Connor StarrHurst

connor.starrhurst@wsu.edu

11596221

STATS 419

Instructor: Monte J. Shaffer

9/6/2020

Read [IRIS](https://www.dropbox.com/sh/8fo64qzdoz3q4i1/AAC86zOqEQyshGjCb6gbq0vba/unit_01_exploratory_data_analysis/week_02/02_IRIS.pdf?dl=0) ([02\_datasets.txt](https://www.dropbox.com/sh/8fo64qzdoz3q4i1/AACQIPnsR6C1qS4zu9swNp8Ba/unit_01_exploratory_data_analysis/week_02/02_datasets.txt?dl=0))

1. Create the "rotate matrix" functions as described in lectures. Apply to the example "myMatrix".

myMatrix = matrix ( c (

1, 0, 2,

0, 3, 0,

4, 0, 5

), nrow=3, byrow=T);

transposeMatrix = function(mat)

{

t(mat);

}

#rotateMatrix90(mat)

#rotateMatrix180(mat)

#rotateMatrix270(mat)

# 3x3 matrix ... ## matrix multiplication

See IRIS Code.txt

2. Recreate the graphic for the IRIS Data Set using R. Same titles, same scales, same colors. See: https://en.wikipedia.org/wiki/Iris\_flower\_data\_set#/media/File:Iris\_dataset\_scatterplot.svg

See IRIS Code.txt

3. Write 2-3 sentences concisely defining the IRIS Data Set. Maybe search KAGGLE for a nice template. Be certain the final write up are your own sentences (make certain you modify what you find, make it your own, but also cite where you got your ideas from). NOTE: Watch the video, Figure 8 has a +5 EASTER EGG.

One of the first things I noticed when looking at the data is that the virginica species is overall a larger plant. It has the largest petals (in length and width) and longest sepals overall. Conversely, the setosa species has the smallest petals (in length and width) with the shortest sepal length.

4. Import "personality-raw.txt" into R. Remove the V00 column. Create two new columns from the current column "date\_test": year and week. Stack Overflow may help: https://stackoverflow.com/questions/22439540/how-to-get-week-numbers-from-dates ... Sort the new data frame by YEAR, WEEK so the newest tests are first ... The newest tests (e.g., 2020 or 2019) are at the top of the data frame. Then remove duplicates using the unique function based on the column "md5\_email". Save the data frame in the same "pipe-delimited format" ( | is a pipe ) with the headers. You will keep the new data frame as "personality-clean.txt" for future work (you will not upload it at this time). In the homework, for this task, report how many records your raw dataset had and how many records your clean dataset has.

The raw dataset had 838 records and my clean dataset has 678 records.

5. Write functions for doSummary and sampleVariance and doMode ... test these functions in your homework on the "monte.shaffer@gmail.com" record from the clean dataset. Report your findings. For this "monte.shaffer@gmail.com" record, also create z-scores. Plot(x,y) where x is the raw scores for "monte.shaffer@gmail.com" and y is the z-scores from those raw scores. Include the plot in your assignment, and write 2 sentences describing what pattern you are seeing and why this pattern is present.

See IRIS Code.txt

Sentences.

6. Compare Will Smith and Denzel Washington. [See 03\_n greater 1-v2.txt for the necessary functions and will-vs-denzel.txt for some sample code and in DROPBOX: \\_\_student\_access\_\_\unit\_01\_exploratory\_data\_analysis\week\_02\imdb-example.  
You will have to create a new variable $millions.2000 that converts each movie's $millions based on the $year of the movie, so all dollars are in the same time frame. You will need inflation data from about 1980-2020 to make this work.

7. Build side-by-side box plots on several of the variables (including #6) to compare the two movie stars. After each box plot, write 2+ sentence describing what you are seeing, and what conclusions you can logically make. You will need to review what the box plot is showing with the box portion, the divider in the box, and the whiskers.