CUNY The City University of New York | School of Professional Studies

Course Name and Number: DATA 607 – Data Acquisition and Management
Credits: 3 cr.
Prerequisite(s): none

How is this course relevant for data analytics professionals?

Data analytics professionals spend *most* of their time getting data and preparing it for analysis. This course 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, unstructured 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

Course Requirements and Interactions

- 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 weekly meeting, you'll need to make available a recorded version of your final presentation before the end of semester.
- 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

Automated Data Collection with R
A Practical Guide to Web Scraping and Text Mining
WILEY
Data Science for Business
What You Need to Know About Data Mining and Data-Analytic Thinking
Foster Provost & Tom Fawcett
O'REILLY
R for Data Science
Hadley Wickham & Garrett Grolemund
O'REILLY
Text Mining with R
A TIDY APPROACH
Julia Silge & David Robinson

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. (abbreviation used: ADC with R)
- *R for Data Science* by Garrett Grolemund and Hadley Wickham. O'Reilly, 2017. Freely downloadable here. Print copies are also available. (abbreviation used: R for DS)
- *Text Mining with R: A Tidy Approach*, Julia Silge and David Robinson. O'Reilly, 2017. Freely downloadable here. Print copies are also available. (abbreviation used: TM with R)
- *Data Science for Business*, Tom Fawcett and Foster Provost, O'Reilly, 2013. (abbreviation used: DS for B)

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 SQL ZOO: SQL Tutorial.
- 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:

DATA 607
Dr. Tati Tchoubar tatiana.tchoubar@sps.cuny.edu

How This Course Works:

Web conferencing meetings 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.

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. For the most part, you can expect me to respond to questions by email within one business day.

Course Schedule

Unit	Topic	Core Readings	Deliverables
Week 1 Aug 27 - Sep 01	Building out your Data Science Development Environment; R: Data Types and Basic Operations	ADC with R, ch 1 *DS for B*, ch 1	Meetup on 08/28, 6:45 p.m. EST Week 1 Assignment
Week 2 Sep 02 – Sep 08	R and SQL	*ADC with R*, ch 7 *DS for B*, ch 2	Meetup on 09/04, 6:45 p.m. EST Week 2 Assignment
Week 3 Sep 09 – Sep 15	R: Character Manipulation and Date Processing	*ADC with R*, ch 8 *DS for B*, ch 3	Meetup on 09/11, 6:45 p.m. EST Week 3 Assignment
Week 4 Sep 16 - Sep 22	R: Exploratory Data Analysis	*R for DS*, ch 7 *DS for B*, ch 4	Meetup on 09/18, 6:45 p.m. EST Project 1
Week 5 Sep 23 - Sep 29	R: Working with Tidy Data	*R for DS*, ch 12 *DS for B*, ch 5	Meetup on 09/25, 6:45 p.m. EST Week 5 Assignment
Week 6 Sep 30 - Oct 6	R: Data Transformations	*R for DS*, ch 5 *DS for B*, ch 6	Meetup on 10/02, 6:45 p.m. EST Project 2
Week 7 Oct 07 - Oct 13	Web Technologies; MongoDB	ADC w R, ch 2-6 *DS for B*, ch 7	No Meetup Week 7 Assignment
Week 8 Oct 14 - Oct 20	Scraping Web Pages	ADC with R, ch 9 *DS for B*, ch 8	Meetup on 10/16, 6:45 p.m. EDT Project 3
Week 9 Oct 21 - Oct 27	Working with Web APIs	ADC with R, ch 9 *DS for B*, ch 9	Meetup on 10/23, 6:45 p.m. EDT Week 9 Assignment
Week 10 Oct 28 - Nov 03	Text Mining	*TM with R*, ch 1-6 *DS for B*, ch 10	Meetup on 10/30, 6:45 p.m. EDT Week 10 Assignment
Week 11 Nov 04 - Nov 10	Recommender Systems	*DS for B*, ch 11 recommenderlab	Meetup on 11/06, 6:45 p.m. EDT Week 11 Assignment

Week 12 Nov 11 - Nov 17	Graph Databases	*DS for B*, ch 12	Meetup on 11/13, 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 Nov 18 - Nov 24	Working with Data in the Cloud; Hadoop and Spark	*DS for B*, ch *13 and 14*	Meetup on 11/20, 6:45 p.m. EDT Work on final projects and presentations Tidyverse Recipes due
Week 14 Nov 25 - Dec 01	Thanksgiving break	*No readings*	*No assignments*
Week 15 Dec 2 - Dec 8	Big Data Analytics	*DS for B*, Appendices A and B	Meetup on 12/4, 6:45 p.m. EDT Work on final projects and presentations
Week 16 Dec 09 - Dec 11	Final Presentations by students	*No readings*	Meetup on 12/11, 6:45 p.m. EDT Final Project 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 SPS - Disability Services

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 SPS - Netiquette Guide

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 SPS - Academic Policies

STUDENT SUPPORT SERVICES

If you need any additional help, please visit Student Support Services