of 7

                     Automatic Zoom                     Actual Size                     Page Fit                     Page Width                                          50%                     75%                     100%                     125%                     150%                     200%                     300%                     400%

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?  
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:  
Notes  
• All discussions, projects, and assignments--unless otherwise noted--are due end of day on Sundays.  
• Each course week will be available on the previous Friday at 6:00 a.m. ET.  
• Course Completion Requirements. To pass this course, you must complete:  
• at least six assignments,  
• three projects,  
• the final project,  
• and make the final project presentation.  
If you cannot deliver your final presentation in our 05/14 Meetup, you’ll need to make available a recorded  
version of your final presentation before 05/14.  
• There are some short ungraded hands on labs that will help you prepare for your weekly programming  
assignments and projects. You don’t need to turn these in.  
• “Discussion”, “Data Science in Context Presentations”, and “TidyVerse Recipes” While this material is  
important, please note that this work only makes up only 23% 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. At the  
same time, if you have limited time for the course, please remember to invest most 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. Data Science in Context presentation are open to a topic of your choosing related to the course  
material and data science in general. They are 5min only and will take place as we dive into the week’s material  
in the weekly meetups. Up to three presentations will take place on each Meetup call. Make sure to schedule your  
presentation early by using the respective Forum in the Discussion Forum.  
• 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.  
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.

• 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:  
• R for Data Science (2e) by Hadley Wickham and Garrett Grolemund. This is the primary text for the course.  
Freely readable here: https://r4ds.hadley.nz/ The first edition is also available in print.  
• Text Mining with R: A Tidy Approach, Julia Silge and David Robinson. O’Reilly, 2017. Freely readable  
https://www.tidytextmining.com/index.html  
• 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  
Print copies for each of these texts are available for purchase on Amazone and elsewhere.  
Recommended Texts:  
• The Language of SQL, 3rd Edition by Larry Rockoff. ISBN: 978-0137632695.  
• Practical SQL, 2nd Edition, by Anthony deBarros.  
• Alternatively, there are many excellent on-line resources, such as http://sqlzoo.net.  
Optional Recommended Text:  
This text book provides good insights into the Data Science discipline in the Business world. Consider using this  
textbook only if available to you for the weekly discussions and general reference. It is not a coding book, but with  
help germinating ideas for some of the projects, including your final project.  
• Provost, F., & Fawcett, T. (2013). Data science for business: [what you need to know about data mining and data-  
analytic thinking]. Sebastopol, Calif., O'Reilly.  
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 PostgreSQL and MongoDB. 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.

Instructor Contact Information:  
Andy Catlin  
andrew.catlin@spsmail.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. You are strongly encouraged to  
attend; all meetups will be recorded. You are not required to attend the meetups, but you are responsible for  
watching the recording if you’re not able to attend.  
We will use Zoom for our meetup call:  
https://us06web.zoom.us/j/99498039934?pwd=Q0R1MVNMbi9oYjRGaEhmVVJCMFEzZz09  
Occasional Weekend Office Hours A few times during the semester, we’ll have optional additional office hours  
on topics of interest, especially around data engineering.  
Regular 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 that might be of interest to other students on the course Slack Channel. Here’s  
the link to join our Slack channel:  
https://join.slack.com/t/data607spring2025/shared\_invite/zt-2yb4r1g5w-  
tcpoJ5OdlC\_WhsWhnKri2w  
For the most part, you can expect me to respond to questions by email or Slack within one business day.

Course Schedule  
Unit Topic Core Readings Assign Projs. Final  
Proj.  
Prop.  
Final  
Proj.  
Final  
Proj.  
Press.  
Discuss Data  
Sc.i in  
Cox.  
Tidy  
Verse  
Wk. 1 Data Ethics; R: Data  
Types and Basic  
Operations  
R for Data Science (2e),  
https://r4ds.hadley.nz/ chapters 1, 2,  
3, 7,14,15,27,28  
50 20  
Wk. 2 R and SQL R for Data Science (2e),  
https://r4ds.hadley.nz/ chapter 8, 9,  
10, 23  
50  
Wk. 3 R: Character  
Manipulation and  
Date Processing  
R for Data Science (2e),  
https://r4ds.hadley.nz/ chapters 14,  
15, 16, 17, 18, 19  
50  
Wk.4 R: Exploratory Data  
Analysis; Data  
Imputation  
R for Data Science (2e),  
https://r4ds.hadley.nz/ chapters 11,  
12, 13, 20  
90  
Wk. 5 R: Working with Tidy  
Data  
R for Data Science (2e),  
https://r4ds.hadley.nz/ chapters 4, 5,  
6  
50 40  
Wk. 6 R: Data  
Transformations;  
Feature Engineering  
Feature Engineering and Selection,  
http://www.feat.engineering/ chapter  
1  
90  
Wk. 7 Web Technologies;  
MongoDB  
R for Data Science (2e),  
https://r4ds.hadley.nz/ chapter 25  
50 20  
Wk. 8 Scraping Web  
Pages  
R for Data Science (2e),  
https://r4ds.hadley.nz/ chapter 26  
70  
Wk. 9 Working with Web  
APIs  
httr quickstart vignette,:  
https://cran.r-  
project.org/web/packages/httr/vignett  
es/quickstart.html  
50 40 25  
Wk. 10 Natural Language  
Processing  
Text Mining w/ R,  
https://www.tidytextmining.com/, ch  
1-4  
50  
Wk. 11 Recommender  
Systems  
Mining Massive Datasets,  
http://www.mmds.org/, ch 9  
50 40  
Wk. 12 Graph Databases Selected readings from web 20  
Wk. 13 Working with Data in  
the Cloud;  
Deployment  
No Readings 90 15  
Spring Break  
Wk. 14 Automated Machine  
Learning  
No Readings 50  
Wk. 15 Final Presentation No Readings 150 30  
Total Points 300 270 20 150 30 140 50 40  
Total Percents 30% 27% 2% 15% 3% 14% 5% 4%

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 BrightSpace, 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/

**Course Learning Materials**

**Required Texts:**

* *R for Data Science* (2e) by Hadley Wickham and Garrett Grolemund.  This is the primary text for the course.  Freely readable here: <https://r4ds.hadley.nz/>  The first edition is also available in print.
* *Text Mining with R: A Tidy Approach*, Julia Silge and David Robinson. O’Reilly, 2017.   Freely readable
* 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>

Print copies of each of these texts is available for download.

**Recommended Texts:**

* Any book on SQL, such as [*The Language of SQL, 3rd Edition*](https://www.amazon.com/gp/product/013763269X) by Larry Rockoff.  ISBN: 978-0137632695.  My favorite text for PostgreSQL is [*Practical SQL, 2nd Edition*](https://www.amazon.com/Practical-SQL-2nd-Beginners-Storytelling/dp/1718501064), by Anthony deBarros.  Alternatively, there are many excellent on-line resources, such as [http://sqlzoo.net](http://sqlzoo.net/).

**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 PostgreSQL and MongoDB. 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.

Discussions (general format for discussion rubrics) Background. Discussions are worth 10 points. Historical grading: 7 points for any reasonable initial post; 3 points for any reasonable response. 0 points if no post made by end of day Sunday. Due date: 0 points if no post made by end of day Saturday. Initial Post (7 points) Provide post that partially addresses discussion topic. [4 points] Provide post that fully addresses discussion topic. [6 points if citation required, otherwise 7 points). When assignment requires a citation: Provide at least citation, using MLA or APA format. Citation should be properly linked from Blackboard [1 point] Extra Credit: Your initial post is on par with professional quality blog posts [2 points] Response (3 points) Provides post that shows you are reflecting on initial post of another student, with some level of explained agreement, disagreement, or additional thoughts [3 points]