CS x415.1 Data Science Principles and Practice

Final Project

Introduction

The purpose of the final project is to exercise and demonstrate your skills on a data science project from start to finish, including:

1. Finding sources of data pertaining to a domain in data science (e.g., marketing, environmental, health care, education, web analytics, economics, etc.)
2. Loading data from different types of sources (e.g., csv, json, web scraping, database, web service api)
3. Transforming data into necessary (usually tidy) format. Dealing with data problems such as outliers, missing data, invalid data.
4. Linking data from various sources using joins
5. Exploratory Data Analysis
6. Modeling
7. Reporting results, including interactive, web-based shiny applications

Your final product (to be turned in) will be a 100% reproducible RMarkdown-generated HTML document, with all data sources available (e.g., as web urls).

You will define and choose your own final project: the dataset(s) you will be working on, defining the problem(s) you will be investigating.

Your grade will be based on how much you demonstrate of the above data science tasks, both breadth and depth (see below). It's your chance to show off what you've learned in this class, and if done right, can serve as a significant data science portfolio project.

Implementation

Not all problems involve significant tasks in all of the above areas. **You do not have to choose a project that involves all of the above in equal, significant depth**. But your project should include all of the above, and significant work in at least two or three of the above areas. It's up to you to decide what to focus on.

For example, baseball statistics are easily loaded into R using data from existing packages. There's not a lot involved with (1) and (2) above. However, does that mean that a baseball or similar project using prepackaged data is not adquate or even desirable? No! Such projects might focus on exploration, modeling, and reporting results. The same for different types of projects.

Get Started Now!

1. Start thinking about what you are interested in: domains and/or data science tasks (1-7 above).
2. I will start a discussion thread on data sources. Start your own investigations and contribute to the discussion
3. Start your project by defining
    1. A problem domain (health care, environmental, etc.)
    2. A problem
    3. One or more datasets that you will be working on
    4. Do some EDA to sharpen up your understanding of (2) above, what and how you plan to present your final results