Econ 5/Poli 5D
Data Analytics/ Social Sciences

Shubhro Bhattacharya
UC San Diego

Lab-3
January 22, 2026

Quick Survey

What did we learn from Lab-2?

Today's Agenda & Motivation

- How economists use large administrative data to study policing
- Why merging datasets is essential in real empirical work
- What today's lab emphasizes: workflow, inspection, merges, and checks
- Then we dive straight into the do-file and fill in the `XX`'s together

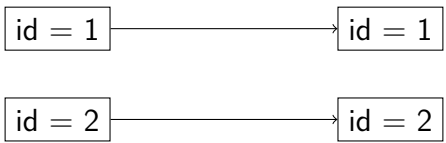
From Raw Stops to Analysis Variables

- Raw data is not analysis-ready: variables live across multiple files
- Typical structure:
  - File A: stop details (who, when, where)
  - File B: search outcomes (searched? contraband found?)
- Research questions require constructing outcomes (e.g., search rates)
- Main skills today:
  1. Merging datasets
  2. Creating analysis-ready variables

Skill 1: Merging Datasets

- Goal: join information from multiple files into one analysis dataset
- **Key step: identify a unique ID** that matches observations across files
- Use `isid <id>` to test if an ID uniquely identifies rows
- Decide merge type: `1:1`, `1:m`, or `m:1` based on uniqueness
- Always inspect `_merge` after merging (what matched vs. didn't)

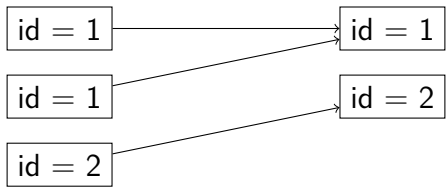
1:1 Merge (One-to-One)



Interpretation

- Each dataset has **exactly one observation per id**
- Safe when ids uniquely identify observations in both datasets
- Example: village-level census + village-level admin data

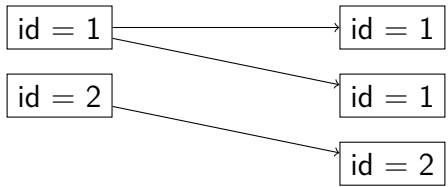
m:1 Merge (Many-to-One)



Interpretation

- Multiple observations in the master dataset match one in using
- Very common and usually safe
- Example: student-level data merged with school-level data

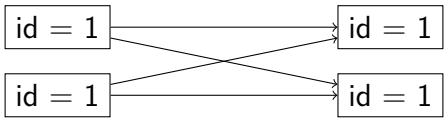
1:m Merge (One-to-Many)



Interpretation

- One observation in master matches many in using
- Less common — often signals a panel or event expansion
- Example: village data merged with yearly outcomes

m:m Merge (Many-to-Many)



Interpretation

- Every observation matches multiple observations
- Causes **data duplication and inflation**
- Almost always a sign of a merge mistake

Tip: **Never** use this!

Skill 2: Making Variables Analysis-Ready (Strings → Numbers)

- Many datasets store key variables as **strings** (e.g., `"TRUE"`, `"FALSE"`)
- For analysis, create **numeric indicators**: `0/1` (or labeled categories)
- Always verify: `tab <var>` and `summ <var>` after recoding
- Rule: **analysis happens on numeric variables**, not raw strings

How to access course materials?

Course materials on my website