School of Professional Studies

Degree Program: MS in Data Science

Course #: DATA 698.001

Semester: Spring 2022

**INSTRUCTOR INFORMATION**

Name: Arthur J. O’Connor

E-mail: Arthur.OConnor@cuny.edu

Phone: Office: 646-664-7267 Mobile: 646-236-4275

Bio: [https://sps.cuny.edu/about/directory/arthur.oconnor](about:blank)

Google Scholar Profile: <https://scholar.google.com/citations?user=8d_pWaUAAAAJ&hl=en>

**Weekly Lecture/Office Hours: Monday at 6:30 pm to 7:30 pm ET**

These one-hour weekly sessions will be held every Monday night at 6:30 pm East Coast time (barring a schedule change, for which you will be notified in advance). The first session (“Lecture/Office Hour #1”) will take place on **Monday January 31, 2022 at 6:30 pm ET**. Link: <https://us02web.zoom.us/j/84921815296?pwd=alk1Q05OTDhGKzNLd1BORStTeE9VUT09>

Meeting ID: 849 2181 5296

Passcode: 520510

As these sessions are recorded for the benefit of students who cannot attend them, please note the following disclaimer:

*Students who participate in this class with their camera on or use a profile image are agreeing to have their video or image recorded solely for the purpose of creating a record for students enrolled in the class to refer to, including those enrolled students who are unable to attend live. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live.*

**COURSE DESCRIPTION**

As an applied degree, the MSDS program culminates in a final Capstone course. Like the finishing stone of a structure, the Capstone or Senior Research Project represents your opportunity to apply what you’ve collectively learned in the Data Science program to a real-world problem or issue that is of professional or personal interest to you. Much like independent study, you choose a topic to research.

Result: Project which you might want to:

* Submit for publication in a journal or personal website
* Post on a social media such as LinkedIn or FaceBook(Meta)
* Upload to a code repository such as GitHub or Bitbucket

…to serve as a sample of your work and/or expertise for potential employers or colleagues in your personal or professional networks

**Group vs Individual Projects**

Unlike other courses in the MSDS degree program – in which students are expected and encouraged to work in teams – most capstone students chose to work of their own topic, rather than in collaboration with others, given the focus on **personal** career goals or interests.

Also, given the diversity of time zones, backgrounds, and personal styles of students, a group project may not be practical. However, students may be allowed to work collaboratively on a group project, at the instructor’s discretion. The project also may be organized in collaboration with a partner organization, for example, a local company, non-profit, or research lab – again at the instructor’s discretion.

To facilitate instructor/student coaching, there is a maximum of only 10 students per section.

**How the Capstone Differs From Other Courses**

In most MSDS courses, students learn to code and use libraries to wrangle the required data sets and perform different statistical and machine learning techniques on defined/discrete exercises.

The Senior Research project has a much broader/multi-disciplinary scope, requiring students to identify a problem or phenomenon and frame it in a practical context, develop a hypothesis or research question that proposes a relationship to variables inherent to that phenomenon, design and implement a model or set of models to measure or explore these relationships, and then interpret, develop and formulate conclusions from the results.[[1]](#footnote-1)

1. **No standard lesson plan or tests**. Each student typically works on her/his own topic, thus each research project is different. There are no common lesson plan, tests, quizzes or exams.

2. **Feedback is different from grading**. In most courses, the primary indicators of student performance are grades. In this course, you can request and receive feedback on your proposals and drafts **multiple** times.

In fact, there are only two points at which you’ll be graded: at mid-term, you will receive a letter grade on the quality of your submitted work, and at the end of the course, you will receive a cumulative grade for your collective performance – based on your proposal, mid-term draft, final project paper and presentation.

3. **Primary learning outcomes**: developing the ability to critically analyze an issue or phenomenon, develop a statistical model or machine learning technique to measure and/or solve it, and articulate your findings clearly, concisely and succinctly.

**LEARNING OUTCOMES**

Since the course’s structure is more of an independent study – each senior research project will be different – there’s no standard/common “lesson plan,” but there are some basic learning outcomes that one can expect from taking this course.

In identifying common traits of successful data scientists, a recent Harvard Business Review article[[2]](#footnote-2) (“Do Your Data Scientists Know the ‘Why’ Behind Their Work?” HBR, May 2019) cited business acumen, relationship-building, and, perhaps most importantly, the ability to explain results in simple, compelling language. Thus, the key objective of this course is about developing the ability to critically analyze an issue or phenomenon, develop an innovative, practical approach to measure and/or solve it, and articulate your findings clearly, concisely and succinctly.

As Data Science diffuses into more disciplines, the ability to express complex analyses clearly, simply and concisely in plain, direct, layman terms/language is becoming critical,[[3]](#footnote-3) with growing demand for “data science citizens” to liaise between technical teams and other stakeholders.[[4]](#footnote-4)

Admittedly, it is challenging it is to clearly articulate complex concepts and methods in plain, simple language. But as data science evolves from a back office to (more visible and highly-paid) front office roles – that’s what Data Scientists increasingly need to do.

At the end of this course, students will learn to:

* Identify a problem or phenomenon and frame it in as a function of a number of factors
* Develop a hypothesis or research question that proposes a relationship to variables inherent to that problem or phenomenon
* Explore and identify data sets to create variables that represent these factors, as well as statistical and/or machine learning method(s) to model (measure and/or predict) this phenomenon
* Design and implement a model or set of models to measure or explore these relationships
* Interpret, develop and formulate/articulate conclusions from the results

**TEACHING CONTENT**

In the first lecture, the class will review this syllabus in detail, with a focus on how to write a quantitative research paper. Below is an outline, with descriptions of the various sections:

* **Abstract**: a short (100 to 150 words) synopsis of what the research explored, its findings, and their significance. Note: **this is a last section you write**, after you’ve completed writing the rest of the paper.
* **Introduction**: The introduction should familiarize your reader with what you are trying to show, as well as the reasons for your research and what value you believe that it has.
* **Literature Review**: In a literature review, you examine prior studies into the causes and factors associated with the phenomenon you want to explore or effect that you want to measure in the study. By reviewing these studies, you see what data sets and models researchers used, and compare and contrast their findings with what phenomena that you’re exploring. A good lit review answers the question: so what makes your study different or more interesting than the current body of knowledge?
* **Research Question and/or Hypothesis**: This describes what problem you’re seeking to examine, phenomenon you want to explore or effect that you want to measure in the study.
* **Data and Variables**: This section describes what data you intend to use, how they were acquired, and how they represent the variables you have chosen.
* **Statistical Methods**: This section describes the methods you used to analyze the data.
* **Discussion of Results**: In this section, you describe the results of your statistical analyses, their significance, and how they compare or contrast with those from other studies.
* **Conclusion**: This section summarizes your final thoughts on your findings and what they show, as well as disclose limitations to the study and suggest future avenues for research.

Some papers vary a bit from this format, but most follow this same general flow. For a more detailed description of an outline and what is included in each section, here’s a pretty good example: <https://library.hccc.edu/c.php?g=452139&p=3154898>

**Statistical Validity**

An important concept in this course, which we will discuss throughout the course, is **statistical validity**, which is about the soundness of the research design and methods. Two dimensions of statistical validity are:

* **Internal Validity**: do the data sets and variables accurate represent the phenomena being explored?
* **External Validity**: can the results of a study can be generalized – not unique to those specific aspects of the analyses conducted (i.e. can one reliably make inferences from the results of the test sample to the population as a whole)?

Some concepts to consider and questions to ask yourself:

* Is the sample of observations selected for the test reflective of the population?
* Are the values of the independent variables not dependent on each other?
* Are there significant confounding or exogenous factors influencing the dependent variable and thus need to be controlled for in the model?

To clarify this last point: when identifying the key variables in your model, have you included control variables to account for other key influences on the dependent variable? Control variables represent major influences or contributing factors that need to be included in order to isolate the relationship between the dependent variable and the independent variables.

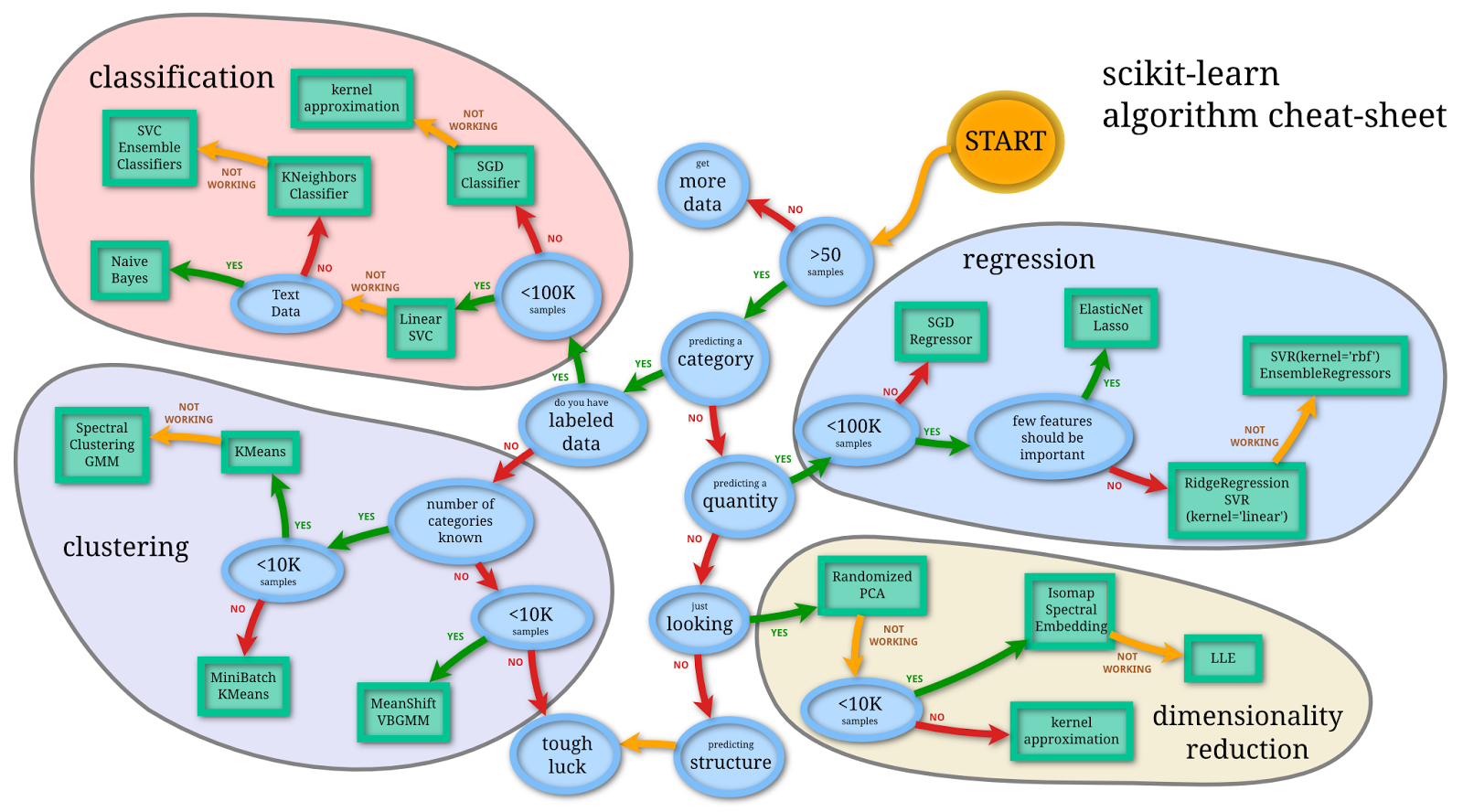
For example, in measuring the effect of a phenomenon on daily changes in a stock price, one would need to control for market movements that also influence price changes, i.e., use a systemic measure for stock market returns, such as the S&P index, to capture movements in the overall equities market.

**Choosing the Right Model**

What types of relationships among variables or end result the researcher seeks to measure or predict will determine what type of statistical or machine learning model is most appropriate for the research project.

For an overview of the main types of techniques and the key differences between statistical and machine learning modelling, the following article in Analytic Steps provides a good synopsis: <https://www.analyticssteps.com/blogs/5-statistical-data-analysis-techniques-statistical-modelling-machine-learning>

In machine learning, many students have found the following Scikit Learn algorithm “cheat-sheet” (below) to help them decide what type of machine learning technique to use, based on the amounts of observation and data types, and relationships the analysis seeks to measure or explore.



As the chart shows, the four main categories of ML models are:

* Regression: predict an outcome, based on its past performance;
* Classification: identify what classes new observations belong, based on results of a training set of data containing observations (or instances) whose category membership traits are known;
* Clustering: determine which observations belong to which clusters, based on their similarities to those in other clusters;
* Dimensionality reduction: simplify the dimensions of a feature set to improve accuracy and reduce noise in the data.

**Useful Tips for this Course**

In selecting a research topic, **choose a topic of domain expertise or experience**. Students with familiarity of the topic’s subject matter often find in much easier to accurately identify meaningful issues and design the research – as well as express themselves clearly and concisely – than those who do not.

**Make sure there are data sets available** and accessible to use in your model that accurately present the variables you intend to use in our research. It’s relatively easy to identify interesting topics; the trick is to identify those that can be measured and tested via statistical means.

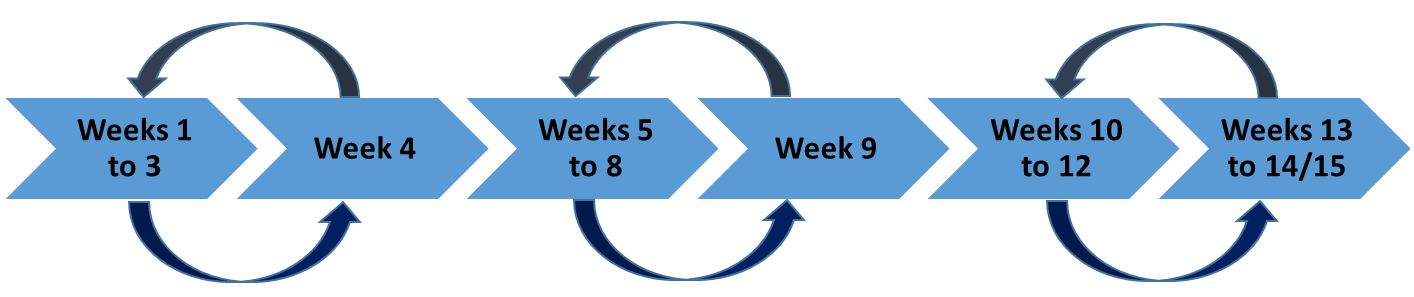
Key to getting your proposal approved is demonstrating that you’ve validated that there are data sets you can use to measure the hypothesis you want to test. ***You don’t want to invest a lot of time and energy developing a proposal – only to discover that you can’t conduct the analysis required for the rest of the research paper.***

**Plan out your work and budget your time**. For the mid-term draft, focus on finalizing the Title, Introduction, Literature Review, Hypotheses, and Data and Methods sections. For the balance of the semester, focus on the most time-consuming aspect of most quantitative research projects (data acquisition/wrangling).

After conducting how exploratory analysis and fine-tuning your models, draft your Discussion of Results, Conclusion, and lastly, the Abstract.

**COURSE SCHEDULE AND DELIVERABLES**

The course has three, iterative phases: 1) research proposal, 2) completed/approved mid-term draft, and 3) completed/approved final paper and presentation – only two of which (mid-term and final paper/presentation) will be officially graded, as shown in the figure below:



The 14/15 week sequence of course learning activities and deliverables are as follows:

**1st Phase/Deliverable: Research Proposal**

You will choose a topic by developing and submitting a short (no less than 3-page) proposal of their research project, which must be approved by the Instructor. This proposal should include a clear and concise articulation of:

* What problem/issue/phenomenon you have chosen to explore – research question(s), theory and hypotheses – and why it is important, significant or meaningful
* What prior research has been done on the topic: explain originality of your research [note: just a few paragraphs on relevant areas of research, not a detailed literature review]
* What research methods you intend to use; what models, variables and data sets will be used to test hypos and theory

**How Results Will Be Measured and Evaluated**

On or before Week 3, you will **e-mail** you proposal draft to your instructor, who will review and provide feedback on the proposal, either approving it on its initial submission or requesting revisions based on a general assessment of the work and/or specific feedback such as questions, comments, and/or suggested changes for further work by you.

In evaluating the proposal for approval, the Instructor will assess:

* Clarity and specificity of the proposed research topic
* Sufficiency of scope and complexity of the issue and proposed analysis
* Availability and statistical validity of the data sets and model(s)

…to help ensure that the project meets basic requirements and can be completed within the semester timeframe. [[5]](#footnote-5)

**2nd Phase/Deliverable: Mid-Term Draft**

Working from the proposal, you begin fleshing out the first sections of the research paper: the Introduction, Literature Review, Hypothesis, Data and Variables and Statistical Methods.

You will draft a medium length (no less than 10-12 pages) paper representing the first half of your research study, which will form the basis of the mid-term letter grade in Blackboard Grade Center.

On or before Week 8, you will **e-mail** your mid-term draft to your instructor, who will review and provide feedback on the proposal, either accepting it immediately or responding with comments requiring more revision to/work on it.

**3rd Phase/Deliverable: Final Paper and Presentation**

Within the last three or four weeks of class, students submit their final drafts, which (like the proposal and mid-term draft) are reviewed and approved by the Instructor.

Within the last two weeks of the class, students will present their research. Final project presentations should be no more than 15-20 minutes. Presentation don’t require the polished delivery and high production values of a [TED Talk](https://www.ted.com/talks), but I expect students to be fully prepared to be concise and hit the highlights/key messages.

In terms of audience, presenters should assume an audience of peers as well as non-technical management. The focus should be on what the research is about and what the results actually mean or reveal, but speakers should also be prepared to respond to technical questions about methodology, if they arise.

While some students produce materials (such as Power Point slides) specifically for their final presentation, such additional materials are not required.

**Note that students have the option to present their final project during a weekly Lecture class or send a recorded video of the presentation to the instructor.**

**Reading/Sources**

Sources (if required) will be identified once the proposal for the Senior Research Project is developed, submitted and approved.

**Transparency**

In the interests of transparency and setting expectations, all proposals and mid-draft drafts, with instructor feedback, will be posted in the Materials section on the Blackboard course site, so that all students can get a better sense of the level of quality and content required to be successful in the class.

**SafeAssign**

The Instructor will submit all proposals to SafeAssign via the DirectSubmit function in Blackboard to assess and ensure the authenticity of student work.

**FEEDBACK AND GRADING**

Given the iterative process of designing, developing and producing a research paper – and unlike other courses in the MSDS program – there are multiple opportunities during in the course for students to solicit feedback on their work product from the instructor, but only two of which students will be formally graded:

* At mid-term, you will receive a letter grade on the quality of your submitted work;
* At the end of the course, you will receive a cumulative grade for your collective performance, based on your proposal, mid-term draft, final project paper, presentation, and, if applicable, your classroom participation.

Grades are given on a letter basis (A, B, C, etc.). Given that each student’s research project is different, and the lack of tests, quizzes or exams in the senior research project course, there are no numerical scores or grade weightings; thus, there is no point scale or rubric to convert numerical scores to letter grades.

In the absence of hard/quantitative measures to assess the quality of a research paper, students must rely on the direction and feedback they receive from their instructor/advisor.

That said, the following criteria will be used to grade the quality of your senior research project. The first set of guidance (“General Criteria”) refers to the overall quality of the project; the second set of quality descriptors apply to specific sections or activities (“Domaine-Specific Criteria”) to help differentiate what is considered “Accomplished” vs. “Proficient” vs. “Needs Improvement.”

**General Criteria**

Significance and Originality of the Topic

* Does your topic represent or address an interesting issue? Will the results have a practical application to a real-world business or social problem?
* Does your research borrow from, or represent a slight variation to, an existing study (perhaps a Kaggle competition)? Or are you designing and creating your own variables based on a novel combination of data sets?

Quality and Sophistication of Analyses

* Do the statistical methods selected accurately test or measure the research question or hypotheses proposed (i.e., quality model design)?
* Do the data sets and variables based them accurately capture/reflect the phenomenon being explored (i.e., does the research study have internal validity)? In this model designed so that the results can be recreated/generalized/applied to other/similar cases (i.e., does the research study have external validity)?
* Does the analysis entail simple descriptive statistics, or does it involve complex classification, regression, or clustering algorithms (sufficient for a data science masters’ thesis)?

Quality of Writing

* Is the final draft in final form (no errors in spelling, grammar, diction, formatting or citations)?
* Does the nature and scope of the research, the design of the analyses, the results and the significance of result clearly and concisely explained in simple, compelling language?

**Domain-Specific Criteria**

|  |  |  |  |
| --- | --- | --- | --- |
| **DOMAIN** | **ACCOMPLISHED** | **PROFICIENT** | **NEEDS IMPROVEMENT** |
| Introduction | The research question is clearly stated, can be answered by the data, and the context of the problem clearly explained. | The research question is unclear and/or not supported by the data. | Research question is ambiguous, unclear, or not stated. |
| Data Display | Includes appropriate, well-labeled, accurate displays (graphs and tables) of the data. | Includes appropriate, accurate displays of the data. | Includes appropriate but no accurate displays of the data. |
| Data Analysis | The appropriate statistical test(s) was used for the data and interpretation was clear. | The appropriate statistical test(s) was used but interpretation was not fully clear or well-articulated. | The incorrect statistical test was used and/or not justified for the data as presented. |
| Conclusion | Conclusion includes a clear answer to the statistical question that is consistent with the data analysis and the method of data collection. | Conclusion includes an answer to the statistical question that is consistent with the data but not with the data collection method. | Conclusion does not include an answer to the statistical question that is consistent with the data analysis. |
| Final Presentation | Speaker articulates the nature of the research and shows the study’s findings clearly, concisely and succinctly | Speaker hits most of the highlights | Speaker does not clearly and concisely explain the nature of the research project nor its findings. |

**PROGRAM LEARING OUTCOMES** (addressed by this course)

Learning outcomes of the Capstone course align with the program learning outcomes, as follows:

* Data Acquisition, Management and Programming: use industry standard data science and analytics packages to collect, describe, clean, format, model, explore and verify structured data, unstructured data and big data.

*Students will apply data wrangling and munging skills they’ve learned, using R and Python*

* Foundational Math and Statistics: demonstrate understanding of linear algebra – differential equations, linear and non-linear programming (NLP), algorithmic search methods for optimization, integer programming (IP) – probability, Bayesian statistics, univariate and multivariate calculus.

*Students will apply what they’ve learned in their probability, statistics, and computational math courses*

* Modeling: Use statistical and machine learning modeling techniques to design, build and test/assess models.

*Students will apply what they’ve learned in terms of statistical data models and algorithmic/machine learning models to design, build, implement, and measure their predictive or descriptive power.*

* Dissemination: Develop/write/present reports to explain/present their models, results, and analyses in plain and easy-to-understand language.

*Students will deliver presentations and report of their project results.*

**Classroom Participation**

Unlike other courses, the Capstone course does not have structured weekly Discussion Board topics. As such the only two forums are the “Getting Acquainted” section for the first weeks of the class, and the “Ask the Instructor” section throughout the course.

**KEY DATES (Subject to change)**

|  |  |  |
| --- | --- | --- |
| **Year** | **Semester** | **Class** |
| 2022 | Spring | DATA 698 |
| **Week** | **Week of** | **Description** |
| 1 | 31-Jan | Introductions |
| 2 | 7-Feb | Proposal development |
| 3 | 14-Feb | Initial drafts due |
| 4 | 21-Feb | Final Review / Approval of proposals |
| 5 | 28-Feb | Weekly check-in on research strategy, data selection, methodology |
| 6 | 7-Mar | Weekly check-in on research strategy, data selection, methodology |
| 7 | 14-Mar | Weekly check-in on research strategy, data selection, methodology |
| 8 | 21-Mar | Initial Mid-Term Drafts due |
| 9 | 28-Mar | Final review / Approval of mid-term drafts |
| 10 | 4-Apr | Weekly discussion / progress report |
| 11 | 11-Apr | Weekly discussion / progress report |
| 12 | 18-Apr | Spring Recess |
| 13 | 25-Apr | Weekly discussion / progress report |
| 14 | 2-May | Weekly discussion / progress report |
| 15 | 9-May | Submission and presentation of Final Project |
| 16 | 16-May | Submission and presentation of Final Project |

**SUBMISSION REQUIREMENTS**

Final project papers are required to be medium length of no less than 15 to 20 pages in length. Papers should be prepared in Microsoft Word format (so that the instructor can provide feedback via the “Track Changes” feature. Text shall be in Times New Roman, 12 point font.

**ACCESSIBILITY AND ACCOMMODATIONS**

The CUNY School of Professional Studies is firmly committed to making higher education accessible to students with disabilities by removing architectural barriers and providing programs and support services necessary for them to benefit from the instruction and resources of the University.

To ensure consistent accommodations with each student, and to inform students of support services available, please feel free to make referrals to the Office of Accessibility: [http://sps.cuny.edu/student\_services/disabilityservices.html](about:blank). Early planning is essential for many of the resources and accommodations provided.

If any student has specific accommodation needs, please contact Christopher Leydon at [disabilityservices@sps.cuny.edu](about:blank) or (646) 664-8616 to discuss these issues.

**ONLINE ETIQUETTE AND ANTI-HARASSMENT POLICY**

The University strictly prohibits the use of University online resources or facilities, including Blackboard, for the purpose of harassment of any individual or for the posting of any material that is scandalous, libelous, offensive or otherwise against the University’s policies. Please see: [http://catalog.sps.cuny.edu/content.php?catoid=2&navoid=205](about:blank)

**ACADEMIC INTEGRITY**

Academic dishonesty is unacceptable and will not be tolerated. Cheating, forgery, plagiarism and collusion in dishonest acts undermine the educational mission of the City University of New York and the students' personal and intellectual growth. Please see: [https://sps.cuny.edu/about/dean/policies/academic-and-student-policies/academic-integrity](about:blank)

**STUDENT SUPPORT SERVICES**

If you need any additional help, please visit Student Support Services: [http://sps.cuny.edu/student\_resources/](about:blank)

1. Based on the philosophy that real learning comes primarily from within: not from your teacher’s neatly presented ideas, but from the connections your own brain forms between them. [↑](#footnote-ref-1)
2. “Do Your Data Scientists Know the ‘Why’ Behind Their Work?” HBR, May 2019, <https://hbr.org/2019/05/do-your-data-scientists-know-the-why-behind-their-work> [↑](#footnote-ref-2)
3. “Analytics translator: The new must-have role,” HBR, Feb 2018, <https://hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role> [↑](#footnote-ref-3)
4. “The Burtch Works Study Data Science & Predictive Analytics, 2020 <https://www.burtchworks.com/wp-content/uploads/2020/08/2020-Burtch-Works-Study-DS-PAP-Slides.pdf> [↑](#footnote-ref-4)
5. As the old saying goes, “A good masters’ thesis is a **completed** masters’ thesis.” [↑](#footnote-ref-5)