Final Project Description and Rubric
COMP 4441 Introduction to Probability and Statistics for Data Science

The final project consists of a project plan, a presentation explaining your original, detailed analysis of a data set, and a paper explaining the analysis and R code for the analysis. When consistent with ethical obligations, the data set should accompany the project. You are permitted to work with a partner on the final project.

The subject you investigate in your final project may be controversial. In fact, projects addressing justice, equity, diversity, and inclusion are encouraged. The methods you use and your data should be treated responsibly and ethically.

The intent of the assignment is for you to assemble a full analysis, including exploratory data analysis, and application of at least one statistical method at or above the complexity of a two-sample method, a Chi-square test, regression, or analysis of variance. 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.

You should use at least one method not covered in class as part of the analysis. Explanation of the method will make up part of the presentation and paper. The explanation should include information about the purpose of the method, the basic principles on which the method is based, the application of the method, diagnostics to assess the successful application of the method, and interpretation of the results of the method. This feature of the project reflects the necessity of learning new skills as a professional in data science.

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.

All students are responsible for submitting a project plan. At most two projects should apply the same new method. If you are working in 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.

The presentations will be at the end of the term. The paper should be a maximum of five pages, double-spaced. If you are working on a team, all members of that team are responsible for submitting a copy of the paper.

Project Plan Rubric

Details	Possible 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**

Presentation Rubric

Details	Possible Points
Data source and definitions explained	10
Main features of data set presented with appropriate graphics	20

Research question presented	10
Method for addressing research question explained	20
Data satisfaction of requirements of method demonstrated	10
Method applied and interpreted correctly	20
Presentation style shows sufficient preparation in organization and familiarity with topics addressed	10
TOTAL	**100**

The following slide descriptions may help you organize your presentation:

Slide 0: Title of Project and Name of Student

Slide 1: The research question that your project is designed to answer

Slide 2: The importance or significance of your project

Slide 3: Description of the dataset used, inputs and outputs

Slide 4: Data Preparation: data cleaning and munging, removal of outliers, data transformation, feature extraction, feature selection, etc.

Slide 5: Data Visualization

Slide 6: Description of what your model does, how the algorithms used work

Slide 7: Major Data Analysis and Modeling

Slide 8-9: Model Evaluation/Model Selection/Model Comparison

Slide 10: Conclusion

Paper Rubric

Details	Possible Points
Data source and definitions explained	10
Main features of data set presented with appropriate graphics	20
Research question presented	10
Method for addressing research question explained	20
Data satisfaction of requirements of method demonstrated	10
Method applied and interpreted correctly	20
Appropriate format, including necessary citations, used	10
TOTAL	**100**

Possible Methods for Use in Project:

- Logistic regression (fitting categorical responses)
- Beta regression
- Robust regression
- Linear discriminant analysis (fitting categorical responses)

- High dimensionality methods (model building when many explanatory variables are present)
  - Lasso and ridge regression
  - Regression trees
  - Principle component analysis
  - Partial least squares
  - Random forests
  - Factor analysis
- Clustering
  - Latent class analysis
- Nonlinear 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
- Mixed models
- ROC curves
- Propensity scores
- Power analysis
- From the textbook
  - Contrasts (Chapter 11)
  - Count data (Chapter 13)
  - Proportion data (Chapter 14)
  - Binary response variable (Chapter 15)
  - Death and failure data (Chapter 16)
- McNemar's test or related methods
- Kendall's tau

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://www.bls.gov/nls/nlsy79.htm, National Bureau of Labor Statistics NLSY79 Longitudinal Survey
- http://archive.ics.uci.edu/ml/datasets.html, University of California, Irvine, Machine Learning Repository
- 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
- The JSE Data Archive
- http://community.amstat.org/stats101/home
- http://wise.cgu.edu/helpful-links/data-sources/, a master list of possibilities
- http://webserv.jcu.edu/math/faculty/TShort/Bradstreet/index.htm, drug development data sets
- 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 U.S. traffic fatalities, 1975–present. (alternative: https://cdan.nhtsa.gov/)
- 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.bjs.gov/ Bureau of Justice statistics
- https://community.amstat.org/dataexpo/home Data Challenge Expo, data and problems
- https://www.pewresearch.org/download-datasets/ Pew Research Center Surveys
- https://vincentarelbundock.github.io/Rdatasets/articles/data.html Summary of R data sets