bit.ly/merge_workshop
The Art of the Merge
Ashley Jester, Tara Das, & Starr Hoffman
Don’t dive right into the merge!

Data Interview Questions

- What is the user’s research question?
- What does each dataset contain?
- Why is the user considering a data merge?
- Where did the datasets come from?
- What is the linking variable across datasets?
- Is a merge possible based on manual comparison of values in linking variable?
- Are there duplicate records in either dataset? Should there be?
- What other processing needs to be done to address the question?
Stata: Merging World Bank Data

Ashley Jester
In this section of the workshop, we’ll cover three scenarios that use data from the World Bank.

For each scenario we will:
- Determine the requirements of the researcher
- Look at the datasets under consideration
- Review the steps necessary for preparing the data for merging
- Merge the data using Stata

At the end, we’ll review what we’ve covered.
World Bank Data

- We’ll be using data from two databases at the World Bank
  - World Development Indicators
  - Health Nutrition and Population Statistics

- Data were accessed on May 8, 2015
  - According to the World Bank documentation, data were last updated on April 14, 2015

- Data are in panel data format
  - Cross-sectional by country
  - Time-series by year
General Stata Information

- Typing “help” plus any command brings up Stata’s syntax information

- You can use either the command line or the graphical user interface (GUI)

- Stata can work with at most two datasets
  - The dataset currently in memory
    - This the “master” dataset
  - The dataset to be accessed
    - This is the “using” dataset

- Use logs to save what appears in the output window

- Use do-files to create a list of commands to execute
Scenario 1

The user has a dataset composed of variables measured for a set of countries over time. She would like to add additional years of observations to this set.

This is not technically a merge but an **append**

- Appends are adding additional observations for the same variables in the original dataset

What does this require?

- Both datasets must contain the same variables
- Variables must share the same:
  - Names
  - Formats
Scenario 1

- Open both datasets to check the variables
  - You need to confirm that both datasets contain the same variables

- Open the do-file to access the commands for executing the append

- Run the append
  - We can also do this using GUI
  - Using the GUI in Stata shows us the syntax required to run the command
  - You can use this to build do-files & learn Stata syntax

- Check the results
  - Use generated variable to as validity check
Scenario 2

The user has a dataset composed of variables measured for a set of countries over time. She would like to add additional variables for the same countries and years.

This is an one-to-one merge
- One-to-one merges join two datasets where there is one - and only one - unique combination of variables for each observation, master & using

What does this require?
- Both datasets must share unique identifiers/key variables
- Unique identifiers/key variables must share the same:
  - Names
  - Formats
Scenario 2

- Open both datasets to check the variables
  - You need to confirm that both datasets share the same unique identifiers/key variables

- Open the do-file to access the commands for executing the merge

- Run the merge
  - Again, we can also do this using the graphical user interface (GUI)

- Check the results
  - Again, use the generated variable to check the validity of the merge
Scenario 3

The user has a dataset composed of variables measured for a set of countries over time. She would like to add additional information for these same countries that does not vary with time.

This is an **one-to-many merge**

- One-to-many merges join two datasets where a single observation in the master dataset maps to many observations in the using dataset.

What does this require?

- Both datasets still must share identifiers/key variables.
- Identifiers/key variables are no longer unique in the many dataset.
Scenario 3

- Open both datasets to check the variables
  - You need to confirm that the one dataset contains the same identifier/key variable that is present in the many dataset

- Open the do-file to access the commands for executing the merge

- Run the merge
  - Again, we can also do this using the graphical user interface (GUI)

- Check the results
  - Again, use the generated variable to check the validity of the merge
Review

- Understanding the requirements of the researcher is key
  - As we will see in the next sections, the software is just a tool

- Communicating with the researcher is essential to understanding the requirements
  - What data does she have?
  - What would she like the data to look like when it’s finished?
  - What steps might be necessary before doing any type of merge?

- Once the merge is completed, you should check for accuracy
  - Keep track of observations
  - Use generated variables to do validity checks
SAS: Merging Hospital Data

Tara Das
Outline

- We’ll review 4 reference questions that use New York City public health and hospital data

- For each question, we’ll -
  - consider the specific research need
  - look at the datasets
  - review data processing steps to prep the datasets for merging
  - merge datasets using SAS

- We’ll conclude with takeaway points
Data background

- New York City and New York State have open data portals (like Data.gov)
- Most hospital performance and outcome measures are reported to New York State by law
- Health Data NY is the state’s open data portal specific to hospital and health data for city and state hospitals.
- **Take Care New York**: New York City’s public health initiative to help all New Yorkers live healthier and longer lives by collaborating with hospitals, city agencies, and community organizations in achieving specific goals.
Hospitals become TCNY partners by implementing at least 4 recommended Community Health Interventions.

Examples:
- Distribute Health Department health promotion campaign literature.
- Screen and counsel patients to quit smoking.
- Reduce percent of HIV infected patients with detectable viral loads.
- Support breastfeeding within the hospital and in the community.
- Promote appropriate and judicious prescribing of opioid analgesics.

The list of TCNY partners is available at http://www.nyc.gov/html/doh/html/about/tcny_partner04.shtml
I have a list of TCNY partners from the Health Department. I have more facilities that need to be added to this list. How can I merge this data?

Overview:

1. Open both datasets
2. Which is the primary dataset?
3. Does any data in the primary dataset need to be updated by the “merge”?
   a. This is an append, not a merge.
4. Confirm that important columns have same headings.
   a. Not all columns need to exist in second dataset.
5. Do append
How do TCNY hospitals compare with non-TCNY hospitals in New York City on supporting breastfeeding?

Breakdown the question:

1. Does the user have a list of TCNY hospitals?
2. Does the user have hospital breastfeeding data?
3. Where did the datasets come from?
4. Let’s look at them and see how they can be merged.
5. We’ll apply the TCNY list to a dataset containing all hospitals and breastfeeding indicators to create a TCNY flag column.
Breastfeeding data for hospitals was downloaded from Health Data NY for year = 2012.


This chart shows percentage of newborns breastfed supplemented with formula, fed any breast milk, fed exclusively breast milk by hospital.

This dataset contains information reported by hospitals required to be compliant with New York State’s Maternity Information Law.
User Question 2: Prepping data for merge
User Question 2: Prepping data for merge

- The dataset has NYC & non-NYC hospitals and one year of data.
- The linking variable will be Hospital name (note: different from column name in TCNY list).
- Hospital names are repeated, but duplicates ok. There are different measures for each hospital.
- Create subset of dataset with only NYC hospitals (filter by hospital county).

- Do a quick eyeball of the merge by comparing datasets - any potential issues?
  - Some matches won’t be identified b/c names vary slightly, abbreviations, extra spaces.
  - Hospital names are tricky - hospitals themselves merge under new names constantly, so need to research sometimes to ensure names refer to same facility.
  - Numeric code identifier is preferable.
  - Hospital names cleaning into separate linking variable

- Prepping data takes time!
  - Any data cleaning or manipulation can be done in Excel.
  - Don’t replace original values, document the changes, and use consistent labels.
Different types of merging

What do you want in the final dataset?? **What is the research question**

1) All observations (if a or b)
2) All observations from A - matched or not (if a)
3) All observations from B - matched or not (if b)
4) Only matching observations (if a and b)
5) Only non-matching observations from A (if a and not b)
User Question 2: Merge

- Import data into SAS
- Sort each dataset by linking variable (has to have same name)
- Merge data
  - NYC Hospital data = a
  - TCNY data = b
  - How do we want final dataset to look like?
    - ***What is the research question***
    - We want to keep all hospital data regardless of match or not with TCNY
    - if a
Is there a difference in hospital-acquired infections (HAI) among TCNY hospitals?

Breakdown the question:

1. Does the user have the HAI dataset?
2. Does the dataset only contain TCNY hospitals?
3. Let’s look at the 2 datasets and see how they can be merged.
4. We’ll apply the TCNY list to the HAI dataset to restrict it to only TCNY.
Hospital acquired infections (HAI) data was downloaded from Health Data NY for year = 2013.

https://health.data.ny.gov/Health/Hospital-Acquired-Infections-Beginning-2008/utrt-zdsi

All acute care hospitals are required to report certain hospital-acquired infections (HAIs) to the New York State Department of Health.
  ○ This includes central line-associated blood stream infections in intensive care units; surgical site infections following colon, hip replacement/revision, and coronary artery bypass graft; and Clostridium difficile infections.
User Question 3: Prepping data for merge
User Question 3: Prepping data for merge

- The dataset has NYC & non-NYC hospitals and one year of data.
- Hospital names are repeated, but duplicates ok. There are different measures for each hospital.
- Can’t subset data to only NYC hospitals (no county column)
- Do a quick eyeball of the merge by comparing datasets
  - Similar issues with question 2 merge
- Should we use hospital name as linking variable?
  - Too many hospital names to manually clean
  - Facility ID would be better linking variable, and is consistent across NYS Health datasets
  - Add facility ID to TCNY list based on breastfeeding dataset (used vlookup function in Excel)
  - Some non-matches due to hospitals merging or closing - removed these (can’t give ID = 0, would match with NYS - All Hospitals in HAI dataset)
User Question 3: Merge

- Import data into SAS
- Sort each dataset by linking variable (has to have same name)
- Merge data
  - NYS Hospital data = a
  - TCNY data = b
  - How do we want final dataset to look like?
    - ***What is the research question***
    - We only care about TCNY hospitals
    - if a and b
I want to do outreach to hospitals to get them to join TCNY.

Breakdown the question:
1. Does the user have a list of hospitals for outreach?
2. Where did the list come from - does it contain non-TCNY hospitals already?
3. Let’s look at the 2 datasets and see how they can be merged.
4. We’ll apply the TCNY list to the hospital list to restrict it to non-TCNY.
User Question 4 - Merge

- Import data into SAS
- Sort each dataset by linking variable (has to have same name)
- Merge data
  - NYC Hospital data = a
  - TCNY data = b
  - How do we want final dataset to look like?
    - ***What is the research question***
    - We only care about non-TCNY hospitals
    - if a and NOT b
Take-aways

- The research question and data are more important than the software
- All merges are not the same - need to understand the question & data
- These scenarios used small datasets
- Prepping the data takes time, and needs vary

- SAS needs the linking variable to be in the same format and with the same heading
  - Easier to merge on numeric variables
  - Ensure that non-linking variables have different column headings
  - Always check dataset observations and SAS logs
  - Easier to control merge with an “if statement”
R: Merging Higher Education & Academic Library Data

Starr Hoffman
Questions for your data:

- What’s your question? What do you want to know?
  - What variables answer this question?
- Look at the codebook
  - Do the dataset years correspond? (2011 vs. 2010-2011)
- Look at your data
- Duplicate entries
  - Institutions with separate/affiliated libraries (law, health, etc.)
  - Branch locations
  - Multiple years of data
- Clean your data
- Review steps
- Merge/append
General R Information

- Use the command "help()" to get information about any function
  - Example: `help(View)`
- R is case-sensitive
- Option of command line (R) or GUI (R Studio)
- R can load many datasets at once; uses syntax to define how they relate
- Save scripts (executable commands) as .txt files
  - Save workspace as .RData (saves all objects and functions)
- Many different ways in R to accomplish the same thing
About the Data

NCES ALS: Academic Library Survey
- https://nces.ed.gov/ipeds/
- Every 2 years since 1996

NCES IPEDS: Integrated Postsecondary Education Data System
- https://nces.ed.gov/ipeds/
- Custom data downloads: http://nces.ed.gov/ipeds/datacenter/
- Annual data from over 7,500 institutions

ARL Statistics: Association of Research Libraries
- http://www.arlstatistics.org/analytics
Question 1

Is there an upward or downward trend in academic library budgets over the past few years?

- Dataset A = 2012 ALS from NCES
- Dataset B = 2010 ALS from NCES

APPEND: Add observations (rows): years of data

- Duplicates are expected (same institution, multiple years (tracked as variable/column)
Question 1

Check your data: number of observations?
Question 2

How do ARL libraries compare with non-ARL libraries on student retention?

- Dataset A = 2011 IPEDS from NCES
- Dataset B = 2010-11 ARL Statistics

- if A (ARLs and non-ARLs)
- if A and B (only ARLs)
- if A and not B (only non-ARLs)
Question 2

What variable do we link on?

…we can merge using a string variable.

Check number of observations.

Create a **status** or **indicator variable** to track the comparison between datasets.
In ARL libraries, is there a correlation between the number of library presentations (instruction) and the institution’s student graduation rate?

- Dataset A = 2011 IPEDS from NCES
- Dataset B = 2010-11 ARL Statistics

- if A (ARL and non-ARL libraries)
- if A and B (only ARL libraries)
- if A and not B (only non-ARL libraries)
Final Takeaways...

- Patron’s question is the most important
- Know where the data came from
- Know what the data means (codebook)
- Clean the data **before** you do the merge
Any questions?

Ashley Jester, ashley.jester@columbia.edu
Tara Das, tdas@columbia.edu
Starr Hoffman, starr.hoffman@columbia.edu