Stat 139: Project Guidelines, Fall 2018

Overview:

The course project is your opportunity to explore a topic of your academic or personal interest. You will work in groups of 2-4 students. If you are looking for group mates and have a project in mind, we recommend that you post your ideas on course website under "Discussions" (make sure to post under the "Project" discussion board).

The project can take two forms: data analysis or theoretical extension. The project is supposed to be interesting for you: choose a topic you like! A final work product of the project will be a paper (5-7 pages of text), which you will hand in during exams period on **Wednesday, December 12, 11:59pm**.

Requirements and Deadlines:

1. Deadline #1: Submit a roughly 1-page project proposal, including the names of your group members, outline of the paper, explanation of what you have done so far, and what is left to do, and description of any challenges you have faced (for example, difficulties obtaining data). Submit this (pdf) to Canvas by **Friday, November 16, 11:59pm**.

2. Deadline #2: Submit an electronic version of your paper (pdf) and source R markdown file on Canvas by **Wednesday, December 12, 11:59pm**.

General Advice:

- The text of your paper (not including tables, graphics, appendices (R code, output, or mathematical derivations), etc...) should be about 5-7 pages in length if you choose single-spaced: there is no required spacing option...choose your favorite. The length may vary widely from project to project.
- Thoughtful graphics 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, hypotheses, 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.

Details on Theoretical Extension Project:

State a question that extends the theoretical material covered in this course, and explore it. Exploring the question might involve performing simulations, reading published statistics papers, writing out some calculations of your own, and/or demonstrating the principles on a small data set (such as the data sets you have used on homework problems).

When selecting a topic, ask yourself whether the question is within the scope of this course and project. We encourage you to choose a topic that you might continue to explore after this course; however, be sure that you will be able to make appropriate progress within the project time frame.

Possible theoretical projects include extended simulations to answer a question that you propose; explanation and application of Bayesian versions of the models discussed in this class (note that you cannot use the same project for this course and another course); a critical review of recent statistics papers on a topic related to this class; or a careful reading of an older, foundational paper that first proposed a method discussed in this course.

Details on Data Analysis Project:

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.

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.

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:

Each member of your group will receive the same grade on the following components:

- **40% - Statistical accuracy and appropriateness of statistical tools**. Does your project demonstrate a solid understanding of the course material? For a data project: Did you make sound statistical choices for your data analysis and justify these choices? For a theory project: Are your statistical claims accurate?

- **40% - Breadth, depth, and motivation**. Have you explained why your topic is important and/or interesting?Did you communicate your results and methods effectively? For a data project: Did you fully explore the question(s) of interest and come to a satisfying conclusion? Did you provide illustrative visuals to supplement any important modeling results? For a theory project: To what extent does your project extend the course material in a thought-provoking way? Did you cite any required references?
- **10% - Overall impression**. Does the project have a professional appearance and a polished look? Was it creative, unique, or insightful? Was it apparent that a lot of effort was put into the project? Was there an overall clear purpose and conclusions?
- **10% - Project Proposal**. Did you submit the proposal on time and include required information? Did you show you had thought about your project in this submission?

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 course website under "Discussions".

- General Social Survey - Indicators of opionions and social measures for US residents through time (collected by UChicago)
  `http://gss.norc.org/get-the-data`
- NHANES - Suvery 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`
- 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
- Interuniversity 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)