Final Project

The focus of the final project is applying analysis techniques we have seen (or will see) in class to answer real world questions. These include ANOVA, linear regression, logistic regression, generalized linear models, principal component analysis, cluster analysis, and discriminant analysis. You do not need to use all of these techniques, but each student must do a substantial analysis with at least one of them.

There will be a group PowerPoint (or PDF) presentation, a written report and peer reviews of other groups' presentations. Each group should agree on who will submit the group assignments (the proposal and the presentation files). Each individual is responsible for turning in their own final individual report and peer reviews. The final project will count for **30% of your course grade**.

Analysis Software

You are not required to use SAS, but you must use software approved by the instructor (so that he can follow the code and check results if need be). SAS, R, and Mathematica are all fine if you can do the analysis you need to do with them. If you would like to use other software, you need to discuss that with the instructor first.

Data Sources

Data should be of general interest so it does not require a lot of additional description to understand. There should be enough data to do meaningful and interesting analyses, and the data should allow for a variety of analysis methods. When looking at data sources, you may also want to consider how much cleaning of the data will be needed to make it usable to analyze. Cleaning may be necessary but is not part of your grade.

If you need ideas for data sources, a number of data sets are available from the following websites (but **you are not limited to the data from these websites**):

http://www.library.illinois.edu/eref/formats/statistics.html
http://lib.stat.cmu.edu/datasets/
http://www.cdc.gov/datastatistics/
http://mathforum.org/workshops/sum96/data.collections/datalibrary/other.resources.html
http://www.statsci.org/datasets.html
http://archive.ics.uci.edu/ml/datasets.html
http://www.reddit.com/r/datasets

Also see the list of data sets to not use posted in compass. If you choose a data set with published results and analyses (e.g. text book, research article, etc.) you must do an analysis of the data which is different from the published analysis. Feel free to discuss data sources you are interested in with the instructor prior to the proposal if you have questions or concerns about the data.

Project Proposal

A group proposal will be due **11:59pm Tuesday March 13**. A single proposal should be submitted by the group. The proposal should be submitted via the course website.

The proposal should include:

1. The names of the group members
2. Title of the project
3. Basic description of data including data source, number of variables in the data sets, variable descriptions (if there are many variables, rough description of the type of information in the data set and description of variables of most interest is fine), and number of data points
4. Description of the questions each member intends to answer and analyses and technique(s) they intend to use

Each group member is expected to do exploratory/descriptive and diagnostic analysis for their portion of the project in addition to the predictive analysis (e.g. linear regression, logistic regression, specific type(s) of generalized linear model, ANOVA, PCA, cluster analysis, discriminant analysis, etc.). Exploratory analysis and diagnostic checking does not need to be mentioned in the proposal, as these are necessary for everyone.

Group members may either use different methods to answer related questions about the data, or answer different questions of interest for different aspects of the data. The group should use a variety of advanced methodologies. No two member can use both the same analytic method and the same question.

Group Presentation

The **group presentations** will be given during the **last two weeks of class (April 25 through May 2)** and slides are due in compass before class on the day your group presents.

The group presentation will be a 20 minute (15-18 minutes of presentation with additional time for questions) in-class PowerPoint presentation.

The presentation should include some background on the data, state the questions you wanted to answer with the analysis, give some highlights of the analyses done, and summarize conclusions (answers to the questions of interest) you were able to make based on the analyses and also provide the real-world takeaways of those results.

Peer Review

Peer review assignments are individual assignments. Students must attend all other groups' presentations and provide peer review of the presentations of other groups. There will be a peer review assignment for each day of presentations. Unavoidable conflicts must be discussed with the instructor as soon as possible. Students do not need to provide a review of their own group's presentation.

Individual Reports, Code, and Data

The **individual reports (and supporting code file)** will be due at **11:59pm on Wednesday May 2** and will be submitted via the course website.

Each student is to do their own report based on their contributions to the project. There should be a written Word or PDF report and a supporting code file(s) for the analysis done.

The written report should include:

- your name, group number, and project title on a **title page**
- an **Introduction** section stating the goals of your analysis and including a description of the data and where it came from-- this should help a reader who hasn't seen the data understand what's in the data and what you want to answer and include relevant descriptive statistics, stats and charts to help motivate the analysis that follows
- **Methods** and **Results** sections including: descriptions of analyses performed and enough information about the initial exploratory analysis, inferential analysis, relevant diagnostic checking and remedies, relevant graphics for the various analyses and comments on interesting features in the graphics that a reader can understand the general process followed, the results and the statistical significance of those results.
- a **Conclusions** section summarizing what can be inferred based on the analysis (e.g. quantitative answers to the questions you intended to address and/or some explanation of why those questions could not be conclusively answered and what further information would be needed or could be explored to answer those questions).
- **Appendices** for extra details of the analysis done, details of diagnostic issues, remedial actions, etc. The material in this section is beyond what is necessary for a reader to understand the basics of the analysis and the takeaway conclusions for the results. It should contain more extensive information about the analysis so that a statistician could follow through the steps that you did, fully understand your analytic choices, and reproduce the analysis if desired. Results in this section must be accompanied by brief explanatory text so it is clear to the reader what the results are and what they mean.

The supporting code should include the code you used to do your analysis and generate the results in your written report. The data should also be submitted if the data file is not too large.

Your final report should be considered the type of report you might submit to your boss or a client if you were a consultant (though the instructor will expect more explanation about the analyses than a client might). You should write in complete sentences, include the results necessary for explaining the analysis, and comment on any results and graphics included in the report (if you have nothing to say about a result or graphic, that result or graphic probably isn't necessary). Appendices should contain additional results (and explanation of those results) you looked at that were not necessary to a client-focused explanation.

Project Grading

- 25% group presentation
- 15% individual peer review
- 60% individual report
    - 10% for general readability and written presentation
    - 20% for initial descriptive and exploratory analysis
    - 30% for inferential analysis (e.g. ANOVA, linear regression, logistic regression, …)