CSC 120: Applied Data Analytics

Fall 2022

Course Information

- Instructor: Dr. Jason Willwerscheid (jwillwer@providence.edu)
- Office: Howley 202
- Office Hours: Tu 10:30a-11:30a; W 12:30p-1:30p; F 10:30a-11:30a; or by appointment
- Course Websites: Canvas and GitHub

Course Catalog Description

Applied data analytics examines how organizations use data to gain insights and make better decisions by utilizing data modeling and statistical methods. Descriptive, predictive, and prescriptive data analytics will be covered, along with techniques for producing effective data visualizations.

Course Objectives

By the end of the course, you will:

- Know how to find relevant datasets in your areas of interest.
- Appreciate that real-world datasets are rarely unproblematic; understand some of the issues that arise in data collection and data cleaning.
- Be comfortable using the statistical programming language R to explore datasets via "wrangling" and visualization.
- Be familiar, at a conceptual level, with basic statistical principles of data modelling, including ethical issues and common misunderstandings.
- Have an in-depth understanding of the uses and abuses of *one* modelling technique (regression).

It's also useful to spell out what the course does *not* aim to do. In particular, you will not necessarily:

- Have technical mastery in R (this is not a programming course).
- Have a detailed mathematical understanding of hypothesis testing, linear regression, etc. (this is not a statistics course).
- Be exposed to a panoply of machine learning techniques such as clustering, support vector machines, and neural networks. These techniques are very powerful, but they require a good statistical foundation if they are to be used responsibly.

Disclaimer

This syllabus is extremely provisional. Although Applied Data Analytics is not new to PC, it is new to me, and I will be teaching it differently from other instructors. What you will find below is the course as I've imagined it, but we will need to be flexible and see how everything goes. **Everything in this syllabus is subject to change,** including (but not limited to) topics, policies, and exam formats.

Course Format

There will be a regular weekly routine:

- M/Tu: Lecture-based, with slides
- W: Discussion-based, with live code demonstrations
- F: Lab (collaborative problem solving)

Course Requirements

- There are no prerequisites for this course.
- A (reliable) laptop is required. You should bring it to class for labs and, optionally, demos, but I will ask you to keep it closed during lectures.
- All required software is freely available online.
- There is no required textbook. All readings will be from materials that are freely available online.

Grade Breakdown

Category	Percentage
Participation	10%
Assignments/Labs	40%
Group Project	20%
Midterm Exam	10%
Final Exam	20%

Attendance and Participation

- Attendance is mandatory, but please do not come to class if you are feeling ill. If you know in advance that you will be absent, I would appreciate notification via e-mail. For multiple or extended absences, I will ask for documentation. One or two undocumented personal health days is fine.
- Participation can take different forms: asking questions; offering feedback; helping others during labs. How you participate is up to you.
- Please refrain from using your laptop and phone during lectures, as it is discourteous to both myself and students around you. Even on days when a laptop is required, please do your best to avoid becoming distracted (texting; online shopping; Facebooking, or whatever kids do these days). If there is an urgent matter that you must attend to, you may silently excuse yourself from the classroom.

Assignments and Labs

- There will be one assignment and one lab per week. **Since the assignments provide necessary background material for the labs, they will be due on Thursday night at 11:59p.** You can expect to finish labs by the end of class on Friday, but I'm happy to give you until the following Thursday to finish up as long as you've used class time productively.
- You are expected to complete assignments and labs on time even in the event of an excused absence, unless we agree in advance that the due date should be extended for you.
- Both assignments and labs will be "submitted" to your personal GitHub repository (more on that soon...). **Everyone will be able to view everyone else's repository.** I am more interested in allowing you to learn from one another than in policing individual effort on homework.
- The grading scheme will be roughly: full marks for satisfactory work completed on time; half marks for satisfactory work completed late; no points for unsatisfactory (shoddy or incomplete) work. I will look at both the last version submitted before the deadline and the final version submitted, so you will never be penalized for revising your work. You may revise all assignments and labs up until the final week of classes.

Group Project

- In small groups (~3 students), you will do an in-depth analysis of a single dataset. We will use group repositories to facilitate collaboration.
- You will be encouraged to locate a dataset that suits your own interests as a group, but I will also provide a list of datasets that you can choose from.
- You will present your findings as a group during the last week of classes, but you should consider it more as a presentation of a work in progress and less as a final, polished piece of work. Ideally, you will be able to further revise based on the feedback you receive.
- I will likely replace one or two of the later labs with group meetings so that we have time to discuss the projects one-on-one (one-on-three?) before presentation.

Exams

- The midterm will be in-class and will be largely conceptual. I might ask you to read a bit of code, but I will certainly not ask you to write any. If there's any math, it will be the kind that doesn't require a calculator. Most (if not all) of the questions will be teased in the homework and lab "Review" sections.
- The final will be a guided data analysis, done individually. It will most likely be a take-home assignment; if so, I will make it about two hours worth of work (or less) and will assign it on the day of our regularly scheduled final exam (but I will give you the full day to do it rather than the scheduled two-hour window).
- You should make every effort to be in class for the midterm and to submit the final on time. If you cannot do so, you must notify me in advance if at all possible, and I will require that your excuse be both legitimate and documented. It will be up to me to decide whether an excuse is legitimate (which is why it's important that we discuss your absence in advance!). Flimsy and undocumented excuses will result in zero marks.

Policies

- The course is 3 credit hours, which means that you are expected to spend about 6 hours outside of class on readings, assignments, etc. If you find that it is taking up much more of your time, then please let me know and I will try to scale back. Since this is a new class, communication is crucial: if you're putting in a lot of time but are still having trouble keeping up, then it's likely that others are as well. Do not allow yourself to suffer in silence.
- **Please do not send me any questions other than administrative ones by e-mail.** I will answer any questions you have about the material during office hours; if you cannot make the drop-in times, then I will work with you to schedule a meeting outside of regularly scheduled office hours, either in person or on Zoom. Painful experience has taught me that "quick questions" are rarely as advertised. Your question will go unanswered; I will feel bad about not answering your question; you will feel bad that I never answered your question. Let's schedule a meeting and not feel bad.
- Masks are optional. I will most likely wear one on lab days, especially when I'm circulating among groups, and you should feel free to wear one as well (during labs or on any other day). In general, please follow the College's COVID-19 protocols. **I reserve the right to modify my mask policy at any time and at my own discretion.**
- Please take a moment to review the College's policy on Academic Integrity. Plagiarism can be a tricky subject in a coding class. It's inevitable that a lot of code will look like a lot of other code since there's often an optimal/conventional way to do things. And indeed, I will encourage you to view each other's assignment submissions to see how others have tackled the same problems. But there are limits. You can consult someone else's code to learn tips and tricks, but you should very rarely just cut and paste, and you should never cut and paste without understanding what a code snippet is doing.
- The usual rules about plagiarism apply to all textual commentary and written assignments.
- If you work with or borrow from someone for an assignment, the courteous and professional thing to do is to give attribution. You can simply add an "Acknowledgements" section at the end of your report, which states, for example, "I worked with X on Problems 1-3," or "My solution to Problem 4 is greatly indebted to Y's most clever and meritorious solution." **You will never be penalized for collaborating on homework, as long as it is acknowledged.**
- The one exception to this collaboration policy is the final exam. This is to be done individually, and there will be zero tolerance for cheating. I will write out a clearer final exam policy at the appropriate time.

Academic Support and Accommodations

- I encourage you to seek out any support and/or accommodations that might be useful to you. See Academic Support Services for a full list of services. In particular, you should be aware of:
  - The Tutoring Center: provides individual and group tutoring services.
  - Accessibility Services: facilitates disability accommodations (for a partial list, see here).
  - The Multicultural Student Success Program: supports students of color and first-generation college students.
  - Student-Athlete Services: supports the academic well-being of student-athletes.
  - Academic Coaching: offers one-on-one coaching and small-group workshops addressing topics such as study techniques, time management, and motivation.
  - English as a Second Language Support: offers support to students who are non-native English speakers.
- Please note that if you require accommodations such as extended time on exams, then you **must** contact Accessibility Services. I am not qualified to grant any such accommodations myself.

Applied Data Analytics
Fall 2022

Exploratory Data Analysis

Introductions	Data science and friends, RStudio, R Markdown, Github	*wk 1*
Raw Data	Relational and non-relational databases, Data types, Base R	*wk 2*
Data Wrangling	Grammar of data manipulation, Summary statistics, `dplyr`	*wk 3*
Data Visualization	*The Visual Display of Quantitative Information*, Base R graphics, `ggplot`	*wk 4*
Data in the World	Missing data and data imputation, Data privacy	*wk 5*
Data Scraping	Scraping legality, etiquette, and technique	*wk 6*
Text Data	String processing, Bags of words	*wk 7*
Data Ethics	Transparency, Interpretability, Reproducibility	*bonus*

Data Modelling

Probability	Random variables, Distributions, Monte Carlo sampling	*wk 8*
Hypothesis Testing	Common hypothesis tests, p-values, Confidence intervals	*wk 9*
Multiple Testing	p-hacking, Publication bias, Crisis of reproducibility	*wk 10*
Linear Regression	Correlation, Linear models, Prediction	*wk 11*
Multiple Regression	Diagnostics, Model fit, Outliers, Causality	*wk 12*
High-Dimensional Data	Collinearity, Feature selection	*wk 13*
Logistic Regression	Modelling binary and count data, Data transformations	*wk 14*

Presentations

Project Presentations	*wk 15*