**Activity: Reproduce or Refute? Building Trust in Data Science**

**Goal:**

Students learn the importance of **clean code**, **documentation**, and **reproducibility** by working in groups to “audit” and rewrite a messy analysis.

**Preparation (Instructor)**

1. Prepare a **small CSV dataset**, e.g., country population, sports data, or COVID stats (10–15 rows).
2. Create a **deliberately messy R script** that:
   * Has no comments
   * Uses inconsistent naming conventions
   * Mixes Base R and Tidyverse without explanation
   * Produces a plot and a summary statistic, but with minimal clarity
3. Upload this messy script and dataset to GitHub or share via Moodle.

**Class Setup**

* **Group size:** 5 students → ~16 groups
* **Time needed:** 50–60 minutes total
* Each group needs **1 laptop** with RStudio installed.

**Steps**

**Step 1 — Introduction (10 min)**

Briefly recap why reproducibility matters.

Show the **Reinhart & Rogoff Excel error** slide and explain how coding mistakes can influence **real-world decisions**.

*Prompt:* “What if someone gave you a dataset and code that generated a key result — how would you check if it’s trustworthy?”

**Step 2 — Hands-on Group Task (30 min)**

Each group downloads the messy R script and dataset.

Their **mission**:

1. **Run the messy code** and confirm it produces a result (e.g., a plot or table).
2. **Refactor the script** following best practices:
   * Use **descriptive object names** (snake\_case).
   * Add **meaningful comments**.
   * Replace Base R with **tidyverse pipes** (|>) where possible.
   * Organize the script into **clear sections** using comments.
3. **Document their workflow**:
   * Write down what steps they took.
   * Note any problems or improvements they made.

**Step 3 — Mini Presentations (10–15 min)**

Each group gives a **1-minute recap** of:

* One improvement they made to the script.
* Why that change makes the code more **readable** and **reproducible**.

Instructor collects key ideas on the board or shared document (e.g., *consistent naming*, *comments*, *project structure*).

**Debrief (5 min)**

* Connect their experience to the big picture:
  + **Code clarity** avoids confusion and errors.
  + Reproducibility is **essential for collaboration** and public trust.
  + Quarto/GitHub will help with these goals later in the course.
* Highlight that today’s lesson mirrors real-world team data science work.

**Deliverables**

* Cleaned R script (one per group).
* A short reflection (homework):

*“Why is reproducibility important in social science research? Give one example where a lack of reproducibility had major consequences.”*

**Why This Works**

* Students **immediately practice coding**, rather than just hearing about it.
* The messy code illustrates **real-world challenges** they will face.
* Group collaboration sets the tone for a course where teamwork matters.
* It connects directly to your slides about **writing good code**, **tidyverse**, and **reproducible research**.