9, 9, 10)

mean(dataset2) # Mean: 5.91
table(dataset2) # Mode: 7 is the most occurring number (3 times)
median(dataset2) # Median: 6.5
range(dataset2) # Range: 1 to 10 = 9
```

1.3 a) The code opens a new window in which you can select the **.csv** file that you want to read into your R session. Using the **colnames()** function, you can see that the **bloodPressure.csv** file contains 6 columns named **Number**, **Age**, **BloodPressure**, **Cholestrol**, **Gender**, **Description**.

1.3 b) You can use a relative path to the file on your computer by providing it directly to the **read.csv()** function (in quotes **'bloodPressure.csv'**). Remember to set your working directory correctly, since R will look inside the working directory folder when it receives such a path. You can also specify a full path (like **'C://path/to/file/bloodPressure.csv'**). With full file paths, R will know exactly where to look, and the location of your working directory does not matter.

```
dataset3 <- read.csv('bloodPressure.csv')
```

1.3 c) The mean age of the respondents ($n = 60$) in the data set is 45.15 years. The minimum age of the respondents is 17, and the maximum age is 69. The age of the respondents spans 52 years. The most occurring age is 39. Twenty-five percent of the respondents is aged below 34.5, fifty percent is aged below 46, and 75 percent is aged below 58.5.

```
mean(dataset3$Age) # Mean: 45.15
table(dataset3$Age) # Mode: 39 is the most occurring number (4 times)
median(dataset3$Age) # Median: 46
range(dataset3$Age) # Range: 17 to 69 = 52

quantile(dataset3$Age, type = 6)
# Minimum: 17
# Lower quartile: 34.5
# Median: 46
# Upper quartile: 58.5
# Maximum: 69
```

1.3 d) Mode: 39

```
getMode <- function(x){
   uniqx <- unique(x)
   uniqx[which.max(tabulate(match(x, uniqx)))]
}

getMode(dataset3$Age) # Mode: 39
```

1.3 e) Mean: 130.62   Range: 121
        Mode: 129   Lower quartile: 118.75
        Median: 131.5   Upper quartile: 145.75

```
mean(dataset3$BloodPressure) # Mean: 130.62
getMode(dataset3$BloodPressure) # Mode: 129
median(dataset3$BloodPressure) # Median: 131.5
range(dataset3$BloodPressure) # Range: 45 to 166

quantile(dataset3$BloodPressure, type = 6)
# Minimum: 45
# Lower quartile: 118.75
# Median: 131.5
# Upper quartile: 145.75
# Maximum: 166
```

1.3 f) These data sets are not skewed.

Explanation: The mean is lower than the median but not than the
mode, and so you cannot conclude that the distribution is skewed
in any direction.

1.3 g) Variance: 1.059
Standard deviation: 1.029

```
var(dataset3$Cholestrol) # Variance: 1.059
sd(dataset3$Cholestrol) # Standard deviation: 1.029
```

1.3 h)
```
# The standard deviation is the square root of the variance
sd(dataset3$Cholestrol) == sqrt(var(dataset3$Cholestrol))
```