Final Project Description and Rubric
Advanced Probability and Statistics for Data Science

The final project consists of a project plan, a presentation explaining a new method incorporated into a full analysis of a question based on a dataset, an brief executive summary of the analysis, a paper explaining the analysis, and R code for the analysis. When consistent with ethical obligations, the data set should accompany the project.

The subject you investigate in your final project may be controversial. The methods you use and your data should be consistent with a responsible application of data analysis and inference.

You may start with a data set that interests you then develop a research question that requires a new method or you may start with a method that interests you then find data and a research question for which it is useful. The analysis should address a question of natural interest regarding the data. This should be your own work rather than a direct replication of an existing analysis. If you would like to do a critical replication of existing work, please discuss this with the instructor.

While several methods not covered in class may be used, the explanation should focus on one method. The explanation of this method should include the principles on which the method is based, a purpose of the method, requirements for application, and basic diagnostics, where applicable. Any other new methods should be explained in the presentation and in the paper to the point that colleagues would understand their application.

Rubric for project plan

All students are responsible for submitting a project plan. If you are working on a team, all members of that team must submit a copy of the project plan. You are not bound to this plan; your final project grade will not depend upon its alignment to the plan. If you change any parts of the plan that you've submitted, you're urged to check back in with the instructor.

Criteria	**Points**
Identify team members, as applicable	5
Method that you will explore	5
Data source or data domain to which you will apply the method	5
TOTAL	**15**

Rubric for paper (maximum of 6 pages)

The executive summary should start the paper. It should be meaningful to someone with familiarity with the subject area from which the data are drawn, but without data science expertise.

The body of the paper should be addressed to a colleague who is interested in your conclusions and also interested in applying the primary method in other contexts in a well-informed process.

Criteria	**Points**
Executive summary explains research question and conclusion	10
Data source and definitions explained	10
Appropriate exploratory data analysis performed	10
Method for addressing research question explained	10
Primary method explained in principle	10
Application of primary method explained	10
Data satisfaction of requirements of method demonstrated	10
Primary new method applied and interpreted correctly	10
Supporting visualizations provided	10
Body of paper appropriate for entry level professional	10
TOTAL	**100**

Rubric for presentation

The presentation should be approximately five minutes long with two minutes for questions.

Criteria	Points
Data source and definitions explained	10
Appropriate exploratory data analysis performed	10
Research question presented	10
Method for addressing research question explained	10
Primary method explained in principle	10
Application of primary method explained	10
Data satisfaction of requirements of method demonstrated	10
Primary new method applied and interpreted correctly	10
Supporting visualizations provided	10
Presentation style shows sufficient preparation in organization and familiarity with topics addressed	10
TOTAL	**100**

Possible methods for use in project:

- Robust Regression
- High Dimensionality methods (model building when many explanatory variables are present)
    - Regression trees
    - Principle Component Analysis
    - Partial Least Squares
    - Random forests
    - Factor analysis
- Clustering
    - Latent Class Analysis
    - DBSCAN
- Non linear methods
    - Further study of generalized linear models or general additive models
    - Spline estimation
    - Principal curves
- Bayesian Analysis (Big topic, scratch the surface)
- Sample Mean Tests for multivariate continuous responses
- Time series data
- Hierarchical Linear Models
- ROC curves
- Propensity Scores
- Methods for data with dependent error structures (geospatial and longitudinal data)
- SEM (Structural Equation Modeling)
- IRT (Item Response Theory)
- Logit models (discrete choice analysis)

Data sources:

"datasets" package in R using: require(datasets), help(package=datasets), then help for individual data set. You can use the data set directly by name. To have it in your environment, use data('data set name')

http://topcoder.com, requires sign in

http://www.bls.gov/nls/nlsy79.htm , National Bureau of Labor Statistics NLSY79 Longitudinal Survey

http://www.kaggle.com/

http://www.amstat.org/publications/jse/jse_data_archive.htm, a collection of data sets curated for statistics education by the American Statistical Association

https://cloud.google.com/bigquery/public-data/

https://www.reddit.com/r/bigquery/wiki/datasets#wiki_datasets_publicly_available_on_google_bigquery

https://research.stlouisfed.org/fred2/ Federal Reserve Data

http://www.kdnuggets.com/datasets/index.html Collected data sets for data analysis and data mining

http://community.amstat.org/stats101/home

http://wise.cgu.edu/helpful-links/data-sources/ a master list of possibilities

https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html index of data sets provided with R

http://www.nhtsa.gov/FARS FARS (Fatal Analysis Reporting System), National Highway Traffic Safety Administration : summary data and raw data for US traffic fatalities, 1975-present.

www.broad.mit.edu/cgi-bin/cancer/datasets.cgi gene expression data sets (more for data mining?)

https://toolbox.google.com/datasetsearch a search tool for data sets

https://www.ipums.org/ large collection of US and international census and survey data. Requires login and data extraction

https://dasl.datadescription.com/ collection of data sets searchable by method.

https://data.cms.gov/ healthcare data