1 Learning Objectives

The learning goals of the Stat 401 data analysis project are

- Formulate clear scientific research questions;
- Explore public data sources and find data that will answer specific questions;
- Recognize limitations of the available data and adapt research questions to the data at hand;
- Conduct a full data analysis with the goal of answering specific research questions;
- Summarize the data analysis process in a concise, organized, clearly written scientific report;
- Develop communication and teamwork skills which will be useful in future careers.

2 General Instructions

Your assignment is to analyze a data set of your choice to best address your scientific questions of interest. Your final analysis should be presented in the form of a 10-minute presentation and written report. The required components, deadlines, and grading for the data analysis project are as follows:

<table>
<thead>
<tr>
<th>Component</th>
<th>Due Date (by 5pm)</th>
<th>Percent of project grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set and research questions</td>
<td>Tue Oct 24</td>
<td>5%</td>
</tr>
<tr>
<td>Data analysis proposal</td>
<td>Tue Nov 7</td>
<td>10%</td>
</tr>
<tr>
<td>Draft report</td>
<td>Tue Nov 28</td>
<td>10%</td>
</tr>
<tr>
<td>Peer assessments</td>
<td>Fri Dec 1</td>
<td>5%</td>
</tr>
<tr>
<td>In-class presentations</td>
<td>Thur Dec 7</td>
<td>20%</td>
</tr>
<tr>
<td>Final report</td>
<td>Mon Dec 11</td>
<td>50%</td>
</tr>
</tbody>
</table>

All components will be turned in via D2L.

You will be working in groups of 2-3 students. Groups are assigned in such a way to provide similar interests within a group. Group members are expected to contribute to the project equally and will evaluate each other at the end of the project. Each group member should contribute to all portions of the data analysis process. All group members will receive the same grade on the project.

You may use any written or online references for this project that you wish (be sure to cite any references used!). Feel free to discuss your project with other students in the
course, but do not ask for help or provide help on the data analysis of another group. If there are questions that you have about the analysis, feel free to contact me either by email or in the office.

Detailed instructions for each component of the project are below.

1. **Data set and research questions.** You are required to select your own data according to the following rules:
   - the data set cannot be part of your current or past research;
   - the data set cannot have been analyzed by other available sources (e.g., textbook, journal article, research groups);
   - the primary response variable must be quantitative, ideally approximately normal (a secondary response variable can be categorical if the explanatory variables for that response are also categorical);
   - the data set should include at least four explanatory variables of interest;
   - the data set should include at least 50 cases ($n \geq 50$).

   My data sources webpage may be a good starting place: [link](http://www.math.montana.edu/shancock/data.html)

   After you upload your data set and research questions to D2L, I will review the data set and questions and approve (or not) the data set. You do not need to wait until the deadline; the earlier you email this portion of your project, the earlier you may start on the analysis.

2. **Data analysis proposal.** Your data analysis proposal should be 3–4 pages, double-spaced, with 12 point font and one inch margins. Sections of the proposal should include: 1. Background, 2. Data set description, 3. Scientific goals and primary questions of interest, 4. Preliminary data exploration, and 5. Analysis plan and modeling. Cite any references used. The goal of the proposal is to give me an idea of the direction of your analysis, and give me an opportunity to give you feedback on your analysis before the draft report is due.

3. **Data analysis draft report.** You will submit an anonymized draft report (see report guidelines below) to D2L. Do not include your names on the draft report. Drafts will be assessed by your peers.

4. **Peer assessments.** We will use D2L to exchange draft reports between groups. Each group will receive two draft reports to assess and provide feedback. An assessment/feedback form will be provided.

5. **Final report.** Your data analysis report may NOT exceed ten pages, including any plots and tables. Your report should be double-spaced with 12 point font and one inch margins. References and an Appendix (if needed) do not count towards the page limit. Adhering to the page limit is part of the project grade.
Your report is graded based on:

- The appropriateness and completeness of your analysis (60%)
- The interpretation of your results (25%)
- Your ability to communicate the results and limitations of your analysis (15%)

Detailed instructions for the format of the report are below.

3 Data Analysis Strategies

- Perform adequate exploratory analysis of the data and provide a complete, yet succinct, presentation of the results including both summary statistics and plots that are relevant to the research questions.
- Clearly state the model building/selection/validation criteria used to address the scientific question(s) of interest.
- Clearly state the statistical model equation before presenting model estimates.
- Perform adequate model diagnostics.
- Provide precise interpretations of the estimated parameters and/or interval estimates in the context of the scientific problem, including assessing statistical and practical significance.

4 Data Analysis Report Guidelines

Your data analysis report should describe the results of your analysis and the conclusions you would reach from those results. This report should look like a formal report to a statistically naive researcher or an interested lay person. Because a statistical analysis aims to answer a scientific question, you should organize your report in the manner which is customarily used in science:

1. **Abstract**: Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Only give the most pertinent estimates, confidence intervals, and p-values, if needed. Don’t give too much detail here, but do note any significant problems that were encountered.

2. **Background/Introduction**: Provide a description of the scientific motivation for the analysis and relevant background literature. You don’t have to go into great detail here, but do give all the facts that entered into your decision process during the analysis. List your specific questions research questions of interest as well as the questions that you were able to answer with available data. Highlight discrepancies between the two categories of questions.
3. Methods:

(a) **Source of the Data**: Describe the source and sampling methods for the data, if known. Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data as well as possible confounding by measured or unmeasured variables. This should not be a detailed presentation of descriptive statistics, however. That will come under Results.

(b) **Statistical Methods**: Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for a potential reader. Include references for non-standard techniques. You may want to describe the software used, and certainly want to describe the methods used for assessing the appropriateness of your models. 2) Explain the basic philosophy behind the analysis techniques in layman’s terms. Explain why you didn’t use more common techniques if necessary.

4. **Results**: Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the reader up to the analyses gradually.

(a) Start off with descriptive statistics. This is an area often given short shrift in previous years. The goal is to describe the basic characteristics of the sample used to address the question, as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling, try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.

(b) Then go to the major models used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, confidence intervals, p-values). Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearity, etc.) Tables can often be used to good effect here. Provide interpretations for all parameter estimates of interest. Describe the use of p-values and confidence intervals if they play an important role in your analysis.

(c) Leave exploratory analyses (if any) for last and highlight the exploratory nature of those analyses. Present the results of your analyses in tables and publishing quality figures.

**DO NOT INCLUDE OUTPUT FROM STATISTICAL PROGRAMS.** (Such means little to me and nothing to a reader). When possible, use words instead of cryptic variable names. Present confidence intervals rather than the values of Z, t, F, or \( \chi^2 \) statistics.
5. **Discussion**: Discuss the conclusions which you feel can be drawn from the analyses. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses.

6. **Appendix**: Anything of an overly technical nature should be put in an appendix *if needed.*

The major theme of the above is to write to the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be summarized in a single sentence (e.g., “We found no evidence to suggest that the final model did not fit the data adequately.”) with relevant diagnostics relegated to an appendix. You are reporting your major results and impressions of the data. If the reader wanted to see every detail, he/she would have to do the analysis himself/herself.

Your report should be well written, organized, and succinct. Givens’ and Hoeting’s “Communicating Statistical Results”, posted on the course webpage, will be helpful.