Syllabus

Help Center

Course Title

Getting and Cleaning Data

Course Instructor(s)

Jeff Leek

Course Description

Before you can work with data you have to get some. This course will cover the basic ways that data can be obtained. The course will cover obtaining data from the web, from APIs, and from colleagues in various formats including raw text files, binary files, and databases. It will also cover the basics of data cleaning and how to make data tidy. Tidy data dramatically speed downstream data analysis tasks. The course will also cover the components of a complete data set including raw data, processing instructions, codebooks, and processed data. The course will cover the basics needed for collecting, cleaning, and sharing data.

Course Content

- Data collection
  - Raw files (.csv,.xlsx)
  - Databases (mySQL)
  - APIs
- Data formats
  - Flat files (.csv,.txt)
  - XML
  - JSON
- Making data tidy
- Distributing data
- Scripting for data cleaning

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.

Weekly quizzes

There are four weekly quizzes. You may begin submitting them as soon as the course opens. Quiz 1 is due at the end of the first week, Quiz 2 is due at the end of the second week, Quiz 3 is due at the end of the third week and Quiz 4 is due at the end of the fourth week. See the Quiz pages for exact due dates. The hard deadline is 5 days after the quiz due date.

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 from your most successful attempt will count toward your grade.

Hard deadlines and soft deadlines for Quizzes 1-3

The reported due date is the soft deadline for quizzes 1-3. You may turn in quizzes 1-3 up to five days after the soft deadline. Each day late will incur a 10% penalty, but if you use a late day, the penalty will not be applied to that day.

****Please Note: There is no grace period for Quiz 4. The hard deadline is firm. You must submit Quiz 4 BEFORE the hard deadline to receive credit****

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 the Course Project.

swirl Programming Assignment (optional)

In this course, you have the option to use the swirl R package to practice some of the concepts we cover in lectures.

Each lesson that you complete in swirl is worth one extra credit point. However, the **maximum number of points you may earn for the assignment is capped at 3**. While these lessons will give you valuable practice and you are encouraged to complete as many as possible, please note that they are **completely optional** and you can get full marks in the class without completing them.

You can find the instructions for how to install and use swirl in the Programming Assignments section of the course under *Week 1*.

The Course Project

The project is available from the first day that the class is open. You may submit the project at any time. It is due BEFORE 11:30 PM UTC on the Sunday that ends the third week of the class. Peer assessment opens immediately after Course Project submissions are due and runs during the fourth week. You are required to review four of your classmates projects. For exact due dates, please click the Course Project link in the left navigation bar. You may not apply Late Days to the Course Project. If you do not submit your work BEFORE the deadline, you will not receive credit or be able to participate in the evaluation phase.

Points and scoring

There are 100 available points for the course. They are broken down as follows

- Quiz 1 = 15 points
- Quiz 2 = 15 points
- Quiz 3 = 15 points
- Quiz 4 = 15 points
- Course project = 40 points

You must receive 70 points to pass the course and achieve the certificate. You must receive 90 points to achieve the certificate with distinction.

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 for Course Project

For many of the Data Science Specialization Course Projects, peer assessment is necessary to evaluate the completion of the assignments. We have created and tested rubrics for these assignments. 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 peer-grades, 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.â 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-ReferencingHandbook.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 â particularly when quoting directly from a source â 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 17 Sep 2014 2:49 AM KST