CONQUERING DATASETS

MONICA LOCKER, ASSESSMENT, TEACHING, AND LEARNING LIBRARIAN
COLLEGE OF THE HOLY CROSS

TOPICS

- Finding Data
- Interpreting Data
- Cleaning Data
- Using Data

FINDING

Like most things we encounter, data falls into different categories, and can be found in different places depending on the category.

- Social science data (demographic trends, public opinion, social issues)
- Economic data (macro-level indicators like GDP, unemployment %)
- Data about businesses/companies
  - Although this is technically "economic," I consider it a separate type of request because the information lives in a different place

FINDING

The first question I ask myself…

Who would want this?

FINDING

- The government
  - Data.gov or a state-level data website
- News organizations or politicians
  - Public opinion data like Pew or Roper
- Multinational NGOs
  - OECD, World Bank
- Private companies
  - Wharton Research Data Services, S&P Capital IQ
- Special interest groups

FINDING

Other techniques:

- Citation hunting in research literature
- Searching for news articles about the topic
- Contacting organizations or special interest groups that have the information
  - Lots of groups employ librarians, who are happy to help their colleagues!

INTERPRETING

Example: World Bank data

data.worldbank.org

INTERPRETING

Example: Equal Employment Opportunity Commission

eeoc.gov/eeoc/statistics/

WHAT IS CLEAN DATA?

From "Tidy Data" (Wickham 2014):

- Each variable is in a column
- Each observation is a row
- Each type of observational data forms a table

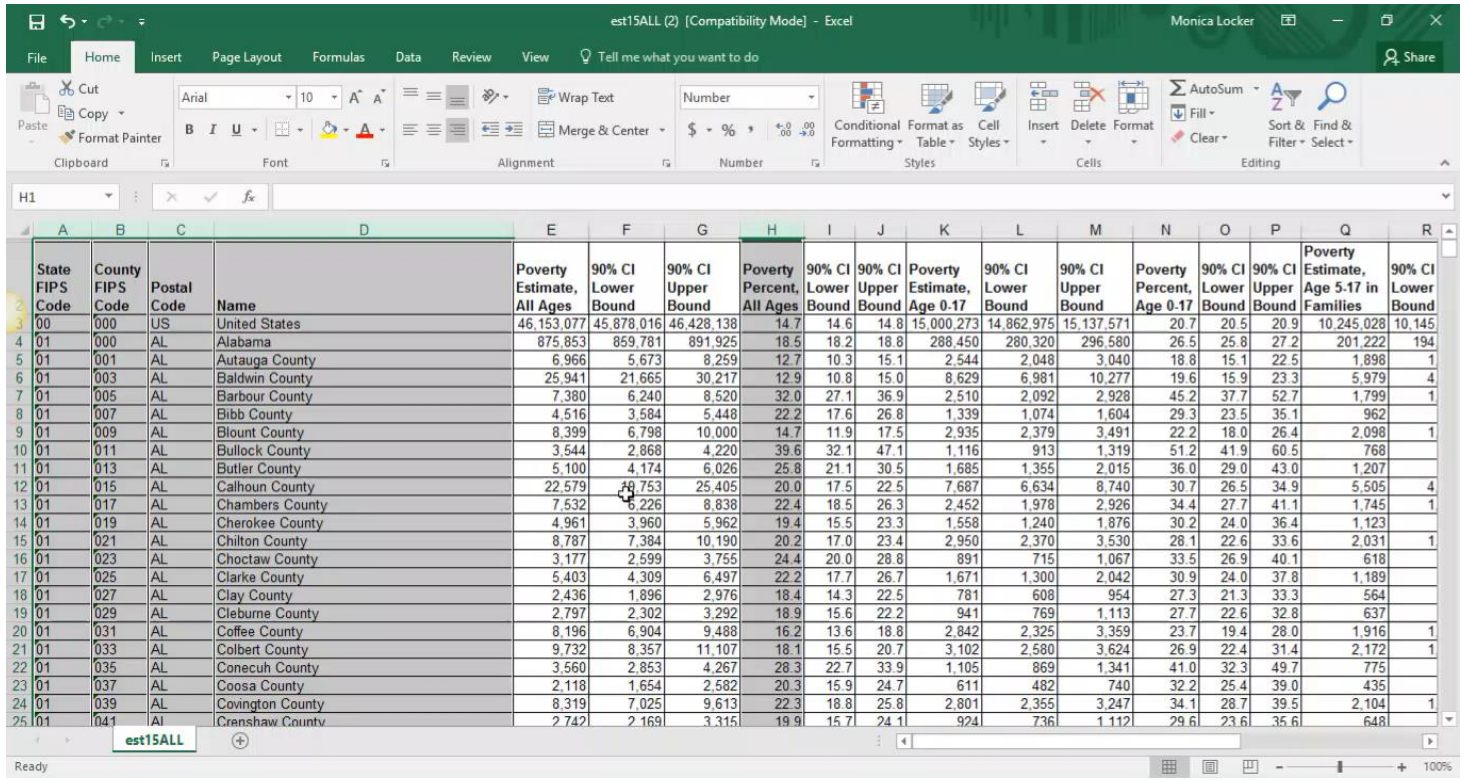
CLEANING DATA

- Always work on a copy
- Make sure the cells (values) are encoded with the correct data type
- Make labels meaningful
- Use the layout that your visualization or analysis tool requires
- Decide what to do with incomplete or missing data points
- Remove commas in numerical fields
- Calculate any fields that you need

EXAMPLE SCENARIO

Example: I'm working on a project that involves comparing poverty trends, immigration trends, and election results in the United States. My unit of observation is going to be the county.

CLEANING DATA



State FIPS Code	County FIPS Code	Postal Code	Name	Poverty Estimate, All Ages	90% CI Lower Bound	90% CI Upper Bound	Poverty Percent, All Ages	90% CI Lower Bound	90% CI Upper Bound	Poverty Estimate, Age 0-17	90% CI Lower Bound	90% CI Upper Bound	Poverty Percent, Age 0-17	90% CI Lower Bound	90% CI Upper Bound	Poverty Estimate, Age 5-17 in Families	90% CI Lower Bound
00	000	US	United States	46,153,077	45,878,016	46,428,138	14.7	14.6	14.8	15,000,273	14,862,975	15,137,571	20.7	20.5	20.9	10,245,028	10,145
01	000	AL	Alabama	875,853	859,781	891,925	18.5	18.2	18.8	288,450	280,320	296,580	26.5	25.8	27.2	201,222	194
01	001	AL	Autauga County	6,966	5,673	8,259	12.7	10.3	15.1	2,544	2,048	3,040	18.8	15.1	22.5	1,898	1
01	003	AL	Baldwin County	25,941	21,665	30,217	12.9	10.8	15.0	8,629	6,981	10,277	19.6	15.9	23.3	5,979	4
01	005	AL	Barbour County	7,380	6,240	8,520	32.0	27.1	36.9	2,510	2,092	2,928	45.2	37.7	52.7	1,799	1
01	007	AL	Bibb County	4,516	3,584	5,448	22.2	17.6	26.8	1,339	1,074	1,604	29.3	23.5	35.1	962	
01	009	AL	Blount County	8,399	6,798	10,000	14.7	11.9	17.5	2,935	2,379	3,491	22.2	18.0	26.4	2,098	1
01	011	AL	Bullock County	3,544	2,868	4,220	39.6	32.1	47.1	1,116	913	1,319	51.2	41.9	60.5	768	
01	013	AL	Butler County	5,100	4,174	6,026	25.8	21.1	30.5	1,685	1,355	2,015	36.0	29.0	43.0	1,207	
01	015	AL	Calhoun County	22,579	19,753	25,405	20.0	17.5	22.5	7,687	6,634	8,740	30.7	26.5	34.9	5,505	4
01	017	AL	Chambers County	7,532	6,226	8,838	22.4	18.5	26.3	2,452	1,978	2,926	34.4	27.7	41.1	1,745	1
01	019	AL	Cherokee County	4,961	3,960	5,962	19.4	15.5	23.3	1,558	1,240	1,876	30.2	24.0	36.4	1,123	
01	021	AL	Chilton County	8,787	7,384	10,190	20.2	17.0	23.4	2,950	2,370	3,530	28.1	22.6	33.6	2,031	1
01	023	AL	Choctaw County	3,177	2,599	3,755	24.4	20.0	28.8	891	715	1,067	33.5	26.9	40.1	618	
01	025	AL	Clarke County	5,403	4,309	6,497	22.2	17.7	26.7	1,671	1,300	2,042	30.9	24.0	37.8	1,189	
01	027	AL	Clay County	2,436	1,896	2,976	18.4	14.3	22.5	781	608	954	27.3	21.3	33.3	564	
01	029	AL	Cleburne County	2,797	2,302	3,292	18.9	15.6	22.2	941	769	1,113	27.7	22.6	32.8	637	
01	031	AL	Coffee County	8,196	6,904	9,488	16.2	13.6	18.8	2,842	2,325	3,359	23.7	19.4	28.0	1,916	1
01	033	AL	Colbert County	9,732	8,357	11,107	18.1	15.5	20.7	3,102	2,580	3,624	26.9	22.4	31.4	2,172	1
01	035	AL	Conecuh County	3,560	2,853	4,267	28.3	22.7	33.9	1,105	869	1,341	41.0	32.3	49.7	775	
01	037	AL	Coosa County	2,118	1,654	2,582	20.3	15.9	24.7	611	482	740	32.2	25.4	39.0	435	
01	039	AL	Covington County	8,319	7,025	9,613	22.3	18.8	25.8	2,801	2,355	3,247	34.1	28.7	39.5	2,104	1
01	041	AL	Crenshaw County	2,742	2,169	3,315	19.9	15.7	24.1	924	736	1,112	29.6	23.6	35.6	648	

LOADING DATA INTO TABLEAU

FOR MORE INFORMATION…

- Henderson, M. E. (2017). *Data management: A practical guide for librarians.* Lanham, MD: Rowman & Littlefield.
- Tableau Free Training Videos: https://www.tableau.com/learn/training
- Coursera: Getting and Cleaning Data, from JHU: https://www.coursera.org/learn/data-cleaning
- Coursera: Introduction to Data Analysis Using Excel, from Rice: https://www.coursera.org/learn/excel-data-analysis
- The Tidyverse: https://www.tidyverse.org/

REFERENCES

- Wickham, H. (2014.) Tidy data. *The Journal of Statistical Software, 59*, 1-23. Retrieved from https://www.jstatsoft.org/article/view/v059i10.
- U.S. Census Bureau, Small Area Income and Poverty Estimates (SAIPE) Program. (2016). SAIPE state and county estimates [Data set]. Retrieved from https://www.census.gov/did/www/saipe/data/statecounty/data/2013.html.
- U.S. Census Bureau, Population Estimates. (2015). Migration/geographic mobility data [Data set]. Retrieved from https://www2.census.gov/programs-surveys/popest/datasets/2010-2015/counties/totals/.
- U.S. Equal Employment Opportunity Commission. (2015). Job patterns for minorities and women in private industry
  (EEO-1) – national aggregate [Data set]. Retrieved from https://www.eeoc.gov/eeoc/statistics/employment/jobpat-eeo1/2015/datasets.cfm.
- The World Bank, DataBank. (2018). World development indicators [Data set]. Retrieved from http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators.