COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Relational Databases and SQL Programming
for Research and Data Science
P8180

CLASS SESSIONS

Tuesdays 5:30pm – 8:20pm
January 18, 2022 – April 26, 2022
Hammer Health Sciences Library, Room LL 204
Zoom link: https://columbiacuimc.zoom.us/j/93571043643 Passcode: P8180-001

INSTRUCTOR

Debby D'Angelo, MS
dd2542@cumc.columbia.edu
Office Hours: by appointment via Zoom videoconference
Zoom link: https://columbiauniversity.zoom.us/j/7397662140

ASSISTANT

Jason Chua, MPH
jac2209@columbia.edu
Office Hours: TBA
Zoom link: TBA

COURSE DESCRIPTION

In this course, you will learn to design and build relational databases in Access and MySQL, to write and optimize queries using the SQL programming language, and to pull data from the Web using APIs. Application of skills learned in this course will be geared toward research and data science settings in the healthcare field; however, these skills are transferrable to many industries and application areas.

You will begin the course examining the pitfalls of using Excel spreadsheets as a data storage tool and then learn how to build properly-designed relational databases to eliminate the issues related to spreadsheets and maintain data integrity when storing and modifying data. You will then learn two aspects of the SQL programming language: 1) the data manipulation language (DML), which allows you to retrieve data from and populate data into database tables (e.g., SELECT, INSERT INTO, DELETE, UPDATE, etc.), and 2) the data definition language (DDL), which allows you to create and modify tables in a database (e.g., CREATE, ALTER, DROP, etc.). You will additionally learn how to optimize SQL queries for best performance, and use advanced SQL functions. Finally, you will learn how to pull data from the web via APIs.

COURSE STRUCTURE

This course is formatted as a flipped classroom, which means that lectures are watched on video in advance, outside of class meetings, and "homework" is done during class meetings under the guidance of the teaching team. This format allows you to learn material at your own pace by re-watching lectures as needed, and makes your time spent on exercises more efficient by allowing you to get clarification from the teaching team in real-time as you work.

The beginning of each class meeting will be a review session dedicated to the current week's lecture material that you watched on video prior to class. A weekly quiz will immediately follow each review session (see "Quizzes" under the "Assessment and Grading Policy" section of the syllabus).

The remainder of the class session will be spent working on programming assignments. During virtual sessions, you will be sent to breakout rooms in Zoom to have the opportunity to work with others on each assignment; during in-person sessions, you may work with others in the classroom. However, you will be responsible for submitting your own individual work for grading. If you have a question as you work, please have someone from your group ask for help, either from the Zoom breakout room or in person, and someone from the teaching team will assist you. Each class session will be recorded and posted to Canvas for you to review later.

COURSE LEARNING OBJECTIVES

Students who successfully complete this course will be able to:

- Describe the limitations of using Excel spreadsheets for data storage
- Design and build fully-functioning, normalized, relational databases using Microsoft Access and MySQL
- Formulate SQL queries to 1) create and modify database tables, and 2) populate and retrieve data from database tables using Microsoft Access, MySQL, R, and SAS
- Optimize SQL queries for best performance
- Extract data from websites via APIs

PREREQUISITES

Students should have taken an introductory biostatistics course and have some familiarity working with data. No prior database development experience or SQL programming experience is required.

TEXTBOOKS

There are no required textbooks for this course; however, the books below are excellent, comprehensive resources on building databases in Access and MySQL. The MySQL book also provides a clear and thorough overview, with examples, on SQL programming syntax.

- Access 2013: The Missing Manual by Matthew MacDonald. ISBN: 9781449357412
- Murach's MySQL, 2nd Edition by Joel Murach. ISBN-13: 978-1890774820

SOFTWARE

We will use **Microsoft Excel**, **Microsoft Access**, **MySQL**, **R,** and **SAS** software in this class. **You must have a laptop that can run Windows, even if it is a Mac**, because Microsoft Access is only available for Windows. MySQL should additionally be installed on Windows to allow these programs to communicate when we connect them later in the semester. Excel, R, and SAS software can be installed on either platform. You may also use classroom computers to complete assignments.

A free download of Microsoft Office is available for students, faculty, and staff (UNI login required):

Students: http://columbiait.onthehub.com
Faculty and Staff: http://cuit.columbia.edu/microsoft-office

MySQL software is also free and can be downloaded from https://dev.mysql.com/downloads/. Installation instructions will be posted to Canvas toward the middle of the semester.

ASSESSMENT AND GRADING POLICY
Final grades will be based on:

In-Class Assignments	20%
Quizzes	20%
Access Database Project	25%
MySQL Database Project	25%
Participation	10%

In-Class Assignments: (20% of final grade) We will cover a total of 20 programming assignments in class, which will be scored out of 100%. The in-class assignments are designed to give practice implementing the concepts covered in lecture, and to gain experience with troubleshooting issues that may arise while programming. You are highly encouraged to take advantage of the flipped classroom format and seek feedback from peers and the teaching team as you work on assignments; however, **you are expected to submit your own unique work for grading**.

Grading rubrics will be posted simultaneously with the assignment so that you may keep the grading scheme in mind as you work. In addition to a completed grading rubric, you will receive detailed feedback from the TAs to help you improve your skills and clarify concepts that you may have missed in the assignment.

Assignments will be due 3 days following class on **Fridays at 5:30pm.** Assignments submitted past the deadline without prior approval will have 10% out of the 100% total deducted per day, and if an assignment is not submitted prior to the next class meeting, no credit will be given.

Occasionally, optional assignments may be offered on topics beyond the scope of the syllabus, which will also be graded out of 100%, and can be used to replace your lowest assignment grade.

Your final in-class assignment grade will be calculated as follows:

In-Class Assignment Average Score (out of 100%) * 20%

Quizzes: (20% of final grade) You will take weekly quizzes in class to assess your comprehension of the lectures you watched to prepare for the current week. Lecture videos will be unlocked for the following week immediately after class on Tuesdays. Each video contains demonstrations that are designed to help build your skills on various topics. You should follow along with each demo, using the files provided, to prepare for the following week's quiz (and ultimately, in-class assignments).

The quizzes will be administered on Canvas and will be published following a review session at the beginning of class, during which you can ask questions to clarify your understanding of the lecture material. Since the quizzes will happen toward the beginning of class, **please join class on time so that you can benefit from the review, and avoid missing the quiz**. No make-up quizzes will be given.

Quizzes will be open-book with approximately 10 questions and will consist of a mix of true/false, multiple choice, fill in the blank, matching, and short answer questions. There will be a **15-minute time limit** for each quiz, and the grades will be scored out of 100%. **You may not communicate with others while taking the quiz.**

Your final quiz grade will be calculated as follows:

Average Quiz Score (out of 100%) * 20%

Access Database Project: (25% of final grade) You will work in teams of 2 to build a small, normalized, relational database in Microsoft Access using the skills you learn during the semester. The database should have a healthcare-related focus, but you may choose another topic with permission. This project will be "scaffolded," with portions of it due at intervals throughout the semester. An overview of the project and its components and deadlines will be available on Canvas. Requirements for individual components of the project will be posted separately.

Your Access Database Project grade will be calculated as follows:

Access Database Project Score (out of 100%) * 25%

MySQL Database Project: (25% of final grade) You will work with your Access Database Project partner to recreate your database in MySQL. An overview of the project and its components and deadlines will be available on Canvas later in the semester.

Your MySQL Database Project grade will be calculated as follows:

MySQL Database Project Score (out of 100%) * 25%

Participation: (10% of final grade) Students who attend class on time, remain for the duration of the class sessions, and ask questions during class can expect to earn full marks for participation. If you expect to be absent or late to class, please let us know in advance. Also, if you have technical problems that keep you from connecting to Zoom, please email us as soon as possible to let us know.

Your Participation grade will be calculated as follows:

Participation Score (out of 100%) * 10%

Letter Grades

A+	Reserved for highly exceptional achievement.
A	Excellent. Outstanding achievement.
A-	Excellent work, close to outstanding.
B+	Very good. Solid achievement expected of most graduate students.
B	Good. Acceptable achievement.
B-	Acceptable achievement, but below what is generally expected of graduate students.
C+	Fair achievement, above minimally acceptable level.
C	Fair achievement, but only minimally acceptable.
C-	Very low performance.
F	Failure. Course usually may not be repeated unless it is a required course.

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath. https://www.mailman.columbia.edu/about/mission-history/public-health-oath

Academic Integrity

Each student in this course is expected to adhere to the Mailman School Honor Code, available online at https://www.mailman.columbia.edu/people/current-students/community-standards/student-honor-code

You are encouraged to utilize the flipped classroom format of this course to teach, and learn from, your peers; however **all assignments and quizzes must consist only of your own individual work. Only the database projects may consist of more than one student's work.**

For assignments, and the database projects, you may never submit as your own work:

- Any part of another student's work (including past or current students in any section of this course)
- Any part of a past or current solution file that was posted in class
- Any part of a file obtained from external sources (e.g., internet sites, files created by other programmers, etc.)

During quizzes, you are not permitted to:

- Compare answers with another student, copy answers from another student, or submit answers that you obtained from an external source
- Communicate with anyone other than the teaching team
- Record questions and/or disseminate questions or solutions to others

Additionally, you may not share course materials such as lecture slides, videos, assignments, database files, or any other resources posted to our course's Canvas site with any individual who is not currently enrolled in this course without permission of the instructor.

Those found to have violated the standards of academic integrity may receive a score of zero on any assessments in question, and may be referred for further disciplinary action.

Disability Access

In order to receive disability-related academic accommodations, students must first be registered with the Office of Disability Services (ODS). Students who have or think they may have a disability are invited to contact ODS for a confidential discussion at 212.854.2388 (V) 212.854.2378 (TTY), or by email at disability@columbia.edu. If you have already registered with ODS, please speak to your instructor to ensure that they have been notified of your recommended accommodations by Meredith Ryer (mr4075@cumc.columbia.edu), Assistant Director of Student Support and Mailman's liaison to the Office of Disability Services.

COURSE SCHEDULE

The following is a tentative course schedule for the Spring 2022 semester. The schedule may change to accommodate the needs of the class.

All course-related documents can be found under the **Files** section of Canvas.
Lecture videos and quizzes can be accessed under the **Modules** section of Canvas.
Assignments should be submitted through the **Assignments** section of Canvas.

Canvas site: https://courseworks2.columbia.edu/courses/150803

Week 1 – Course Introduction, Limitations of Excel for Data Storage

1/18/22

Learning Objectives:

- Define course objectives and logistics
- Describe limitations of using Excel spreadsheets for data storage

Lectures (In Class):

- Course/Syllabus Overview
- Limitations of Excel for Data Storage

Quiz: Introduce Yourself!

Assignments:

- In-Class Assignment 1 (Excel) – Due Friday, 1/21 at 5:30pm
- In-Class Assignment 2 (Excel) – Due Friday, 1/21 at 5:30pm

Week 2 – Relational Databases, Table Design

1/25/25

Learning Objectives:

- Define database, relational database
- Build database tables containing:
  - variable names that follow good naming conventions
  - correct data types assigned to each variable
  - correct identifiers selected as primary key
  - indexes on variables, where appropriate
  - validation rules applied to variables, where appropriate
  - a combo box linked to a lookup table
  - no HIPAA identifiers
- Connect tables to create relationships that:
  - are joined on the correct fields
  - have referential integrity enforced
  - show the correct relationship type (e.g., 1-to-1, 1-to-many)
  - have cascade update/delete selected appropriately

Lectures Due Today: Week 2 Lecture Videos

Quiz: Quiz for Week 2 Lectures

Assignments:

- In-Class Assignment 3 (Access) – Due Friday, 1/28 at 5:30pm
- Database Project Part 1: Proposal – Due Friday, 2/4 at 5:30pm

Week 3 – Data Anomalies, Normalization Rules, Access Query Builder, Table Joins

2/1/22

Learning Objectives:

- Describe insertion, deletion, and update anomalies
- Build database tables normalized up to 3rd Normal Form
- Compare normalization rules to "Tidy Data" rules
- Create select queries using the Access query builder
- Differentiate query results produced with outer vs. inner joins

Lectures Due Today: Week 3 Lecture Videos

Quiz: Quiz for Week 3 Lectures

Assignments:

- In-Class Assignment 4 (Access) – Due Friday, 2/4 at 5:30pm
- In-Class Assignment 5 (Access) – Due Friday, 2/4 at 5:30pm
- Database Project Part 1: Proposal – Due Friday, 2/4 at 5:30pm

Week 4 – SQL Data Manipulation Language (DML)

2/8/22

Learning Objectives:

- Describe the types of operations performed in the SQL DML
- Write SQL queries containing the following clauses: SELECT, FROM, WHERE, GROUP BY, HAVING, ORDER BY, INSERT INTO, UPDATE, DELETE
- Write SQL queries containing outer and inner joins
- Write SQL queries containing a subquery
- Add aggregate functions and functions that calculate across columns to SELECT queries
- Compare the use of joins vs. subqueries when multiple tables are involved in one query
- Apply SQL Style Guide conventions to SQL query syntax

Lectures Due Today: Week 4 Lecture Videos

Quiz: Quiz for Week 4 Lectures

Assignments:

- In-Class Assignment 6 (Access) – Due Friday, 2/11 at 5:30pm
- In-Class Assignment 7 (Access) – Due Friday, 2/11 at 5:30pm
- Database Project Part 2: Access Tables – Due Friday, 2/25 at 5:30pm

Week 5 – SQL Data Definition Language (DDL), SQL in R

2/15/22

Learning Objectives:

- Describe the types of operations performed in the SQL DDL
- Write SQL queries to create and alter database tables
- Write SQL queries using the 'sqldf' package in R

Lectures Due Today: Week 5 Lecture Videos

Quiz: Quiz for Week 5 Lectures

Assignments:

- In-Class Assignment 8 (Access) – Due Friday, 2/18 at 5:30pm
- In-Class Assignment 9 (R) – Due Friday, 2/18 at 5:30pm
- Database Project Part 2: Access Tables – Due Friday, 2/25 at 5:30pm

Week 6 – Data Entry Forms

2/22/22

Learning Objectives:

- Create new data entry forms
- Add objects to forms
- Modify form and object properties
- Add subforms and ensure their correct connection to main form
- Add buttons to forms to open/close forms and navigate between records
- Write VBA code to alter form and object properties, and to perform data validation between different variables
- Describe the purpose of Data Entry Mode
- Apply good design practices in form layout and formatting (e.g., labels, colors, object spacing and alignment)

Lectures Due Today: Week 6 Lecture Videos

Quiz: Quiz for Week 6 Lectures

Assignments:

- In-Class Assignment 10 (Access) – Due Friday, 2/25 at 5:30pm
- Database Project Part 2: Access Tables – Due Friday, 2/25 at 5:30pm

Week 7 – Reports

3/1/22

Learning Objectives:

- Design a report to show summary of records without grouping
- Design a report to show summary of records with grouping
- Apply good design practices in report layout and formatting (e.g., labels, colors, object spacing and alignment)

Lectures Due Today: Week 7 Lecture Videos

Quiz: Quiz for Week 7 Lectures

Assignments:

- In-Class Assignment 11 (Access) – Due Friday, 3/4 at 5:30pm
- Database Project Part 3: Access Queries – Due Friday, 3/18 at 5:30pm

SPRING BREAK – NO CLASS

3/8/22 No class meeting. Have fun, rest, and enjoy!

Week 8 – Introduction to MySQL, SQL DML in MySQL

3/15/22 Learning Objectives:

- Locate relevant parts of MySQL Workbench interface
- Open and save new SQL scripts in MySQL
- Run queries using SQL DML in MySQL

Lectures Due Today: Week 8 Lecture Videos

Quiz: Quiz for Week 8 Lectures

Assignments:

- In-Class Assignment 12 (MySQL) – Due Friday, 3/18 at 5:30pm
- Database Project Part 3: Access Queries – Due Friday, 3/18 at 5:30pm

Week 9 – SQL DDL and SQL DML in MySQL

3/22/22 Learning Objectives:

- Create a new schema in MySQL
- Add and modify tables within a schema
- Set primary keys, foreign keys, and indexes within tables
- Produce EER diagram to show table relationships
- Populate and retrieve data using SQL DML
- Develop a "front end" interface in Access for MySQL tables
- Apply a password to the Access "front end"

Lectures Due Today: Week 9 Lecture Videos

Quiz: Quiz for Week 9 Lectures

Assignments:

- In-Class Assignment 13 (MySQL) – Due Friday, 3/25 at 5:30pm
  Database Project Part 4: Access Data Entry Forms and Reports – Due Friday, 4/1 at 5:30pm

Week 10 – Views, Case Statements, Triggers, Connecting Access to MySQL

3/29/22

Learning Objectives:

- Define view and describe its uses
- Use a case statement within a SELECT query to conditionally recode a variable
- Apply triggers to validate data
- Establish ODBC connection between Access and MySQL

Lectures Due Today: Week 10 Lecture Videos

Quiz: Quiz for Week 10 Lectures

Assignment:

- In-Class Assignment 14 (MySQL) – Due Friday, 4/1 at 5:30pm
- In-Class Assignment 15 (Access, MySQL) – Due Friday, 4/1 at 5:30pm
- Database Project Part 4: Access Data Entry Forms and Reports – Due Friday, 4/1 at 5:30pm

Week 11 – Common Table Expressions (CTEs), Temporary Tables and Window Functions

4/5/22

Learning Objectives:

- Define a CTE and its uses
- Use a CTE instead of a subquery
- Define a temporary table and its uses
- Use a temporary table instead of a subquery
- Apply different window functions for data wrangling

Lectures Due Today: Week 11 Lecture Videos

Quiz: Quiz for Week 11 Lectures

Assignment:

- In-Class Assignment 16 (MySQL) – Due Friday, 4/8 at 5:30pm
- Database Project Part 5: MySQL Tables, Views & Access Front End – Due Friday, 4/22 at 5:30pm

Week 12 – More Joins, Query Optimization

4/12/22

Learning Objectives:

- Join tables in SQL queries with USING statement
- Perform self-joins, full joins, and cross joins in SQL queries
- Formulate SQL queries to retrieve data efficiently
- Compare speed of data retrieval when formulating a query in multiple ways

Lectures Due Today: Week 12 Lecture Videos

Quiz: Quiz for Week 12 Lectures

Assignments:

- In-Class Assignment 17 (MySQL) – Due Friday, 4/15 at 5:30pm
- In-Class Assignment 18 (MySQL) – Due Friday, 4/15 at 5:30pm
- Database Project Part 5: MySQL Tables, Views & Access Front End – Due Friday, 4/22 at 5:30pm

Week 13 – Pulling Web Data Using APIs

4/19/22

Learning Objectives:

- Define API
- Compare Web data retrieval via API vs. retrieval via Web scraping
- Extract data from the Web via API using the httr package in R

Lectures Due Today: Week 13 Lecture Videos

Quiz: Quiz for Week 13 Lectures

Assignment:

- In-Class Assignment 19 (R) – Due Friday, 4/22 at 5:30pm
- In-Class Assignment 20 (MySQL) – Due Friday, 4/22 at 5:30pm
  Database Project Part 5: MySQL Tables, Views & Access Front End – Due Friday, 4/22 at 5:30pm

Week 14 – Database Project Working Session

4/26/22

- During this class meeting, database project teams will be given time to work on the database project and receive feedback from the teaching team and peers

Assignment:

- Database Project Part 6: Final Submission with Corrections – Due Friday, 5/6 at 5:30pm

Database Project Due

Friday, 5/6/22

Due Today **(by 5:30pm)**:

- Database Project Part 6: Final Submission with Corrections