DA 101 Final Project - Fall 2021

Description

The purpose of the final project is to combine the core skills you have gained in the class in an application of the data cycle producing a short polished report on a question of your choosing, ideally something that you're passionate about or is relevant to your life or interests (some suggestions are offered toward the end).

This final project will incorporate all of the steps involved in the Data Analysis Cycle. You will be in charge of stating an interesting question that doesn't duplicate previous projects or labs, conducting exploratory analyses using skills from the entire semester, building a model that you will interpret, then communicating your key findings in a polished, professional narrative.

Remember that the data cycle is an iterative, not linear, process. You may find that your original question must be revised to match the available data or that your data will need restructuring to match the question effectively. Further, your first prediction/explanation model might not be your final model. Consider revising the model to solidify your ability to interpret a clear finding.

Finally, the final project will be an opportunity to practice your communication with non-technical collaborators or audience, as your conclusions should be stated in plain language that could be understood by the general public.

Final Project Timeline

Date	Description
Due Friday Nov 19 by 5:00 pm	1 page proposal document (see full description below)
Due Wednesday Dec 1 by class time	Codebook and documentation prepared for your dataset in a google doc to share, printed and ready to share and discuss in class with assigned reading.
Due Sunday Dec 5 by 11:59 pm	1 page progress report
TBA Week of Dec 6-10	Team Presentation on preliminary results. *Projects may be yet unfinished, but you should have made significant progress, and have a presentation to share on your topic and what you've learned.*
Dec 13 LAST DAY OF CLASS!	Final thoughts on becoming a data analyst, final questions
Due at Final Exam Time (see below)	Turn in two things by 11:00 am on the assigned final deadline: **Individual Team/Self Assessment forms** & **A knitted html file for the written Project Report**

Choosing Your Data and Question(s)

The **question** is open for you to determine. The **datasets available** include those that you are already familiar with and several more, and represent the key areas in which you may choose to focus your "discipline" in a Data Analytics Major:

- **Anthropology and Sociology**
- **Biology**
- **Economics/Business**
- **Physics**
- **Political Science**
- **Psychology**
- **Philosophy**
- **Sports Analytics**
- **Arts & Humanities**
- **Design your own domain area!**

Question-asking

We can ask questions that are **closed** (yes or no answer; are the two things *different* or not?) or **open** (require more thought and explanation; how *much* does something change? *How* is it related to something else?). Likewise, our data analyses can serve one or more purposes as we move through the data analysis cycle: **descriptive** (describes different measures of the data), **exploratory** (looking for patterns or unknown relationships in the data); **inferential** (using a sample to tell us something about a larger population); **predictive** (use relationships in the current/past data to predict the future); or **causal** (what happens to one variable when one or more other variables change?).

Mostly, choose something that you are curious about and will allow you to demonstrate the listed concepts for the project. *Have some fun with it!*

Where can I find datasets?

- Denison Libguides: http://libguides.denison.edu/data-analytics
- Google Dataset Toolbox: https://toolbox.google.com/datasetsearch
- US Government open data: https://www.data.gov/
- Health data: https://healthData.gov/
- Lots of data: https://data.world/
- Kaggle: https://www.kaggle.com/datasets
- United States Politics: https://voteview.com/data
- Substance Abuse and Mental Health Administration: https://www.samhsa.gov/data/
- Ecological data: https://ecologicaldata.org/find-data
- Sports data: https://community.amstat.org/sis/sportsdataresources
- US Dept. of Education - https://collegescorecard.ed.gov/data/
- US Dept. of Energy - https://www.osti.gov/dataexplorer

How do I choose data and organize my project?

Your project must analyze a real-world data set, using the course concepts and materials. You should choose a topic and dataset that interests you, and come up with a main question and hypothesis that is answerable. The goal of the project includes demonstrating the breadth of techniques we

learned as part of the data analytics cycle this semester: from data cleaning, visualization, summary stats, hypothesis testing, regression, and interpretation. Do not include raw computer output in your writing: Readability is a major component of reports, so use headings, tables, figure captions, etc. to make it easy to follow.

The variables to include in your model are open for you to determine, as is the interpretation of the model. **The entire final project can be completed solo or as a team of 2 or 3, it is your choice. Be sure to include each team member's name on the final report. By including all names I will assume that each team member contributed equally and fully to the final product (everyone contributes to technical and non-technical aspects) and are deserving of the same grade.** If you choose to work by yourself, you will still be expected to produce a full report.

Progress checkpoints

Consistent work and progress on your final projects, one the topic is chosen, will ensure a strong end to the semester. The final weeks in lab will be dedicated to "project sprints" - in other words, using the time to work ahead on your projects as much as possible. You will still need to work on the project outside of class time to complete it to a high standard. Ask questions of your professor and TAs early and often!

You will receive feedback from your instructor at the project proposal and progress report stages, and these assignments should also help you focus your thoughts as your explore the dataset.The final week of class, each team will share out their progress in a peer-evaluated presentation, which you can use as a final way to fine-tune your method, message, and presentation, prior to completing the final report.

These checkpoints are *graded* and are part of your grade for the final project.

Checkpoint 1: Project Proposal

Friday Nov 19 by 5:00 pm : 1 page document stating your team, dataset (with url), and main questions, uploaded to Notebowl.

This 1-page proposal document will contain:

- *team member names*
- *dataset you will use (and hyperlink to cite source)*
- *a list of the big question(s) you will investigate*
- *5-7 sentences explaining your rationale, why is this dataset and question interesting? How will the data dictate your process for problem solving?*

Checkpoint 2: Codebook and Documentation

Due Wednesday Dec 1 by class time

Codebook and documentation explaining your dataset.

You may choose to focus on the variables that you will be including in your analysis, especially if you are using a particularly large dataset with many columns.

One copy turned in per group or solo analyst

Below is a suggested format to use to make sure the useful information is communicated to a reader or potential data user (you too!) needed to interpret and understand it. As you edit, remove any instructional or template text and replace with your own. I recommend starting a google folder where you can keep this and other files to help you (& your team) keep track of progress, and update as needed. If you have questions particular to your dataset or another idea for how to present the codebook and documentation, please discuss in advance with your professor.

Dataset Name

Data source: Name of data provider, website, and link

Original data collectors: Names of people or organization

Data size: e.g. number of KB or MB

Special permissions: Note if the data is freely available, published, or has restrictions for use.

1-2 short paragraphs describing the data table(s) you plan to use. Your description ideally should also introduce the main questions you are considering answering, and to the best of your knowledge so far, the relevant columns (variables) you will be focusing on. It should also describe what each row represents, as well as any potentially "tricky" or difficult aspects of the dataset - for example, how is missing data denoted in the dataset? If the dataset already had a well-defined codebook or metadata provided that can give additional information, please link it here.

Name_of_your_datafile.csv

Please note that if you have multiple data tables, for your project, you will need to provide more than one of these tables, one for each dataset. You should make a note about which column in each dataset contains common information, and represents the "link" or the "key" between them.

Column Name	Variable Definition	Units	Data Type	Variable Codes and definitions	Missing value codes
The name of the column, exactly as it appears in your .csv file	Describe what each variable represents, and if known, how it was measured.	if units needed; grams, days...	integer, decimal, ordinal, nominal, character	if not a continuous variable, you could list the acceptable values	if there is missing data, indicate what is used, e.g. blank, NA, NULL, etc.

Add more rows as needed...

For **Variable Definition**, If you don't know or don't have information on how to interpret a variable, or want to give a word of caution, say so here. If the variable has any special considerations or challenges inherent to its measure, you may note that here, too.

For **Variable Codes and Definitions**, If you don't need to specify acceptable values, you can just fill this in with NA (for not applicable).

Checkpoint 3: Progress Report

Sunday Dec 5 by 11:59 pm

The report should demonstrate significant coding and data exploration or analysis progress since the previous week. You should be beginning to move beyond initial data wrangling and exploration at this point.

One copy turned in per group or solo analyst

The progress report is the final "check-in" before preparing your presentation and final html report that you are required to have with your professor. I encourage you to use this progress report as a way to reflect on your work so far, and if you have been able to create visuals, summary statistics, and statistical tests that can answer the questions you posed at the beginning of your progress. Consider at which stage in the data analytics cycle you are currently at, and what you'll still need to accomplish between now and the final deadline to have a final project you can be proud of, that meets the requirements stated in the prompt below (scroll all the way down to see them).

Your progress report can be relatively short, but it should contain enough detail to demonstrate significant progress on your coding and analysis since the previous week. More detail will also help you gain useful feedback and help at this stage, setting you up better for writing your final report. Below you can find what must be included in your progress report.

- 1 paragraph on where you are currently at in your data analysis
- At least 1 visual using your dataset
- Share something that you have learned
- Share something that you have found difficult, confusing, or haven't figured out yet

Presentation of Preliminary Results

(Individual and Group grade)

Week Dec 6-10, exact dates TBA

Team Presentation on preliminary results.

Your projects do not need to be finished, but you should have made some significant progress at this point, and have some preliminary graphs and statistics to share.

Group presentations will take place **during class time on December 6-10**. Presentation order will be randomly determined using R. The presentations will be given a group grade (for the presentation as a whole) and an individual grade (each person's contributions to the presentation itself, presentation style: speaking audibly, clearly explaining concepts and the data or analysis). Each individual in the classroom will also receive a grade for participating in the presentation days by attending and asking questions to their classmates who are presenting, and by contributing to a "peer review" process, on some subset of their classmate's presentations (TBA), to help give your peers feedback they can use to finish their final report.

Presentations should be about 7 minutes each, and should describe:

- What dataset you used, and which aspects/variables in the data were of interest
- What main question or set of questions did you seek to answer?
- At least one graph and finding from your data exploration phase
- What was your approach for analysis?
- At least one graph and finding from your analysis so far (even if you think it didn't work out, and your analysis will change still, that's OK because it's part of the process!).
- Your main findings so far, or what you think your next steps will be.

You can use any style of presentation you would like, as long as you hit the main points listed above in some way. For example, you may choose to make a powerpoint presentation, or to walk your classmates through your code or html report so far that you show on the screen, you can use the whiteboard or a tablet to sketch your findings or to show some new code or function you learned, or you can print handouts for the class to use. You could use some combination of these things, or get creative (video? song?). The main thing is for your presentation to be engaging and clear for others to understand who aren't familiar with your dataset, while explaining the progress that you've made so far. *Each member of the team should contribute equally to the presentation* during class (in other words, **everyone talks**).

Final presentation ordering TBD

Final Written Report

Due Date for Section 01 (MWF 8:30am): **Thursday 16 December at 4pm**
Due Date for Section 04 (MWF 11:30am): **Monday 20 December at 4pm**

A knitted and polished html file reporting the results of your final data analysis project and, turned in individually, a team/self assessment form

Roughly, 3-5 written pages (though this is hard to measure in an html document, so consider it a *guideline*. Think about this report as a "final takeaway" of all the skills you've learned in class over the semester. Below is a rough structure of your final written report. **Here you should use code folding so this section mirrors a ready to deliver report with clear section headers and interpretations of any statistical or graphical output (like several of our previous projects), but it will also be easy for your instructor to see the code, if needed.**

Introduction

- Provide a one or two paragraph introduction, professionally written, that gives an overview of the essentials someone needs to know to make sense of the data you show. You must cite and link to your dataset. text here

Ethical Consideration

- Provide one paragraph discussing the stakeholders in your data analysis and your ethical concerns or responsibilities using the data and in your analysis. Everyone has ethical considerations, no matter what the dataset or subject matter.

Data Explanation and Exploration

- Provide some details describing the data you are working with. What are the observations? The key variables you will be looking at? Are there any particular challenges in the data you will need to work through or be aware of during analysis?
- Provide two polished visuals that describes the data in a way relevant to your question (descriptive, not related to your statistical model specifically–not a scatterplot). Write text that describes the data and what the visuals tell you about your data or decisions you will need to make for the analysis.

Statistical Analysis and Interpretation

- Provide at least two distinct statistical models (for example; multivariate regression and t-test) that you interpret correctly and fully in the text.
- Provide at least three polished visuals that specifically support and validate the model(s) you have developed (e.g., residual and regression line/scatter, histogram showing normality of data or residuals, etc.), or help to communicate your main result. Visuals should have captions and be referred to clearly in your text, and they should not all be the same (e.g., not three scatterplots).
- Text should fully explain what you show and your findings, to someone who is unfamiliar with your data, code, and models, in terms of the data and in plain language.

Conclusions

- Provide one or two paragraphs concluding about the data: what does it tell us, what are the limitations to this data/model, and what is one future direction you could envision for future data analysts or data collectors?

- Find at least one reference that is relevant to or supports your insights, and cite it in this section. You may cite a reference by linking directly to it in your Rmd `[text here](link here)`, and listing the full citation below the conclusions section. Please ask me if you aren't sure how to cite references.