Data Guide from Grad Student perspective: linking data to research projects.

Purpose

We can access and see the numbers. But it is important to learn beyond the numbers. Without knowing how the data is collected, organized, or limited, it is difficult to “own” the data in your research projects.

Written by Theodore Kim tk46@rice.edu

AI Disclosure

During the preparation of this work, the author used [Gemini] to check grammar and improve readability. The final text was thoroughly reviewed, edited, and approved by the author

1. Census

https://www.census.gov/data/academy/courses/an-intro-census-data.html
There is self-paced learning tool from US Census.

Overview

What kind of data source/ data set they have?

-The U.S. Census Bureau conducts over 130 different surveys and programs each year, alongside the once-a-decade decennial census. (answer by Google AI)

Key survey we should know:
Decennial Census: Takes place every 10 years to count every resident in the United States.
American Community Survey (ACS): Survey data. Collect detailed social and economic data.
These two data are more likely to be used for economic, social, demographic variables. It is good idea to check other research articles to see what they have used.

What are their methodology?

https://www.census.gov/programs-surveys/acs/methodology.html
What would be bias? Can we trust this measurement? Acknowledging the limitation of the measurement is important because that would be referred for data interpretation and also provide future research implications.

Column variables and questions in the U.S. Decennial Census change significantly every 10 years (YR2020/ YR2030).

The Economic Census is the official five-year measure of American businesses and the economy. There will be more covered information regarding economic indicators at years ending in "2" and "7" (such as 2017 and 2022).

Also, take a look at the difference between 1-year estimates and 5-year estimates. (https://www.census.gov/programs-surveys/acs/guidance/estimates.html). Consider which option is best used for your research projects.

Where can we get the data?

US Census website; https://data.census.gov/
IPUMS; https://www.ipums.org/
Personally, IPUMS is easier to conduct. You can select each geographical level and columns. Also, IPUMS will export to have the first row with column names and first column with ID, which is suitable for data analysis without rearranging tables. However, after talking with Kelley Center

for Government Information and Civic Engagement, IPUMS has a very high entry barrier. Users need to know basic knowledge about the US Census survey options and differences.
***ACS data are in https://www.nhgis.org/ IPUMS NHGIS**

How to go beyond national/state-level data or go to panel data?

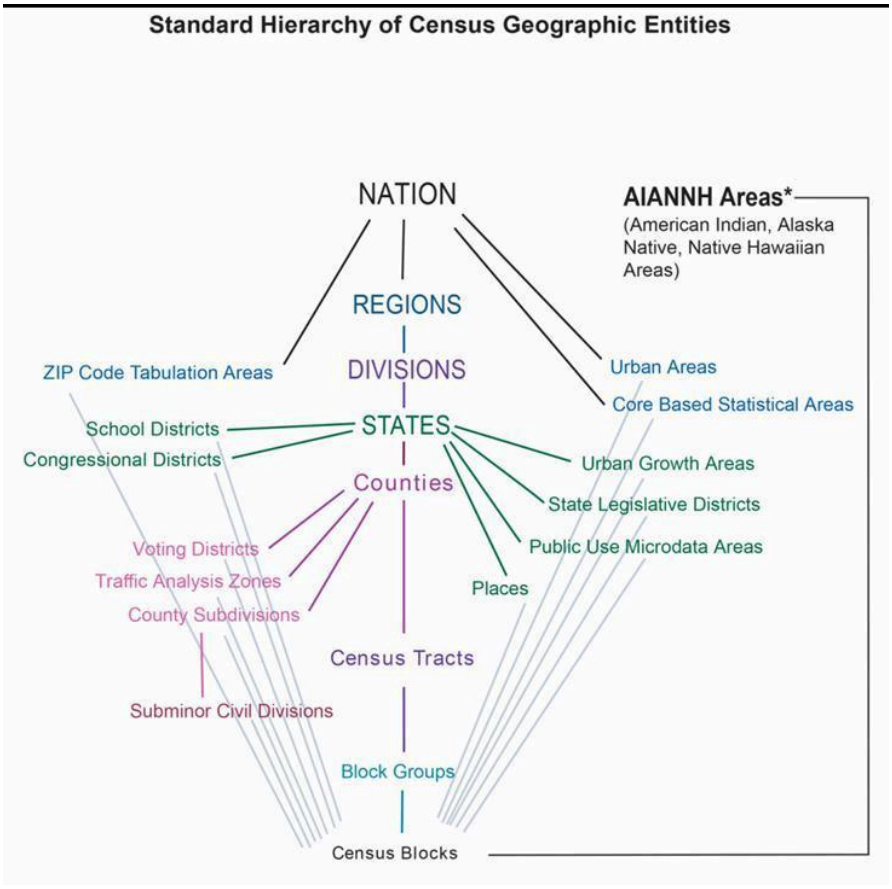
Unit of Analysis Challenge

Year Match

Survey Year v. Reported Year
The reported year is when the U.S. Census Bureau publishes or releases the data, while the survey year is the specific time period the data covers. Always check the survey year.

Geographical Match

Given the data, what is the unit of analysis?
There are multiple ways to divide US geography. Zip code is used by postal service. Census tract is mainly used by Census and policymakers. For economic indicator there is MSA or CBSA.



[image from : https://mcdc.missouri.edu/geography/sumlevs/]

Cite:

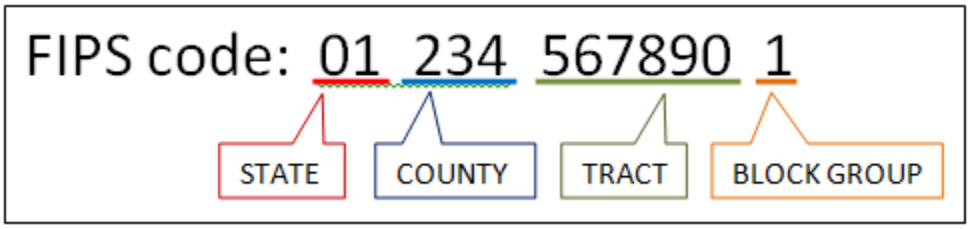
- https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html
- https://www.census.gov/newsroom/blogs/random-samplings/2014/07/understanding-geographic-relationships-counties-places-tracts-and-more.html

Knowing this is important because it tells you how to merge and concat each different dataset. Sometimes, we may want to learn voting behavior (voting district) and social environment (census tract). We may want to see small businesses performance (zip code), state policy (state), and local conditions (census tract).

From macro perspective, we may need to utilize certain conditions. County-level data include a lot of rural areas creating a lot of zeroes for business or entrepreneurship research. We may need to choose MSA or CBSA.
"CBSA stands for Core Based Statistical Area. It is a collective term defined by the U.S. Office of Management and Budget (OMB) and used by the U.S. Census Bureau to describe a geographic region containing at least one urban core of 10,000 or more people, plus adjacent counties with close economic and social ties." (Google AI)

States and Counties FIPS code

https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt
One way to merge across multi-level geographic information is through FIPS code.
11-digit FIPS codes include census tract information, combining the 5-digit state and county code with a 6-digit census tract identifier.



[image from: https://kodingdiary.wordpress.com/category/machine-learning/]

If you are able to get 11 digits, then state-level and county-level information can be merged using the first 5 digits. Currently (Aug 2026), only United States Geological Survey (USGS) is using GNIS code. Other database in US Census are using the FIPS Code.
Also, learn and get a qualitative evidence through interview with officials if policy, support, funding, and program is determined by census tract. For instance, I learned that Houston City Hall approval is based on the crime rate at census tract-level. They are actually looking at the census tract.

Census Tract Rule

The Census Bureau periodically updates census tract boundaries to account for demographic and population shifts. It changes every ten years. When an existing tract experiences significant population growth, it is retired and split into multiple smaller tracts
Reference: https://youtu.be/5ff8ezDmGOM?si=M7G_S7IL2DfRfSke

Column Name Challenge [IPUMS specific]

Column names are coded differently in NHGIS code style. Especially, this is relevant when downloaded from https://www.nhgis.org/
The image below is showing the population by race and ethnicity.

```
Data Type (E):
    Estimates

        NAME_E:      Area Name

    Table 1:     Hispanic or Latino Origin by Race
    Universe:    Total population
    Source code: B03002
    NHGIS code:  AF2U
        AF2UE001:    Total
        AF2UE002:    Not Hispanic or Latino
        AF2UE003:    Not Hispanic or Latino: White alone
        AF2UE004:    Not Hispanic or Latino: Black or African American alone
        AF2UE005:    Not Hispanic or Latino: American Indian and Alaska Native alone
        AF2UE006:    Not Hispanic or Latino: Asian alone
        AF2UE007:    Not Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
        AF2UE008:    Not Hispanic or Latino: Some other race alone
        AF2UE009:    Not Hispanic or Latino: Two or more races
        AF2UE010:    Not Hispanic or Latino: Two or more races: Two races including Some other race
        AF2UE011:    Not Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races
        AF2UE012:    Hispanic or Latino
        AF2UE013:    Hispanic or Latino: White alone
        AF2UE014:    Hispanic or Latino: Black or African American alone
        AF2UE015:    Hispanic or Latino: American Indian and Alaska Native alone
        AF2UE016:    Hispanic or Latino: Asian alone
        AF2UE017:    Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
        AF2UE018:    Hispanic or Latino: Some other race alone
        AF2UE019:    Hispanic or Latino: Two or more races
        AF2UE020:    Hispanic or Latino: Two or more races: Two races including Some other race
        AF2UE021:    Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races
```

<Image above is showing YR 2015>

```
Data Type (E):
    Estimates

        NAME_E:      Area Name

    Table 1:     Hispanic or Latino Origin by Race
    Universe:    Total population
    Source code: B03002
    NHGIS code:  ADK5
        ADK5E001:    Total
        ADK5E002:    Not Hispanic or Latino
        ADK5E003:    Not Hispanic or Latino: White alone
        ADK5E004:    Not Hispanic or Latino: Black or African American alone
        ADK5E005:    Not Hispanic or Latino: American Indian and Alaska Native alone
        ADK5E006:    Not Hispanic or Latino: Asian alone
        ADK5E007:    Not Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
        ADK5E008:    Not Hispanic or Latino: Some other race alone
        ADK5E009:    Not Hispanic or Latino: Two or more races
        ADK5E010:    Not Hispanic or Latino: Two or more races: Two races including Some other race
        ADK5E011:    Not Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races
        ADK5E012:    Hispanic or Latino
        ADK5E013:    Hispanic or Latino: White alone
        ADK5E014:    Hispanic or Latino: Black or African American alone
        ADK5E015:    Hispanic or Latino: American Indian and Alaska Native alone
        ADK5E016:    Hispanic or Latino: Asian alone
        ADK5E017:    Hispanic or Latino: Native Hawaiian and Other Pacific Islander alone
        ADK5E018:    Hispanic or Latino: Some other race alone
        ADK5E019:    Hispanic or Latino: Two or more races
        ADK5E020:    Hispanic or Latino: Two or more races: Two races including Some other race
        ADK5E021:    Hispanic or Latino: Two or more races: Two races excluding Some other race, and three or more races
```

<Image above is showing YR 2016>

Since the column name is coded differently, I used to check every column name and rename the column name in order to concatenate. Nowadays, you can ask AI to match the column names and code.

Race and Ethnicity Category Challenge

When we do the survey or submit applications, there is a section on racial or ethnic group. We can see similar thing in US Census data. Depending on research how you conduct, some may only focus on “ethinic group: Hispanic or Latino”. Other may focus on skin color, so they use “Race: White/Black”.

It is necessary to keep an eye on these two columns (ethnic group and racial group). If you want to identify “African American”, it is important to consider whether it is “Non-Hispanic & Race Black”.

Importantly, it is hard to assume that “category White” represents the image of “European White”. By definition, Census “white” includes “A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.”[1]. Why? Check out the Dow v. US supreme court case Sept. 15 1915 (https://en.wikipedia.org/wiki/Dow_v._United_States). This is why I don’t rely on “White” information for economic performance regarding economic equality. Also, they announced “The 2030 U.S. Census will officially introduce a distinct Middle Eastern or North African (MENA) category”. I expect that there may be some changes in results.

Furthermore, the Asian population is complicated. US Census defined that “Asian – A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan,

[1] https://www.census.gov/topics/population/race/about.html

the Philippine Islands, Thailand, and Vietnam." There are various ethnic backgrounds in Asia and different immigration histories in the US. It is hard to generalize based on the Asian group survey information. This is also similar to the Hispanic population.

Disclosure Avoidance Rule

"The Census Bureau has always faced a challenge: how to publish useful statistics while protecting the confidentiality of the data and the people behind the numbers."

You may encounter the information shown as 'D'. This is **Cell Suppression**. They do not disclose that data point because of privacy issue. For instance, if there is only one Asian supermarket owners in one county, then it is easy to identify in US Census (income, revenue, location, zip code, loan application, etc.) This is the concern of privacy.

Cell suppression is a disclosure avoidance technique that protects the confidentiality of individual survey units by withholding the values of certain cells within a table from release (https://www.census.gov/topics/business-economy/disclosure/about.html)

Recently, we have seen major changes with Disclosure Avoidance Rule. There are two articles regarding the rule changes.

June 2026 [Trump privacy restrictions may reduce Census Bureau data : NPR]
The US Government ordered regarding data privacy. It is likely to be dffiicult to gather neighborhood-level information or rural area information in the future.

New information Aug 2026 [Understanding the New Disclosure Avoidance Policy: What It Means for Data, Confidentiality and You]

The administrative order allows two methods of disclosure avoidance:

1. *Coarsening.* Reducing the detail of data (for example, grouping ages into ranges or rounding). The new order names coarsening as the preferred disclosure avoidance method for all statistical products.
2. *Suppression.* Withholding certain data from publication. This is permitted only as a last resort.

Under the administrative order, a third method of disclosure avoidance called noise infusion is no longer allowed.

3. *Noise infusion.* Adding random changes to data values. This includes statistical methods such as record swapping and the use of synthetic data.

… Coarsening and suppression can reduce the amount and detail of data published, especially for small geographic areas or population groups. These are the places where confidentiality risks are highest.

What is impacted YR2030?

- *2030 Census*. Decennial census data are the hardest to protect due to the data's scale and detail. This new order has shifted the research agenda, but our key objectives remain the same: better assessment and communication of disclosure risks and better understanding of data users' needs. We plan to release a demonstration data product in mid-2027 for public feedback that will inform decisions for census testing and ultimately the 2030 Census.
- *ACS*. We're researching new disclosure avoidance methods. Transitioning to methods compliant with the new order will occur gradually, with a full transition by 2029.
- *Demographic surveys*. Many of our demographic surveys use noise infusion methods. We will perform disclosure risk assessments for these products and research and identify new methods that are compliant with the order.
- *Economic indicators*. Most of our economic indicators are protected by suppression and won't be affected. The Quarterly Financial Report is affected and will switch back to suppression.
- *County Business Patterns*. These data have historically used noise infusion, and we are exploring alternatives.
- *Longitudinal Employer-Household Dynamics (LEHD)*. These products use state unemployment insurance data as inputs. Products produced with noise infusion required by states in our data-sharing agreements will continue until these agreements are updated.

2. Business Information: What kind of business data I can see?

ABS, ZBP (CBP)

Business information is vulnerable to the disclosure avoidance. Especially, in the small town and rural area, small numbers of businesses are out there.

Ironically, when I want to study about disinvestment and economic development for certain areas, I face a lot of limitations to understand what is going on out there and what is really working for those underserved communities.

I use CBP over ZBP because zip-code level data hardly shows details like business income or racial categories. CBP is more likely to provide detailed categories. However, it is still limited to understand rural counties.

Also, ABS information is based on employer businesses, which means “a business that has paid employees and reports an annual payroll to the federal government”. For minority group or underserved population, they are more likely to rely on “Nonemployer Business: A business that has no paid employees and earns at least $1,000 in annual receipts.” and “A legal category for an unincorporated business owned by a single individual.”

The ABS is an annual survey that collects demographic characteristics from employer businesses. However, the ABS excludes the collection of demographic data from nonemployer businesses. The NES-D was developed to produce similar estimates as ABS on owner demographics for nonemployer businesses. The NES-D is not a survey; rather, it leverages existing individual-level administrative records to assign demographic characteristics to the universe of nonemployer businesses. [Nonemployer Statistics by Demographics (NES-D) Data]

X. Data Coding

During the non-AI era (*Back in my day!*), I used Python for data wrangling, matching, and cleaning. I used R for detailed cleaning for certain tasks: datetime and character matching. I use the main regression analysis, table, and graph with STATA. I did three separate programs because, in reality, certain program library/coding is better at certain tasks. Also, depending on RAM or data storage, I need to use whatever is available.

I did my data wrangling mainly during YR 2021-2024 which is before major AI development. But I don't think using more than two programming languages is no longer necessary. Nowadays, code writing and 'translating' can be done through asking AI. For instance, "translate this R code to Python code" and do trial and error. I am aware that ChatGPT is, for instance, giving a random code library. I would recommend finding the most commonly used and confirmed code library.

Also, there are more advanced ways to construct data analysis. I would like to leave room for future researchers.