Stat 414 Project

Overview:
In this project, your group will select or collect a multilevel data set of interest to you, thoroughly analyze the data using methods from Stat 414 (or new methods that expand upon ideas from Stat 414), and present your results in both a written report and a brief presentation. The data set can come from research you have conducted, friends, professors who have collected scientific data, reputable internet sites, etc.

Groups: 1-3 people, to be formed on your own.

The dates below are tentative, let me know if you need more time/want to meet in person or over zoom to discuss your project etc. Finding a good dataset is worth the most time in this project!

Grading:		Points	Tentative Due Date
Part I:	Proposal and Data Assembly	10	Oct 22
Part II:	Exploratory Data Analysis	10	Nov. 5
Part III:	Initial Modeling Results	10	Nov. 19
Part IV:	Final Report	50	Dec. 7
	Final Presentation	20	Dec. 7
		100	

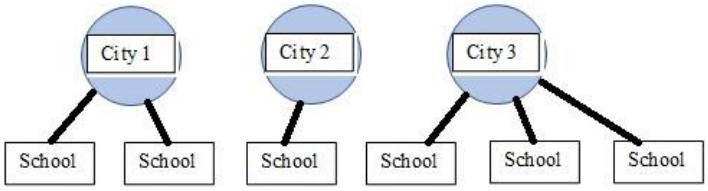
Data sources:
Be sure that your data is rich enough so that there are opportunities for model fitting choices, controlling for covariates, discovering interesting interactions, and generally providing interesting answers to real, compelling research questions (using or expanding upon methods from Stat 414). **Your data should have a multilevel structure**, meaning there are "level 1" and "level 2" units (even level 3) and that you have variables for the Level 1 and the Level 2 observations (e.g., hurricane names and age of rater, percentage black in a county and percentage black in the state, age of VRBO listing and average income in the county). (If your data has a non-normal response variable, we can work with that too.)

There is a link to possible data sources in Canvas that you may use as a launching pad if you're searching for data. You may also want to find some analyses with citations to relevant papers as a starting point, but I will ask that your project is original in terms of the analyses you perform and references you find.

Part I: Proposal

1. Identify the important **research question(s)** which will guide your project (e.g. Do youth who participate in physical exercise class have lower BMI? Does this vary by neighborhood?) – and describe why your chosen project is interesting to you. Provide rationale for each variable included in your proposed data set (e.g., "We may subgroup by gender." " We need to control for diet.")
2. Describe the data source you will be using. If available, include a link to the raw data.
3. Identify the proposed Level 1 and Level 2 units. Ideally your Level 2 units can be conceptualized as a random sample from a larger population and there are quick a few of them (so treating them as fixed effects could be a bit annoying) and you have several observations within each Level 2 unit (does not have to be balanced). Also list the expected number of units at each level
   Example: Observational units = secondary school students
   Level 1 = students (100 students from each neighborhood), Level 2 = neighborhood (30 neighborhoods)

Consider including a schematic of the multilevel structure?



4. Complete a *variable chart* (similar to the one below) for your anticipated variables. A typical list will include 6-10 variables. List the variable, whether or not the variable is quantitative or categorical, the units of measurement for each variable (if appropriate). For any variable whose definition is unclear, provide a short definition. As an example, if body mass index (BMI) were the response variable, attending a physical exercise class daily and age could be possible predictors (especially if I want to have age-adjusted effects of exercise class), the first few lines of the chart might read:

Name	Variable Role	Type	Values	Units
BMI	Response	quantitative	> 0	kg/m^2
attend PE class daily	Level 1 Predictor	categorical	yes, no	NA
age	Lever 1 Predictor	quantitative	12 to 18	Years
Public or private	Level 2 (school) predictor	Categorical	Public or Private	NA

Definitions: BMI = body mass index = weight / height squared. A measure of body fat.

Part II: Exploratory Data Analysis

1. **Exploratory data analysis**. Calculate or prepare appropriate numeric and graphical summaries for all relevant variables. Summarize the data using methods that are appropriate to the data type.

   (a) Most likely you will need to "clean" your dataset first. Make note of any problematic data and observations that need to be removed. Discuss implications about any decisions you make about missing data (e.g., narrowing of the population to which you can generalize).

   (b) Univariate descriptive statistics (e.g., five-number-summaries for continuous variables; tables of counts and proportions for categorical variables) and graphs for all relevant variables in your data set.

   (c) Explore the relationships between important *pairs* of variables both graphically and numerically. Depending on the type of your response and explanatory variables, you may consider graphs such as boxplots, scatterplots, spaghetti plots, and segmented bar charts, and you may consider summary statistics (like mean, median, standard deviation) by group, correlations, regression equations, and two-way tables with proportions. At this stage of the project the graphs can be loose with titles and labels, but for your final paper it is essential that your figures have (meaningful) captions and axis labels!

2. **EDA Report**. In no more than 3 pages, summarize the main findings of your exploratory analysis, referring to specific plots and summary statistics where necessary. Based on your exploratory analysis, are there any variables that you are now skeptical of including in final models or any that seem like they will be especially important? Has this exploration expanded any of your primary research aims?

- Begin with a short paragraph introducing your project and primary research questions. (This introduction will be expanded into several paragraphs for the final paper.)
- Use your graphical and numerical summaries to tell a story, supporting your conclusions with summary statistics. Weave numerical summaries seamlessly into your text, and refer to graphs where appropriate. Integrate your output with your discussion.
- Write well! Complete sentences, good flow, proper grammar, the works…
    - Aim your report at audience familiar with 313/324-level statistics, but may be a little rusty. Also, they have no specific knowledge on your research topic, but they have the ability to catch on quickly. Explain your terms clearly.
    - Give concise but precise statements interpreting summary statistics, etc. – in the context of your data set and research questions you pose. Avoid vague terms like "this data," "these results," etc. Also avoid cryptic variable names that you may have used in your statistical software.

Part III: Modeling Results

Based on the results of your exploratory data analysis in Part II, you will begin to fit several models, trying to ultimately settle on a single model (or models) to address your research questions. Then you will prepare an Initial Modeling Report with Annotated Appendix as described below:

(a) The **Initial Modeling Report** should be 2-3 pages, organized into the following sections:

- Model building process. Describe the steps you took, and the reasons for those steps.
- Current model. Describe your tentative final model at this point – tell which features you like, and *provide interpretations for key parameters*.
- Concerns and future plans. Describe concerns you have with your current models and additional data analysis plans you have.

Don't go overboard here, we don't have enough time to "find the best model," just want to get a sense of some reasonable models and how you are comparing them.

(b) Your **Annotated Appendix** should follow these guidelines:

- Definitions of important variables and the source of the data.
- Commented, reproducible output so that I can trace how you constructed your final data set, what the results of your exploratory data analyses were, and what plots and analyses you generated. I should be able to take your data file, go through your data organization and variable creation steps, and ultimately generate your same models.
- Include preliminary models that you considered along the way.
- Any plot or table referenced in the main body should be labeled (e.g., Figure 1). Tables and figures that are informative but were not referenced specifically in the main report. Include a short annotation – one or two sentences on what they show.

Part IV: Final Presentation and Final Report

Your Part IV score will be based on (a) the quality of your team's Final Report, (b) the quality of your team's Final Presentation, as determined by your classmates and me, and (c) your individual contribution to your team, as assessed by all group members. The audience for the Final Report and Final Presentation is anyone who has taken Stat 414 but might not be informed about your particular project.

Final Report

Your report should be a thoughtful, concise, polished, document, no longer than 8 pages. Relevant tables and/or figures should be formatted neatly into your report (because they count as part of your 8

page maximum). (If you submit as a word file, make sure image are "in line with text" so they don't move around if I add comments…) Be sure to label and reference your graphs and tables so they are interpretable on their own. An annotated appendix containing less relevant figures and tables along with important documentation and output should be attached to the end of your report (see below for more details). **Upload a copy of your final data file.**

1. **Introduction** A few paragraphs that contain background information, motivation for your research, and a statement of your research goals. Be sure to incorporate any supporting references into the text. The purpose of the background is to place your work in the greater context of the literature in the area you are investigating. Then you should explicitly identify a hypothesis that you will investigate with your analysis.

2. **Data source/Methods** Three to five paragraphs (or fewer) that…

    - Briefly describe your data, where it came from (source), definitions of important variables, and how it was collected
    - Indicate any modifications made to the data, recoding, or decisions about missing data
    - Briefly describe the methods you used (e.g., multilevel regression) in your analysis
    - Do <u>not</u> report results in this section!

    <u>Note:</u> If you are using a method not covered in Stat 414, you may choose to expand Data course/Methods a bit to describe your statistical method.

3. **Results** The meat of your report, which should include…

    - A general description of your data (completed via your exploratory data analysis)
    - A description of the null model, intraclass correlation coefficient and how relates to your context and the relevance of the multilevel structure to your data
    - A description of the results from your analyses, including interpretations of parameter estimates, tests, and confidence intervals in context.
    - Tables that summarize results and figures that illustrate results. These tables and figures should be well-labeled, numbered (e.g., Figure 1), and have a good, descriptive caption. Each report should have a minimum of two plots; rarely are residual plots part of the main body of the report unless they are an integral part of the story.
        - Especially effective graphs compare your data to the model and discuss how the model does and does not capture important features of the data. At least consider showing the "effects plots" in addition to the raw data graphs.
    - You should *interpret* tests, confidence intervals, and coefficients in this section, but you should <u>not</u> editorialize here! Save that for the Discussion.

4. **Discussion** A few paragraphs that:

    - Begin with an accurate summary statement; describe how the results help answer your research questions and what was most interesting from your analysis. In fact, the first paragraph of the Discussion is very important – in professional journals, it is often the first and sometimes the only paragraph that is read in a paper. After the first sentence highlights primary results, the remainder of the first paragraph might compare your results to others in the literature or include interesting secondary results.
    - Discuss possible implications of the results in the context of the research question.

- Make a statement regarding potential confounding variables in your study.
- Make a statement about the generalizability of your results. Don't give generic statements of possible causation and generalizability, but thoughtfully discuss relevant issues – confounding variables, representativeness of the sample, etc.
- Identify any limitations of your study. Discuss the potential impact of such limitations on the conclusions.
- Identify strengths and weaknesses of your analysis.
- Make suggestions for future research. Identify important next steps that a researcher could take to build on your work.
- Do not include test statistics or p-values in this section.

5. **Annotated Appendix**

- Tables and figures that are informative but were not referenced specifically in the main report. Include a short annotation – one or two sentences on what they show.
- Annotated output so that I can trace how you constructed your final data set, what models you ran to produce the results quoted in your report, and what intermediate models you also considered.
- Description of statistical modeling steps that were not included in the main body of your report. Possible entries here include:
  - How you handled missing data
  - Evaluation of assumptions.
  - Outlier analysis and how you decided to deal with any outliers along with rationale for your decision.
  - Describe hypotheses testing you performed during model building and how you decided on the explanatory variables you ultimately included in your final model.
  - Assessment of the final model.
- How you went from the model output in R to interpretations in your report (e.g. exponentiate coefficients, then take inverse)
- Anticipate questions someone might have after reading your report, and make sure those questions can be answered with information in the appendix.
- A citation for any reference article(s) you included in your proposal. Also include a link, if appropriate.

Final Presentation

Each group will upload 5-minute Powerpoint (or equivalent) presentation describing your findings and any new methods used. This can be done with FlipGrip or YouTube (let me know if you need more details on using either. You can make the presentation in Flipgrid or it should allow you to upload a clip, e.g., if you want to screen share from different locations in Zoom)) .

- You can use your Final Report as an outline for your talk; the Results section should comprise the biggest chunk of your presentation, and you can probably skip the Methods section (unless you're doing something unusual).
- Use plots to tell your story as much as possible. When you present a graph, be sure to orient listeners to what variables are on each axis, and what the main point of the graph is.

- Avoid big chunks of output.
- You will be assessed based on: organization, verbal presentation, ability to use statistical terminology correctly and confidently, use of graphs to tell your story, success in anticipating and answering audience questions, and ability to hold the audience's interest.

Also consider submitting your project to the Undergraduate Statistics Research Project competition (next deadline Dec. 22). I'm very happy to help with this process…(their format is a bit different…) https://www.causeweb.org/usproc/usresp