PSYC 417: Final Project

Instructions: The final project tests your understanding of everything we have covered in class. It is fairly open-ended: You choose what dataset to work on and what analyses you want to run. You should have a clearly defined research question from a dataset of your choosing and apply techniques we've learned this semester to draw insights from the data. **You should have at least two predictors and one outcome variable.**

Data: You have two options for your data set:

1.) If you work in a research lab, you may use data from one of your projects (so long as they can be analyzed with the techniques we've covered in class).
2.) You can select any publicly available data set. Options include*:
    - Kaggle: https://www.kaggle.com/datasets/
    - OpenNeuro: https://openneuro.org
    - OSF: https://osf.io
    - Google doc with a long list of open datasets: https://docs.google.com/spreadsheets/d/1ejOJTNTL5ApCuGTUciV0REEEAqvhI2Rd2FCoj7afops/edit#gid=0

*You can look at other databases as well. You **do not** have to analyze psychology or neuroscience data for the final project. Please make the question relevant to your interests or career goals! You cannot use one of the datasets we used in class since the code provided in answer keys would give you an unfair advantage. Do not use an online resource with code already provided. We will see the original source and want this to be your own work.

Notes: All work **has to be your own**. Do not copy a classmate (e.g., by using the same data set as a friend). See the rubric for scoring criteria. You can make the project as simple or complex as you wish, so long as you cover everything on the rubric (e.g., have two+ predictors) and only use the techniques that we learned in class. When in doubt, check with us.

Turning it in: You should turn in all necessary files:

1.) Submit all of your code via .qmd and rendered .html file and submit the dataset as a .csv. If you are using your own data and have IRB concerns, please let us know.
2.) Submit a written report (ideally a word document or .pdf) describing the data/problem and your solution, including plots. There are no specific length requirements, but 5-10 double spaced pages is a decent estimate (including figures if text is wrapped around them).
3.) YOU MAY ONLY USE TECHNIQUES YOU LEARNED IN THIS CLASS. A modest point deduction may be made for unique or technical code external to the methods we've applied in 417. (Especially ones common to AI coding tools.)

Rubric (100pts)

Data Cleaning and Preprocessing – Show that you understand the principles of tidy data.	
3pts	- Successfully load the data into R and load all required packages. - Filter/manipulate the data, and/or construct new features/variables as needed.
3pts	- Data submitted successfully with project as a .csv
3pts	- All other files submitted, including .qmd, .html, and written reports.
General coding/data science ability – Show that you understand the fundamentals of programming we covered this semester by applying them to your problem.	
6pts	Illustrate that you understand conditional statements by implementing them somewhere in the code. (Can be in the bonus.)
6pts	Illustrate that you understand loops by implementing them somewhere in the code. (Can be in the bonus.)
6pts	Illustrate that you understand how to define and use functions by implementing them somewhere in the code. (Can be in the bonus.)
10pts	Create at three different APA-formatted figures relevant to your hypotheses. (It is recommended that one of these be a conceptual diagram.)
10pts	Model the relationship (e.g., using linear or logistic regression) between some subset of predictors (two+) and an outcome.
10pts	Use either bootstrapping or machine learning to supplement your analysis. This can be in the spirit of prediction (if machine learning) or in the spirit of a sensitivity analysis (if bootstrapping). If applicable, comment on if your results are consistent.
Analysis and report – Write a brief report summarizing the results. This should have the components of an APA paper (e.g., Introduction, Method, Results, Discussion), but does not need to meet any length or strict APA formatting requirements (just make an effort).	
7pts	Introduction: Describe the problem, define variables and provide relevant background knowledge, identify your research question and hypotheses in terms of parameters.
7pts	Method: Describe the data (e.g., how it was acquired, the variables and their measurement scales), describe what techniques you're using (e.g., regression) and why you chose the analysis plan you did.
13pts	Results: Describe and report the results in APA format (e.g., t/F values, p-values, estimates, 95% confidence intervals). Interpret what your estimates mean (e.g., change in y as x changes). Present your figures and supplementary analyses here.
13pts	Discussion: Summarize what you did. What have your learned about your data and what conclusions can you draw? What are the limitations and future directions?
3pts	References: You should have an APA-formatted references page with at least 3 sources.
Bonus Points – We're going to be pickier than normal about these.	
3pts	Build a function in R that computes the correlation between two numeric variables. You may use the cov() and sd() commands to do this. Apply this to two variables in your dataset and compare the results to the cor() command in R.
2pts	Clean, concise, and heavily commented report showing thorough understanding of coding hygiene and reproducibility.