Exploratory Data Analysis

Roger D. Peng

Course Description

This course covers the essential exploratory techniques for summarizing data. These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data. We will cover in detail the plotting systems in R as well as some of the basic principles of constructing data graphics. We will also cover some of the common multivariate statistical techniques used to visualize high-dimensional data.

Course Content

- Making exploratory graphs
- Principles of analytic graphics
- Plotting systems and graphics devices in R
- The base, lattice, and ggplot2 plotting systems in R
- Clustering methods
- Dimension reduction techniques

Lecture Materials

Lecture videos will be released weekly and will be available for the week and thereafter. You are welcome to view them at your convenience. Accompanying each video lecture will be a PDF copy of the slides and a link to an HTML5 version of the slides.

Course Textbook

The book *Exploratory Data Analysis with R* covers the lecture material in this course.

Assessments

Quizzes

There will be two quizzes. The quizzes will open on the first day of the course. Quiz 1 is due at the end of the first week, and Quiz 2 is due at the end of the second week. Please refer to the individual Quizzes to see the exact date and time that each Quiz is due. Quizzes are accessed by clicking the Quizzes link in the left navigation bar.

Course Projects

The two course projects will be assessed via peer assessment. In these projects, you will be asked to construct or reproduce certain plots, the purpose of which is to make you familiar with various plotting options. You will be evaluated on the plot that you produce and the code that you write to construct the plot. Course projects evaluated via peer assessment will make use of your GitHub account. Course Projects are accessed by clicking the Course Projects link in the left navigation bar.

Points and Scoring

- Quiz 1: 20%
- Quiz 2: 20%
- Course Project 1: 25%
- Course Project 2: 35%

You must earn 70% to pass the course and earn a certificate. Students who earn 90% and above will receive a certificate with Distinction.

Background lectures

Background lectures about the content of the course with respect to other quantitative courses, course logistics, and the R programming language are provided as reference material. It is not necessary to watch the videos to complete the course, however they may be useful for explaining background, the grading schemes used, and how to use R.

Quiz Scoring

You may attempt each quiz up to 3 times. The score of your most successful attempt will count toward your grade.

Hard deadlines and soft deadlines

The reported due date is the soft deadline for each quiz. You may turn in quizzes up to five days after the soft deadline. The hard deadline is **five days** after the Quiz is due at 23:30 UTC. Each day late will incur a 10% penalty; if you use a late day, the penalty will not be applied to that day.

Late Days for Quizzes

You are permitted 5 late days for quizzes in the course. If you use a late day, your quiz grade will not be affected. Late days may not be used for Course Projects.

Typos

- We are prone to a typo or two - please report them and we will try to update the notes accordingly.
- In some cases, the videos may still contain typos that have been fixed in the lecture notes. The lecture notes represent the most up-to-date version of the course material.

Differences of opinion

Keep in mind that currently data analysis is as much art as it is science - so we may have a difference of opinion - and that is ok! Please refrain from angry, sarcastic, or abusive comments on the message boards. Our goal is to create a supportive community that helps the learning of all students, from the most advanced to those who are just seeing this material for the first time.

Peer Assessment of Course Projects

For the course projects, peer assessment is necessary to evaluate the completion of the assignments. We have created and tested rubrics for each assignment. They are not perfect and will not be perfectly applied. However, we believe that the feedback from peer assessment adds value above simple multiple choice assessments.

- We have tried to make the criteria as objective as possible. Do your best to apply them to the best of your abilities.
- If you have questions or suggestions about the rubrics, please report them in the forum, "Rubric Issues".
- If you disagree with the scores you received through peer review, you may report those issues in the "Grading Issues" forum. Please note that it will be impossible for us to revise course project grades assigned via peer assessment, but we will attempt to use reports to improve future versions of the rubric.

Plagiarism

Johns Hopkins University defines plagiarism as "...taking for one's own use the words, ideas, concepts or data of another without proper attribution. Plagiarism includes both direct use or paraphrasing of the words, thoughts, or concepts of another without proper attribution." We take plagiarism very seriously, as does Johns Hopkins University.

We recognize that many students may not have a clear understanding of what plagiarism is or why it is wrong. Please see the following guide for more information on plagiarism:

http://www.jhsph.edu/academics/degree-programs/master-of-public-health/current-students/JHSPH-StudentReferencing_handbook.pdf

It is critically important that you give people/sources credit when you use their words or ideas. If you do not give proper credit -- particularly when quoting directly from a source -- you violate the trust of your fellow students.

The Coursera Honor code includes an explicit statement about plagiarism:

I will register for only one account. My answers to homework, quizzes and exams will be my own work (except for assignments that explicitly permit collaboration). I will not make solutions to homework, quizzes or exams available to anyone else. This includes both solutions written by me, as well as any official solutions provided by the course staff. I will not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.

Reporting plagiarism on course projects

One of the criteria in the project rubric focuses on plagiarism. Keep in mind that some components of the projects will be very similar across terms and so answers that appear similar may be honest coincidences. However, we would appreciate if you do a basic check for obvious plagiarism and report it during your peer assessment phase.

It is currently very difficult to prove or disprove a charge of plagiarism in the MOOC peer assessment setting. We are not in a position to evaluate whether or not a submission actually constitutes plagiarism, and we will not be able to entertain appeals or to alter any grades that have been assigned through the peer evaluation system.

But if you take the time to report suspected plagiarism, this will help us to understand the extent of the problem and work with Coursera to address critical issues with the current system.

Technical Information

Regardless of your platform (Windows or Mac) you will need a high-speed Internet connection in order to watch the videos on the Coursera web site. It is possible to download the video files and watch them on your computer rather than stream them from Coursera and this may be preferable for some of you.

Here is some platform-specific information:

Windows

The Coursera web site seems to work best with either the Chrome or the Firefox web browsers. In particular, you may run into trouble if you use Internet Explorer. The Chrome and Firefox browsers can be downloaded from:

- Chrome: http://www.google.com/chrome
- Firefox: http://www.mozilla.org

Mac

The Coursera site appears to work well with Safari, Chrome, or Firefox, so any of these browsers should be fine.

Created Tue 26 Feb 2013 10:18 AM KST

Last Modified Wed 24 Jun 2015 4:35 AM KST