MSDS 691
Relational Databases
Bootcamp 2019

Instructor: Nick Ross

Time and Location: All locations and times can be found on the MSDS Calendar. Exams are Friday from 9-10.

Contact Information: ncross@usfca.edu

Office Hours: Given that this is a bootcamp class I will generally be around and will not be holding "official" office hours. Feel free to reach me by slack if you need specific assistance.

Overview and Objectives

- **All HW submissions should be online. 10AM submission time.**
- **Assign groups the first day and have them send an email with their group.**

This class provides an intro to accessing and analyzing structured data using SQL (pronounced "Sequel") and Pandas. Despite the best efforts of software engineers, being able to manipulate and transform data is a foundational data science technique and these two tools are the most common solutions to these problems.

There are a number of different SQL variants; in this class we will focus on learning PostgreSQL, which is often just called "Postgres." Much of what is covered in this class will apply to other versions of SQL, such as MySQL, RedShift or Vertica. PostgreSQL specific topics, of which there are a few, will be called out during the class.

For Pandas we will be using whatever is the most recent version, though anything past version 0.23 should be fine.

There is an official textbook for this course (*Data Management for Analytics*). Many students, however, find using some of the resources below to be useful. None are required.

- *Learning SQL* by Alan Beaulieu: Is an easy read and covers the basics. However it focuses specifically on MySQL and doesn't cover some topics which we cover in this course.
- *PostgresSQL Documentation*: The PostgreSQL documentation is incredible and you should use it as a resource.
- *Learning Postgres* by Salahaldin Juba, Achim Vannahme, and Andrey Volkov. This is another resource which covers much of what we use in this course – some students enjoy their writing style.
- *PostgreSQL (2nd Edition)*: Korry Douglas. And another PostgreSQL resource.
- *Python for Data Analysis*: Wes McKinney. The author of this book is the person who wrote Pandas. I find it to be fun to read but organized in a way that is difficult to use as a reference.

Learning Objectives

At the end of this 7 week course, students will:

- Understand data storage in Relational Databases, including comparisons to other database types.
- Write SQL queries to extract and transform data:
  - Grouping, ordering, sorting, collapsing, joins and aggregating
  - Use string, math and date functions to manipulate data contained within tables
  - Apply window functions to simplify queries
- Use server specific functions to navigate database information tables
- Define different data types
- Identify when to properly use indexes
- Define and classify tables as being in 1st, 2nd or 3rd normal forms
- Basic efficiency: understand query performance and how to increase efficiency
- Use large-query best practices to write maintainable, easy-to-read queries
- Write code in Pandas to extract and transform data:
  - Grouping, ordering, sorting, collapsing, joins and aggregating
  - Use string, math and date functions to manipulate data contained within tables
  - Apply window functions to simplify data management

Syllabus

- Week #1
  - Class #1 (7/10)
    - SQL Mod #1: Rows and Columns
    - SQL Mod #2: Basic Manipulations
    - HW assigned: 0A, 0B, 0C, 1A
  - Class #2 (7/12)
    - SQL Mod #3: Subqueries
    - Pandas Mod #1: Introduction
    - HW Due: 0A, 0B, 0C
    - HW assigned: 1B, 2A

- Week #2
  - Class #3 (7/16)
    - Pandas Mod #2: More Manipulations
    - SQL Mod #5: Aggregations
    - HW Due: 1A
    - HW Assigned: 2B, 3A
  - Class #4 (7/19)
    - Pandas Mod #3: Aggregations
    - SQL Mod #4: Rabbit Hole[1]
    - HW Due: 1B, 2A
    - HW Assigned: 3B, 4B
  - Exam #1 (7/19): Classes #1 to #3
- Week #3
  - Class #5 (7/23)
    - SQL Mod #6: Dates and Types
    - SQL Mod #7: Averages
    - HW Due: 2B, 3A
    - HW Assigned: 4A
  - Class #6 (7/26)
    - SQL Mod #8: Joins
    - Pandas Mod #4: Joins
    - HW Due: 3B, 4B
    - HW Assigned: 6A, 6B
  - Exam #2 (7/26): Class #1 to #5
- Week #4
  - Class #7 (7/30)
    - SQL Mod #9: Advanced Joins
    - SQL Mod #10: Information Schema and Views[2]
    - HW Due: 4A
    - HW Assigned: 7A, 7B
  - Class #8 (8/2)
    - SQL Mod #11: Window Functions
    - Pandas Mod #5: Window Functions

[1] Only if time permits.
[2] If time permits

    - HW Due: 6A, 6B
    - HW Assigned: 8A, 8B
  - Exam #3 (8/2): Classes #1 to #7

- Week #5
  - Class #9 (8/6)
    - SQL Mod #12: Performance
    - SQL Mod #4 & #10 if necessary
    - HW Due: 7A, 7B
  - Class #10 (8/9)
    - Review
    - HW Due: 8A, 8B
  - Exam #4 (8/9): Classes #1 to #10

Grading

The MSDS program considers a grade of "A" to represent exceptional work with respect to both the instructor's expectations and peer student achievements. A grade of "B" represents the expected outcome, what is called "competence" in a business setting. A "C" grade represents achievements lower than the instructor's expectations for competence in the subject. A grade of "F" represents little or no work in the course.

Grades are determined using the table below with brief explanations of each component afterward.

Professionalism	10%
Homework	10%
Exams (4)	80%
Total	100%

Homework

Most of the homework is self-explanatory. Importantly, unless explicitly stated, all homework is to be done in groups. A **hard copy** of the assignment should be turned in at **the start** of class. One assignment should be turned in per group and assignments turned in after I start lecturing receive, at most, $\frac{1}{2}$ credit.[3]

The purpose of the homework is to provide exercises to assist in your learning. They are graded very, very, very lightly. Within each assignment there are three sections of questions:

[3]Things I do not care about: group dynamics. If you are in a group and the person assigned to the turn the homework in fails to do so on time, then all are penalized. I also do not care about printer issues.

(a) First Five, (b) Main Problems and (c) Additional Problems. Only the First Five and the Main Problems need to be turned in. The additional problems are there for you and its your choice if you wish to complete them. Answer keys will be provided (electronically) for all assignments.

To maximize your likely success I recommend doing the following: Before meeting with your group make sure you can do the First Five on your own. Attempt to do all of the Main Problems and *then* meet with your group. Undertake the additional problems if you have time.

Important: do the homework by hand and *then* check with your computer. Many students rely on the computer to "figure things out". The computer will not be available for any exam (or most interviews).

Exams

There are four exams during this five weeks of this course. All exams are similar: they are closed book, closed note, no computer. A description of a dataset will be given to you and questions, similar to what you'd find in the homework will be asked. You can get a sense of what the exams are like by looking in the Appendix of the textbook. These were exams given in previous years. Though keep in mind that the content will be different.

While some more theoretical questions may be asked, the vast majority of the exam is applying the concepts from class and the homework.

The exam will have time pressure. I'm a firm believer in a "if you *know* the material, you will be fast." In order to prepare for the exam you should be able to do the homework problems with ease.

Professionalism

As you can see from the final grade table, quite a bit of the grade involves being a professional. In this class, that means:

- Showing up on time
- Showing up prepared
- Contributing to class
- Being responsible to your group when doing group work
- Behaving toward your peers and professors in a way consistent with what would be expected in a work environment

Students who show up on time, work through the example problems and pay attention in class have a tendency to do extremely well in this class. The opposite if also true; students who fail to make an effort, fail to contribute in class and fail to pay attention have a tendency to do poorly in this class.

An approximation of your participation grade may be made by following the rubric below:

Grade	Attendance / Promptness	Participation / Professionalism
A	Prompt and Complete attendance each day	Full attention (e.g. no playing on the phone) and **frequent, informed** contribution to class discussion.
B	Prompt and Complete attendance each day	Full attention (e.g. no playing on the phone) and **regular** contribution to class discussion.
C	Prompt and Complete attendance each day	General attention and occasional contribution to class.
D/F	Frequently late / tardy	Marginal Attention (texting, playing on a computer) and little or no classroom contribution.

Table 1: Professionalism Grading Rubric

Odds and Ends:

- Attendance is required. Unless a student gives me prior warning, all absences are considered unexcused.
- The quickest way to reach me is via slack / email.
- Class participation is strongly encouraged and can significantly help your grade.
- Cheating is not tolerated. At all. Unless an assignment is clearly designated as group work, I expect it to be done alone. USF's honor code can be found online and I expect it to be followed. Disciplinary action will be taken against any student found violating this code.
- As a Jesuit institution committed to cura personalis - the care and education of the whole person – USF has an obligation to embody and foster the values of honesty and integrity. USF upholds the standards of honesty and integrity from all members of the academic community. All students are expected to know and adhere to the University's Honor Code. You can find the full text of the code online at www.usfca.edu/academic_integrity. The policy covers:
  - Plagiarism – intentionally or unintentionally representing the words or ideas of another person as your own; failure to properly cite references; manufacturing references.
  - Working with another person when independent work is required.
  - Submission of the same paper in more than one course without the specific permission of each instructor.
  - Submitting a paper written by another person or obtained from the internet.
  - The penalties for violation of the policy may include a failing grade on the assignment, a failing grade in the course, and/or a referral to the Academic Integrity Committee.

- If you are a student with a disability or disabling condition, or if you think you may have a disability, please contact USF Student Disability Services (SDS) at 415 422-2613 within the first week of class, or immediately upon onset of disability, to speak with a disability specialist.

  If you are determined eligible for reasonable accommodations, please meet with your disability specialist so they can arrange to have your accommodation letter sent to me, and we will discuss your needs for this course. For more information, please visit: http://www.usfca.edu/sds or call (415) 422-2613.

- All students are expected to behave in accordance with the Student Conduct Code and other University policies (see http://www.usfca.edu/fogcutter/). Open discussion and disagreement is encouraged when done respectfully and in the spirit of academic discourse. There are also a variety of behaviors that, while not against a specific University policy, may create disruption in this course. Students whose behavior is disruptive or who fail to comply with the instructor may be dismissed from the class for the remainder of the class period and may need to meet with the instructor or Dean prior to returning to the next class period. If necessary, referrals may also be made to the Student Conduct process for violations of the Student Conduct Code.

- As an instructor, one of my responsibilities is to help create a safe learning environment on our campus. I also have a mandatory reporting responsibility related to my role as a faculty member. I am required to share information regarding sexual misconduct or information about a crime that may have occurred on USFs campus with the University. Here are other resources:

  - To report any sexual misconduct, students may visit Anna Bartkowski (UC 5th floor) or see many other options by visiting our website: www.usfca.edu/student_life/safer.
  - Students may speak to someone confidentially, or report a sexual assault confidentially by contacting Counseling and Psychological Services at 415-422-6352.
  - To find out more about reporting a sexual assault at USF, visit USFs Callisto website at: www.usfca.callistocampus.org.
  - For an off-campus resource, contact San Francisco Women Against Rape (SFWAR) (415) 647-7273 (www.sfwar.org).