Stat 139: Final Homework Guidelines, Fall 2025

Overview:

The final homework is your opportunity to explore a topic of your academic or personal interest, using methods learned in this class. You will work in groups of 3-5 students. There will be a Google Form to help you link up with group mates and/or find a topic.

This must be an applied data analysis of your choosing. It should be on a topic that is of interest to you, so choose one that you will enjoy working on! The final deliverable will take the form of a mini-paper, which you will hand in at the end of the semester..

Milestones:

1. Milestone #1: Find a team and/or ask us to help link you with team members.
2. Milestone #2: Submit 'Project Proposal' which should include: the names of your group members, title of the project, variables you expect to measure and the data source, a brief explanation of what analyses you expect to perform, and a description of any challenges you have faced (for example, difficulties obtaining data)..
3. Milestone #3: Submit an 'EDA and Baseline Model' which should include: A description of the data and the source, some EDA/visuals of your data, and the baseline model you will use to test your main hypothesis of interest.
4. Milestone #4: You must have at least one check-in with a TF who has been assigned to you to go over your planned analyses.
5. Milestonee #5: Your final "paper"! This should follow the typical format of a scientific paper with sections: (i) Introduction and Motivation, (ii) Data and EDA, (iii) Methods, (iiv) Results, and (v) Conclusion and Discussions (note, the Data/EDA section is somewhat unique to this project). The paper should also be an a format appropriate for submission to an actual scientific journal in a field related to the project you have chosen. You may include an appendix to the paper with other relevant results from R (like extra models, graphs, and assumption checking not mentioned directly in the paper itself). Submit the electronic version of your paper (pdf) and source R code on Gradescope.

General Advice:

- Thoughtful graphics and thoughtfully summarized modeling results are often more illuminating than tables or pages of R output.
- Your grade does not depend on whether your original hypotheses are correct - you will probably learn more if your hypotheses are incorrect!

- Your research question, motivation, initial applied hypotheses (not formal statistical ones), methods, assumptions, results, limitations, and conclusions & further discussions should be presented clearly in your paper, if applicable (you may not need a separate section for each of these topics, and not all of these topics are relevant for all projects). Your paper should also include a short discussion of any challenges you faced. Be sure to cite sources, if applicable, and include a reference list.

- You should use a suitable template from a scientific journal in a related field. If you want to practice with LaTeX you can use a LaTeX template, though this is not required. The final paper should be submitted as a .pdf.

More Details:

State an applied research question and answer it by analyzing a data set, using the tools emphasized in this course. Focus on assessing assumptions, choosing appropriate tools, and evaluating the validity of your results. **The focus of this project should NOT solely be on building a best predictive model, but instead focus on the interpretations and relationships in the data set. This is not a machine learning class!**

If you are in a field other than statistics, already working on applied research, or considering future applied work, we encourage you to choose a topic that relates to your interests. However, if you use a data set that has been analyzed before by you or others (that you know of), the research question that you address for this project must be new. If the data set is associated with a published paper, for example, you might explore a different model or ask a different question. The data must be accessible to the teaching staff.

When choosing a data set, ask yourselves: Can you address your research question using this data set? Are the statistical tools you'll need within the scope of this course?

This is not a project about conducting experiments or surveys. We strongly encourage you to use data that already exists. If you do generate data yourself, keep in mind the general principles we've discussed.

Grading:

The milestones will be graded as part of the homeworks on which they are assigned. The final paper will be graded as a homework, on which all team members will receive the same grade.

Previous Project Topics:

- Predicting Congressional Vote Alignment with an Incumbent President Based on Funding and Stance on Issues
- Cancer in the US: Rates, Under-diagnosis, and Cost
- How Does a Team's Play-by-Play Performance Affect the Final Score in the NFL?
- Is Bigger Always Better?: An Exploration of Films' Budgets and IMDB Scores
- Does Post-earnings-announcement Drift Counter Public Equity Market Efficiencies?
- Mean Reversion in the Stock Market.
- Who's Masking?: Exploring Mask-Wearing Behavior by U.S. County
- Sentiment Score Trends in Selected Children's Fantasy Literature
- Homelessness in America: Putting Literature to the Test
- Optimizing the Energy Competition: Modeling Daily Energy Consumption in Harvard's Houses

Data Resources:

You may find the following data resources helpful (some links may be dead or redirected). Feel free to share other useful sources that you come across by posting them on the Ed Discussion board.

- Data is Plural `https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid=0`
- Cambridge, MA data `https://data.cambridgema.gov/browse`
- Harvard Open Data Project `https://www.hodp.org/data`
- Data.gov `https://data.gov/`
- General Social Survey - Indicators of opinions and social measures for US residents through time (collected by UChicago)
  `http://gss.norc.org/get-the-data`
- NHANES - Survey of Americans regarding Health and Nutrition along with some biomedical indicators (collected by a branch of the CDC)
  `https://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm`
- Kaggle - a repository of user-posted data sets for data science exploration and competitions
  `https://www.kaggle.com/datasets`
- Opportunity Insights, from Raj Chetty
  `https://opportunityinsights.org/data/`
- Google Dataset Search - a repository of public data sets housed by Google.
  `https://toolbox.google.com/datasetsearch`
- *Data Analysis Using Regression and Multilevel/Hierarchical Models*
  http://stat.columbia.edu/∼gelman/arm/
- StatLib at CMU - including the Data and Story Library
  http://lib.stat.cmu.edu/datasets/
- A range of data resources for academic community
  `http://datalib.edina.ac.uk/catalogue/all`
- Various Sports Data Sets:
  `http://it.stlawu.edu/~rlock/sports.html`
- Government data (from more than 70 agencies)
  `https://www.usa.gov/statistics`
- 100+ Interesting Data Sets for Statistics
  `http://rs.io/100-interesting-data-sets-for-statistics/`
- Real estate sales data in the US (it is free, although registration is required to get a full access)
  `https://www.redfin.com/`
- Data available though Harvard Library (click “Data” on the left and explore!)
  `http://library.harvard.edu/`
- Aid Data - Open Data for International Development
  `http://www.aiddata.org/content/index`
- Center for Economic Policy Research - ceprDATA
  `http://ceprdata.org/`

- Correlates of War
`http://www.correlatesofwar.org/`
- European Union - EUROSTAT
`https://ec.europa.eu/eurostat/data/database`
- FBI - Crime Statistics
`https://www.fbi.gov/services/cjis/ucr`
- Federal Reserve Economic Data (FRED)
`http://research.stlouisfed.org/fred2/`
- Gapminder
`https://www.gapminder.org/data/`
- IMF Data and Statistics
`http://www.imf.org/external/data.htm`
- Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan
`http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp`
- IQSS Dataverse Network
`https://dataverse.harvard.edu/`
- Journal of Peace Research - Replication Data
`http://www.prio.no/Journals/Journal/?x=2&content=replicationData`
- Paul Hensel's International Relations Data Site
`http://www.paulhensel.org/data.html`
- Peace Research Institute, Oslo, Norway
`http://www.prio.no/Data/`
- Polity IV - Political Regime Characteristics and Transitions
`http://www.systemicpeace.org/polity/polity4.htm`
- Princeton University - Economics Data Links
`http://library.princeton.edu/catalogs/articles.php?subjectID=109`
- Resources for Economists (RFE) - Data
`http://rfe.org/showCat.php?cat_id=2`
- UC-San Diego - Data and Statistics for Political Science
`http://ucsd.libguides.com/content.php?pid=62534&sid=567117`
- United Nations - National-by-Nation Data
`http://data.un.org/`
- World Bank Data - free and open access to data about development in countries around the globe
`http://databank.worldbank.org/data/home.aspx`
- World Health Organization (WHO) - Global Health Data
`https://www.who.int/gho/database/en/`
- World Justice Project - Rule of Law Index
`http://worldjusticeproject.org/rule-of-law-index`
- Resources on Duke University web-site
`https://stat.duke.edu/resources/datasets`
- Data surfing on the WWW, from Robin Lock
(http://it.stlawu.edu/∼rlock/datasurf.html)