Week 1 - Homework

STAT 420, Summer 2017, Dalpiaz

Directions

- Be sure to remove this section if you use this `.Rmd` file as a template.
- You may leave the questions in your final document.

Exercise 1

For this exercise, we will use the `diabetes` dataset from the `faraway` package.

(a) Install and load the `faraway` package. **Do not** include the installation command in your `.Rmd` file. (If you do it will install the package every time you knit your file.) **Do** include the command to load the package into your environment.

(b) Coerce the data to be a tibble instead of a data frame. (You will need the `tibble` package to do so.) How many observations are in this dataset? How many variables? Who are the individuals in this dataset?

(c) Which variables are factor variables?

(d) What is the mean HDL level (High Density Lipoprotein) of individuals in this sample?

(e) What is the standard deviation of total cholesterol of individuals in this sample?

(f) What is the range of ages of individuals in this sample?

(g) What is the mean HDL of females in this sample?

(h) Create a scatter plot of HDL (y-axis) vs weight (x-axis). Use a non-default color for the points. (Also, be sure to give the plot a title and label the axes appropriately.) Based on the scatter plot, does there seem to be a relationship between the two variables? Briefly explain.

(i) Create a scatter plot of total cholesterol (y-axis) vs weight (x-axis). Use a non-default color for the points. (Also, be sure to give the plot a title and label the axes appropriately.) Based on the scatter plot, does there seem to be a relationship between the two variables? Briefly explain.

(j) Create side-by-side boxplots for HDL by gender. Use non-default colors for the plot. (Also, be sure to give the plot a title and label the axes appropriately.) Based on the boxplot, does there seem to be a difference between HDL level and gender? Briefly explain.

Exercise 2

For this exercise we will use the data stored in `nutrition.csv`. It contains the nutritional values per serving size for a large variety of foods as calculated by the USDA. It is a cleaned version totaling 5138 observations and is current as of September 2015.

The variables in the dataset are:

- `ID`
- `Desc` - Short description of food
- `Water` - in grams
- `Calories` - in kcal
- `Protein` - in grams
- `Fat` - in grams

- Carbs - Carbohydrates, in grams
- Fiber - in grams
- Sugar - in grams
- Calcium - in milligrams
- Potassium - in milligrams
- Sodium - in milligrams
- VitaminC - Vitamin C, in milligrams
- Chol - Cholesterol, in milligrams
- Portion - Description of standard serving size used in analysis

(a) Create a histogram of Calories. Do not modify R's default bin selection. Make the plot presentable. Describe the shape of the histogram. Do you notice anything unusual?

(b) Create a scatter plot of calories (y-axis) vs protein (x-axis). Make the plot presentable. Do you notice any trends? Do you think that knowing only the protein content of a food, you could make a good prediction of the calories in the food?

(c) Create a scatter plot of Calories (y-axis) vs 4 * Protein + 4 * Carbs + 9 * Fat + 2 * Fiber (x-axis). Make the plot presentable. You will either need to add a new variable to the data frame, or, use the I() function in your formula in the call to plot(). If you are at all familiar with nutrition, you may realize that this formula calculates the calorie count based on the protein, carbohydrate, and fat values. You'd expect then that the result here is a straight line. Is it? If not, can you think of any reasons why it is not?

Exercise 3

For each of the following parts, use the following vectors:

```
a = 1:10
b = 10:1
c = rep(1, times = 10)
d = 2 ^ (1:10)
```

(a) Write a function called sum_of_squares.

- Arguments:
    - A vector of numeric data x.
- Output:
    - The sum of the squares of the elements of the vector. $\sum_{i=1}^{n} x_i^2$

Provide your function, as well as the result of running the following code:

```
sum_of_squares(x = a)
sum_of_squares(x = c(c, d))
```

(b) Write a function called rms_diff.

- Arguments:
    - A vector of numeric data x.
    - A vector of numeric data y.
- Output:
    - $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$

Provide your function, as well as the result of running the following code:

```
rms_diff(x = a, y = b)
rms_diff(x = d, y = c)
rms_diff(x = d, y = 1)
rms_diff(x = a, y = 0) ^ 2 * length(a)
```