STAT210 Project Instructions

You will conduct a statistical study on a topic of your choice. This task will require you to find a dataset, pose interesting research questions, analyse the data, present your results orally to the class, and hand in a written report describing your study and its findings. The project is an opportunity to show off what you've learned about data analysis, visualization, and statistical inference.

You should pose the problem that you want to solve as precisely as possible at the outset. Next, identify the population you want to describe, and think about how you will obtain relevant data. You should also make a hypothesis, *a priori* (before you analyse the data), about the results you expect to see.

All analysis must be completed using the R programming language via RStudio, and your write up must be an R Markdown document. To help you get started we provide a template Rmd file (stat_inf_project.Rmd).

Places to Find Data

Finding the right data to answer your particular question is part of your responsibility for this assignment. Public data sets are available from hundreds of different websites, on virtually any topic. Be creative! Go find the data that you want!

Below is a list of places to get started, but this list should be considered grossly non-exhaustive:

- Gapminder (www.gapminder.org/data/)
- Kaggle (www.kaggle.com/datasets)
- Data.gov (catalog.data.gov/dataset)
- U.S. Bureau of Labor Statistics (www.bls.gov)
- U.S. Census Bureau (census.gov)

Keep the following in mind as you select your topic and dataset:

- You need to have enough data to make meaningful inferences. There is no magic number of individuals required for all projects. But aim for at least 100 observations and make sure there are a combination of numerical and categorical variables (at least 5 variables in total).

Data

You must finalize and submit your data file to us via Moodle. Your technical report should import this data into RStudio using the `read.csv()` command.

- The data must be in CSV format (`.csv`). This means that the first row should be a comma-separated list of variable names, and the rest should be rows of data.
- Name all variables helpfully and contextually, e.g., use `Airport` and `WaterTemp`, and not `A` and `B`. Similarly, for the category names, use whole words and phrases, not cryptic codes, e.g., use `Male` and `Female`, not `1` and `2`. A binary variable `isFemale` can be coded 0 for male, and 1 for female (and then is self-documenting). A variable `sex` coded 1 and 2 is just asking for trouble.

Technical Report

Your technical report will be an annotated R Markdown file (`.Rmd`) that contains your R code, interspersed with explanations of what the code is doing, and what it tells you about the problem.

Content

You should **not** need to present *all* of the R code that you wrote throughout the process of working on this project. Rather, the technical report should contain the *minimal* set of R code that is necessary to understand your results and findings in full. If you make a claim, it *must* be justified by explicit calculation. A knowledgeable reviewer should be able to compile your `.Rmd` file without modification, and verify every statement that you have made. All of the R code necessary to produce your figures and tables *must* appear in the technical report. In short, the technical report should enable a reviewer to reproduce your work in full.

Format

Your technical report should follow this basic format:

1. Introduction: an overview of your project. In a few paragraphs, you should explain *clearly* and *precisely* what your research question is, why it is interesting, and what contribution you have made towards answering that question.
2. Data: a brief description of your data set. What variables are included? Where did they come from? What are units of measurement? What is the population that was sampled? How was the sample collected?
3. Exploratory data analysis: Perform exploratory data analysis (EDA) that addresses the research questions you outlined above. Your EDA should contain numerical summaries and visualizations. Each R output and plot should be accompanied by a brief interpretation.
4. Statistical Inference: Perform inference via hypothesis testing that addresses the research question you outlined above. Each R output and plot should be accompanied by a brief interpretation.

    - State hypotheses
    - Check conditions
    - State the method(s) to be used and why and how
    - Perform inference
    - Interpret results
    - If applicable, state whether results from various methods agree

    It is your responsibility to figure out the appropriate methodology. What techniques you use to conduct inference will depend on the type of data you're using, and your sample size. You should conduct at least two hypothesis tests, and report the associated p-value and the conclusion. If your data fails some conditions and you can't use a theoretical method, then you should use an appropriate simulation based method.

    - If you **can** use both theoretical and simulation based methods, then choose one and stick with it. You don't have to do both. However if you **can't** use both, then you need to decide which is appropriate.
    - It's essential to make sure the method you're using is appropriate for the dataset and the research question you're working with.

5. Conclusion: a summary of your findings and a discussion of their limitations. First, remind the reader of the question that you originally set out to answer, and summarize your findings. Second, discuss the limitations of your work, and what could be done to improve it. You might also want to do the same for your data.

Additional Thoughts

The technical report is *not* simply a dump of all the R code you wrote during this project. Rather, it is a narrative, with technical details, that describes how you addressed your research question. You should *not* present tables or figures without a written explanation of the information that is supposed to be conveyed by that table or figure. Keep in mind the distinction between *data* and *information*. Data is just numbers, whereas information is the result of analyzing that data and digesting it into meaningful ideas that human beings can understand. Your technical report should allow a reviewer to follow your steps from converting data into information.
You will not receive extra credit for simply describing your data *ad infinitum*. For example, simply displaying a table with the means and standard deviations of your variables is not meaningful. Writing a sentence that reiterates the content of the table (e.g. “the mean of variable x was 34.5 and the standard deviation was 2.8…”) is equally meaningless. What you should strive to do is interpret these values in context (e.g. “although variables x_1 and x_2 have similar means, the spread of x_1 is much larger, suggesting…”).
Your report should be submitted via Moodle as an R Markdown (`.Rmd`) file and the corresponding rendered output (`.html`) file.

Presentation

An effective oral presentation is an integral part of this project. One of the objectives of this class is to give you experience conveying the results of a technical investigation to a non-technical audience in a way that they can understand.

You will make a 10-minute oral presentation to the class. You should make (good) slides. Your goal should be to convey to your audience a clear understanding of your research question, along with a basic understanding of your analysis, and how well it addresses the research question you posed.

You should prepare electronic slides for your talk.

Assessment Criteria

Your project will be evaluated based on the following criteria:

- General (10%): Is the topic original, interesting, and substantial – or is it trite, pedantic, and trivial? How much creativity, initiative, and ambition did the student demonstrate? Is the basic question driving the project worth investigating, or is it obviously answerable without a data-based study?

- Data (10%): Are the variables chosen appropriately and defined clearly, and is it clear how they were measured/observed? Is there sufficient data to make meaningful conclusions?

- Analysis (50%): Are the chosen analyses appropriate for the variables under investigation? Are the summary statistics calculated correctly? Is each plot and R output accompanied by a narrative? Is the hypothesis stated clearly and matches the research question? Are the conditions checked in context of the data (not just a generic bullet point list of the conditions, but reasoning through them for the given dataset)?

- Technical Report (20%): How effectively does the written report communicate the goals, procedures, and results of the study? Are the claims adequately supported? How well is the report structured and organized? How well is the report edited? Are the statistical claims justified? Are text and analyses effectively interwoven in the technical report? Clear writing, correct spelling, and good grammar are important.
- Oral Presentation (10%): How effectively does the oral presentation communicate the goals, procedures, and results of the study? Do the slides help to illustrate the points being made by the speaker without distracting the audience? Do the presenters seem to be well-rehearsed? Does she appear to be confident in what she is saying? Are her arguments persuasive?