Course: **Data Acquisition & Management (AIM 5001)**
Credits: 3 Credits / Graduate
Pre/Coreqs: N/A
Instructor: James Topor
Instructor Contact: james.topor@yu.edu

COURSE OVERVIEW

Data Acquisition and Management focuses on the data structures, data design patterns, algorithms, methods, and best practices for the pre-modeling phases of data science workflows, including problem formulation, gather, analyze, explore, model, and communicate, analytics programming focuses on the gather, analyze, and explore workflow steps. This comprises the "data wrangling" work which is where most data scientists spend the majority of their time. Because data science is iterative, this preparatory work informs the modeling phase. Often, the creation and validation of new models requires going back for additional data, different data transformations, and exploration of data distributions. In short, every effective data scientist needs to master analytics programming. Course topics include reading from or writing to databases, text files, and the web; shaping data into "tidy" data frames, exploratory data analysis, data imputations, feature engineering, and feature scaling.

COURSE LEARNING OUTCOMES

By the end of this course, students will be able to:

- Obtain data from structured and unstructured data sources.
- Transform, modify and explore data as needed to support and validate modeling operations.
- Engineer data features based on business and modeling constraints
- Perform basic Exploratory Data Analysis
- Create high quality explanatory narratives and visualizations in support of reproducible analytical work

REQUIRED MATERIALS

- Larry Rockoff, *The Language of SQL, 2nd Edition*. Addison-Wesley (2017).
- S. Juba, A. Volkov, *Learning PostgreSQL 11, 3rd Edition*. Packt Publishing (2019).
- Wes McKinney, *Python for Data Analysis*, 2nd edition, O'Reilly (2018).
- Mark Pilgrim, *Dive Into Python 3*, freely available web-based content: http://diveintopython3.problemsolving.io/table-of-contents.html
- Zheng, Alice and Casari, Amanda, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly (2018).

Web-based readings and videos on related topics will also be assigned.

Relevant Software, Hardware, or Other Tools:
We will make use of the PostgreSQL relational database platform, the MongoDB NoSQL database platform, and the Neo4J Graph Database platform. We will make use of Python via the freely available Anaconda environment, including Jupyter Notebooks and the Spyder IDE. Students are also welcome to use Google's Colab platform where feasible. 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.

ASSIGNMENTS & GRADING

Approach to Assignments. All Python-based projects and assignments are to be written in IPython (Jupyter) notebooks and submitted via Canvas. Non Python-based assignments will be submitted directly within the AIM 5001 Canvas portal.

Evaluation Criteria. All course projects will be evaluated like work assignments from a demanding employer. The primary evaluation basis is adherence to the deliverables stated in each assignment's functional requirements. To achieve a top grade, students must also adhere to best practices for software engineering principles, including reproducibility; following appropriate coding guidelines; and DRY. Furthermore, assignments must be clearly and concisely written using proper English language grammar and should present relevant supporting text in a logical flow. Presentations should include an appropriate level of detail for their intended audience.

Assignments	Grading
Discussions / Weekly Response Assignments (14 x 10 Pts) The fourteen module-specific discussions will focus primarily on use cases related to the topics covered within the associated module. Students will prepare short responses to discussion questions, which will be used to prompt group discussion.	13%
Assignments (10 x 100pts, lowest Assignment grade dropped) On most weeks when projects are not due, there will be short-form ("mini-project") assignments to help reinforce the current learning material. These assignments may include completing tasks using course analytical tools. Some assignments may require working in small groups.	25%
Projects (3 x 100pts) Students will work individually and in teams on three data acquisition and management projects. At the end of the course, each student will have a portfolio of increasingly complex projects ready to show an employer.	24%
Midterm Exam (100 Points) Students will address a series of practical challenges derived from the content of AIM 5001 Course Modules 1 through 7.	7%
Final Project (150pts) and Presentation (50pts) Working individually or as part of a small team, students will create a formal proposal that specifies one or more research questions to be answered based on data students have chosen to work with. They will then attempt to answer the research questions described in their proposal using the skills they've developed during the semester. Students will present their final projects to their peers for feedback.	20%
Final Exam (100 Points) Students will address a series of practical challenges derived from the content of AIM 5001 Course Modules 1 through 14.	11%

- All projects and assignments, unless otherwise noted, are due end of day on Sundays.
- Each week's materials will be made available via Canvas no later than the previous Friday at 6:00 a.m. ET.
- **Course Completion Requirements**: As a prerequisite to passing this course, you must complete all four projects (including the final), and make the final presentation during

the final class session. Failure to either submit any one of the four projects or present your final project will preclude you from achieving a passing grade in this course. Please note that completion of the four projects is not the sole determinant of whether you will receive a passing grade: however, failure to submit any one of the four will prevent you from achieving a passing grade.

- **Discussions / Weekly Response Assignments**: While this material is important, please note that this work only makes up 13% of your grade. Please do the readings, and participate in the discussions and any discussion-related group assignments. 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 your work can be reproduced by others. 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.
- **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! Failure to provide proper citations for any third party components of the content you submit will be treated as a violation of the Katz School's **Student Code of Conduct** and will be treated accordingly.
- **Late work policy**. Please note: **Assignments, discussion responses, exams, and projects cannot be accepted after their due dates for any reason**. Any assignment, discussion, exam, or project that is not submitted before its associated deadline will automatically be assigned a grade of **ZERO**. You will enhance your chances for success in this class 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 *Data Acquisition & Management*, you'll need to demonstrate work above and beyond what is expected.

GRADING SCALE:

Quality of Performance	Letter Grade	Range %	GPA/ Quality Pts.
Excellent - work is of exceptional quality	A	93 – 100+	4
	A-	90 - 92.9	3.7
Good - work is above average	B+	87 - 89.9	3.3
Satisfactory	B	83 - 86.9	3
Below Average	B-	80 - 82.9	2.7
Poor	C+	77 - 79.9	2.3
	C	70 - 76.9	2
Failure	F	< 70	0

How This Course Works:

Online Live Sessions are held every week on ***Tuesdays from 5.40 p.m. to 7:00 p.m. ET***, with the exception of Katz School official holidays. You are strongly encouraged to attend these weekly classes since each will include opportunities for hands-on learning via discussions and case studies as well as a presentation / demonstration of many of the concepts you will need to use for any assignment or project due that week. You are also required to bring your laptop to these Live Sessions as this will serve to facilitate the hands-on learning segments. Class dates can be found in the Course Schedule shown on the following page.

Office Hours can be scheduled by appointment. If you need extra help and are willing to invest the time and effort to be successful, your instructor will make time available to help you. ***But…***you should not be asking for extra help on a project or assignment the day before or the day it is due, since this will indicate that you are not investing the time and effort needed to be successful in the course.

You are encouraged to ask questions on Canvas where other students will be able to benefit from your inquiries. For the most part, you can expect your instructor to respond to questions asked either via email or via Canvas within one business day.

KATZ SCHOOL CLASS ATTENDANCE POLICY

Students are expected to attend all scheduled classes in their entirety. Students who fail to fulfill this requirement will receive an academic penalty appropriate for the course work missed.

Students may not miss 30% or more of their scheduled class. If a student misses 30% or more of a course during the semester, they will receive a final grade of "F." This grade will be reflected on the student's official university transcript.

For programs within clinical components students may not miss 20% or more of any course, clinical or not. At the Katz School, this pertains only to students in the Speech Language Pathology program. If a student misses 20% or more of a course during the semester, they will receive a final grade of "F." This grade will be reflected on the student's official university transcript.

If the student is absent because of a disability which is documented with the Office of Disability Services at Yeshiva, falls ill or there are other extenuating circumstances, the student must inform the instructor in advance. The instructor may require appropriate documentation to make any exception to this policy.

COURSE SCHEDULE

Students should expect to spend a minimum of 9 hours each week outside of the classroom sessions on the materials, assignments, discussions, and projects required for this course.

Module	TOPIC	SCHEDULE OF MAJOR ASSIGNMENTS
Module 1 Jan 19 – Jan 24 Class: T Jan 19	Intro to SQL + PostgreSQL	M1 Assignment
Module 2 Jan 25 – Jan 31 Class: T Jan 26	SQL Aggregation & Grouping + Principles of Database Design	M2 Assignment
Module 3 Feb 1 – Feb 7 Class: T Feb 2	Python Basics: Syntax, Data Types, Objects, Control Flow	M3 Assignment
Module 4 Feb 8 – Feb 14 Class: T Feb 9	Python Data Structures, Comprehensions, & Functions	M4 Assignment
Module 5 Feb 15– Feb 21 Class: T Feb 16	Text Processing	**Project 1 Due**
Module 6 Feb 22 – Feb 28 Class: T Feb 23	NumPy: Numerical Python ** ***Final Project Requirements Distributed*** **	M6 Assignment
Module 7 Mar 1 – Mar 7 Class: T Mar 2	Pandas Series & Dataframe Objects	**Project 2 Due**
Module 8 Mar 8 – Mar 14 Class: T Mar 9	Exploratory Data Analysis + Creating Visualizations in Python	M8 Assignment **Midterm Exam**
Module 9 Mar 15 – Mar 21 Class: T Mar 16	Working with Web Data	M9 Assignment **** 1st Draft of Final Project Proposal Due ****
Module 10 Mar 22 – Mar 28 Class: T Mar 23	Data Preparation & Feature Engineering	**Project 3 Due**
Mar 29 – Apr 4	***** NO CLASSES: UNIVERSITY CLOSED FOR PASSOVER *****	N/A
Module 11 Apr 5 – Apr 11 Class: T Apr 6	Data Reshaping & Aggregation in Pandas	M11 Assignment **Final Project Proposal Due**
Module 12 Apr 12 – Apr 18 Class: T Apr 13	Text Mining	M12 Assignment
Module 13 Apr 19 – Apr 25 Class: T Apr 20	NOSQL Databases: MongoDB	M13 Assignment
Module 14 Apr 26 – May 2 Class: T Apr 27	Graph Databases: Neo4J	No assignments or projects due: Work on Final Projects + prep for Final Exam
Module 15 May 3 – May 9 Class: T May 4	Final Project Presentations + Writeups Due **** Final Project Presentations Tuesday May 4 ****	**Final Exam** **** Final Project Writeups Due Friday May 7 ****

ONLINE LEARNING POLICIES

Online Learning Formats
Your course consists of two online learning formats:

- **Synchronous Learning**: Live real time sessions using Zoom (webinar system). During these sessions, we will be able to see and talk with each other. Attendance is required.
- **Asynchronous Learning**: Pre-created content such as videos, assignments, links and articles. There will also be the use of community and collaboration tools like discussion boards and group tools. These sessions are not in real-time but rather involve engagement over the course of each week.

Online Learning Engagement Policy
A successful online class only happens when there is an active community. Students are required to attend both the weekly live synchronous sessions and participate in other community building activities such as the discussion boards.

Netiquette
Netiquette is a set of rules for behaving properly in an online course. Often the anonymity of online courses can cause a lapse in judgement when learners are excited or passionate about a subject. This can lead to statements that could be demeaned as offensive. You are all adults and are treated as such. However, it is still important to talk about these issues. The following bullet points cover some basics communicating in an online course:

- Be sensitive to the fact that there will people with different cultural and linguistic backgrounds, as well as different political and religious beliefs.
- Use good taste when composing your responses in Discussion Forums. Swearing and profanity is also part of being sensitive to your classmates and should be avoided.
- Don't use all capital letters when composing your responses as this is considered "shouting" on the Internet and is regarded as impolite or aggressive.
- Be respectful of your others' views and opinions. Avoid "flaming" (publicly attacking or insulting) them as this can cause hurt feelings and decrease the chances of getting all different types of points of view.
- Be careful when using acronyms. If you use an acronym it is best to spell out its meaning first, then put the acronym in parentheses afterward, for example: Frequently Asked Questions (FAQs). After that you can use the acronym freely throughout your message.
- Use good grammar and spelling (avoid using text messaging shortcuts).
- If you aren't sure what someone meant, consider asking for clarification.
- Remember that your peers are not required to respond to your specific post, so don't be offended if your question goes unanswered.

UNIVERSITY POLICIES & RESOURCES

ACCESSIBILITY AND ACCOMMODATIONS
The Office of Disability Services collaborates with students, faculty and staff to provide reasonable accommodations and services to students with disabilities. Students with disabilities who are enrolled in this course and who will be requesting documented disability-related accommodations should make an appointment with the Office of Disability Services, (646) 592-4132, rkohn1@yu.edu, during the first week of class. Once you have been

approved for accommodations, please submit your accommodation letter to ensure the successful implementation of those accommodations. For more information, please visit: http://yu.edu/Student-Life/Resources-and-Services/Disability-Services/

ACADEMIC INTEGRITY

The submission by a student of any examination, course assignment, or degree requirement is assumed to guarantee that the thoughts and expressions therein not expressly credited to another are literally the student's own. Evidence to the contrary will result in appropriate penalties.

Academic integrity is a set of responsibilities and standards to facilitate high academic quality and rigor with the purpose of clarifying expectations and student conduct. The submission by a student of any coursework, or degree requirement is assumed to guarantee that the thoughts and expressions therein not expressly credited to another are literally the student's own. Examples of violations on academic integrity are, but not limited to:

- Cheating
- Plagiarism
- Dishonesty
- Assisting or attempting to assist another student in an act of academic dishonesty
- Providing papers, essays, research, or other work to aid another student in Intentional Misrepresentation
- Engaging in unauthorized cooperation with other individuals in completing assignments or examinations
- Submitting the same assignment, in part or whole, in more than one course, whether at YU or another institution, without prior written approval from both faculty members.

For more information, visit http://yu.edu/registrar/grad-catalog/

STUDENT SUPPORT SERVICES

If you need any additional help, please visit Student Support Services: http://yu.edu/academics/services/