DA 101: Introduction to Data Analytics

Denison University, Granville OH
Sarah R. Supp
supps@denison.edu

Preferred: Zip file of materials is provided here: DA101_Example_materials.zip

Knowledge Areas that contain competencies (knowledge, skills, and dispositions) covered in the course

Knowledge Area	Total Number of Contact Hours
Analysis and Presentation (AP)	10
Computing and Computer Fundamentals (CCF)	2
Data Acquisition, Management, and Governance (DAMG)	28
Data Mining (DM)	15
Data Privacy, Security, Integrity, and Analysis for Security (DP)	5
Professionalism (PR)	12
Programming, Data Structures, and Algorithms (PDA)	13
Software Development and Maintenance (SDM)	2

Where does the course fit in your undergraduate Data Science curriculum?

This is the required introductory course for the Data Analytics major. Non-majors take the course to fulfill general education requirements in quantitative reasoning and interdisciplinarity or to gain skills to help with future career goals. It has no prerequisites and is normally taken by first year and sophomore students. We typically teach seven sections per year, each capped at 20 students.

Is this course from or used in other curricula/majors?

DA 101 was created specifically for the Data Analytics major. It is designed to be a broad introduction to the concepts of programming, statistical reasoning, data communication, ethical

analysis, and current trends in the field of data analytics and data science more generally. It is taken by both majors and non-majors.

What is covered in the course?

Many of the most pressing problems in the world can be addressed with data. We are awash in data and modern citizenship demands that we become literate in how to interpret data, what assumptions and processes are necessary to analyze data, as well as how we might participate in generating our own analyses and presentations of data. Consequently, data analytics is an emerging field with skills applicable to a wide variety of disciplines. This course introduces analysis, computation, and presentation concerns through the investigation of data driven puzzles in a wide array of fields – political, economic, historical, social, biological, and others. No previous experience is required.

What is the format of the course?

The class is in-person with three 50-minute class sessions and one 3-hour lab per week. Each class session uses a combination of lecture, hands-on problem solving, and student-led discussion to encourage engagement and practice with the concepts and skills. Lab sessions engage students in collaborative problem-solving with real datasets on a variety of topics (e.g., natural science, social science, humanities, sports, and business), with the goal of producing a complete report that works through the entire data analytics cycle, while teaching and practicing new coding, statistical, ethical, and communication skills.

How are students assessed?

The class is typically taught as a series of projects (individual- and team-based) to provide hands-on practice and to assess learning. Students use RStudio to generate RMarkdown code, analyses, and explanations, knitted as html reports. Projects are introduced weekly in the lab, and students spend approximately 6-10 hours per week completing them. This means that while they get started in lab where they are able to work together and ask questions, completing the lab report is done as homework. At the end of the term, students work together in teams of 1-3 (their choice) to identify a dataset, ask a research question, and synthesize their skills to create a report and oral presentation, complete with statistical and visual interpretation of the data.

No cumulative exams are given in the course, but low-stakes bi-weekly quizzes provide opportunities to assess student learning throughout the semester.

Students also engage with primary or secondary literature related to broad topics in data analytics each week, and may spend 1-3 hours on reading and preparing for in-class discussion. Topics may include but are not limited to: bias or missing data in datasets, careers in data science, ethical dilemmas, statistical interpretation, diversity in data analytics, and data visualization and data art.

Course tools and materials

The course is taught using mainly free online open source resources.

- Recommended Textbook: **R for Data Science**, Garrett Grolemund and Hadley Wickham, ISBN 978-1491910399 or freely available online https://r4ds.had.co.nz/ ● The course uses R via RStudio, which is freely available
- Data sets used include those provided freely as data packages in for R, open data from various online sources, and data from sources to which Denison subscribes

Why do you teach the course this way?

This course is taught in a way that emphasizes hands-on practice, explicitly teaching

collaboration strategies, because the role of data analysts is inherently interdisciplinary and frequently requires the ability to work closely with others while considering a problem from multiple perspectives. Because data analytics is a constantly evolving field, including the foundational skills in coding and statistical analysis, it is critical for students to learn to feel confident in searching for help themselves, and in their ability to learn new skills on their own.

As an introductory course, teaching with freely available tools and resources allows students to become proficient in analysis skills and quantitative reasoning early in their college careers, and for them to be able to apply those skills (as majors or non-majors) in other courses, work experiences, or internships they may have in the future, without the need for special permissions or licensing. Using real datasets in class and discussing data analytics in the context of current events provides ample opportunities for students to connect their learning to their own lives and goals, and to explore critical issues of bias, ethics, and communication.

Students typically consider the course challenging and time consuming, though it is also frequently reviewed very positively, and student achievement in the final course assessment is often high. The final project is often viewed as very rewarding by students, because they have an opportunity to make their own choices about the dataset, a question they want to answer, and which methods they want to practice.

Body of Knowledge coverage

KA	Sub-domain	Competencies Covered	Hours
AP	Foundational considerations	data visualization; using color; accessibility; perceiving bias; guidelines and standards	5
AP	Visualization	historical and contemporary examples of visualization; characteristics of effective visualization; suitability of different techniques for different data and different users; inference based on visualization; role of color; chart types; given a set of data for a particular purpose, identify and implement an effective visualization	5
CCF	File systems	compare and contrast different approaches to file organization; appreciate the importance of good file organization; metadata	2
DAMG	Data acquisition	select data sources; security and privacy standards and best practices	3
DG	Working with various types of data	write programs to perform basic operations on data of each type; summary statistics; data representation; text data processing	8
DG	Data integration	understand the challenges brought by heterogeneous data sources; merge datasets together on common variable(s)	2

DG	Data reduction and compression	data sampling approaches; data filter techniques	2
DG	Data transformation	simple function transformation methods and their applications	2
DG	Data cleaning	evaluate and improve data quality; data cleaning; aware of the harm of data quality problems; appreciate and implement data cleaning	9
DG	Data privacy and security	complete CITI SBE modules for human subjects research; explain basic data privacy and security guidelines, especially as dictated by the university; appreciate ethical implications of the harm of data governance policies and actions; aware of the harm of data loss due to security and privacy failures; Maintain the upmost ethical standards regarding legal and social responsibility for data	2
DM	Proximity Measurement	correlation coefficient; describe and compare measurement concepts and their relevance to different kinds of data - continuous, discrete, nominal, ordinal; select metrics appropriate for comparing various kinds of data	5
DM	Data Preparation	Gathering data, its relationship to problem solving, importance of expert knowledge and being open to the views of experts; Sources of data including databases and online information sources; adequacy of data for particular purposes; Ethical considerations around obtaining and using data; privacy concerns around collocating data; concerns around potential bias in data; Munging data - dealing with errors in data, gaps in data, cleansing data, validating data, profiling data, transforming data, and joining datasets as appropriate; quality considerations; Methods of dealing with dataset issues such as imbalance; automated and manual approaches and trade-offs between these	10
DP	Social Responsibility	Tradeoffs between the right to privacy and need of transparency; Ethical responsibilities about disclosing, transmitting, and sharing information obtained from analytics tools; Aware of data sensitiveness when data is processed as an input.; Identify scenarios where data cleaning must be considered before processing information.	5
PR	Continuing professional development	Recognise that data science is a rapidly changing field where keeping current, as well as knowing how to stay current, are vital.	1

PR	Communication	Evaluate aspects of the technical literature relevant to data science; Produce presentations for a range of audiences; Reflect positively on the significance of new learning and new experiences; Recognise strengths and weaknesses regarding knowledge	3
PR	Teamwork	Team selection, the need to complement abilities and skills; dynamics of teams and team discipline; Elements of effective teams; Outline steps that could be taken to deal	2

		with conflict situations; Set aside unimportant differences when working with others; Demonstrate appropriate levels of flexibility.	
PR	Privacy and confidentiality	Include and maintain privacy and confidentiality to ensure confidence in data science activities.	1
PR	Ethical considerations	Illustrate a range of situations in which a data scientist may venture beyond their range of competence and identify steps to mitigate; Techniques for establishing lack of bias in data sets; Alert to the deep ethical issues in data gathering and use; Aware of issues of bias and seek to remove these; Self-directed and self motivated in the advancement of data science.	5
PDA	Programming	Core coding concepts including variables and primitive data types, expressions and assignments, conditional and iterative control structures, functions and parameter passing; documentation; Decomposition to break a program into smaller pieces; Types of errors, how they occur, and how to handle them; Strategies for debugging; Read, write and debug programs that include core concepts and practices; Trace the execution of code segments and articulate summaries of their computation; Use consistent documentation and program style standards that contribute to the readability and maintainability of software.	3
PDA	Algorithms	Simple numerical algorithms(average, min, max, or mode, etc); Properties of graphs: connectedness, betweenness, centrality, etc.; Be aware that there are multiple ways to address a problem; Recognize that method has implications for efficiency.	10

SDM	Software design and development	Coding standards; data lifecycle; Recognize the value of a team built on respect, diversity, and collaboration; Work with others by demonstrating good listening skills, the ability to present an idea, and the ability to negotiate; Approach data and software projects with a lifecycle mindset	2

Additional topics

Introduction to basic statistics

- summary statistics
- distributions
- hypothesis testing
- t-test
- linear regression
- model selection
- p-hacking

Other comments