FINAL PROJECT INSTRUCTIONS

STAT/CS 287
JAMES BAGROW

- Please read and follow these instructions **completely**
- Attendance to all final talks is mandatory. Barring an emergency or previous arrangements, your entire final project **will not be graded** if you miss your presentation or anyone else's.
- Don't forget to come to office hours if you need help or have questions!
- Note: I will post ***followup instructions*** for preparing the slides for your final presentations.

Contents

1. Grading criteria

25% in each:

- Motivation/Purpose
  - Why is your project interesting and/or important. Work very hard to get this (citations to literature and published statistics are expected in order to successfully motivate your project)
- Structure/Organization of code/Data provenance:
  - Good filenames, good variables, clean/simple and well-documented code
  - Good record keeping
- Analysis Scope & Depth
  - How many **questions** did you ask, how thoroughly did you try to answer them
- Presentation & Appearance
  - Writeup quality: well organized, proper spelling and grammar
  - Good quality figures, well chosen ways to customize your data views for your project
  - Quality of **final presentation** including keeping to the time limit

A. Joining data

For the final project graduate students should consider it mandatory to combine 2+ datasets (barring arrangements made with me). By "joining" data I mean attempting to explain a feature in one dataset by comparing it with information in another dataset (almost always from a different source). For example, combining demographic information from census records with stock market data over similar time periods, or economic indicators with sports data.

- Graduate students: It will be effectively **impossible to get an A on your final project** if you do not go through the effort of joining disparate datasets.

Come to office hours if you need to discuss this.

2. Project organization and upload format

Please upload your project as a compressed (zipped) directory named `FINAL_NETID.zip` (example: `FINAL_JBAGROW.zip`.) This directory will contain the following files and directories:

1. `data/` and its contents - Folder storing the raw/original data
2. `stats/` and its contents - Folder containing files from intermediary analysis (optional)
3. `readme.txt` or `readme.pdf` - File describing all the other files in your submission
4. `ideas.txt`/`ideas.pdf`, `notes.txt`/`notes.pdf`, `etc`. - File(s) describing your research process, ideas you tried along the way, etc. Think of these as a logbook or journal. Should be described in the readme file.
5. Python `.py` files - Your download scripts, analysis, and figure-producing code, as appropriate. (*Show your work!*) Also described in the readme file.
6. `report_NETID.pdf` (example: `report_JBAGROW.pdf`)- **Your project report**. This is where 90-95% of your grade comes from. This is your most important file. Please upload a pdf, not a Word file.
7. `slides_NETID.pdf` (example: `slides_JBAGROW.pdf`). If you want to prepare your slides in Powerpoint or Keynote, please export them to pdf.

(Trailing `/`'s distinguish a directory name: `data/` is a directory named "data"; the `/` is not part of the name of the directory.)

In practice, you will have the freedom to organize your research according to your own personal preferences, but for these projects please follow the above standard. **Deviations from these requirements risk points loss** :(

Data and stats files should be included in their respective directories. However, if your data files are larger than 100 MB total when compressed, please remove them before zipping your upload (the `data` directory should be empty but still present). Keep these data files around as I may ask you for them in the future.

- Please do not use **absolute paths** in your code. For example: `open("/Users/jsmith/Documents/Course_work/Senior_year/Fall/DataScience/important_input.txt")`. This will not run on machines with different directory structures. Instead, set your Python **working path** to the appropriate location and open the file from there.

3. Report format and structure

Here is how to organize your **report**:

1. **Heading**: title, name, date
2. **Abstract**: 100-200 word summary of your project
3. Short **Data** or **Materials** section: describe dataset(s), examples of file formats, ... Basically tell me a brief story of how you got the data, processed it for bad values, merged multiple datasets together, etc.
4. **Results**. This is the bulk of the report, the story of your project. It should contain lots of text and figures with short blocks of code needed to generate plots, summary statistics, etc.
5. **Discussion** or **Conclusions**. Some 1-3 paragraphs summarizing what you've learned and especially what promising future work can be pursued with these data.

This layout (title, abstract, methods and materials, results, discussion) is the typical format for many (if not most) scientific papers!

The contents of this report are 90-95% of your project grade. As a written artifact, your report should be thorough but readable. I anticipate receiving formal, professional scientific or technical document, with **bibliographic citations**, references to supporting figures and tables ("Figure 4 shows that …") and so forth.

- Refer to specific parts of your code throughout your report. Example: "We found a significant trend between *x* and *y* ($p < 0.001$; `stat_tests.py, lines 100-104`) but not between *a* and *b* ($p > 0.05$; `stat_tests.py, lines 110-115`)." Please use a monospaced font (such as Courier) for all code references.

4. Presentation information

Instead of an exam during the assigned final exam period, each student will give **a brief 2-minute presentation** on his/her final project. (Teams of size n can speak for up to 2n minutes with all members taking turns to speak.) Include your slides with the final upload (see Sec. 2).

- Aim for approx. 4-5 slides detailing what motivated your research, what data you used and, briefly, any steps you took to clean or transform it, what your main discoveries from the research were, and what future directions are worth taking.
- **Talk order will not be alphabetical**, so don't expect to breeze in halfway through the final exam period because your name is Bob Zinser. Of course, Bob might as well not bother, because if he's late his final project won't be graded at all!
- Make sure you can finish your talk in two minutes. It is very unprofessional to run out of time, so you will be graded on this!

I will post ***followup instructions*** for preparing the slides for your final presentations.

The final project is a major part of your final course grade. Do not skimp on it! On the whole, aim for a **thorough** project and you will do fine.

- Avoid just downloading one CSV file and making 2-3 histograms, or not doing any data joining/merging - *You will have a bad time!*
- Consider running your text through a spell checker.
- There are many online sources for advice on giving talks well. Take the time to study these and rehearse if you need to. Keep the talk focused on accurately representing your final project and make sure you can finish in two minutes.