OPIM 5603 — Statistics in Business Analytics
Fall 2019, University of Connecticut

Homework 2 - v2

Instructions: Please complete the following questions and submit them as an RNotebook (as an `Rmd` file) via the submission link on HuskyCT. You must submit the assignment by the time and due date listed on the course syllabus. Failure to submit a file by the deadline will result in a score of 0 on the assignment.

Set the heading of the RNotebook as an `html_document`. Add your name and a date to the header as well. The solution to each problem should be a separate section (specified by #), and each subproblem should be set as a subsection (specified as ##). For example, for Problem 2, you should have a section titled Problem 2, specified by:

```
# Problem 2
```

in your RNotebook. Also, for subproblem b in Problem 2, you should have a subsection, specified by:

```
## Problem 2b
```

As with all course material, the problems appearing in this homework assignment are taken from the instructor's real-world experiences, from other courses taught at the University of Connecticut, and from the sources listed in the course syllabus.

Note that R code submitted should work independent of the data that sits in the data structure. For example, suppose there was a vector `r_vec` with the values $(1, 2, 6)$ and the problem asks for you to create R code to create a vector `answer` which doubles each element of `r_vec`. The answer

$$\texttt{answer} \leftarrow \text{c}(2, 4, 12)$$

would be given no credit. The answer

$$\texttt{answer} \leftarrow 2^{*}\texttt{r\_vec}$$

would be an appropriate answer. If you have any questions, please submit them via email to the instructor and/or the teaching assistant prior to submitting your solution.

Problem 1 (40 points)

By default, `R` provides several packages that are downloaded and loaded. One data set is called `women`, which is a data frame and is available in the base package of `R`, i.e., it is automatically available.

a. How many variables are in this data frame? What does each variable represent? Don't forget to include units.

b. Calculate the body mass index (Weight in lbs*703/height in inches^2) for every observation and create a variable BMI which records the output.

c. You can add a variable `x` that you want to name `name` to a dataframe `df` by writing

`df[name] = x`.

Add the BMI to `women` with an appropriate column name.

d. Calculate the arithmetic mean, standard deviation, skewness, and kurtosis of the BMIs without using any package or functions, except for `sum()` and `length()`, togehter with standard mathematical operators.

e. Is the distribution of BMIs skewed? If so, in what direction? Explain.

f. Is the distribution of BMIs leptokurtic, mesokurtic, or platykurtic? Explain.

Problem 2 (40 points)

In fantasy sports, participants select collections of players that accumulate points throughout a sport game. In this problem we will be exploring the distribution of points that players accumulate and their projections.

a. Consider the data set `filtered_qb.csv` which contains records from top quarterback from last NFL season. Load the data into `R`.

b. Calculate the arithmetic mean, standard deviation, and skew for both the `projection` variable and the `actual` variable. Are there any noticable differences in the central tendency, variability, or shape of the two variables?

c. `R` has a built-in function called `fivenum` which calculates the five number summary. It returns a vector of length five that reports the five number summary. Create a vector called `our.five.sum` which stores the five number summary for the `actuals` variable. Print the median of this variable by indexing the appropriate value in `our.five.sum`.

d. `R` has tremendous plotting capabilities.

- Use the function `boxplot()` to create two box plots, one for each variable (`projections` and `actuals`). You need only provide one argument, which is the variable.
- Use the function `plot()` to create a scatter plot showing the relationship between `projections` and `actuals`. Provide just two inputs to the functions, the two variables of interest. Does it look like the projections are fairly accurate?
- Use the function `hist()` to create two histograms, one for each variable (`projections` and `actuals`). You need only provide one argument, which is the variable.
- The functions for plotting variables have many arguments that are customizable. Use the `help()` function to explore some of these. Specifically, for histograms, explore the optional argument `breaks`, set it to 10 and to 30 to see the differences in the `actuals` histogram.

Problem 3 (20 points)

From `ISwR` package, use the `cars` data frame for the following questions.

a. How many rows are in this data frame?

b. Create a copy of the `cars` data frame and assign it to `copy_cars`.

c. Divide the dist with speed in `copy_cars` and assign it to a new variable called time within `copy_cars`. Round this to 2 nearest decimals.

d. `R` has a function `head()` which displays the top rows of a dataframe. Print the top 14 rows of `copy_cars` using the `head()` function.

e. Calculate the median of the time variable using the function `median()`.

f. Print $Q1$ and $Q3$ by extracting the value in the relevant position of the five number summary.