

Exam Review Guide

Data Wrangling and Visualization

Prof. Jack Reilly

Fall 2026

Midterm Exam Guidelines

There are two exams that take place around the mid-point of the semester: a pen-and-paper **core exam** and a computer based **practicum** exam.¹

- On the “**core**” **exam**, you are permitted the full class time (although the exam will likely not take the whole time). You need to bring a pen and/or pencil. If you bring a pencil for scratch work, please make sure to also bring and write your answers in **pen**, not pencil. You may also bring a **cheat sheet** — a single 8.5×11 sheet of paper with whatever you want written (or typed) on it.
- On the **take-home “practicum” exams**, you are permitted as much time as you like between when the exam is handed out and when it is due. The take-home **is open-book, open-note, and open-internet**, but it **is not open-human**. You *are* allowed to ask AI questions and use responses, but you must keep a record of what you ask and turn it in with your exam. **Do not ask AI to complete the test for you.**
- You can ask the professor questions.
- On both tests, all material covered in the class, in reading, or in homework is fair game. The take-home will lean toward **computation**, while the in-class portion will lean toward **conceptual questions**.

Core Themes & Big Ideas

- Reproducible research and workflows
- Open science and transparency
- Data provenance, lineage, and quality
- Tidy data vs. “messy” data

¹A second **practicum** exam takes place towards the end of the semester.

- Exploratory Data Analysis (EDA)
- “80% of work is data cleaning and wrangling”
- Code tradeoffs: correctness, efficiency, interpretability

Data Wrangling & Management Concepts

DATA STORAGE & FILE TYPES

- **Open formats:** CSV, TSV, JSON
- **Closed formats:** Stata, SPSS, SAS, etc.

DATA TYPES

- Boolean (T/F)
- Numerical (integers, doubles, etc.)
- String

DATA CLEANING

- Missing data (NA / nulls)
- Outliers
- Inconsistent formatting (dates, strings, units)

DATA TRANSFORMATION

- Recoding values
- Binning variables
- Filtering & subsetting data
- Append / concatenate / stacking (row-binding)

MERGING DATA

- Merges & joins
 - Inner join
 - Left join
 - Right join
 - Full outer join

AGGREGATION & STRUCTURE

- Collapsing / aggregation / grouping / summarization
- Reshaping data: wide vs. long formats
- Survey weighting

Tools, Languages, & Work Environments

- R and RStudio
- Base R tools
- Tidyverse (dplyr, tidyr, ggplot2, etc.)
- Markdown, R Markdown, Quarto
- git / GitHub (version control, branches, pushes)
- Scripting & code organization
- File management & organization
- Notebooks (Quarto Notebooks, R Markdown)

Visualization Concepts & Design Principles

- Options for contrast: position, length, color, size, shape, angle
- Human visual perception capabilities
- Data-ink ratio
- “Chartjunk”
- Where good graphics go bad: aesthetics, data, perception

COMMON GRAPH TYPES

Univariate

- Histogram
- Density plot
- Bar chart

- Line graph
- Dot plot
- Boxplot
- Violin plot

Bivariate

- Scatterplots
- Sliced scatterplots

VISUALIZATION DESIGN

- Axis design: labels, ticks, breaks
- Overplotting, transparency, jittering
- Visual hierarchy & emphasis
- Avoiding misrepresentation / misleading visuals
- Evaluation and critique of visualizations

Workflow & Best Practices

- Modular, readable code
- Documenting code with comments
- Version control and clear commit messages
- Reproducibility: full pipeline from raw data to outputs
- Collaboration and shared workflows

Final Exam Guidelines

In addition to the final project, there is a practicum at the end of the semester.² The structure for this practicum is the same as the first, and all material is cumulative. You will need to know how to work with all of the above concepts, and may also need some of the following.

²There is no second core exam.

Data Workflow

- Code for portability
 - Projects
 - renv
 - environment control
- Flow control
 - loops (for and while)
 - lists
 - calling other scripts from your script
- Accessing remote data
- Data API calls (ie, census)

Network Data

DESCRIBING NETWORKS

- Nodes (vertices, points, actors, sites)
- Edges (ties, links, bonds, relations, arcs)
- Ego
- Alters
- First & Second Order Ties
- Dyads & Triads
- Directed & Undirected Networks
- Weighted & Unweighted Networks
- Centrality
 - Degree
 - Geodesic (shortest) paths
 - Betweenness

ANALYZING AND UTILIZING NETWORK DATA

- Network data
 - Nodelist
 - Edgelist
 - Adjacency matrix
- Network visualization focal points
 - Key actors/links
 - Relationship Strength
 - Structural Properties
 - Communities
- Sources of network map illustration, contrast, and differentiation
 - Color
 - Position
 - Size
 - Shape
- Principles of network aesthetics
- iGraph library

Geographical Data

CARTOGRAPHY

- What is a map?
 - Models
 - “Map is not territory”
- Elements of maps
 - Point
 - Line
 - Polygon
 - Shading
 - Labels
 - Scale
- Coordinates & Coordinate Systems
- Map Projection
 - Theorema Egregium
- Areal Units
 - Boundaries
 - Modifiable Areal Unit Problem
- Color and Contrast
 - Bertin’s visual variables: position, size, shape, value, hue, orientation, texture
 - Mapping of visual variables and encoding to graphical element types
 - Color theory & color schemes: sequential, diverging, qualitative
 - Hue, saturation, and value
 - Induction
- Overall map design principles

GEOGRAPHICAL DATA

- GIS and R
- Types of data
 - Vector
 - Raster
- Spatial data joins
- Geographical R libraries