Statistics and Data Analysis
Homework 1: Descriptive Statistics

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. Find the values of the following expressions:

   (a) $1 + (2 - 9 \times 6) + \pi^2$.

   (b) $\dfrac{\sqrt{7800} + 10^{1.8}}{27}$.

   (c) $\lceil 10.68 \rceil + \max\{\sqrt{98}, \pi^2\}$.

2. Define the Boolean variables a, b, and c as

$$a = \begin{cases} \text{TRUE} & \text{if } 3 > 2 \\ \text{FALSE} & \text{otherwise} \end{cases}, b = \begin{cases} \text{TRUE} & \text{if } 2 < 4 \\ \text{FALSE} & \text{otherwise} \end{cases}, \text{ and } c = \begin{cases} \text{TRUE} & \text{if } \sqrt{15} = 4 \\ \text{FALSE} & \text{otherwise} \end{cases}.$$

   Moreover, let

$$d = \begin{cases} \text{TRUE} & \text{if } a, b, \text{ and } c \text{ are all TRUE} \\ \text{FALSE} & \text{otherwise} \end{cases}.$$

   Find the Boolean values of the following expressions:

   (a) (a AND b) OR (c AND d).

   (b) a AND (b OR c) AND d.

   (c) a AND b OR c AND d.

3. Consider the variable `teamHeight` defined as

```
teamHeight <- c(178, 172, 175, 184, 172, 175, 165, 178, 177, 175,
                180, 182, 177, 183, 180, 178, 179, 162, 170, 171)
```

   Assume that 1 foot is 30 centimeters and 1 inch is 2.5 centimeters.

   (a) How tall is the 6th team member in feet and inches?

   (b) How tall are all the team members in feet and inches?

   (c) Find the indices of those team members who are shorter than six feet.

   (d) Find the average height, in centimeters, for those members who are shorter than six feet.

4. Consider the variable `teamHeight` again.

   (a) Draw a histogram with the default number of classes. For each class, include the lower bound and exclude the upper bound. Which class has the highest frequency? What is that frequency?

   (b) Draw a histogram with ten classes $[160, 162.5)$, $[162.5, 165)$, ..., and $[182.5, 185)$. Which class has the highest frequency? What is that frequency?

5. Consider the variable `teamHeight` again and the histogram with 10 classes you just depicted.

   (a) Draw a pie chart for the 10 classes. Do not worry about the labels.

   (b) Draw a pie chart only for those classes which covers at least 2 members. Use frequencies to be the labels of the slices.

   (c) Draw a bar chart only for those classes which covers at least 2 members. Use class midpoints to be the labels of the bars.

6. Load the data set "SDA-Fa14_data_wholesale.txt" by executing the statements

```
W <- read.table("SDA-Fa14_data_wholesale.txt",
                header = TRUE)
ws <- data.frame(Channel = W$Channel, Region = W$Region, Fresh = W$Fresh)
```

   Browse through the data for a while. Whenever you want to extract a column as a vector, type `ws$Channel`, `ws$Region`, or `ws$Fresh`.

   (a) Draw a histogram for all the fresh food sales. Is there any extreme values?

   (b) Identify the index of that extreme value.

   (c) Draw a histogram by excluding those extreme values you found in Part (a).

7. Consider the data frame `ws` again:

   (a) Some people believe that customers at Lisbon in average consume more fresh food than those not at Lisbon. Based on the sample data, is that belief correct?

   (b) For the two channels (1 for hotel/restaurant/café, 2 for retail stores), whose average fresh food sales is higher?

   (c) Among the six channel-region combination, whose average fresh food sales is the highest?

   (d) Draw a bar chart for the six numbers you get from Part (c).

8. Consider the variables `price` and `size` defined as

```
size <- c(75, 59, 85, 65, 72, 46, 107, 91, 75, 65, 88, 59)
price <- c(315, 229, 355, 261, 234, 216, 308, 306, 289, 204, 265, 195)
```

   and the variable `TeamHeight`:

   (a) Find the sample variance and standard deviation for `price`.

   (b) Find the sample variance and standard deviation for `size`.

   (c) Find the sample coefficients of variation for `price` and `size`. Which variable has higher variability?

   (d) Find the variance and standard deviation for the population data `TeamHeight`.

   Note. The R functions `var()` and `sd()` find sample variances and standard deviations. How would you convert their outputs to population variances and standard deviations?

9. Consider the wholesale data set.

   (a) For sales data collected from channel 1 and region 1, calculate the means, medians, and sample variances for milk sales.

   (b) For sales data collected from channel 1 and region 1, draw a histogram for milk sales data with the default number of classes and class intervals.

   (c) For each of the six channel-region combination, calculate the sample correlation coefficient between fresh food sales and milk sales.

   (d) Draw scatter plots for the channel-region combinations with the highest and lowest correlation coefficients.