Final Project - DS 710

The final project for this class is your opportunity to apply what you have learned in this course to answer a question that interests you, by collecting and analyzing real-world data from Twitter.

What you'll be submitting

Five things, to GitHub via a pull request.

1) A **1 to 2-page** executive summary which reports your question, analysis, and results non-technically.

- It must include at least 1 figure, which may be embedded with the text or included on a second page. This figure must have been generated in R or Python by you, and the code should appear in your submitted code files.
- In .docx or .pdf format -- pdf is strongly preferred. Submit no other format for the summary.

2) A Python notebook containing the Python code you used to gather data from Twitter, and process it for analysis in R.

- Do not include your consumer key, consumer secret, access token, or access secret.
- This should be a clean, commented, final version of the code. It must run correctly from top to bottom with no errors, and have sane run counts for cells. Make it pretty!

3) A .csv or .txt file containing the data from 100 tweets, along with any variables you computed about the tweets (such as sentiment score or number of exclamation points).

- You should do a sanity check on this output. For example, if you searched for tweets containing the phrase "data science", there should not be any tweets with the word "data" but no "science." If you excluded tweets by @McDonalds, there should not be any tweets by @McDonalds.
- The point of this file is to allow us to do a sanity check on your data and the variables you computed.

4) A .csv or .txt file containing your parsed data for analysis in R.

- The point of this file is to allow us to check that the format of your output file matches what your Python code produces, and what your R code reads in. So, if you have lots of data (that's great!), we encourage you to truncate this file to the first 1000 lines of data, rather than creating a .zip file with all of the data.
- If the data you read into R matched the format of your data file for part 3, you may omit this.

5) An R script containing the R code you used to analyze the data from Python.

- This should be a clean, commented, final version of the code.
- You are encouraged to submit a .pdf file generated from R Markdown. Alternatively, a .r file is also acceptable.
- R output from hypothesis tests should be included as results in your .pdf file generated from R Markdown, or as comments in your .r file.

Submit your project to GitHub.

Notes

- Read the feedback on your final project proposal.

A detailed checklist for the project is available on the next three pages.

Executive Summary (40 points)

	What to include
Length	• 1-2 pages
Introduction	• Clearly explains the question of interest, and why/to whom it is interesting.
Data Collection and Analysis	• Clearly explains what keywords/features used to collect data, and why these keywords/features are appropriate to address the question of interest. • States when data were collected, and whether the REST or Streaming APIs (or both) were used. • Method(s) of analysis are appropriate to the question of interest and explained in a non-technical way. • Includes at least 1 hypothesis test, and the conclusion is explained correctly and in a non-technical way. ○ Does **not** include R output from hypothesis tests. That's too technical for an executive summary.

Figures	• Includes at least 1 graph which was made by you in R or Python.
	  ○ You may include tables if appropriate, but tables are not graphs.
	  ○ Pie charts and word clouds may be included, but they do not count toward the 1 graph minimum.
	  ○ If you are making your figure(s) in R, you are strongly encouraged to use ggformula.
	• Figures are appropriate to the data and question of interest.
	• Well-integrated with discussion of analysis and/or results. (For example, "As shown in Figure 1, …")
	• Legends or captions used appropriately.
	• Color used appropriately.
	• Font size and line widths chosen so that figures are legible when page is viewed at 100% Zoom.

Results/Conclusion	• Explains results clearly and accurately in a non-technical way. • Conclusion relates results to larger question or implications.

Writing Style	• Readable and interesting for a reader who does not know computer programming or statistics. ○ You can refer to technical topics (for example, "Using a t-test, I found strong evidence that…"), but don't get into the nitty-gritty here. • Professional spelling and grammar.

Data files (10 points)

	What to include
Parsed data file	• Data file is in a .csv or .txt format. • Format of data file is consistent with Python code (no editing by hand was necessary). • Format is consistent with R code (no editing by hand is necessary to run R code for this data file).
100 tweets file	• Includes username, text of tweet, any other variables you gathered that you used in your analysis, and any variables you computed. • Do a sanity check on your data. For example, if you searched for tweets containing the phrase "data science", there should not be any tweets with the word "data" but no "science." If you excluded tweets by @McDonalds, there should not be any tweets by @McDonalds. • Do a sanity check on the variables you computed. For example, if you counted exclamation points, the value of the number should match the number exclamation points displayed in the tweet. If you did a sentiment analysis, the tweets that are classified as having positive sentiment should (in general) be more positive than the tweets that are classified as having negative sentiment.

Python and R Code (50 points)

	What to include
Python and R Code	• Code is consistent with analyses described in the executive summary.

	• Clean, final version of code: When run by the reader, code produces no error messages, and all output is relevant to the analysis in the executive summary. • Evidence of complex thinking or problem-solving in both Python and R. • Functions created for effective task management AND/OR evidence of effort put into writing efficient code. • Comments used appropriately to make code readable. • R output from hypothesis test(s) is included as comments in the R code. • DOES NOT include consumer key, consumer secret, access token, or access secret.