

DAT-102 Final project specifications

project goal Experientially acquire and/or magnify sample-based data inquiry skills including study design, implementation, sharing, and documentation

overview Formulate a compelling inquiry question about a unit of analysis of interest to you and one which can be studied safely and easily by CCAC students (such as a Vehicle, or Student, or Building). Design a unit sampling methodology and rationale. Implement your methodology, digitize the data, analyze the results, draw preliminary conclusions, and document your work thoroughly for future students.

Final project requirements

The following are the base level requirements for the Spring 2019 final project in data 102. All of these specs are negotiable through discussion with the instructor. Customization and creative modification of all requirements in the spirit of experimentation and exploration is strongly encouraged.

Project design

1. **Define a single unit of analysis**, such as a Book or a CCAC Student or Colleague, or Bicycle, or Road.
2. **Formulate a compelling inquiry question** (or small set of questions) related to this identified unit of analysis and compose a brief (1 paragraph) rationale for how you arrived at this question. Consider your own background, interests, hobbies, career/job, values, and curiosities.
3. **Generate a classification tree** whose root is the complete population of all instances of your unit of analysis (e.g. all books cataloged by the Library of Congress) and whose sub-branches depict sub-populations of your unit of analysis (e.g. books with call numbers starting with CR (heraldry)).
 - Formulate your tree iteratively over several drafts. Include all drafts, no matter how rough, with your final project documentation.
 - At the conclusion of your project, create a final copy of your classification tree digitally or neatly by hand (and scanned). draw.io is a google drive based tool that's excellent for making diagrams of all kinds.
 - Include at least 5 sub-categories of your unit of analysis
4. **Choose a sub-population** or a pair of sub-populations of your unit of analysis which you plan to study in your project. Provide a brief rationale for why this sub-population is appropriate.
5. **Define a variable** or set of variable values which you will gather about your chosen unit of analysis. (e.g. for our Library unit of analysis, we gathered 1) total numbered pages, 2) density of images, figures, or tables, 3) most recent publication year, 4) line density per page.

Create data instruments and procedures

6. **Create a spreadsheet to serve as a data gathering tool. Each row should capture data about a single instance of your unit of analysis. 1 or more columns should be dedicated to each of your chosen variables.**
7. **Add a data dictionary** in a second tab in your spreadsheet from the previous requirement which includes the following:
 - The name of each variable
 - Data type (integer, text, date, count, etc.)
 - Possible values or range of values, especially if coded (i.e. M = male, F=female, O=Other)

- Tips for accurate measurement

8. Develop a sampling procedure and write it out in step-by-step format such that another DAT-102 student has a reasonable chance of correctly implementing your very same sampling procedure on a sub-population of their choosing. Include how to randomize your selection, how to gather each variable, and what to do in foreseen unusual circumstances that might arise in the data gathering process. Include the number of units you plan to sample and any relevant calculations you used to arrive at this quantity.

9. Locate and **cite 1 or more peer-reviewed journal articles** relating to your area of inquiry. Some topics will lend themselves to more relevant academic literature than others. If yours is one that falls outside of academia's scope, choose a reliable and meaningful non-academic source and include a citation and access instructions in your final project.

10. Formulate a hypothesis or set of hypothesis concerning the outcome of the data you are about to gather. Draw hypothetical box-and-whisker charts, distributions, etc. Compose a paragraph defending your hypothesis.

Carry out your study & make claims

11. **Carry out your study:** gather your data, record that data digitally in your spreadsheet.

12. Carry out basic descriptive statistics on each of your variables of interest: mean, median, mode, quartiles, standard deviation (if appropriate). Summarize the differences seen across sub-populations using these descriptive statistics.

13. Generate compelling visualizations describing your results: box-and-whisker charts, neat hand-drawn figures, etc.

14. If you feel comfortable with the concepts, **apply basic statistics means testing** (e.g. interpreting the p-value generated from a t-test) comparing measured values between two sub-populations. See chapter 3 on confidence intervals in our Lock 5 statistics textbook, and chapter 6 for Inferences on means and proportions.

15. **Extract data-backed claims** that relate to your chosen inquiry question using your gathered data. A claim that suggests that not enough data was gathered, or that the data gathered is inclusive (or even worthless) is entirely appropriate.

Documentation and sharing

15. **Write a short letter to a future DAT-102 student** who may wish to continue your study detailing the following:

- Your experience in gathering the data (did anything unexpected happen?)
- Likely sources of error or bias in your sample: defend your approach to randomly drawing from the sub-population of interest.
- Revisions you would suggest making to your own study now that you've carried it out once.

16. Assemble a **list of resources** you used for each part of your study, including references to sections in our statistics textbook that were helpful in your analysis, any and all internet sites, etc.

17. Neatly assemble all project related files in your online shared directory in the Spring 2019 cohort work directory.