DATA (CDAD-UH 1001Q)

SPRING 2022

Instructor:	Azza Abouzied azza@nyu.edu	Lectures:	MW 11:50 – 13:05 @ C2 W006.
		Zoom	https://nyu.zoom.us/j/96525338713

Course Description: Data is everywhere. We have massive data sets keeping track of humanity's everyday minutiae from babies born to calories consumed, friends made to crimes committed. How can we use these data to make useful predictions and gain insights into ourselves and humanity in general? This course introduces the basics of *data-driven thinking* and covers topics such as wrangling, exploration, analysis, prediction and storytelling through data visualization.

Course Page: The course is organized as a weekly living online book that can be accessed here:
https://azzadev.github.io/databook

Course textbook:

- John W. Foreman. **Data Smart: Using Data Science to Transform Information into Insight**. 1st. Wiley, Nov. 2013.

Getting help: We will be using slack to handle all communications. We will use it as a shared class forum and a means for you to directly contact me or other students. Feel free to post questions, thoughts, findings on the class forum. These offline discussions are as important as in-class ones and a means for you to get quick feedback from your peers.

You can join the slack workspace using htis link and then make sure that your are on the private spring 2022 channel:
datacoreworld.slack.com

Learning Outcomes:

- Solve real-life problems through *data-driven thinking*

- Transform data into valuable insights, decisions and products with the help of mathematics and statistics

- Communicate effectively using the language of data-driven analysis: explain how data was collected, sampled, cleaned, transformed, analyzed and visualized

- Have a personal toolbox of data analysis techniques that can be applied to different problems including optimization, regression, and statistical learning

- Accurately critique analysis outcomes, predictions, journalistic and scientific presentations of data

- Understand the limitations and biases of data-driven techniques

To be an expert at data-driven thinking, you need more than a single course and you will need to drastically expand the small toolbox we will build at the end of this class. You will also need to gain lots of experience solving many problems. The most important learning outcomes are that you become sufficiently excited about data-driven thinking that you apply it when suitable to your own personal or professional problems even after the course concludes and that you can detect when data-driven methods are being used nefariously to manipulate you!

Teaching Methodologies:

- *Lectures*: Class lectures will introduce a variety of data analysis techniques through **case studies**. Through in-class exercises, we will demonstrate and apply these techniques to real data. Lectures will provide a gentle introduction to the mathematics and statistics behind many techniques.

- *Readings*: The course textbook Foreman, *Data Smart: Using Data Science to Transform Information into Insight* is a reference book on different techniques. The course schedule lists sections (exercises) of the course text-book that you should read (attempt to complete) for each week. Required readings will include chapters from reference textbooks, research papers, news articles and blog posts. Some assigned readings are audibles. By keeping on top of the readings, you will make the best use of lecture time: you can clarify concepts you found difficult to understand and you can better participate in class discussions and exercises. Optional readings might include more advanced or related concepts, which you should refer to if you find a particular topic interesting or if it covers techniques you may need for your project.

- *Problem Sets & Project*: You will complete four problem sets in groups. There is little to *no programming* in this class. We use *spreadsheets*! We also use free, easy to use tools such as Tableau and Trifacta. Why? Because computational and data-driven thinking is not about the tools, machines or programming languages. It is about breaking down problems into simple logical processes. Anyone can engage in and benefit from data-driven thinking from a personal trainer to a Chief Technical Officer (CTO). Most problem sets require you to apply techniques already demonstrated in class to a different problem and data set. Through technical reports and visualizations you will improve your technical writing and learn how to effectively communicate analysis outcomes as well as the uncertainty associated with such outcomes.

Software Requirements: Install a recent Excel copy and free community versions of Tableau (for data visualization) and Trifacta (for preparing and cleaning data). Through the class, you will learn how to use these tools.

Class Deliverables:

Problem Set 1 (15%)	Personal Data Collection & Visualization
Problem Set 2 (15%)	Good & Evil Visualization
Problem Set 3 (15%)	Optimization & Prescriptive Data Analytics
Problem Set 4 (15%)	Prediction with Statistical Models?
Project Proposal, Report & Presentation (30%)	Pick a real-life problem and apply data-driven techniques to solve the problem.
Class Participation (10%)	Attend all lectures, read assigned material before class and participate in discussions that occur both in class and on the slack channel.

Grading Policy: In general, a 90% or above is within the A range, 80%-90% is within the B range and 70%-80% is within the C range. You have 100 hours of lateness forgiveness that you can use throughout the course for any problem set or project submission deadline.

Course Schedule:

Week	Lectures, Tutorials & Assignments
	Think
1	**Overview & Why a deeper analysis of data matters?**
	Assignment: Data on a personal level (part 1) - Daily journal of data exposure (1 week).
	See
2-4	**Visualization**
	Tableau Tutorial
	Assignment: Visualization Design: visualize the personal data (1 week)
	Assignment: Good & Evil visualizations (2 weeks)
	Project: Team Formation for Class Projects

	Describe
5-6	**Probability, Descriptive Statistics & Hypothesis testing**
	Basic statistics with Excel Tutorial
	Outliers & Liars
	Project: Project proposals (1-2 pages)
	Science & Data
	Process
7-9	**Data Wrangling**
	Trifacta Tutorial
	IBM NLP and Pattern Discovery Tutorial
	Project: Midterm milestone report (1-2 pages and mockup)
	Prescribe
10-11	**Optimization: Prescriptive Analytics**
	Assignment: Optimization - The best treat box that you can afford! (2 weeks)
	A/B Testing
	Predict
12-13	**Prediction: Regression and Other Techniques**
	Bayesian Techniques
	Biases and the ethics of AI decision-making
	Project: Milestone report 2 (1-2 pages and online draft)
	Assignment: Statistical prediction (2 weeks)
	Present
14	**Project Presentations**
	Project: Final project deliverable and report

Course Readings:

Course readings will be assigned on a weekly basis and communicated directly on the course book: `https://azzadev.github.io/databook`. Here is a sample of possible reading materials that may be assigned.

- Julia Angwin et al. **Machine Bias: Investigating the algorithms that control our lives**. May 2016. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

- Jeff Larson et al. **How We Analyzed the COMPAS Recidivism Algorithm**. May 2016. URL: `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

- Wikipedia. **Simpson's Paradox**. 2016. URL: `http://bit.ly/1OHFSOk`.

- Lars Backstrom and Jon Kleinberg. **Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook**. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing*. CSCW '14. Baltimore, Maryland, USA: ACM, 2014, pp. 831–841. URL: `http://doi.acm.org/10.1145/2531602.2531642`.

- Charles Duhigg. **How Companies Learn Your Secrets**. In: *The New York Times Magazine* (Feb. 2012). URL: `http://nyti.ms/18LN5uz`.

- Kyle Rush. **Optimization at the Obama campaign: a/b testing**. [Online, accessed 08-Nov-2016]. Dec. 2012. URL: `http://kylerush.net/blog/optimization-at-the-obama-campaign-ab-testing/`.

- Charles Duhigg. **What Does Your Credit-Card Company Know About You?** In: *The New York Times Magazine* (May 2009). URL: `http://nyti.ms/1Lwmum3`.

- Edward R. Tufte. **The Visual Display of Quantitative Information Paperback**. 2nd. Graphics Press, May 2001.

- Edward R. Tufte. **Visual and Statistical Thinking: Displays of Evidence for Making Decisions**. Graphics Press, Apr. 1997. URL: `http://www.cc.gatech.edu/~stasko/7450/Papers/shuttle.pdf`.

- William S. Cleveland and Robert McGill. **Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods**. In: *Journal of the American Statistical Association* 79.387 (1984), pp. 531–554. URL: `http://www.jstor.org/stable/2288400`.

- Hal R. Varian. **Letters to the Editor**. In: *The American Statistician* 26.3 (1972), pp. 62–66. URL: http://www.jstor.org/stable/2682871.

- RadioLab. **A Very Lucky Wind**. URL: http://www.radiolab.org/story/91686-a-very-lucky-wind/.

- Jiajie Zhang and Donald A. Norman. **The Representation of Numbers**. URL: http://bit.ly/2fPXAkU.

- Kate Crawford. **Artificial Intelligence's White Guy Problem**. [Online, accessed 08-Nov-2016]. June 2016. URL: http://nyti.ms/28VgTst.

- Lauren Kirchner Jeff Larson Surya Mattu and Julia Angwin. **How We Analyzed the COMPAS Recidivism Algorithm**. May 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

- Wikipedia. **Tay (bot)**. [Online; accessed 08-Nov-2016]. 2016. URL: https://en.wikipedia.org/wiki/Tay_(bot).

- WildML. **Deep Learning for ChatBots**. [Online, accessed 08-Nov-2016]. Apr. 2016. URL: http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/.

- Amit Datta, Michael Carl Tschantz, and Anupam Datta. **Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination**. In: *Proceedings on Privacy Enhancing Technologies* 2015 (1 Apr. 2015), pp. 92–112. URL: http://dx.doi.org/10.1515/popets-2015-0007.

- Matthew Kay, Cynthia Matuszek, and Sean A. Munson. **Unequal Representation and Gender Stereotypes in Image Search Results for Occupations**. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: ACM, 2015, pp. 3819–3828. URL: http://doi.acm.org/10.1145/2702123.2702520.

- Judea Pearl. **Comment: Understanding Simpson's Paradox**. In: *The American Statistician* 68.1 (2014), pp. 8–13. URL: http://dx.doi.org/10.1080/00031305.2014.876829.

- This American Life. **No Coincidence, No Story!** Mar. 2013. URL: https://www.thisamericanlife.org/radio-archives/episode/489/no-coincidence-no-story.

- Cynthia Dwork et al. **Fairness Through Awareness**. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: ACM, 2012, pp. 214–226. URL: http://doi.acm.org/10.1145/2090236.2090255.

- Edward Segel and Jeffrey Heer. **Narrative Visualization: Telling Stories with Data**. In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (Nov. 2010), pp. 1139–1148. URL: http://dx.doi.org/10.1109/TVCG.2010.179.

- Donald A. Norman. **Things That Make Us Smart: Defending Human Attributes In The Age Of The Machine**. Basic Books, Apr. 1994. URL: http://bit.ly/2gatpJc.

Academic Integrity: As set forth in NYU Abu Dhabi's Academic Integrity Policy, the relationship between students and faculty at NYU Abu Dhabi is defined by a shared commitment to academic excellence and is grounded in an expectation of fairness, honesty, and respect, which are essential to maintaining the integrity of the community. Every student who enrolls and everyone who accepts an appointment as a member of the faculty or staff at NYU Abu Dhabi agrees to abide by the expectation of academic honesty. The full policies and procedures relating to Academic Integrity may be found on the NYUAD Student Portal.

Wellness: As a University student, you may experience a range of issues that can interfere with your ability to perform academically or impact your daily functioning, such as: heightened stress; anxiety; difficulty concentrating; sleep disturbance; strained relationships; grief and loss; personal struggles. If you have any well-being or mental health concerns please visit the Counseling Center on the ground floor of the campus center from 9am-5pm Sunday - Thursday, or schedule an appointment to meet with a counselor by calling: 02-628-8100, or emailing: nyuad.healthcenter@nyu.edu. If you require mental health support outside of these hours call NYU's Wellness Exchange hotline at 02-628-5555, which is available 24 hours a day, 7 days a week. You can also utilize the Wellness Exchange mobile chat feature, details of which you can find on the student portal. If you need help connecting to these supports please contact me directly.