**Homework Assignment: Week 1**

Assignment Due: Sunday, January 9, 2022 by Midnight CT

Submit electronic solution files to the drop box in BrightSpace

Answer each question. You may add your answers, when appropriate, to this document and submit the document to the drop box.

1. Classify each of the following data as qualitative, quantitative, or logical. Explain your choice of classification.

1. Your Graceland ID number
   1. This is qualitative data because it identifies an entity. It appears quantitative, but it does not represent a numerical quantity.
2. Wind speed
   1. This is quantitative data because it is a numerical measurement. It is most likely continuous quantitative data because wind speed can be any value of an interval.
3. The binary computer output 0110
   1. This is logical data because the computer output represents a series of logical statements (True, False).

2. Consider the datasets below. If the dataset follows the principles of Tidy Data, label it “Tidy Data.” If it does not follow the principles of Tidy Data, state the principle (or principles) not followed, and reformat the data so that it follows the Tidy Data principles.

The following dataset does not follow the principles of Tidy Data. The ‘values’ column has multiple types of observational units. It can be confusing to tell what the value is depicting.

a) name, names, values

"Phillip Woods" "age” 45

"Phillip Woods" "height" 186

"Jessica Cordero" "age" 37

"Jessica Cordero" "height" 156

Name Age Height

"Phillip Woods" 45 186

"Jessica Cordero" 37 156

The following dataset does follow Tidy Data principles.

b) country year Accidents

1 Afghanistan 1999 745

2 Afghanistan 2000 2666

3 Brazil 1999 37737

4 Brazil 2000 80488

5 China 1999 212258

6 China 2000 213766

3. On page 8 of the document *A Short R Tutorial* (see the Resource Module in BrightSpace), the data class **Factor** is defined (page 7). Factors are particularly difficult data types and we will explore some of their properties later in the course. At this point, write a definition of a **Factor** variable (or vector) type and provide an example, different from the one in this document.

A Factor variable is one that categorizes an observation based on its fixed number of levels. An example can include categorizing students by major – CSIT, Math, Economics, English, Business, Art. With this list of majors, there are 6 levels that a student (an observation) can be categorized by.

4. If the following R code was executed, what would the final output be?

The final output will be -425.

1 a <− 20

2 b <− -19

3 c <− 3

4. t <− a\*(b-c) + 5\*c

5. The vector **shots** has the following elements (75, 81, 68, 75, 72). If the following R code was executed, what would the final output be?

The final output will divide each entry by 3.

25.00000 27.00000 22.66667 25.00000 24.00000

1 shots/3

**For questions 6 – 10, you will need to write code in R to solve the problem. You are asked to submit your R code for grading…I do not need to see the actual answers, as I will run the code myself. You may provide one R file that contains all of the solutions if you desire, just be sure that the solution are identified by problem number.**

6. R has many built in datasets. One of the most popular is the iris dataset. In fact, the iris dataset is used so often for example calculations that working with it is considered a rite of passage for all data scientists. (For this problem you may want to use the str() function.)

a) How many observations and how many variables are in the iris dataset?

b) What data types are in the iris data set?

c) Find the sixth observation from the top of the data frame.

d) Find the tenth observation from the bottom of the data frame.

e) Find the sum of the Petal.Length.

7. What are the shortest and longest Sepal.Width? (You may want to use the order[] function.)

8. The interquartile range (IQR) is a measure of the spread in data. The IQR is defined as the third quartile minus the first quartile: . Find the IQR for Petal.Legth.

9. A rough estimate of petal area can be attained by taking the Petal.Width times the Petal.Length. Create the variable Petal.Area which contains the area of each petal represented in the dataset.

10. For this problem you will need to gather some demographic data from five friend and/or family members. Collect data on the person’s height, eye color, and number of pets.

a) Create four vectors: name, height, eye.color, and pets, and populate the vectors with the data you have collected.

b) Bind these vectors together in a date frame named myPeople.

c) Find the total number of pets and the range of the number of pets.

d) Find the mean and median height of the persons in the dataset.

e) Determine the number of levels of eye color in the dataset.