📊 COGS 109: Group Project 📊

⚠️This is a living document! I will be adding details / examples / FAQs as the course unfolds.

Important Links & Key Dates

1. Group assignments are here!
    a. Please reach out to your group members ASAP and establish a method of communication. If you don't hear back within ~1 week, or a member has dropped the course, please let me know.
2. Please submit your project checkpoints here:
    a. Checkpoint 1 and self-grading rubric due Mon, Oct 20 @ 11:59pm SEE GUIDELINES
    b. Checkpoint 2 and self-grading rubric due Mon, Nov 3 @ 11:59pm SEE GUIDELINES
    c. Checkpoint 3 and self-grading rubric due Mon, Nov 24 @ 11:59pm SEE GUIDELINES
3. Wed, Dec 3: We will hold an in-class project expo, where you will evaluate other groups' projects and have your own project evaluated by your peers. Please submit your evaluations for two groups (each member of your group must submit their own individual feedback) by @ 11:59pm.
4. Wed-Fri, Dec 3-5: Final group project presentations (book your presentation date). The booking page will open Friday, Nov 14.
5. Wed, Dec 10: Please submit your final group presentation slides by @ 11:59pm.

Project goals

The final project is a chance to address a real-world question using data analysis and modeling techniques that you've learned from class.

Projects are both collaborative and individual:

- Each group (2–3 students) will work together on the same dataset.
- Every student must:
    - contribute substantially by being each responsible for 3 model variants within one model type,
    - be able to explain the entire project (including a general overview of their teammates' analyses) during the final group presentation.

To help you scaffold progress and ensure steady progress on the project, we will have 3 project checkpoints so you can get feedback from the teaching team.

The final project will culminate in:

1. a class-wide project expo on Dec 3, as well as
2. a final group presentation (Dec 3-5) to Prof. Lai + one member of the teaching team.

Grading

The group project is worth **16pts** of your final grade and contains the following elements:

Project component	Points	Individual or group grade?
Project checkpoints	6pts = 2pts each * 3 checkpoints	Individual
Project expo - presentation	1pts = 0.5pt each * 2 presentations	Group (unless absent)
Project expo - evaluation	1pts = 0.5pt each * 2 evaluations	Individual
Final group presentation	8pts	Group + individual components
Total	16pts	

The details of each element are under each corresponding section below.

Group Logistics

The project will be completed in **groups of 2–3**. You may choose your own group (provided that you are enrolled in the same **discussion section**) or be assigned to a group by the teaching team.

If you haven't already, please complete the pre-course survey to indicate group preferences and preferred group members.

- *If you choose your own group*, you are responsible for handling any interpersonal issues.
- *If you are assigned to a group*, the teaching team can help mediate if problems arise.
- Grouping will be based on section enrollment, programming experience, and preferred programming language.
- All group members should be familiar with and able to explain each other's analyses.
- Workload will scale with group size, since each person will be responsible for analyzing three variants of one model type.

Project Checkpoints

For all project checkpoints:

1. Report format: 5-10-min recorded Zoom video with slides, designed to scaffold progress toward your final presentation.
2. Grading: Project checkpoints are *self-graded* using a Google Form. In the form, you will:
    a. Describe your contribution to the project's progress so far.
    b. Assign yourself a grade (0–2 points) based on a rubric.
3. Because contributions may differ, checkpoint grades can vary across group members.

Checkpoint 1 and self-grading rubric due Mon, Oct 20 @ 11:59pm (Week 4)

1. Identify a problem or question
    a. Choose a **real-world question** (hypothesis- or data-driven) that can be addressed with data.

    b. Explain **why it is interesting or important** (social relevance, scientific curiosity, practical application, etc.).
2. Identify or create a relevant dataset
    a. Describe the dataset you will use. If possible, explain **how the data were collected**.
    b. Provide details about the key features of the data, including:
        i. Number of observations (*n*)
        ii. Number and type of features (predictors)
        iii. Possible sources of **noise or bias** in the dataset
3. Choose your modeling approaches
    a. Select a relevant method from class
    b. **Each student** in the group must choose **one (1) model type** (e.g., Linear Regression, LDA, KNN, Decision Trees, Random Forest, etc.) to be responsible for. Please clearly indicate who is in charge of what type of model.
        i. No need to specify the 3 variants at this point.
        ii. Ideally, each person should work on a different model type. However, if two individuals are particularly interested in the same model type, they may do so, provided that their three variants are sufficiently distinct from each other.
    c. Justify why the method(s) is appropriate for your question.
4. Describe any difficulties you ran into
5. Please fill out the self-grading rubric (Google Form) and submit your group's update video (Canvas)!

Checkpoint 2 and self-grading rubric due Mon, Nov 3 @ 11:59pm (Week 6)

These guidelines are **finalized** as of: Thursday, Oct 23.

To keep the video update portion succinct (~10 minutes), each group member should create and present the following slides:

1. (Optional) **Briefly describe** any updates or changes since the last checkpoint (1 slide)
    a. Have you adjusted your model type(s) or variants?
    b. If so, explain what changed and why.
2. Identify model variants (1 slide)
    a. **Restate your modeling goal.** What is the **outcome** you are predicting? Is it continuous (quantitative) or categorical (qualitative)? (Your model choice should be appropriate for the outcome).
    b. For each group member, propose **three (3)** variants of the model you would like to explore.
        i. **Example:** A student chooses to investigate their data with variants of linear discriminant analysis (LDA).
            1. Variant 1: Change number of predictors → see how accuracy and separation change.

        2. Variant 2: Standardize data → see how scaling affects the discriminant boundaries.
        3. Variant 3: Compare LDA vs. QDA → see impact of linear vs. quadratic boundaries.
    c. Your project mentor will be giving feedback on this, so don't worry if you're not sure what counts as a variant (we also encourage you to approach us early with questions or if you need advice!)
3. Exploratory data analysis (EDA) (1 slide)
    a. (Optional) Briefly summarize any interesting observations from EDA
    b. Describe how you split the model into train / test set
4. Reporting results (1-2 slides)
    a. Each student should:
        i. Fit and evaluate at least one (1) model variant per person
        ii. In your presentation, describe the model variant you chose to fit. Include any relevant equations without cluttering the slide.
        iii. No need to do cross-validation yet (we will do that when more model variants are fit)
        iv. Use appropriate model assessment metrics to evaluate performance, for example:
            1. All model types: training and test error, Python statsmodels output tables
            2. Classification models: confusion matrices, ROC curves, bar plots of model accuracy and naive baseline accuracy
            3. Regression models: residual plots, report R^2, Adjusted R^2, RMSE
        v. Include at least 2 figures or tables per person (examples above)
        vi. Summarize main conclusions from the model. Interpret parameters if fitting a parametric model (e.g., log-odds and odds ratio interpretation for logistic regression).
5. (Optional) Describe any difficulties you ran into (1 slide)

- Please fill out the self-grading rubric and submit your group's update video! (Both Google Forms).

Checkpoint 3 and self-grading rubric due Mon, Nov 24 @ 11:59p (Week 9)

These guidelines are NOT yet finalized. Please check back for updates.

To keep the video update portion succinct (~10 minutes), each group member should create and present the following slides:

1. (Optional) Briefly describe any updates or changes since the last checkpoint (1 slide)
    a. Have you adjusted your model type(s) or variants?
    b. If so, explain what changed and why.
2. Reporting results (2-3 slides)
    a. Each student should:

i. Fit and evaluate at least two (2) model variants per person (including the one from Checkpoint 1).
ii. In your presentation, describe the model variants you chose to fit (both of them). Include any relevant equations without cluttering the slide.
iii. Perform cross-validation on each model variant and report the one with the lower CV error.
iv. Use appropriate model assessment and selection metrics to evaluate performance, for example:
    1. All model types: training and cross-validation error vs. model variants (lines on same plot), Python statsmodels output tables
    2. Classification models: confusion matrices, ROC curves, bar plots of model accuracy and naive baseline accuracy
    3. Regression models: residual plots, report R^2, Adjusted R^2, RMSE
v. Include at least 2 NEW figures or tables per person (examples above). Figures should be different from those in Checkpoint 2.
vi. Summarize main conclusions from the model variants. Interpret parameters if fitting a parametric model (e.g., log-odds and odds ratio interpretation for logistic regression).
vii. Model comparison
    1. Compare your two model variants using appropriate model comparison metrics.
    2. Summarize which variant performs best and why.

3. (Optional) Describe any difficulties you ran into (1 slide)

- Please fill out the self-grading rubric (Google Form) and submit your group's update video (Canvas)!

In-Class Project Expo (Wed, Dec 3 @ 9-9:50am)

On Wednesday, December 3, we will hold a project expo in class where you will have the chance to share your work with peers, practice presenting, and give/receive feedback.

Expo presentation

- Each group gives a 7-minute presentation followed by 3 minutes of Q&A (a shortened version of your final group presentation).
- Each group will be evaluated by two peer groups.
- Grading: All group members receive the same score for the expo presentation (unless someone is absent). Everyone in the group is responsible for understanding the whole project—communicate with your group so that you can explain any part if asked!

Evaluating other groups' projects

- Each group will also evaluate two other groups' presentations during the expo.

- Grading: Evaluations are completion grades. Each student submits their own evaluations, and scores may differ across group members depending on whether the evaluations are completed.

Final Group Presentation (Dec 3-5)

Your group project will culminate in a final presentation to Prof. Lai + one member of the teaching team.

Final Presentation (8 points total)

- Live 10-min oral presentation with content slides, followed by 5-min of Q&A time where Prof. Lai + one member of the teaching team will ask you questions about your project. Be prepared to answer any questions about project motivation, theory and assumptions behind your chosen models, and your key results and interpretations.
- Grading: Based on clarity, quality, and content of the presentation + Q&A performance (see rubric below). All group members receive the same score (unless someone is absent). As with the expo, each student is expected to understand the entire project.

You are always welcome to ask for feedback on your project and presentation at any time before it's due! Reminder that this is a great way to utilize discussion section and office hours 🙂

The final presentation will be worth 8 out of the final project's 16 points.

Criteria	Points	Description
Clear analysis questions	0.5	★ 0.5: Questions and/or hypotheses were clear, well-motivated, and relevant to the dataset. ★ 0: Unclear or poorly defined questions.
Well-motivated methodological approach	1	★ 1: Methods were appropriate for the question, with a clear rationale for why they were chosen. ★ 0: Methods were unclear, inappropriate, or unjustified.
Model analyses and interpretation of results or findings	2	★ 2: Clear, accurate presentation of results with interpretation tied back to research question(s). ★ 1: Results presented but interpretation incomplete or somewhat unclear. ★ 0: Results insufficient or unrelated to the research question.
Limitations and future directions	0.5	★ 0.5: Thoughtful discussion of project limitations and/or possible extensions, new questions, or next steps. ★ 0: No discussion of limitations or future directions.
Clear delivery and engagement	1	★ 0.5: Presentation was well-prepared, clear, and engaging (good pacing, organized, not read word-for-word). ★ 0: Disorganized, unclear, rushed, or disengaging.
Presentation visual aids	1	★ 0.5: Slides were clear, uncluttered, and effectively used clearly-labeled figures/visuals to support points.

		★ 0: Slides were cluttered, hard to read, or visuals were unclear or did not support presentation.
Q&A	2	★ 2: Group answered questions accurately and confidently, demonstrating understanding of all aspects of the project. ★ 1: Able to answer some questions, but with gaps in clarity or depth. ★ 0: Unable to answer questions or showed lack of understanding.

Model types

Even though some of you may already have experience with a wide range of models, please choose a model type from the list below. If we happen to run out of time and do not cover a specific model in detail, you are still welcome to investigate it for your final project.

If you're struggling to come up with three variants of a model to test, AI tools are a helpful source of inspiration. For example, you might ask: "*What are examples of model variants for KNN?*"

A list of model types to choose from:

Supervised Learning (labeled data)

📉Regression (continuous outcomes)

- Linear family
  - Simple Linear Regression
  - Multiple Linear Regression
  - Polynomial Regression
  - Splines
  - Generalized Additive Models (GAMs)
- Non-linear / flexible regressors
  - Decision Trees
  - Random Forests
  - K-Nearest Neighbors (KNN, regression version)
  - Simple Neural Networks (for regression)
  - Convolutional Neural Networks (CNNs, regression version)

📉Classification (categorical outcomes)

- Linear decision boundaries
  - Logistic Regression
- Linear decision boundaries
  - Linear Discriminant Analysis (LDA)
- Quadratic / more flexible boundaries
  - Quadratic Discriminant Analysis (QDA)
- Probabilistic generative classifiers

    - Naive Bayes
- Non-linear / flexible classifiers
    - Decision Trees
    - Random Forests
    - K-Nearest Neighbors (classification version)
    - Simple Neural Networks (for classification)
    - Convolutional Neural Networks (classification version)

Unsupervised Learning (unlabeled data)

📈Dimensionality Reduction

- Principal Components Analysis (PCA)
- Factor Analysis

📈Clustering

- K-Means
- Hierarchical Clustering

Datasets

Here are some suggested places to find datasets, compiled by our wonderful TA, Jiesen.

General

https://www.kaggle.com
https://github.com/awesomedata/awesome-public-datasets
https://archive.ics.uci.edu/datasets/
https://huggingface.co/datasets

Neuroscience

https://crcns.org/data-sets
https://neurodata.io/ocp/
https://nimh-dsst.github.io/OpenCogData/
Allen Institute: https://alleninstitute.org/bigneuron/data/
EEG dataset: https://github.com/meagmohit/EEG-Datasets
EMG dataset: https://archive.ics.uci.edu/ml/datasets/EMG+data+for+gestures

Government/economics/politics

https://data.gov/
https://data.imf.org/en/Datasets
https://data.worldbank.org/
https://www.oecd.org/en/data/insights/data-explainers/2024/09/api.html

MISC

Kitti Computer Vision: http://www.cvlibs.net/datasets/kitti/
Coco Images dataset: http://cocodataset.org/#home
Smart City Dataset: https://smartcities.data.gov.in/