TL;DR

May be too long, but please do read

Total

10 pts

Showcase your inner data scientist

TL;DR

Pick a dataset, any dataset…

…and do something with it. That is your final project in a nutshell. More details below.

May be too long, but please do read

The final project for this class will consist of analysis on a dataset of your own choosing. The dataset may already exist, or you may collect your own data using a survey or by conducting an experiment. You can choose the data based on your interests or based on work in other courses or research projects. The goal of this project is for you to demonstrate proficiency in the techniques we have covered in this class (and beyond, if you like) and apply them to a novel dataset in a meaningful way.

The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather let me know that you are proficient at asking meaningful questions and answering them with results of data analysis, that you are proficient in using R, and that you are proficient at interpreting and presenting the results. Focus on methods that help you begin to answer your research questions. You do not have to apply every statistical procedure we learned (and you can use techniques we haven't officially covered in class, if you're feeling adventurous). Also, critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the statistical analysis should be discussed here.

The project is very open ended. You should create some kind of compelling visualization(s) of this data in R. There is no limit on what tools or packages you may use, but sticking to packages we learned in class (`tidyverse`) is required. You do not need to visualize all of the data at once. A single high quality visualization will receive a much higher grade than a large number of poor quality visualizations. Also pay attention to your presentation. Neatness, coherency, and clarity will count. All analyses must be done in RStudio, using R.

Data

In order for you to have the greatest chance of success with this project it is important that you choose a manageable dataset. This means that the data should be readily accessible and large enough that multiple relationships can be explored. As such, your dataset must have at least 50 observations and between 10 to 20 variables (exceptions can be made but you must speak with me first). The dataset's variables should include categorical variables, discrete numerical variables, and continuous numerical variables.

If you are using a dataset that comes in a format that we haven't encountered in class, make sure that you are able to load it into R as this can be tricky depending on the source. If you are having trouble ask for help before it is too late.

Note on reusing datasets from class: Do not reuse datasets used in examples, homework assignments, or labs in the class.

Below are a list of data repositories that might be of interest to browse. You're not limited to these resources, and in fact you're encouraged to venture beyond them. But you might find something interesting there:

- TidyTuesday (https://github.com/rfordatascience/tidytuesday)
- NHS Scotland Open Data (https://www.opendata.nhs.scot/)
- Edinburgh Open Data (https://edinburghopendata.info/)
- Open access to Scotland's official statistics (https://statistics.gov.scot/home)
- Bikeshare data portal (https://www.bikeshare.com/data/)
- UK Gov Data (https://data.gov.uk/)
- Kaggle datasets (https://www.kaggle.com/datasets)
- OpenIntro datasets (http://openintrostat.github.io/openintro/)
- Awesome public datasets (https://github.com/awesomedata/awesome-public-datasets)
- Youth Risk Behavior Surveillance System (YRBSS) (https://chronicdata.cdc.gov/Youth-Risk-Behaviors/DASH-Youth-Risk-Behavior-Surveillance-System-YRBSS/q6p7-56au)
- PRISM Data Archive Project (https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/fenway.html)
- Harvard Dataverse (https://dataverse.harvard.edu/)
- If you know of others, let me know, and we'll add here…

Deliverables

Question 1. Proposal - due 11/28/2022 (1:00pm).

Question 2. Presentation - due 12/05/2022 (1:00pm).

Proposal

This is a draft of the introduction section of your project as well as a data analysis plan and your dataset.

- **Section 1 - Introduction:** The introduction should introduce your general research question and your data (where it came from, how it was collected, what are the cases, what are the variables, etc.).

- **Section 2 - Data:** Place your data in the `/data` folder, and add dimensions and codebook to the README in that folder. Then print out the output of and codebook to the README in that folder. Then print out the output of `glimpse()` or `skim()` of your data frame.
- **Section 3 - Data analysis plan:**
    - The outcome (response, Y) and predictor (explanatory, X) variables you will use to answer your question.
    - The comparison groups you will use, if applicable.
    - Very preliminary exploratory data analysis, including some summary statistics and visualizations, along with some explanation on how they help you learn more about your data. (You can add to these later as you work on your project.)
    - The method(s) that you believe will be useful in answering your question(s). (You can update these later as you work on your project.)
    - What results from these specific statistical methods are needed to support your hypothesized answer?

Each section should be no more than 1 page (excluding figures). You can check a print preview to confirm length.

The grading scheme for the project proposal is as follows. Note that after you receive feedback for your proposal you can improve it based on the feedback and re-submit it. If you re-submit, your final score for the proposal will be the average of two scores you receive (first and second submission).

Total	10 pts
Data	2 pts
Proposal Writing | 4 pts	
Workflow, organization, code quality	2 pt
Teamwork	2 pt

Presentation

10 minutes maximum (including Q&A), and each team member should say something substantial.

Prepare a slide deck using the template in your repo. This template uses a package called `xaringan`, and allows you to make presentation slides using R Markdown syntax. There isn't a limit to how many slides you can use, just a time limit (10 minutes total: ideally, 7-minute talk / 3-minute Q&A). Each team member should get a chance to speak during the presentation. Your presentation should not just be an account of everything you tried ("then we did this, then we did this, etc."), instead it should convey what choices you made, and why, and what you found.

Before you finalize your presentation, make sure your chunks are turned off with `echo = FALSE`.

Presentations will take place during the last day of the class (12/05/2022).

You will watch presentations from other teams in your workshop and provide feedback in the form of peer evaluations. The presentation line-up will be generated randomly.

The grading scheme for the presentation is as follows:

Total	**20 pts**
Time management: Did the team divide the time well amongst themselves or got cut off going over time?	2 pts
Content: Is the research question well designed and is the data being used relevant to the research question?	3 pts
Professionalism: How well did the team present? Does the presentation appear to be well practiced? Did everyone get a chance to say something meaningful about the project?	3 pts
Teamwork: Did the team present a unified story, or did it seem like independent pieces of work patched together?	3 pts
Content: Did the team use appropriate statistical procedures and interpretations of results accurately?	3 pts |
Creativity and Critical Thought: Is the project carefully thought out? Are the limitations carefully considered? Does it appear that time and effort went into the planning and implementation of the project?	3 pts |
Slides: Are the slides well organized, readable, not full of text, featuring figures with legible labels, legends, etc.?	3 pts |

Repo organization

The following folders and files in your project repository:

- `presentation.Rmd` + `presentation.html` : Your presentation slides
- `README.Rmd` + `README.md` : Your write-up
- `/data` : Your dataset in CSV or RDS format and your data dictionary
- `/proposal` : Your project proposal

Style and format does count for this assignment, so please take the time to make sure everything looks good and your data and code are properly formatted.

Tips

- You're working in the same repo as your teammates now, so merge conflicts will happen, issues will arise, and that's fine Commit and push often, and ask questions when stuck.

- Review the marking guidelines below and ask questions if any of the expectations are unclear.

- Make sure each team member is contributing, both in terms of quality and quantity of contribution (we will be reviewing commits from different team members).

- Set aside time to work together and apart (physically).
- When you’re done, review the documents on GitHub to make sure you’re happy with the final state of your work. Then go get some rest!
- Code: In your presentation your code should be hidden (`echo = FALSE`) so that your document is neat and easy to read. However your document should include all your code such that if I re-knit your R Markdown file I should be able to obtain the results you presented.
    - Exception: If you want to highlight something specific about a piece of code, you’re welcomed to show that portion.
- Teamwork: You are to complete the assignment as a team. All team members are expected to contribute equally to the completion of this assignment and team evaluations will be given at its completion - anyone judged to not have sufficient contributed to the final product will have their grade penalized. While different teams members may have different backgrounds and abilities, it is the responsibility of every team member to understand how and why all code and approaches in the assignment works.

Marking

Total	30 pts |
Proposal	10 pts
Presentation	20 pts

Criteria

Your project will be assessed on the following criteria:

- Content - What is the quality of research and/or policy question and relevancy of data to those questions?
- Correctness - Are statistical procedures carried out and explained correctly?
- Writing and Presentation - What is the quality of the statistical presentation, writing, and explanations?
- Creativity and Critical Thought - Is the project carefully thought out? Are the limitations carefully considered? Does it appear that time and effort went into the planning and implementation of the project?

Team peer evaluation

You will be asked to fill out a survey where you rate the contribution and teamwork of each team member out of 10 points. You will additionally report a contribution percentage for each team member. Filling out the survey is a prerequisite for getting credit on the team member evaluation. If you are suggesting that an individual did less than 20% of the work, please provide some explanation. If any individual gets an average peer score indicating that they did less than 10% of the work, this person will receive half the grade of the rest of the group.