INFO 511 Foundations of Data Science
Final Project Information

This course requires a Final Project to be completed that demonstrates student's learning. There will be MILESTONES that both check knowledge and ensure timely completion of the project. The course cannot be passed without completion of Milestones and Final Project and Presentation.

The Final Project has three aspects:

Final Project in GitHub: The Final Project GitHub will be a repository of the work completed for every aspect necessary to complete the project. Main aspects of the Final Project in GitHub is to demonstrate knowledge and skills used, understanding of reproducibility and collaboration, and addresses the era of Open Science, which focuses on data projects being open, accessible, and reproducible.

Final Project Written Report: Whether you select a traditional research report, data visualization, or GtiHub design, you will be required to complete a maximum of three-page report to demonstrate your data communication skills. Key aspects of data communication must be in the final report. The information is to include the dataset used (cited), any and all data processing and manipulation and why, and shall include a general question (or research question), methods and analysis section, and finally a conclusion with next steps for recommended research.

Final Presentation: A final presentation is expected for each team or individual. Each person is expected to present in front of the class and will be graded on conveying data information for a broad audience. The presentation may not be any longer than 8 minutes (you will be stopped if over time).

Final Project Options:

Final Projects may take one of three focus areas, though all of the options, each requiring proper data science processes, will have overlap. The focus selected should be relevant to the student's personal growth and exploration, though it is expected that you demonstrate advanced knowledge in all aspects.

All options will require students to complete and demonstrate a formal question, use of method/data processing, analysis, and conclusion and all must include citations (data source must be cited). Each option will include the methodologies (data processing) that are discussed in class. This means that students must all demonstrate EDA and wrangling and notate what they do to the data. Though, for example, selecting data visualization focus option means that the final report out has an emphasis on data viz, though it will have a purpose (a question) that drives the report. Similarly, for all other focus options, it is expected that data visualization will be part of the EDA and processing and will be well documented within the reports or data management documentation. As such, each focus

area will have overlap, though the final project to be turned in and presented is how the main focus is demonstrated.

1) **Research/ Formal Reporting Focus:**
   In this focus area, you will create a formal research question or data question and use data to answer the question. This focus area is ideal for anyone wanting to start working on a thesis, a publication, or interested in long-term research. Formal and well written reporting in data science is an advanced skill and highly recommended. Students are required to state a research question, identify a data set and use the key features/ variables to answer the question. A clear and well explained methods and analysis section are required, along with a formal conclusion (was the question answered and how…) and followed by a formal next step in research recommendation.
2) **Data Visualization:**
   A main focus in data visualization will demonstrate the student's addressing a question through data visualization. The final report may have an infographic image or other format of advanced data visualization. The final report must also contain a 1-2 page informative report. This report should indicate the research question, address the data set that was used to answer/demonstrate information about the question, and all processing, etc. will be included. A concise written findings will complement the data visualization to explain how the data answered the question.
3) **Data Management:**
   As data knowledge and data management are key components to data science, the focus area for data management will encompass the creation of a data management platform with detailed documentation. The data management may be through GitHub or the creation of a data dashboard. As noted, this option still requires the exploration of a data set and details of the data exploration and data management processes will be required for the final reporting as well as the detailed documentation for the development of the GitHub or dashboard build. Selection of this focus area requires the student create more than a basic GitHub repo. The page must be advanced and creative and have a purpose and documentation to the process. This applies to any dashboard that is built. A data set must be identified and the reason for data management stated (research question or purpose), and documentation for data analysis and citations as well as build details are required.

Final Project Report: DUE May 06, 2025 50 points

The Final Project Report is a written report that demonstrates the student's data communication and general data science and statistical knowledge.
Objectives: The student will demonstrate what they have learned in the course, with a major focus on their coding (demonstrated in Final Project GitHub), their understanding of using data preprocessing and exploratory analysis, completing a statistical model that

aligns with the research question, writing up results, completing meaningful visuals, and being able to convey the information for a technical and non-technical audience. The latter part means that the report should be understood by your professor and technical users, but not overly convoluted so that a non-technical person misses the point of the report. Being able to coney information can be a difficult task in data science, and learning to do this early on is important.

Final Project Report Requirements: (The things the report will be graded on)

- Clearly state research question or research purpose:
    - The final report must have a clearly stated research question or detailed purpose. Consider this the *thing* you are wanting to learn from the data. The question may be specific and addresses an area of your research domain, or it may be exploratory in nature. Finally, it may be an exploration of a data set such as a comparison and exploration of processes that will be listed in the data management option. Either way, the question and purpose must be clear.
    - The question or purpose is guided by a theory, policy, framework, or common discipline initiative. Exploration is a common aspect of data science, however, as we discussed for this course specifically, it is ideal to understand the reasons *why* you are exploring the data. For example, you might consider is there a common business or finance guidance that drives certain questions and data exploration? Is there a policy or a known theory that gives reason to explore the relationship between specific features/variables? Or is there a new way to consider managing data or is there a new dataset that warrants cataloguing and feature/variable exploration, or would you like to increase your advancement of data management and building GitHub and websites related to data? These are all viable examples and should be clearly indicated in your project.
- Process and analysis/Methods section: Details on the methods used and why. Should include topics covered in this course.
    - Data: Accessing data, where did the data come from? Why this data? How was it accessed? Were considerations made for the security or privacy of the data—i.e., was IRB necessary/involved, was it secure, is it public and anonymize? Who owns the data and how was it originally collected (this latter part can be brief but do let the audience know why the data was originally collected (e.g., the data is from the National Center for Educational Statistics and the dataset is administrative data, from the Education Demographic and Geographic Estimates (EDGE) program (NCES, 2020). The data was collected in 2020 and is part of the United States Census (NCES, 2022). In this case you would state the year the data was collected, who collected it researcher, government agency, etc., what was the reason the data was collected (for a research project, public health initiative, etc.) .
        - Cite the data! And each citation in the reference section.

        - Data preprocessing, cleaning and EDA steps taken: Explain the summary statistics, cleaning process, how missing data was handled, and any variable issues that needed to be addressed.
        - Visualization (EDA) may not be necessary unless there was something salient about why you need to share the information. It is recommended that data visualization is saved for the final analysis.
    - Model selection needs to be clearly stated and why. Use models that were taught in class. Please note, if you select a more advanced model, you will be graded on that model being conducted and reported accurately. It is wise to select a simple question that can be addressed by the models taught in class and complete the final project in a highly professional manner. Please also account for your time. More complicated models and work require more technical reports and have the potential for more risks that can eat up time very quickly. Do not over complicate this, just make it very very good.
- Results section: make sure the audience understands the whether the question was answered (either supported or opposed but make it clear).
    - Visuals to support the results. If necessary, an appendix may be used without negatively affecting the maximum page amount. Make sure if visualization is included in the appendix, it is necessary.
- Next steps/future research recommendation: This is a short section where critical skills used to analyze the data is conveyed to make recommendations for future research/exploration. Data science is rarely completed, there are always more questions that the data reveals and answers that the data can support. Once you answer the research question, think about what the data "showed" and make a recommendation for next steps for future data scientists to conduct.
- Conclusions: The conclusions for this project will be a succinct summary of the project.
        - Essentially, it is a professional summarization of the entire scope of work that takes the technical out of the report and makes it accessible for a non-technical audience. Think of it like reporting to the CEO or what a journalist will read to report on your information, or how you would explain your work for a conference proposal.
- In-text citations: Please stay consistent in your citation. Please make sure it is noted in Milestone 1 which citation style is used and do it well.
- References: all in-text citations will have a corresponding reference page. This page will not be included in the maximum page count. Meaning, you may have additional pages that are used for references.
- Generative AI Tool use acknowledgement: Please include this in the final submission and document throughout the entire project process how generative tools were used.
- Appendix: This is optional. The appendix is a good place to add extra visuals that were used during the EDA and cleaning phase to explain why something was done. It

is also a good place to explain additional information in graphical or table details about the data. Please be sure to follow consistent citation appendix protocols.

- *Researcher* Bio-Sketch: Many research publications, grants, and for professional purposes, will require a bio-sketch. The bio-sketch tells a little about individual's professional skills and their role on the project. Each person on the project will have a bio-sketch with that includes their name, degree plan, year in their degree plan, and their role on this project. This will not count towards your maximum page count.
- Peer review recommendations response page: this page does not count towards your 3-page maximum but is required. You must list the recommendations that your peer made and respond their comments/recommendations.
  Responses may be simple as:
  - Recommendation accepted and changes/alterations made…
  - Recommendations rejects and why…

Final Project Page Length: The Final Project is to be completed as a professional report, whether it takes a business or research focus and should not be more than 3 pages long in technical/scientific reporting. As detailed above, exceptions for page length are made for References, Appendices (visuals), Generative Tool Use, Bio-Sketch, and Review Responses. The main body of the report must be contained within 3 pages. Points will be taken off for reports with information that extends beyond 3 pages. The page length is for you to practice professional and comprehensive writing and reporting of the most relevant information.

The Grading of the Final Report: The Final Report will be graded based on all elements being included in the final report.

Final Project Presentation: Due April 15, 2025 30 points

*All presentations are due on the same day. Students will be randomly selected and told when they will present, but alterations cannot be made to the presentation to remain fair on timing.

Each person will present on their project for 5-6 minutes. Each person is expected to have a part in the presentation. There are a lot of people in this class and a timer will be set to ensure presenters stay on time. Points are taken off for going over, we cannot be disrespectful of our colleagues' time, so plan ahead and practice!

The presentation can be in a conference style, reporting to the board of directors, or to your research colleagues, the freedom in the scenario is yours. The presentation must convey the information for a general audience to understand what the project entails and why and what the outcomes are.

For help in how to think about the presentation, you may want to plan for presenting the findings to stakeholders. Start with the question, and include the steps taken to answer the question (data, methods, model). Consider how to tailor the message to both technical and non-technical audiences.

Main point is the information must be conveyed succinctly and with professionalism.

The Final Project Presentation should include:

- The question should be stated and the *why* for the question. Make sure it aligns with the Final Project Report. Make it very succinct.
- The dataset that was used for the project should be discussed. The slide should have a citation of some identifying information about where the data was accessed. If privacy, IRB, or other considerations were necessary, this needs to be stated.
- Data Process: Details for cleaning and EDA: you can have a lot of information on a slide and yet only report the most important details. For example, you may show summary statistics and visual checks, and a list of the steps taken, but talk about them briefly and focus on areas that were most salient such as focusing on how the missing data was handled or how you addressed outliers if those were the most important issues.
- Make sure the features/variables are known to the audience. Meaning, make sure the predictors are indicated, and the outcomes are indicated. X variable was used to predict Y outcome.
- Modeling, methods, and results: Make sure the features/variables are discussed in the appropriate manner.
- A visual that conveys the results
- Conclusion and next steps recommendation

MILESTONES:

Milestone 1: 1-Page Proposal - 15 points

For Milestone 1, you will turn in a 1-page proposal that details the focus area you will take for your final project (research report, data visualization, or data management).
The plan is detailed, demonstrating clear thought about what the project will entail and the feasibility to complete the project successfully by the due date.
It is stated if it is a team or an individual project.
The research question is stated.
Frameworks, theory, policy, etc. is stated as the guiding principle for the project. See me if this is unusual for your discipline and we can discuss.
The citation style is stated.
Dataset that is likely to be used, or how data will be found is stated.
Given the question, it is clear the type of data or variables/features the data will need to have. This information may include details about the data being publicly available data that is public health, business, education, agricultural data, etc. If you are using faculty/mentor data, it must be noted.

If IRB or clearance to access the data is required must be stated and how to handle this will also be stated, such as IRB is required, or a secure cloud is required to access data.

Milestone 2: 1-Page Data, Methods, Analysis Reporting Update - 15 Points

Milestone 2 requires you to demonstrate that the data was identified, downloaded, and the methods have started (EDA, processing, etc.). Issues with data and how it was handled is detailed. Missing data and use cases required by owner of data is detailed. This means that information that is pertinent to using the data is conveyed in the report update so that the instructor is made aware of how the data is being handled.

Milestone 3: Final Draft Turned in for Peer Review – 20 points

A final draft is required to be turned in and will be shared with a colleague. You will also receive a colleague's draft and will review it.
In data science, peer review, quality checks, etc. are common and good practice. In research, peer review is salient for publication and moving research forward.
For the project, you will demonstrate your peer review skills. You will be given a colleague's final draft and will make recommendations that are detailed and meaningful for the betterment of your colleague's report. You may comment on coding and make recommendations for tidier workflows or better coding functions. You may ask questions that are not clearly answered about how the data was handled. Meaning, if the report states the combination of variables/features to create a new one that is not clearly stated as to why it was done is a good place to ask a question or make a comment. Your job is to help your colleague produce quality report.

Milestone 4: Return Peer Review – 20 points

This step is timed so that you do not delay returning your colleague's reviewed work.
Each person will receive their report back with the comments and recommendations and will be required to make the changes or consider why they should not. Not all comments and recommendations are necessary, though it is important to consider if it would strengthen your work. This decision making process is a difficult one in research and data science, and it is important to make a response to the recommendations—recommendations accepted and changes made, or recommendations rejected and why. This final documentation will be included in your final submission, though it will not count towards your final report page limit.