Course Description

This course is about manipulating data to prepare it for analysis using Python tools. SQL and NoSQL technologies are referred to, to some extent and in basic way. It's expected that ElasticSearch will be the NoSQL store that we'll be using to provide access to some assignment data this term.

The data used for assignments varies in structure and volume. It includes poorly structured and user-generated content data. The data are stored in various formats, and are about things like customer transactions, air transportation, corporate email, and customer evaluations of hospitality experiences.

In this course students are required to "peer review" their classmates' contributions to the course. Students' peer reviews count towards their final grade in the course.

Course Perspective and Pedagogical Philosophy

This isn't a course about Python, SQL, or NoSQL, per se. It's about getting data into a desired condition to be analyzed. It's about formulating, implementing, and testing solutions to the problem of getting data from a current state to a required state. So, code, datasets, and storage technologies aside, it's about practical data preparation problem-solving.

In the opinion of many, a good way to learn many things is by *doing* them. The famous mathematician and educator Paul Halmos once said that to do is to know, and to talk is *not* to teach. I'll add to this that a book is not by itself a course. There are readings in this course, as there are in other MSPA courses. The readings provide some reference material and can fill in content gaps with respect to the course learning objectives. For some tasks in this course students may need to go beyond the required and recommended readings by looking elsewhere, like online. This is a key aspect of practical problem-solving, a critical skill for data scientists and for predictive analytics professionals.

In this course the problems to be solved consist of getting data from "State A" to "State B." They often need to be further define these problems so as to open up potential solutions, and ways to apply solutions using (usually) Python code that provides results. Students in this course are not provided with the code to do the assignments, or with complete solutions to them. General feedback, tips, and "hints" on assignments are provided, by me (your instructor), our TA, and also by your fellow 420 students as participants in this course's learning community.

Course Learning Goals

These are the common 420 course goals. They are achieved by completing the course readings and doing the assignments. You will be the ultimate, and most important, judge of whether you attained them.

- Define key terms, concepts and issues in data management and database management systems with respect to predictive modeling
- Evaluate the constraints, limitations and structure of data through data cleansing, preparation and exploratory analysis to create an analytical database.
- Use object-oriented scripting software for data preparation.