Choose Your Own Hypothesis Project

Overview

Students contracting for an **A** or **B** grade will complete an independent research-style project called the **Choose Your Own Hypothesis** assignment. This project integrates the major skills of the course — data sourcing, cleaning, analysis, and uncertainty estimation. You'll build a GitHub repo you can use as part of your portfolio.

Your goal is to:

- Identify a real-world question that can be answered with data,
- Develop a **testable hypothesis**,
- Analyze your data using tools from the course, and
- Quantify **uncertainty** in your results — including at least one metric where the **Central Limit Theorem (CLT)** does not apply, using **bootstrapping**.

Students contracting for an **A** will also present their results in a short talk during the final week of class.

Timeline

Week	Milestone	Deliverable
9	Project introduction	In-class overview, examples, and brainstorming
12	Topic check-in	Completed **Project Scaffold Table** (see template)
13	Descriptive statistics checkpoint	Markdown file or notebook with initial exploration and visualization
14	Final analysis	Complete GitHub repository with all files and uncertainty estimation; A-contract students present in class

Step-by-Step Project Guide

Step 1: Determine your topic and data source

Choose a topic that matters to you and that aligns with the university mission of *advancing the common good*. Your question should be **quantifiable** and **answerable with data**.

Examples

- *Are electric vehicles listed for sale at higher prices than comparable gas vehicles?*
- *Is there a systematic difference in Yelp reviews across cuisines?*
- *Do public transit delays increase during extreme temperature days?*

Deliverable: a short paragraph describing your idea and where the data will come from.

Clearly, I will be taking an expansive view of what constitutes "common good."

Here are two good sources of data:

- A repository of Awesome Public Data Sets
- A repository of Open Data Sets

Step 2: Acquire and document your data

Identify and acquire data that can test your hypothesis. You may use public datasets, APIs, or your own data collection.

Your data documentation should include:

- The **source** (URL, API, or dataset name)
- The **unit of analysis** (what one row represents)
- The **key variables** and their meanings
- Any **cleaning, filtering, or transformations** you perform

Locally, store your data in a `data/` folder in your GitHub repository. Don't commit this data to the repo. Instead, provide a link in the README where I can download your data.

Example: `data/used_cars.csv` — each row is one car listing; includes `price`, `mileage`, `model`, and `state`.

Step 3: Formulate your hypothesis

Write a single, testable hypothesis using **the Only One Test framework**. Clearly state what you expect and why. Identify:

1. **Test statistic** — what numerical value captures your comparison (e.g., mean difference, median ratio, correlation).
2. **Null hypothesis** — what it would mean if there were no real effect.
3. **Alternative hypothesis** — your prediction.

You will carry out your analysis using a permutation test.

Example:

- *Null:* There is no difference in mean odometer readings between Priuses sold in California and Missouri.
- *Alternative:* Priuses sold in Missouri have higher odometer readings on average.

Step 4: Identify at least one metric for bootstrap uncertainty

In addition to your permutation test, select two metrics and create bootstrap uncertainty intervals for those metrics. For at least one metric, make sure **the CLT does not apply**. The easiest way to do this is to choose a value that is not a mean or proportion.

Use **bootstrapping** to estimate a **confidence interval** for this metric.

Examples

- Median listing price difference between vehicle types
- Correlation between sentiment and review length
- Proportion of listings under $10,000 in two regions

Your report should explicitly explain:

- Why the CLT does not apply, and
- How you generated bootstrap samples and estimated uncertainty.

Step 5: Conduct your analysis

Using the tools from class, create a notebook to carry out the steps of your analysis:

- Load and clean your dataset
- Perform descriptive statistics and visualization
- Run your **permutation test** to evaluate your hypothesis
- Estimate **uncertainty** using bootstrapping (or another resampling method)

Each stage should be reproducible and documented in Markdown, with comments explaining what you did and why.

Step 6: Interpret and communicate

Your final deliverable should *tell the story* of your analysis:

- What question did you ask?
- What data did you use?
- What patterns or relationships did you find?
- How certain are you about those results?
- What would you recommend or conclude?

Your GitHub repository should read as a transparent, self-contained record of your work.

Repository Structure

This is one potential structure of the project.

```
project-name/
│
├── data/              # Raw data files or links
├── analysis.ipynb     # Main analysis code with narrative
├── README.md          # Overview of question, data, and findings
└── results/           # Optional: figures or output tables
```

Project Scaffold Table (Week 12 Check-In)

Use this table to submit your project plan for instructor feedback.

Element	Your Plan
Topic / Question	What real-world question are you investigating, and why does it matter?
Hypothesis	What do you expect to find? State a clear, testable hypothesis.
Outcome / Metric / Test Statistic	What variable(s) or statistic(s) will you analyze?
Units of Analysis	What does each observation represent?
Data Source(s)	Where will your data come from? Include links or access notes.
Why this data works	How is the dataset appropriate for your question?
Uncertainty Metric	Which variable will you use bootstrapping for?
Null Hypothesis	In plain English, what would it mean if there were no effect?

Items to Think About

Category	Description
Question & Hypothesis	Clear, specific, and testable.
Data & Documentation	Data are relevant, sufficient, and well-documented.
Analysis Quality	Correct, reproducible code; sound logic; descriptive statistics and visualization.
Uncertainty Estimation	Includes bootstrap or non-CLT interval with correct interpretation.
Interpretation & Communication	Clear conclusions, well-written, and connected to the hypothesis.
GitHub Repository	Organized, readable, and complete with appropriate structure and commits.
(A-track only)	Clear, concise oral presentation.

🧭 Tips for Success

- Start simple — a well-executed small question is better than an ambitious one you can’t finish.
- Write your hypothesis before analyzing.
- Use plain language when interpreting results.
- Label your figures clearly.
- Commit regularly and explain what changed in each commit.
- If your repository is private, add your instructor's GitHub account to the repo as a collaborator.
- Remember: uncertainty and clarity are the goals, not perfection.