Final Project Guidelines

Guidance concerning your BIOS 669 project for Spring 2024

Instead of an in-class final examination, at the end of the semester I will have you turn in a final project of your own choosing. This project should be non-trivial and should illustrate application and mastery of skills learned in this course (as well as its prerequisites). Basically, it should involve meaningful work with data.

Since the final project for BIOS 669 is totally your choice, you should ideally find something that will be an interesting learning experience for yourself. On the other hand, if you've got a really busy semester, then you might go for expediency and choose a project that lets you piggyback on work that you are already doing anyway, such as your Master's or undergraduate honors thesis, PhD dissertation, or current project for your GRA. If you are working on a thesis this semester, it's likely that you can take a part of that and make it into your 669 project.

What you will turn in

is an organized zip file that contains a document that tells me how I should approach the zipped materials. Such materials should definitely include a brief paper, SAS code, and SAS logs (which should be clean as usual!), and the zipped materials could also include data in various forms, copies of the surveys/forms used to collect the data (or links to those forms online), codebooks that describe the data (or links to those codebooks online), RTF or PDF files, HTML files, a Python program used for web scraping, etc. When possible (and if it doesn't violate confidentiality restrictions), I would like you to turn in the data with your project, or possibly point me to where I can find the data on the web. Being able to see/use the data would be very helpful to me in evaluating your work. But I understand that this will not always be possible.

Each project is to be individually done. You can discuss your work with others, but turn in your own project.

Your grade

in the course will not be determined by your final project, but the project will definitely affect your final grade, especially if you are making a push for an H or A. The project will be worth the equivalent of five regular assignments, or about one-sixth of your course grade. I will evaluate your work based on both ambition and execution, and remember that it's the last piece of your work that I will see, so it could have a huge impact on my impression of what you have learned and what you can do. Here are some areas where you could excel:

- Choosing an ambitious but realistic project so that you can really do what you set out to do.
- Choosing and executing appropriate programming techniques for the project work.
- Writing very clear, readable, understandable code.
- Applying principles covered in the course, such as always checking derived variables.
- An excellent write-up of your work, including descriptions of your goals, the data you are using, the methods you applied, and your results and findings.
- Creating effective displays (tables and graphs) to show your discoveries.

Project Proposal

Please feel free to discuss your project idea or project work with me at any time. By **March 26, 2024**, you will email me a brief (one paragraph to one page) description of your proposed project. You are welcome to submit this earlier as well. It is also acceptable to submit a "backup" option in your proposal, especially if you submit your proposal early. Once you and I agree your proposal is realistic and appropriate for a final project, I'll provide my approval for your topic via email. If unexpected issues arise that require major changes to your approved proposal, please notify me so we can agree on the appropriate-ness of the new idea.

Due Date and Presentations

The project will be due through Canvas by **11:59 PM Tuesday, April 30th**. During our scheduled exam time (8 AM on Friday, May 10th), we will convene in our normal classroom and anyone who likes can make a brief (5-10 minute) presentation based on their project. Choosing to make such a presentation will definitely have a positive effect on your course grade.

Miscellaneous project ideas:

- Combine data that's "at different levels" – county data + state data, for example – in order to answer a certain question or questions.
- Compare SAS and SQL for working with some data.
- As I mentioned above, pick a data-intensive part of your work for some other purpose such as your Master's or undergraduate honors thesis or your PhD dissertation.
- Follow up on a newspaper or magazine article that seems to be based on an analysis of publicly available data. Can you reproduce the work?
- Write up and show me how you solved a genomics data problem.
- If there's a topic that's not covered in this iteration of BIOS 669 but you think I should cover it in the future, do something in that area to give me a head start on future inclusion.
- Do you have access to some kind of streamed sensor data? If so, explore what you can do with that data – sampling, filtering out redundancy, etc. – to make it usable. Can you write an application that provides a useful report on a regular basis?
- Do a data linking exercise, where you have data that you think belongs together but you don't have quite the right identifiers and must do some sort of a fuzzy or approximate match to combine records.
- Through an NIH website, obtain public use (limited access) data sets from an NIH-funded study and use it to answer a question of interest to you or to try to reproduce some reported analyses on that data. (Our METS data, for example, is listed as available on the National Institute of Mental Health, or NIMH, data website.) Such data is usually available for research or educational purposes, such as this course, if you fill out a form providing that information.
- Write an ambitious macro or series of macros that you have been inspired to write by the accomplished student-written macros that you've used in this course.
- Past projects that can be used as examples: past_669_projects compiled for 2024.pdfDownload past_669_projects compiled for 2024.pdf
- Ideas for data sources: possible data links for BIOS 669 2024.pdfDownload possible data links for BIOS 669 2024.pdf

A special note concerning use of data with weights: Much publicly available data seems to come with weights (since it is collected according to a complex sample survey design), and the data must be analyzed using those weights in order for the analysis to be optimal. If you elect to use such data, I do expect you to use the weights during your analysis. Usually you can find good information about how

to use the sampling/weighting variables for a particular study. The NHANES study provides especially good help online – sample code etc.

Additional thoughts

It will make me very happy if you take the opportunity to use tools from this course for your project whenever they make sense. Specifically, your project work could possibly benefit from the following:

- Data checking and cleaning (and if you find very bad problems correct them with a detailed comment in the log that explains exactly what you are doing)
- Following the twenty analysis data set guidelines (see that set of notes)
- Checking any derived variables as we did in our variable-checking exercise
- Using the exclusions macro to analyze exclusions as you make your analysis data set
- Using the comparative table macro to compare your fundamental groups, if this concept applies to your data (RPTC)
- Producing a codebook to describe your data (MTDC)

What will make me even happier is if I see that you've used this course as a jumping off point for meaningful work with data. Some examples:

- One student wrote a web scraping program in Python (more sophisticated than this course's introduction to web scraping) to gather data on professional hockey players, and then he analyzed that data with SAS.
- Another student downloaded "green taxi" data from New York City and used a New York Times API with some SAS tools to find out the boroughs in which those taxi rides began and ended. He then answered some interesting questions with his data.

I'm also VERY interested in the idea and practice of "reproducibility". I hope that some practices you've used in this class – saving your programs, using meaningful program names, putting the standard footnote on every piece of output so that you can always tell where it came from, making every program self-contained in terms of having LIBNAME statements, etc. - will help your own work be reproducible by yourself and others, if necessary. Other people are talking about reproducibility in the context of both ethical research (Can these results be believed? Does your analysis really show what you are alleging?) and scientific progress (Again, can these results be believed? Did you choose appropriate techniques for the data and did you apply them correctly?). If people who write papers make their data available, that's great. If they make both their data and their analysis programs available, that's even better - though even this practice doesn't guarantee reproducibility. I think it would a FANTASTIC learning experience for you to try to reproduce with SAS the work (tables and figures) that someone did for a journal article where they provide the data and possibly the analysis programs to support their paper. I know that more and more journals are asking authors to do this.