ANA 515 Assignment 2, Loading, Saving, and Describing Data

Due end of week 2

Points = 65

Using the dataset that you found, create an R Markdown document with the content listed below.

Be sure to set the echo option to TRUE so all of your code will also print out to the output document of your choice. Review this resource for more on R Markdown: <https://rmarkdown.rstudio.com/lesson-1.html>.

Put the following information in your R Markdown file:

**Section 1:** Description of the data. (10 points)

Describe the dataset you are using, both in terms of the content (what is this data measuring? how was it collected? what kinds of research questions are you hoping to use it to answer?) and in terms of its format (what type of file is it saved in? what if it is in a flat file, is it fixed width or delimited? if it is delimited, what is the delimiter? if it is binary, what is the program that would normally be used to open it?).

This will be a paragraph in your output file.

**Section 2**: Reading the data into R. (5 points)

Include code that reads the data into R and assigns it to a dataframe object that you can use later in the document. This will be your first code chunk. Explain with a #comment which R function you used to read in the data (e.g., read\_csv) and which package it came from (if it was not a base R function).

If there were any special options you needed to use (e.g., skip to skip some rows without data), list those and explain why you used them.

**Section 3:** Clean the data. (10 points)

Include some code to clean the data (e.g., rename columns, convert any dates into a “Date” format), **at least one cleaning function**. You can filter to certain rows if you would like, but do not filter out missing values, as we’ll want to learn more about those later. You could also choose to subset/filter your dataset to have fewer columns (must keep at least 4 columns).

**Section 4:** Characteristics of the data. (20 points)

Use inline code to put a sentence in your output document that reads: (5/20 points)

This dataframe has ... rows and ... columns. The names of the columns and a brief description of each are in the table below:

Include a table (using Markdown directly or kable from the knitr package) with 2 columns: (15/20 points)

1. the column name in the dataframe
2. a very brief description of what each column measures

**Section 5**: Summary statistics. (20 points)

Pick three columns of the dataframe. You can either create a subset or filter. (5/20 points)

Use a summary function to get the following summaries of these columns: (10/20 points)

(1) minimum value

(2) maximum value

(3) mean value

(4) number of missing values

If there are missing values, make sure you use the appropriate options in summarizing these values to exclude those when calculating the minimum, maximum, and mean.

Assign the result of this summary to a new object (<- give it a name), and make sure the summary shows up in your final, rendered output document. (5/20 points)