DS 201-Principles of Data Science

Syllabus Fall 2021

Instructor: Prof. Christian Lopez, 569 Rockwell Integrated Science Center, lopezbec@lafayette.edu
Class: MWF 3:10 PM- 4:00 PM, RISC 360
Office Hours: Tuesdays & Thursdays 2:00 PM -4:00 PM
Moodle Website: https://moodle.lafayette.edu/course/view.php?id=20760
Prerequisite: An introductory statistics course AND an introductory computing course

Course Description

This is a survey course that will introduce the principles of data science. Specifically, it will cover how to (i) collect and manage large sets of data, (ii) summarize and visualize data to convey information in a meaningful way, and (iii) implement basic machine learning models; all this while doing so in a thoughtful manner taking ethics and privacy into consideration.

We will be using example datasets from multiple domains and problems to study the effectiveness of different techniques and approaches. This course can be viewed as a fusion between a computing course focused on programming and algorithms, and a statistics course focused on estimation and inference. Hence, the introductory statistics and computing course prerequisites. While in this course we will be using multiple programming languages (e.g., Python3, R), this is not a programming class. Hence, students are not required to be "experts" in any language, instead, they are expected to have a basic understanding of programming constructs (e.g., functions, loops, variables), which will allow them to quickly learn new languages.

Student Learning Outcomes

Upon completion of this course, students will be able to:

- Extract, transform, and load datasets.
- Use SQL to create tables and insert, modify, retrieve, and delete data.
- Use a database composed of at least 2 tables.
- Perform Exploratory Data Analysis using multiple techniques and tools.
- Create effective visualizations.
- Understand the advantages and disadvantages of different visualization approaches.
- Apply machine learning methods and assess the quality of model output (predictions).
- Develop scripts to build program pipelines.
- Effectively communicate results.
- Work effectively and synergically in teams on data science projects.
- Discuss the ethical and privacy considerations in the decision to collect, store, use, and/or display a piece (or sub-pieces) of data.

Expectations

Every week, we will have a series of activities interwoven with lectures. In some instances, we will adopt a flipped-classroom approach where prior to the class, you will gain exposure to new concepts and material by watching lecture videos, doing guided readings, and/or completing the required activities; then you will do active work that calls for the analysis and/or the application of the concepts and material learned. Be advised, it is expected that students spend at least 2 hours of work outside of class for every hour of class (e.g., 2.5hrs. of class ~ 5 hrs. of work outside of class).

There will be discussions, dialogues, and exercises led by the faculty. You are expected to actively participate in these discussions during the class sessions, as well as the online forums, and work on the particular

exercises that allow you and your student colleagues to learn by doing, learn by observing the results of others, and to learn from one another while trying out new ideas. You will learn by analyzing datasets, interacting with your professor and other students; and engaging with your professor's instruction as well as external media.

Finally, there will be a Q&A forum/web service named Piazza (which you will be able to gain access to via Moodle). In Piazza, you will post all the questions regarding assignments, the class, and or class activity. When using this system (and any other tool), you need to follow the Students Code of Conduct. Hence, you should not provide the solution nor ask for the solution of any assignment (or parts of it). I know this is a bit complicated with the type of assignments/class we have but try to ask questions about the underlying principles of the assignment so you can better understand its fundamentals (I will let you know asap if a question or answer is not appropriate).

What you learn will depend directly on your willingness to participate, be involved, and complete assignments and exercises. Therefore, try different things even if you think they might "fail," ask questions—to faculty and to each other.

Textbooks

For this course will use the book:

- Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press. [https://ebookcentral-proquest-com.ezproxy.lafayette.edu/lib/lafayettecol-ebooks/detail.action?docID=5345177]
- Wilke, C. O. (2019). Fundamentals of data visualization: a primer on making informative and compelling figures. O'Reilly Media. [https://serialmentor.com/dataviz/]

There are free electronic versions of these books (e-book) available through the library, online, or by contacting the professor.

There are many good books about Data Science, Data Visualization, and Programming for Data Science in R and Python. We will cover material from different textbooks that are freely available online, such as:

- VanderPlas, J. (2016). *Python data science handbook: essential tools for working with data*. O'Reilly Media, Inc. [https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb#scrollTo=02mdomAyhkg4]
- Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".[https://r4ds.had.co.nz/]
- Janssens, J. (2014). *Data Science at the Command Line: Facing the Future with Time-tested Tools*. " O'Reilly Media, Inc.". [https://www.datascienceatthecommandline.com/index.html]
- Downey, A. B. (2011). *Think stats*. " O'Reilly Media, Inc.". [https://greenteapress.com/wp/think-stats-2e/]
- Loukides, M., Mason, H., & Patil, D. J. (2018). *Ethics and Data Science*. O'Reilly Media. [https://www.amazon.com/Ethics-Data-Science-Mike-Loukides-ebook/dp/B07GTC8ZN7]
- Ng, A. (2017). *Machine Learning yearning*, deeplearning.AI *[https://www.deeplearning.ai/machine-learning-yearning/]*
- Géron, A. (2019) 2nd Ed. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.(ISBN-10: 1491962291). [https://ebookcentral-proquest-com.ezproxy.lafayette.edu/lib/lafayettecol-ebooks/detail.action?docID=5892320] (available through Lafayette's Library)

Grading Scheme

The grading breakdown of this class is shown below.

Attendance	5%
Participation	5%
Self-Assessments	10%
Labs/Small Projects	35%
Homework	15%
Final Exam	10%
Final Project	20%

Grading Scale

Typically, grades are assigned as follows from your final numerical grade:

A: 93-100	B+: 87-89	C+: 77-79	D+: 67-69	F: 0-59
A-: 90-92	B: 83-86	C: 73-76	D: 63-66	
	B-: 80-82	C-: 70-72	D-: 60-62	

Assignments Types

Attendance: The success of this course is highly dependent on your involvement with the class material and classmates. As such, if no advance notice is given (i.e., 24hrs), you will be considered absent if you arrive at a class session more than 10 minutes late or not at all, and your attendance grade will be reduced proportionally. If you have a valid excuse that prevented you from providing advance notice, I will consider it on a case-by-case basis (again, **open and honest communication will take you a long way in this class**).

Participation: To get a perfect participation grade come to each class prepared and on time, contribute to class discussions during class or via the forums at least once a week, allow other students to contribute, and maintain a high level of respect and professionalism with your peers. Participation includes asking or answering questions, expressing an opinion about a topic of discussion, and involvement in group projects or group activities. All sincere efforts to participate are admired, so don't worry, just speak up and post. You are even welcome to express an opinion different than mine. All types of participation count except participation that shows you failed to prepare for class or are disrespectful to others. Finally, to continuously improve the course and ensure students are making the most out of it, there will be an anonymous feedback repository where students are encouraged to provide constructive feedback regarding the things they liked or would like to improve from the class, activities, lectures, etc.

***Self-Assessments*:** The material in this class is cumulative, so it is essential to stay current in the class. To help you stay current, there will be weekly short self-assessment quizzes posted on Moodle. Some of the quizzes will test your knowledge of the previous day or week. The intention here is to encourage you to actively review the material each day!

***Homework*:** As in any other learning endeavor, practice is important. The homework, as well as the Labs/Small Project, are meant to help you practice and gain proficiency. Homework includes completing the assigned problems. The answers will be provided after the deadline for submitting the homework has passed. You will receive the points for a problem that you complete in an acceptable way. If you do obtain outside assistance (e.g., friend, the internet, book, etc.) in completing the problem, be sure that you **do NOT copy** the answer. Rather, learn from outside assistance, and then do the problem on your own. Directly copying an answer would constitute cheating. Similarly, avoid asking/responding to specific questions about an assignment in Piazza (e.g., what is the code for completing X). Instead, ask/respond to questions that help you understand the underlying principles of the assignment.

***Labs/Small Projects*:** The labs are like small projects which will provide you an opportunity to show your knowledge and skills to solve a practical Data Science problem, in some cases while working with

teams. Lab assignments will be given every 2 to 3 weeks, and some of our class time might be dedicated to working on these assignments. We will discuss common issues or interesting observations from the labs. Notice that class time will often not be enough to complete the lab. This means that you should plan on working on them outside of class.

Final Exam: The final exam will be composed of an "in-class" portion and a take-home portion. Exam absences will receive a score of zero for the exam (unless a dean's excuse is given). Before the written portion of the exams begins, you are required to close your course materials and put them and your phone in front of the class to avoid the temptation to look at them during the exam. The final exam will be given via Moodle. It will be close a book, close notes, close phone, and close Moodle material exam. That is, you should only have Moodle's exam window open. The exam will consist of a series of questions and short-code questions. The final exam will cover all the content of the class. For the take-home portion, you will be given some requirements that you would need to satisfy. Both portions of the exam will have a time limit proportional to their complexity, for the take-home portion, you can use any material you want but be advised that trying to find a solution or part of it online might consume valuable time.

Final Project: The final project, will be a group project where students will turn in a tutorial (i.e., text, figures, code) that will walk users through the entire data science pipeline: data curation, parsing, and management; exploratory data analysis; hypothesis testing and machine learning to provide analysis; and then the curation of a message or messages covering insights learned during the tutorial. Students may choose an application area and dataset(s) that are of interest to them; please feel free to be creative about this! The tutorial should be self-contained, a mix of Markdown text and Python or R code, and delivered as a GitHub statically hosted Page.

Late Submission of Course Work Policy

If a student submits an assignment after the due date without having made arrangements with the instructor at least 24 hrs. before the deadline (i.e., you did not reach out a day before the assignment was due), you will get no points for that assignment (0 points). Hence, try to work on all your assignment at least a day before the deadline, so if something happens, you can reach out, or if you know you won't be able to complete on time, you can let me know 24hrs prior.

Class Participation Rubric

	Strong Work	Needs Development	Unsatisfactory
Listening	Actively and respectfully listens to peers and instructor	Sometimes displays lack of interest in comments of others	Projects lack of interest or disrespect for others
Preparation	Arrives fully prepared with all assignments completed, and notes on reading, observations, questions	Sometimes arrives unprepared or with only superficial preparation	Exhibits little evidence of having read or thought about the assigned material
Quality of Contributions	Comments are relevant and reflect an understanding of assigned text(s) or assignments; previous remarks of other	Comments sometimes irrelevant, betray lack of preparation, or indicate lack of attention to previous remarks of other students	Comments reflect little understanding of either the assignment or previous remarks in a seminar

	students; and insights about assigned materials		
Impact on Class	Comments frequently help move the class conversation forward	Comments sometimes advance the conversation, but sometimes do little to move it forward	Comments do not advance the conversation or are actively harmful to it
Frequency of Participation	Actively participates at appropriate times	Sometimes participates but, at other times, is "tuned out."	Seldom participates and is generally not engaged

Class participation deserving 100% participation grade will be strong in most categories; participation that is strong in some categories but needs development in others will receive an 80%; a grade of 60% reflects a need for development in most categories; 40% of work is typically unsatisfactory in several categories, and 0% of work is unsatisfactory in nearly all categories.

Technology

In this class, students will be introduced to several programs and applications to help their learning journey (e.g., SoloLearn, Kahoot). Technology in the classroom should enhance the learning environment for all students. The use of technology for purposes defined by the College as academic dishonesty is prohibited. In the event that students receive permission in advance to digitally record a class (audio or video), the material should not be posted to the internet for public access, unless a prior agreement has been made with me. The use of technology in my classes should reflect two key values:

- **That we are here for a common purpose – education**. The use of technology in the classroom by the faculty member and the students should always support student learning. If you are using your phone, tablet, or computer in class, be prepared to show me how you are using the technology to support your learning.

- **That the classroom should be a place of mutual respect**. Students need to respect my efforts to create a classroom environment and to organize the course in ways that support the learning of all students. Students also need to respect their fellow classmates and their classmates' rights not to be distracted from participating fully in the classroom.

Diversity and Inclusiveness

Lafayette College is committed to creating a diverse community: one that is inclusive and responsive and is supportive of each and all of its faculty, students, and staff. The College seeks to promote diversity in its many manifestations. These include but are not limited to race, ethnicity, socioeconomic status, gender, gender identity, sexual orientation, religion, disability, and place of origin.

The College recognizes that we live in an increasingly interconnected, globalized world, and that students benefit from learning in educational and social contexts in which there are participants from all manner of backgrounds. The goal is to encourage students to consider diverse experiences and perspectives throughout their lives. All members of the College community share a responsibility for creating, maintaining, and developing a learning environment in which **difference is valued, equity is sought, and inclusiveness is practiced.**

It is the mission of the College to advance diversity as defined above. The College will continue to assess its progress in a timely manner in order to ensure that its diversity initiatives are effective.

Learning Needs and Accessibility:

Lafayette College is committed to creating a learning environment that meets the needs of its diverse student body. If you anticipate or experience any barriers to learning in this course, you are welcome to discuss your concerns with me. If you have a disability or think you may have a disability, please meet with

the Office of Accessibility Services, to begin this conversation or request an official accommodation. If you have already been approved for accommodations through the Office of Accessibility Services, please meet with me so we can develop an implementation plan together.

Religious Observances:

If you plan to be absent from class due to the observance of a religious holiday, please communicate this to me before the end of the second week of class. You will need to get a Dean's Excuse for religious purposes. This dean's excuse needs to be approved before the "drop/add" deadline (see academic calendar).

Tentative Schedule

This is a tentative schedule, subject to change. Check Moodle for the most up to date information:

Week	Date		Class	Topics/Activities
1	Mon	30,Aug	1	> Syllabus & Intro to DS
	Wed	01,Sep	2	>What is Data Science?
	Fri	03,Sep	3	>What is Data Science? (cont)
2	Mon	06,Sep	4	>The DS Process- Understanding Problem
	Wed	08,Sep	5	> The DS Process-Understanding Data
	Fri	10,Sep	6	>The DS Process-Understanding Data (cont)
3	Mon	13,Sep	7	> Dealing with Data
	Wed	15,Sep	8	> Intro to SQL
	Fri	17,Sep	9	> Intro to the Command Line
4	Mon	20,Sep	10	>Intro to Linux and Servers
	Wed	22,Sep	11	>Shell Scripts and Pipelines
	Fri	24,Sep	12	>Intro to Data Visualization
5	Mon	27,Sep	13	> Data Visualization Principles
	Wed	29,Sep	14	> Data Visualization Principles (cont)
	Fri	01,Oct	15	> Assertion Evidence Approach
6	Mon	04,Oct	16	>Data Viz in Python
	Wed	06,Oct	17	>Data Viz in R
	Fri	08,Oct	18	> Statistics Recap
7	Mon	11,Oct	19	**FALL BREAK**
	Wed	13,Oct	20	> Intro to ML
	Fri	15,Oct	21	> Intro to ML (cont)
8	Mon	18,Oct	22	> Intro to Linear Regression
	Wed	20,Oct	23	>Intro to Multivariate LR
	Fri	22,Oct	24	> Overfitting and Underfitting
9	Mon	25,Oct	25	> Regularization
	Wed	27,Oct	26	> Intro to Logistic Regression
	Fri	29,Oct	27	>Intro to Neural Networks
10	Mon	01,Nov	28	> Neural Networks with Keras
	Wed	03,Nov	29	> Training and Testing ML Models
	Fri	05,Nov	30	> Training and Testing ML Models (cont)

11	Mon	08,Nov		> ML model performance metrics
	Wed	10,Nov		>Intro to Deep Learning
	Fri	12,Nov		> Unsupervised Machine Learning
12	Mon	15,Nov		> Unsupervised Machine Learning (Cont)
	Wed	17,Nov		>Ethics in Data Science
	Fri	19,Nov		>Ethics in Data Science (cont)
13	Mon	22,Nov		>Ethics in Data Science (cont)
	Wed	24,Nov		**THANKSGIVING BREAK**
	Fri	26,Nov		**THANKSGIVING BREAK**
14	Mon	29,Nov		Final Project Prep time
	Wed	01,Dec		
	Fri	03,Dec		
15	Mon	06,Dec		Final Exam Prep
	Wed	08,Dec		
	Fri	10,Dec		

- **Check E-Mail and Moodle Daily.** Information about the class, including assignment updates and schedule changes, will be posted to Moodle and/or sent by e-mail. Not reading your e-mail or checking Moodle will not be accepted as a reason for me to accept a late assignment or your absence from a class activity.

Communication

My preference is for you to address me as either Professor Lopez or Dr. Lopez. If you have a preference regarding how you would like to be addressed, please let me know.

If you need to schedule a meeting or have a request of me that will require time outside of class, please be sure to follow up any conversation we might have about the request immediately before, during, or after class with an e-mail to confirm that I have placed the request on my calendar. Because class time can be busy, by the time I return to my office, there is a chance I will have been distracted and forget our conversation.

Students often worry about how to e-mail a professor. I recommend reading some guidelines/advice at http://web.wellesley.edu/SocialComputing/Netiquette/netiquetteprofessor.html.

Academic Integrity

All students are expected to abide by the Student Code of Conduct including policies around academic integrity whether we are in a face-to-face or remote classroom environment. Please be sure to review the Student Code of Conduct through this link HERE.

At Lafayette College, all course materials are proprietary and for class purposes only. This includes posted recordings of lectures, worksheets, discussion prompts, and other course items. Such materials should not be reposted. Online discussions should also remain private and not be shared outside of the course. You must request my permission before creating your own recordings of class materials, and any recordings are not to be shared or posted online even when permission is granted to record. If you have any questions about the proper usage of course materials feel free to ask me. Also, if you have any concerns with being recorded during the course please let me know.

Federal Credit Hour Statement

The student work in this course is in full compliance with the federal definition of a four-credit hour course. Please see Registrar's Office website for the full policy and practice statement.

COVID-19, Masks, and Final Notes

We all can agree that these are difficult and challenging times for everyone, and unfortunately, even more for some people than for others. What we need to remember is that we are all humans and we are all in this together. As such, we need to strive to practice compassion and empathy as much as we can with everybody! While this syllabus sets forward the rules and policies for this class, given the current circumstances, I am very flexible with everything, the rule of thumb is to: reach out to me beforehand so we can work something out. Remember, I am here to help you succeed!

Wearing a mask is known to reduce the transmission of SARS-CoV-2, the virus responsible for COVID-19. Regardless of your vaccination status, to protect the health of our class and until further notice, the College policy is that masks must be worn during all indoor class sessions. Masks should be worn properly over the nose and mouth and secured on the chin. Food and drink must also be eaten outside of the classroom. Students who show up to class without a mask will be asked to return to class wearing one in order to protect the health of our classroom community. In the event that you do not have access to a mask to wear during the class session, please let me know and we'll make sure that you will be able to obtain one.

As humans, we ALL have the extraordinary ability to learn, understand, and deal with new and difficult situations. This is the definition of intelligence, and everybody can adapt to new circumstances and develop new skills with time and effort. This means that you CAN learn new things, you CAN really change your "basic intelligence", and your intelligence is something about you that you CAN change. Having a "growth mindset" and acknowledging that you can become better at anything if you put time and effort, and that failure is just part of the learning process, is fundamental for your success. In this class, I will reward perseverance and effort, which you can show with your participation, 1-1 meetings, MSG, online forums, etc.