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

Four things, to GitHub via a pull request.

1) A **1-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 your parsed data for analysis in R.

4) 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.
- It can be a .r file, or a .docx or .pdf file generated from R Markdown. R Markdown is super cool.
- R output from hypothesis tests should be included as comments.

Submit your project to GitHub.

Notes

- Large data files (over 50 MB or so) may be put in a zipped folder--but please don't put code files/executive summary there.
- Read the feedback on your final project proposal.
- Note that if you are doing a 1-sample test of proportions versus 50%, with no sentiment analysis or other analyses that go "above and beyond", you will be under a stronger burden of proof to convince us that their results are actionable.
- If you are gathering data about a company, please consider saving twitter usernames along with whatever other information you save (e.g. tweet text or hashtags). This way, if you want to filter out or focus on specific users in some way, you have the data. There's no real way to go back and repeat a search, since the week window for tweet availability is moving

A detailed checklist for the project is available on the next two pages

Executive Summary (40 points)

	What to include
Length	• Summary is one page (not including figures), with figures possibly pushing text onto a second page.
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.
Figures	• Includes at least 1 graph which was made by you in R or Python. (Note: You may include tables if appropriate, but tables are not graphs.) • Does **not** include R output from hypothesis tests. That's too technical for an executive summary. • 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.

Parsed Data file (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). • Do a sanity check on your data. If you say you searched for tweets containing the phrase “data science”, there should not be any tweets with the word “data” but no “science.”

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.