

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 (7 x 30)	21%
Projects (4 x 100)	40%
Final Project (1 x 140)	14%
Final Project Presentation (1 x 40)	4%
Discussion Participation (15 x 10)	15%
Data Science Context Presentations (2 x 30)	6%
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.
- 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 all five projects (including the final), and make the final presentation. If you cannot deliver your presentation in the 12/22 Meetup, you'll need to make available a recorded version of your final presentation before 12/22.
- There will also be **short ungraded hands on labs** most weeks that will help you prepare for your weekly programming assignments and projects.
- **"Discussion" and "Data Science Context PresentationS".** While this material is important, please note that this work only makes up 21% of your grade. Please do the readings, and participate in the discussions and any discussion-related group assignments, and make your two short presentations. *If you are participating and turning in your work on time, you'll receive the full 21% here.* At the same time, if you have limited time for the course, please remember to invest the majority of your efforts not in the discussion and data science context work, but 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 timely, tested, and reproducible. Cutting corners—as long as they are documented at the time of submission—is also acceptable.

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 on your own contributed work!

- **Late work policy.** Late projects cannot be accepted once solutions have been posted; you will have the option of creating an alternative project in lieu of the assigned project. Late projects can only receive a maximum score of 80%. You will be much more successful if you start early, and turn in your work on time (even if it's not perfect!).
- Students that complete all work in a satisfactory and timely manner will earn a maximum grade of A-. To earn a grade of A in MSDA 607, you'll need to demonstrate work above and beyond what is expected.

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. Errata here: <http://www.r-datacollection.com/errata/errata.pdf>.
- *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>.
- Any book on R, such as *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:

Andy Catlin
andrew.catlin@sps.cuny.edu
616-638-8344

How This Course Works:

Meetups take place every week on Thursdays from 6:30 p.m. to 7:30 p.m. ET. Please see course outline below for specific dates. You are strongly encouraged to attend; all meet-ups will be recorded.

Office Hours can be scheduled by 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 0 Aug 26 – Aug 28	Building out your Data Science Development Environment	<i>Automated Data Collection with R</i> , chapter 1.	Environment Setup <i>No assignment</i>
Week 1 Aug 29 – Sep 4	R: Data Types and Basic Operations	<i>Data Science for Business</i> , chapter 1	Meetup on 9/1, 6:30 p.m. EDT <i>No assignment</i>
Week 2 Sep 5 – Sep 11	R and SQL	<i>Automated Data Collection with R</i> , chapter 7. <i>Data Science for Business</i> , chapter 2	Meetup on 9/8, 6:30 p.m. Week 2 Assignment
Week 3 Sep 12 – Sep 18	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 9/15, 6:30 p.m. Week 3 Assignment
Week 4 Sep 19 – Sep 25	R: Exploratory Data Analysis	<i>Data Science for Business</i> , chapter 4	Meetup on 9/22, 6:30 p.m. Project 1
Week 5 Sep 26 – Oct 2	R: Working with Tidy Data	<i>Data Science for Business</i> , chapter 5	Meetup on 9/29, 6:30 p.m. Week 5 Assignment
Week 6 Oct 3 – Oct 9	R: Data Transformations	<i>Data Science for Business</i> , chapter 6	Meetup on 10/6, 6:30 p.m. Project 2
Week 7 Oct 10 – Oct 16	Web Technologies; MongoDB	<i>Automated Data Collection with R</i> , chapters 2 through 6. <i>Data Science for Business</i> , chapter 7	Meetup on 10/13, 6:30 p.m. Week 7 Assignment
Week 8 Oct 17 – Oct 23	Scraping Web Pages	<i>Automated Data Collection with R</i> , chapter 9. <i>Data Science for Business</i> , chapter 8	Meetup on 10/20, 6:30 p.m. Project 3
Week 9 Oct 24 – Oct 30	Working with Web APIs	<i>Automated Data Collection with R</i> , chapter 9. <i>Data Science for Business</i> , chapter 9	Meetup on 10/27, 6:30 p.m. Week 9 Assignment
Week 10 Oct 31 – Nov 6	Text Mining	<i>Automated Data Collection with R</i> , chapter 10. <i>Data Science for Business</i> , chapter 10	Meetup on 11/3, 6:30 p.m. Week 10 Assignment
Week 11 Nov 7 – Nov 13	Recommender Systems	<i>Data Science for Business</i> , chapter 11	Meetup on 11/10, 6:30 p.m. Week 11 Assignment
Week 12 Nov 14 – Nov 20	Graph Databases	<i>Data Science for Business</i> , chapter 12	Meetup on 11/17, 6:30 p.m. Project 4; Final Project Proposals due
<i>Nov 21 – Nov 27</i>	<i>Thanksgiving Break</i>	<i>No readings</i>	<i>No assignments</i>
Week 13 Nov 28 – Dec 4	Working with Data in the Cloud	<i>Data Science for Business</i> , chapter 13	Meetup on 12/1, 6:30 p.m. Work on final projects and presentations
Week 14 Dec 5 – Dec 11	Hadoop and Spark	<i>Data Science for Business</i> , chapter 14	Meetup on 12/8, 6:30 p.m. Work on final projects and presentations
Week 15 Dec 12 – Dec 18	Big Data Analytics	<i>Data Science for Business</i> , Appendices A and B	Meetup on 12/15, 6:30 p.m. Work on final projects and presentations
Final Exam Period Dec 19 – Dec 22		No readings	Meetup on 12/22, 6:30 p.m. 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/