Stat7350: Final Project

Due: April 15, 2019

The final project is a continuation of what you have learned in the class, and the skills you have used to complete Assignments 1-3. The major difference is that here you are required to find your own dataset to analyze. In some way this assignment is the reverse of what biologists typically do, since here you will searh for a story after choosing the data, rather than the other way around (note, nothing is preventing you from exploring multiple datasets and choosing the one that is more interesting . . .)

Part I: Find a dataset to analyze

This can be almost anything that you choose, as long as it is suitably complex to conduct the requirements for Part II below. Hints about datasets: The best approach is probably to start with a problem that interests you, and then look for data. Alternately, you can start with a dataset. Here are a few places you can find data:

- Registry of Open Data on AWS.
- Data.gov, U.S. open government data *this is where I found the bee dataset for assignment 2*
- [SQLShare](https://sqlshare.escience.washington.edu/sqlshare/: public scientific datasets. Some require considerable knowledge to interpret, others are easier to understand. You can select "All datasets" and then filter by keyword, or you can select a tag from among those in the left column.

Part II: Conduct an exploratory analysis

Using the techniques of exploratory data analysis, interact with your data (as in Assignment 1). Filter, summarize, and plot the data in different ways to identify interesting patterns. You should aim to have at least three different figures that present different aspects of the data to tell a story. This means that you will have to find a suitably complex dataset, or multiple related datasets.

In parallel with the figures, you should conduct a data analysis to tell a complete story. I am not looking for complicated statistics here. I expect things like linear models (`lm` or `aov`, `t.test`, `wilcox.test`), corrlation tests (`cor.test`), chisquare test (`chisq.test`). You are of course welcome to use more complicated statistics if you desire. When necessary, assumption tests should be included (and reported in the appropriate section, see below).

Some of the work done in this section will not be directly included in your final report, though you should include a reproducible RMD file as a "Supporting File". It should be clear from this file how to reproduce your results. Give clear and explicit instructions for obtaining the data and for running the analysis and creating the figures included in the report. It may also include additional figures or analyses, you do not need to edit it down to include only what you include in the main report.

Part III: Write a research report

Section 1: Introduce the dataset and the questions that you will answer with your report

This is the first major section of the report. Clearly state the context and motivation for this project. Why was this data collected? What is already known about this topic? Why is it important to know more? At the end of the introduction, you should clearly state the main question as a single sentence with an unambiguous answer.

Section 2: Analysis Methods

You should describe the dataset that you used, including exact URLs. The data must be real — neither you nor someone else may make up the data.

You should then write a complete and clear description of the analysis you performed. A complete methods section will include enough information that anyone with reasonable intellect could reproduce your analysis given just the data. Note that this does **not** mean you should include code, but rather should describe the dataset you are working with (e.g., what are the variables) and explain how conducted analyses to answer the questions identified in the inroduction.

This section should tie your progmamming analysis to your research questions, indicating exactly what results would lead you to what conclusions.

Section 3: Results

This should be similar to the results section of a manuscript. It will be a narrative discussion that answers the question you posed through text, figures, and report of statistical analyses you conducted. All included figures should use principles of effective display.

Note that you should pick a dataset that allows you to say something definitive!. Focus in particular on the results that are most interesting, surprising, or important. Discuss the consequences or implications. Interpret the results: if the answers are unexpected, then see whether you can find an explanation for them, such as an external factor that you analysis did not account for. If there are limitations to the data you should note them here (but you should pick a dataset that doesn't have too many limitations).

Note that all figures in scientific papers should contain legends. Here is some good advice about writing legends. There's also some here.

Section 4: Conclusion & Reflection

Summarize what you learned about this dataset in approximately one paragraph.

Summarize what you learned from this assignment in approximately one paragraph. What do you wish you had known before you started? What would you do differently? What advice would you offer to future students embarking on this project?

Evaluation

Components of the assignment will be graded on a 3-point scale. Each aspect has different weights based on the amount of work expected to go into that section. The general rubric is here:

Weight	Topic	Excellent: 3	Satistfactory: 2	Needs Work: 1
4	Datatset	Data set chosen is appropriate (complex enough to be interesting, simple enough to be interpretable	N/A	Data set chosen is either too simple or too complicated for this purpose.
2	Introduction	Provides background research into the topic and summarizes important findings from the review of the literature describes problem to be solved; justifies the study; explains the significance of the problem.	Provides background research into the topic and describes the problem to be solved	Provides background research into the topic but does not describe the problem to be solved; insufficient or nonexistent explanation of details.
2	Methods	Ideas are arranged logically to fully address the problem statement and desired outcomes. Reader can easily follow the line of reasoning. An analysis plan is fully outlined.	Ideas flow and are usually linked to each other. The reader can follow the line of reasoning most of the time. An analysis plan is mostly outlined.	The writing is not logically organized and does not address the problem statement. The reader cannot identify a ine of reasoning. An analysis plan is not included or is inappropriate.
2	Results/Discussion	Addresses the topic with clarity; organizes and synthesizes information; and draws conclusions	Addresses the topic; lacks substantive conclusions; sometimes digresses from topic of focus	Presents little to no clarity in formulating conclusions and/or organization
5	Presentation: graphs	Graphs carefully tuned for desired purpose. One graph illustrates one point	Graph well chosen, but with a few minor problems: inappropriate aspect ratios, poor labels.	Graphs poorly chosen to support questions.
2	Statistical analysis	Statistical analysis supports the stated results. All necessary information is provided, including the specific tests used and resulting effect size, degrees of freedom, p-values. Where appropriate assumptions have been stated and tested.	Statistical analysis supports most of the stated results. The majority of ecessary information is provided. Where appropriate, assumptions have generally been stated and tested.	Statistical anaysis and reporting of results is almost entirely lacking.
1	Organization & Style	Well planned structure, written in an engaging, iteresting style. Strong paragraph structure. No grammatic errors. Reference have a consistent format.	Some evidence of organization, most paragraphs have topic sentences with supporting details. Style is competent though not engaging or inventive. Few grammatical errors. References are mostly in a consistent format.	Organization is unpredictable, paragraphs poorly structured. Lacks control over sentence structure, difficult to follow. Many grammatical errors. Referencing is inconsistent.
1	Ease of access for instructor, compliance with course standards	Access as easy as possible.	Satisfactory.	Not an earnest effort to reduce friction and comply with conventions.
1	Achievement, mastery, cleverness, creativity	Student has gone beyond what was expected and required, e.g., extraordinary effort, additional tools not addressed by this course, unusually sophisticated application of tools from course.	Tools and techniques from the course are applied very competently and, perhaps,somewhat creatively. Chosen task was acceptable, but fairly conservative in ambition.	Student does not display the expected level of mastery of the tools and techniques in this course. Chosen task was too limited in scope.