WEB SCRAPING AND APIS

There are now five rules that apply to all projects so far:

a) Follow instructions *precisely*. If I do not tell you what to write on a particular line, leave it blank.
b) Do not use any functions or approaches to problems that we have not yet learned in this course.
c) All code must be *scalable by sample size* unless specifically noted otherwise.
d) Any code using *magrittr* should contain a max of one verb per line.
e) Always use the standard heading set up to this point unless otherwise specified: **R Studio API Code, Libraries, Data Import and Cleaning, Analysis, Visualization**

Part 1 – Set up a new R Studio Project

Part 2 – Facebook API Page Comparisons

1. **week11-facebook.R**
    a. Download per-post like data from the most recent 200 posts each from the Facebook pages of 1) the Society for Industrial and Organizational Psychology and also 2) the Society for Personality and Social Psychology.
    b. Create a tibble called *facebook_tbl* with each of the 400 like counts as rows, one column representing data source (SIOP vs. SPSP) and the other representing like count. Thus you should ultimately have a 400 row x 2 column table.
    c. Conduct an independent-samples t-test comparing the number of likes per post across the two pages, and calculate a d-value summarizing the effect using **cohen.d()** from *effsize.*
    d. Create a single plot displaying both density distributions (violin plots) and jitterplots, displaying the two groups side-by-side within that plot.
    e. *Note 1*. Just copy-paste in your Facebook token. I will replace it with mine when I test your code.
    f. *Note 2*. Do not copy/paste any group id numbers; grab the numbers you need by line number and column name.
    g. *Hint*. You might not want to use the tibble to run the t-test; there are easier ways.

Part 3 – Google Scholar Web Scrape

2. **week11-scholar.R**
    a. There is no API available to scrape Google Scholar, so you'll need to do this by hand. Choose someone's Google Scholar page – it does not matter who you choose, as long as they have at least 10 papers listed. Using R code alone, read this page and create a single tibble called *profile_tbl* containing columns representing the following information: name of article, author list, year, and citation count. Thus, a person with 20 papers listed should have 80 pieces of data: 20 rows x 4 columns.
    b. Calculate a correlation between year and citation count.
    c. Finally, create a scatterplot showing this relationship, superimposing a regression line.
    d. *Note.* Don't worry about anything past their first 20 citations.