

DATA 607 Data Acquisition and Management

CUNY SPS Master of Science in Data Science

Spring 2021

Instructor: Andrew Catlin

Class Meetup: Wednesdays 6:45pm

Office Hours: by appointment

Email: andrew.catlin@sps.cuny.edu

Course Description

In this course students will learn about core concepts of contemporary data collection and its management. Topics will include systems for collecting data (real time, sensors, open data sets, etc.) and implications for practice; types of data (textual, quantitative, qualitative, GIS, etc.) and sources; an overview of the use of data, including what and how much should be collected and the distinction between data, information, and knowledge from a data-centric point of view; provenance; managing data with and without databases; computer and data security; data cleaning, fusing, and processing techniques; combining data from different sources; storage techniques including very large data sets; and storing data keeping in mind privacy and security issues. Students will be required to create a working system for a large volume of data using publicly available data sets.

Course Learning Outcomes:

By the end of the course, students should be able to:

- Load data into R from various data sources, including CSV files, Excel spreadsheets, relational databases, APIs, and web pages.
- Perform various data cleansing and transformation work, including splitting, combining; resampling; variable creation; data aggregation; sorting and filtering data; strategies for working with outliers and missing data; data visualization and analysis in support of data cleansing activities.
- Understand different information architectures, data types, and data structures.
- Understand relational and non-relational database design and querying.
- Provide context for data science

Program Learning Outcomes addressed by the course:

- Business Understanding. Apply frameworks and processes to build out data analytics solutions from understanding of business goals.
- Data Culture. Embody and champion the highest standards for the ethical and moral use of data; understand issues related to data privacy and data security.
- Solid foundational data programming skills, using industry standard tools, essential algorithms, and design patterns for working with structured data, unstructured data and big data.
- Data understanding. Collect, describe, model, explore and verify data.
- Data preparation. Selecting, cleaning, constructing, integrating, and formatting data.

How is this course relevant for data analytics professionals?

Most data analytics professionals spend most of their time getting data and preparing it for analysis. This is the course that teaches these key skills, as we work with both structured and unstructured data.

Grading

Assignment	Percent
Assignments (6 x 50)	30%
Projects (3 x 90)	27%
Final Project Proposal (1 x 20)	2%
Final Project (1 x 120)	12%
Final Project Presentation (1 x 30)	3%
Discussion Participation (14 x 10)	14%
Data Science in Context Presentation (1 x 50)	5%
TidyVerse recipes	4%
Quizzes	3%
TOTAL	100%

Grade Distribution

Quality of Performance	Letter Grade	Range %	GPA
Excellent - work is of exceptional quality	A	93 - 100	4
Excellent	A-	90 - 92.9	3.7
Good - work is above average	B+	87 - 89.9	3.3
Satisfactory	B	83 - 86.9	3
Below Average	B-	80 - 82.9	2.7
Poor	C+	77 - 79.9	2.3
Poor	C	70 - 76.9	2
Failure	F	< 70	0

Notes

- All projects and assignments, unless otherwise noted, are due end of day on Sundays. Late projects are not accepted. However, there are eight assignments and four projects assigned, and your final grade is based on your six highest-scoring assignments and your three highest-scoring projects.
- **Late projects are not accepted.** However, there are eight assignments and four projects assigned, and your final grade is based on your six highest-scoring assignments and your three highest-scoring projects.
- Each course week will be available on the previous Friday at 5:00 p.m. ET.
- **Course Completion Requirements.** To pass this course, you must complete at least six assignments, three projects, the final, and make the final presentation. If you cannot deliver your presentation in our 05/13 Meetup, you will need to make available a recorded version of your final presentation before 05/13.
- There will also be **short ungraded hands-on labs** most weeks that will help you prepare for your weekly programming assignments and projects.
- **“Discussion”, “Data Science in Context Presentations”, and “TidyVerse Recipes”.** While this material is important, please note that this work only makes up less than one quarter of your grade. Please do the readings and participate in the discussions and any discussion-related group assignments, make your Data Science in Context presentations, and participate in the creation and editing of TidyVerse recipes on the shared

GitHub site. If you are participating at a reasonable level and turning in your work on time, you will receive the full 23% here. At the same time, if you have limited time for the course, please remember to invest the majority of your efforts in completing the projects and assignments. The assignments merit close attention because they will help you to be successful on the projects.

- **Reproducibility Requirement, Testing Requirement, But Not Perfection!** Students are responsible for providing all code and data so that I can test your work. If you turn in code that does not run, you will not receive credit, unless you also include an explanatory note at the time of submission. At the same time, you do not need to turn in perfect code. Generous partial credit will be given for deliverables that are timely, tested, and reproducible. Cutting corners—as long as they are documented at the time of submission—is also acceptable.
- **Groupwork** is encouraged on most projects and assignments and required on Project 3. Effective virtual collaboration is highly valued in the data science marketplace; because of its interdisciplinary nature, much of the work that needs to be done requires more than one person, and increasingly often at multiple locations.
- **Earning a Grade of A.** If you complete the course work correctly and on time, you will comfortably pass the course. A grade will be reserved for students that go above and beyond, such as consistently taking on challenge assignments.
- **Policy on Sharing and “Stealing” Code.** In this course, you may collaborate, and you may take base code from whatever sources you wish. But you must document what you started with, and what you added, so you are graded only on your own contributed work!

Course Learning Materials



Textbooks:

- R for Data Science by Garrett Grolemund and Hadley Wickham. O'Reilly, 2017. Freely readable here: <http://r4ds.had.co.nz/>. Print copies are also available.
- Text Mining with R: A Tidy Approach, Julia Silge and David Robinson. O'Reilly, 2017. Freely readable •
- Data Science for Business, Tom Fawcett and Foster Provost, O'Reilly, 2013.
- Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining, Simon Munzert, Christian Rubba, Peter Meißner, and Dominic Nyhuis. Wiley, 2015. Important errata here: <http://www.r-datacollection.com/errata/errata.pdf>.
- Max Kuhn and Kjell Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models (Chapman & Hall/CRC Data Science Series) 1st Edition, 2019. Freely readable at <https://bookdown.org/max/FES/intro-intro.html> Recommended Texts:
- Any book on SQL, such as The Language of SQL by Larry Rockoff. ISBN: 978-1435457515. Alternatively, there are many excellent on-line resources, such as <http://sqlzoo.net>.
- Another excellent R book with a more statistical bent is R for Everyone by Jared Lander. ISBN: 978-0321888037.

Relevant Software, Hardware, or Other Tools: We will make use of the R programming environment and the RStudio IDE. We will use other open-source software, including (your choice of) MySQL or PostgreSQL, MongoDB, Neo4J, Hadoop, and Spark. Details for obtaining and installing the appropriate software will be provided in the course materials. All of the software will work on (or from) both PCs and Macs.

How This Course Works:

Meetups take place every week on Wednesdays from 6:45 p.m. to 7:45 p.m. ET. Please see course site for specific dates. You are strongly encouraged to attend; all meetups will be recorded.

Andy Catlin will have **Thematic Workshops** most Sunday afternoons. These are optional, ungraded enrichment opportunities for interested students. Students from any section are welcome to participate in either or both of the optional Thematic Workshops.

- **Data Engineering Workshop**, most Sundays 2:00 p.m. – 3:00 p.m. ET. In addition to the coverage of SQL in Week 2, Andy Catlin will have additional office hours for students who want to improve their SQL skills, and get practice working with cloud-based database tools on Amazon Web Services, including such as EC2 instances, RDS, DynamoDB, and RedShift. While we will use AWS free tier services as much as possible, interested students will be responsible for their own cloud computing costs. To get the most benefit out of the workshop, you should plan to (optionally) spend an additional 3 to 4 hours per week on related assignments: we will leverage several of DataCamp’s SQL courses, as well as some lab materials provided by AWS.
- **Data Visualization and Storytelling Workshop**, most Sundays 3:00 p.m. – 4:00 p.m.
- You will work more with ggplot2, practice using Tableau, and learn about the basics of Data Visualization best practices. Later in the program, you will learn much more about Data Visualization in your DATA 608 course.

Regular Office Hours can also be scheduled by e-mail appointment. If you need extra help and are willing to invest the time and effort to be successful, I will make the time to help you. But...you should not be asking for extra help on a project the day before it is due, since this indicates that you are not investing the time and effort to be successful.

You are encouraged to ask questions on the “Ask Your Instructor” forum on the course discussion board where other students will be able to benefit from your inquiries. I can set up a GoToMeeting session for screen sharing. For the most part, you can expect me to respond to questions by email within one business day.

Schedule

Note: Schedule is subject to change.

Dates	Topic	Core Readings	Deliverables
Jan-29 to Feb-07	Building out your Data Science Development Environment;	Data Science for Business, chapter 1	Week 1 Assignment
Feb-08 to Feb-14	R: Data Types and Basic Operations	Data Science for Business, chapter 2	Week 2 Assignment
Feb-15 to Feb-21	R and SQL	Data Science for Business, chapter 3	Week 3 Assignment
Feb-22 to Feb-28	R: Character Manipulation and Date Processing	Data Science for Business, chapter 4	Project 1
Mar-01 to Mar-07	R: Exploratory Data Analysis; Data Imputation	Data Science for Business, chapter 5	Week 5 Assignment
Mar-08 to Mar-14	R: Working with Tidy Data	Data Science for Business, chapter 6	Project 2
Mar-15 to Mar-21	R: Data Transformations; Feature Engineering	Data Science for Business, chapter 7	Week 7 Assignment
Mar-22 to Apr-04	Web Technologies; MongoDB	Data Science for Business, chapter 8	Project 3
Apr-05 to Apr-11	Scraping Web Pages	Data Science for Business, chapter 9	Week 9 Assignment; Tidyverse Recipes Initial Post due
Apr-12 to Apr-18	Working with Web APIs	Data Science for Business, Chapter 10	Week 10 Assignment; Tidyverse Recipes Initial Post Peer Grading due
Apr-19 to Apr-25	Text Mining	Data Science for Business, chapter 11	Week 11 Assignment
Apr-26 to May-02	Recommender Systems	Data Science for Business, chapter 12	Project 4; Final Project Proposals due; Data Science in Context presentations due for students opting to make recorded versions
May-03 to May-09	Graph Databases	Data Science for Business, chapters 13 and 14	Work on final projects and presentations; Tidyverse Recipes Extension Peer Grading due
May-10 to May-16	Working with Data in the Cloud; Hadoop and Spark	Data Science for Business, Appendices A and B	Work on final projects and presentations
	Big Data Analytics; Automated Machine Learning	No readings	Final Project Presentations

Accessibility and Accommodations

The CUNY School of Professional Studies is firmly committed to making higher education accessible to students with disabilities by removing architectural barriers and providing programs and support services necessary for them to benefit from the instruction and resources of the University. Early planning is essential for many of the resources and accommodations provided. Please see: http://sps.cuny.edu/student_services/disabilityservices.html

Online Etiquette and Anti-Harassment Policy

The University strictly prohibits the use of University online resources or facilities, including Blackboard, for the purpose of harassment of any individual or for the posting of any material that is scandalous, libelous, offensive or otherwise against the University's policies. Please see: http://media.sps.cuny.edu/filestore/8/4/9_d018dae29d76f89/849_3c7d075b32c268e.pdf

Academic Integrity

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the educational mission of the City University of New York and the students' personal and intellectual growth. Please see: http://media.sps.cuny.edu/filestore/8/3/9_dea303d5822ab91/839_1753cee9c9d90e9.pdf

Student Support Services

If you need any additional help, please visit Student Support Services: http://sps.cuny.edu/student_resources/