

Course Name and Number: IS 607 – Data Acquisition and Management

Credits: 3 cr.

Prerequisite(s): none

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.

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.

Assignments and Grading:

Assignments (6 x 50)	30%
Projects (3 x 100)	30%
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%
TOTAL	100%

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

Notes

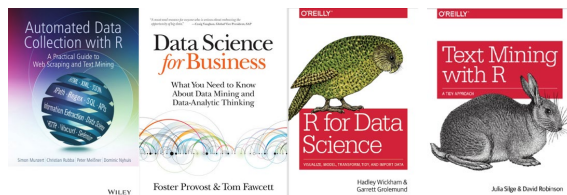
- All projects and assignments, unless otherwise noted, are due end of day on Sundays.

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.

- 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 05/15 Meetup, you'll need to make available a recorded version of your final presentation before 05/15.
- 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'll 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 don't 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'll comfortably pass the course. A grades 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



Required Texts:

- *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>.
- *R for Data Science* by Garrett Grolemund and Hadley Wickham. O'Reilly, 2017. Freely downloadable 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 downloadable here: <http://tidytextmining.com/>. Print copies are also available.
- *Data Science for Business*, Tom Fawcett and Foster Provost, O'Reilly, 2013.

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.

Contact Information:

607/001

Andy Catlin

andrew.catlin@sps.cuny.edu

616-638-8344

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 meet-ups will be recorded. **Office Hours** can be scheduled by e-mail appointment. If you need extra help and are willing to invest the time and effort to be successful, I'll make the time to help you. But...you should not be asking for extra help on a project the day before it's due, since this indicates that you're 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.

Unit	Topic	Core Readings	Deliverables
Week 1 Jan 28 – Feb 03	Building out your Data Science Development Environment; R: Data Types and Basic Operations	<i>Automated Data Collection with R</i> , chapter 1. <i>Data Science for Business</i> , chapter 1	Meetup on 01/30, 6:45 p.m. EST Week 1 Assignment
Week 2 Feb 04 – Feb 10	R and SQL	<i>Automated Data Collection with R</i> , chapter 7. <i>Data Science for Business</i> , chapter 2	Meetup on 02/06, 6:45 p.m. EST Week 2 Assignment
Week 3 Feb 11 – Feb 17	R: Character Manipulation and Date Processing	<i>Automated Data Collection with R</i> , chapter 8. <i>Data Science for Business</i> , chapter 3	Meetup on 02/13, 6:45 p.m. EST Week 3 Assignment
Week 4 Feb 18 – Feb 24	R: Exploratory Data Analysis	<i>Data Science for Business</i> , chapter 4	Meetup on 02/20, 6:45 p.m. EST Project 1
Week 5 Feb 25 – Mar 03	R: Working with Tidy Data	<i>Data Science for Business</i> , chapter 5	Meetup on 02/27, 6:45 p.m. EST Week 5 Assignment
Week 6 Mar 04 – Mar 10	R: Data Transformations	<i>Data Science for Business</i> , chapter 6	Meetup on 03/06, 6:45 p.m. EST Project 2
Week 7 Mar 11 – Mar 17	Web Technologies; MongoDB	<i>Automated Data Collection with R</i> , chapters 2 through 6. <i>Data Science for Business</i> , chapter 7	Meetup on 03/13, 6:45 p.m. EDT Week 7 Assignment
Week 8 Mar 18 - Mar 24	Scraping Web Pages	<i>Automated Data Collection with R</i> , chapter 9. <i>Data Science for Business</i> , chapter 8	Meetup on 03/20, 6:45 p.m. EDT Project 3
Week 9 Mar 25 – Mar 31	Working with Web APIs	<i>Automated Data Collection with R</i> , chapter 9. <i>Data Science for Business</i> , chapter 9	Meetup on 03/27, 6:45 p.m. EDT Week 9 Assignment
Week 10 Apr 01 – Apr 07	Text Mining	<i>Automated Data Collection with R</i> , chapter 10. <i>Data Science for Business</i> , chapter 10	Meetup on 04/03, 6:45 p.m. EDT Week 11 Assignment
Week 11 Apr 08 – Apr 14	Recommender Systems	<i>Data Science for Business</i> , chapter 11	Meetup on 04/10, 6:45 p.m. EDT Week 12 Assignment
Week 12 Apr 15 – Apr 18	Graph Databases	<i>Data Science for Business</i> , chapter 12	Meetup on 04/17, 6:45 p.m. EDT Project 4; Final Project Proposals due Data Science in Context presentations due for students opting to make recorded versions
Week 13 Apr 19 – Apr 28	Spring Recess	No Readings	No Assignments
Week 14 Apr 29 – May 05	Working with Data in the Cloud; Hadoop and Spark	<i>Data Science for Business</i> , chapters 13 and 14	Meetup on 05/01, 6:45 p.m. EDT Work on final projects and presentations Tidyverse Recipes due
Week 15 May 06 – May 12	Big Data Analytics	<i>Data Science for Business</i> , Appendices A and B	Meetup on 05/08, 6:45 p.m. EDT Work on final projects and presentations
Week 16 May 13 – May 15		No readings	Meetup on 05/15, 6:45 p.m. EDT 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/