

Final Project Guidelines

Throughout this course we explore a wide variety of data acquisition, data management, data manipulation, and analysis methods. An essential component of analytics and machine learning work is knowing when and how to apply a particular type of methodology, algorithm, or software tool when presented with a new challenge. For your **Final Project** of this course you are responsible for selecting your own data sources and defining your own research question(s). Once you've identified the data sources you are interested in working with, you should then define one or more research questions you will attempt to answer as part of your work for the Final Project. The data you select **must** serve to answer one or more formal research questions that you define for purposes of framing your **Final Project** work.

Your **Final Project** is comprised of three separate deliverables:

1. A formal Final Project Proposal;
2. Your Final Project writeup + Python code (in the form of a Jupyter or Colab Notebook);
3. A “live” presentation of your work during our final Live Session.

A summary of the schedule and scoring for these deliverables is provided below.

Deliverables Schedule

Deliverable Date Points

Proposal	10%
Final Project	75%
Final Project Presentation	15%

Final Project Checklist

To receive full credit for the Final Project, you will need to deliver on all of the items mentioned in the checklist shown below. **Please read carefully through this checklist before you make your project proposal.** You are (within these checklist constraints) strongly urged to limit scope and make the necessary simplifying assumptions so that you can deliver your work on time.

- ☐ **Proposal** describes your motivation for performing this analysis.
- ☐ **Proposal** describes from where you plan to source your data.
- ☐ Your project has a recognizable and reproducible “data science workflow.” [Example: First the data is acquired and explored, then necessary transformations and clean-up are performed, then the analysis and presentation work is performed]
- ☐ You are **NOT** allowed to use any dataset that has been provided to you in the course. You may not use a data set from UCI or Kaggle. You are also **NOT** allowed to use any data set that is embedded within any

Python library (e.g., any of the scikit-learn data sets).

- ☐ Your data must have both numeric and categorical features.
- ☐ **Project** includes statistical analysis and graphics that describe and/or validate your data (e.g., EDA). You should clearly explain your data cleaning process. You must address how you deal with outliers, NaNs, and other discrepancies. Make sure to explain to test different imputation alternatives and explain your choice.
- ☐ **Project** includes at least three data reshaping, aggregation, and/or grouping operations. Using tools we have explored such as `pivot_tables`, `cross_tab`, etc.
- ☐ **Project** includes the use of data-appropriate data preparation and feature engineering and feature selection methods.
- ☐ **Project** includes at least five graphics that support your conclusion(s).
- ☐ **Project** includes at least one statistical analysis that supports your conclusion(s).
- ☐ **Project** includes at least one Python feature that we did not cover in class, e.g., something you discovered during your coursework or that you found to be necessary for completing your research. It should also include at least one Python model we have not discussed, such as KNN, SVM, etc.
- ☐ **Project** includes data scaling, training, testing, and cross-validation of at least four separate types of models that we have discussed. The models can include but are not limited to multiple regression, binary logistic regression, Poisson regression, Negative Binomial, etc. You should make sure to discuss the data structure and potential issues such as imbalanced data and how you dealt with these issues.
- ☐ **Project** includes interpretation and inclusion of categorical and numeric variables (although they may be in different models)
- ☐ **Project** includes a clear presentation, commentary, and explanation of EVERY step.
- ☐ **Project** includes an implementation of at least two (2) models that we have not covered in class. It must also include an ensemble approach (separate from the two models previously mentioned in this directive).
- ☐ **Project** includes analysis of models using metrics for both regression models (R-square, AIC, etc) and classifiers models (confusion matrix, ROC curve, etc)

- ☐ **Presentation.** Was the presentation delivered in the allotted time (8 to 10 minutes)?
- ☐ **Presentation.** Did you show (at least) one challenge you encountered in code and/or data, and what you did when you encountered that challenge? If you didn't encounter any challenges, your assignment was clearly too easy for you!
- ☐ **Presentation.** Did the audience come away with a clear understanding of your motivation for undertaking the project?

- ☐ **Presentation.** Did the audience come away with a clear understanding of at least one insight you gained or conclusion you reached or hypothesis you “confirmed” (rejected or failed to reject...)?
- ☐ **Code and data.** Have you delivered the submitted code and data where it is reproducible and self contained—preferably in a Jupyter or Colab Notebook on GitHub? Can your results be fully reproduced with what you’ve delivered? You won’t receive full credit if your code references data on your local machine.
- ☐ **Code and data.** Does all of the delivered code run without errors?
- ☐ **Deadline management.** Were your draft project proposal, project, and presentation delivered on time? Please turn in your work on time! You are of course welcome to deliver ahead of schedule.

Policy on Online References

Remember that you can take work that you find elsewhere as a base to build on, but you need to acknowledge the source, so that your grade is based upon what you actually contribute, **not** on what you start with.

Proposal Guidelines

Your first deliverable for this Project is the **Final Project Proposal**. The Proposal for the Final Project will be submitted in the form of a formal research proposal document. . Furthermore, you need to ensure that the Project you are proposing will satisfy all of the requirements specified in the **Final Project Checklist**. Your proposal must include each of the sections outlined below and must be submitted in the form of a Jupyter or Colab Notebook that has been posted to your GitHub repository.

Introduction (2 Points)

This section should provide some context for the basis of the research questions you plan to answer without actually describing the research questions themselves. For example, if your research questions are focused on a health related issue, you might provide a brief summary of how many people are affected by that issue each year either regionally, nationally or globally, including any infection rate or mortality statistics you were able to gather. Basically, in the Introduction you are trying to make the reader understand why the research questions you are going to propose are relevant and should be of interest to them.

Research Questions (3 Points)

Provide a single succinct sentence describing each of your research questions. Then provide a paragraph or two explaining how the results of your research might be used/implemented in the "real world".

Data to be Used (1 Points)

Clearly identify the sources of your data and explain the methods you will use to collect the data from those sources, e.g., "Data will need to be collected from this source via scraping of a web page.", etc.

Approach (4 Points)

Explain how you plan to manage the data you are collecting: e.g., will you be storing it within some sort of database, etc.? Explain what types of statistical analysis you plan to utilize to help answer your research questions. Explain any graphics you plan to generate to help answer your research questions. The reader should come away with a clear understanding of how you plan to proceed with your work. Keep in mind that

since this is a proposal, you need to be able to convince the reader that your proposed project is: a) realistic; and b) feasible within the time allotted for the project. Also, be sure to clearly articulate the roles and responsibilities of each team member for your Project work.

Your Final Project Proposal Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and subsection headers, and be free of spelling and grammar errors.

Upload / submit your Jupyter Notebook within the provided Final Project Proposal via submission portal.

Proposal Approval

Once you've submitted your proposal its content will be reviewed for purposes of determining whether or not what you have proposed is acceptable as a Final Project. If so, you will be approved to start work on your Final Project. If not, you will receive feedback regarding any issues that need to be addressed before you can receive approval for your Project. You will be able to re-submit your Proposal as many times as necessary to achieve the required approval.

Your Second deliverable for this Project is your Final Project Jupyter or Colab Notebook. It should contain a combination of Python code cells and explanatory narratives contained within **properly formatted Markdown cells**. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

- 1) **Abstract (5 Points):** Use 250 words or less to summarize your problem, methodology, and major outcomes.
- 2) **Introduction (5 Points):** Describe your project, including the scientific or business motivation for the research question you have chosen to answer. This section should summarize the content of your Final Project Proposal, so be sure to explain your research question, describe the source and content of the data set you have chosen to work with, and summarize your approach to meeting the requirements for the Project.
- 3) **Research Approach (5 Points):** Explain + present the end-to-end methodology you made use of for all aspects of your Final Project work, including your EDA, data preparation, and investigative analysis work. Be sure to include a description of your data management strategy as part of your narrative.
- 4) **Exploratory Data Analysis (20 Points):** Explain + present your EDA work including any conclusions you draw from your analysis, including any preliminary predictive inferences. This section should include any Python code used for the EDA.
- 5) **Data Preparation (10 Points):** Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering techniques you have applied to the data set. This section should include any Python code used for Data Preparation.
- 6) **Prepped Data Review (5 Points):** Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.
- 7) **Investigative Analysis & Results (40 Points):** Explain + present your investigative analysis work, including any Python code used as part of that process. Provide and explain your answers to your

research questions. This should include all your models and their explanations (and metrics assessments).

8) Conclusions (10 Points): Summarize your work and clearly state the conclusions of your research. Were you able to answer the research questions you originally posed in your Proposal? Comment on any potential future extensions of the work you've completed for the Project.

Your Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and subsection headers, and be free of spelling and grammar errors. Furthermore, your Python code should include succinct explanatory comments.

Upload / submit your Notebook within the provided Final Project submission portal or submit it on Github and just provide the link.

Your third deliverable for this Project is an approximately 8-10 minute live presentation of your work. Your presentation should include a brief overview of your research questions and the data you selected to work with, your EDA work, a high-level explanation of your data preparation + feature engineering process, a discussion of your approach to answering your research questions, the results of your research, and your conclusory statements.