Course Syllabus

95-868: Exploring and Visualizing Data

Mini 3 2021

Instructor: David Choi

davidch@andrew.cmu.edu

Office hours: TBD

Course Description:

This course covers the fundamentals of statistical exploration and visualization of data. We will fit models and produce specialized graphs to explore data in a detailed and statistics-oriented manner. This course also serves as a crash course in R, a widely used statistical programming language.

In this class, students will learn:

1. How to use R to perform basic data tasks such as filtering, aggregating, and organizing data sets, and for production of graphics
2. How transformations, model fits, residuals, and simulation can be used to explore and check assumptions about data

Prerequisites

A first course in statistics is required, such as either 95-796 or 90-711.

Attendance

Attendance is not mandatory, but is strongly encouraged as having people in attendance usually leads to a better lecture. The lectures will be recorded for viewing.

We will have review sessions during weeks 1-4 of the class. They are strictly optional, and we will record them. During the review session, we will provide coding exercises that you can try, and then I'll discuss the solutions. (We will probably only record the part of the review session where I discuss the solutions)

How does this 95-868 compare to 94-842?

94-842 (R for Policy Analytics) is an excellent course which teaches R at a gentler pace than this one. Some students prefer to take 94-842 before this one (or just by itself), while others prefer to jump straight into this course. Both types of students have done equally well in the past. It really depends on whether you tend to get comfortable with new software tools quickly, or slowly.

There is about 1-2 weeks of overlap between the two courses at the beginning, and then they diverge. 94-842 teaches how to write careful and thoughtful reports, where the goal is to answer a predetermined question (probably using linear/logistic regression). 95-868, on the other hand, teaches how to explore data, which may suggest new questions that you would not have even thought of beforehand. This is sometimes called hypothesis generation. Often, you will discover that the data is complex in ways which go beyond the basic linear model.

In summary, 94-842 is a bit gentler in terms of teaching R, but in terms of teaching statistics, one is confirmatory while the other is exploratory: both are important, complementary, and may feel “advanced” in their own ways.

Textbooks

Helpful references for using R. These aren't necessary but might be helpful references for you in the future.

1. R for Data Science, by Hadley Wickham (try this one first, it’s the most modern)
2. R Graphics Cookbook, by Winston Chang
3. R for Everyone, by Jared Lander
4. R Cookbook, by Paul Teetor

Coursework, late policy, and grades

Your grade in this course will be based on 5 homework assignments and 1 mini-project.

The mini-project will be similar to a homework assignment, except that 1) the questions will be more open-ended, and 2) they will require knowledge from all weeks of the course, whereas each homework only covers the most recent lecture material.

HW: 70% Mini-project: 30%

For your homework assignments, each student has 5 late days. Each late day is a 24 hour extension. You may use them at your discretion, to cover travel for interviews, illness, or

general business. Otherwise, late homework will not be accepted. Email to let me know that you are using a late day.

Note 1: You may use more than one late day for a single assignment if you want to.

Note 2: late days cannot be used for the mini project

Grades will be curved according to Heinz college standards.

Email

Homework questions should be sent at least 24 hours before the deadline to assure a timely response.

Collaboration

You are encouraged to discuss general approaches and clarification questions with your fellow students. However, you should do your homework yourself.

Do not look at (or copy) another student's homework.

Do not copy from another student's homework.

If you receive any help from another student or outside the class (such as stackexchange or other forums or websites), you must clearly identify where you received help. The expectation is that your grade must reflect the work that you alone did.

Tentative Schedule

Week 1:
Mon: Introduction to R, Rstudio, and RMarkdown,
Wed: Data cleaning and aggregation
Fri: Review session
HW 1 released Wed, due in 1 week

Week 2:
Mon: Graphics part 1
Wed: Graphics part 2
Fri: Review session
HW 1 due Wed 6pm
HW 2 released Wed, due in 1 week

Week 3:
Mon: Averages and sample sizes, part 1 (p-values)
Wed: Averages and sample sizes, part 2 (confidence intervals)
Fri: Review session
HW 2 due Wed 6pm
HW 3 released Wed, due in 1 week

Week 4:
Mon: Univariate distributions part 1 (quantiles and QQ plots)
Wed: Univariate distributions, part 2 (residuals and transforms)
Fri: Review session
HW 3 due Wed 6pm

Week 5:
Mon: Functions of one variable, (splines and cross-validation)
Wed: Interactions and multivariate models, part 1 (interactions)
No Review Session
HW 4 released Mon, due in 1 week

Week 6:
Mon: Interactions and multivariate models, part 2 (stepwise variable selection)
Wed: Logistic regression and generalized linear models
No Review Session
HW 4 due Mon 6pm
Mini project released Mon, due in 9 days

Week 7:
Mon: Clustering and Heatmaps
Wed: Interactive graphics with shiny
Mini-project due Wed 6pm