Data Mining and Analysis

Summer 2018

Instructor: Katharine Jarmul
Email: kjarmul@jou.ufl.edu
Twitter: @kjam

About the Instructor: Katharine Jarmul is a data scientist and educator based in Berlin, Germany. Originally from Los Angeles, California, she first began working with Python for data analysis in 2008 at the Washington Post. Since then, she has worked at large and small companies working primarily on data extraction, cleaning and insights. She co-authored the O'Reilly book *Data Wrangling with Python* and has a M.A. in journalism and a M.S. in education.

Office Hours: Office hours will be held virtually via Slack (https://slack.com/is) as necessary. In order to schedule office hours, please email or message me privately -- give at least 48 hours notice if possible and include at least two available meeting times (for your schedule) and appropriate time zone details. I reside in Berlin, Germany (CET), so please allow up to 24 hours for proper receipt and response of your messages. There will be **2 required office hours per semester**, one which must be scheduled in the first two weeks of course and another project check-in within the final 4 weeks of the course. Scheduling them more often is encouraged, as this can be a great way to review concepts, ask questions and confirm understanding.

Course Website: http://elearning.ufl.edu

Course Communication: For most communication, the course will revolve around our Slack team, with a mixture of group channels for things like #help and more general chats hosted in channels like #general. If your question is more involved than a simple chat message (i.e. more than one paragraph of text), please use email instead (kjarmul@jou.ufl.edu). Be sure to include the course number and a relevant topic in the title of the email.

Course Description: Data Mining and Analysis provides a hands-on overview of data analysis and mining techniques for managing large datasets. Students will learn how to clean and analyze realistic datasets using tools like Python and databases. The course offers an introduction to data mining theory and statistics, while focusing on practical data mining applications and use cases, such as basic machine learning and statistical modeling.

Course Objectives:

By the end of this course, students will:

- Analyze summary statistics of datasets to describe qualities such as correlation, distribution and data quality.
- Describe and differentiate between different machine learning techniques such as classification, regression and unsupervised learning techniques.
- Utilize statistical findings to determine actionable consumer insights and client recommendations.
- Evaluate features within a dataset for information gain and entropy.
- Extract, Load and Transform data in tabular format using Pandas and NumPy (ETL/ELT).
- Create informative visualizations of dataset statistics found via data mining.
- Apply machine learning theory to several datasets using supervised and unsupervised learning methods and evaluate the effectiveness of the techniques used.
- Develop questions and create experiments on datasets guided by business questions and actionable outcomes.
- Critique data mining practices with regard to ethics, transparency and data privacy.

Course Goal:
Why is this course important? Data is plentiful and data collection practices at most companies mean it is a growing resource. But what should be done with the collected data? How can we analyze it to learn more about our

audiences and products? What inferences can we learn by investigating the statistics of our datasets? Finally, how can we do so while still respecting data science ethics and data privacy? This course covers those topics as we dive into both theory and practice of data mining and analysis.

Expectations:
Throughout this course, students will build skills for both data analysis and programming.

In order to cover all of these topics, students will need to apply themselves thoroughly to the coursework and bring a willingness to try new things and a curiosity for data. In this course, students will learn and apply self-guided learning techniques, such as how to debug without an expert by their side and using StackOverflow and group chats. Throughout the course, students will be relied upon to ask questions and help others who are stuck. These are great skills beyond the scope of programming and data mining which will help students succeed in most data analysis tasks they perform or advise in the workplace.

This course requires students to perform a pre-class assessment and have a laptop with an operating system which allows them to install applications and programs (i.e. Administrative access). If you are running Windows, you will need Windows Vista or later. If you are running OS X, you will need 10.8 or later (Mountain Lion). If you are running Linux, please insure you can install Python 3.

Ownership Education:

As graduate students, you are not passive participants in this course. This class allows you to not only take ownership of your educational experience but to also provide your expertise and knowledge in helping your fellow classmates. Your Slack team should be treated as a place where you can and should pose questions to your classmates when you have a question as it relates to an assignment or an issue that has come up at work. Your classmates along with your instructor will be able to respond to these questions and provide feedback and help. This open communication and accountability also allows everyone to gain the same knowledge in one location rather than the instructor responding back to just one student which limits the rest of the class from gaining this knowledge.

Required Text: Data Science for Business by Foster Provost and Tom Fawcett, O'Reilly Media, 2013

Required Installations: You will need to have Python and several other libraries installed on your computer. I will also provide a shared server for some exercises (code assignments); but it is highly recommended you set up your local computer to run all programs for testing, project work and your own use. If you have not used Python before, I recommend following the Python 3.6.X installation instructions here:

- MacOSX (https://www.python.org/downloads/mac-osx/)
- Windows (https://www.python.org/downloads/windows/)
- Other Platforms (https://www.python.org/download/other/)

If you have never used Python for data science before, I also ask that you install Anaconda (https://www.continuum.io/downloads) for managing packages and different Python versions.

To properly install Python 3.6+, here are some outlines for each operating system:

- Windows Vista or later
- Apple OS X 10.8 or later (Mountain Lion)
- Linux: Please ensure you can install via normal package manager (or builds)

If you run into trouble during any installation, feel free to email -- however, I encourage you to first try searching and solving your problem. Becoming more familiar with the inner workings of your computer and how to fix computer problems is a great first step in learning to program and a skill you will hone throughout this course.

Additional Readings:

Listed in the course schedule and in each weekly module on Canvas

Prerequisite knowledge and skills:
Students will have successfully completed the Introduction to Programming with Data course (using Python), which includes introduction to topics such as data types, data cleaning and analysis, documentation, testing, programming basics and SQL. Students will already have been exposed to tools and libraries such as Pandas, Jupyter notebook and matplotlib.

A review of basic statistics will be helpful in following along with the course, so a review of one or more of the following is a pre-assessment that should be completed before the first week of the course:

- Basic Statistics: https://www.khanacademy.org/math/statistics-probability/probability
- (Choose at least 5 sections from) An Overview of Statistics: https://www.khanacademy.org/math/statistics-probability
- Introduction to Statistics (from Facebook): https://classroom.udacity.com/courses/st101

If you prefer a book to review, please use Neil J. Salkind's Statistics for People who (Think They) Hate Statistics.

Teaching Philosophy:
In the field of data science, there is both pure theory and pure practice. In this course, we will approach both with vigor, but focus on practical applications of theory. Throughout this course, we will touch upon the deeper theories and academic approaches to data science and machine learning; however, the course will have a strong emphasis on practical use cases and projects. I believe this allows you to quickly apply and excel at data science to help you do your work, while still allowing for questions, growth and curiosity towards the academic field. This approach will be reflected in our "Weekly Readings" which will blend the research in the field with the daily applications.

Instructional Methods:
This course will involve several different instructional methods as a way to address different learning styles and approaches. If you find your learning style is not adequately addressed, please feel free to offer feedback via email at any time. The methods are as follows:

- Video lectures
- Required readings
- Discussion threads, posting and voting
- Asynchronous group discussions (via Slack)
- Coding projects (alone and in small assigned groups)
- Peer and self-assessments and code reviews

Course Policies:

Attendance Policy:

Due to its online nature, the course will not have in-person meetings or attendance in a classic sense. Students are required to:

- Set up two scheduled office hours throughout the course
- Check Slack for regular updates at least twice a week
- Check the course discussion board and participate in postings, voting and discussion threads at least once a week
- View and read all required content by the due date
- Send required projects in by their respective due dates
- Respond to group coordination and emails in 24 hours or less

If a student fails to meet the above attendance requirements, there will be a reduction in the participation portion of the student's grade. I encourage students to install necessary applications (such as Slack) on their mobile device or set up alerting to ensure you can promptly respond to fellow students and instructor messages without long delays.

Late Work and Make-up Policy:

Deadlines are critical to this class. All work is due on or before the due date. Pre-approved extensions for deadlines will only be permitted for emergencies. Minor inconveniences such as technical issues, family vacation or minor illness are not valid reasons for extensions. With this in mind there will be penalties for late work. NO LATE ASSIGNMENTS WILL BE ACCEPTED FOR FULL CREDIT without prior arrangements that are acceptable to the instructor, unless the lateness is due to an excused absence such as illness or catastrophic emergency that can be documented. This is true for all assignments, discussion boards, papers, case studies, etc. Late penalties are as follows:

Assignments <= one hour late:	15% penalty
Assignments > an hour late, but <= 12 hours late:	25% penalty
Assignments > 12 hours late, but <= 24 hours late:	50% penalty.
Assignments > 24 hours late, but <= 48 hours late:	70% penalty.
Assignments > 48 hours late: 0 points (no credit, or	100% penalty).

If you have an emergency or pre-approved schedule when you will be unavailable to complete assignments (such as offline or limited access to internet), you must let the instructor know at least 3 days in advance.

Requirements for class attendance and make-up exams, assignments, and other work in this course are consistent with university policies that can be found in the online catalogue at:
https://catalog.ufl.edu/ugrad/current/regulations/info/attendance.aspx

Emergency and extenuating circumstances policy: Students who face emergencies, such as a major personal medical issue, a death in the family, serious illness of a family member, or other situations beyond their control should notify their instructors immediately.
Students are also advised to contact the Dean of Students Office if they would like more information on the medical withdrawal or drop process: https://www.dso.ufl.edu/care/medical-withdrawal-process/ .
Students MUST inform their academic advisor before dropping a course, whether for medical or non-medical reasons. Your advisor will assist with notifying professors and go over options for how to proceed with their classes. Your academic advisor is Natalie Lee, and she may be reached at natalielee@jou.ufl.edu .

Coursework:
Most non-coding coursework will be submitted via Canvas. There are several other services we will use throughout the course for submissions, including:

● Group Project	Gitlab
● Code Assignments	Canvas & Jupyter Server

The weekly coursework deadlines and where to submit work will be posted to Canvas and updated as needed throughout the course.

The Jupyter Server requires a GitHub login (which you need to supply the first week of course). It is located at https://mmc6936.kjamistan.com/ (https is required).

Deadlines:
This class, like others, involves many deadlines. Here is a reminder. The new lecture starts on Mondays:

Group Tasks	11 AM EST Sundays the week assigned
Course Discussions	11 PM EST Thursdays the week assigned
Code Assignments	11 PM EST Sundays the week assigned
Final Coding Project	11 AM EST last Wednesday of the semester

Grading:
Your work will be evaluated according to this distribution. There will be opportunities in some of the assignments for extra credit. Those opportunities will only count for students with regular on-time completion of other assignments (i.e. >= 75% of work turned in on-time and complete).

● Reading Reactions / Chat Participation	20%
● Assignments	35%
● Group Contribution	10%
● Final Project	35%

The final grade will be awarded as follows:

A	100%	to	93%
A-	< 92%	to	90%
B+	< 90%	to	87%
B	< 87%	to	83%
B-	< 82%	to	80%
C+	< 80%	to	77%
C	< 77%	to	73%
C-	< 72%	to	70%
D+	< 70%	to	67%
D	< 67%	to	63%
D-	< 62%	to	60%

F < 59% to 0%

https://catalog.ufl.edu/ugrad/current/regulations/info/grades.aspx

I will round grades to the nearest half-percent, meaning a 92.50-92.99 will result in an A; whereas a 92.00-92.49 would remain an A-.

Weekly Lectures:
This course will have weekly video lectures shared via Canvas. These videos will be a mixture of content produced by the instructor, as well as other free online videos that explain and demonstrate the course content for the week. I ask that you watch all videos and complete all reading before attempting the assignment and group tasks -- even if the content is review for you. It's likely there are some tips and other pointers covered that you will be assessed on at a future point.

It is highly recommended you watch and read each module in order. I have arranged the topics so they build upon one another and reference previously covered content. This will help your learning progress intuitively. If you find yourself asking numerous unanswered questions, keep track of them. If they remain unanswered at the end of the module, please post them to the module discussion board or group chat. This helps me know you are watching and following along and allows me and your fellow students to engage and react by sharing our thoughts, answers and related questions.

Assignments:

Data Science Journal

Throughout this course, I ask that you keep a data science journal where you write down things that you learn and questions you have as you explore data mining, analysis and machine learning. Your format should follow that outlined in this blog post (https://linbug.github.io/self-improvement/personal%20tracking/imposter%20syndrome/2017/09/30/How-I-hacked-my-imposter-syndrome-using-personal-tracking/), where you document what the problem, question or confusion is. It is okay if you don't always have an answer for what you thought would happen, the exact category or what you learned, but I would like you to take time researching each issue (time capped at an hour or less).

Although you will not be explicitly turning these journal entries in, they can be used as extra credit (especially if you had difficulty with the weekly code assignment). You will also be using two of these entries at the end of the course in your final assignment -- so having many to choose from is ideal. In addition to their usefulness as points in this course, this process will be helpful for your growth as a data mining and analysis practitioner. It will show you what you have learned as you look back through earlier entries and give you a standard process for managing the inevitable “Oh no, I don’t know this” moments.

Course Discussion:

Each week we will cover a new set of materials with video content, code practice and examples and reading materials. Both I and the weekly moderator (see following *Weekly Moderator* section) will pose several questions about the assignments -- both questions that can be answered by the content and those that require you to form some opinions and thoughts related to the content. Your responses to at least 2 of these prompts are required.

I expect your response to be well-thought out and presented coherently. Your response to each prompt should be minimally one paragraph (4-7 sentences) and maximally 3 paragraphs. Your initial responses will be due by Thursday (11PM EST); but will be evaluated at the beginning of the following week -- giving you time to continue responding if another prompt or comment inspires a more thoughtful or interesting response. Your Communicative & Collaborative section of the rubric requires a response to your classmates -- please take time to review and respond to at least one classmate during the module before 11PM EST Sunday evening (of that same module).

Weekly (co-)Moderator: The weekly moderator assignment will be rotated so that all students have a chance to minimally co-moderate. The moderator will be evaluated on the questions and prompts chosen to share with the group; and the ability to keep the conversation focused and interesting -- encouraging students to keep writing. The moderator will also be in charge of finding one interesting related piece of content (video, article, tweet, GitHub repository, blog post) to share with the class. A list of suggested blogs for useful content and several newsletters are included in the Canvas resources for the course. The weekly moderation will be evaluated according to the moderation rubric and be part of the percentage of the reading reaction grade. In the event there are more than 12 students in class, there may be some modules that have more than one weekly co-Moderator.

Discussion Post Rubric (for more details on several of these rubric choices, I recommend reading the 6+1 Writing Traits Rubric included in the Canvas materials -- from which several of these points are adapted)

	Exceptional (90-100)	**Proficient (80-89)**	**Basic (70-79)**	**Poor (<70)**
Ideas	Main idea is clear; relevant details support and add value to the conversation; the topic is relevant to the discussion at hand and supports the main idea.	Main idea is clear; the supporting details are present, but some details don't add value; the topic is related, but doesn't add compelling information to the conversation.	Main idea is unclear or unsupported by details. The topic is related, but not relevant.	Main idea is missing as are supporting details. Topic is missing or irrelevant.
Organization	The writing is clear, focused and organized. Details are added in a logical order with attention to transitions between items.	The writing is clear, somewhat focused and shows some organization. Details are added in a supporting order, but	The writing is unclear or unfocused. There is basic organization (i.e. paragraphs) but little attention to order of	The writing is unclear, unfocused and lacks basic organization. Details may be missing or presented with no attention to order or

		transitions might be missing.	details or presentation of support.	transitions.
Word Choice	The writing uses insightful words, with technical words and concepts from the module used appropriately and when they add meaning.	The writing uses clear terms and technical words from the reading; several perhaps misused or added without meaning.	The writing does not include more than one technical word from the module or misuses nearly all when present.	The writing misuses all technical words included, or they are completely missing.
Conventions	The writing follows good conventions for conversational posts, with few if any grammatical or spelling errors.	The writing follows most writing conventions, and would require at least one round of edits and responses to be exceptional.	The reader is distracted by grammatical and spelling errors. Minimal 2 rounds of edits to be exceptional.	The writing is rife with grammatical and spelling errors.
Communicative & Collaborative	The writing promotes conversation and responses. The writer incorporates other responses and ideas, while still communicating their* own ideas.	The writing allows for conversation and responses, however it doesn't explicitly incorporate questions, mentions or others' ideas.	The writing includes one or fewer references to others' responses or references to others are not relevant. The writing is more statement and fact oriented and lacks attention to ongoing conversation.	The writing lacks any references towards others' ideas or responses.

*they is used to better generalize for non-cis pronouns

Moderator Rubric

	Exceptional (90-100)	**Proficient (80-89)**	**Basic (70-79)**	**Poor (<70)**
Leadership	The moderator demonstrates leadership qualities such as passion, open-mindedness, authenticity and inspiration in their* prompt and in follow up responses. They are patient and encouraging with responses.	The moderator demonstrates at least one quality of leadership in their* prompt. They encourage at least two responses.	The moderator writes a prompt, but it doesn't clearly demonstrate leadership qualities. They encourage at least one response.	There is no prompt, or the prompt lacks any qualities. The few (if any) responses lack encouragement or thoughtfulness.
Management	The moderator posts at least two prompts in a timely and organized fashion. The follow-up posts and responses show ability to encourage others' participation and a value for others' feedback.	The moderator posts at least one prompt before the deadline (but not earlier). The follow up posts are not encouraging or are too late (i.e. Sunday night before the due date) to promote conversation.	The moderator posts only one prompt. The follow up is either lacking encouragement or too late.	The moderator fails to either post prompts or responses in a timely fashion.

Assignments

Assignments will be submitted via the Jupyter Server or Canvas. The specifics (including data, problem set or challenge) of each assignment will be available on the Monday of the module week. The due date for each assignment will be Sunday by 11PM EST. Many assignments will have some element of code included as well as some writing (usually documentation, problem solving reflection and / or debugging notes); however several assignments are focused solely on theory, research and business practices and will require you to perform some outside research and turn in a written report.

I encourage you to evaluate each piece submitted with the following rubric *before* actually submitting the work. Consider this rubric a checklist for items that each assignment will require. Information about PEP-8 and how to easily lint (or clean up) your code for issues are available in the Canvas resources. Note, that several pieces of the rubric (Documentation, Testing) are not always required for the assignment (you can see the specifics in the Rubric assigned to the assignment in Canvas).

The written report assignments which don’t have a code section will be graded based on the discussion rubric.

Code Assignment Rubric

	Exceptional (90-100)	**Proficient (80-89)**	**Basic (70-79)**	**Poor (<70)**
Clear, legible code, logic	The code follows PEP-8 standards and is legible. Variable and function names are clear and meaningful. The logic, data types and layout are	The code follows most PEP-8 standards and is somewhat legible. Variable and function names or logic (regarding data types or	The code follows some PEP-8 standards. Variables and functions are often named unclearly. Some logic is applied to data types, but at	The code shows no regard for PEP-8, legibility or normal logic flows.

	easy-to-follow.	organization) are sometimes unclear.	times it might be unclear.	
Code or Project Structure	The code (or repository) has been organized in a manner to allow readability and easy extension. If it is a repository, the folders and files follow common conventions and the structure itself is well-documented.	The code (or repository) is readable and at least somewhat extensible (ability to import and use). If a repository, most folders and files follow common conventions and the structure is documented.	The code is readable but would take work to improve for extensibility. If a repository, the folders and files follow some conventions, but not enough to easily share with others. The structure may or may not be documented.	The code is disorganized and not very legible. It might function, but it is not easily shared, readable or extensible. If a repository, little organization exists and common conventions are not followed.
Correct Functionality / End Result	The code offers exceptional or above-and-beyond comprehension and functionality for the end result. This can be in the form of a “one step further” approach or applying new data to ensure general usability.	The code functions properly and returns the proper result.	The code has at least one error in the result, but also at least one proper step or intermediate result is achieved.	The code is missing or achieves improper or false results.
Library Knowledge	The code shows a clear	The code uses outside	The code utilizes at least	The code uses no outside

	understanding of when and how to apply outside or standard libraries. The author has taken time to learn and apply outside libraries for the problem at hand.	libraries demonstrated in the module appropriately. Only one or two sections of the code could benefit from more library utilization.	one outside library, but is not effective or efficient in its use. There are more than two sections of code which could benefit from more library utilization.	libraries or applies them to irrelevant or improper uses.
Mathematical / Statistical Reasoning	The code demonstrates a strong grasp of mathematical and statistical logic. The data types and output chosen reflect appropriate mathematical reasoning and can be strongly supported by work in the field.	The code demonstrates a basic grasp of mathematical and statistical logic. Most data types and chosen output reflect mathematical reasoning.	The code shows little grasp of mathematical or statistical logic. The data types chosen show a rudimentary understanding of the mathematical concepts.	The code has few (if any) elements showing any mathematical or statistical reasoning. Data types are misused or rarely used.
Documentation	The documentation is clear, concise and covers relevant topics. There is both module level documentation and class / function level. If necessary, the inline	The documentation is clear, concise and covers most relevant topics. There is complete module level documentation although not all classes and functions are	The documentation is unclear or too short or long for manageable reading. Several modules, functions or classes lack documentation.	The documentation is incomplete, missing or illegible for at least one if not all module, class, function and inline code sections.

	documentation clarifies the logic choices.	documented.		
Debugging Notes	The author has included appropriate debugging notes in the ReadMe either in the form of a narrative (how I solved X?) or an FAQ. These notes are ready for public consumption and use.	The author has included some debugging notes in a file. They have at least one outside reference but are difficult to follow.	The author has included only one note or reference re: debugging and it is not included in a separate document or resource.	There are no debugging notes or they are indistinguishable from other notes.

Final Project

The final project will be a group project with group work (in the form of peer review, code review, issue assignment and delegation) being a large part of the tangible grade. Of course, a functional product is also a requirement and non-functional code will be returned for further work. There will be several stages to this project and reviews along the way, to ensure the groups are working well and progress is steady and obvious (I am hoping this also discourages procrastination).

Final projects will be presented and reviewed in the final week of the course and will uploaded via GitLab. The code and notebooks as well as any additional resources shared in the GitLab project will be reviewed using the Code Assignment review rubric. The overall project must meet the defined requirements (posted on Canvas), and the group work will be reviewed using the following “Data Science Team” rubric. For the rubric, team members will be evaluated based on the overall team performance (see “As A Team” criteria) as well as individually (see “As a Member” criteria). These evaluations will be based on review of the GitLab project as well as peer and individual reviews.

Teams will be assigned by the instructor after a survey. If you have concerns about your team, a particular team member or your ability to contribute to the team, please contact the instructor immediately. You will be evaluated based on your ability to work together as well as the individual strengths you bring to the table; so I encourage you to attempt to resolve team issues internally before bringing them to the instructor (i.e. as you would for a true Data Science team if the instructor was instead the CEO/CTO/CIO).

Data Science Team Rubric

	Exceptional (90-100)	**Proficient (80-89)**	**Basic (70-79)**	**Poor (<70)**
As a Team: Collaborative & Communicative	The team communication is collaborative and timely. Issues that are brought to light are quickly addressed. Peer feedback is constructive and positive.	The team communication is collaborative or timely. Issues are eventually addressed. Peer feedback is lacking constructive or positive elements.	The team communication lacks timeliness or collaboration. Peer feedback is missing or overwhelmingly negative.	Team communication is infrequent, incomplete and lacking collaboration.
As a Team: Use of Tools	The team utilizes all tools available within GitLab in appropriate capacity; including Issues, planning, and documentation.	The team uses most of the tools available in an appropriate capacity.	The team uses at least one tool in an appropriate capacity.	The team does not utilize the tools available or uses them in unclear, inappropriate ways.
As a Team: Accepts Challenges &	When faced with inevitable challenges, the team is	When facing challenges, the team is responsive; but	When facing challenges, the team is unresponsive	The team has little to no ability to delegate or

Able to Delegate	adaptive, uses issues to report and inform other members and appropriately and clearly delegates work and review.	may not always utilize issues or communication to handle problems or delegate issues.	or unclear. One or two members end up carrying the group to the finish.	respond to challenges or does so in an incomplete or untimely fashion.
As a Member: Responsible	The team member takes on tasks and issues and helps delegate what they cannot do well. They ensure the work assigned to them is completed in a timely manner.	The team member responds to delegated or assigned tasks (and might at times volunteer). They ensure their work is completed in time.	The team member responds to some, but not all, delegated or assigned tasks. The work is completed mostly on time.	The team member shows little to no responsibility towards their team or tasks.
As a Member: Responsive	The team member is both timely and open-minded when responding to others’ mentions, comments and feedback.	The team member is often timely and open-minded when responding to others; with some visible deviations.	The team member is at least twice not on time or responsive to others feedback, comments or mentions.	The team member is often unresponsive to other members of the team.
As a Member: Productive	The team member commits meaningful and contributive code and comments frequently.	The team member commits and comments at least once a week. Some contributions or comments lack thought or attention to	The team member commits and comments at least every other week. The contributions are mostly helpful.	The team member does not contribute as a productive member via code and / or comments.

		detail.		

Final Presentation + Findings Rubric

	Exceptional (90-100)	**Proficient (80-89)**	**Basic (70-79)**	**Poor (<70)**
Presentation: Organized	The presentation has a clear beginning, middle and end with attention paid to building on previous concepts and findings.	The presentation builds on previous concepts, but has an unclear beginning or end.	The presentation either does not build on concepts or findings, or does so in a disjointed manner.	The presentation pays little attention to organization with little or no building on previous findings and concepts.
Presentation: Audience -Appropriate	The presentation content and language used shows attention and care given the audience. Time is spent on topics engaging to decision makers and peers.	The presentation content and language is appropriate for the given audience. Topics are related to decision makers and peers.	The presentation uses some appropriate language and content, but also includes material too easy or too difficult for the target audience.	The presentation content and language do not fit the audience. The topics are either not present or do not include topics for decision makers or peers.
Presentation: Data Visuals	Data visualization is featured prominently in the presentation and the quality of the visuals	Data visualization is included in the presentation and the visuals add to the content. The visualizations	Data visualization is included in the presentation, but at least one visual is challenging to understand or	Either data visualization is not included in the presentation, or a majority of the visuals are ineffective,

	improves the content. The visualizations are clear and effective at communicating the findings. The visuals are included in an engaging way and the audience wants to further explore the data based on these visuals.	are effective at communicating the findings. The visuals are included appropriately, but perhaps not in an engaging way.	ineffective at communicating the findings.	inaccurate or unethical.
Findings: Clear and well-written*	The presentation deck and supplemental final findings materials are well-written and follow convention. The materials make the findings clear and easy-to-understand. Technical and mathematical words are used appropriately and add meaning.	The presentation deck and supplemental final findings materials are written clearly and follow most conventions. There are a few unclear places or words used improperly, but the overall usage and clarity is not hindered by these mistakes.	The presentation deck and supplemental final findings are lacking clarity or are poorly written. Technical and mathematical words are sometimes misused.	The presentation deck and supplemental final findings materials are incomplete or are unclear or error-prone.
Findings: Statistically Accurate & Reproducible	The findings and visualizations included in the material and	The findings and visualizations included in the material and	The findings and visualizations included in the material and	The statistics and visuals are either missing or are have major

	slide deck are accurate and easily reproduced. They show an understanding of the statistical principles and are well chosen to convey the outcome.	slide deck are accurate and able to be reproduced.	slide deck have minor inaccuracies or are not able to be reproduced.	inaccuracies.
Findings: Actionable	The findings have a clear actionable result, even if this result is in the form of a new design plan. The importance and relevance of the findings and the topic as a whole is made clear to the audience.	The findings have a clear and actionable result, but perhaps lacking a few areas of follow through or plan. The importance of the findings is referenced, but perhaps not made clear to the audience.	The findings have either an unclear result or little engagement in “what happens next.” The importance of the findings is either missing or unclear.	There is little or no attention paid to what to do with the findings. The importance of the findings is lacking clarity or is entirely missing.

* What does well-written mean? See Discussion Posts Rubric: Organization, Word Choice and Convention

Any clarification needed on the rubrics should be done before the end of the first week of class. If you have questions, suggestions or need help, please message the instructor or post in group chat to clarify the problem or question immediately.

University Policies

University Policy on Accommodating Students with Disabilities:

Students requesting accommodation for disabilities must first register with the Dean of Students Office (http://www.dso.ufl.edu/drc/). The Dean of Students Office will provide documentation to the student who must then provide this documentation to the instructor when requesting accommodation. You must submit this documentation prior to submitting assignments or taking the quizzes or exams. Accommodations are not retroactive, therefore, students should contact the office as soon as possible in the term for which they are seeking accommodations.
Students with Disabilities who may need accommodations in this class are encouraged to notify the instructor and contact the Disability Resource Center (DRC) so that reasonable accommodations may be implemented. DRC is located in room 001 in Reid Hall or you can contact them by phone at 352-392-8565.
University counseling services and mental health services:

**Netiquette: Communication Courtesy:
All members of the class are expected to follow rules of common courtesy in all email messages, threaded discussions and chats.
http://teach.ufl.edu/wp-content/uploads/2012/08/NetiquetteGuideforOnlineCourses.pdf

Class Demeanor:
Mastery in this class requires preparation, passion, and professionalism. Students are expected, within the requirements allowed by university policy, to attend class, be on time, and meet all deadlines. Work assigned in advance of class should be completed as directed. Full participation in online and live discussions, group projects, and small group activities is expected.

My role as instructor is to identify critical issues related to the course, direct you to and teach relevant information, assign appropriate learning activities, create opportunities for assessing your performance, and communicate the outcomes of such assessments in a timely, informative, and professional way. Feedback is essential for you to have confidence that you have mastered the material and for me to determine that you are meeting all course requirements.

At all times it is expected that you will welcome and respond professionally to assessment feedback, that you will treat your fellow students and me with respect, and that you will contribute to the success of the class as best as you can.

Getting Help:
For issues with technical difficulties for E-learning in Canvas, please contact the UF Help Desk at:

- Learning-support@ufl.edu
- (352) 392-HELP - select option 2
- https://lss.at.ufl.edu/help.shtml

** Any requests for make-ups due to technical issues MUST be accompanied by the ticket number received from LSS when the problem was reported to them. The ticket number will document the time and date of the problem. You MUST e-mail your instructor within 24 hours of the technical difficulty if you wish to request a make-up.

Other resources are available at http://www.distance.ufl.edu/getting-help for:

- Counseling and Wellness resources
  http://www.counseling.ufl.edu/cwc/Default.aspx
  352-392-1575
- Disability resources
- Resources for handling student concerns and complaints
- Library Help Desk support

Should you have any complaints with your experience in this course please visit http://www.distance.ufl.edu/student-complaints to submit a complaint.

Course Evaluation:
Students are expected to provide feedback on the quality of instruction in this course based on 10 criteria. These evaluations are conducted online at https://evaluations.ufl.edu
Evaluations are typically open during the last two or three weeks of the semester, but students will be given specific times when they are open. Summary results of these assessments are available to students at https://evaluations.ufl.edu/results

University Policy on Academic Misconduct:

Academic honesty and integrity are fundamental values of the University community. Students should be sure that they understand the UF Student Honor Code at http://www.dso.ufl.edu/students.php

The University of Florida Honor Code was voted on and passed by the Student Body in the Fall 1995 semester. The Honor Code reads as follows:

Preamble: In adopting this Honor Code, the students of the University of Florida recognize that academic honesty and integrity are fundamental values of the University community. Students who enroll at the University commit to holding themselves and their peers to the high standard of honor required by the Honor Code. Any individual who becomes aware of a violation of the Honor Code is bound by honor to take corrective action. A student-run Honor Court and faculty support are crucial to the success of the Honor Code. The quality of a University of Florida education is dependent upon the community acceptance and enforcement of the Honor Code.

The Honor Code: “We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honesty and integrity.”

On all work submitted for credit by students at the University of Florida, the following pledge is either required or implied:

"On my honor, I have neither given nor received unauthorized aid in doing this assignment."

For more information about academic honesty, contact Student Judicial Affairs, P202 Peabody Hall, 352-392-1261.

ACADEMIC HONESTY

All graduate students in the College of Journalism and Communications are expected to conduct themselves with the highest degree of integrity. It is the students’ responsibility to ensure that they know and understand the requirements of every assignment. At a minimum, this includes avoiding the following:

Plagiarism: Plagiarism occurs when an individual presents the ideas or expressions of another as his or her own. Students must always credit others' ideas with accurate citations and must use quotation marks and citations when presenting the words of others. A thorough understanding of plagiarism is a precondition for admittance to graduate studies in the college.

Cheating: Cheating occurs when a student circumvents or ignores the rules that govern an academic assignment such as an exam or class paper. It can include using notes, in physical or electronic form, in an exam, submitting the work of another as one's own, or reusing a paper a student has composed for one class in another class. If a student is not sure about the rules that govern an assignment, it is the student's responsibility to ask for clarification from his instructor.

Misrepresenting Research Data: The integrity of data in mass communication research is a paramount issue for advancing knowledge and the credibility of our professions. For this reason any intentional misrepresentation of data, or misrepresentation of the conditions or circumstances of data collection, is considered a violation of academic integrity. Misrepresenting data is a clear violation of the rules and requirements of academic integrity and honesty.

Any violation of the above stated conditions is grounds for immediate dismissal from the program and will result in revocation of the degree if the degree previously has been awarded.

Students are expected to adhere to the University of Florida Code of Conduct https://www.dso.ufl.edu/sccr/process/student-conduct-honor-code

Although it should not need to be said, I will state that any student failing to abide by the Code of Conduct towards any other student or faculty member will be immediately dismissed from the course communications and placed in mediation.

If you have additional questions, please refer to the Online Graduate Program Student Handbook you received when you were admitted into the

Program.

Schedule

Course Introduction:

Course Introduction Video:

- Welcome to Data Science & Data Mining! https://mediasite.video.ufl.edu/Mediasite/Play/3a3590a70cea438689d249eddc48b18c1d
- Introduction Discussion: What is Data Mining?

Course Syllabus Video:

- Course Syllabus Video https://mediasite.video.ufl.edu/Mediasite/Play/463e10a3f53b4472b6099adc748147c11d

Statistics Orientation Videos:

Please complete one of the following introduction to statistics video courses that are free online. This will help you get a jumpstart on the course and allow some things to be easier repetition and review rather than immediately diving in with no background.

- Basic Statistics: https://www.khanacademy.org/math/statistics-probability/probability
- (Choose at least 5 sections from) An Overview of Statistics: https://www.khanacademy.org/math/statistics-probability
- Introduction to Statistics (from Facebook): https://classroom.udacity.com/courses/st101

If you prefer a book to review, please use Neil J. Salkind's Statistics for People who (Think They) Hate Statistics.

Tasks:

- Course Introduction Discussion
- Send instructor your GitHub username
- Install Slack and join the team via email invite
- Clone or Download the Course Repository:

https://github.com/kjam/uf-data-mining-and-analysis
- Please fill out the initial skills survey https://goo.gl/forms/WcbUTL0OwjCSgjwd2

Week One: Introduction to Data Science Modeling

Learning Objectives:
- Students will select possible use cases for machine learning.
- Students will outline goals of machine learning and data science in a business setting.
- Student will review Python with data programming techniques.

Watch:
- Review of concepts from Introduction to Programming with Data https://mediasite.video.ufl.edu/Mediasite/Play/723e9400d3f6464b8fc94677fc5cf2d01d
- What is a Data Scientist? https://www.youtube.com/watch?v=uy1_hccQDSI&list=PLAwxTw4SYaPk41og7PER4HBpGciPw6n3x&index=2 and https://www.youtube.com/watch?v=9PIqjaXJo7M&list=PLAwxTw4SYaPk41og7PER4HBpGciPw6n3x&index=4
- What does a Data Scientist do? https://www.youtube.com/watch?v=vowXaEDh1uk&list=PLAwxTw4SYaPk41og7PER4HBpGciPw6n3x&index=5 and https://www.youtube.com/watch?v=hDxjx7SMPPk&index=13&list=PLAwxTw4SYaPk41og7PER4HBpGciPw6n3x
- Google Cloud AI: The 7 Steps of Machine Learning: https://www.youtube.com/watch?v=nKW8Ndu7Mjw (related article: https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e)

Required Readings:

- Data Science for Business, Chapter 1
- Everyday Examples of ML / AI: https://www.techemergence.com/everyday-examples-of-ai/
    - Please read at least ONE of the cited examples (via links) in this article.

- Quora: Is Data Science Really a Rising Career?
  https://www.quora.com/Is-data-science-really-a-rising-career

- Resources:
  - How to Fork a Repo:
    https://help.github.com/articles/fork-a-repo/

Assignments:

- Complete module one checklist: installing Python, Jupyter and IPython packages, signing up for GitHub and StackOverflow and generating an SSH key and putting it in your GitHub profile (https://help.github.com/articles/connecting-to-github-with-ssh/)
  - NOTE: you should likely already have this done from the prior course
- Module one discussion posts
- Code Assignment: Programming with Python Review

Week Two: Aggregate Statistics for Data Analysis

Learning Objectives:

- Students will apply data mining techniques to a large dataset.
- Students will analyze and present summary statistics on a dataset.
- Students will determine use cases for data mining strategies.

Watch:

- Refresher (Introduction to Programming with Data course): Statistics with NumPy
  https://mediasite.video.ufl.edu/Mediasite/Play/e33b47805fb84c4c9c62f724a8b420791d
- Pandas statistical data mining
  https://mediasite.video.ufl.edu/Mediasite/Play/693966e777f54089b3664f0b54c9cad11d
- Pandas Group & Transform
  https://mediasite.video.ufl.edu/Mediasite/Play/257986831bea4b959da1cbcc2fe958331d
- What is Data Mining?
  https://www.youtube.com/watch?v=R-sGvh6tI04

Required Readings:

- Mining Massive Datasets (Chp 1): http://infolab.stanford.edu/~ullman/mmds/ch1.pdf
- Python Data Science Handbook
    - Grouping: https://jakevdp.github.io/PythonDataScienceHandbook/03.08-aggregation-and-grouping.html
    - Pivot Tables: https://jakevdp.github.io/PythonDataScienceHandbook/03.09-pivot-tables.html
- Try out some functions from Panda's Apply operations https://chrisalbon.com/python/data_wrangling/pandas_apply_operations_to_groups/

Assignments:

- Group questionnaire & Reflection: https://goo.gl/forms/KSSOQOARwMlceQTt2
- Code Assignment: Pandas Grouping and Filtering
- Module two discussion

Week Three: Data Mining with Pandas and SQL

Learning Objectives:

- Students will apply map-reduce technique using Pandas and SQL.
- Students will select summary statistics when given a business use case and dataset.
- Students will structure an unstructured dataset for future analysis when given a business motivation.

Watch:

- Inserting Data with SQL: https://www.khanacademy.org/computing/computer-programming/sql/sql-basics/p/creating-a-table-and-inserting-data
- Aggregating Data with SQL: https://www.khanacademy.org/computing/computer-programming/sql/sql-basics/p/aggregating-data

- Data Mining in SQL
  https://mediasite.video.ufl.edu/Mediasite/Play/aa48281afe8940c2a90987d841c90ffc1d
- MapReduce in Data Mining and Pandas
  https://mediasite.video.ufl.edu/Mediasite/Play/aca66b4006a441fda70c98951d13786d1d

Required Readings:

- Data Science for Business, Chp 2
- Introduction to MapReduce:
  http://infolab.stanford.edu/~ullman/mmds/ch2.pdf (If anything after 2.3 is a bit confusing, feel free to skim / skip).
- SQL Aggregation:
  https://swcarpentry.github.io/sql-novice-survey/06-agg/

Assignments:

- Team Communication Guidelines and Roles
- Team GitLab Setup
- Code Assignment: Data Mining with SQL and Pandas
- Module three discussion

Week Four: Predictive Analytics

Learning Objectives:

- Students will determine information gain and entropy of a given dataset.
- Students will categorize dependent and independent variables when given a dataset and use case.
- Students will find significant features which allow for information gain given a particular business problem and dataset.

Watch:

- Information Gain with Pandas
  https://mediasite.video.ufl.edu/Mediasite/Play/f671e7bcbd714b968b1b1da39082da941d
- Finding Correlation and Significant Variables / Features with Pandas
  https://mediasite.video.ufl.edu/Mediasite/Play/c15279508d7248b798d8af107c70fb251d

- Introduction to Decision Trees, Entropy and Information Gain https://www.youtube.com/watch?v=iZYv1WdWwQo&index=114&list=PLAwxTw4SYaPkQXg8TkVdIvYv4HfLG7SiH (#114 to #166)

Required Readings:

- Data Science for Business, Chp 3
- Investigate our dataset using https://pair-code.github.io/facets/
- https://www.oreilly.com/ideas/identifying-viral-bots-and-cyborgs-in-social-media
- http://www.unofficialgoogledatascience.com/2016/10/practical-advice-for-analysis-of-large.html

Assignments:

- Group Assignment: Initial Final Project Research & Possible Questions
- Code Assignment: Information Gain and Correlation with Pandas
- Module four discussion

Week Five: Introduction to Machine Learning

Learning Objectives:

- Students will differentiate between supervised and unsupervised machine learning models and use cases.
- Students will determine possible applications of machine learning to real world problems.
- Students will critique application of a variety of models to a given dataset, defending their choice for best model.

Watch:

- Overview of ML Methods and Models https://mediasite.video.ufl.edu/Mediasite/Play/445ac07aac0a4162b7152db203edd9411d
- Playing with Tensorflow (playground.tensorflow.org) https://mediasite.video.ufl.edu/Mediasite/Play/925516700b714ad4aec662c0a8a6a1511d
- Introduction to ML (MIT): https://www.youtube.com/watch?v=h0e2HAPTGF4

Required Readings:

- Data Science for Business, Chp 4-5
- Visual Introduction to Machine Learning: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- Algorithms Tour @ StitchFix: http://algorithms-tour.stitchfix.com/

Assignments:

- Group Assignment: Business Needs & Impact Analysis
- Code Assignment: Introduction to Machine Learning Theory and Application
- Module five discussion

Week Six: KNN and Clustering

Learning Objectives:

- Students will apply K-Nearest-Neighbors on a dataset to determine useful clusters.
- Students will distinguish groupings in a dataset using unsupervised clustering methods.
- Students will test different values of K when using KNN and determine best fit with the dataset.

Watch:

- Clustering & KNN (Harvard): https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=c322c0d5-9cf9-4deb-b59f-d6741064ba8a (From CS109: http://cs109.github.io/2015/pages/videos.html)
- Clustering Documents with Scikit-Learn https://mediasite.video.ufl.edu/Mediasite/Play/89ec41b9171340349a4f9db7a49d32161d

Required Readings:

- Data Science for Business, Chapter 6
- Clustering (Scientific Marketer): http://scientificmarketer.com/2007/02/clustering-considered-harmful-i-outline.html

- K-Means Clustering with Scikit-Learn: https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html

Assignments:

- Group Assignment: Data Science Project Proposal
- Group Assignment: Creating and Assigning Tasks and Issues in GitLab & Kanban usage
- Code Assignment: KNN with Python
- Module six discussion

Week Seven: Linear Models

Learning Objectives:

- Students will apply linear models such as linear regression to a dataset.
- Students will select features from a dataset for feature engineering.
- Students will judge what features increase accuracy of their model.

Watch:

- Least Squares Linear Regression https://www.youtube.com/watch?v=yMgFHbjbAW8
- Forecasting with Linear Regression: https://www.youtube.com/watch?v=ZaxpCw6lCe4
- Feature Engineering on Lobste.rs Dataset https://mediasite.video.ufl.edu/Mediasite/Play/9aa098a0234146c8a97a6dfb315a11711d
- Linear Regression on Story Score https://mediasite.video.ufl.edu/Mediasite/Play/41f574509a444482955280d4bc41f6bf1d

Required Readings:

- Data Science for Business, Chapter 7-8
- Feature Engineering: https://jakevdp.github.io/PythonDataScienceHandbook/05.04-feature-engineering.html
- Comparing Different Scalars

http://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

- Extra: Linear Regression (in depth) https://jakevdp.github.io/PythonDataScienceHandbook/05.06-linear-regression.html

Assignments:

- Group Assignment: Progress on research goals and business impact via GitLab
- Code Assignment: Linear Models
- Module seven discussion

Week Eight: Introduction to Bayesian Statistics

Learning Objectives:

- Students will summarize Bayes law in relation to a given use case.
- Students will apply a Bayesian model to a dataset and evaluate its effectiveness.
- Students will judge whether a given use case is a good fit for Bayesian modeling.

Watch:

- Workshop on Bayesian methods in Python: https://www.youtube.com/watch?v=TpgiFIGXcT4
    - To follow along, the notebooks are available online here: https://hub.mybinder.org/user/allendowney-bayesmadesimple-icon4kpf/tree or you can install them yourself from the repository here: https://github.com/AllenDowney/BayesMadeSimple
    - NOTE: it is okay to skip coding alongside and just watch the answers, but please try and type them as he does and run them in the binder or on your own notebook.
- A visual guide to Bayesian Thinking: https://www.youtube.com/watch?v=BrK7X_XlGB8
- Building a Bayesian text classifier https://mediasite.video.ufl.edu/Mediasite/Play/dc7525b12e4a485fb53030fc18cf6aaf1d

Required Readings:

- Data Science for Business, Chapter 9
- Introduction to Bayesian Inference: http://nbviewer.jupyter.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Ch1_Introduction_PyMC3.ipynb

Assignments:

- Group Assignment: Continued progress on project via GitLab
- Data Science Assignment: Probability-Based Modeling
- Module eight discussion

Week Nine: Machine Learning with NLP

Learning Objectives:

- Students will apply NLP statistical models such as TF-IDF to several corpora.
- Students will utilize pre-processing methods to prepare text data for analysis.
- Students will distinguish and categorize text using machine learning.

Watch:

- Preprocessing the Lobste.rs Text data https://mediasite.video.ufl.edu/Mediasite/Play/8ca04844316449fba996be01998660841d
- Introduction to Word2Vec and Sense2Vec with SpaCy https://mediasite.video.ufl.edu/Mediasite/Play/fad768505168474891a00688d26a3e3d1d
- Tag predictor for Lobste.rs dataset https://mediasite.video.ufl.edu/Mediasite/Play/9e3b6c25c1f84d4685fc8c0c4462c7351d
- Tariq Rashid: Topic Modeling - https://www.youtube.com/watch?v=Bxlzbck51SU

Required Readings:

- Data Science for Business, Chapter 10
- Introduction to NLP Using Spacy:

http://nicschrading.com/project/Intro-to-NLP-with-spaCy/

- Preprocessing Text for Machine Learning: https://machinelearningmastery.com/clean-text-machine-learning-python/
- Extra (Optional Reading) - Word Vectors for Document Similarity: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Assignments:

- Group Assignment: v 0.10 release
- Code Assignment: Training a Text Classifier
- Week nine discussion

Week Ten: Model Interpretability

Learning Objectives:

- Students will select important features for a model using information gain theory and validating with the learned model.
- Students will judge interpretability of a series of example models, evaluating using Miller's Law.
- Students will deconstruct a complex model into a simple and interpretable model.

Watch:

- Why Interpretability? And How? https://mediasite.video.ufl.edu/Mediasite/Play/b1cc59fb78ac4b87859dfb9ad6f5f8531d
- Introspecting our Score and Topic predictors https://mediasite.video.ufl.edu/Mediasite/Play/a0be1cfca53248358b127adb206d86ba1d
- Interpretable Machine Learning Using LIME https://www.youtube.com/watch?v=CY3t11vuuOM

Required Readings:

- Data Science for Business, Chapter 11
- An Introduction to LIME: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

- Ideas on Interpreting Machine Learning: https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning

Assignments:

- Group Assignment: Incorporate v 0.1 Feedback
- Code Assignment: Interpreting Your Models
- Week ten discussion

Week Eleven: Ethics, Privacy in Data Mining and Machine Learning

Learning Objectives:

- Students will identify ethical concerns in data science use cases.
- Students will evaluate data privacy in relation to data mining theory.
- Students will utilize anonymization techniques on a dataset.

Watch:

- Privacy & Ethics in ML / Data Mining https://mediasite.video.ufl.edu/Mediasite/Play/6c2fab3114a94c2dacbcd2deec9da0591d
- How I'm fighting Bias in Algorithms: https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms
- Ethical Machine Learning: https://www.youtube.com/watch?v=hDgXIUM3Rmw
- Privacy vs. Privatization: https://www.youtube.com/watch?v=JIo-V0beaBw
- You are the Product: https://www.youtube.com/watch?v=TAUTmxo1sCY

Required Readings:

- Data Science for Business, Chapter 12
- The End of Privacy: https://www.economist.com/node/202103
- Introduction to Differential Privacy: https://recherche.orange.com/en/differential-pivacy-or-how-to-anonymize-datas-while-managing-its-usage/
- Prejudice in Semantics: https://joanna-bryson.blogspot.de/2017/04/we-didnt-prove-prejudice-i

s-true-role.html

- Teaching Robots Right from Wrong: https://www.1843magazine.com/features/teaching-robots-right-from-wrong

Assignments:

- Group Project: v 0.2
- Code Assignment: Privacy & Ethics Written Assignment (Canvas)
- Week eleven discussion

Week Twelve: Final Project and What's Next?

Learning Objectives:

- Students will share and analyze what they have learned throughout the course and in the production of their final project.
- Students will evaluate others learning via peer evaluation.
- Students will judge topics we have learned and their applicability in their field.
- Students will create client-oriented presentations to effectively communicate and present findings.
- Students will determine future topics in machine learning and data science which may be of interest to them.

Watch:

- What happens next? https://mediasite.video.ufl.edu/Mediasite/Play/c15f81540c0c46baa5a53865ea8611591d
- The Birth of a Word: https://www.ted.com/talks/deb_roy_the_birth_of_a_word
- Data Mining for Good: https://media.ccc.de/v/30C3_-_5405_-_en_-_saal_g_-_201312291730_-_data_mining_for_good_-_patrick#t=0
- One video on a topic you enjoy on https://www.youtube.com/user/PyDataTV

Required Readings:

- (at least) One data science or machine learning article from:

  - FiveThirtyEight
  - KDNuggets
  - O’Reilly
  - Data Elixir
  - Data Machina
  - Medium
- Share a fun article you found on a topic we covered in class along with a synopsis
- Data Science Journal Sharing (Canvas)

Assignments:

- Final group evaluations
- Group project presentation and review
- Data Science Journal Reflection & Responses

Disclaimer:

This syllabus represents my current plans and objectives. As we go through the semester, those plans may need to change to enhance the class learning opportunity. Such changes, communicated clearly, are not unusual and should be expected.