COMPREHENSIVE PROJECT

There are four rules that apply to this assignment:

a) Follow instructions *precisely*.
b) All code must be *scalable by sample size* unless specifically noted otherwise.
c) Use standard headings: **R Studio API Code, Libraries, Data Import and Cleaning, Analysis, Visualization.** Delete any headings without code under them.
d) All code *must be refactored* to avoid future code rot (i.e., minimize unnecessary code complexity).

Project A

Your supervisor is interested in doing an NLP study using Twitter data but is not sure which hashtag would be best for analysis. Develop an app in a folder called **projectA**. Once completed, include a URL to the live app in a comment. Ensure your web application has the following characteristics:

1. Pulls the last 500 tweets from the hashtags #COVID, #COVID19, #COVID−19, and #COVID_19. Do **NOT** do this interactively; these tweets should be pulled **one time** for each person who opens your application (i.e., tweets should be pulled fresh from Twitter, but only once per user).
2. Allows the user to select between these four hashtags, and then for whichever hashtag the user selects, displays a single word cloud containing its fifty most common n-gram lemmas.
3. In a summary table, displays how many tokens with over 5 mentions overlap between each pair of hashtags. For example, if "scary" appears 10 times for COVID, 7 times for COVID-19, and 3 times for COVID_19, this token should be counted in the COVID-COVID-19 comparison only. This summary table should thus have 6 rows of comparisons and one column of summary statistics.
4. In a summary bar chart, show the 20 tokens most often appearing across all four hashtags (e.g., if "health" appears 10, 20, 15, and 1 time across the four hashtags, the number 46 would be used to determine placement on this chart).

Project B

You are interested in seeing what research has been done in COVID19 in the field of psychology, so you decide to start searching on Google Scholar with this term: **"covid-19" source:psychology**.

1. In a file **projectB.R**, scrape **all** results that come up from this search, across all pages.
2. Convert these results into a data frame containing four columns: article titles, author lists, journal title, year, and a link to each article. Some journal titles will be partial or missing, which is okay.
3. Display a figure showing publication counts per year in the ten most popular outlets.

Project C

In any dataset of your choosing, identify a prediction problem of your choosing involving psychological variables **that would benefit from machine learning** versus the use of ordinary least squares (including logistic) regression. In **projectC.Rmd**, import the data, clean the data, and run the analysis necessary to address this prediction problem. Include comments in markdown explaining where the data came from, why you chose this dataset, what question you intended to answer, why machine learning is appropriate in this situation, why you chose the analytic strategy you did, what you found, and why you interpreted it that way.