INST 447 – Data Sources and Manipulation
Section 0101,
Van Munching Hall (VMH), Room 1203
Tue/Thu. 12:30 PM– 1:45 PM

Instructor: Yla Tausczik
E-mail: ylatau@umd.edu
Phone: (301) 405-2058 (I rarely check voicemail, email instead)
Office: 2118C Hornbake, South Building
Office Hours: Tuesday/Thursday 2:00pm-3:30pm, By Appointment.

"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. In addition you must bring a charged laptop to class with this software and/or find a "lab partner" to work with in class who brings a charged laptop.

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, plotnine, etc.)
- Jupyter Notebooks (http://jupyter.org/install). We will use Jupyter Notebooks for composing data science reports.

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. You should finish the reading by **Thursday before class** of the assigned week.

COURSE ACTIVITIES

Class Activities: This course is set up as a "lab class" for data science. Class time will be focused getting dirty with data rather than lecture. Each week there will be an in class exercise that asks you to develop skills related to the topic of the week (e.g. aggregation, regular expressions). Each exercise will focus on a data set and one or more research questions. In order to do the exercises successfully it will be important to do the readings before class as well as to attend class. You can work on these as a group in class, but all work and code must be independently created. Exercises will be graded as credit/no credit. You can drop your two lowest scores.

Homework: There will be a total of 4 homework assignments. These are your opportunity to apply concepts learned in class to real problems and data sets. These assignments will be approximately 2 page reports composed using Jupyter notebooks. You are expected to work *individually* to answer the specific problems that are assigned. Completed assignments will be submitted via Canvas/ELMS. For each assignment you must turn in a copy of your Jupyter notebook as both a .ipynb and a .html file. These act as mini-data science project reports.

Group Project: Over the course of the semester you will also define and complete your own data science project. You will identify data set(s) and research question(s); follow the steps in the data science pipeline to extract insights from data; and report on your results. The only requirements for this project is that 1) it must be centered on data, 2) it must tell us something interesting, and 3) to get an A you must go beyond what we learn in class. You will work on this project as a group of 3 to 4 members. Early in the semester you will turn in a project proposal that outlines your goals for the project. All projects must be approved by me. At the end of the semester you will present your results in class and write up your results as blog post composed in Jupyter notebook. These will be published online.

Exams: There will be one midterm and one cumulative final each worth 25% of your final grade. These exams provide an opportunity for you to test your understanding of the concepts, techniques, and problems associated with data manipulation. In order to learn and understand the material fully it is important to review and revisit it multiple times.

Grading:

Class Activities 10%
- 12 in class exercises (drop 2)

Homework 20%
- 4 programming assignments

Group Project 20%
- Project proposal (2%)
- Project presentation (5%)
- Project report (13%)

Exams 50%
- Midterm (25%)
- Final (25%)

Grades will be assigned based on the total percent of points earned, using the following rubric. Grades will be rounded to the nearest 10^{th} 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. 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. Late assignments will be penalized by 10% if they are turned in within one week of the due date and 50% if they are more than one week late. With the exception of group project presentations, which cannot be turned in late. All work must be turned in by the last scheduled day of class Monday December 10th, 2018.

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.

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.

OFFICE HOURS

Please visit me during office hours. This is an opportunity to ask questions about the material covered in the reading materials or in lecture. 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 come to office hours so that we can figure out the problem early.

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/).

COURSE SCHEDULE

Week	Week	Due (Exercises due Thursdays Assignments due Tuesdays)
1	Thinking about Data	Exercise 1
2	Data Structures	Exercise 2
3	Data Storytelling	Exercise 3
4	Data Wrangling	Exercise 4 Project Proposal
5	Data Cleaning	Programming Assignment 1 Exercise 5
6	Data Ingestion	Exercise 6
7	Data Storage	Programming Assignment 2 Exercise 7
8	**Review and Midterm (Thu. Oct. 18)**	
9	Advanced Data Ingestion: APIs	Exercise 8
10	Text Analysis	Exercise 9
11	Time Series	Programming Assignment 3 Exercise 10
12	Advanced Data Ingestion: Web Scraping	Programming Assignment 4 Exercise 11
13	Advanced Data Cleaning	
14	Applications	Exercise 12
15	Presentation & Review	Project Presentation Project Report
Final (Monday Dec 17 1:30-3:30pm)		

This schedule is for planning purposes and may change. See ELMS/Canvas for current information and deadlines.