Data Manipulation and Analysis

Overview   This course introduces students to the emerging field of Data Science. Instructional units cover the standard practices for effective data manipulation, analysis and interpretation as well as necessary concepts in the three disciplines involved (mathematics, statistics and computing.) Numerous examples of typical scenarios are provided. The emphasis on this course is in the application of the concepts rather than the theory. In the second semester, students will work in teams on large projects in which they will use programming to analyze large datasets and create models. The students will summarize their findings for each project in a written report and will also present them orally.

Objectives

- Create models driven by large amounts of data
- Interpret graphs and tables, and make sense of data by visualizing it
- Create interactive presentations for visualizing data
- Automate routine data manipulation tasks, such as formatting, collection and storage
- Describe modeling techniques for clustering, regression and classification of data.
- Explain typical trade-offs that data scientists must make when analyzing data.
- Demonstrate effective communication skills, through team working, oral presentations and good written communication.

Assessment   In the first semester, formative assessment includes worksheets, several practice activities for each lesson, and unit quizzes. Summative assessments include small programming projects at the end of each unit. In the second semester, projects are assessed on accomplishment, originality, code sophistication, presentation and team work. Milestones are set for each project, so that teams receive feedback on progress before the final evaluation of each project.

Course Essentials

Equipment	Cost/Unit
Classroom set of computers	$0 if you already have some, $500-600 per computer if you need to purchase
Prerequisite	Intro to Computational Thinking (LSU Pathway)

Outline of First Semester

Unit 1: Review of programming	Syntax. Semantics. Conditions. Loops. Lists and other data structures. Vectorized operations.
Unit 2: Data preparation	Small data and big data. Sources of data. Variables, observations and values. Importing, cleaning and reshaping data. Dealing with missing values. Exporting data.
Unit 3: Data manipulation	Filtering and sorting rows and columns. Transforming values. Programmatically adding and modifying rows and columns. Grouping observations and summarizing data. Relational operations: joining data from multiple tables. Parsing data. Processing strings and dates.

Outline of Second Semester

Unit 4: Data visualization	Static graphs and charts. Interactive graphs. Visualization for exploration vs. visualization for presentation.
Unit 5: Exploratory data analysis	Students create projects that use descriptive statistics and visualization to explore large datasets of interest.
Unit 6: Modeling techniques	Students create projects where they use simple regression models to analyze large datasets of interest. An overview of clustering and classification techniques is included in the lessons, in order to allow advanced students to use those techniques in their projects.