Signature Assignment: Research Journal Report[1]

Course Learning Objectives

Course: Introduction to R Programming

Course Cross-listed Codes: ANTH155, DATA-R155, SOC155, BIOL155, PSYC155, and ANTH580

CLO: Design R Reports that explain code and results for transparency and reproducibility.

Module Learning Objective:

- MLO 1: Manage R files and data objects created in RStudio (load, save, and appropriately share).
- MLO 2: Write code in RStudio to solve sample problems on your chosen dataset(s).
- MLO 3: Conduct fundamental analyses and design visualizations using RStudio.

Assignment: Research Journal Report

Submission: Upload your final rendered report (HTML or PDF) to Canvas → Signature Assignment Module. No ZIPs or links; if your data is needed for review, provide a URL in the Dataset Citation & License section.

Final Due: Monday, December 8th (Noon)

Resubmission Window (no penalty): Until Friday, December 12th, 11:59 pm. No work accepted after December 12th.

Purpose

You will create a research journal–style report that documents how you explored, cleaned, and lightly analyzed an open dataset (or two) using tools and techniques learned across the term. Your report should read like an internal research notebook:

[1] Created 9/21/2025; This is a working document. Some content may change before the end of the semester.

before each code chunk, state your goal/question, explain your reasoning, show your attempts and checks, cite resources you used to learn, and reflect on what you discovered and how you'd do it next time.

What you will produce

A single Quarto or R Markdown report that renders to HTML or PDF and includes:

1. Narrative, code, and results are presented together, allowing readers to follow your thinking *and* verify what happened.
2. Weekly sections (see Scaffold Checkpoints) that show your progress on the same dataset(s) as you learn new techniques.
3. In-place AI Use Logs anywhere you consulted AI or other resources.
4. Dataset Citation & License information (author, year, title, URL, license—if available) and a short description of the dataset and why you chose it.

   Important: Do not paste or upload your actual dataset into AI tools. When using AI for learning, ask for generic examples, record them in your log, then write your own code for your data.

Report organization

1. Title & Overview (what this report is and why you chose the dataset)
2. Dataset Citation & License (author, year, title, URL, license + short description)
3. Data Overview (rows, columns, types; which required types are present; plan if multiple datasets)
4. Week-by-Week Journal Sections
   - Categorical (Oct 6)
   - Numeric & Integer (Oct 13)
   - Cleaning with dplyr (Oct 20, Oct 27, Nov 3)
   - Chaining with %>% (Nov 10)
   - Random Selection & IDs (Nov 17)
   - Visualization (Dec 1)
5. Summary of Key Decisions (e.g., NA handling, recoding, data preparation choices)
6. References & Resources (help pages, tutorials, AI Use Logs are embedded at point of use)

   For each section in the Week-by-Week: clearly state the goal, explain your reasoning, show code attempts and checks, and reflect on what worked, what didn't, and how you verified the results.

Scaffold checkpoints (weekly, due by **12:00 pm Noon** the day of class)

Checkpoints are formative: Earn *up to* 2 points each toward your learning or technical skill point categories (outside the 20 points for the final). Submit each checkpoint in Canvas by 12:00 pm. Checkpoint credit will be tallied in Canvas in your Learning Skills or Technical Skills categories (not in the 20-point Signature Assignment grade).

Sep 22 — Data Requirements & Ethics

Explore dataset options:

- **Focus:** Finding candidate datasets & matching required column types.
- **Submit in-class activity:** 2–3 candidates with a brief note on available column types and openness.
- **Journal cues:** What intrigues you? What questions might be answerable? What challenges do you foresee?

How to pick a dataset (quick checklist):

1. Use open-source data that others (including your instructor) can access. No protected human subjects data.
2. Your project across the term must involve these column types (multiple datasets are allowed if needed):
    a. **Numbers**: both integer and numeric (decimal), and explicitly consider discrete vs. continuous behavior where applicable.
    b. **Binary**: 0/1 or TRUE/FALSE.
    c. **Categorical**: factor or character treated as categories.
    d. **Free-text**: character strings.
3. Any file type is allowed as long as it yields a dataframe with the required column types.
4. You may analyze multiple datasets separately if one dataset doesn't contain all the required types.

Dataset Citation & License (You will write this where you first load the data):

1. Title of the dataset
2. Creator(s)/Organization (who collected/maintains it)
3. Year (of publication or version/date stamp)
4. Version or DOI (if available)
5. URL (direct page where the dataset can be accessed)

6. License name (and link), e.g., CC BY 4.0, CC0, ODbL; if none stated → choose a different dataset for this assignment
7. What it contains (2–3 sentences: subjects, variables, time period, geography)
8. How it was collected (method, instrument/source, who collected it)
9. Why you chose it + fit to course requirements (which required types it has: integer, numeric/decimal—discrete vs. continuous, binary, categorical, free-text)

   Quick rule: No license ≈ not open. For this project, pick a dataset that clearly states its license (or is in the public domain).

Module 5: Sep 29 — Dataset declaration

Focus: Declare your final dataset(s).
Submit in-class activity: Current notebook with memoing. Draft outline sections:

7. Title & Overview (what this report is and why you chose the dataset)
8. Dataset Citation & License (author, year, title, URL, license + short description). Include a code chunk that loads the dataset.
9. Write a short data overview of your data (what do the rows, columns represent?).

Note: If you have more than one dataset, do steps 2 and 2 twice.

Module 6: Oct 6 — Categorical (base R)

Focus: Identify and explore categorical variables, including subsetting, extraction, and insertion.
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. Rename columns and variables in rows where helpful; convert classes as needed; articulate your NA handling choices and why they fit your context.

Module 7: Oct 13 — Numeric & integer (base R)

Focus: Identify and explore numeric and integer variables, including subsetting, extraction, and insertion using logical operators.
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results.

Module 8: Oct 20 — Cleaning with *dplyr* (1)

Focus: Selecting and filtering of both categorical and numeric data
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. Rename columns where helpful; convert classes as needed; articulate your NA handling choices and why they fit your context. Consider potential recoding plans.

Module 9: Oct 27 — Cleaning with *dplyr* (2)

Focus: Mutate, ifelse
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. Demonstrate recoding.

Module 10: Nov 3 — Cleaning with *dplyr* (3)

Focus: Summarizing and group by column(s).
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. Create category counts, basic summaries, notes on outliers or unusual values and what they might mean for later steps.

Module 11: Nov 10 — Chaining (%>%) and tidying your code

Focus: Refactor earlier steps into clear pipelines for readability.
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. Before/after comparison and a short memo on what the pipeline improves for you as a reader.

Module 12: Nov 17 — Random selection & unique IDs

Focus: Practice random selection of rows and assigning random unique IDs; include a brief fairness/anonymization statement.
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. Replace IDs with random unique codes and keep the crosswalk

private. Select a random subset of rows (Application: focus group, machine learning or other modeling).

Module 13: Dec 1 — Visualization (ggplot2)

Focus: Visual design and interpretation.
Submit: Current notebook with memoing. Before each code chunk, clearly state the goal, explaining what the code will do. Show code attempts and outcome checks. After each section, explain the outcome, and reflect on what worked, what didn't, and how you verified the results. At least one univariate and one bivariate plot with captions tied to your data context. Explain what a reader should take away and how the plot supports your goals.

Module 14: Dec 8 — Final submission (graded, 20 pts)

Focus: Compose Outline sections: Summary of Key Decisions (conclusion) and References & Resources.
Submit to Canvas by 12:00 pm (Noon): Your rendered report (HTML or PDF). It should include all sections above, with narrative, code, results, and in-place AI Use Logs.

Collaboration policy

This is an individual submission. You may discuss ideas with classmates and compare approaches, especially since many of you will be using different datasets. Your final code, analysis, and writing must be your own.

Academic integrity & AI use

- You may use any resource (help pages, documentation, textbooks, classmates as discussion partners, online tutorials, generative AI for learning), but all learning must be documented, and the final code applied to your data must be your own.
- Generative AI–written code is not permissible for this assignment. Instead, use AI to ask questions and request generic examples. Document those examples and then write your own code for your dataset.
- In-place AI Use Log (insert at the point of use, not as an appendix):
    - Tool used (name/version if shown)
    - What you asked (your exact prompt or summary)
    - Short example the AI gave that does not use your data (quote minimally)
    - What you learned, in your own words

    - Your own code that applies the idea to your dataset
- Privacy: Do not upload type or paste your actual dataset or details about your data into AI tools.

Rubric Checklist

CLO 3 — Transparent reporting & research journaling (6 pts)

- Goals stated before code; reasoning + verification are visible (2)
- In-place AI Use Logs where applicable; sources cited (2)
- Weekly sections are coherent, and decisions (e.g., NA handling, recoding) are explained (2)

CLO 4 — Problem-solving code on your dataset(s) (5 pts)

- Techniques from each week are correctly applied to your data with at least one self-check per technique (3)
- At least one univariate and one bivariate plot with meaningful captions tied to your context (2)

CLO 1 — Managing files & data objects (5 pts)

- Data load succeeds; all required column types are addressed (including a note on discrete vs. continuous where relevant) (3)
- NA handling and data prep choices are documented and justified (2)

CLO 2 — RStudio operations & packages (4 pts)

- Evidence of RStudio skills for writing/running chunks and consulting Help (2)
- Use of dplyr/ggplot2/readr as taught (2)

Total: 20 points