BAIS:3250
Homework 5

Total Points: 50

Specifications

- For this homework, you should submit an R script file with your codes for all questions
  - Name your file "***hawkid*_homework5.R**"
  - Add a comment with your name at the top of the file
  - Add `rm(list=ls())` at the top of your file to clear the workspace
  - Add comments denoting each question number
  - You may add other comments for clarification
- The solution for each question must be generated as R variables or output files with specific names as instructed
  - All solutions should be generated by running your code without any customization or modification by the instructor
  - Your code should assume that all necessary files are in the current working directory. Do not include and `setwd()` commands
  - Load required packages with the `library()` command. Your script should not include any unnecessary packages or `install()` commands

---

Carefully review "food.txt" (the data file) and "food_description.txt" (the data description). Then answer the following questions:

1. (8 points) Use `read.csv()` to load the data into R, applying the optional arguments to denote missing values and treat all text as factors. **Note:** There are no UTF-8 or special characters in the data

   Use R code to produce the following output:

   - `food`: data frame created from "food.txt"
   - `food$Total`: create a new integer column that stores the number of foods that each respondent has eaten before
   - `food$Continent`: edit/collapse the levels of the factor so that there are only 2 (North America and Other)

2. (12 points) Is there a relationship between a respondent's demographics and whether they've tried specific foods?

   - `caviar`: proportion table that calculates the marginal proportion of North American vs. Non-North American respondents that have (or have not) eaten caviar, relative to the total number of respondents from that location. **Note:** respondents with an unknown location will be excluded from the results
   - `alligator_test`: variable that stores the results of a Chi-squared test examining whether having eaten alligator and gender are independent
   - Add a comment to your script interpreting the result of the Chi-squared test. Use the following format/template:

The p-value of the Chi-squared is XX. Thus, we YY reject the null hypothesis that having eaten alligator and gender are independent.

Replace XX with the p-value from the Chi-squared test. Replace YY with "can" or "cannot".

3. (4 points) Is there a relationship between a respondent's age and the total number of foods they've tried?

    - `age_total`: numeric variable that stores the correlation between respondent age and total number of foods tried
    - Add a comment to your script interpreting the correlation coefficient. Use the following format/template:

      The Pearson correlation coefficient is XX. Thus, these variables have a YY, ZZ linear relationship.

      Replace XX with the correlation coefficient. Replace YY with "weak" or "strong". Replace ZZ with "positive" or "negative".

4. (10 points) Use the `RSocrata` package to collect data on the Growth and Survival Carapace Widths experiment that investigates the effect of density on the growth of red king crabs (https://www.opendatanetwork.com/dataset/noaa-fisheries-afsc.data.socrata.com/vsba-nbxa). Specifically, use the SoQL query language to retrieve the columns for date, habitat, treatment type, and carapace width. Only retrieve data for the 2011 experiment. (**Hint:** review the API documentation for the dataset to make sure you have the correct endpoint and field names)

    - `crab`: data frame that matches the specifications above. **Note**: If `carapace_width` is not read into R as a numeric data type, transform it to a number after loading the data.

5. (16 points) Is there a difference in crab size for the two experimental treatments?

    - `treatment_summary`: `dplyr` summary table that calculates the median carapace width for each experimental treatment, plus the number of crabs in each group. The summary table should have 3 columns called "treatment", "total_crabs" and "median_size"
    - `treatment_test`: variable that stores the results of a t-test examining whether the average carapace width is significantly different for crabs raised in each treatment condition ("20" represents crabs raised in a tank with 20 other crabs, "5" represents crabs in a tank with 5 other crabs)
    - Add a comment to your script interpreting the results of the t-test. Use the following format/template:

      The p-value of the t-test is XX. Thus, we YY reject the null hypothesis that average carapace width is equal for the treatment conditions.

      Replace XX with p-value. Replace YY with "can" or "cannot"

    - T-tests rely on an assumption of normality. Create a "Base R" histogram displaying the distribution of carapace width over all crabs