**Data Science for Social Scientists**

Psyc 546, Spring 2023

Homework Assignment 7

**Due Date**: March 22nd (by 8:15 PM)

**Reminder**: See the assigned readings, Canvas resources, and the lecture slides for a tutorial on how to use R to perform the various functions included in the homework assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Homework 7 – Submission Portal on Canvas**.

Your submitted R script file should contain code to answer all of the questions below. Please use comments (e.g., #Question 1) to label the code for each question.

1. Create the following character string. Then, add special characters to the string so that a blank line separates the two statements. That is, your final answer should still just be a single character string object, but special characters should be added so that when the string is written out to the console, there should be a blank line between the two statements. [1 point]



1. Import the names\_data.csv file from Canvas. Then, using a stringr function, create a new variable in the data frame called full\_name that combines the last name column with the first name column. A comma should separate the two names in the full\_name variable. [1 point]
2. First, utilize a stringr function that is able to count the number of people in the names\_data file that live in a state that starts with the letter N. Then, construct a line of code that calculates the proportion of the people in the data file that live in a state that starts with the letter N. [1 point]
3. Utilize a stringr function that is able to count the number of people in the names\_data file that have a phone number ending in either 7, 8, or 9. [1 point]
4. Use the janeaustenr package for this question. First, assign all of Jane’s books to an object except “Emma”. That is filter out the book Emma but keep the other five. Then, make the object in tidytext format by unnesting it based on words and removing stop words. Finally, use ggplot2 to create a geom\_col() figure of the most frequent words and paste the figure below. The plot should consist of the following elements: [2 points]
   1. Only words with a frequency count above 400 should be included
   2. The words should be reordered so that the most frequent word is the top horizontal bar
   3. The x-axis is labelled “Frequency Count”
   4. The y-axis is labelled “Most Frequent Words”
   5. The following title is included: “Text Mining Jane Austen (Minus Emma)”
5. Question 6 uses “sentences” that loads up along with the stringr package. First, coerce “sentences” to be a data frame and then assign it to an object. Then, create a new column/variable in the data frame named “char\_count” that counts the number of characters in each sentence. Finally, perform a descriptive analysis of your choice to find the following values for char\_count: [2 points]
   1. Mean: 39.38
   2. SD: 4.86
   3. Median absolute deviation: 4.45
   4. Min: 26
   5. Max: 57
6. Use the gutenbergr package to download “The Great Gatsby” by F. Scott Fitzgerald. For reference, it has an ID of 64317. Assign it to an object called great\_gatsby. Then, make it in tidytext format by unnesting it based on words and removing stop words from it.

Using the “bing” sentiment lexicon, join together the great\_gatsby tidytext object with the bing lexicon. Make sure that this join method gets applied to the great\_gatsby object so that it should have a third column with the sentiment values (i.e., positive, negative). Finally, create a pie chart that visualizes the amount of negative vs. positive words in the book. Make sure the main title for the pie chart is “Great Gatsby Sentiments”. Paste the figure below. [2 points]

Chart, pie chart

Description automatically generated