Assignment 5

STA 141A

Due **November 20 at 10:30am**.

Assemble your answers into a report. Please do not include any raw R output. Instead, present your results as neatly formatted tables or graphics, and write something about each one. You must **cite your sources**. Your report should be **no more than 10 pages** including graphics, but excluding code and citations. Think carefully about what information is important to include.

When you are finished, submit a digital copy on Canvas. The digital copy must contain:

- Your report as a PDF file, with your code in the appendix.
- Your code as 1 or more R scripts.

Your submission will be graded according to the STA 141A grading standards, which are available on Canvas.

Description

In this assignment, you'll clean and extract features from a messy, expanded version of the Craigslist data set. In the messy data set, each Craigslist post is in a separate text file. Use a text editor to examine the files and determine the format. Some good text editors are:

- Windows: Notepad++
- OS X: TextEdit, TextWrangler, TextMate
- All: Visual Studio Code, Sublime Text, SciTE
- All (advanced): Vim, NeoVim, Emacs

The messy Craigslist data set is available as `messy_cl.zip` on Canvas. **Get started early. Ask questions on Piazza and in office hours to get help.**

Potentially useful functions/packages:

- `readLines`, `lapply`, `sapply`, `list.files`
- Package `stringr`, especially `str_trim`, `str_split_fixed`, `str_detect`, `str_match`, `str_replace_all`, `str_squish`, `str_remove_all`
- Package `lubridate`

Questions

1. Write a function `read_post` that reads a single Craigslist post from a text file. Your function should have a parameter `file` that controls which file is loaded. Your function may also have other parameters as you see fit. In your code, use comments to document the parameter(s) and return value of your function. Read Question 2 and 3 before starting this question. You do not need to write anything in your report for this question.

2. Write a function `read_all_posts` that uses `read_post` (from Question 1) to read all information from all posts in a directory and return them in a single data frame. Make sure the columns of the data frame have appropriate R types. Your function should have a parameter `directory` that controls which directory the posts are loaded from. Your function may also have other parameters as you see fit. In your code, use comments to document the parameter(s) and return value of your function. You do not need to write anything in your report for this question.

3. In your report, briefly discuss the design of `read_post` and `read_all_posts`: How does your `read_all_posts` function use your `read_posts` function? Did you have to make any design changes to `read_post` so that it works well with `read_all_posts`? How did you choose what the rows and columns should be in the data frame returned by `read_all_posts`? Is your choice of rows and columns convenient for further string processing? You may want to answer this last part after attempting Questions 4-8.

4. Extract the rental price from the title of each Craigslist post. Do all of the titles have prices? How do these prices compare to the user-specified prices (the price attribute)?

5. Extract the deposit amount from the text of each Craigslist post. Is there a relationship between rental price and deposit amount?

6. Extract a categorical feature from each Craigslist post (any part) that measures whether the apartment allows pets: cats, dogs, both, or none. Are there any apartments that allow some other kind of pet? For apartments that allow pets, make a graphic that shows how pet deposits are distributed and discuss what the graphic suggests about pet deposits.

7. Extract a categorical feature from each Craigslist post that measures whether each apartment has some kind of heating: a heater, a fireplace (including wood-burning stoves), both, or neither of these. Also extract a categorical feature from each Craigslist post that measures whether each apartment has air conditioning. Is air conditioning more common than heating? Do apartments with air conditioning typically have heating? Do apartments with heating typically have air conditioning?

8. Craigslist has an optional feature to hide email addresses and phone numbers from web scrapers like the one that scraped this data set. Do people seem to use this feature? How can you tell? Explain your strategy and your conclusions.