USP/PLSI 492: Research Methods

Instructor: Pietro Calogero

Getting Census data and prepping it for use in GIS: a real-world problem with Richmond, CA

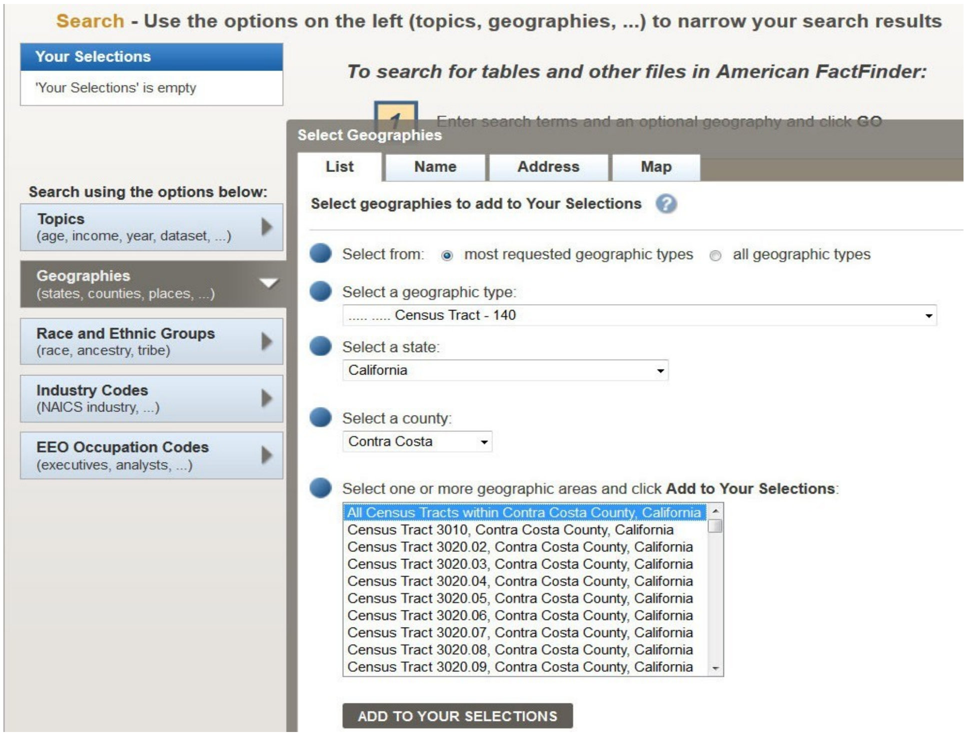
Initially I wrote this tutorial using San Francisco as my example. However, that was misleadingly simple because the borders of U.S. Census tracts and county boundaries align. Therefore, with San Francisco—which is both a city and a county—you never have to deal with census tracts that overlap the border.

Richmond, CA, however, is only one city within Contra Costa County. And the borders of Richmond are disturbingly weird. The U.S. Census did not bother trying to get their tract-boundaries to align with the Richmond city borders at all. Census tracts are designed to have roughly the same population as each other, and designed to be fairly regularly-shaped. So in western Contra Costa County, they often overlap between city of Richmond, adjacent cities such as El Sobrante, and unincorporated county area.

Urban planners should be concerned about this because of the economic, racial, and health-inequities between Richmond and the wealthy communities on the west side of San Francisco Bay. Right away, we face much greater methodological challenges in studying Richmond, because erratic political boundaries make geopsatial analysis much more challenging. And that is why we should focus on Richmond, not already-gentrified enclaves like San Francisco.

1. Get the data.

Go to American FactFinder 2, Advanced tab, and set your filters to get data from the geographic area of your study, which whatever characteristics most closely correspond to your survey-design. Since Richmond does not correspond to census-tracts, and yet I want to work with tract-level data in Richmond, I go to the next-larger geographic unit: I get the Census Tract data for Contra Costa Co:



Search - Use the options on the left (topics, geographies, ...) to narrow your search results

Your Selections

'Your Selections' is empty

To search for tables and other files in American FactFinder:

1 Enter search terms and an optional geography and click GO

Search using the options below:

Topics (age, income, year, dataset, ...)

Geographies (states, counties, places, ...)

Race and Ethnic Groups (race, ancestry, tribe)

Industry Codes (NAICS industry, ...)

EEO Occupation Codes (executives, analysts, ...)

Select Geographies

List | Name | Address | Map

Select geographies to add to Your Selections

Select from: ● most requested geographic types ○ all geographic types

Select a geographic type: .... Census Tract - 140

Select a state: California

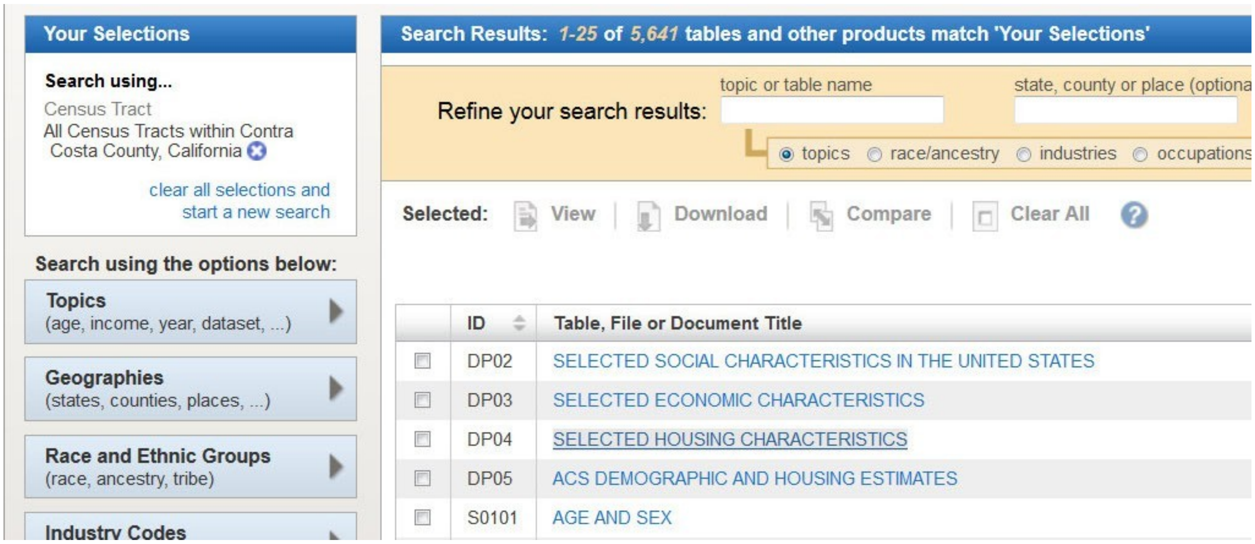
Select a county: Contra Costa

Select one or more geographic areas and click **Add to Your Selections**:

All Census Tracts within Contra Costa County, California
Census Tract 3010, Contra Costa County, California
Census Tract 3020.02, Contra Costa County, California
Census Tract 3020.03, Contra Costa County, California
Census Tract 3020.04, Contra Costa County, California
Census Tract 3020.05, Contra Costa County, California
Census Tract 3020.06, Contra Costa County, California
Census Tract 3020.07, Contra Costa County, California
Census Tract 3020.08, Contra Costa County, California
Census Tract 3020.09, Contra Costa County, California

ADD TO YOUR SELECTIONS

Select the table or tables that most closely correspond to the whole population (universe) that you are seeking to represent with survey data. As in this Richmond example, you may find that you need to go to a larger geographic unit, a larger population, and then refine the data by various criteria—including geography—until you can get a demographic profile of the population from which you want to do your sample-survey. In this case I am interested in housing data, and I know that what I want is in Table DP04. Therefore, with just one filter—Census tracts of Contra Costa County—one of my first hits is the table I want. However, you may want to set many more filters so that you are not picking through 5,641 tables.



Your Selections

Search using...
Census Tract
All Census Tracts within Contra Costa County, California

clear all selections and start a new search

Search using the options below:

Topics
(age, income, year, dataset, ...)
Geographies
(states, counties, places, ...)
Race and Ethnic Groups
(race, ancestry, tribe)
Industry Codes

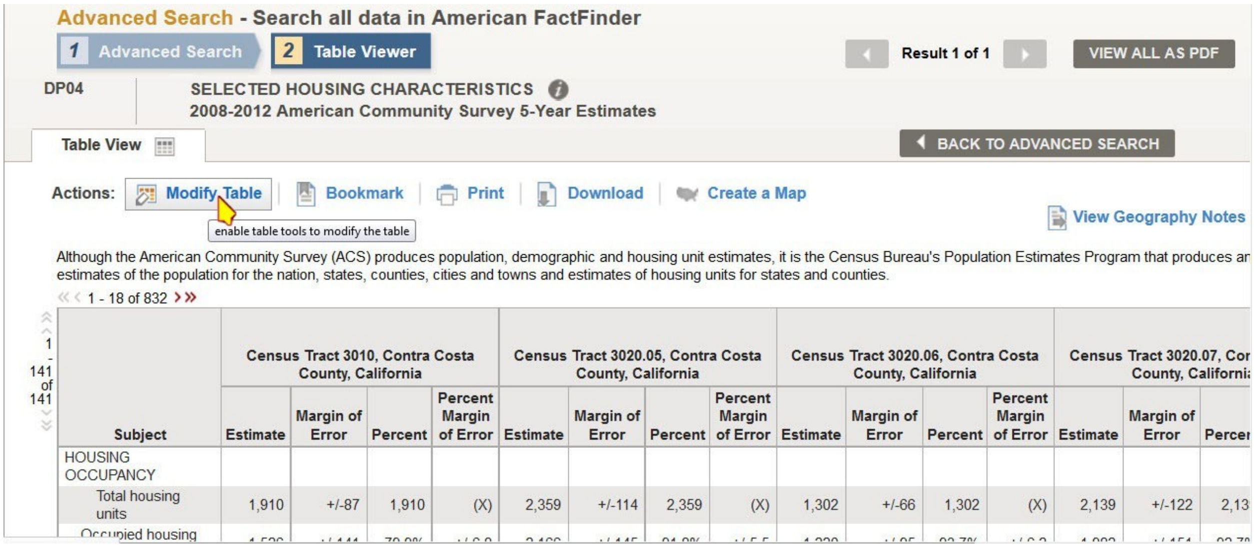
Search Results: 1-25 of 5,641 tables and other products match 'Your Selections'

Refine your search results: topic or table name | state, county or place (optiona

● topics ○ race/ancestry ○ industries ○ occupations
Selected: View | Download | Compare | Clear All

	ID	Table, File or Document Title
□	DP02	SELECTED SOCIAL CHARACTERISTICS IN THE UNITED STATES
□	DP03	SELECTED ECONOMIC CHARACTERISTICS
□	DP04	SELECTED HOUSING CHARACTERISTICS
□	DP05	ACS DEMOGRAPHIC AND HOUSING ESTIMATES
□	S0101	AGE AND SEX

Once you have found a table that you think will have your data, click the hyperlink of the table name itself so that you can preview the data. Note: if you are getting ACS data, rather than straight-up decennial census data, the ACS data will be estimated values. And it will include margins of error, and percent margins of error for each data-point. That means that ACS data will show 4X the number of data-columns. To reduce this before downloading, click the "Modify Table" link in the Actions bar.



Advanced Search - Search all data in American FactFinder

DP04 SELECTED HOUSING CHARACTERISTICS
2008-2012 American Community Survey 5-Year Estimates

Table View
BACK TO ADVANCED SEARCH
Actions: Modify Table | Bookmark | Print | Download | Create a Map
enable table tools to modify the table
View Geography Notes

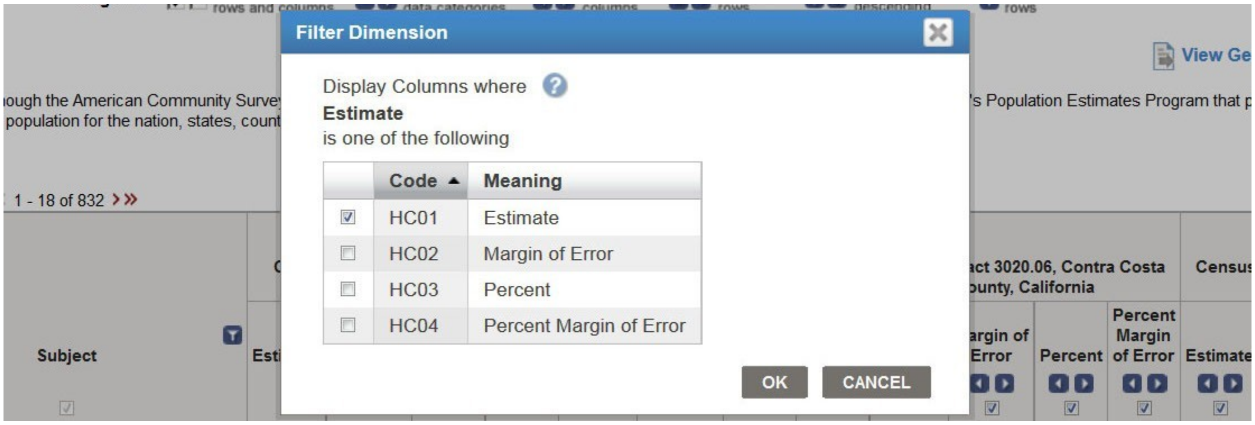
Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Program that produces ar estimates of the population for the nation, states, counties, cities and towns and estimates of housing units for states and counties.

1 - 18 of 832

1 - 141 of 141

	Census Tract 3010, Contra Costa County, California				Census Tract 3020.05, Contra Costa County, California				Census Tract 3020.06, Contra Costa County, California				Census Tract 3020.07, Cor County, Californi:		
Subject	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error	Percent	Percent Margin of Error	Estimate	Margin of Error	Percer
HOUSING OCCUPANCY															
Total housing units	1,910	+/-87	1,910	(X)	2,359	+/-114	2,359	(X)	1,302	+/-66	1,302	(X)	2,139	+/-122	2,13
Occupied housing															

Then, in the top left cell of the table itself (above the word **Subject**), a small blue funnel will appear. Click on this funnel, and check only the "Estimates" column. This will hide the "Estimate Margin of Error", "Percent", and "Percent Margin of Error" columns, and exclude them from the data-download. That reduces the amount of data you will need to sift through by 75%.



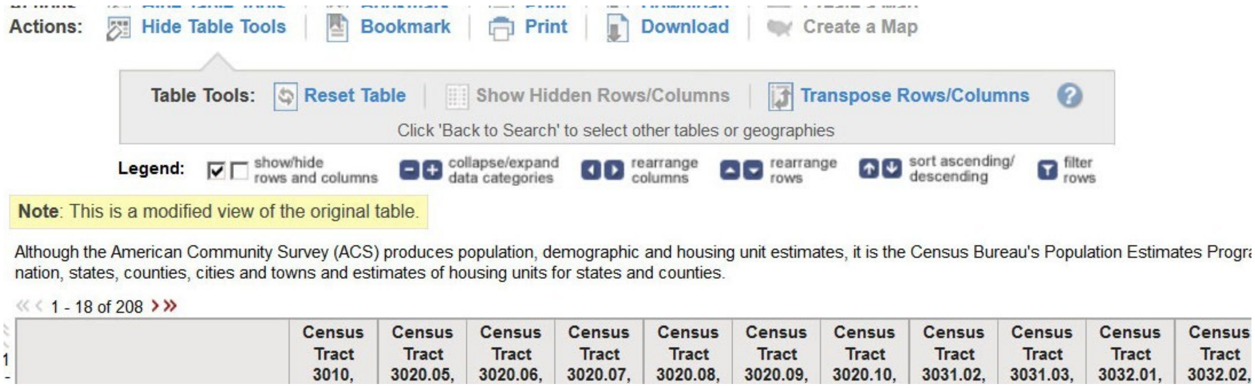
Filter Dimension

Display Columns where
Estimate
is one of the following

	Code ▲	Meaning
☑	HC01	Estimate
□	HC02	Margin of Error
□	HC03	Percent
□	HC04	Percent Margin of Error

OK   CANCEL

Note that before I reduce the desplay to “HC01/Estimate” only, there are 832 columns in this table...



Actions: Hide Table Tools | Bookmark | Print | Download | Create a Map

Table Tools: Reset Table | Show Hidden Rows/Columns | Transpose Rows/Columns

Click 'Back to Search' to select other tables or geographies

Legend: ☑□ show/hide rows and columns; collapse/expand data categories; rearrange columns; rearrange rows; sort ascending/descending; filter rows

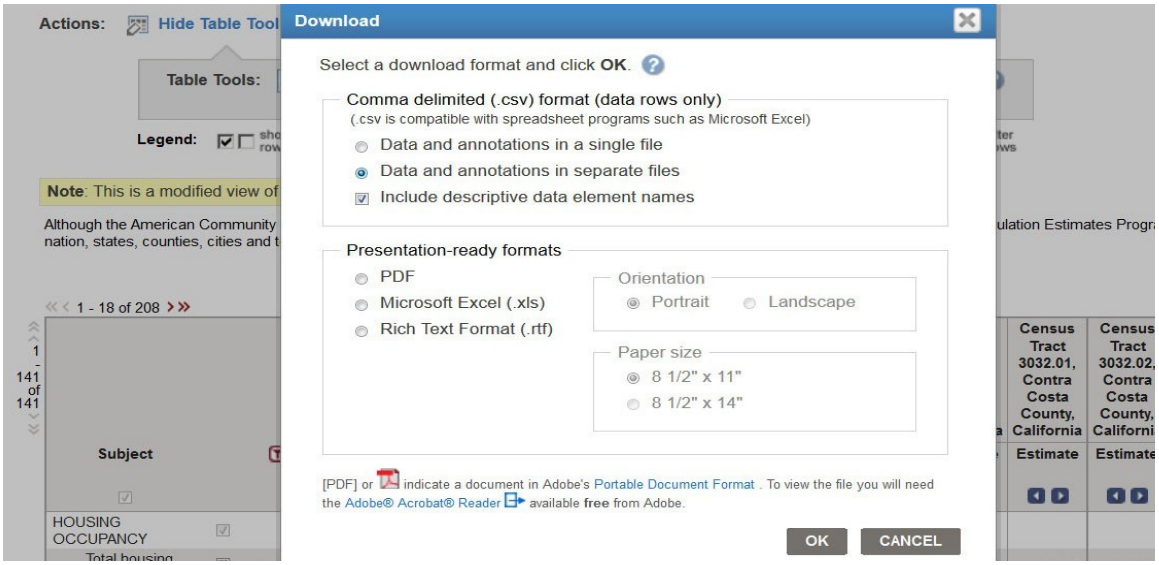
Note: This is a modified view of the original table.

Although the American Community Survey (ACS) produces population, demographic and housing unit estimates, it is the Census Bureau's Population Estimates Progra nation, states, counties, cities and towns and estimates of housing units for states and counties.

1 - 18 of 208

Once I filter the view to “Estimates” only, the number of columns drops to 208.

Select “Download…” And then choose the .CSV option. That will be a text file with values separated by commas. American FactFinder also offers an option to download the file in “Presentation-ready formats”. Do not succumb to the temptation of convenience! If you pick any of these options, it will strip off geocodes and it will be far more difficult to join this tabular data with TIGER geospatial data.



Download

Select a download format and click **OK**.

Comma delimited (.csv) format (data rows only)
(.csv is compatible with spreadsheet programs such as Microsoft Excel)

○ Data and annotations in a single file
● Data and annotations in separate files
☑ Include descriptive data element names

Presentation-ready formats

○ PDF
○ Microsoft Excel (.xls)
○ Rich Text Format (.rtf)

Orientation
● Portrait   ○ Landscape

Paper size
● 8 1/2" x 11"
○ 8 1/2" x 14"

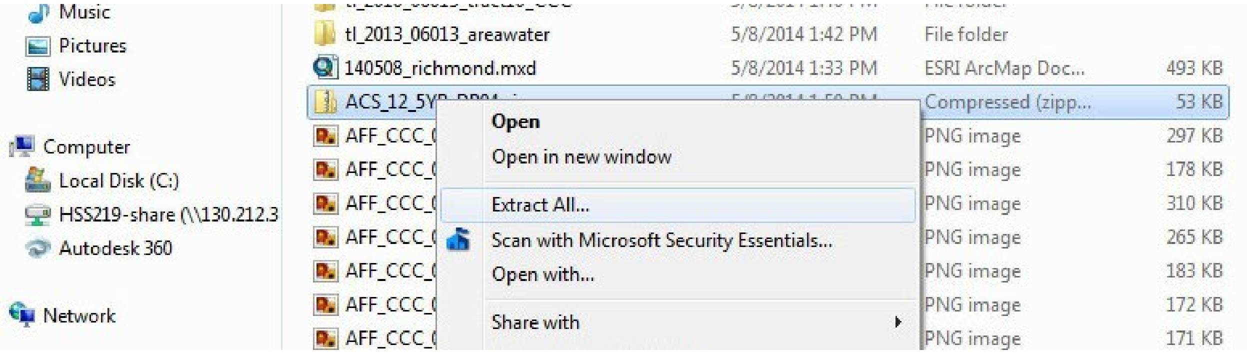
[PDF] or indicate a document in Adobe's Portable Document Format . To view the file you will need the Adobe® Acrobat® Reader available **free** from Adobe.

OK   CANCEL

Actions: Hide Table Tool
Table Tools:
Legend:
Note: This is a modified view of
1 - 18 of 208
Subject
HOUSING OCCUPANCY
Census Tract 3032.01, Contra Costa County, California
Census Tract 3032.02, Contra Costa County, Californi
Estimate
Estimate

DO choose “separate files”, and DO choose to “include descriptive data”.

Save the file at a PATH/NAME that you will be able to find again. And remember to extract it from the .zip archive:


Music
Pictures
Videos
Computer
Local Disk (C:)
HSS219-share (\\130.212.3
Autodesk 360
Network
tl_2013_06013_areawater 5/8/2014 1:42 PM File folder
140508_richmond.mxd 5/8/2014 1:33 PM ESRI ArcMap Doc... 493 KB
ACS_12_5YR_DP04... Compressed (zipp... 53 KB
AFF_CCC_ PNG image 297 KB
AFF_CCC_ PNG image 178 KB
AFF_CCC_ PNG image 310 KB
AFF_CCC_ PNG image 265 KB
AFF_CCC_ PNG image 183 KB
AFF_CCC_ PNG image 172 KB
AFF_CCC_ PNG image 171 KB
Open
Open in new window
Extract All...
Scan with Microsoft Security Essentials...
Open with...
Share with

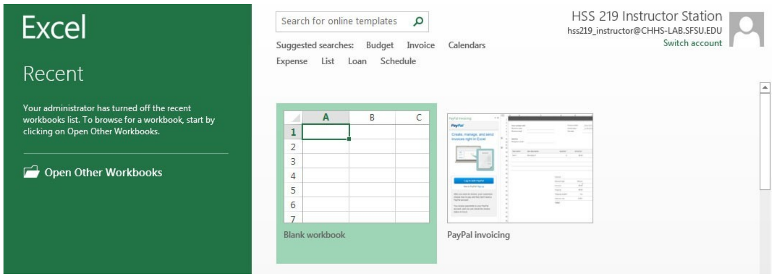
2. Prepare and arrange the data in Excel.

ArcMAP and QGIS can import .csv files. But the file you download from FactFinder will need to be cleaned up in two ways, which I call "winnowing" and "relabeling". Before we winnow and relabel, though, make sure you learn how to bring CSV data into Excel in a way that does not corrupt the data.

2a. Import a CSV into Excel without losing leading zeroes.

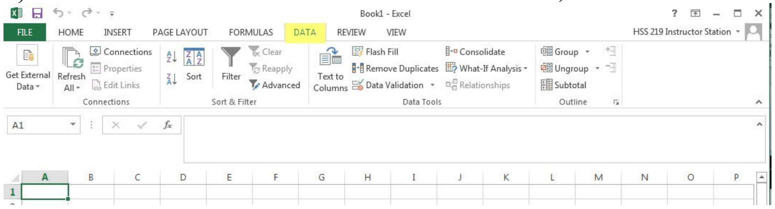
Importing CSV data into Excel is also a little tricky. If you open a FactFinder CSV file directly into Excel, it will strip off any leading zeros in any data. In California, the GEO.ID for census tracts begins with a zero, such as: 06075010100.

So: a) Open a **blank** Excel spreadsheet.

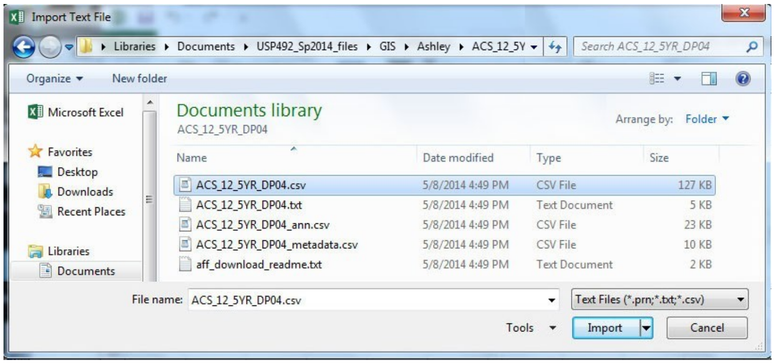

Your administrator has turned off the recent workbooks list. To browse for a workbook, start by clicking on Open Other Workbooks.

Excel
Recent
Open Other Workbooks
Search for online templates
Suggested searches: Budget Invoice Calendars
Expense List Loan Schedule
HSS 219 Instructor Station
hss219_instructor@CHHS-LAB.SFSU.EDU
Switch account
A
B
C
1
2
3
4
5
6
7
Blank workbook
PayPal invoicing

b) Under the Data tab > Get External Data subtab, Select "From text."



c) Choose the Path and file. The largest CSV is usually the right one.

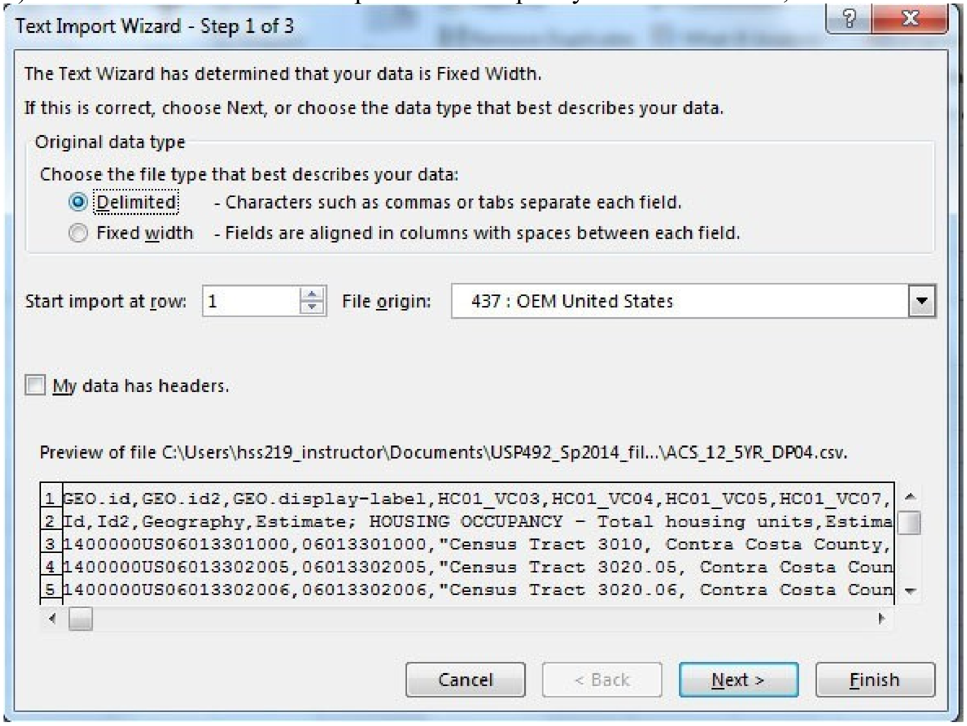


Name	Date modified	Type	Size
ACS_12_5YR_DP04.csv	5/8/2014 4:49 PM	CSV File	127 KB
ACS_12_5YR_DP04.txt	5/8/2014 4:49 PM	Text Document	5 KB
ACS_12_5YR_DP04_ann.csv	5/8/2014 4:49 PM	CSV File	23 KB
ACS_12_5YR_DP04_metadata.csv	5/8/2014 4:49 PM	CSV File	10 KB
aff_download_readme.txt	5/8/2014 4:49 PM	Text Document	2 KB

File name: ACS_12_5YR_DP04.csv

Text Files (*.prn;*.txt;*.csv)

d) This will invoke the Data Import Wizard. Specify "Delimited" data, not fixed-width.



The Text Wizard has determined that your data is Fixed Width.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

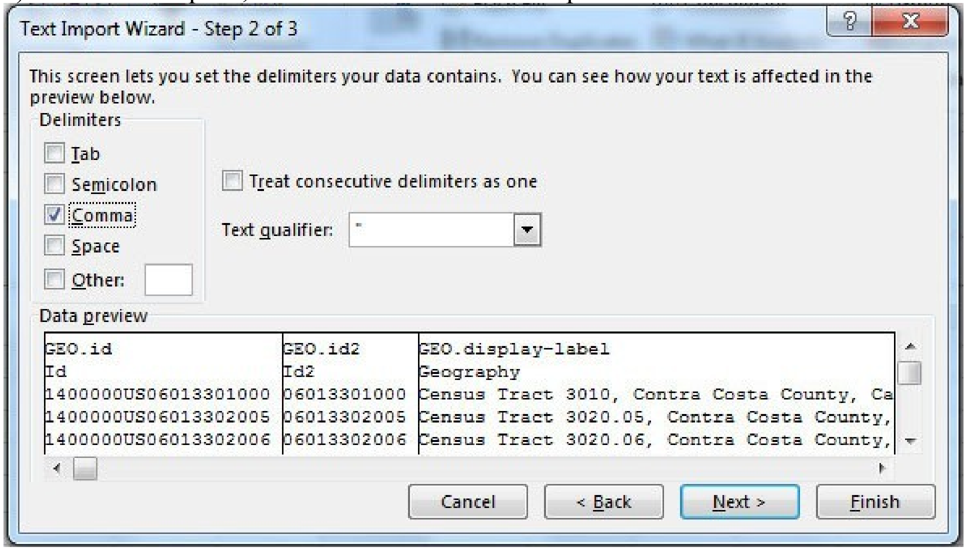
Start import at row: 1 File origin: 437 : OEM United States

□ My data has headers.

Preview of file C:\Users\hss219_instructor\Documents\USP492_Sp2014_fil...\ACS_12_5YR_DP04.csv.

```
1 GEO.id,GEO.id2,GEO.display-label,HC01_VC03,HC01_VC04,HC01_VC05,HC01_VC07,
2 Id,Id2,Geography,Estimate; HOUSING OCCUPANCY - Total housing units,Estima
3 1400000US06013301000,06013301000,"Census Tract 3010, Contra Costa County,
4 1400000US06013302005,06013302005,"Census Tract 3020.05, Contra Costa Coun
5 1400000US06013302006,06013302006,"Census Tract 3020.06, Contra Costa Coun
```

e) In the second pane, Choose "commas" as the separator and uncheck all the others.



This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

□ Tab

□ Semicolon

☑ Comma

□ Space

□ Other:

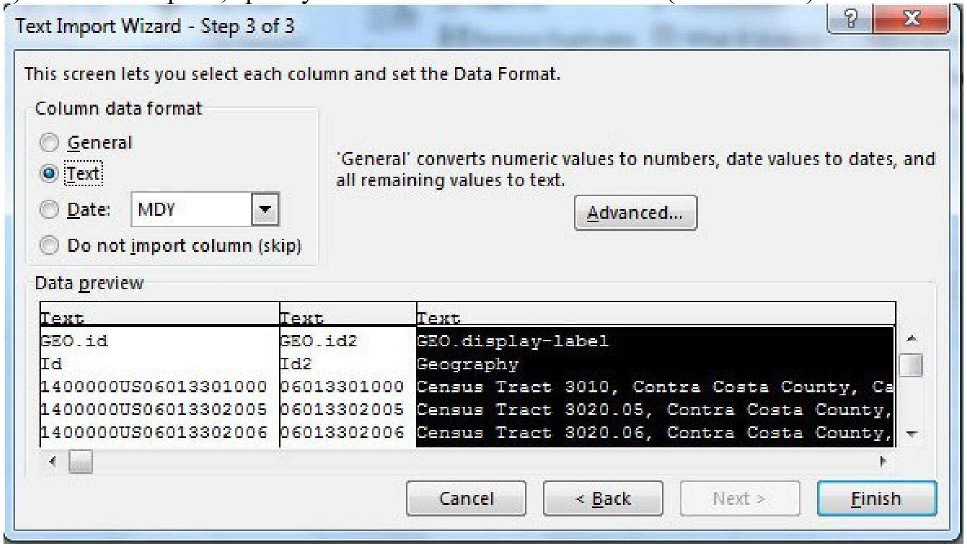
□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
GEO.id                GEO.id2      GEO.display-label
Id                    Id2          Geography
1400000US06013301000  06013301000  Census Tract 3010, Contra Costa County, Ca
1400000US06013302005  06013302005  Census Tract 3020.05, Contra Costa County,
1400000US06013302006  06013302006  Census Tract 3020.06, Contra Costa County,
```

f) In the third pane, specify the first three columns as Text (not General).


Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

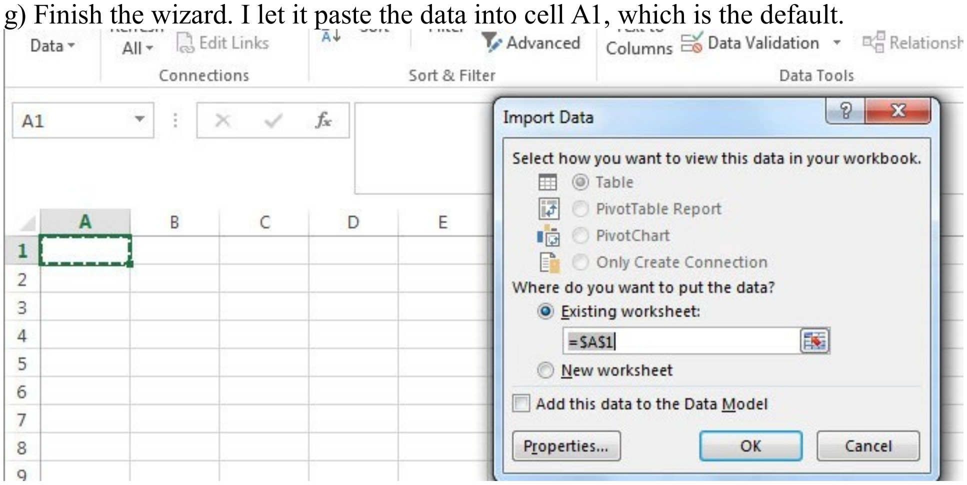
○ General
● Text
○ Date: MDY
○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Data preview

```
Text                  Text         Text
GEO.id                GEO.id2      GEO.display-label
Id                    Id2          Geography
1400000US06013301000  06013301000  Census Tract 3010, Contra Costa County, Ca
1400000US06013302005  06013302005  Census Tract 3020.05, Contra Costa County,
1400000US06013302006  06013302006  Census Tract 3020.06, Contra Costa County,
```

g) Finish the wizard. I let it paste the data into cell A1, which is the default.



Import Data

Select how you want to view this data in your workbook.

● Table
○ PivotTable Report
○ PivotChart
○ Only Create Connection

Where do you want to put the data?

● Existing worksheet:
=A1
○ New worksheet

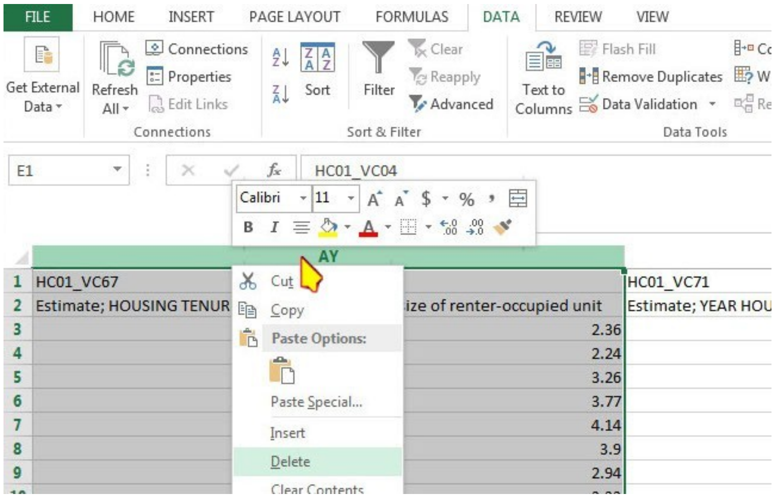
□ Add this data to the Data Model

Check that the data imported cleanly.

2b. Winnowing the data: pulling the needle out of the haystack.

Table DP04 has more than 100 columns of data, covering all sorts of housing characteristics. I only want to analyze a few characteristics, so I delete most of the columns. What I keep are:

- Column B: GEO.id2, which is the "correlation column" is will use to join this data to a TIGER file.
- Column C: this names the Census tract. I keep this as a verification-check that the Table-Join works correctly.
- Columns D: the total number of housing units in each Census Tract.
- Any remaining columns that include data I will analyze.



E1 HC01_VC04

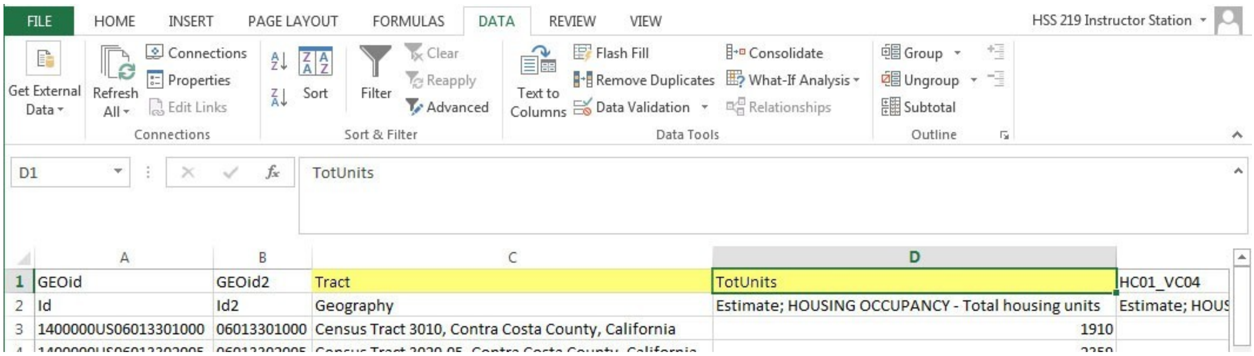
		AY	
1	HC01_VC67		HC01_VC71
2	Estimate; HOUSING TENUR	ize of renter-occupied unit	Estimate; YEAR HOU
3		2.36	
4		2.24	
5		3.26	
6		3.77	
7		4.14	
8		3.9	
9		2.94	

One more thing: Excel creates three-sheet Workbooks by default. Just to reduce confusion later, go to the bottom left of the window-border select Sheet2, R-click, Delete; and Sheet3, R-click, Delete. You can eve rename the Sheet1 as "Data" or something to remind you of which sheet to look for as you import this into other programs.

2c. Re-labeling the headers

In this third Excel step we need to do two things: edit the header-row (Row 1) so that ArcMAP can read the data, and also edit the header labels so that we ourselves can understand what data we are looking at.

To make the headers readable by ArcMAP, any punctuation-marks must be removed (except under_score). Rename "GEO.id2" to "GEOid2". Rename "GEO.display-label" to "tract_num".



D1 TotUnits

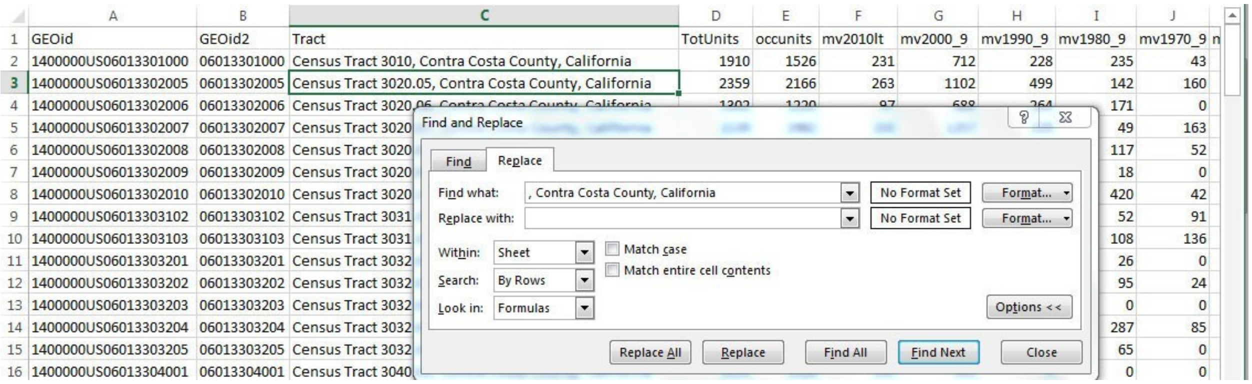
	A	B	C	D	
1	GEOid	GEOid2	Tract	TotUnits	HC01_VC04
2	Id	Id2	Geography	Estimate; HOUSING OCCUPANCY - Total housing units	Estimate; HOUS
3	1400000US06013301000	06013301000	Census Tract 3010, Contra Costa County, California	1910	

To make the remaining headers human-readable, rename them using the descriptive information in Row 2. [Back when you were downloading this CSV from FactFinder, the Row 2 information was included when you check the box "Include descriptive data element names".]
Replace the Row 1 codes with "dehydrated" but meaningful terms like "totalunits", "medvalue10", or "mvin2012". Max length: 12 characters. Don't start with a number. Don't use punctuation marks other than underscore: "_". Don't use spaces. Try to create the shortest possible name that you can intuitively understand.

Finally, I delete Row 2. Now, this table only has one header row (Row 1) and the rest of the rows are data-records for each census tract.

2d. Use find/replace to shorten descriptive data.

The data-column that includes the tract-number in FactFinder files has a lot of text. Use the "Find/Replace" function to shorten the name:



	A	B	C	D	E	F	G	H	I	J
1	GEOid	GEOid2	Tract	TotUnits	occunits	mv2010lt	mv2000_9	mv1990_9	mv1980_9	mv1970_9
2	1400000US06013301000	06013301000	Census Tract 3010, Contra Costa County, California	1910	1526	231	712	228	235	43
3	1400000US06013302005	06013302005	Census Tract 3020.05, Contra Costa County, California	2359	2166	263	1102	499	142	160
4	1400000US06013302006	06013302006	Census Tract 3020.06, Contra Costa County, California	1302	1220	97	688	264	171	0
5	1400000US06013302007	06013302007	Census Tract 3020						49	163
6	1400000US06013302008	06013302008	Census Tract 3020						117	52
7	1400000US06013302009	06013302009	Census Tract 3020						18	0
8	1400000US06013302010	06013302010	Census Tract 3020						420	42
9	1400000US06013303102	06013303102	Census Tract 3031						52	91
10	1400000US06013303103	06013303103	Census Tract 3031						108	136
11	1400000US06013303201	06013303201	Census Tract 3032						26	0
12	1400000US06013303202	06013303202	Census Tract 3032						95	24
13	1400000US06013303203	06013303203	Census Tract 3032						0	0
14	1400000US06013303204	06013303204	Census Tract 3032						287	85
15	1400000US06013303205	06013303205	Census Tract 3032						65	0
16	1400000US06013304001	06013304001	Census Tract 3040						0	0

Find and Replace

Find what: , Contra Costa County, California

Replace with:

Within: Sheet

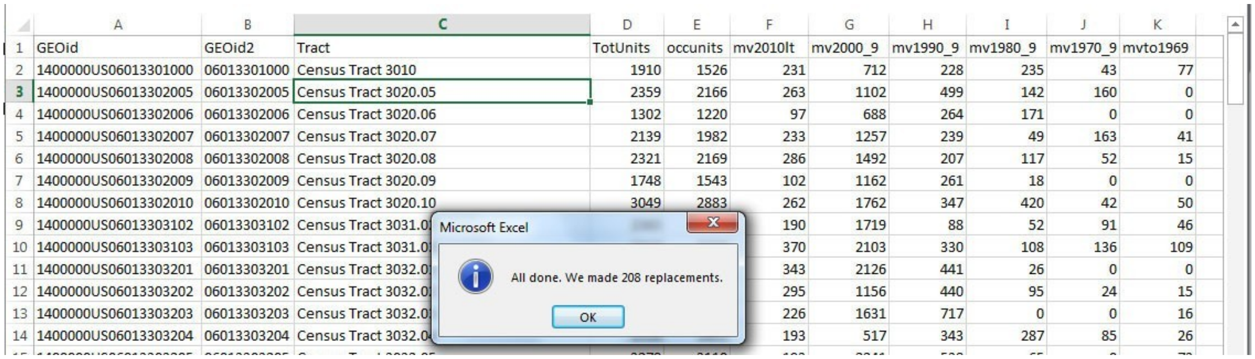
Search: By Rows

Look in: Formulas

□ Match case

□ Match entire cell contents

I left the “Replace with” field blank, so that Excel deleted all of:
, Contra Costa County, California



	A	B	C	D	E	F	G	H	I	J	K
1	GEOid	GEOid2	Tract	TotUnits	occunits	mv2010lt	mv2000_9	mv1990_9	mv1980_9	mv1970_9	mvto1969
2	1400000US06013301000	06013301000	Census Tract 3010	1910	1526	231	712	228	235	43	77
3	1400000US06013302005	06013302005	Census Tract 3020.05	2359	2166	263	1102	499	142	160	0
4	1400000US06013302006	06013302006	Census Tract 3020.06	1302	1220	97	688	264	171	0	0
5	1400000US06013302007	06013302007	Census Tract 3020.07	2139	1982	233	1257	239	49	163	41
6	1400000US06013302008	06013302008	Census Tract 3020.08	2321	2169	286	1492	207	117	52	15
7	1400000US06013302009	06013302009	Census Tract 3020.09	1748	1543	102	1162	261	18	0	0
8	1400000US06013302010	06013302010	Census Tract 3020.10	3049	2883	262	1762	347	420	42	50
9	1400000US06013303102	06013303102	Census Tract 3031.0			190	1719	88	52	91	46
10	1400000US06013303103	06013303103	Census Tract 3031.0			370	2103	330	108	136	109
11	1400000US06013303201	06013303201	Census Tract 3032.0			343	2126	441	26	0	0
12	1400000US06013303202	06013303202	Census Tract 3032.0			295	1156	440	95	24	15
13	1400000US06013303203	06013303203	Census Tract 3032.0			226	1631	717	0	0	16
14	1400000US06013303204	06013303204	Census Tract 3032.0			193	517	343	287	85	26

Microsoft Excel

All done. We made 208 replacements.

2e. Save as both an .XLS and a .CSV format

Just to cover your bases, save in both formats. ArcMAP can read Excel spreadsheets directly. QGIS cannot.

Here is another tip: Save the file with a very short filename. When you do a table join in GIS, it will concatenate the filename and the column-name together. So choose the shortest possible filename that still makes sense to you. In this case, “DP04.csv” would be a good name, because it reminds me where this data came from.