INST 447 – Data Sources and Manipulation
Section 0101, Online

Instructor: Yla Tausczik
E-mail: ylatau@umd.edu
Phone: (301) 405-2058
Office: 2118C Hornbake, South Building
Office Hours: Mondays and Tuesdays 3-4:30pm
Office Hours Location: Office hours will be online via WebEx at https://umd.webex.com/meet/ylatau

TA: Zhuoni Jie
Email: zjie@umd.edu
Office: 4117A Hornbake, South Building
Office Hours: Wednesdays and Thursdays 3-4:30pm
Office Hours Location: Office hours will be online via WebEx at https://umd.webex.com/meet/zjie or in person by appointment only

"There's the joke that 80 percent of data science is cleaning the data and 20 percent is complaining about cleaning the data" (Kaggle founder Anthony Goldbloom, The Verge 2017). Data science involves the transformation of structured and unstructured data into insights using data analytic methods. Tasked with these transformations, data scientists, must acquire skills to masterfully ingest, process, clean, wrangle, reformat, store and summarize many different forms of raw data. Raw data are often large, complex, biased and messy. Data scientists must also learn how to identify imperfections, biases, and other problems in data and correct for these problems.

This course will introduce basic concepts in data manipulation including data formats and structures (e.g. data frames, csv, xml, json); data ingestion; data cleaning and validation (e.g. missing values, recoding, visualization); data wrangling (e.g. aggregation, subsetting, merging, reshaping); and data storage and standards. Throughout the course students will be encouraged to critically think about data. Students will be asked to consider the origins of the data, sources of bias in the data, the best ways to summarize and represent the data; the meaning of data analytic results; and how best to present results to decision makers. Through homework assignments, projects, and in-class activities, you will practice working with these techniques and develop data analytic skills.

LEARNING OBJECTIVES

After successfully completing this course you will be able to:

- Identify imperfections, biases, and other problems in data sets
- Clean up, standardize, and normalize data to prepare for data analysis
- Extract data from a variety of data types and formats

- Collect large data sets through scalable, automated means, such as web scrapers
- Transform data among a variety of formats and standards
- Explain ethical and equity issues with the collection and use of data

COURSE MATERIALS

Software: The following software is necessary for you to successfully complete the homework, exams, and project for this course. Every student will need access to this software at home.

Required:

- Python 3 (https://www.python.org/downloads/). We will primarily use Python for data manipulation.
  - Pandas Data Analysis Library (pandas)
  - Other supplementary libraries (e.g. numpy, matplotlib etc.)
- Jupyter Notebooks (http://jupyter.org/install). We will use Jupyter Notebooks for completing labs and creating data science reports.
- OpenRefine (http://openrefine.org/). We will use OpenRefine to perform data cleaning steps.

Optional:

- Microsoft Excel, Open Office Calc, or Google Spreadsheets. Microsoft Excel is available for Macintosh through the university's TERPware website (https://terpware.umd.edu). Open Office Calc is a free software spreadsheet application available online (https://www.openoffice.org/product/calc.html). Google Spreadsheets can be found on Google Drive (https://www.google.com/drive/). You may find it helpful to inspect some of your data as spreadsheets (if it is small enough).

Readings: Completing the required reading for the class is essential to understanding the core concepts of data processing and manipulation. In order to learn, you must review the material multiple times. The required reading will consist of tutorials, book chapters, articles and papers and will be posted on ELMS/Canvas.

COURSE ACTIVITIES

Lecture Videos: Each unit will have 2-5 short video lectures that will introduce key concepts, new technologies, and new methods. These videos will provide you with the background that you need to know to get the most out of the labs.

Labs: This course is set up as a "lab class" for data science. Course time will be focused getting dirty with data. Each unit there will be a lab that asks you to develop skills related to the topic of the unit (e.g. aggregation, regular expressions). Each lab will focus on a data set and one or more operations. In order to do the labs successfully it will be important to do the readings and watch the videos. You can work on these individually or in pairs, but all work and code must be independently created. If you work with someone else on the lab you need write down the name of your partner in your submitted file.

Programming Assignments: There will be a total of 5 programming assignments. These are your opportunity to apply concepts learned in videos and labs to real problems and data sets. These assignments will be approximately 2 page reports composed using either a word processing application (e.g. Google Docs, Word) or Jupyter notebooks. You are expected to work *individually* to answer the specific problems that are assigned. Completed assignments will be submitted via Canvas/ELMS. These act as mini-data science project reports.

Quizzes: In order to learn and understand the material fully it is important to review and revisit it multiple times. Each unit will have a quiz to be taken after completing the readings, watching the videos, and finishing the lab. These quizzes will be primarily about the material learned in this unit, but will also include one or two review questions from prior units. Quizzes will help you review the material learned in the unit and identify areas that you may still have some confusion. You will be able to drop your lowest scoring quiz.

Exams: There will be one midterm and one cumulative final each worth 20% of your final grade. These exams provide an opportunity for you to test and demonstrate your understanding of the concepts, techniques, and problems associated with data ingestion, cleaning, manipulation and storage.

Grading:

Labs (X 10)	20%
Programming Assignments (X 5)	20%
Quizzes (X 10)	20%
Exams	40%

- Midterm (20%)
- Final (20%)

Grades will be assigned based on the total percent of points earned, using the following rubric. Grades will be rounded to the nearest 10th of a percent. Please come and talk to me early if you are think that there might be a problem.

A	90.0-100% (A- 90.0-92.9%)
B	80.0-89.9% (B+ 87.0-89.9%); B- 80.0-82.9%)
C	70.0-79.9% (C+ 77.0-79.9%; C- 70.0-72.9%)
D	60.0-69.9% (D+ 67.0-69.9%; D- 60.0-62.9%)
F	0-59.9%

COURSE POLICIES

Excused Absences: If an assignment due date or exam is a religious holiday for you, please let me know at least one week in advance, so an alternate due date can be set. Missed quizzes and exams with an excused absence must be made up within 2 weeks of the original deadline. Notification and documentation of an excused absence must be provided as soon to the missed deadline as possible. Missed assignments, quizzes, or exams without a documented, excused absence cannot be made up and will receive a score of 0.

Late Work: Timely submission of the completed assignments is essential. The due date of each assignment will be stated clearly in the assignment description on ELMS/Canvas. Late assignments will be penalized by 10% if they are turned in within one week of the due date and 50% if they are turned in within three weeks of the due date. No credit will be given for assignments that are more than three weeks late except for excused absence. All work must be turned in by the last scheduled day of classes Tuesday May 12th, 2020.

Regrading: Fairness in giving grades is very important to me, at the same time both our time is best spent on helping you learn the material. Regrading of assignments, quizzes, and exams must be turned in within one week of receiving the graded work. They must be submitted as a written document in which you include the graded work, an explanation of what you believe was missgraded, and an explanation for why you think it should be given a different score. For any regrade requests, the entire assignment will be regarded and your score may go up or down.

Extra Credit: I very rarely offer extra credit opportunities. I believe that the labs, quizzes, programming assignments, and exams are the best way to practice the course objectives and to show mastery of the material. If you are having difficulty scoring well on these assignments I'm happy to work with you during office hours to help you study more effectively and to improve your grades. In addition, if you can demonstrate on the comprehensive final exam that you have learned more than your grade reflects I will raise your final grade to be within one letter grade of your grade on the final exam (before any curve is applied). For example, if you scored 90.0% on the final before any curve you would receive at least a B- in the class regardless of your other grades.

Other Policies: Other policies relevant to undergraduate courses are found here: http://ugst.umd.edu/courserelatedpolicies.html. Topics that are addressed in these various policies include academic integrity, student and instructor conduct, accessibility and accommodations, attendance and excused absences, grades and appeals, copyright and intellectual property.

GETTING HELP

ELMS/Canvas Discussion Threads: One of the most difficult parts of an online course, is that it takes extra effort to create a community. For this reason, I am requiring that all content related questions be asked using ELMS/Canvas discussion threads rather than by email to the TA or me. My goal is to create a community in these discussion threads in which you can ask and answer each other's questions. Research has demonstrated that learning communities are invaluable resource for students trying to learn new skills; they can provide you with information and understanding, they can provide you with the opportunity to help others giving you the chance to practice and demonstrate mastery of new skills, they can help provide motivation and a sense of belonging. The only rule is that you cannot post questions asking for answers to assignments/labs/quizzes or post answers to assignments/labs/quizzes; you must ask questions about the underlying concepts. **In addition, the TA or I will respond to all questions at least once per weekday (excluding spring break).**

Office Hours: Please visit me and/or the TA online during office hours if you want extra help. We won't give you the answers to the assignments, but we will go over the material with you and help answer your questions. This is an opportunity to ask questions about the material covered in the reading materials or videos. If you are having trouble in the course please talk to me as soon as possible. If you do poorly or lower than you expected on the first exam, it is imperative that you attend office hours so that we can figure out the problem early.

Email: Feel free to email me or the TA about personal matters only (e.g. excused absences, grades). We won't respond to emails about content questions, these should be posted on ELMS/Canvas discussion threads. I will respond to your emails within 48 hours. However, I usually do not respond to emails in the evenings or on the weekends.

ACADEMIC DISHONESTY

Cheating in any form (copying, falsifying signatures, plagiarism, etc.) will not be tolerated. It will result in a referral to the Office of Student Conduct irrespective of scope and circumstances, as required by university rules and regulations. There are severe consequences of academic misconduct, some of which are permanent and reflected on the student's transcript. If you have any questions regarding the University's policies on scholastic dishonesty, please see http://osc.umd.edu/OSC/Default.aspx.

It is very important that you complete your own assignments, and do not share files (excluding raw data), partial work or final work. For this class **I consider sharing partially processed data to constitute copying** and is not allowed for any of the exercises, homework assignments, or exams.

University of Maryland Code of Academic Integrity

The University of Maryland, College Park has a nationally recognized Code of Academic Integrity, administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. For more information on the Code of Academic Integrity or the Student Honor Council, please visit http://shc.umd.edu/SHC/Default.aspx.

ACCOMMODATIONS

Please come and see me as soon as possible if you think you might need any special accommodations for disabilities. In addition, please contact the Disability Support Services (301-314-7682 or http://www.counseling.umd.edu/DSS/). Disability Support

Services will work with us to help create appropriate academic accommodations for any qualified students with disabilities. If you experience psychological distress during the course of the semester you can get professional help at the Counseling Center (301-314-7651 or http://www.counseling.umd.edu