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.

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

Assessments

Quizzes

- There are four weekly quizzes.
- You must earn a grade of at least 80% to pass a quiz
- You may attempt each quiz up to 3 times in 8 hours.
- The score from your most successful attempt will count toward your final grade.

Course Project

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. You will be graded by your peers on a series of yes/no questions related to the project. You will be required to submit:

1. a tidy data set as described below
2. a link to a Github repository with your script for performing the analysis

3. a code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md.

Finally, you are required to review four of your classmates' projects. You must earn a grade of at least 80% to pass 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.

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

Grading Policy

You must score at least 80% on all graded assignments (Quizzes & Project) to pass the course.

Your final grade will be calculated as follows:

- Quiz 1 = 15%
- Quiz 2 = 15%
- Quiz 3 = 15%
- Quiz 4 = 15%
- Course project = 40%
- swirl Programming Assignment (optional) - 0%

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.

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 JHU referencing guide for more information on plagiarism.

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.