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 (8 x 25)	**20%**
Projects (4 x 100)	**40%**
Final Project (1 x 200)	**20%**
Final Project Presentation (1 x 50)	**5%**
Discussion Participation (10 x 15)	**15%**
TOTAL	**100%**

Notes

- All projects and assignments, unless otherwise noted, are due end of day on Sundays.
- Each course week will be available by 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 12/17 Meetup, you'll need to make available a recorded version of your final presentation before 12/17.
- There will also be **short ungraded quizzes** each week that will help you prepare for your weekly programming assignments.
- **"Data Science Context" Discussions**. While this material is important, please note that this work only makes up 15% of your grade. Please do the readings, and participate in the discussions and any discussion-related group assignments. *If you are participating and turning in your work on time, you'll receive the full 15% 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 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.

Quality of Performance	Letter Grade	Range %	GPA/ Quality Pts.
Excellent - work is of exceptional quality	A	93 - 100	4.0
	A-	90 - 92.9	3.7
Good - work is above average	B+	87 - 89.9	3.3
Satisfactory	B	83 - 86.9	3.0
Below Average	B-	80 - 82.9	2.7

Poor	C+	77 - 79.9	2.3
	C	70 - 76.9	2.0
Failure	F	< 70	0.0

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.

Recommended Texts:

- Any book on SQL, such as *The Language of SQL* by Larry Rockoff. ISBN: 978-1435457515
- 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, the RStudio IDE, and MySQL. In addition, we will use MongoDB, Neo4J, and Hadoop. All are open source. Details for obtaining and installing the appropriate software will be provided in the first week's course materials. All of the software will work on both PCs and Macs.

Contact Information:

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

How This Course Works:

Meetups take place roughly every other week on Thursdays from 8:15 p.m. to 9:15 p.m. ET. Please see course outline below for specific dates. You are strongly encourage to attend; all meet-ups will be recorded.

Office Hours (cell phone or using GoToMeeting): take place on most Thursdays when there are not meetups scheduled, from 6:15 p.m. to 6:55 p.m. ET, or by appointment. You're encouraged to schedule an appointment, but you can try to call anytime. If you need extra help and are willing to invest the time and effort to be successful, I'll invest the additional 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 24 to 48 hours. If you do not hear back from me within 48 hours of sending an email, please resend your message.

This course is conducted entirely online. Each week, the student will have various resources made available, including weekly readings from the textbooks and additional readings provided by the instructor. Most weeks will have homework assignments to be submitted. Students are expected to complete all assignments by their due dates.

In addition, there will be five projects (including your final project and project presentation) assigned during the semester. Further details on each of these projects are available in Blackboard.

Course Outline:

Unit	Topic	Core Readings	Deliverables
Week 1 **Aug 27 – Aug 30**	**Building out your Data Science Development Environment**	*Automated Data Collection with R*, chapter 1.	**Environment Setup**
Week 2 **Aug 31 – Sep 6**	**Review of SQL**	*Automated Data Collection with R*, chapter 7.	**Office Hours on 9/3, 6:15 p.m.** **Week 2 Assignment**
Week 3 **Sep 7 – Sep 13**	**R: Data Types and Basic Operations**	tbd	**Meetup on 09/10, 8:15 p.m.** **Week 3 Assignment**
Week 4 **Sep 14 – Sep 20**	**R: Character Manipulation and Date Processing**	*Automated Data Collection with R*, chapter 8.	**Office Hours on 09/17, 6:15 p.m.** **Week 4 Assignments**
Week 5 **Sep 21 – Sep 27**	**R: Exploratory Data Analysis**	tbd	**Meetup on 09/24, 8:15 p.m.** **Project 1**
Week 6 **Sep 28 – Oct 4**	**R: Working with Tidy Data**	tbd	**Office Hours on 10/01, 6:15 p.m.** **Week 6 Assignments**
Week 7 **Oct 5 – Oct 11**	**R: Data Transformations**	tbd	**Meetup on 10/08, 8:15 p.m.** **Project 2**
Week 8 **Oct 12 – Oct 18**	**Web Technologies**	*Automated Data Collection with R*, chapters 2 through 6.	**Office Hours on 10/15, 6:15 p.m.** **Week 8 Assignment**
Week 9 **Oct 19 – Oct 25**	**Scraping Web Pages**	*Automated Data Collection with R*, chapter 9.	**Meetup on 10/22, 8:15 p.m.** **Project 3**
Week 10 **Oct 26 – Nov 1**	**Working with Web APIs**	*Automated Data Collection with R*, chapter 9.	**Office Hours on 10/29, 6:15 p.m.** **Week 10 Assignment**
Week 11 **Nov 2 – Nov 8**	**Text Mining**	*Automated Data Collection with R*, chapter 10.	**Meetup on 11/05, 8:15 p.m.** **Week 11 Assignment**
Week 12 **Nov 9 – Nov 15**	**MongoDB**	tbd	**Office Hours on 11/12, 6:15 p.m.** **Week 12 Assignment**
Week 13 **Nov 16 – Nov 22**	**Graph Databases**	tbd	**Meetup on 11/19, 8:15 p.m.** **Project 4**

Unit	Topic	Core Readings	Deliverables
Week 14 **Nov 23 – Nov 29**	**Thanksgiving Break**	No readings	No assignments
Week 15 **Nov 30 – Dec 6**	**Hadoop**	tbd	**Meetup on 12/03, 8:15 p.m.** **Final Project Proposals Due**
Week 16 **Dec 7 – Dec 13**	**Working with Data in the Cloud**	tbd	**Offfice Hours on 12/10, 6:15 p.m.** Work on final projects and presentations
Final Exam Period **Dec 14 – Dec 17**	**Work on final projects**	No readings	**Meetup on 12/17, 8:15 p.m.** **Final Projects and Presentations Due**

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/