

# DATA 115: Introduction to Data Analytics

## Fall 2022

---

### Instructor Information

---

Instructor: Sergey Lapin  
Office: WSU Everett Room 405  
Office hours: M,W 11 am - noon  
Phone: 425-405-1780  
Email: [slapin\(at\)wsu.edu](mailto:slapin(at)wsu.edu)  
Classroom: EVRT 358; Clark Hall 151; Engineering and CS Bldg 125  
Class Time: Tu,Th 10:35-11:50

---

### 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.

---

## Course Materials

---

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 GitHub and 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<sup>1</sup>.

**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, tibble, and tidy.

---

## 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.

---

<sup>1</sup>In 2018, Kaggle surveyed 23,859 data scientists and found that R was second to Python in terms of most commonly used languages: [link](#).

---

## Class Communication

---

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](mailto: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.

---

## Learning Outcomes

---

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

---

## Motivating Questions

---

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 five 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 Monday, 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 but at the end of the semester your lowest two scores will be dropped. 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.
- **Final Project:** The final assessment in the course will be a group project, where each team will complete a data analytics task from beginning to end. 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.
- **Personal Dataset:** Beginning in the first week of class, each student 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. The purpose of this is both to have an example of completed work for internship and other professional applications as well as to allow students to have access to a meaningful starting point to apply techniques that will be encountered later in the major.

---

## 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.

---

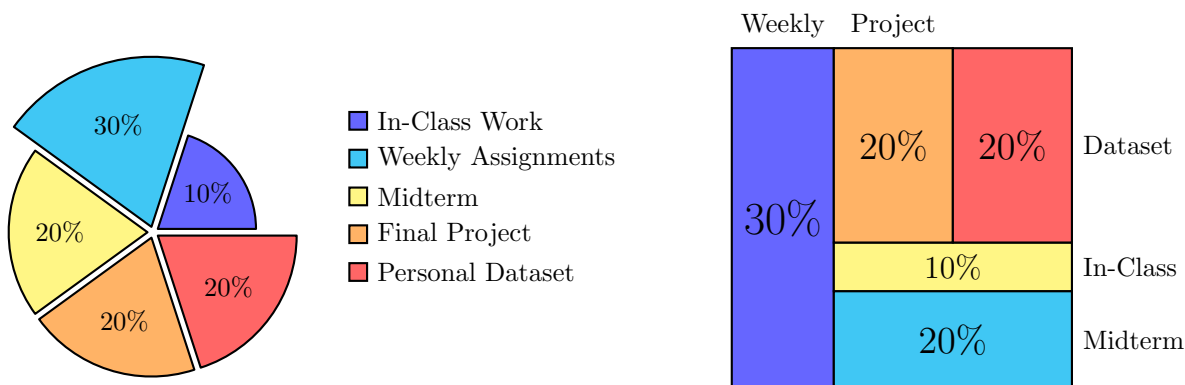
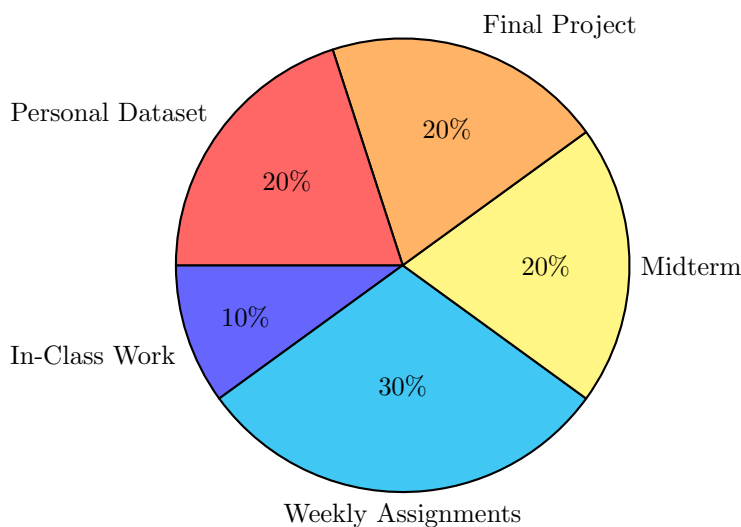
## Grading Policy

---

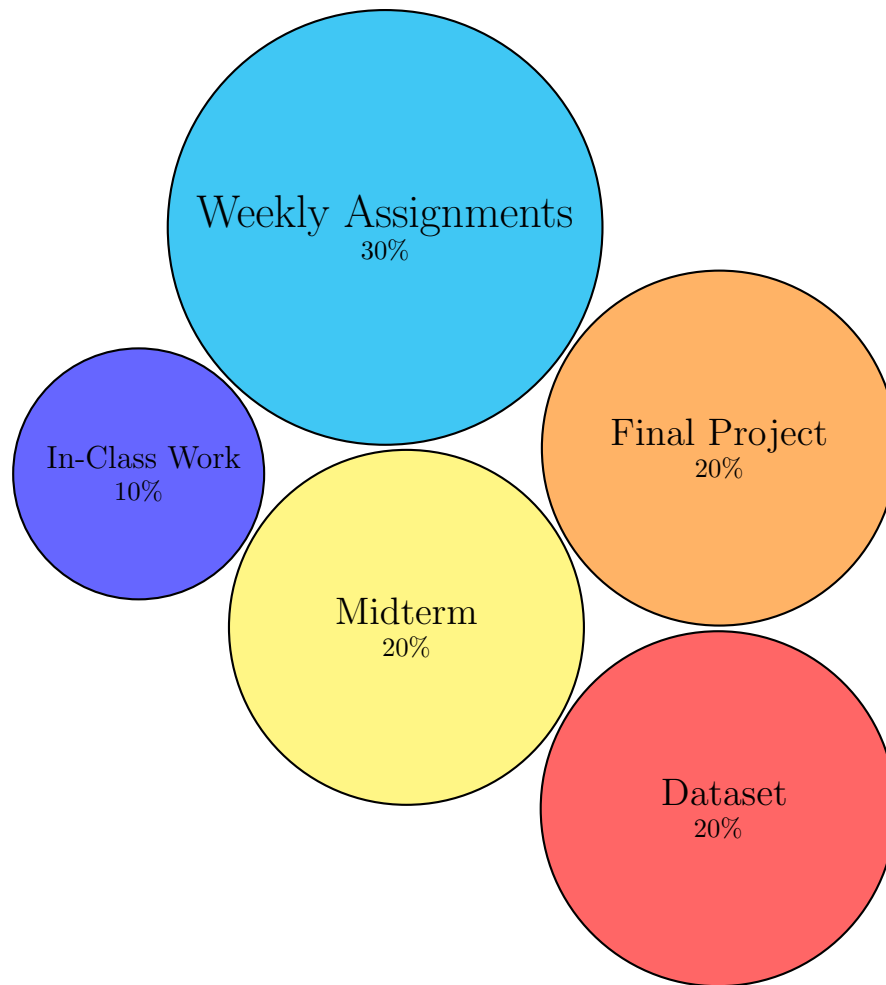
The breakdown of grade components and letter grades will be as follows <sup>2</sup> (the first weekly assignment will ask you to evaluate the visualizations ☺):

Assignment	Percentage
Participation	10%
Weekly Assignments	30%
Midterm	20%
Final Project	20%
Personal Dataset	20%

Grade	Percentage	Grade	Percentage
A	95-100	C+	77-79.99
A-	90-94.99	C	73-76.99
B+	87-89.99	C-	70-72.99
B	83-86.99	D	60-60.99
B-	80-82.99	F	0-59.99



<sup>2</sup>We reserve the right to be more lenient



---

## Weekly Topics

---

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.
- **Discussions:** The difference between Data, Analytics, and Insight

### 2. Week 2

- Importing, processing, and cleaning data.
- Pivot tables in Excel
- **Discussions:** College Mobility

### 3. Week 3

- Introduction to R and case studies
- Functions, scripts, and IDEs for programming.
- Installation Day
- **Discussions:** MLB Strike Zone Accuracy

4. Week 4
  - Data Visualization
  - **Discussions:** 538 Election Forecast Visualizations
5. Week 5
  - Exploratory Data Analysis (basic stats)
  - **Discussions:** The fallacy of interpreting deaths and driving distances
6. Week 6
  - Exploratory Data Analysis (multivariate)
  - **Discussions:** Is sushi healthy? (and spurious correlations)
7. Week 7
  - Probability Distributions
  - Model Fitting
  - Probability Distributions
  - **Discussions:** How coronavirus grounded the airline industry
8. Week 8
  - Data Cleaning in R
  - Midterm Exam
9. Week 9
  - Linear Regression
  - **Discussions:** Regression and NFL results
10. Week 10
  - Multiple Regression
  - Logistic Regression
  - **Discussions:** The COMPAS recidivism algorithm
11. Week 11
  - Clustering and Classification
  - Dimension Reduction
  - **Discussions:** Halloween candy power rankings
12. Week 12
  - Network Models
  - Null Models and Social Networks
  - Veteran's Day
  - Final projects start
13. Week 13
  - Professional practices
  - Careers
  - **Discussions:** How does your state make electricity?

14. Week 14

- Ethics in Data Science
- Copyright and Licensing
- **Discussions:** Data anonymity and privacy concerns

15. Week 15

- Presentations of personal dataset projects

16. Final Group Presentations Wednesday 05/04 10:30-12:30

---

## University Policy Statements

---

### Academic Integrity Statement

Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course. Students who violate WSU's Academic Integrity Policy (identified in Washington Administrative Code (WAC) 504-26-010(4)) will fail the assignment, will not have the option to withdraw from the course pending an appeal, and will be reported to the Center for Community Standards. Multiple violations of the policy will cause you to fail the course.

Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students, WAC 504-26-010(3). You need to read and understand all of **the definitions of cheating**. If you have any questions about what is and is not allowed in this course, you should ask course instructors before proceeding.

If you wish to appeal a faculty member's decision relating to academic integrity, please use the form available at [communitystandards.wsu.edu](https://communitystandards.wsu.edu). Make sure you submit your appeal within 21 calendar days of the faculty member's decision.

### WSU Reasonable Accommodation Statement

Reasonable accommodations are available for students with documented disabilities or chronic medical or psychological conditions. If you have a disability and need accommodations to fully participate in this class, please visit your campus' Access Center/Services website to follow published procedures to request accommodations. Students may also contact their campus offices to schedule an appointment with a Disability Specialist. All disability related accommodations are to be approved through the Access Center/Services on your campus. It is a university expectation that students visit with instructors (via email, Zoom, or in person) to discuss logistics within two weeks after they have officially requested their accommodations. For more information contact a Disability Specialist on your home campus:

- Pullman, WSU Global Campus, Everett, Bremerton, and Puyallup: 509-335-3417 Access Center (<https://www.accesscenter.wsu.edu>) or email at [access.center@wsu.edu](mailto:access.center@wsu.edu).



## Religious Accommodation Statement

Washington State University reasonably accommodates absences allowing for students to take holidays for reasons of faith or conscience or organized activities conducted under the auspices of a religious denomination, church, or religious organization. Reasonable accommodation requires the student to coordinate with the instructor on scheduling examinations or other activities necessary for course completion. Students requesting accommodation must provide written notification within the first two weeks of the beginning of the course and include specific dates for absences. Approved accommodations for absences will not adversely impact student grades. Absence from classes or examinations for religious reasons does not relieve students from responsibility for any part of the course work required during the period of absence. Students who feel they have been treated unfairly in terms of this accommodation may refer to Academic Regulation 104 – Academic Complaint Procedures.

---

## Expectations for Student Effort

---

Students should expect to spend a minimum of 9 hours per week, engaged in the following types of activities: reading, listening to/viewing media, discussion, or conversation in the LMS or other academic technology, conducting research, completing assignments and reviewing instructor feedback, studying for and completing assessments, etc

## Safety and Emergency Notification

Classroom and campus safety are of paramount importance at Washington State University, and are the shared responsibility of the entire campus population. WSU urges students to follow the “Alert, Assess, Act,” protocol for all types of emergencies and the “**Run, Hide, Fight**” response for an active shooter incident. Remain ALERT (through direct observation or emergency notification), ASSESS your specific situation, and ACT in the most appropriate way to assure your own safety (and the safety of others if you are able).

Please sign up for emergency alerts on your account at MyWSU. For more information on this subject, campus safety, and related topics, please view the **FBI’s Run, Hide, Fight video** and visit the **WSU safety portal**.

Full details can be found at <https://provost.wsu.edu/classroom-safety/>

## Discrimination and Harassment Policy Statement

Discrimination, including discriminatory harassment, sexual harassment, and sexual misconduct (including stalking, intimate partner violence, and sexual violence) is prohibited at WSU (See **WSU Policy Prohibiting Discrimination and Harassment** (Executive Policy 15) and **WSU Standards of Conduct for Students**).

If you feel you have experienced or have witnessed discriminatory conduct, you can contact the WSU Compliance & Civil Rights (CCR) and/or the **WSU Title IX Coordinator** at 509-335-8288 to discuss resources, including confidential resources, and reporting options. (Visit [ccr.wsu.edu](http://ccr.wsu.edu) for more information).

Most WSU employees, including faculty, who have information regarding sexual harassment or sexual misconduct are required to report the information to CCR or a designated Title IX Coordinator or Liaison. (Visit [ccr.wsu.edu/reporting-requirements](http://ccr.wsu.edu/reporting-requirements) for more info).

## Online Discussion Policy

The essence of education is exposure to diverse viewpoints. In your discussion posts you’ll meet students with vastly different opinions and backgrounds. You’re encouraged to disagree with the substance of others’ ideas and opinions but do so with an active sense of respect for one another, and without losing focus on the topic at hand. Personal attacks, inflammatory statements, flaming, trolling, and disruption of the discussion do not have a place in academic discourse. Postings must comply with University policy on use of computing resources, including those regarding harassment and discrimination, as well as conform to the **WSU Community Standards**.

Your instructors will promote high-quality academic discussions by removing any posts they view as disruptive of the educational process and alerting students whose posts have been removed that they have violated course expectations. Students who continue to misuse the discussion boards after a warning may be subject to removal of access rights, course failure, and referral to the Office of Community Standards. [Visit netiquette guidelines.](#)