STRING MANIPULATION

There are four 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. This means your code should work equally well on a dataset with N=10 as N=1000.
d) Any code using *magrittr* should contain a max of one verb per line.

This week you'll be working with a list containing every unique citation made by every I/O psychologist currently working in a research institution in the United States in any article or book chapter they've ever published.

Part 1 – Set up a new R Studio Project with one R script called week6.R

Part 2 – Data Import and Cleaning

1. **Lines 1-3:** Write a comment that says: **R Studio API Code**, and set the wd as usual.
2. **Line 5:** Write a comment that says: **Data Import**
3. **Line 6**: Import any library you need for functions in the Data Import section
4. **Line 7:** Convert the text file citations.txt into a vector of strings called *citations.*
5. **Line 8:** Create a new vector called *citations_txt* containing only non-blank lines from *citations.*
6. **Line 9:** Using *citations* and *citations_txt,* print to console the number of blank lines eliminated.
7. **Line 11:** Write a comment that says: **Data Cleaning**
8. **Line 12-14:** Import any additional libraries you need for functions in the Data Cleaning section
9. **Line 15:** Using the *sample*() function, display a random draw of 10 citations to the console. You may want to run this command several times to get a sense of what cites look like in this file.
10. **Line 16:** Using an appropriate library and function, create a new tibble called *citations_tbl* containing two columns: *line* which indicates line number in the source file and *cite* which contains the citation text you just imported. Begin a series of pipes with this command. For all remaining code, write functions within that pipe to capture information **assuming correct APA style**. For each line, also convert extracted text to an appropriate type.
11. **Line 17:** Remove all quotations marks, including double and single, from all citations.
12. **Line 18:** Create a new variable called *year* that contains the year of publication.
13. **Line 19:** Create a new variable called *page_start* that contains the first page of each citation.
14. **Line 20:** Create a new variable called *perf_ref* that contains TRUE for any citation in which there is a reference to the word "performance", regardless of capitalization, and FALSE when there is not.
15. **Line 21:** Create a new variable called *title* that contains the citation title.
16. **Line 22-as many needed:** Create a new variable called *first_author* that contains the last name and any initials of the first author of each citation. Be careful with the number of initials and match regardless of any extra or missing spaces. Also, you can ignore hyphenated names. This is more complicated than the previous matching, so you might want to split it across lines.

Part 4 – Submission