Data 601: Introduction to Data Science

Fall 2019

General Information

Meeting Times and Location

Wednesday, 7:10pm-9:40pm, at

Instructor	Email	Office Location & Hours
Ben Payne	benpayne@umbc.edu	Prior to class or by appointment in the atrium of Albin O. Kuhn Library or at Subway

I check my email daily and generally will respond to questions in the evening hours. When questions reference code that you have written, please include the Jupyter notebook as an attachment. If I do not respond to your email within 24 hours, please remind me that you are waiting for a response.

I am available before class for questions and help. Outside of that timeframe, please email me to schedule an appointment.

Description

The goal of this class is to give students an introduction to and hands on experience with all phases of the data science process using real data and modern tools. Topics that will be covered include data formats, loading, and cleaning; data storage in relational and non-relational stores; data analysis using supervised and unsupervised learning using Python; data visualization; and scaling up for Big Data.

Prerequisite

Students must be enrolled in the Data Science Program. Other students may be admitted with instructor permission. Students are expected to have experience with programming.

Course Learning Objectives

Upon completion, students will understand:

- Understand issues relating to acquisition, cleaning and loading of data.
- Be able to perform data analysis using Python.
- Understand the basics of how data can be presented and visualized.
- Understand issues involved when the analysis scales up to Big Data.
- Understand pros and cons of the different data analysis methods
- Be able to provide estimates of requirements for storage and compute associated with analysis

Course Materials

Optional Texts (not required)

- "Python Data Science Handbook" by Jake VanderPlas. O'Reilly Media
- "Data Wrangling with Python: Tips and Tools to Make Your Life Easier" by Jacqueline Kazil and Katharine Jarmul. O'Reilly Media
- "Think Like a Data Scientist: Tackle the data science process step-by-step" by Brian Godsey. Manning Publications

Please review options at the UMBC library . You do not need to buy any books.

Recommended Software and Hardware

All software used in this course is free.

- Web browser capable of running Jupyter Notebooks.
- Jupyter Notebook software using Anaconda.
- A laptop. Electrical outlets are available in the classroom. UMBC Wi-Fi is available.
- Ability to connect your laptop to the classroom projector. VGA and HDMI plugs are available. You can also use the classroom computer to display your content to the class.
- Paper and pen or pencil for in class exercises.

Course Format and Assignments

Students will complete assigned homework, readings, essays, quizzes, two projects, and a final project. This course incorporates a variety of hands-on labs and practical exercises to engage students and prepare them for challenges they may encounter in the workplace.

Students will occasionally present their solutions to homework assignments in class. Projects will also involve presentations.

The final project will provide students opportunity to showcase what they have learned in a format similar to what they will encounter in a professional work setting.

Course Communication

Email is preferred. If you are sending code to be reviewed, provide the Jupyter notebook as an attachment. Emails should not exceed 1MB in size.

Course Syllabus

Subject to revision; 20190601

wk	Class	reading	topic	homework
1	Jan 30		Overview of Data Science; Data structures *Python and Jupyter; CSV, JSON, XML*	essay on "50 years of data science" or "Very Short History of Data Science"
2	Feb 6	AtBSwP ch 1; LtPwP ch 1	Python in Jupyter *Lists, dictionaries; series, dataframes*	List rotation, count words and characters
3	Feb 13		Getting data *Science; Scraping; negotiation*	HTML in XML
4	Feb 20		Data cleanup *Missing data, outliers, encoding, smoothing data*	Ragged CSV
5	Feb 27	AtBSwP	Automation; Reports *Create PDFs, HTML; Excel from Python; email*	
6	Mar 6		Project 1 presentations: data characterization Submit proposal on (week 4 – 1 day)	Nothing
7	Mar 13	ItP ch 1	Math *Statistics, Probability*	Dice roll
8	Mar 20		No class – Spring Break	Nothing
9	Mar 27		Time series analysis; linear regression	imputation
10	Apr 3		Clustering; regression *K-means, linear regression; document analysis*	
11	Apr 10		Project 2 in-class exercise to reproduce results	nothing
12	Apr 17		Property graphs	
13	Apr 24	AoAiSaE ch1	Scaling and estimation	
14	May 1		Concurrency, Elasticity	
15	May 8		Cost/benefit analysis, Ethics	
16	May 22		Final Project Presentations: merged data	

Automate the Boring Stuff with Python, A. Sweigart | Learning to Program with Python, R. L. Halterman | Introduction to Probability, C. M. Grinstead | Mining Massive Datasets, J. Leskovec and A. Rajaraman and J. D. Ullman | Art of Approximation in Science and Engineering, S. Mahajan

Grading Criteria

Students are expected to participate in class discussions. Missed lectures result in an attendance score of zero for that lecture. Students will occasionally present results of analysis to the entire class. Student participation in class is expected, so watching recorded lectures outside of class does not count towards attendance.

Assignments are graded on a 100 point scale. Homework assignments are typically due 24 hours prior to the start of class.

Course work	Grade Distribution
Attendance and Participation	10%
Homework	30%
Project 1	20%
Project 2	20%
Final Project and Presentation	20%

Final Grade will be computed as follows:

90-100%	A
80 to 89%	B
70 to 79%	C
60 to 69%	D
<60	F

Course Materials

All Course Materials created by the instructor are covered by a "Creative Commons - Attribution 4.0 International" license. Recordings of lecture materials and voice recordings of the instructor will be made during class.

All assignments, quizzes, essays, project proposals, and projects should be submitted via Blackboard.

When submitting a Jupyter notebook to Blackboard, include the assignment requirements as a comment at the top of the notebook. Name the .ipynb file to be reflective of the purpose of the assignment.

If your assignment uses a file found on the Internet, specify the complete URL in the submitted notebook. If your approach to an assignment relies on local file(s) for input to the Jupyter Notebook, please include the input file(s) in your submission to Blackboard. Blackboard allows multiple files to be uploaded per assignment. If a large file (more than 1 MB) needs to be submitted, provide a link to that data hosted on UMBC's box.com service. If the large file is associated with a Jupyter notebook submitted via Blackboard, include the link to UMBC's box.com in the notebook as a comment. For small data sets (less than or equal to 1 MB), students submit code and data in Blackboard.

Essays should be submitted to Blackboard as plain text (not DOCX or PDF or images). Essays are constrained by word count; check the rubric for specific guidance. Essays should include a title. Essays should include citations. All text (including the title and citations) is included in word counts.

Do not include your name on materials submitted via Blackboard, including essays, code, and projects. Anonymous grading is enabled. Resubmissions of previously graded work are not allowed.

Course Policies

UMBC provides a range of writing assistance, which can be found in the following:

- The Writing Center: http://lrc.umbc.edu/tutor/writing-center/
- Research Guides & Tutorials: http://lib.guides.umbc.edu/tutorial

Failure to follow guidelines for each assignment, including the required format, style, length, and submission may result in at least one-letter-grade reduction (10%) on the assignment depending on the type or number of transgressions.

Each assignment is provided with a required outcome, desired output format, and grading rubric. Read the rubric to earn the best score for your assignment. If you have questions, email the instructor.

Notify the instructor if you suspect you will not be able to complete the homework by the specified deadline. Late/incomplete assignments will be accepted if an extension has been agreed to in advance. Late submissions incur 10 percentage point reduction in score, with the reduction doubling each week. Emergency situations will be handled on a case by case basis with appropriate justification or documentation.

Academic Integrity

If work produced by someone else is used, cite the contribution. If a website is referenced, specify the full URL of the source. If a student's work from previous or concurrent courses is used, specify that in the assignment submission. Uncited references for software and essays is plagiarism.

Discussion of assignments with other students is not allowed until all work has been graded. Submit your own independent work. Sharing of work is not acceptable. Explaining your approach to other students must wait until everyone has a grade for their assignment.

By enrolling in this course, each student assumes the responsibilities of an active participant in UMBC's scholarly community in which everyone's academic work and behavior are held to the highest standards of honesty. Cheating, fabrication, plagiarism, and helping other to commit these acts are all forms of academic dishonesty, and they are wrong. Academic misconduct could result in disciplinary action that may include, but is not limited to failure, suspension or dismissal.

Refer to the UMBC policy on Academic Integrity:
http://catalog.umbc.edu/content.php?catoid=17&navoid=879#academic-integrity.

Student Academic Misconduct in the Grad School is handled by the Associate Dean (AD) of the Grad School. The first step is for the instructor to consult with the AD. If the instructor and AD determine that a less serious infraction has occurred, then the AD provides written authority to the instructor to resolve the matter. The instructor then notifies the students in writing, and the students must be allowed the opportunity to provide an explanation. The student has the right to appeal the decision directly to the AD, and the student must be informed of this right. If an appeal is requested, then the AD will convene a Grievance Committee under the policy established by University of Maryland Graduate School, Baltimore (UMGSB).

Incidents are tracked by UMBC. Repeated infractions incur more severe penalties.

Student Disability Services (SDS)

UMBC is committed to eliminating discriminatory obstacles that may disadvantage students based on disability. Services for students with disabilities are provided for all students qualified under the Americans with Disabilities Act (ADA) of 1990, the ADAAA of 2009, and Section 504 of the Rehabilitation Act who request and are eligible for accommodations. The Office of Student Disability Services (SDS) is the UMBC department designated to coordinate reasonable accommodations that would allow students to have equal access and inclusion in all courses, programs, and activities of the University.

If you have a documented disability and need to request academic accommodations, please register with the Office of Student Disability Services (SDS) as soon as possible. To begin the registration process please visit the SDS website and review the registration information, including disability documentation guidelines and how to submit the SDS registration form online using the confidential data management software called Accommodate https://sds.umbc.edu/accommodations/registering-with-sds/.

Once accommodations have been approved, you and your instructors will be notified via an emailed accommodation letter from the SDS office. Both the SDS office and Shady Grove's Center for Academic Success(CAS) will work with you to ensure you receive the approved accommodations. If you have any questions or concerns, please contact the Office of Student Disability Services via disAbility@umbc.edu or phone at 410-455-2459. Please note that accommodations are not retroactive and begin once SDS sends an approved accommodation letter.

For more information on the services CAS provides, please contact Mary Gallagher (maryg@umd.edu) or visit https://shadygrove.umd.edu/student-services/center-for-academic-success.

Disclosure of Sexual Misconduct and Child Abuse or Neglect

As an instructor, I am considered a Responsible Employee, per UMBC's Policy on Prohibited Sexual Misconduct, Interpersonal Violence, and Other Related Misconduct. While my goal is for you to be able to share information related to your life experiences through discussion and written work, I want to be transparent that as a Responsible Employee I am required to report disclosures of sexual assault, domestic violence, relationship violence, stalking, and/or gender-based harassment to the University's Title IX Coordinator.

As an instructor, I also have a mandatory obligation to report disclosures of or suspected instances of child abuse or neglect.

The purpose of these reporting requirements is for the University to inform you of options, supports and resources; you will not be forced to file a report with the police. Further, you are able to receive supports and resources, even if you choose to not want any action taken. Please note that in certain situations, based on the nature of the disclosure, the University may need to take action.

[source]

Week	Reading	Exercises	Homework
Date	Subject		