HUDM 5026 - Introduction to Data Analysis and Graphics in R
HW 06 – Multivariate Plotting and More on Data Transformations

Instructions.

- Use R Markdown to create an html document with the homework tasks.
- You are encouraged to discuss problems with classmates, but all work you submit must be your own.
- As always, any plots should have appropriate axis and overall labels.

PART I (MULTIVARIATE PLOTTING)

1. Use `read_csv()` from **tidyverse** to import the data file named "decath.csv", which you can find in the homework folder at the Files section of Canvas. These data include the scores for the 33 competitors in the men's decathlon at the 1988 Olympic Games. The events are 100 meters (100), long jump (long), shotput (poid), high jump (haut), 400 meters (400), 110-meter hurdles (110), discus throw (disq), pole vault (perc), javelin (jave) and 1500 meters (1500)
    - Is it a tibble? Describe how you can tell? (Hint, after loading tidyverse, check out `is.tibble()`.
    - Use `print()` to print out the entire data set. (Hint, see arguments `n =` and `width =` .)
2. Use `apply()` to calculate the mean and standard deviation for scores for each event.
3. Use `cov()` and `cor()` to calculate sample covariance and correlation matrices for the `decath` data. Round the correlation matrix to two decimal places and discuss any pairwise correlations that have absolute value larger than 0.5. When I say to discuss them, I mean you should look at the two events and discuss why it does or does not make sense that they would be strongly positively or negatively related.
4. Install package **corrplot**. Then, research and use function `corrplot()` to visualize the correlation matrix with ordering based on hierarchical clustering (i.e., with argument `order = "hclust"`). Describe any groupings of events and discuss why you think those events are intercorrelated. Does it make sense or not and why?
5. Use base R to create the following four plots.
    (a) scatterplot of discus vs shot put
    (b) scatterplot of shot put vs javelin
    (c) scatterplot of discus vs javelin

(d) histogram of shot put scores

6. Use `mfrow` or `mfcol` to create a two-by-two plotting array and then plot all four of the above plots in one window.

7. Use function `ggMarginal` in package **ggExtra** to create a scatterplot of discus vs shot put with marginal density curves.

PART II (MORE ON DATA TRANSFORMATIONS)

1. Import or load the acupuncture data we used last week and, once again, call it `acu`.

2. Create a new version of the data called `acu2` that are sorted by treatment group, sex, and age, in that order, so that the group variable has all 0s first and all 1s next, the sex variable is sorted 0/1 within levels of the group variable, and age is sorted last within combinations of the two categorical variables.

3. Create a subset of the original data called `acu3` that only includes participants who were in the acupuncture group and were over 30 years of age.

4. Create a subset of the original data called `acu4` that only includes participants who had baseline severity at or above the 40th percentile of baseline severity scores.

5. Use `which()` to note which rows of the original data correspond with participants that are female and 40 years of age or older.