DATA 115: Introduction to Data Analytics

Spring 2024

Instructor Information

Instructor: Sergey Lapin
Office: WSU Everett Room 405
Phone: 425-405-1780
Email: slapin(at)wsu.edu
Class Time: MWF, 11:10-12

Course Description

This course provides an introduction to the field of data analytics. As befits a rapidly developing, interdisciplinary subject, we will draw on recent and relevant materials from statistics, mathematics, and computer science, as well as many application domains. Motivated by natural questions that arise in simple data examples, we will cover many of the basic techniques for working with data including sourcing raw data, cleaning and processing, exploring and analyzing, and finally presenting conclusions. In order to provide a foundation for later courses, we will also explore initial examples of many of the core topics that will be encountered. You will have plenty of opportunities to work with real data and the R programming language.

In addition to familiarizing you with basic tools and methods, this course will provide a broad exposure to the diverse types of data analytics projects that are being conducted around the world. A key component of the course will be critically analyzing published data analytics works and discussing their strengths and shortcomings. Finally, as data driven practices are becoming common in many career fields, we will focus on professional development topics such as presentation skills and examples of the ethical and legal issues that can arise in modern data analysis projects.
There is no required textbook to purchase for this course, as we will be making use of open source and online materials all semester. The first three textbooks below are available as free .pdfs and will be referenced during lectures and assignments. You may find the texts on the remainder of the list to be useful additional supplementary resources but access to them will not be required. In addition to these texts, shorter weekly readings will be posted to the course Canvas page. Data sets and programming scripts for the course will be uploaded to Canvas.

- **Open resources:**
  - *Introduction to Statistical Learning* (James, Witten, Hastie, and Tibshirani)
  - *R for Data Science* (Wickham and Grolemund)
  - *ggplot2: elegant graphs for data analysis* (Wickham)

- **Additional resources:**
  - *Doing Data Science* (O'Neil and Schutt)
  - *Introduction to Statistical Investigations* (Tintle, Chance, Cobb, Rossman, Roy, Swanson, and VanderStoep)
  - *The Art of Data Science* (Peng and Matsui)
  - *Data Science from Scratch* (Grus)
  - *An Introduction to Data Science* (Stanton)
  - *The Visual Display of Quantitative Information* (Tufte)
  - *The Elements of Data Analytic Style* (Leek)
  - *Introduction to Probability* (Grinstead and Snell)
  - *Weapons of Math Destruction* (O'Neil)

---

**Software**

At the beginning of the course we will focus on manipulating and extracting details from data using spreadsheet programs like Excel. Later segments of the course will provide training and examples using the R programming language, which you will install on your own computer or access through a cloud platform. R is an open source language with a robust ecosystem of packages for data analytics that has become one of the most common languages for a broad range of data-related tasks\(^1\).

No previous programming experience or knowledge of statistical software tools will be assumed. We will start with the basics of R in Week 3 and build up familiarity with standard data analytics libraries like dplyr, ggplot2, knitr, and tidyr.

---

**Attendance and Participation**

Due to the pace of the course and the range of topics that we will cover this term, daily attendance will be essential for your success. Although it is not officially a part of the course grade, missing class could adversely affect your grade by impacting your understanding of the material.

---

\(^1\)In 2018, Kaggle surveyed 23,859 data scientists and found that R was second to Python in terms of most commonly used languages: [link](#).
We will use the Canvas forums for course discussions. This is a great place to ask questions from your peers, as well as to get feedback on your ideas. Announcements and other official communications will be posted on Canvas as well as sent to your official WSU email accounts. You should check these messages regularly to stay informed about upcoming due dates and updates to the syllabus.

I am accessible by email at slapin@wsu.edu. Please include “DATA 115” in the subject line for any messages concerning the course. I will commit to responding within 48 hours but this does mean that queries sent immediately before a deadline may not receive substantive responses in time to be directly helpful, so please plan ahead.

Students who successfully complete the course will be able to:

• Describe different types, uses, and structure of data sets
• Perform basic procedures to obtain, process (clean), and store data
• Understand and compute simple summary statistics and statistical models
• Construct simple scripts for processing, analyzing, and visualizing data
• Conduct exploratory data analysis
• Apply elementary (supervised and unsupervised) learning techniques
• Analyze published data analytics work across multiple application domains

Professional Preparation

– Work collaboratively on data analytics projects
– Present data preparation processes and the results of analyses
– Understand legal and ethical ramifications of data-driven projects
– Curate and store data sets
– Use common programming tools and computational platforms

The following questions (or more importantly, your ability to provide reasonable answers to them) represent a useful way for you to evaluate your progress in the course. Many of our weekly topics are devoted to providing context for these questions (adapted from a similar course taught at Denison University), in the sense that the material that is presented in lecture, as well the relevant assignments and readings, should prepare you to be able to successfully provide thoughtful and detailed responses.

1. What is data, what are the types of data, why are they important, and what can I do with them?
2. How do I acquire, store, and access data?
3. How can I clean data and put it into a usable format?
4. How should I handle missing data?
5. What should I do about outliers?
6. How can I visualize data for better understanding?
7. How do I summarize and report univariate data?
8. How can I extrapolate historical data into predictions about the future?
9. How can I extrapolate many variables at once?
10. How can I analyze social interactions?
11. How can I use data to sort objects into classes?
12. What are the limits to the conclusions that can be drawn from data?
13. What are the ethical, legal, and social considerations of data acquisition, storage, and analysis?
14. What are good standards for sharing code, maintaining data, and reporting results?
15. What are best practices for presenting quantitative results to audiences?

Assignments and Assessments

There will be four main types of graded assignments in this course.

- **Participation and in-class work:** In addition to participation in lecture and discussion, each week you will be responsible for completing one or more short readings on a relevant data analytics topic. We will discuss these readings in small groups during the Friday class meetings and each group will be required to write a short response summarizing the discussion or to complete a corresponding worksheet. Occasionally we will have short individual quizzes in the Friday meeting period which will also count in the participation grade.

- **Weekly Assignments:** Each Wednesday, a problem set will be assigned, covering the course material for the forthcoming week. These will usually be a mixture of direct questions about the lecture material and opportunities for you to apply the methods we discuss to real data. Individual responses to the assignment will be due at midnight the following Wednesday. No late work will be accepted. Written assignments must be submitted as .pdf files.

- **Midterm Exam:** There will be a single midterm exam during the 8th week of class, covering the material that we will have encountered to that point. This exam will be open notes and completed electronically through Canvas.

- **Personal Dataset and Final Project:** The final assessment in the course will be an individual or a small (max 3) group project, where each team will complete a data analytics task from beginning to end. Beginning in the first week of class, each student (or group) will be encouraged to begin gathering and curating a data set that is of particular interest to them, with the goal of having a polished repository by the end of the term. More details will be discussed later in the semester but this will provide you with flexibility to tackle a topic of broader depth than those encountered in the weekly assignments. In addition to completing a writeup of the data processing steps and conclusions, each group will give a presentation to the rest of the class describing their results and findings during the finals period. Groups and project topics will be finalized during Week 12 of the course.

Collaboration Policy
For the assignments you are encouraged (and sometimes required) to work with other students in the class. However, the work that you submit should be your own and in particular should be written in your own words and communicate your own understanding of the solution. If you do collaborate, please list the names of the other students you worked with on your submission. You may be asked to explain your work in person to obtain full credit. Obtaining solutions from external sources like chegg or coursehero for course problems will be considered a violation of the academic integrity policy with consequences described below.

**Grading Policy**

The breakdown of grade components and letter grades will be as follows² (the first weekly assignment will ask you to evaluate the visualizations ⊗):

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>10%</td>
</tr>
<tr>
<td>Weekly Assignments</td>
<td>30%</td>
</tr>
<tr>
<td>Midterm</td>
<td>20%</td>
</tr>
<tr>
<td>Final Project</td>
<td>40%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>Percentage</th>
<th>Grade</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>94-100</td>
<td>C+</td>
<td>77-79.99</td>
</tr>
<tr>
<td>A-</td>
<td>90-93.99</td>
<td>C</td>
<td>73-76.99</td>
</tr>
<tr>
<td>B+</td>
<td>87-89.99</td>
<td>C-</td>
<td>70-72.99</td>
</tr>
<tr>
<td>B</td>
<td>83-86.99</td>
<td>D</td>
<td>60-60.99</td>
</tr>
<tr>
<td>B-</td>
<td>80-82.99</td>
<td>F</td>
<td>0-59.99</td>
</tr>
</tbody>
</table>

²We reserve the right to be more lenient.
The following outline describes the preliminary plan for our class. An updated version will be posted on the course Canvas page and updated throughout the semester.

1. Week 1
   - Course outline and syllabus
   - Introduction to modern data analytics with examples.

2. Week 2
   - Importing, processing, and cleaning data.
   - Practice working in Excel

3. Week 3
   - Introduction to R and case studies
   - Functions, scripts, and IDEs for programming.

4. Week 4
   - Data Visualization

5. Week 5
   - Exploratory Data Analysis (basic stats)

6. Week 6
   - Exploratory Data Analysis (multivariate)
7. Week 7
   • Probability Distributions
   • Model Fitting

8. Week 8
   • Data Cleaning in R
   • Midterm Exam

9. Week 9
   • Linear Regression

10. Week 10
    • Multiple Regression
    • Logistic Regression

11. Week 11
    • Clustering and Classification
    • Dimension Reduction

12. Week 12
    • Network Models

13. Week 13
    • Professional practices
    • Careers

14. Week 14
    • Ethics in Data Science
    • Copyright and Licensing

15. Week 15
    • Presentations of projects