### **DATA 318 Spring 2022**

**Final Project**

Congratulations! You’ve almost made it through DATA 318. Here is the final boss:

1. Choose a target variable (categorical or quantitative) that you would like to model. This variable can have something to do with economics, politics, sports, education,… basically anything that interests you. For this final project, your models should be **predictive, not just descriptive**.
2. Find data for your variable, along with **5 or more** “candidate” predictor variables. These cannot be “canned” datasets that have already been cleaned and organized. For example, you can use data pulled directly from the U.S. Bureau of Labor Statistics website, but not the same data that has been cleaned and organized and posted on a .edu page. Kaggle is not an accepted source (though often you may be able to find the original source with some looking).
3. Clean (if necessary) and summarize (numerical and visual summaries) your data, then build predictive models (**at least 3 different types**) for your target variable. Analyze the model’s performance and summarize your findings. This should all be done in a RMarkdown file so that Dr. Tanner may re-create your analysis with little trouble.
4. Prepare a written report (4-6 page) with the content from #3. This report should tell the story of your data and the approach that you took to model your target variable.
5. A brief (one-page max) proposal is due **Thursday April 7 at 5pm.** The proposal should address: What is your target variable? Why is this interesting? What is your data source (provide citations)? What is your plan for analysis (which models are you planning on using and how will they be evaluated)?
6. You may turn in an **optional draft** (must be complete) by **Thursday April 21 at 5pm.** If you turn in a draft, I will provide you with feedback on Monday April 25.
7. You will present your project to your peers during the final exam period **Thursday, April 28 at 11am-1pm**.
8. You will submit your written report, data, and RMarkdown file by **Thursday, April 28 at 11am.**
9. You may work alone or in pairs.

### Final Project Rubric (125 points)

Scope (15 points)

* Appropriateness and depth of data for final project - 15 points

RMarkdown Workbook (35 points)

* Import, organize and clean data (if necessary) - 5 points
* Exploratory Analysis - 5 points
  + Numerical and graphical summaries of the data.
* Model Creation - 15 points
  + At least 3 appropriate model used
  + Correct use of models
* Model Assessment - 5 points
  + Use cross-validation/test set to evaluate model performance
* Holistic/Professionalism - 5 points
  + Clearly labeled and easy to follow the flow of data & computations

Report (50 points)

* Introduction - 5 points
  + What are your project goals? Why is this interesting/important?
  + What variables did you use? Why might they be important?
  + How could your model be used to make predictions?
* Data Source - 10 points
  + Where did your data come from?
  + How was it collected?
  + Is it reliable? Are there potential sources of error?
  + Proper citation (using your favorite citation convention). More than just the link. To cite an R package, use `citation(“packagename”)` for the citation details.
* Exploratory Analysis - 10 points
  + How are the predictors related to the target variable?
  + Use graphs to tell the story of the data. Don’t leave graphs alone, discuss them.
* Model and Interpreting model - 20 points
  + Discuss design decisions involved in creating your models.
  + Discuss and compare the model performances.
  + Classification - Discuss the tradeoffs between types of errors and which is more important to minimize for your context
  + Regression - Discuss the typical error in the context of the problem. How does this compare to what you would expect from a subject matter expert?
  + Discuss some predictions that your model makes for some chosen inputs (and the expected accuracy of those predictions)
  + Discuss possible enhancements to the model (that could be done)
* Holistic/Professionalism - 5 points
  + Grammar, Spelling, Layout, Readability, Clean Graphics
  + Basically does it look like something that you would give to a boss.

Presentation (25 points)

* Introduction - 10 points
  + What are your project goals? Why is this interesting/important?
  + What variables did you use? Why might they be important?
  + Where did your data come from? What is the original source?
* Tell the story of the model - 10 points
  + Discuss design decisions involved in creating your models.
  + Discuss and compare the model performances.
  + Discuss possible enhancements to the model (that could be done)
* Holistic/Professionalism - 5 points
  + Project voice, face audience, clear and informative visuals

Proposal doc: ​​<https://docs.google.com/document/d/1ZeKwi2r2nA5YE_3BCzaLNxzb-DJ1tXOktkveh3hHxKg/edit?usp=sharing>