Introduction to IPUMS and NHGIS

Katie Genadek
Minnesota Population Center
University of Minnesota
kgenadek@umn.edu

The IPUMS projects are funded by the National Science Foundation and the National Institutes of Health

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

My name is Katie Genadek and I am a research associate and the director of outreach dissemination for all the MPC data projects. My plan is to give you an overview of the data we have available and what we do and why it's so special. Then, I'll show you how we get the data. Next, I'll walk you through three research questions and how I would handle them on the website and what I'd do with it. Then, I'll have you work through more exercises, so you can play around on the computer with the data.

To give you a bit more detail about the variety of data projects at the MPC, this is a screenshot of popdata.org, from which you can link to any of our data projects. The Minnesota Population Center is a misleading name. We are a population center like many other population centers around the country, but we are different in that we make all of this data available freely to users around the world. These are all of the MPC data projects that you can find on popdata.org.

- I am going to focus primarily on IPUMS USA today, specifically the American Community Survey (or ACS) because that's what a lot of people use. I am also going to talk about NHGIS (the National Historic Geographic Information System), another tool for accessing ACS data as well.
- We have the world's largest selection of international data available—currently data from 78 countries.
- Current Population Survey (CPS) data from the population survey is similar to the ACS in that is given out every single year. There is more information on employment. It's a smaller sample size so if you guys are doing more local things it can be harder to use.
- Finally there's a bunch of different survey data both international and local to the US, and a lot of historical data.

The Minnesota Population Center

- University-wide interdisciplinary research
  - Members from 23 departments
- Education and Training
  - Demography curriculum
  - Professional and research training
  - Seminars and workshops
- Revolutionary Data Tools
- Home of the IPUMS and other MPC data projects
- More than 200 employees

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

The MPC is an interdisciplinary research center. We have an education component, but primarily we are the home of the IPUMS and other MPC data projects. We currently have more than 200 employees. So we are pretty big— especially for a university center that's kind of rare. A lot of the people who work there are students (about 60), but everyone else works full time.

What the MPC data projects do:

- Standardizes data across years (integration)
- Provides excellent documentation
- Provides "bonus" calculated fields
- Provides data in multiple formats
- Makes data freely available online
- Provides user support

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

What we do is we standardize data across years. This is our bread and butter—we call it integration.

- We take data sets and we make them so they are easy to use over time and over place.
- We provide excellent documentation. That's a big part of what we do.
- As we learn the data, we try to make anything we find out about available on the web for you. So we load in the metadata for you, then we provide "bonus" calculated fields. We do things to make it easier for you to use like:
  - Putting the CPI right in the database so you can take it.
  - Calculating within the household number of children.
- We provide data in multiple formats. Via our export system, you can get the data in SAS , SPSS, Stata, and Excel.
- We also have an online analysis system that we will use today so you can look at the data online.
- This is all free. We are primarily funded through the National Institute of Health (NIH) and started with funding from the National Science Foundation (NSF) and have a few other supporters as well.
- Finally we provide user support. I manage a user support team. There's a group of people who answer emails for you within 24-48 hours. The easiest way to get help is to email ipums@umn.edu. We are happy to look into other data too, especially in the social science data world.
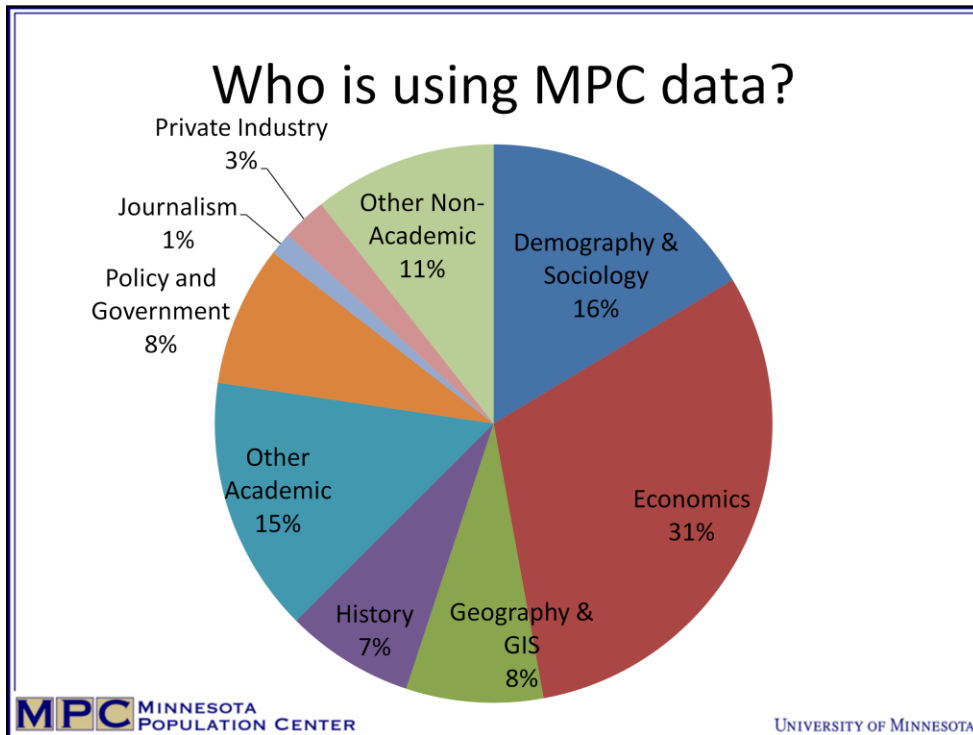
Who is using MPC data?

- Academic researchers, which makes sense because we started at the university.
- Economists; this is where I come from. I started using this data when I was working on my Master's Degree.
- Demography and sociology.
- Policy and government and journalism is our biggest group of growing users.
- People now have more computing power which is needed to handle this big data. With more computing power, it is possible and profitable to use this big data in many other settings.

SUMMARY DATA

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

I'm going to cover the summary data first and then get on to the microdata.

Summary Data

Age	Both sexes	Male	Female
Total population	281,421,906	138,053,563	143,368,343
Under 5 years	19,175,798	9,810,733	9,365,065
5 to 9 years	20,549,505	10,523,277	10,026,228
10 to 14 years	20,528,072	10,520,197	10,007,875
15 to 19 years	20,219,890	10,391,004	9,828,886
20 to 24 years	18,964,001	9,687,814	9,276,187
25 to 29 years	19,381,336	9,798,760	9,582,576
30 to 34 years	20,510,388	10,321,769	10,188,619
35 to 39 years	22,706,664	11,318,696	11,387,968
40 to 44 years	22,441,863	11,129,102	11,312,761
45 to 49 years	20,092,404	9,889,506	10,202,898
50 to 54 years	17,585,548	8,607,724	8,977,824
55 to 59 years	13,469,237	6,508,729	6,960,508
60 to 64 years	10,805,447	5,136,627	5,668,820
65 to 69 years	9,533,545	4,400,362	5,133,183
70 to 74 years	8,857,441	3,902,912	4,954,529
75 to 79 years	7,415,813	3,044,456	4,371,357
80 to 84 years	4,945,367	1,834,897	3,110,470
85 to 89 years	2,789,818	876,501	1,913,317
90 years and over	1,449,769	350,497	1,099,272

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

This is what summary data looks like. This sex/age breakdown is from the 2000 census for the entire U.S. This is what we picture when we get summary data: data displayed in tables. A little bit different than microdata. The summary data is aggregated, while the microdata includes the individual records allowing for analysis for more refined categories. For example, this table shows age grouping in specific age groupings. What if you wanted to look at children under 18? With the microdata, you can do that analysis.

But before getting to the microdata, I want to give you an overview of our other data. NHGIS is similar to FactFinder but in addition we have block group and block data, which I believe they do not have in FactFinder. And we also have the GIS files available for all of these areas. The biggest thing is we go back to 1790.

We have U.S. Census aggregate data from 1790 to 2013, population and agriculture stuff, and all of the electronic summary files of the 2010 census. And finally we have the American Community Survey Data. This is some of our newest data. We have all the one, three, and five year data files that come out and make them available here.

More NHGIS Data

County Business Patterns, 1970 – 2002

Many other unique datasets
- FDIC bank deposit data, 1920 – 1936
- U.S. unemployment data, 1937
- Censuses of Religious Bodies, Farm Real Estate Values and more

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

We also have some more interesting stuff—county business patterns and other unique historical data.

This is what it looks like. Here is a screen shot of all of our ACS data. We now have 2013 ACS data. You can get this for the one year, three year, and five year data—right there in all the different tables. There's a lot of data in there.

GIS Shapefiles

- GIS boundary files, 1790-2012
  - Nation, State and County boundaries, 1790-2012
  - Census Tract boundaries, 1910-2012
  - Metropolitan region boundaries, 1950-2012
  - All other released Census geographies, 1980-2012
    - Blocks
    - Block Groups
    - School Districts
    - Urban Areas
    - Places
    - County Subdivisions
    - Congressional Districts
    - American Indian Areas
    - and several more

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

For the boundary files, we have everything all the way back to 1790 and going to the most current stuff, down to the block. When you get your data you can automatically attach a state file.

This site was used regularly during the government shutdown, since users could not download shapefiles from government websites. Many people continued to use this site because it's very convenient.

Time Series Tables

- Harmonized census tables
- Multiple years of data in one file
- 197 tables

We also have time series tables for when you want to look at areas over time. I've learned more and more about geography in the last year or two working with geographers. It's crazy because geography changes as population changes; some geographic areas stay the same, but others are moving. So we've gone through and tried to make the most comparable geography over time. Basically, we layered geographies on top of each other to get a comparable area. This can be really nice and it's the easiest way to look at change over time.

Learn More About NHGIS

- Online tutorials
  http://youtu.be/P1znKKm8vX4

- Online workshop materials
  https://www.pop.umn.edu/data-user-resources/data-support

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

To learn more about NHGIS we have online tutorials and workshop materials.

MICRODATA

MPC MINNESOTA
POPULATION CENTER

UNIVERSITY OF MINNESOTA

On to the microdata…

Microdata versus Summary Data

Microdata	Summary Data
■ Shows full range of responses for individuals	■ Premade or published tables of aggregate characteristics
■ Enable custom tables and sophisticated analyses	■ Enable examination of small geographic areas
■ Suppression: geography, truncation, and item level suppression	■ Suppression: limited content, grouped intervals, and cell suppression

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Microdata shows the full responses for individuals. This allows you to create custom tables, to do sophisticated analyses. People use this to run regression. However, unlike the aggregate data, where you can get data for small geographic areas (like the block level or for a zip code or a county), with microdata, you are limited to the Public Use Microdata Areas (or PUMAs), which are areas of 100,000 people.

Microdata provides information about individuals and households.

This looks a bit intimidating—don't worry, this is what the data we work with looks like but users get data formatted for use in a stats package. Each line is an individual, and each column or group of columns is a variable, organized in households. Microdata enable the creation of custom tables and tabulations and sophisticated statistical analyses.

This is what microdata looks like. This is a person in microdata (single row). And then you have a bunch more people (more rows). And within these people all these different columns are different variables. This is taken from the 1910 census. You can see here this first person with 1000 under "Relation to head" is the man in the household (signified by the 1); 2 is the wife and 4 signifies children. The marital status there is 22 for people who are married. The first person in the household is literate; the next ones are not. This is what microdata looks like in the underneath terms. However, we make it a lot easier to use through statistical packages.

What is IPUMS?

Integrated - consistent codes, labels, and documentation

Public Use - anonymized, downloadable

Microdata - individual-level

Series - pooled data over time and place

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

We take this microdata and we integrate it. That's what the I in IPUMS stands for. We have been around since the early 90's. We predate the iPod. Part of the reason it's IPUMS is because it's for Integrated and Public Use Microdata Series.

The Public Use Microdata Sample (PUMS) was made available by the Census starting in 1960 and they went back and did 1940 and 1950. The Census Bureau released Decennial census data every ten years. We integrated them into the Integrated Public Use Microdata Series by making consistent code labels and documentation.

INTEGRATION

I'm going to explain to you exactly what we do with integration here.

Variable Harmonization
Marital Status

Bangladesh 2011	Mexico 1970	Kenya 1999
1 = Unmarried	1 = Married, civil & relig	1 = Never married
2 = Married	2 = Married, civil	2 = Monogamous
3 = Widowed	3 = Married, religious	3 = Polygamous
4 = Divorced/separated	4 = Consensual union	4 = Widowed
	5 = Widowed	5 = Divorced
	6 = Divorced	6 = Separated
	7 = Separated	
	8 = Single	

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

I'm going to use an example from IPUMS international because it's a lot more fun.

We have data from Bangladesh. In 2011, they asked people in Bangladesh if they were unmarried, married, widowed, or divorced/separated.

In Mexico in 1970 you can see there were eight different codes for different kinds of marital status. They break down married into three different kinds: 1) civil & religious, 2) civil, and 3) religious. They have a consensual union and widowed.

You can also see right away that the codes are different than Bangladesh. For example, 1 is unmarried in Bangladesh when 8 is single for Mexico. So if you want to do analyses of these two samples, you have to recode the data. And that's what we're going to do for you.

Finally for Kenya in 1999, we have polygamous marriage identified.

Translation Table

Input

Bangladesh 2011	Mexico 1970	Kenya 1999
1 = Unmarried	1 = Married, civil & relig	1 = Never married
2 = Married	2 = Married, civil	2 = Monogamous
3 = Widowed	3 = Married, religious	3 = Polygamous
4 = Divrc or separated	4 = Consensual union	4 = Widowed
	5 = Widowed	5 = Divorced
	6 = Divorced	6 = Separated
	7 = Separated	
	8 = Single	

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

We use something called a translation table to do this so you don't have to do it in code. Here's Bangladesh, there's Mexico, and there's Kenya. We get all these raw codes and we use this as input data.

Translation Table

Harmonized		Input		
Code	Label	Bangladesh 2011	Mexico 1970	Kenya 1999
		1 = Unmarried	1 = Married, civil & relig	1 = Never married
		2 = Married	2 = Married, civil	2 = Monogamous
		3 = Widowed	3 = Married, religious	3 = Polygamous
		4 = Divrc or separated	4 = Consensual union	4 = Widowed
			5 = Widowed	5 = Divorced
			6 = Divorced	6 = Separated
			7 = Separated	
			8 = Single	

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

Then we create a harmonized output code.

Translation Table

Harmonized		Input		
Code	Label	Bangladesh 2011	Mexico 1970	Kenya 1999
		1 = Unmarried	8 = Single	1 = Never married
		2 = Married		
			2 = Married, civil	
			3 = Married, religious	
			1 = Married, civil & relig	
				2 = Monogamous
				3 = Polygamous
			4 = Consensual union	
		4 = Divrc or separated		
			7 = Separated	6 = Separated
			6 = Divorced	5 = Divorced
		3 = Widowed	5 = Widowed	4 = Widowed

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

We match and line up the codes for each survey.

Translation Table

Harmonized		Input		
Code	Label	Bangladesh 2011	Mexico 1970	Kenya 1999
1 0 0	Single	1 = Unmarried	8 = Single	1 = Never married
2 0 0	Married or in union	2 = Married		
2 1 0	Married, formally			
2 1 1	Civil		2 = Married, civil	
2 1 2	Religious		3 = Married, religious	
2 1 3	Civil and religious		1 = Married, civil & relig	
2 1 4	Monogamous			2 = Monogamous
2 1 5	Polygamous			3 = Polygamous
2 2 0	Consensual union		4 = Consensual union	
3 0 0	Divorced or separated	4 = Divrc or separated		
3 1 0	Separated		7 = Separated	6 = Separated
3 2 0	Divorced		6 = Divorced	5 = Divorced
4 0 0	Widowed	3 = Widowed	5 = Widowed	4 = Widowed

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

And we do that by recoding and moving these around so that now everyone who has never married is getting a code of 100. Instead of writing code to do this, we are doing it for you, and we are doing it in a very systematic way so that we can see exactly where the codes are going. Again, here's all the different kinds of marriage, and you can tell right here we try to do this without losing any information.

So if you just want to look at "married", you can use any one with a 200, but we use 210 (that second digit) to show formal marriage, and 220 for a consensual union. Our third digit shows an even at a finer version of marriage.

Harmonized

Code	Label
1	Single
2	Married or in union
3	Divorced or separated
4	Widowed

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

At the end, the user sees the final set of codes. So you don't have to deal with all that other messy code. And if you just want to have your general code, you can just use this. So if you just want to look at married across, we make it easier for you to do that. But if you're really interested in polygamous marriage, for example, you can do that by using the more detailed codes shown on the previous slide.

Integrating Documentation

- Sample Descriptions
- Variable Descriptions
  - Availability by Sample
  - Universes
  - Comparability
  - Allocation and Imputation Flags
  - Questions and Instructions to Respondents
  - Instructions to Enumerators

In addition to integrating the data (that's our big thing), we also have a lot of metadata and we integrate documentation so it's all the same. We have sample description, and we also have a lot on variables.

- We have the availability.
- We state the universe: who gets asked a question for every single question in every single sample.
- We provide comparability statements, to guide using data across time and place, especially when there's a question change.
- And the biggest thing is we have all the questions and instructions to respondents. So this way you can actually look and easily compare how the question is asked across time which can greatly impact the results.
- The other big thing we have is allocation and imputation flags, if you really start digging down into the data. Especially in some place like ACS, they've changed how they've allocated and included things over time. You may see something in your data that looks like this big change, when really they've just changed how they are allocating missing values.

Additional Enhancements to the Data

- Integrated Variables

- Geographic Areas

- Consistent industrial and occupation coding schemes

- Constructed family interrelationship variables

MPC MINNESOTA POPULATION CENTER — UNIVERSITY OF MINNESOTA

In addition to integrative variables, we've made integrated geographic areas.

- PUMAs have changed over time. For example, for San Diego county, there were 22 PUMAs based on the 2010 Census, while there were only 16 using the 2000 census geographies. We try to combine them to create metro areas, to create cities, and to create counties in the U.S. We do this internationally as well.
- We have consistent industrial and occupational coding schemes (based on international and U.S. coding schemes), which are applied over time so you can look at occupational data over time with the same codes.
- Finally we make constructed family interrelationship variables. This is again related to who we are as a center. As I mentioned, we are run by historians but we work with genealogy, so we really care about these and really wanted these. We go into the data for you and count the number of children. We give you the location of childrens' parents because you can have multiple families within households. We've predicted who we think the step mom is and who is not, and things like that over time. This kind of information is really useful for looking at families.

www.ipums.org/usa

This is our website: www.ipums.org/usa. We are working to make it look fancier, but the data is there. It's not super obvious, but you want to start with "Select Data" and start digging into our data there.

Microdata Data:

IPUMS-USA

I'm going to tell you about what's actually in the data, now that we've integrated it.

Starting with the U.S.A…

- We have U.S. decennial censuses going back to 1850. We don't have 1890 because those records were burned in a fire. Some data we entered ourselves. Other people throughout the U.S. entered other data and we collected it and created files. Some data were originally on microfiche reels and were entered back though the WPA program during the 30s.

- We have a complete-count dataset for 1880. This is done in collaboration with the Church of Jesus Christ of Latter-day Saints, the Mormon Church. They have been working on the 1880 census for genealogy purposes for a long time but were having a little trouble. So, we merged forces and helped them and they helped us by making it available to researchers.

- We've linked samples from 1850-1930.

- In December 2014, we received a full count dataset for 1940. This is the most recent census data available. Census data becomes public 72 years after the surveys were filled out. We worked with ancestry.com on this as they digitized the data and we paid them to digitize a little more data and now we have this data available for researchers.

- Finally we have samples from Puerto Rico and the American Community Survey starting in 2000 through 2013. When the Census Bureau releases the data, we generally have it up within a week or two.

This is our first MPC data project and we have over 45,000 users using this data.

The American Community Survey

- Replaced the long form of the Decennial Census
  - Demonstration stage: 2000 to 2004
  - Full implementation 2005, group quarters added 2006
- Rolling sample design

Microdata samples:

- Full survey responses for 1% of US population
- Yearly samples, multi-year samples

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

The American Community Survey replaced the long form of the Decennial Census back in 2000.

- The demonstration stage was 2000 to 2004. They actually did testing before that but we don't have data from that. There are slightly smaller samples in 2000, 2004, but full implementation began happening in 2005 with reports added in 2006 (e.g. institutions).
- Unlike the decennial census, this is a rolling sample design. This is the biggest difference between the ACS and the decennial census long form. Questions are nearly identical, they add some things, take some things away. But the biggest difference is this rolling sample and we currently do not know the specific date when you took it. You could've taken it anytime throughout that year. I'm pushing hard to have this information added.
- The microdata are 1% of the population. They give the survey to about 3% of the population a year. They get about a percent and a half back of the population, which is what is used to make the Factfinder numbers. And then make a 1% version available in microdata. It's okay if it's slightly different than you're the Factfinder results because it actually is a little different.
- There are also multi-year samples. For the microdata, multiple years doesn't really get you that much. The years are just put together with the weights adjusted and the income adjusted with the most recent years. We don't get any finer levels of geography. The only advantage is the larger sample size.

Census and ACS Household-Level Variables

- Geography
- Group quarters
- Housing characteristics
- Household economic characteristics
- Household appliances

For household level variables, the following are included geography, household status, household income, and questions about what type of appliances you have in the home, Internet, computer, air conditioner, etc.

Census and ACS Individual-Level Variables

- Age
- Race
- Marital Status
- Birthplace
- Nationality
- Family Interrelationship

- Education
- Work
- Income
- Disability
- Migration Status
- Commuting
- Health Insurance

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

For individual-level variables, the basic characteristics are included, anything you'd expect in a census: marital status, birthplace, nationality, family interrelationship, education, work, income, disability, migration status, commuting, and now health insurance.

Geography Limitations

- No confidentiality restrictions for samples prior to 1950 – no geographic limitation
- Samples from 1950-1970
  - Inconsistent geographic identifiers
- Recent samples:
  - State
  - Some Metropolitan Areas
  - County Groups
  - Public Use Microdata Areas (PUMAs)

Prior to 1950, you can see exactly where people lived; no confidentiality restrictions there. For the samples from 1950 to 1970, the geographic identifiers are pretty limited. For more recent samples, you can get big metropolitan areas (most of them); you can get groups of counties in 1980, and more recently you can get PUMAs.

What are PUMAs?

- Public Use Microdata Areas (PUMAs)
- Comprised of approximately 100,000 persons
- Boundaries do not always align with jurisdictional boundaries
- Detailed contents and maps available
- GIS shape files for PUMAs available

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

What are PUMAs?

- PUMAs are a made-up geographic area comprised of approximately 100,000 people.
- Boundaries do not always align with jurisdictional boundaries. They often do, especially in California.
- PUMAs change about every 10 years. Following the decennial census, we reevaluate the PUMAs. Demographers do this, which is why they also vary state to state.
- There are some guidelines, and I believe that for the new one in 2013 (really came out in 2012), they're supposed to be related to MSAs. Sometimes they are built in different boundaries, but it doesn't always work.
- We have detailed content and maps available. We created an online interactive map so you can zoom in and see your PUMAs and how they change from 2000 to 2010 to 2012.
- We also have GIS files available for PUMAs. So you can map them yourself in addition to looking at ours.

Sample Expansion

- 1850 full count dataset – in collaboration with the Church of Jesus Christ of Latter-day Saints

- 1960 data recovery and 5% sample creation

- 1790-1930 full count data

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

More of what's going on at MPC…

- We have another project with the Mormons, we're almost done with 1850.
- We have a 1960 data recovery and 5% sample creation that'll be coming out within the next month. It's been ongoing that we've had records that were lost. That was in collaboration with the Census Bureau.
- We also have 1790 to 1930 full count microdata that we are working on. This was in collaboration with ancestry.com. People were scraping their website for data and it was hard for them and they were frustrated, so they worked with us to create this for researchers. They don't have names, so you can't use this for genealogy, but you can use it for research.

Obtaining Data:

DATA EXTRACTION

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

In addition to making all this data possible across time, we have a pretty cool data extraction system.

Online Extraction System

- Users create custom data files
  - Pick any samples of interest
  - Pick any variables of interest
- Creates custom syntax for reading the data files into SPSS, Stata, SAS, and CSV
  - Labels variables and values within the data
- Codebook available
- Record of extract is preserved on user account

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

You can create a custom data file looking at any samples and variables you want. Instead of taking the ACS 1% PUMS file and downloading the whole thing and reading it in, you can get only what you want. If you just want your one ACS file and you just want three things (like your PUMA, age, sex, race, and how people are commuting), you can just grab those. So you can get as many or as few variables and samples as you want.

- And then we create a custom data file for you with SPSS, Stata, SAS, or CSV version or Excel.
- All of your variables are labeled and all of your values are labeled within your data; it's all pretty clean.
- You have a custom codebook attached to it that you can look at and it comes with your file.
- And you have a record of your extract.
- We make the data extract for you and hang it on the website. You have it for about 72 hours or up to 5 days that you can download it and you have a record of it so you can go back and recreate it or add variables to it. So you don't have to worry about getting it right once; you can just keep doing it.
- We email you when your data is ready. If you're just looking at one or two years of ACS, then it can be ready in about a minute or two. If you grab a bunch of data or a bunch of people are using it, it can take a half hour.

Online Extraction System

- Additional Features:
  - Case Selection
  - Attach Variables
  - Custom Sample Size
- Learn one data set, learn how to use all MPC microdata
  - Documentation systems
  - Extract systems
  - Data files produced

Additional features…

- We also have case selections, so if you're only interested in looking at children or women or states, then you can just go in and select that one state.

- You can attach variables to people within the household. This is a little complicated. But let's say you're interested in looking at spouse age differences. And you're really interested in the set of people who are more than ten years apart. Within our system you can attach a spouse's age to each other, so that they are just on one line within your data analysis package. You can easily create an age difference without having to loop through stuff and do that yourself.

- Finally you can create a custom sample size. So if you just want to quickly grab of small number of cases, or you want to grab a bunch of people from every year or not that many, then we will do that within our system and re-weight it for you.

From one data set you can learn how to use them all.

Online Analysis System

- High-speed tabulation software developed at UC-Berkeley
- Allows for analysis of microdata without statistical package
- All analysis performed online
- Can analyze multiple years of data
- Help guides on webpage

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

I mentioned the online analysis system that we are going to look at today. There's a high-speed tabulation software that looks at all this data online. This allows you to look at microdata without a statistical package and all the analyses are performed online—it's incredible. And you can analyze multiple years of data, so you can look at change over time within it. This is what a lot of our journalists use. There are help guides on the webpage.

Obtaining Data:

USING MPC DATA

MPC MINNESOTA
POPULATION CENTER

UNIVERSITY OF MINNESOTA

Video Tutorials
http://www.youtube.com/user/MPCIPUMS
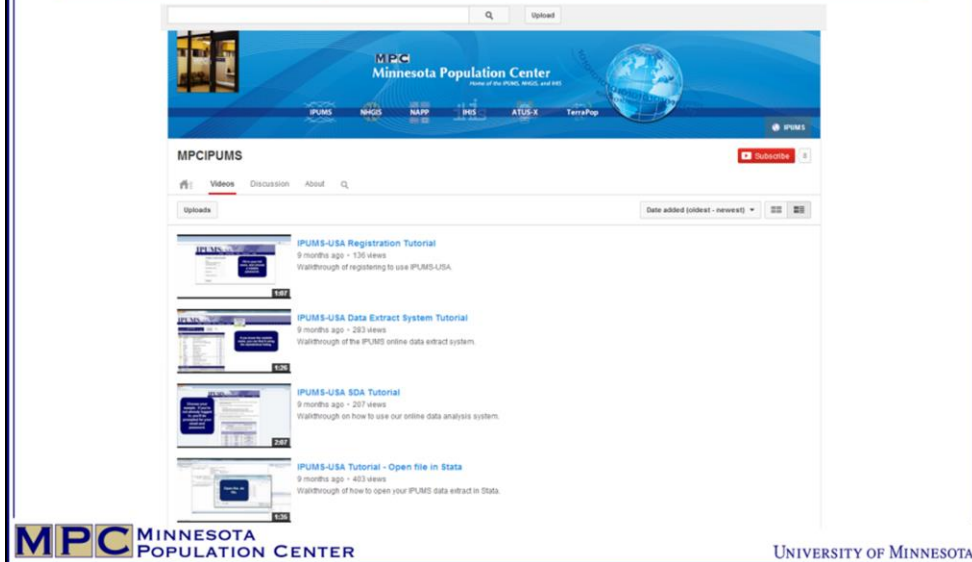
We have Youtube videos at this link that are short and include a lot of different tutorials and different aspects of our website.

Why people use MPC data

- Comprehensive online documentation

- Integration makes analyzing change over time possible

- Data analysis system allows you access the data and analyze it online

- All of the data are available for **free** online

- User support is available by e-mail to help you as needed

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Why people use this: You can look at it online and it's all free.

Using MPC data

- Cite the data – check homepages for citation information

- Enter your work in our bibliography
  - IPUMS Research Award

- Data Questions
  - User forum (http://answers.popdata.org/)
  - Email IPUMS user support (ipums@umn.edu)

- Data Errors
  - Get an IPUMS mug!

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

When you use the data, we request two things: cite the data and enter your work into our bibliography.

- Citing the data can be a little hard with reports, but we have a short citation and a long citation.
- Our bibliography is how we try to keep track of everything that's going on, and we have a Research Award available.

If you have data questions, you can email us. We also have a user forum where There are tons of commonly asked questions.

Finally, if you find data errors, you get an IPUMS mug. We encourage you to look for errors in the data. Please let us know and we'll reward you with a mug.

Using the extract system and documentation

- Variable Description
- Online Analysis

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

Before we start, we are going to watch a couple tutorials.

- This first one shows how to navigate variable descriptions because I think it is important to learn about the data before using it: **https://www.youtube.com/watch?v=3Xq1jl9IoV4**
- This second tutorial walks through the process of using the online analysis tool: **https://www.youtube.com/watch?v=E4_YY2xr3Lc**

Analyzing IPUMS Data

- Register for IPUMS-USA
- Three research questions
- Hands-on workshop exercise
- Further workshop materials:

https://www.pop.umn.edu/data-user-resources/data-support

MPC MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Before you can access the data, you need to register for IPUMS USA. It's an instant registration. After answering a couple questions, you automatically have access to the data. Almost all of our projects are like that. After you register, we'll go through the three research questions.

Create Login

To access IPUMS, you need an account.

In google, type ipums, or go to **usa.ipums.org.**

Click on iPUMS usa. From there go to IPUMS Registration (far left).

Click on Apply for access. Enter your email under Request an Account and follow the steps. Write in whatever you'd like for Research Projects (and only has to be 25 words, not 50 like it indicates).

Example research questions, methods, answers

- Research Question: How has the racial composition of California changed since 1980?
  - How was race measured in each year?
  - What are options for comparing race over time?

MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

This question is a new one for me because I think it's healthy to see what happens when you're doing this. There are definitely funny things with the microdata that come up when you're using it and every time I use it I learn something new about it and I use it every day.

The first question is, "How has the racial composition of California changed since 1980?" Right off the bat, the first things I'm thinking are, how was race measured in each year (because since 1980 there have been a lot of changes in measurement), and what options do we have for comparing it over time?
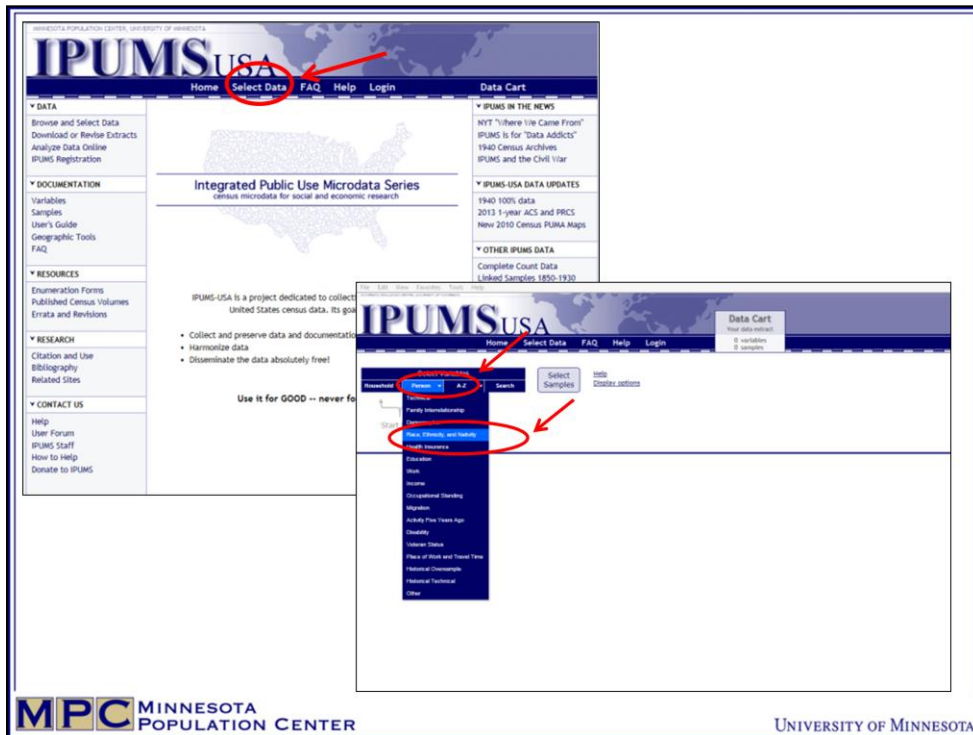
So I am going to go to the website and…

- Click <u>Select Data</u>. Again, this is the place where we just dig in and learn a lot.
- Go to <u>Person</u>
- and then <u>Race, Ethnicity, and Nativity</u>

- Go to <u>Race</u> because that's what we are wondering about.
- Then let's just go straight to the <u>Codes</u> page and look.

And you see here how the race variable has been coded across samples.

Now, let's choose the samples we want.

- Go to Select samples at the top of the page.
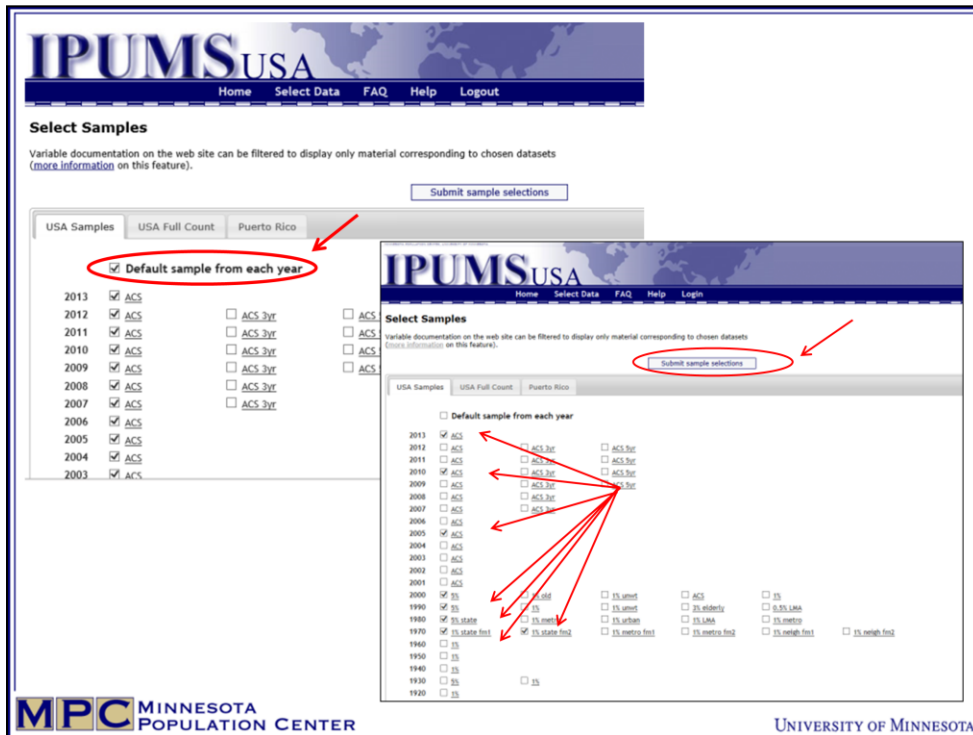
Our question asks about over time since 1980, instead of taking all the samples, which is the default…

- Uncheck default sample for each year
- Choose the ones of interest. Let's take 2013, 2010, 2005, 2000, 1990, 1980, 1970 1% state fm1, and 1970 1% state fm2.
- To complete the selection, click on Submit sample selections

Now, let's look at the Codes tab again.

The list now reflects the specific samples chosen. You can see right away a big difference on the general codes. Starting in 2000 they added two major races or three or more.

For more details, click on Detailed Codes and you can see that things have changed quite a bit.

By scrolling down, you'll see that, starting in 2000, they started getting tribal identities for American Indians.

Code	Label	2013 acs	2010 acs	2005 acs	2000 5pct	1990 5pct	1980 5pct	1970 sta2	1970 sta1
643	Nepalese	X	.	.	.	.	.	.	.
650	Other Asian or Pacific Islander (1980)	.	.	.	.	.	X	.	.
651	Asian only (CPS)	.	.	.	.	.	.	.	.
652	Pacific Islander only (CPS)	.	.	.	.	.	.	.	.
653	Asian or Pacific Islander, n.s. (1990 Internal Census files)	.	.	.	.	.	.	.	.
660	Cambodian	X	X	X	X	X	.	.	.
661	Hmong	X	X	X	X	X	.	.	.
662	Laotian	X	X	X	X	X	.	.	.
663	Thai	X	X	X	X	X	.	.	.
664	Bangladeshi	X	X	X	X	X	.	.	.
665	Burmese	X	.	.	.	X	.	.	.
666	Indonesian	X	X	X	X	X	.	.	.
667	Malaysian	X	X	X	X	X	.	.	.
668	Okinawan	.	.	.	.	X	.	.	.
669	Pakistani	X	X	X	X	X	.	.	.
670	Sri Lankan	X	X	X	X	X	.	.	.
671	All other Asian, n.e.c.	X	X	X	X	X	.	.	.
672	Asian, not specified	.	X	X	X	.	.	.	.
	Two or more Asian races								
673	Chinese and Japanese	X	X	X	X	.	.	.	.
674	Chinese and Filipino	X	X	X	X	.	.	.	.
675	Chinese and Vietnamese	X	X	X	X	.	.	.	.
676	Chinese and Asian write_in; Chinese and Other Asian	X	X	X	X	.	.	.	.
677	Japanese and Filipino	X	X	X	X	.	.	.	.
678	Asian Indian and Asian write_in	X	X	X	X	.	.	.	.

Scrolling down further, there were more variations. For example, Asian or Pacific Islander. So there's a lot of change over time.

55

We know the variable of Race, so now we can specify the geographic area: California

- Right now, the general codes are sufficient, so click on General codes.
- To go back to the variable list click on Group: Race, Ethnicity, and Nativity— PERSON at the top
- Geographic variables can be found under Household, by selecting Geographic.

- Then go to __STATEFIP__ and click on __Codes__
- Notice that California is __06__ so we can put that in.

To get to the online analysis tool, we need to go to the homepage and go to Analyze Data Online

Since we are interested in data for California from 1980 to present, choose United States, 1850-2013 and you will be prompted to log in with your IPUMS username and password.

After logging in, the first thing we want to do is specify California by clicking on Household Geographic, which will open up a list so you can choose statefip.

After you see statefib in the "Selected" box, click on View and you'll be able to see California is 6.
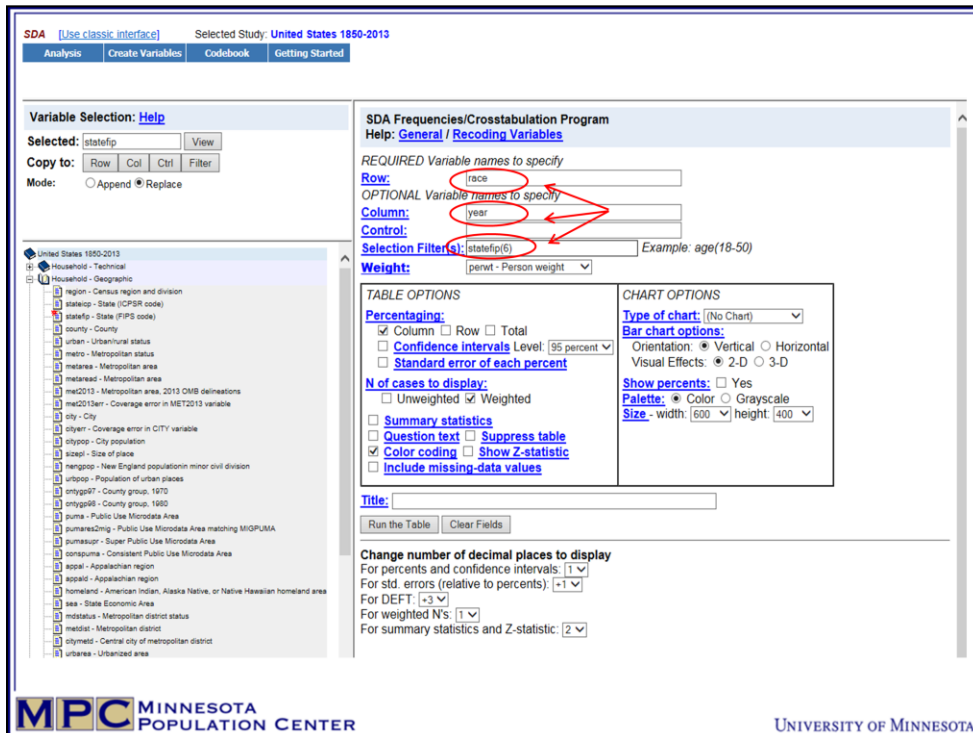
I'm ready to specify my variables for analysis.

- Because of my prior experience, I know that the race variable is "race" so I put it in Row
- I want to know change over time, so I specify "year" in Column
- Since I am interested in California, I'm going to limit the statefip information by using the information I previously looked up by putting "statefip(6)" in Selection Filter(s)

These selections will create a race by year tabulation for California.

I'm not going to worry about what's in the TABLE OPTIONS or CHART OPTIONS sections right now. We'll go back to that, because generally I run a table once and then I go figure out what I need to do.

To run the data and get the table, click on Run the table. And you can see the result.

I should mention that by default, the data are weighted by the perwt to make the sample representative. If it was perfectly even, everyone would represent 100 people, because we have 1% of the population. But from a sampling perspective that doesn't always work because there is non-response from people and they know they don't get enough rural people. So in rural areas I might be worth 120 people vs. 100 people. And city people are sometimes valued less than that. They vary.

All the way back to 1850 is beyond my interest, so let's limit the years. I'm going to go back to the tab where I specified my selections. Under Selection Filter, Add "year(1980, 1990, 2000, 2010)". Click Run the Table.

Frequency Distribution					
Cells contain: -**Column percent** -Weighted N	year				
	1980 1980	1990 1990	2000 2000	2010 2010	*ROW* *TOTAL*
1: White	**84.6** 20,072,800.0	**69.2** 20,541,976.0	**59.3** 20,105,863.0	**62.5** 23,332,784.0	**67.4** *84,053,423.0*
2: Black/Negro	**7.7** 1,819,300.0	**7.4** 2,183,831.0	**6.6** 2,223,673.0	**6.0** 2,249,521.0	**6.8** *8,476,325.0*
3: American Indian or Alaska Native	**1.0** 238,100.0	**.8** 249,925.0	**.9** 305,769.0	**.8** 280,771.0	**.9** *1,074,565.0*
4: Chinese	**1.4** 333,600.0	**2.4** 700,284.0	**2.9** 967,973.0	**3.4** 1,287,458.0	**2.6** *3,289,315.0*
5: Japanese	**1.1** 269,600.0	**1.1** 312,871.0	**.8** 283,868.0	**.8** 283,659.0	**.9** *1,149,998.0*
6: Other Asian or Pacific Islander	**3.1** 745,800.0	**6.1** 1,818,375.0	**7.5** 2,529,225.0	**9.3** 3,480,212.0	**6.9** *8,573,612.0*
7: Other race, nec	**1.0** 233,600.0	**13.1** 3,883,581.0	**17.0** 5,747,850.0	**13.1** 4,893,038.0	**11.8** *14,758,069.0*
8: Two major races	**.0** .0	**.0** .0	**4.8** 1,611,365.0	**3.8** 1,409,391.0	**2.4** *3,020,756.0*
9: Three or more major races	**.0** .0	**.0** .0	**.3** 103,734.0	**.4** 132,529.0	**.2** *236,263.0*
COL TOTAL	*100.0* *23,712,800.0*	*100.0* *29,690,843.0*	*100.0* *33,879,320.0*	*100.0* *37,349,363.0*	*100.0* *124,632,326.0*

race

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

You can see in 1980, 85% of people identified as White, and in 2010 it was 63%. When you're looking at this off the bat, you can see when the Two major races join in, Other race, nec category gets very large. The Other race, nec in 1980 is 1%, and is up to 17% in 2000 and back to 13.1% in 2010. What's missing here? Hispanic. This is a census data thing. They don't have Hispanic in Race, but they're piloting it right now to think about it in 2020. So we need to incorporate Hispanic.

So back we go to the tab where we specified the selections. There are a couple of things you can do here. You can recode race to account for Hispanic.

First, let's check out the Hispanic variable. Click on Person – Race, Ethnicity, and Nativity, then click hispan – Hispanic origin. Make sure hispan is shown in the box above between Selected and the View button. Then click View, which opens the Hispanic origin table.

With that information in mind, we can go back to our selections and type "hispan" in the Row field and then click Run the Table. The new Frequency Distribution chart shows big growth in Hispanic origin. You can see Not Hispanic goes from about 80% to 62% of the population of California. There's a big growth in Mexican, but not much change in Puerto Rican.

Now, going back to the selection page, just doing this on the fly… I think what we should do… Race by Hispanic origin. In <u>Row</u> field type "<u>race</u>", in <u>Control</u> field type "<u>year</u>", in <u>Column</u> field, type "<u>hispan(r: 0; 1-4)</u>" 'r' stands for recode and we are collapsing the different types of Hispanics into one category with the 1-4 part. Click <u>Run the Table</u>.

Statistics for year = 1980(1980)

race	hispan 1 (0)	hispan 2 (1-4)	ROW TOTAL
1: White	82.8 / 15,843,800.0	92.5 / 4,229,000.0	84.6 / 20,072,800.0
2: Black/Negro	9.3 / 1,788,200.0	.7 / 31,100.0	7.7 / 1,819,300.0
3: American Indian or Alaska Native	1.0 / 190,500.0	1.0 / 47,600.0	1.0 / 238,100.0
4: Chinese	1.7 / 331,100.0	.1 / 2,500.0	1.4 / 333,600.0
5: Japanese	1.4 / 266,400.0	.1 / 3,200.0	269
6: Other Asian or Pacific Islander	3.5 / 679,500.0	1.5 / 66,300.0	745
7: Other race, nec	.2 / 42,100.0	4.2 / 191,500.0	233
COL TOTAL	100.0 / 19,141,600.0	100.0 / 4,571,200.0	23,712

Statistics for year = 1990(1990)

race	hispan 1 (0)	hispan 2 (1-4)	ROW TOTAL
1: White	77.1 / 17,075,721.0	46.0 / 3,466,255.0	69.2 / 20,541,976.0
2: Black/Negro	9.4 / 2,092,168.0	1.2 / 91,663.0	7.4 / 2,183,831.0
3: American Indian or Alaska Native	.9 / 198,497.0	.7 / 51,428.0	.8 / 249,925.0
4: Chinese	3.1 / 693,355.0	.1 / 6,929.0	2.4 / 700,284.0
5: Japanese	1.4 / 308,347.0	.1 / 4,524.0	1.1 / 312,871.0
6: Other Asian or Pacific Islander	7.9 / 1,744,262.0	1.0 / 74,113.0	6.1 / 1,818,375.0
7: Other race, nec	.2 / 46,670.0	50.9 / 3,836,911.0	13.1 / 3,883,581.0
COL TOTAL	100.0 / 22,159,020.0	100.0 / 7,531,823.0	100.0 / 29,690,843.0

Cells contain: -Column percent -Weighted N

Color coding: <-2.0 <-1.0 <0.0 >0.0 >1.0 >2.0 Z
N in each cell: Smaller than expected | Larger than expected

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

Here is the result.

In 1980, about 92% of people identified as Hispanic identified their race as White. This changes a lot. So in 1990, it's 46%.

We had a question about allocation. That is, what percentage of the responses had to be determine based on other data because the information was missing on the survey. The data quality flag for this variable is called QHISPAN.

Go back to the selection page. In Row field, type "qhispan" and then click Run the Table

Statistics for year = 1980(1980)

Cells contain: -Column percent -Weighted N		hispan 1 (0)	hispan 2 (1-4)	ROW TOTAL
qhispan	0: Not allocated	97.9 18,732,700.0	98.7 4,511,700.0	98.0 23,244,400.0
	1: Allocated from information for this person or from relative, this household	1.7 322,000.0	1.1 48,200.0	1.6 370,800.0
	2: Allocated from nonrelative, this household	.0 3,900.0	.0 1,100.0	.0 5,000.0
	4: Allocated	.4 82,400.0	.2 10,200.0	.4 92,600.0
	COL TOTAL	100.0 19,141,600.0	100.0 4,571,200.0	100.0 23,712,800.0

Color coding: <-2.0 <-1.0 <0.0 >0.0 >1.0 >2.0 Z
N in each cell: Smaller than expected   Larger than expected

Statistics for year = 1990(1990)

Cells contain: -Column percent -Weighted N		hispan 1 (0)	hispan 2 (1-4)	ROW TOTAL
qhispan	0: Not allocated	96.6 21,394,763.0	98.2 7,397,039.0	97.0 28,791,802.0
	4: Allocated	3.4 764,257.0	1.8 134,784.0	3.0 899,041.0
	COL TOTAL	100.0 22,159,020.0	100.0 7,531,823.0	100.0 29,690,843.0

Color coding: <-2.0 <-1.0 <0.0 >0.0 >1.0 >2.0 Z
N in each cell: Smaller than expected   Larger than expected

Statistics for year = 2000(2000)

Cells contain: -Column percent -Weighted N		hispan 1 (0)	hispan 2 (1-4)	ROW TOTAL
qhispan	0: Not allocated	96.0 21,970,562.0	95.4 10,476,613.0	95.8 32,447,175.0
	4: Allocated	4.0 922,339.0	4.6 509,806.0	4.2 1,432,145.0
	COL TOTAL	100.0 22,892,901.0	100.0 10,986,419.0	100.0 33,879,320.0

Color coding: <-2.0 <-1.0 <0.0 >0.0 >1.0 >2.0 Z
N in each cell: Smaller than expected   Larger than expected

Statistics for year = 2010(2010)

Cells contain: -Column percent -Weighted N		hispan 1 (0)	hispan 2 (1-4)	ROW TOTAL
qhispan	0: Not allocated	97.8 22,738,474.0	98.7 13,914,283.0	98.1 36,652,757.0
	4: Allocated	2.2 518,897.0	1.3 177,709.0	1.9 696,606.0
	COL TOTAL	100.0 23,257,371.0	100.0 14,091,992.0	100.0 37,349,363.0

Color coding: <-2.0 <-1.0 <0.0 >0.0 >1.0 >2.0 Z
N in each cell: Smaller than expected   Larger than expected

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

The result shows that, in 1980, a small percentage had to be allocated for hispanics and non-hispanics. In 1990, more non-hispanics were allocated than hispanics. In 2000, it's a little higher with 4%. And then in 2010, about 2%. That's a pretty small allocation. If you pulled this out of the data in an extract, you would have that again on the person record, so you would know if that record was allocated and you could decide if you wanted to include those or do something with them.

Example research questions, methods, answers

- Research Question: In the San Diego metro area in 2013, how many people are living with disabilities?
  - What is the Metro area code of San Diego?
  - How is disability measured?
  - Who is in the universe for the disability question?

Back to my questions… In the San Diego metro area in 2013, how many people are living with disabilities? What is the metro area code for San Diego? Can we identify this metro area? How is disability measured? Who is in the universe for the disability question?

Being aware of the universe is a really big deal. For example, income will have six or seven 9's for people under 16. If you do not handle this situation, you'll have a lot of really wealthy children in your sample, and you'll think, "the incomes are so high this can't be right," and it's *not* right! You have to see who's in the universe.

First, I'm going to change my samples. <u>Change Samples</u> and grab <u>2013</u> and click <u>Submit sample selections</u>.

And we're going to look at the disability variables. Go to Person and click Disability.

In 2008, questions about disability were added to the ACS. You can see there are a whole bunch of variables. The X's indicate that they're available in the sample.

You see hearing difficulties, vision difficulties, independent living difficulties, cognitive difficulties, etc. Let's go with the independent living variable. Click on DIFFMOB next to "Independent living difficulty."

The variable description helps you determine if it answers your research question. Here we see that DIFFMOB indicates whether the respondent has any physical, mental, or emotional condition lasting six months or more that makes it difficult or impossible to perform basic activities outside the home alone. This does not include temporary health problems, such as broken bones or pregnancies.

The questionnaire text provides more details. To view the questionnaire, click on Questionnaire Text and then specify our dataset ACS(2013)

17. a) Is this person deaf or does he/she have serious difficulty hearing?
[ ] Yes
[ ] No

b) Is this person blind or does he/she have serious difficulty seeing even when wearing glasses?
[ ] Yes
[ ] No

[G] - Answer question 18a-c if this person is 5 years old or over. Otherwise, SKIP to the questions for PERSON 2 on page 12.

18. a) Because of a physical, mental, or emotional condition, does this person have serious difficulty concentrating, remembering, or making decisions?
[ ] Yes
[ ] No

b) Does this person have serious difficulty walking or climbing stairs?
[ ] Yes
[ ] No

c) Does this person have difficulty dressing or bathing?
[ ] Yes
[ ] No

[H] - Answer question 18 if this person is 15 years old or over. Otherwise, SKIP to the questions for Person 2 on page 12.

19. Because of a physical, mental, or emotional condition, does this person have difficulty doing errands alone such as visiting a doctor's office or shopping?
[ ] Yes
[ ] No

MPC MINNESOTA POPULATION CENTER — UNIVERSITY OF MINNESOTA

So there are a bunch of different questions.

Question 19: *Because of a physical, mental, or emotional condition, does this person have difficulty doing errands alone such as visiting a doctor's office or shopping? Yes or no?*

Question 18b: *Does this person have serious difficulty walking or climbing up stars?*

Question 18c: *Does this person have difficulty dressing or bathing?*

If you're interested, you can dig into this more.

With this information in mind, we can go back to the IPUMS Online Data Analysis System from the Home page and click on Analyze Data Online.

Let's scroll down and grab the 2013 ACS because it's faster when using one dataset.

I'm going to look at DIFFMOB in the U.S. to start before digging into San Diego. Type "diffmob" into the Row field and click Run the Table.

Here's this N/A for Not in Universe. Right away we can see this zero and we can look at who's in that. Also 4.9%, which includes those people not in the universe, have independent living difficulties.

Let's check out the universe statement. Go back to IPUMS-USA and DIFFMOB tab and click Universe.

Here we see ACS, PRCS: Persons age 16+. So again, all of the children are not in the universe.

I'm going to drop those by using Selection Filters. Back in the Online Analysis system, type "age(16-100)" into Selection Filter(s) and click Run the Table.

SDA 3.5: Tables

2013 ACS sample

May 27, 2015 (Wed 10:15 AM CDT)

Variables					
Role	Name	Label	Range	MD	Dataset
Row	diffmob	Independent living difficulty	0-2		1
Weight	perwt	Person weight	1.00-2,112.00		1
Filter	age(16-100)	Age	0-95		1

Frequency Distribution		
Cells contain: -Column percent -Weighted N		Distribution
diffmob	1: No independent living difficulty	93.9 235,528,264.0
	2: Has independent living difficulty	6.1 15,381,714.0
	COL TOTAL	100.0 250,909,978.0

Allocation of cases (unweighted)	
Valid cases	2,541,519
Cases excluded by filter or weight	591,276
Total cases	3,132,795

Datasets	
1	/share/htdocs/us2013a
2	/sda_support/newvars/us2013a

CSM, UC Berkeley

Now, 6.1% in the U.S. having independent living difficulties.

Now for San Diego. Go back to IPUMS-USA and click Disability-Person to get back to the variable list. To get the geographic list of variables, go to Household and then click Geographic.

There are two variables: 1) MET2013, Metropolitan area, 2013 OMB delineations, and 2) MET2013ERR, which is the coverage error. This is where geography gets hard. We're layering PUMAs within metro areas. But we're going to use this MET2013 variable. Click on it.

Based on the description, this variable uses the official 2013 delineations for the metropolitan statistical areas. There's lots of information, you can scroll down to "Crosswalk between 2013 MSAs and 2010 PUMAs" or to "MET2013 Omission and Commissions Errors by MSA."

But all I really care about is if we can identify San Diego. Click on Codes.

From this code list, I search for San Diego (or scroll down). So the code for
"San Diego-Carlsbad, CA" is 41740.

With this number, I go back to the Online Analysis System and I add "met2013(41740)" to the Selection Filter(s) and click Run the Table.

Now, we see 5.1%, slightly less than the U.S.

Using the extract system and documentation

- Extract System

Onto my third research question. To answer this question, we're going to use the extract system and documentation. We've been using the online analysis, but let's say you want to do your analysis in SPSS or Wxcel. We're going to watch our video on the Extract System:
https://www.youtube.com/watch?v=pOjEWoIcdHI

Example research questions, methods, answers

- Research Question: How do people in CA commute to work?
  - What information does the ACS have on commuting?
  - Which respondents are asked about commuting?
    - What variables do we need to restrict the sample to these respondents?
  - What is the state code for CA?

MPC MINNESOTA POPULATION CENTER   UNIVERSITY OF MINNESOTA

Here's my research question: How do people in California commute to work? What information does the ACS have on commuting? Which respondents are asked about commuting? Do you have to be working? Do you have to be working right now? What variables do we need to restrict the sample to these respondents? What is the state code for CA? We already did that.

I'm going back into my data searching back on the home page, to Select Data
And then click Change Samples.

I'm going to go to my Default sample from each year so I can see what's going on. Then click Submit sample selections. Then go to Person and select Place of Work and Travel Time.

So we see "means of transportation to work," "carpooling," "vehicle occupancy," "travel time to work," "travel departure for work," "time of arrival to work." So these are all variables available and we can look at them.

Click on TRANWORK.

Tranwork seems to be a big one: *TRANWORK reports the respondent's primary means of transportation to work on the most recent day worked (1970), or over the course of the previous week (the 1960 and 1980-2000 censuses, the ACS, and the PRCS). The primary means of transportation was that used on the most days or to cover the greatest distance.*

Click Codes.

So here are the codes available: *auto, motorcycle, bus, railroad, ferryboat, bicycle, walked only, worked at home, etc.*

Let's see who was asked this question. Click Universe.

ACS, PRCS: Persons age 16+ who worked last week, 1960: Persons age 14+ who worked last week, reported, 1970: Persons age 14+ who worked last week, 1980-2000: Persons age 16+ who worked last week.

So pretty similar over time.

Let's go check out the Questionnaire Text.

Click on ACS(2013) because that is our sample and look at question 31. How did this person usually get to work LAST week? *If this person usually used more than one method of transportation during the trip, mark (X) the box of the one used for most of the distance. How many people, including this person, usually rode to work in the car, truck, or van LAST week?*

The results from this question will provide the data I need.

So let's say I want to create an EXTRACT. Click on Place of Work and Travel Time – PERSON to go back to the list of variables.

Click on TRANWORK, CARPOOL, RIDERS, TRANTIME, DEPARTS, ARRIVES to make an extract for all these variables.

In addition to transportation from work, is there anything else we want? Let's grab an age variable and sex variable. Click on Person and choose Demographic. Select SEX and AGE. So now I have some variables, some samples I collected. Let's look at what we have. Click on VIEW CART.

There're some preselected variables automatically in here, which includes the serial weight, the year. You can get rid of them if you want or keep them. Now, I'm going to create my data by clicking on Create data extract. Here we have Samples and Variables. Here's this Data format. To change the data format, click on Change in the Data format row.

This is where you choose your format. Let's ask for a csv. Click on Comma delimited (.csv) and then hit Submit.

I forgot, we need to specify California, so I'm going to grab the state variable. Click on Change in the Variables row.

Remember, the geographic data is part of the household information. Click on Household and go to Geographic. And then grab this state variable: STATEFIP, then click on VIEW CART.

Go back to Create data extract.

And now I can select the state of CA by clicking on Select cases and then
STATEFIP and Submit and then choosing California and Submit again.

Add a description under <u>Describe your extract</u>. Let's use <u>"CA transportation"</u> and then click <u>Submit extract</u>.

I submit this extract and there we go. We have Extract number 87. The Command Files include specific programs to read in the csv file in SPSS, SAS, or STATA. Here's a custom codebook. You can Revise it, and you have your Description.

And then we're waiting for data, you'll be notified by email when it's ready. We wait for the data to come up, and then we can try opening it.

While we are waiting, I'll show you how to revise it. Click on <u>revise</u>. Let's change the samples. Click on <u>Change</u> in the <u>Samples</u> row.

I'm going to change this and just grab one sample. Unclick default sample from each year and select 2013 ACS. Then select Submit sample selections. And click Create data extract.

Let's make sure that we are including California only. Click Select cases. Since STATEFIP is selected, click on Submit.

Make sure 06 California is highlighted and click Submit. And click Submit extract.

After selecting Submit extract, this page comes up.

We can check the data through the Online Analysis tool. Go back to the Home page and select Analyze Data Online and select United States, 1850-2013.

Make sure Household-Geographic is selected.

Type "tranwork" in the Row section. Type "year" in the Column section. Type "age(16-100) statefip(6) year (1960, 1980, 1990, 2000, 2013)" in the Selection Filter(s) section, and click on Run the Table.

Looking at California data while waiting for the data to come up…

Here are all of our N/As. So we need to go back and add in "tranwork (10-100) to the Selection Filter(s) section. Make sure "tranwork" is still in the Row section and "year" is still in the Column section. And click on Run the Table.
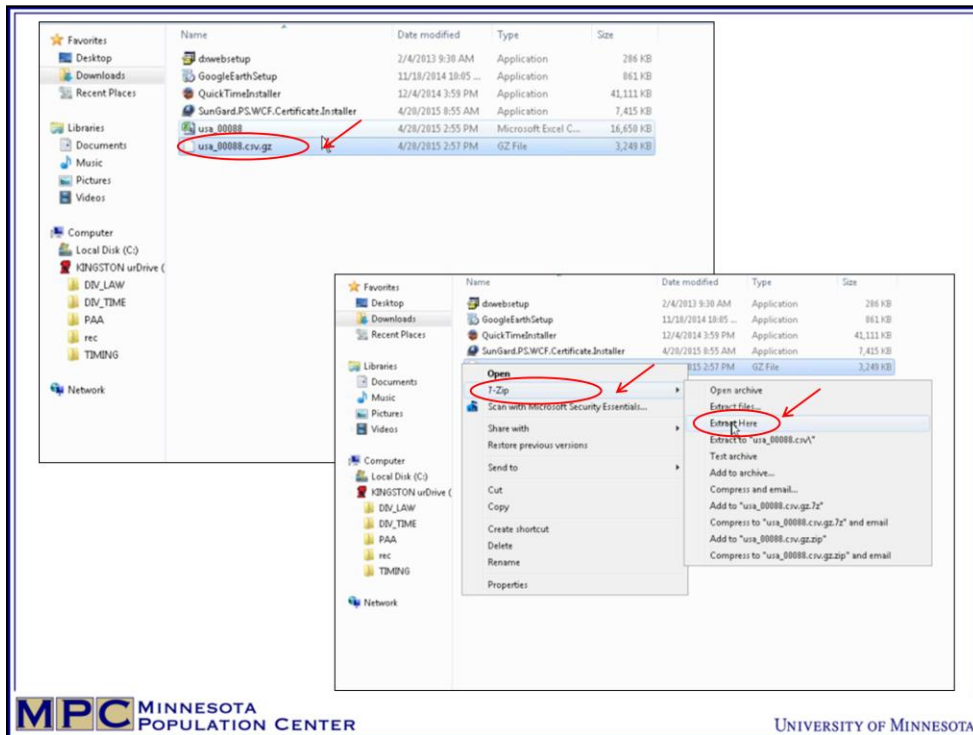
While waiting for the chart to come up, let's see if the data have been extracted. Go back to IPUMS-USA.

The data are ready. We want to download the file. Right click CSV. Go to save link as and we're going to save it and you have to unzip it.
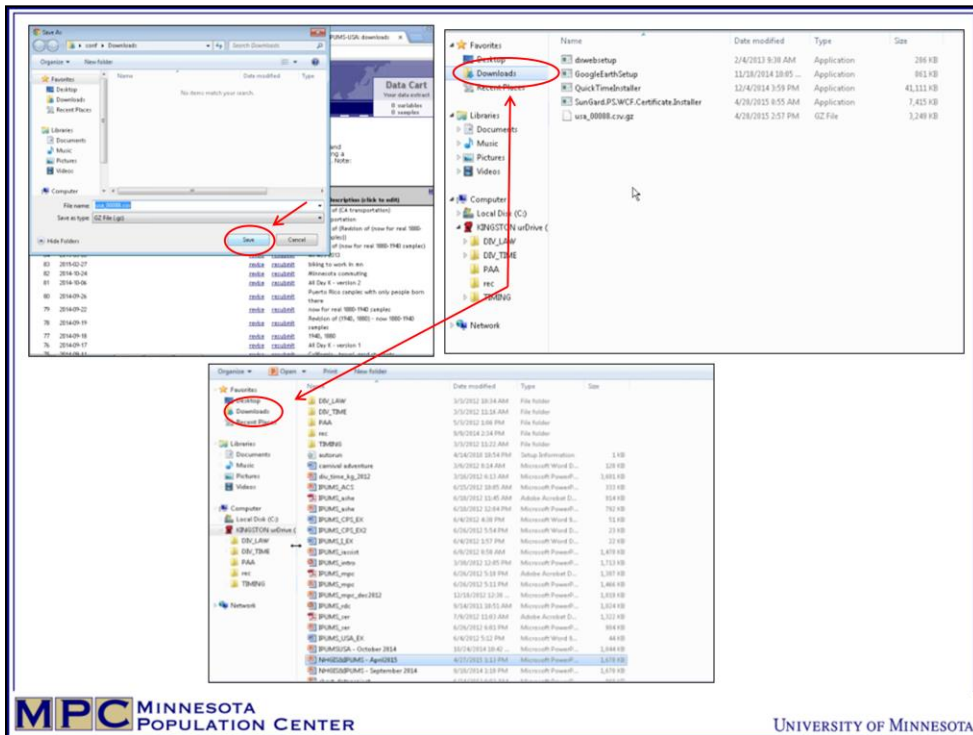
Press <u>Save</u> and save it to <u>Downloads</u>
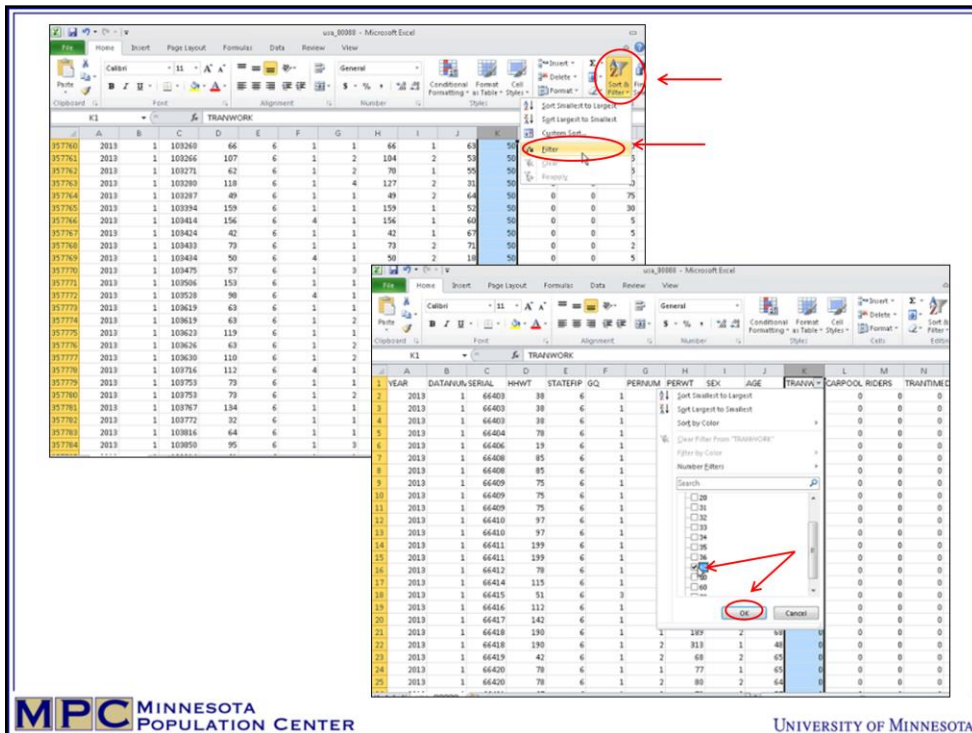
Go to downloads. Select the downloaded GZ File and right click, and to 7-ZIP, and then Extract here. Now we have it!

Here are the data! Each row represents a completed survey and each column shows the variable. For example, here is TRANWORK.

Since Bike Month is next month, let's find out how many people bike to work. The code is 40 for bike to work. I'm going to sort on that. Go to Sort& Filters and select Filters, then check box "40".

Based on this list of 40's, that's a lot in a sample, but that doesn't tell us that much without applying the weights.

The Minnesota Population Center table showing transportation to work (tranwork) by year, with Column percent and Weighted N values:

tranwork		1960	1980	1990	2000	2013	ROW TOTAL
		1960	1980	1990	2000	2013	
	10: Auto, truck, or van	.0 / 0	.0 / 0	86.1 / 11,992,565.0	86.6 / 12,600,690.0	84.1 / 14,092,623.0	63.1 / 38,685,878.0
	11: Auto	77.4 / 4,305,838.0	72.1 / 7,611,100.0	.0 / 0	.0 / 0	.0 / 0	19.4 / 11,916,938.0
	14: Truck	.0 / 0	10.0 / 1,051,500.0	.0 / 0	.0 / 0	.0 / 0	1.7 / 1,051,500.0
	15: Van	.0 / 0	2.7 / 283,800.0	.0 / 0	.0 / 0	.0 / 0	.5 / 283,800.0
	20: Motorcycle	.0 / 0	1.1 / 118,400.0	.5 / 73,437.0	.3 / 39,116.0	.3 / 56,789.0	.5 / 287,742.0
	30: Bus or streetcar	7.8 / 432,654.0	5.2 / 550,400.0	.0 / 0	.0 / 0	.0 / 0	1.6 / 983,054.0
	31: Bus or trolley bus	.0 / 0	.0 / 0	4.1 / 566,735.0	3.8 / 547,038.0	3.7 / 621,346.0	2.8 / 1,735,119.0
	32: Streetcar or trolley car	.0 / 0	.0 / 0	.2 / 21,320.0	.1 / 1,214.0	.1 / 22,742.0	.1 / 65,276.0
	33: Subway or elevated	.0 / 0	.4 / 37,200.0	.5 / 73,840.0	.7 / 102,643.0	1.1 / 179,575.0	.6 / 393,258.0
	34: Railroad	.3 / 15,130.0	.1 / 15,100.0	.1 / 15,392.0	.3 / 38,003.0	.4 / 68,862.0	.2 / 152,487.0
	35: Taxicab	.1 / 7,663.0	.1 / 6,600.0	.0 / 5,993.0	.1 / 8,107.0	.0 / 7,269.0	.1 / 35,632.0
	36: Ferryboat	.0 / 0	.0 / 0	.0 / 4,524.0	.1 / 7,362.0	.1 / 8,973.0	.0 / 20,859.0
	40: Bicycle	.0 / 0	1.2 / 123,980.0	1.0 / 136,058.0	.8 / 117,382.0	1.1 / 189,154.0	.9 / 566,494.0
	50: Walked only	6.4 / 354,133.0	4.6 / 482,300.0	3.4 / 471,965.0	2.8 / 408,140.0	2.7 / 454,770.0	3.5 / 2,171,308.0
	60: Other	2.0 / 113,982.0	.8 / 80,600.0	.8 / 105,963.0	.8 / 109,536.0	1.0 / 174,878.0	1.0 / 684,959.0
	70: Worked at home	6.0 / 333,230.0	1.8 / 194,300.0	3.3 / 459,862.0	3.8 / 552,822.0	5.3 / 879,731.0	3.9 / 2,419,945.0

Let's go back to the table we were waiting for in the Online Analysis Tool.

First, looking at biking to work, it's pretty big, 1.2% of the population in 1980. And now, almost 200,000 people in California are biking.

Looking at the 60's, 80's 90's, the Subway must have been introduced in California.

A huge number Worked at home. Even from 2000 to 2013, that's super surprising.

Here you have it, you can now use data.

Now, for more practice…

Go to www.pop.umn.edu and click on Data User Resources and User Support and Training Materials.

If you scroll down here, there are a lot of options.

- Under the Online Module column are exercises for how to use the Online Analysis Tool that we have been working with today.

- Under the Print Version columns are exercises teaching you how to use each of the data projects listed in each row. Our focus today has been on IPUMS-USA and there are others listed here.

- Under the SAS, SPSS, and Stata are exercises walking you through how to use the data in those packages.

While we are trying to change this website, these exercises may be helpful to you. For example, let's click on NEW beside IPUMS-USA.

After clicking NEW, you get to this page. It's a nice walk-through method. It's a walk-through activity, which is pretty helpful.

So I am happy to take questions or help you individually as you work through any of these exercises, or if you want to ask me about something completely different, we can do that, too.