Applied Data Science

Olga Scrivner

“You can have data without information, but you cannot have information without data.”
(Daniel Keys Moran, science fiction writer)

Course Description

The goal of this course is to develop practical skills needed to perform applied data science research. The course is organized around each stage of the data science workflow (setting expectations, exploratory data analysis, modeling, interpreting, and communicating results) and covers algorithms, best practices, and evaluation criteria. Both good and bad applications examples will be discussed to help the student develop a deeper understanding and intuition about the choice of algorithm or visualization for the data task, the development of the best practices, and the methods for evaluating results of different approaches. Lectures and reading will provide students with theoretical foundation for research and hands-on practice and assignments will help develop practical skills.

Required Reading


Weekly reading will be provided via Canvas.

Prerequisites

Prerequisites: R and Python Basics

Course Goals

- The students will learn about the theoretical framework of data science and its tools.
- The students will learn effective strategies for integrating, analyzing, and presenting data.
- The students will learn how to use questions to guide themselves through data acquisition, integration, analysis, and presentation using data visualization and story and how to talk about data.
- The students will be able to use the data science workflow as a cognitive model to assess what needs to be done and know how to proceed when taking on a new data assignment.
Course Learning Objectives

At the end of this course students will be able to:

- Apply quantitative modeling and data analysis techniques to the solution of real world problems, communicate findings, and effectively present results using data visualization techniques.
- Apply principles of Data Science to the analysis of problems.
- Use Python and R to solve real-world problems.
- Learn the fundamentals of data analytics and the data science pipeline
- Apply statistical methods, regression techniques, and machine learning algorithms to make sense out of data sets.

Course Structure

Weekly lectures, readings, quiz and hands-on R practice with DataCamp platform. The instructor will provide students with a free academic access to DataCamp. Students are required to obtain digital copies of a required books and do weekly reading.

Course Communication Channels

Piazza, Zoom, Canvas, DataCamp

Virtual Classroom

During the semester you are invited to participate in four virtual classroom [see syllabus and announcement for more details].

Grading Policy

No late assignments are accepted unless an official documentation is provided (e.g., medical documentation). Business trips, job interview, vacations are not considered as valid excuses.

- **30%** Midterm exam
- **5%** Proposal
- **20%** Final Project
- **10%** 2 Assignments
- **35%** Weekly Quizzes and Homework

Assessments

Each week will have a quiz based on lectures and reading or a short homework to assess learned skills. There will be two graded practices from DataCamp.
Final Project

Proposal (1 page, double spaced). First select a fairly challenging data set for a final project. You are required to select your own data. The proposal should include the following: (1) description of the dataset (provenance, size, structure); (2) specific hypotheses (questions); (3) plan for future EDA analysis and modeling for your final report.

The written final report (10 pages including figures, double spaced, APA style references) in a professional journal APA style that includes an introduction to the problem (hypotheses of interest), a brief description of the methods used to collect and clean the data and the measurements obtained, a result section summarizing the results of the analyses deemed appropriate for the final project, and a discussion section that relates the analyses to the hypotheses stated in the introduction. Also include references in APA style. You will decide which programming language to choose (R and/or Python). You are required to create a github project for your final paper where you provide 1) code: Python Jupyter notebook and/or R Markdown files, 2) figures, 3) data. Note screenshots of code are not allowed in the paper. Figures must follow APA style. Workshop on APA will be offered.

Assignments

2 hands-on assignments will be done using the DataCamp platform for interactive coding. You will be required to sign up for a free account on DataCamp using your IU email. For each completed practice you will be required to provide a statement of completion from DataCamp.

Student Integrity Policy

Students are expected to conduct themselves in a manner befitting their status as a student of a respected and distinguished institution of higher education. In college courses, we are continually engaged with other people’s ideas: we read them in texts, hear them in lecture, discuss them in class, and incorporate them into our own writing. As a result, it is very important that we give credit where it is due. Plagiarism is using others’ ideas and words without clearly acknowledging the source of that information. See http://www.indiana.edu/~wts/pamphlets/plagiarism.shtml for help in addressing plagiarism in your own work.

Students with Disabilities Accommodations

Arranging Accommodations: Once you have submitted a request for service, provided appropriate documentation verify your disability, and met with your DSS Coordinator, it is time to arrange your reasonable accommodations with your instructors.

Schedule

The schedule is tentative and subject to change.
<table>
<thead>
<tr>
<th>Week</th>
<th>Theory and Tools</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1 [Jan 7]</td>
<td>Data Science Workflow&lt;br&gt;<strong>Virtual Classroom</strong> - Jupyter Notebook and R Markdown</td>
<td>quiz</td>
</tr>
<tr>
<td>Week 2 [Jan 14]</td>
<td>Data: Structure, Types</td>
<td>quiz</td>
</tr>
<tr>
<td>Week 3 [Jan 22]</td>
<td>Data Cleaning (OpenRefine)</td>
<td>homework</td>
</tr>
<tr>
<td>Week 4 [Jan 28]</td>
<td>Data Wangling and Data Management</td>
<td>quiz</td>
</tr>
<tr>
<td>Week 5 [Feb 4]</td>
<td>Exploratory Data Analysis [practical stats]&lt;br&gt;<strong>Virtual Classroom</strong> - GitHub Desktop and RPubs</td>
<td>quiz</td>
</tr>
<tr>
<td>week 6 [Feb 11]</td>
<td>Exploratory Data Analysis [visual analysis, outliers]</td>
<td>homework</td>
</tr>
<tr>
<td>week 7 [Feb 18]</td>
<td>Modeling [regressions]</td>
<td>quiz</td>
</tr>
<tr>
<td>week 8 [Feb 25]</td>
<td>Models Evaluation</td>
<td>quiz</td>
</tr>
<tr>
<td>week 9 [March 4]</td>
<td>Social Web</td>
<td>mid-term practice &amp; test</td>
</tr>
<tr>
<td>week 10 [March 10]</td>
<td>Spring Break</td>
<td></td>
</tr>
<tr>
<td>week 11 [March 18]</td>
<td>Clustering and Dimension Reduction</td>
<td>Proposal - 1 page</td>
</tr>
<tr>
<td>week 12 [March 25]</td>
<td>Principles of Analytic Graphics&lt;br&gt;<strong>Virtual Classroom</strong> - Shiny Workshop</td>
<td>quiz</td>
</tr>
<tr>
<td>week 13[April 1]</td>
<td>Data and Code Management [YesWorkflow]</td>
<td>quiz</td>
</tr>
<tr>
<td>week 14 [April 8]</td>
<td>Case Studies&lt;br&gt;<strong>Virtual Classroom</strong> - APA style in DS</td>
<td>homework</td>
</tr>
<tr>
<td>week 15 [April 15]</td>
<td>Case Studies</td>
<td></td>
</tr>
<tr>
<td>week 16 [April 22]</td>
<td>Final project</td>
<td></td>
</tr>
<tr>
<td>Finals Week [April 29]</td>
<td></td>
<td>Final Project Due</td>
</tr>
</tbody>
</table>