Syllabus for Statistics 222: Statistics MA Capstone Project University of California, Berkeley, Spring 2015

M/W 9-11 am Lecture, 340 Evans

Instructor: Dr. Cari Kaufman
Office: 315 Evans Hall
Office Hours: TBA
e-mail: cgk@stat.berkeley.edu

GSI: Jarrod Millman
Office Hours: TBA
email: millman@berkeley.edu

About This Course

In this course you will develop a portfolio of data analysis projects, culminating in a collaborative group project in which you'll work with an industry partner. The course is somewhat unique in that it's organized around data sets and not around a prespecified set of lecture topics. However, along the way, we'll cover a variety of methods used in modern applied statistics, we'll explore computational tools for working with non-standard data formats and doing reproducible research, and we'll practice the communication skills you'll need to be a successful practicing statistician.

Course Structure

In the first part of the course, everyone will be working on the same data analysis projects, although you will have more and more latitude for choosing research questions and methods as the course goes on. For each of these projects, you'll be assigned a team of 3-4 students. Although you'll each turn in a separate project report, you will be working together, as described below. In the last five weeks of the course, you'll again be working in a team, and you'll also have an industry partner who is providing the data set. For this project, each team will turn in one final report.

Here is how a typical two-week module will operate during the beginning of the course.

- First Monday: Introduction of the data set for this module and the main questions we want to answer. Lecture covering *some* of the methods or skills you'll need to answer these questions. Before Wednesday's class, you'll download the data and do some preliminary work with it.

- First Wednesday: Data debriefing. First with your teammates and then as a class, you'll discuss what you're finding with the data so far. What strategies are working or not working? What methods are appropriate for answering the questions we've posed with this data? What else do you need? In particular, we'll identify what additional topics (statistical or computational) you need to learn to carry out an effective analysis. Within your teams, you'll divide these up and assign research questions to each member of the team. (Example research questions: What is the EM algorithm? What is the basic functionality of the ggplot2 R package?)

- Second Monday: Presentation of topics identified last week, first with your team and then in a review as a class. I'll lead the class discussion, but you should be prepared to talk in front of the class. It'll be helpful for both the team and class discussions if you prepare some materials to help you present what you found. This could include notes, a few slides, code for an example, etc. For the rest of this week, you'll continue working on your data analysis.

- Second Wednesday: I'll wrap up with a lecture that covers any topics that haven't yet been adequately discussed so far. Your individual reports will be due the following Monday.

Here is our preliminary schedule:

	Mon	Wed	Data set
January		21	
	26	28	Randomized response survey of "study drug" usage
February	2	4	
	9	11	Airline on-time performance: a large SQL database
	16	18	
	23	25	Twitter (3 weeks; first week will be obtaining the data)
March	2	4	
	9	11	
	16	18	Choice of competition from Kaggle
April	30	1	
	6	8	Industry projects
	13	15	
	20	22	
	27	29	
May	4	6	(RRR week) Final reports due Friday May 8th

I'll have more to say about the final projects when we get closer. Briefly, you will continue working in teams, with two teams assigned to each of three industry partners. For this project, your team only needs to write *one* final report, not one for each team member, and the reports will be evaluated by the instructor with input from the industry partners.

Textbooks and Other Resources

There is no required textbook for this course. I will be pointing you to a lot of different resources as we go, and you will also be collecting them yourself and sharing them with the class.

Grading

Grading will be made up of three components: individual reports (50%), final report (40%), and participation (10%).

Reports: Your individual reports will be graded on a 0-5 scale by the GSI, with late reports automatically losing 1 point for each day late. I'll give you a rubric with more details about what we're expecting here. It's very important that you NOT turn in commented code as your report. Your report should be about the substantive aspects of the problem, the statistical methods you used, and your results. Annotated code should be submitted in a technical appendix.

Final reports: Your final reports will be graded on a 0-100 scale by the instructor. This will be similar in structure to the reports you've been turning in individually, but we'll be expecting more detail. Again, I'll give you a rubric ahead of time.

Participation: This will include an evaluation by the instructor as well as your team members after each project. We'll have a discussion about what your joint expectations are of each other as teammates.

Academic Integrity

Any work submitted by you and that bears your name is presumed to be your own original work that has not previously been submitted for credit in another course. You will work in teams on each data analysis project, but for these both the code and writeup must be your own. In particular, discussing your code with another student is acceptable, whereas simply giving him or her your own code is not. This is different from the final project, in which what you turn in will be jointly produced by the entire team, and sharing of code is encouraged. If you are not clear about the expectations for completing any particular assignment, be sure to seek clarification from the instructor. Any evidence of cheating or plagiarism will be subject to disciplinary action.