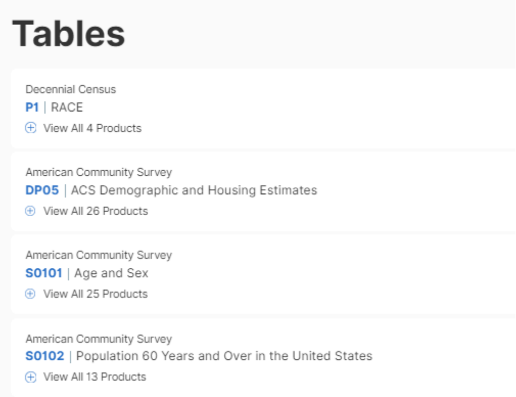
CSC 400, Census Data Directions, Spring 2024

The United States Census is a classic example of dealing with quintessential big data. Every ten years, the Census Bureau collects data on over 300 million people in the United States, growing every time they do it. In addition to general data on every American household, they also collect additional, more detailed data. They used to do this in the form of a long-form survey that selected households received every decade, but now they sample fewer people more often to collect more timely data on a whole host of additional variables. Both kinds of data collection results are available to the public and to researchers in various forms. In doing so, the Census must preserve the privacy of respondents according to the law.

In this assignment, you will look at several forms of data available from the Census for similar search terms and analyze what methods are used in each case to make the data available ethically and preserve privacy.

Start by going to https://data.census.gov/

Select your search terms. I suggest starting with a location of your choice (for example, Buffalo, NY) and seeing what data is available. A list of available tables (of data) will come up.



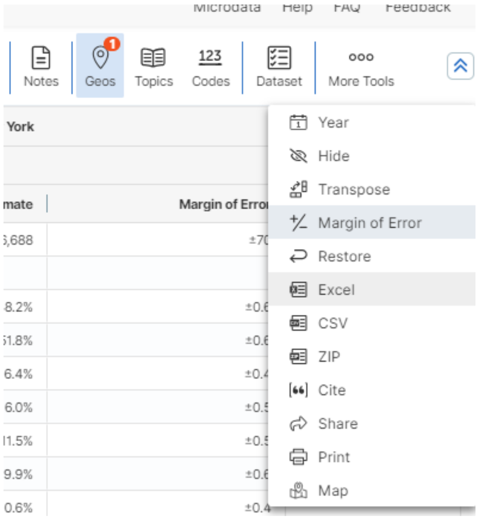
Tables

Decennial Census
P1 | RACE
⊕ View All 4 Products

American Community Survey
DP05 | ACS Demographic and Housing Estimates
⊕ View All 26 Products

American Community Survey
S0101 | Age and Sex
⊕ View All 25 Products

American Community Survey
S0102 | Population 60 Years and Over in the United States
⊕ View All 13 Products



The top table here is from the decennial census and is a summary of the racial breakdown of Buffalo. One important question to consider with this data is the form of the data. Is this available in individual responses or in another form? How does this form of the data help to preserve privacy?

If we add race to the search terms, we can narrow the search further. Compare the results you get from the American Community Survey. What kind of information is available? (This is sampling data rather than census data, so you should expect some differences in the presentation.)

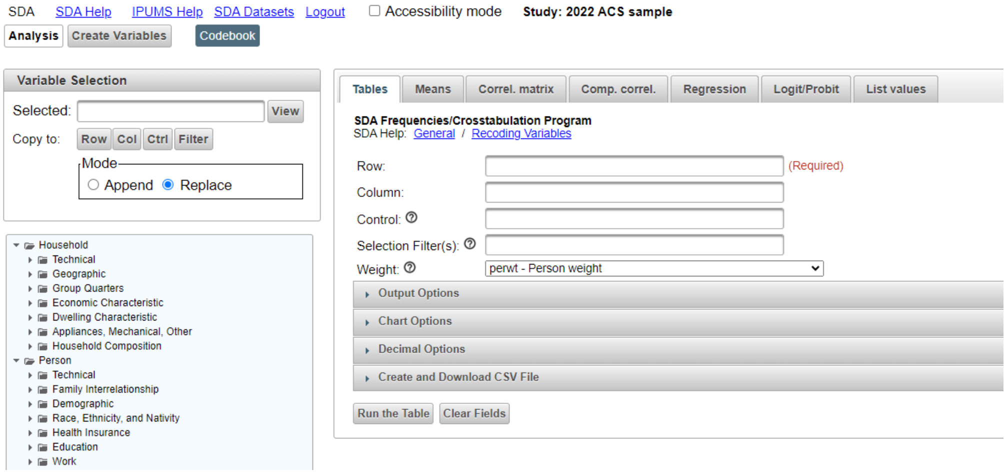
Download some of the data for further inspection and analysis. You can download the data in Excel or CSV format under More Tools on the top left of the screen.

Use the tools on the website to create plots or maps of the data you are looking at (map is on the same menu, along with citation tools you can use to properly cite the data for this assignment).

Look at the codes used in the data set (also visible in the previous screenshot). How are these codes used to help preserve privacy?

Next, go to https://www.ipums.org/ and select the USA site. You can download data here or analyze it online. To access data, you will need to create an account. Use your school account and indicate that it is for a course. Read the sign up carefully. How does it attempt to address data privacy and security through the sign-up process?

When you get into the online analysis tool, the screen you get looks like this:

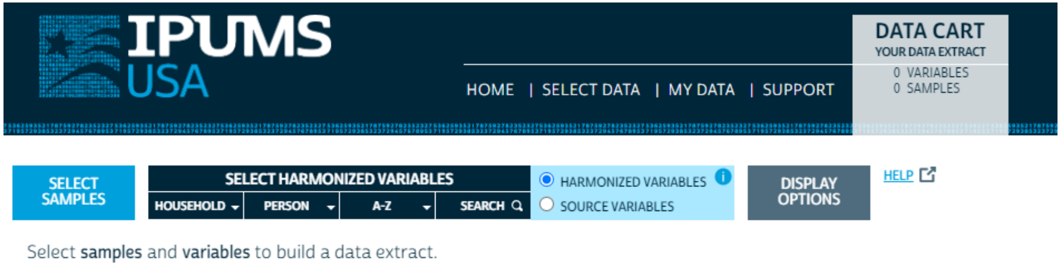


Study: 2022 ACS sample

SDA Frequencies/Crosstabulation Program

Weight: perwt - Person weight

I used the 2022 ACS sample for simplicity. Do a search for data similar to what you were looking at on the main census site. Experiment with the analysis tool. Talk about what you tried and the outcomes. You can also download data from the survey directly for analysis offline.

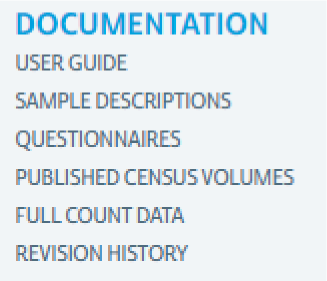


Select **samples** and **variables** to build a data extract.

Compare the format of the data available here to the data available on the main Census site. Does the level of detail differ? Do you get individual observations or only summaries?

In your report on what you found, relate what you see to databases and data warehousing, as well as OLAP. Be sure to consider aspects of Privacy, Ethics and Security when handling large amounts of data. How do Census methods satisfy these requirements? What kind of laws are in place to protect Census data?

Be sure to look at the documentation on both sites for additional details about data handling methods.



Write a report of 4-6 pages about what you found and responding to the questions posed in this document. Be sure to clearly state what your search and analysis terms were and properly cite the data (using the online citation tools provided). This assignment is worth **40 points**.