Statistics 222 - Statistics Master's Capstone
Spring 2014

Description:
This course is project-based and designed to expose students to practical data science in a project-based setting. The course will introduce frameworks for research code development and data analysis. The topics will include statistical issues and data preparation, tools to facilitate reproducible computational science, such as version control, data structuring, and scripting, as well as legal issues in research sharing. Students will be expected to present their work in class and turn in a final project.

Prerequisites:
A basic familiarity with statistical methods is assumed and computational experience at the level of Stats 243 Introduction to Statistical Computing, or equivalent.

Textbook:
There is no textbook for this class. We will use supplemental material that will be made available on bspace.

Lectures:
MW 9am – 11am in 340 Evans.

Lecturer:
Victoria Stodden, <victoria@stodden.net>
Office hours: Monday 1pm - 2pm in 309 Evans; Tuesday 11-1pm in 340 Evans

GSI:
Christine Ho <christineho@berkeley.edu>
Office Hours: TTh 9am - 11pm in 444 Evans

Grading:
There are three reports and two in-class presentations. The presentations will occur at midterm and at the end of the semester. Students may work in groups up to 4 students (each project will be evaluated with consideration given to the number of students in the group. For example a group of 4 students is expected to be 4x as productive as a student working alone on a project). The first report will present the datasets and research questions as a proposal, the midterm report and presentation will present initial completed results, and the final report and presentation will present the completed report.

We also require a weekly checkin in the form of a written status report to myself or the GSI to report on progress and troubleshoot on an ongoing basis. This will be due each Sunday evening before 12am.

research proposal	10%	due Monday February 3
midterm report	20%	due Wednesday March 19 (no extensions)
midterm presentation	10%	
peer review / weekly updates	10%	
final report	30%	due Wednesday May 7 (no extensions)
final presentation	20%	

Course Description:

We will introduce the practical aspects of data science including computational environments that include version control, provenance and workflow systems, and reproducibility of computational findings. Software includes GitHub, python, IPython notebooks, and LaTeX. We will also have guest speakers from time to time.

The main focus of the course is the final project. Students will not be expected to incorporate every topic covered in class, but rather use what is useful in producing reliable statistical analyses. The final project will be based on a dataset and seek to answer well-formed research questions. The work will be presented and written up into a final project report, including the code and data developed in the course of the research. Students may work alone on their final project, or may choose to work in groups of up to four students. The amount of work per student should be roughly the same regardless of the size of group you are in, and project grading will be adjusted accordingly.

Several research datasets are suggested for projects. If you would like to use a different one please discuss it with me before you turn in your preliminary report.

historical stock prices	http://finance.yahoo.com/q/hp?s=^GSPC+Historical+Prices
kiva.org microloan data	http://build.kiva.org/docs/data/snapshots
top reddit posts	https://github.com/umbrae/reddit-top-2.5-million
NFL Game Metadata	csv file
ResearchCompendia.org	(more on this in class)

I will discuss these datasets in more detail in class, as well as providing more information on what comprises a suitable research question or questions. Roughly, it should be challenging enough to be interesting and nonobvious, and should be proportionately ambitous relative to the number of members of your project group.

At all steps help will be available, on both statistical issues and technical issues, and on the final report. Feedback will be given on presentations.

Tips on final project (adapted in part from Gary King's "Publication, Publication" which is available on space):

1. Your final paper should address a substantive problem and contain one or a few clear points; one point with several supporting points is better than a lot of unrelated points. Your point should unambiguously answer the question: Whose mind are you going to change about what?

2. Unlike almost all previous papers you may have written, do not allocate space in your paper in proportion to how much work you put in accomplishing each task. The point of this paper is to make your scholarly point. This paper should not be about you or a report of what you did; it should be about what you contribute to our collective knowledge about the world. Space in your paper should be allocated in proportion to how much of a contribution it makes to changing the minds of someone in the literature about something important.

3. Papers should be no longer than about 20 pages per student (double-spaced, one-inch margins, 12pt, including figures, tables, and references). Think in terms of a short research note, not a full-length article. If you can do it in 10 pages, so much the better.

4. We provide a formal way to provide you some advice along the way: In class, you will turn in a very early draft of your paper with the tables and figures in near final form but relatively little text. You'll also turn in a "replication data set," (to be discussed) just as faculty routinely do. We will then give this to another student, who will try to replicate your results (without talking with you). That student will then write a memo about your paper. In science, we compete to advance knowledge about the world, not to tear each other down. Thus, the purpose is to improve the student's work.

Honor Code:

The student community at UC Berkeley has adopted the following Honor Code: "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." The hope and expectation is that you will adhere to this code.

Collaboration and Independence: Reviewing lecture and reading materials and studying for exams can be enjoyable and enriching things to do with fellow students. This is recommended. However, unless otherwise instructed, homework assignments are to be completed independently and materials submitted as homework should be the result of one's own independent work.

Cheating: A good lifetime strategy is always to act in such a way that no one would ever imagine that you would even consider cheating. Anyone caught cheating on in this course will receive a failing grade in the course and will also be reported to the University Center for Student Conduct.

Plagiarism: To copy text or ideas from another source without appropriate reference is plagiarism and will result in a failing grade for your assignment and usually further disciplinary action. For additional information on plagiarism and how to avoid it, see, for example: http://gsi.berkeley.edu/teachingguide/misconduct/prevent-plag.html

Academic Integrity and Ethics: Cheating on exams and plagiarism are two common examples of dis- honest, unethical behavior. Honesty and integrity are of great importance in all facets of life. They help to build a sense of self-confidence, and are key to building trust within relationships, whether personal or professional. There is no tolerance for dishonesty in the academic world, for it undermines what we are dedicated to doing - furthering knowledge for the benefit of humanity. Your experience as a student at UC Berkeley is hopefully fueled by passion for learning and replete with fulfilling activities. And we also appreciate that being a student may be stressful. There may be times when there is temptation to engage in some kind of cheating in order to improve a grade or otherwise advance your career. This could be as blatant as having someone else sit for you in an exam, or submitting a written assignment that has been copied from another source. And it could be as subtle as glancing at a fellow student?s exam when you are unsure of an answer to a question and are looking for some confirmation. One might do any of these things and potentially not get caught. However, if you cheat, no matter how much you may have learned in this class, you have failed to learn perhaps the most important lesson of all.