DS 2500 - Spring 2026
Team Project Specifications/Requirements

This is an end-to-end data science project on meaningful data, using the concepts and tools we have covered in DS2500 this semester.

Total Points: 2,600 pts

Important dates

You can find each milestone description under each assignment on Canvas and in the documents linked below.

Milestone	Date	Notes
Kickoff lecture (in-class)	2/6 lecture	Instructors introduce the project specifications and requirements during the lecture, you can begin thinking about ideas and teaming.
Milestone 1: Team formation and topic report	**2/20 11:59pm ET**	Submit your team formation and topic report PDF on Canvas. A template can be found in the assignment for this on Canvas. For random team assignment, fill this Google Form by February 11 11:59pm ET.
Milestone 2: Proposal	**3/10 11:59pm ET**	Submit your proposal PDF on Canvas.
Milestone 3: Progress report	**3/31 11:59pm ET**	Submit your progress report PDF on Canvas
Milestone 4: Presentation (in-class)	**4/13 11:59pm ET** (slide submission) 4/14 and 4/17 (presentations)	Submit your presentation slides (PDF) as a team on Canvas. You will present your project in class on week 15. All members must be present in class and participate in the 5-minute presentation.
Milestone 5: Final report and deliverables (written report, code, data, presentation materials)	**4/21 11:59pm ET**	Submit a final report of your project (PDF), code, and data as a team on Canvas. More details can be found in the document on final deliverables.
Peer evaluations	**4/22 11:59pm ET**	Submit peer evaluations on your own on Canvas.
Individual reflection (optional; extra credit)	**4/24 11:59pm ET** (optional)	Submit individual reflection on your own on Canvas.

Instructions and templates

The following links take you to documents with detailed instructions and templates for each milestone. Make sure to read these carefully and early on!

- Milestone 1: Team formation & topic report
- Milestone 2: Proposal
- Milestone 3: Progress report
- Milestone 4: Presentation
- Milestone 5: Final report and deliverables

Project teams

Team size: You'll work in a team of 2-4.

Team collaboration

- We expect a substantive project that incorporates the contributions of all teammates.
- We strongly suggest that each member of the team takes on a distinct hypothesis-based analysis task within the project. For instance, you might collaborate on data acquisition and data cleaning, and then define at least one interesting question per person to investigate using some combination of data analysis, statistics or machine learning. This way, each team member will be responsible for some dedicated part of the data analysis and the corresponding summary and visualization.
- Your team members are all responsible for arranging your collaboration and meetings.
- All team members must be present and participate in the project presentation.

Finding team members

- You are free to find team members on your own. Please note that all team members must be from the same lecture section.
- If you would like to be *randomly assigned to a team*, please fill this Google Form by **Wednesday, February 11 11:59pm ET**. If you fill the Google Form, we expect that you will not join a team outside of the one that we assign you. We will send your team assignments via email by next evening (i.e., February 12) to give you plenty of time to work on the first milestone.

Project requirements

- Project scope
  - We expect a substantive project that incorporates the contributions of all teammates.
  - As a rule of thumb, you should expect this project to be about twice as large/complex as your DS2001 project / Fundies project.
- Communication
  - We encourage you to discuss project ideas with instructors and TAs in lab or office hours.

    - Make sure to be in close contact with your teammates over the course of the semester.
- Deadlines
    - We've built in interim milestones throughout the semester to help your team stay on schedule and get feedback on your progress. Make sure you use these feedback opportunities as they're designed to help you succeed.
    - Please make sure to submit your materials for each milestone on time. **You may not submit your project late. No late submissions will be accepted for any of the milestones.**
    - The deadlines listed above are the last and final project deadlines.
- Dataset
    - You must identify a real dataset(s) that is both substantial, and ideally, of interest to you and your teammates.
    - We highly encourage looking for data from sources other than Kaggle. Kaggle datasets often have questionable provenance, may include synthetic data, and are typically already pre-processed for a specific challenge.
    - We encourage that you obtain your own raw data through some combination of APIs, web scraping, by combining data from multiple repositories, etc. - the more innovative the data acquisition is, the more likely it is that your project will be unique and answer interesting questions that others haven't already done.
    - It is important to ensure that your data collection meets ethical standards and doesn't violate any rules set out by the data provider.
    - We have provided a recommended list of resources below for finding good datasets and ideas for this project.
- Final report
    - You must consider, and include in your final report, the ethical considerations involved in how that data was collected and how bias may be reflected.
    - You'll have to submit a final report with your code, data (if possible), and presentation materials as project artifacts or deliverables.

Project resources and ideas

Data source examples

- TidyTuesday datasets
- https://archive.ics.uci.edu/
- https://data.cdc.gov/
- https://data.census.gov/
- https://github.com/fivethirtyeight/data
- https://data.worldbank.org/
- https://data.gov/
- Search on Github for data

Project topic examples

The examples below are just meant to help you with ideas, however, you should strive to develop a topic area that is likely to be unique.

- Stock market trend prediction

- House price prediction
- Medical diagnosis prediction
- Wildfire cause prediction
- Movie recommendation system
- Game outcome prediction
- Sport outcome prediction
- Public transit usage trend analysis

Past DS 2500 project summaries

- **Wildfires in California**: This project explores the causes of wildfires in California using data from 1921 to 2023, focusing on how they relate to factors like duration, fire intensity, and frequency. After cleaning the dataset and training the model with relevant features, as well as additional ones derived from existing data, such as fire intensity and area-to-perimeter ratio, we applied K-Nearest Neighbors and Random Forest classifiers to predict wildfire causes. Accuracy improved significantly from 20 to 30 percent to over 63 percent after grouping causes into three broad categories: Human Activity, Industrial, and Lightning. Our analysis showed that wildfire frequency has increased over time and that human-driven causes have become more dominant. This project shows how data science can be used in real-word scenarios like wildfire management by helping us make more informed decisions to prevent future fires and protect the environment.

- **Eviction rate analysis:** Our project investigates the complex factors contributing to eviction rates across the United States by analyzing demographic and economic data. Using data from The Eviction Lab and the U.S. Counties Database from 2015, we explored the relationships between eviction rates and race, income, poverty levels, rent, and property values. Our analysis found a correlation between race and eviction filings, particularly a higher eviction rate in areas with a greater percentage of Black renters, highlighting systemic disparities. Other factors like median income, poverty rate, and average rent showed inconsistent relationships with eviction rates. Our work and findings suggest that structural inequality may play a large role in eviction trends rather than financial indicators, and emphasizes the importance of complex approaches to addressing housing instability.

- **Diabetes analysis**: Our project examines the links between instances of diabetes and other health/demographic factors within Iraqi patients. The literature on diabetes prevalence and its causes in Iraq is comparatively lacking, so we chose to analyze a dataset from Mendeley Data which provides the medical data and laboratory results of 1000 Iraqi patients across 3 diabetes disease classes (diabetic, non-diabetic, pre-diabetic). Our primary method of analysis was a K-Nearest Neighbors Classifier; this yielded results which suggest that individuals with lower HbA1c levels and younger ages are more commonly associated with the non-diabetic class. Also analyzed using a logistic regression model, but concluded that there is significant room for improvement in our model, particularly when it comes to identifying positive cases. Drawing from the conclusions of our project, we could do further analysis into the levels of correlation between different factors and the instances of diabetes, and then go a step further to analyze the correlation between those different factors.

- **NFL and weather**: Investigating the factors that most-heavily influence National Football League results is crucial for coaches to develop effective training programs. With fixtures

across the country during multiple meteorological seasons, identifying the extent to which weather conditions are predictors of game outcomes is important. We hypothesized that weather would not have a major impact on results, but predicted that more extreme conditions may grant additional home field advantage to most-accustomed players. We utilized a K-Nearest Neighbors classifier to build a prediction model, before adapting the model through a multiple linear regression approach. We also performed linear regression and feature scaling normalization to effectively display the relationship between features (namely temperature and wind speeds) and game outcomes. From the initial model, our results implied temperature alone was a poor predictor of match scores, but accuracy and F1 scores of 0.58 and 0.25, respectively, qualify the model's efficacy. Our data indicated a marginal negative correlation between temperature and point differential, suggesting closer games occurred in warmer conditions. Although our model's predictive power was limited, learning that weather conditions alone are not strong predictors of game results is valuable information. It is also worth noting there are a multitude of factors that impact NFL games, so future work ought to continue the study and identify the most influential factors to support the development of effective practice techniques.

Evaluation

Your entire group will receive a grade based on your team & topic submission, proposal, project deliverables, and presentation. Your individual grade may be adjusted up or down based on your group members' individual reflections (evaluations) and feedback from your group member(s) on your contribution level.

Group grade

Factor	Weight	Points	Description
Team formation & topic	10%	260	Team formation & topic is graded for complete submission on time.
Proposal	20%	520	● The proposal must satisfy requirements specified in the proposal document. ● The proposal must be submitted on time.
Progress report	10%	260	● Progress report must satisfy requirements specified in the progress report document. ● The progress report must be submitted on time.
Presentation	20%	520	● Presentation must satisfy the requirements ● All team members must contribute in presenting the materials
Final report (with code & data)	40%	1,040	● Scope and substance of the project is appropriate for DS2500 and for the size of the team

			● Project must satisfy data/documentation/function/ visualization/submission requirements ● A written report must cover all sections enumerated in this document. ● All sources and datasets must be cited. Visualizations included in the project will be evaluated along the same guidelines as the homeworks: clear, easy to follow, and make good use of labels, legends, titles, sizes, and colors
TOTAL	100%	2,600	

Individual grade

- Each team member will submit their own individual reflection on Canvas at the end of the semester.
- Your grade will be the same as your group's overall grade with a possible adjustment based on the feedback from teammates.
- The peer feedback itself will not be graded, but failure to submit the feedback for ALL teammates will have a negative impact on your grade.
- ***You can earn extra credit by reading the peer feedback and writing a short individual reflection.***

Reporting non-contributing members

If you encounter issues with teammates not contributing equitably:

- Contact the instructors (Deahan and Ben) via email if problems arise
- Do not wait until the end-of-semester reflections to report significant issues
- Early communication allows for intervention and documentation of contribution problems
- If teammates report that a member has not contributed adequately to the project, that member's individual grade will be adjusted accordingly.