Homework 00

STAT 430, Fall 2017

Due: Friday, September 8, 11:59 PM

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.

Solution:

```
library(faraway)
```

(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?

Solution:

```
library(tibble)
diabetes = as_tibble(diabetes)
diabetes
```

```
## # A tibble: 403 x 19
##       id  chol stab.glu   hdl ratio glyhb   location   age gender height
##  * <int> <int>    <int> <int> <dbl> <dbl>     <fctr> <int> <fctr>  <int>
##  1  1000   203       82    56   3.6  4.31 Buckingham    46 female     62
##  2  1001   165       97    24   6.9  4.44 Buckingham    29 female     64
##  3  1002   228       92    37   6.2  4.64 Buckingham    58 female     61
##  4  1003    78       93    12   6.5  4.63 Buckingham    67   male     67
##  5  1005   249       90    28   8.9  7.72 Buckingham    64   male     68
##  6  1008   248       94    69   3.6  4.81 Buckingham    34   male     71
##  7  1011   195       92    41   4.8  4.84 Buckingham    30   male     69
##  8  1015   227       75    44   5.2  3.94 Buckingham    37   male     59
##  9  1016   177       87    49   3.6  4.84 Buckingham    45   male     69
## 10  1022   263       89    40   6.6  5.78 Buckingham    55 female     63
## # ... with 393 more rows, and 9 more variables: weight <int>,
## #   frame <fctr>, bp.1s <int>, bp.1d <int>, bp.2s <int>, bp.2d <int>,
## #   waist <int>, hip <int>, time.ppn <int>
```

```
?diabetes
```

We find there are 403 observations and 19 variables that describe African Americans from central Virginia.

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

Solution:

```
any(is.na(diabetes$hdl))
```

```
## [1] TRUE
```

```r
anyNA(diabetes$hdl)
```

```
## [1] TRUE
```

```r
mean(diabetes$hdl, na.rm = TRUE)
```

```
## [1] 50.44527
```

Notice that we need to deal with some missing data. We only remove observations with missing data from the variable of interest. Had we instead removed any observation with missing data, we would have less data to calculate this statistic.

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

Solution:

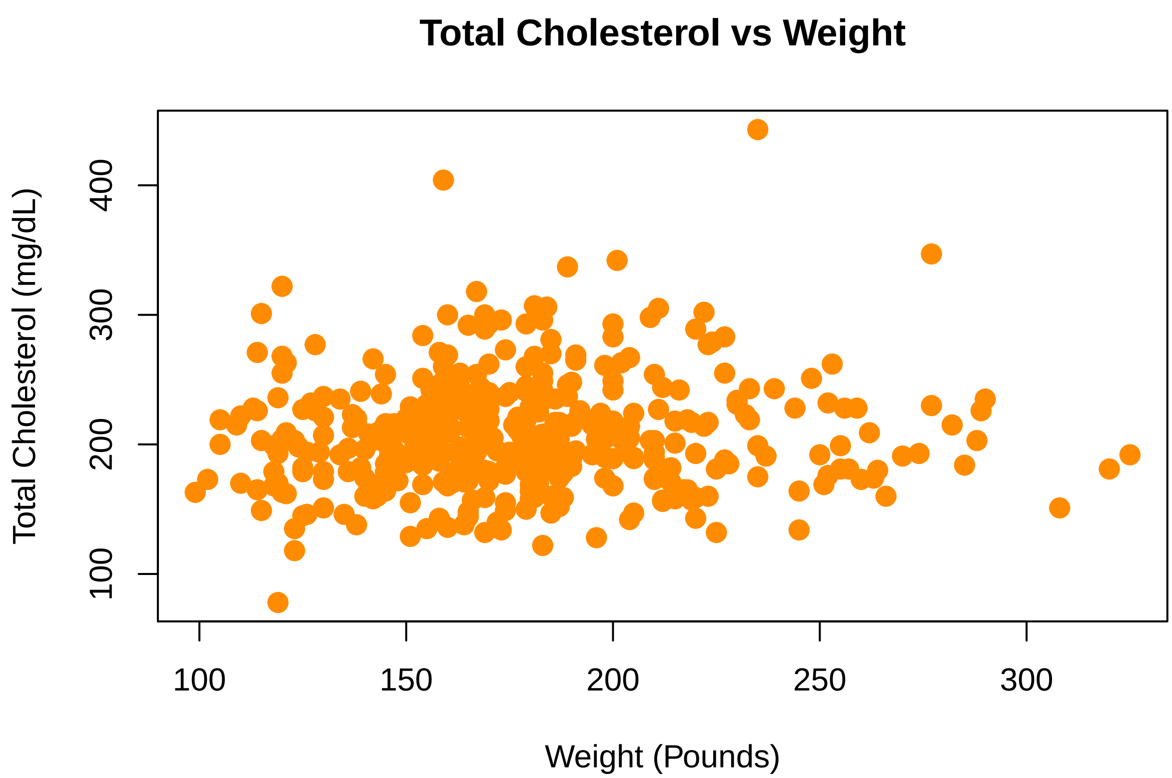
```r
mean(subset(diabetes, gender == "female")$hdl)
```

```
## [1] 52.11111
```

(e) 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.

Solution:

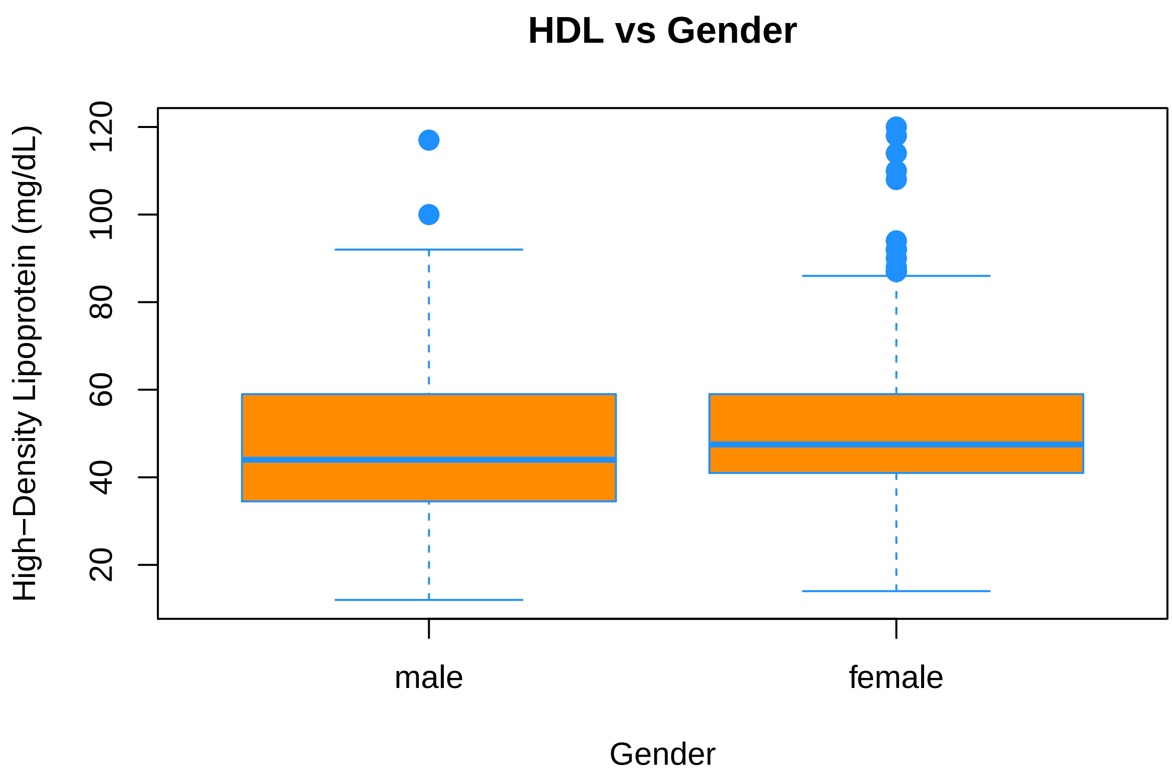
```r
plot(chol ~ weight, data = diabetes,
     xlab = "Weight (Pounds)",
     ylab = "Total Cholesterol (mg/dL)",
     main = "Total Cholesterol vs Weight",
     pch  = 20,
     cex  = 2,
     col  = "darkorange")
```



Overall, we see very little trend. Average total cholesterol seems nearly constant for different weights.

(f) 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 in HDL level between the genders.? Briefly explain.

```
boxplot(hdl ~ gender, data = diabetes,
     xlab = "Gender",
     ylab = "High-Density Lipoprotein (mg/dL)",
     main = "HDL vs Gender",
     pch  = 20,
     cex  = 2,
     col     = "darkorange",
     border = "dodgerblue")
```



Aside from slightly less variation among females, there seems to be very little difference in HDL level between the genders.

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:

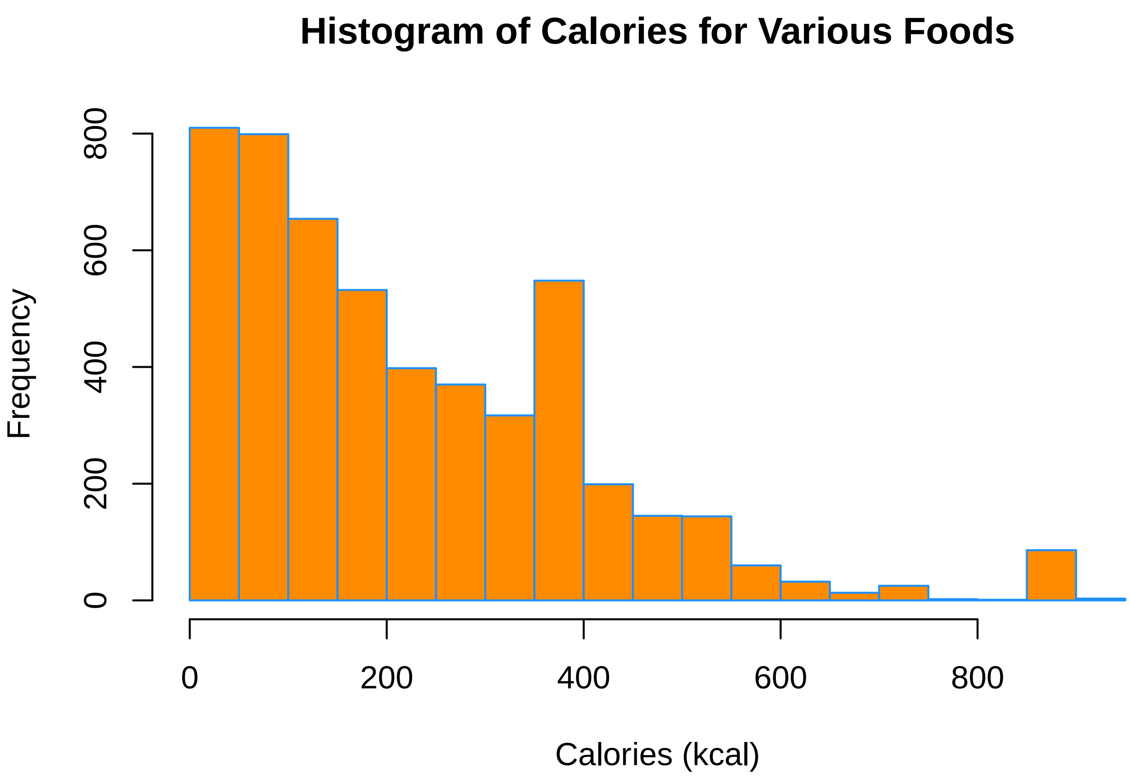
- `ID`
- `Desc` - Short description of food
- `Water` - in grams
- `Calories` - in kcal

- `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?

Solution:

```r
library(readr)
nutrition = read_csv("nutrition.csv")
hist(nutrition$Calories,
     xlab = "Calories (kcal)",
     main = "Histogram of Calories for Various Foods",
     border = "dodgerblue",
     col  = "darkorange")
```



The distribution of `Calories` is right-skewed. There are two odd spikes, one around 400 kcal and one past 800 kcal. Perhaps some foods are being rounded to 400, or portion sizes are created with 400 kcal in mind. Also, perhaps there is an upper limit, and portion sizes are created to keep calories close to 900 but not above.

(b) 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?

Solution:

```
plot(Calories ~ I(4 * Protein + 4 * Carbs + 9 * Fat + 2 * Fiber), data = nutrition,
     xlab = "Protein (grams)",
     ylab = "Calories (kcal)",
     main = "Calories vs Protein",
     pch  = 20,
     cex  = 1,
     col  = "darkorange")
```

The result is *not* a straight line. There could be any number of reasons:

- There are actually additional components that make up food energy that we are not considering. See Wikipedia: Food Energy.
- Rounding
- Measurement error

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))
```

Solution:

```
sum_of_squares = function(x) {
  sum(x ^ 2)
}

sum_of_squares(x = a)
```

```
## [1] 385
```

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

```
## [1] 1398110
```

(b) Write a function called `rms_diff`.

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

If the vectors have different lengths, the shorter vector should be repeated until it matches the length of the longer vector.

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)
```

Solution:

```
rms_diff = function(x, y) {
  sqrt(mean((x - y) ^ 2))
}

rms_diff(x = a, y = b)
```

```
## [1] 5.744563
```

```
rms_diff(x = d, y = c)
```

```
## [1] 373.3655
```

```
rms_diff(x = d, y = 1)
```

```
## [1] 373.3655
```

```
rms_diff(x = a, y = 0) ^ 2 * length(a)
```

```
## [1] 385
```

Notice the value `385` appears again!