Applied Data Analysis
Winter 2024
Project Guidelines

Project Type and Scope

The project is an important component of this course. It is an opportunity for you to work on a data analysis project of your choice. Also keep in mind that course projects are one of the favorite topics during a job interview.

Your project should significantly advance your skills beyond what is covered in class.

For some examples of data sets you could use, look at the Ed Discussion thread for sharing interesting data sets. There are also examples of past projects on Canvas.

There are many topics of active research, and you will likely find research articles containing sophisticated solutions for these problems. Given the available time, however, you cannot expect to develop a state-of-the-art solution during this course project. Use this opportunity to delve deep into a problem of your choice, apply what you know so far in data analysis, study the literature and learn new techniques relevant to your problem, develop a serious solution, evaluate its performance, and present your experience and results. Along the way, look for some insights that future interviewers may find unexpected or interesting—"Out-of-state students have higher graduation rates than in-state students" or "Grade Point Average is a poor predictor of career success."

Please reach out for feedback during the quarter, if you have questions as your project evolves.

Time Commitment

The project has a significant impact on your grade, and you should plan to work accordingly. There is no mid-term or final exam in this course. Plan to spend a minimum of 50 hours on your project during the quarter.

Most projects will require several iterations of: reading up on the existing literature, thinking about the problem, learning how to use relevant software libraries, writing & executing code, and preparing a presentation & report. Depending on the specific project, you will spend more time in one activity than another.

Evaluation

The deliverables will be evaluated on the ESNU scale (**E**xcellent, **S**atisfactory, **N**eeds Improvement, **U**ngradable). Criteria for Excellent and Satisfactory are listed below each deliverable.

Deliverables

There are five deliverables:

- Major Deliverables (2):

    i. Project Proposal

ii. Final Report

- Minor Deliverables (3):

  i. Mid-Quarter Presentation

  ii. Final Presentation

  iii. Peer Review

Mid-Quarter Presentation / Pitch

The mid-quarter presentation / pitch will be in front of the class on Monday, February 5th. This should be no more than 5 minutes long, describing the project objectives, work done so far, and challenges/obstacles. If you need to miss this class, please contact me to make other arrangements.
You should offer constructive feedback to other students on their proposals and presentations.
This is not meant to evaluate how much work has been done so far. It's an opportunity to talk about your project and learn what projects other students are doing.
Upload your slides as a pdf to the GitHub Classroom repo here: https://classroom.github.com/a/gL_Op5KG

Excellent:

- Provide context for the problem and why the solution could be impactful
- Describe data that will be used and has identified data sources (or roadblocks)
- Describe work completed to date
- Identify challenges/roadblocks (non-data)
- Ends at or before 5 minutes

Satisfactory:

- Provide context for the problem but does not identify why the solution could be impactful
- Complete two out of the following three items:
  - Describe data that will be used and has identified data sources (or roadblocks)
  - Describe work completed to date
  - Identify challenges/roadblocks (non-data)
- Does not end before 5 minutes, but is nearly finished

Project Proposal

Submit your proposal as a pdf on GitHub Classroom by 5:30pm, Monday, February 5th. It should be 3-4 pages long and include the following:

- Title.
- Brief description of what you want to do, including why it is useful, the data and software you will use, and the software you will write, if any.

- A **detailed description of the related work**. You should search for research papers and projects that solve the same or a similar problem. For the closest two or three such papers, you should describe what methods they used for obtaining the data, preprocessing the data, learning models, choosing metrics, and evaluating their results. You should also report their results, and what implications their work has on your project. (Learn as much as you can from such related work as it will give you an idea of what you need to teach yourself—beyond the material covered in class—for your project.)
- Brief plan of action, including any insights you have, the various steps of the project, the software libraries or packages you will be using, and the software you will be developing on your own.
- Brief description of how you will evaluate your work. There may be existing techniques or results you could compare to, or you could test how well your solutions performs on test data, or how well it models the available data.

Upload your proposal to the Github Classroom repo here: https://classroom.github.com/a/0FPwlCn2

Excellent:

- Provide context for the problem and why the solution could be impactful
- Describe data that will be used and has identified data sources (or roadblocks)
- Describe software that you will use and write (packages, work beyond off-the-shelf libraries)
- Description of 2-3 papers and projects that solve a similar problem, including:
  - Data used
  - Data preprocessing
  - Models considered and used
  - Metrics for performance evaluation
  - Evaluation
  - Results
  - Implications for your project (approaches/techniques, anticipated challenges, benchmarks)
- Plan of action
  - Current state of your project
  - Remaining steps to complete it
  - Anticipated roadblocks
- Explain the metrics and benchmarks you will use to evaluate your model

Satisfactory:

- Provide context for the problem but may not identify why the solution could be impactful
- Describe data that will be used but may not have identified data sources (or roadblocks)
- Description of at least 1 paper or project that solves a similar problem, with at least six of the following:

  - Data used
  - Data preprocessing
  - Models considered and used
  - Metrics for performance evaluation
  - Evaluation
  - Results
  - Implications for your project (approaches/techniques, anticipated challenges, benchmarks)

- Plan of action
  - Current state of your project
  - Remaining steps to complete it
- Explain the metrics you will use to evaluate your model

Final Report

Submit your final report as a pdf on Github Classroom by 5:30pm, Monday, March 4th. The report should resemble a professional paper and include—

- An executive summary describing the work you have done.
- An introduction describing the problem and its significance.
- A detailed description of the related work.
- A detailed description of your solution and the work you have done.
- Include both what worked and what didn't, particularly if you have insights into the reason why.
- A detailed presentation of the results you have obtained and its analysis.
- Suggestions for future work along the same lines.
- Description of your effort, including the relative effort on different activities. What did you have to learn for the project? What skills did you already possess? Did you need to learn how to use particular libraries? Did you spend more time reading research papers, or fine tuning your parameters?
- Bibliography.

Please write your report in clear, concise English, and include clearly captioned, helpful figures where appropriate. Unclear or sloppy reports will affect your grade.
Also upload all your source code and data to your GitHub Classroom repo for the report. In your repo, include a readme file containing—

- A description of the directory structure, and a brief description of the purpose of each file and the number of lines of code in it.
- A list of files you are using that contain code not written by you, or not written as part of this project (maybe it was part of another project).

- A description of any data files you use for your work.

- Any other relevant information about your code.

Your code should be documented such that one can get a reasonable idea of how it works. For example, each function should have a document string in the recommended style of the programming language.
Upload your report (as a pdf), code, and data (or links to data) to the Github Classroom repo here:
https://classroom.github.com/a/HzVeCAyc

Excellent:

- Clear, concise Executive Summary that draws the reader in and conveys the gist of what your project accomplished and why it matters

- Clear, concise introduction of the problem and its significance

- Description of 2-3 papers and projects that solve a similar problem, expanding what you included in the proposal. This should include:

  - Data used
  - Data preprocessing
  - Models considered and used
  - Metrics for performance evaluation
  - Evaluation
  - Results
  - Implications for your project (approaches/techniques, what you learned and used, benchmarks)

- Description of solution:

  - The data you used, including sources
  - Software that you used and wrote (packages, work beyond off-the-shelf libraries)
  - Description and analysis of results, including summary metrics and comparison against benchmarks

- Description of the data you would have liked to have in an ideal world, but couldn't get

- Description of what you would do in the next 1-3 months to continue your project if you had more time; what questions come next?

- Description of your effort, including relative effort on different activities

- Description of what you tried that did not work, what trade-offs were made, and how was the final solution chosen

- Description of what you found interesting or challenging about the project

- Clear and complete bibliography

Satisfactory:

- Executive Summary is clear, but may not be crisp or concise
- Introduction describes problem but may not establish context
- Description of related work may not meaningfully expand what was included in the proposal
- Description of your solution covers at least 2 of the 3 items below:
  - The data you used, including sources
  - Software that you used and wrote (packages, work beyond off-the-shelf libraries)
  - Description and analysis of results, including summary metrics and comparison against benchmarks
- Descriptions of at least 4 of the 5 items below:
  - The data you would have liked to have in an ideal world, but couldn't get
  - What you would do in the next 1-3 months to continue your project if you had more time; what questions come next?
  - Your effort, including relative effort on different activities
  - Approaches and methods that did not work, trade-offs made, and choice of final solution
  - What was interesting or challenging about the project
- Bibliography is sufficient to locate referenced material

Final Presentation

On Monday, March 4th, 5:30pm, during our Final Exam session, you will deliver a 10 minute presentation to the rest of the class describing the motivation/usefulness of your project, data used, your approach, unexpected challenges and modifications, and any results or conclusions you have reached.
Upload your slides (as a pdf) to the GitHub Classroom repo here: https://classroom.github.com/a/jfb5sde-

Excellent:

- Provide context for the problem and why the solution could be impactful
- Describe data that was be used and the data sources
- Describe your approach to the problem
- Describe work completed
- Identify challenges/roadblocks (non-data) and how you handled them
- Share at least one meaningful result
- Share at least one meaningful conclusion
- Less than 10 minutes

Satisfactory:

- Provide context for the problem but does not identify why the solution could be impactful
- Describe data that was be used but may leave out the sources
- Complete two out of the following three items:
  - Describe your approach to the problem
  - Describe work completed
  - Identify challenges/roadblocks (non-data) and how you handled them
  - Identify at least one result or conclusion (maybe not both, maybe not meaningful)
- Does not end before 10 minutes, but is nearly finished

Peer Review

Submit your peer reviews on Github Classroom by 11:59 pm, Thursday, March 7th. You will have the reports and presentations of three other students shared with you Monday, March 4th, by 11:59pm. For each report & presentation, prepare a review (about one page) of the project. This should not be a summary of their project. This should be your views on what was done well or poorly, whether or not the conclusions are accurate or meaningful, possible improvements to the existing project, and possible extensions of the project for future consideration. This will not affect the grade of the author of the report and presentation, but will impact the grade of the reviewer. I am looking for constructive, insightful feedback, that can be returned to the author.
Upload your peer reviews (as pdfs) to the GitHub Classroom repo here: https://classroom.github.com/a/mF0j5EOg

Excellent: For each project shared with you, complete the following items. This should be of high enough quality to share with the project author.

- Describe at least one thing that was done well
- Describe at least one thing that you think could have been done better (technically)
- Express and briefly defend an opinion on whether the conclusions are meaningful
- Provide at least one improvement that could be made to the project (conceptually)
- Provide at least one possible future extension of this project

Satisfactory: For the projects shared with you, complete 4 of the following 5 items. This should be of high enough quality to be useful, but might not be at the standard to share with the project author.

- Describe at least one thing that was done well
- Describe at least one thing that you think could have been done better (technically)
- Express and briefly defend an opinion on whether the conclusions are meaningful
- Provide at least one improvement that could be made to the project (conceptually)
- Provide at least one possible future extension of this project