Streamlining ACS Data: The Case for Tidy Data

Tidy Data from this paper:

Wickham, H. . (2014). Tidy Data. *Journal of Statistical Software*, *59* (10), 1–23. https://doi.org/10.18637/jss.v059.i10

Presented by Anna Vasylytsya (U.S. Census Bureau, xD)

Introduction

- Emerging Technology Fellow - xD
- US Census Bureau
- Background in Data Science and Public Policy Evaluation

"xD is an emerging technologies group that's advancing the delivery of data-driven services through new and transformative technologies."

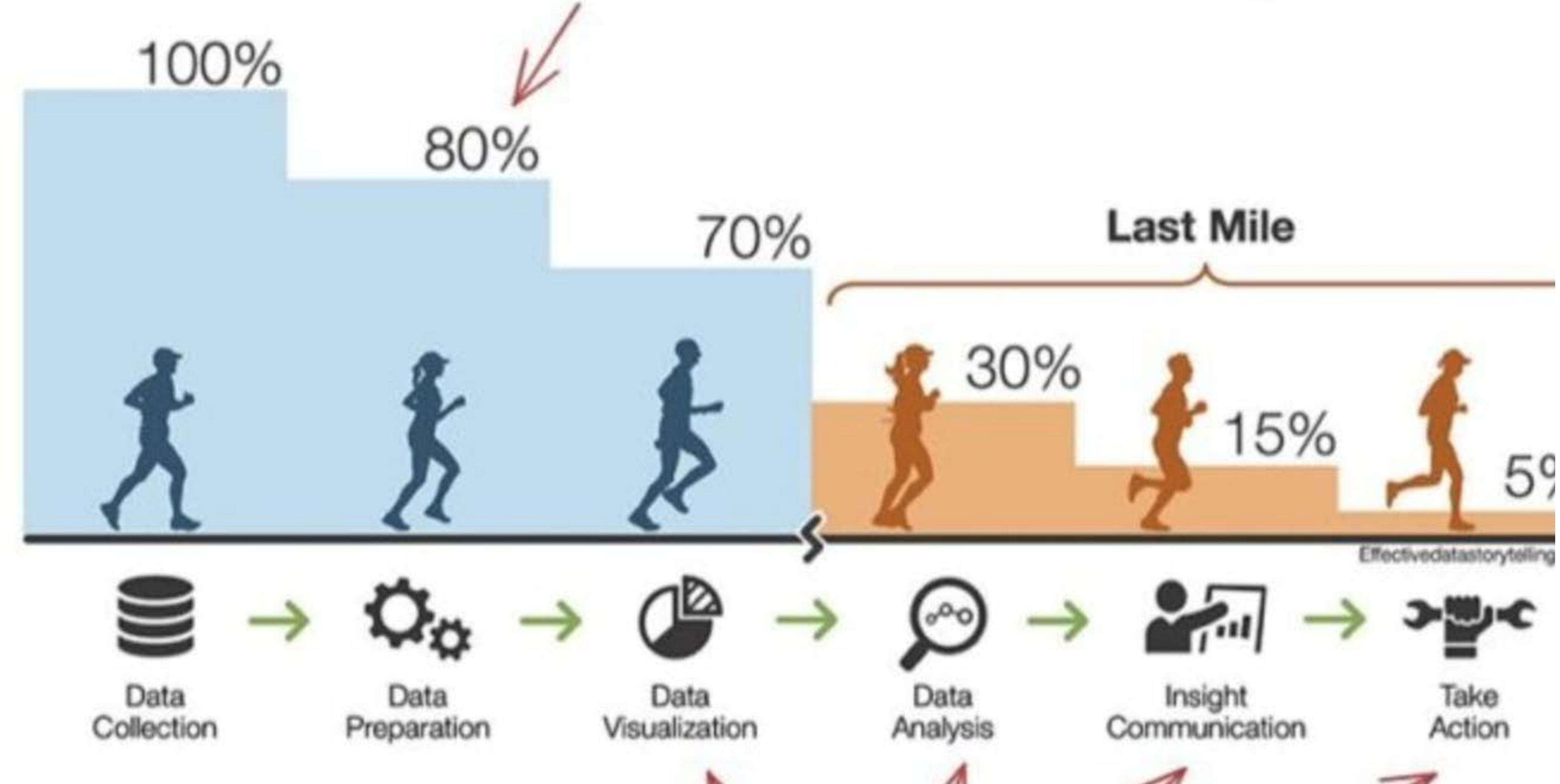
www.xD.gov

Agenda

1. Define tidy data
    a) Messy data
2. Look at Census data and its format
    a) Census data in open-source packages
3. Conclusion

There are advantages in adopting the tidy data standard for public data.

Both in data.census.gov and in the Census API



what causes the problem

where people experience the problem

Tidy data is **independent** of programming language/tool.

Data formatted in a tidy way can be used by Excel, Python, R, STATA, SAS, etc.

What is tidy data?

“Tidy data is a standard way of mapping the meaning of a dataset to its structure.”

Three characteristics of tidy data

- A **variable** contains all values that measure the same underlying attribute (e.g. age, population).

- An **observation** contains all values measured on the same unit (e.g. person, day, household, state, etc.)

- Every **value** belongs to a **variable** and an **observation** (structured in a table).

Three characteristics of tidy data structure

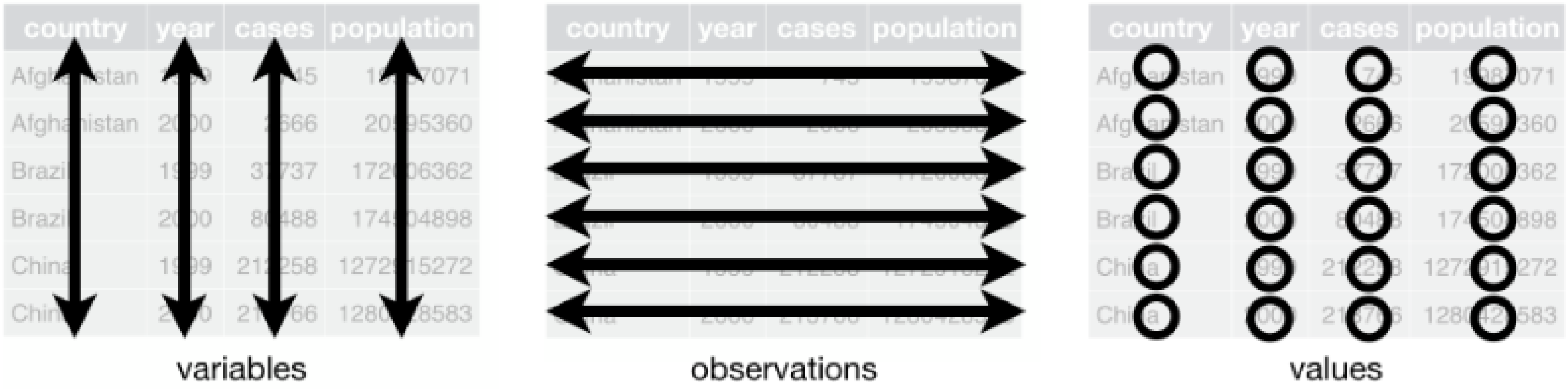


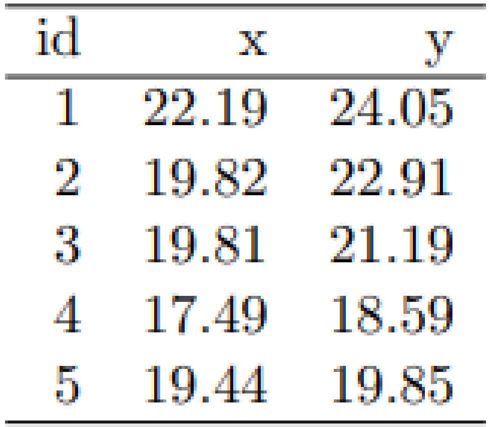
Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

Figure source: R for Data Science. Chapter 12. https://r4ds.had.co.nz/tidy-data.html

Tidy data can be wide or long

Tidy if “x and y represent length of left and right arms”

Also known as **wide** data

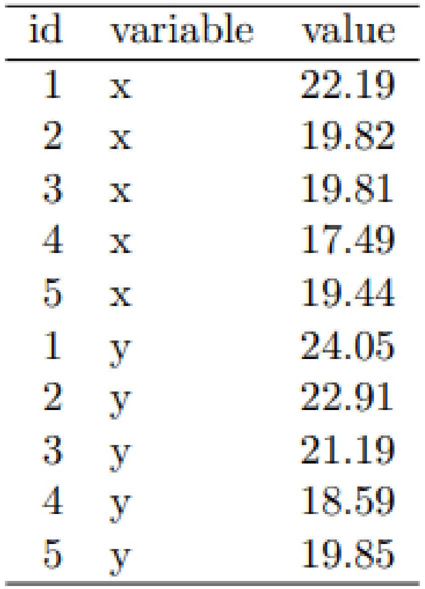


id	x	y
1	22.19	24.05
2	19.82	22.91
3	19.81	21.19
4	17.49	18.59
5	19.44	19.85

Unit of observation = id

Tidy if “x and y represent measurements on day 1 and day 10” respectively

Also known as **long** data



id	variable	value
1	x	22.19
2	x	19.82
3	x	19.81
4	x	17.49
5	x	19.44
1	y	24.05
2	y	22.91
3	y	21.19
4	y	18.59
5	y	19.85

Unit of observation = id + variable

Data tidying = structuring datasets to facilitate analysis and ML modeling

Three characteristics of tidy data

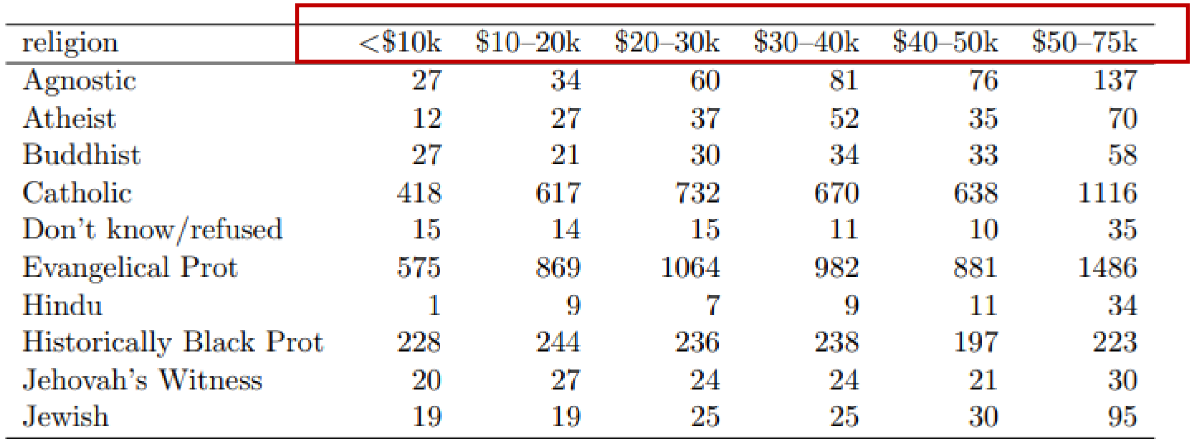
1. Each **variable** forms a **column**
2. Each **observation** forms a **row**
3. Each type of observational unit forms a **table**

- For folks that know database schema design, this is more or less the **third normal form**.

“**Messy data** is any **other** arrangement of the data.”

Column headers are values, not variable names

Messy

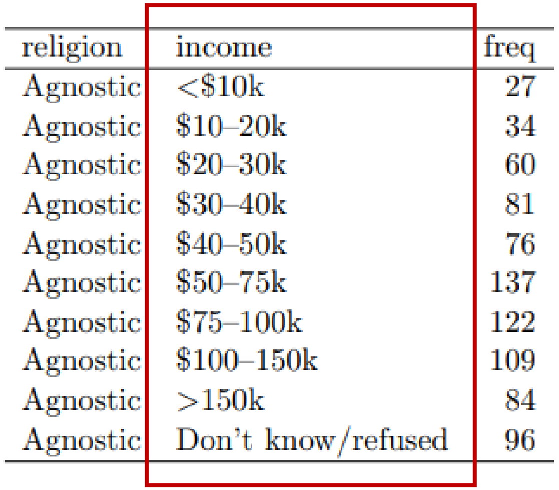


religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

The above table is from a Pew Research Center report.
** Note: this is an appropriate format for a data visualization, not great to data analysis.

Tidy



religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. column has been renamed to income, and value to freq.

* Examples from Wickham (2014).

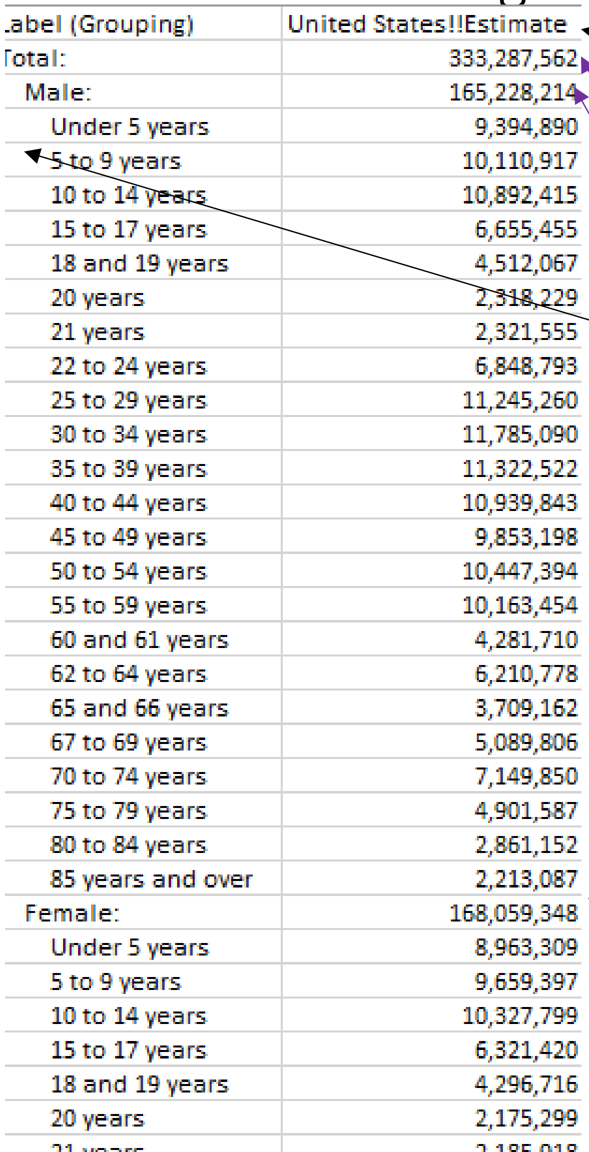
Examples

Data in this section is publicly available on data.census.gov or the Census API

Public ACS 1-year Table B01001 - 2022

download from data.census.gov

Original



.abel (Grouping)	United States!!Estimate
Total:	333,287,562
Male:	165,228,214
Under 5 years	9,394,890
5 to 9 years	10,110,917
10 to 14 years	10,892,415
15 to 17 years	6,655,455
18 and 19 years	4,512,067
20 years	2,318,229
21 years	2,321,555
22 to 24 years	6,848,793
25 to 29 years	11,245,260
30 to 34 years	11,785,090
35 to 39 years	11,322,522
40 to 44 years	10,939,843
45 to 49 years	9,853,198
50 to 54 years	10,447,394
55 to 59 years	10,163,454
60 and 61 years	4,281,710
62 to 64 years	6,210,778
65 and 66 years	3,709,162
67 to 69 years	5,089,806
70 to 74 years	7,149,850
75 to 79 years	4,901,587
80 to 84 years	2,861,152
85 years and over	2,213,087
Female:	168,059,348
Under 5 years	8,963,309
5 to 9 years	9,659,397
10 to 14 years	10,327,799
15 to 17 years	6,321,420
18 and 19 years	4,296,716
20 years	2,175,299
21 years	2,185,018

Geography identifier in column name

Indentation is challenging work with programmatically

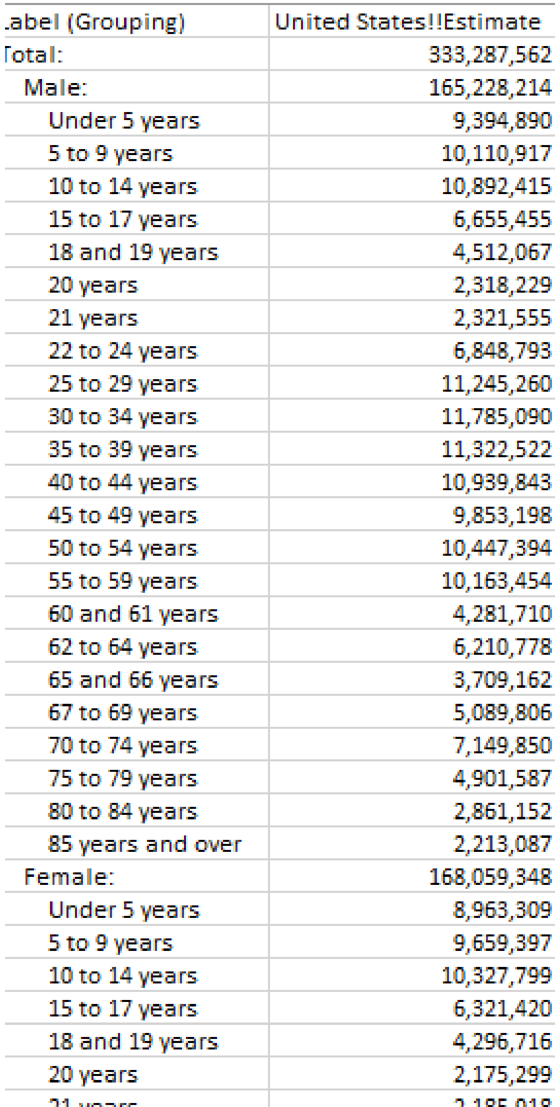
Three rows of totals = triple counting if column is summed

Public ACS 1-year Table B01001 - 2022

download from data.census.gov

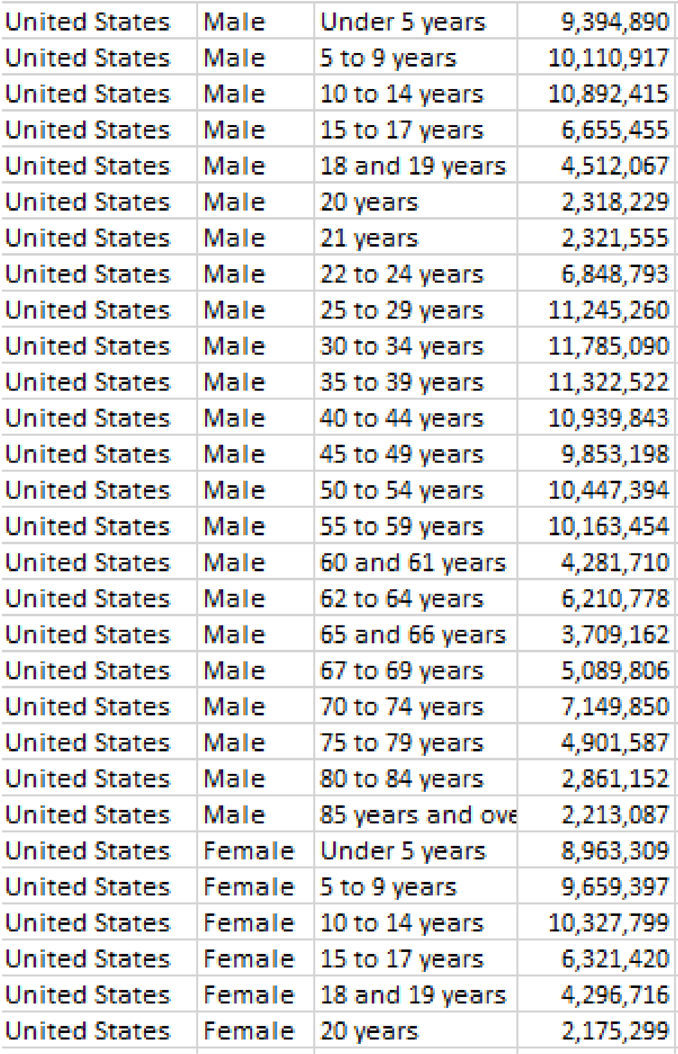
How I would restructure the original table to make it tidy:

Original



Label (Grouping)	United States!!Estimate
Total:	333,287,562
Male:	165,228,214
Under 5 years	9,394,890
5 to 9 years	10,110,917
10 to 14 years	10,892,415
15 to 17 years	6,655,455
18 and 19 years	4,512,067
20 years	2,318,229
21 years	2,321,555
22 to 24 years	6,848,793
25 to 29 years	11,245,260
30 to 34 years	11,785,090
35 to 39 years	11,322,522
40 to 44 years	10,939,843
45 to 49 years	9,853,198
50 to 54 years	10,447,394
55 to 59 years	10,163,454
60 and 61 years	4,281,710
62 to 64 years	6,210,778
65 and 66 years	3,709,162
67 to 69 years	5,089,806
70 to 74 years	7,149,850
75 to 79 years	4,901,587
80 to 84 years	2,861,152
85 years and over	2,213,087
Female:	168,059,348
Under 5 years	8,963,309
5 to 9 years	9,659,397
10 to 14 years	10,327,799
15 to 17 years	6,321,420
18 and 19 years	4,296,716
20 years	2,175,299
21 years	2,185,018

Tidy



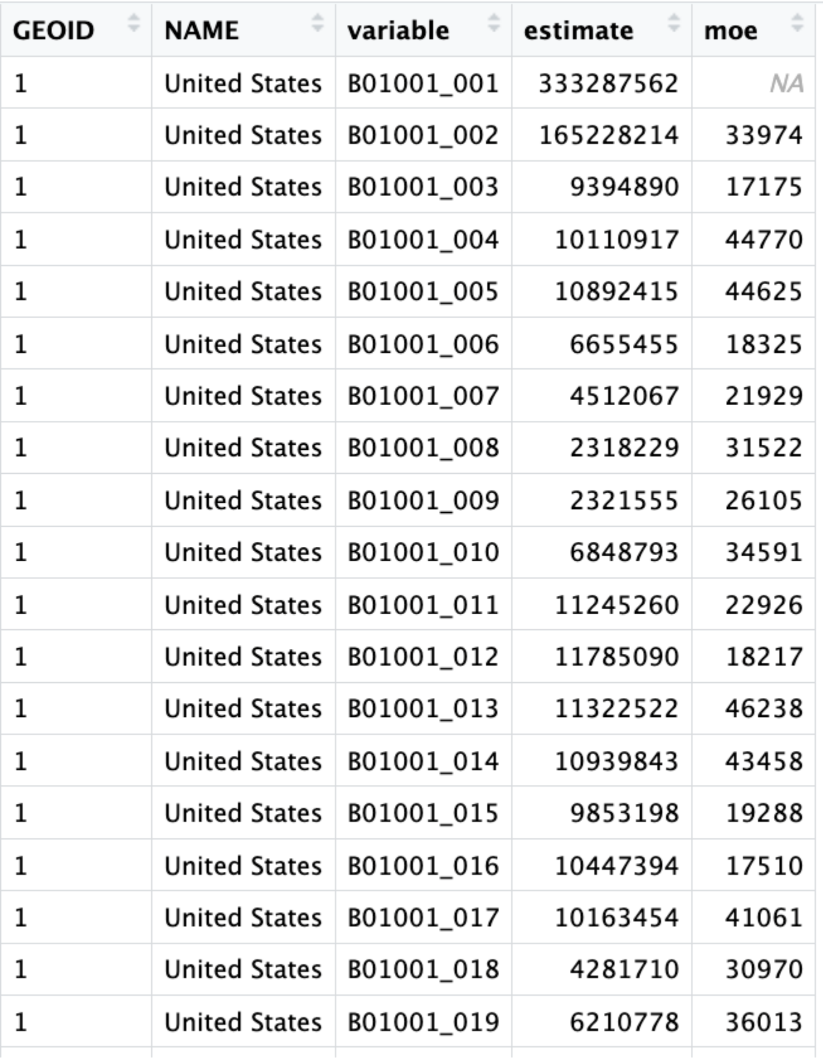
United States	Male	Under 5 years	9,394,890
United States	Male	5 to 9 years	10,110,917
United States	Male	10 to 14 years	10,892,415
United States	Male	15 to 17 years	6,655,455
United States	Male	18 and 19 years	4,512,067
United States	Male	20 years	2,318,229
United States	Male	21 years	2,321,555
United States	Male	22 to 24 years	6,848,793
United States	Male	25 to 29 years	11,245,260
United States	Male	30 to 34 years	11,785,090
United States	Male	35 to 39 years	11,322,522
United States	Male	40 to 44 years	10,939,843
United States	Male	45 to 49 years	9,853,198
United States	Male	50 to 54 years	10,447,394
United States	Male	55 to 59 years	10,163,454
United States	Male	60 and 61 years	4,281,710
United States	Male	62 to 64 years	6,210,778
United States	Male	65 and 66 years	3,709,162
United States	Male	67 to 69 years	5,089,806
United States	Male	70 to 74 years	7,149,850
United States	Male	75 to 79 years	4,901,587
United States	Male	80 to 84 years	2,861,152
United States	Male	85 years and ove	2,213,087
United States	Female	Under 5 years	8,963,309
United States	Female	5 to 9 years	9,659,397
United States	Female	10 to 14 years	10,327,799
United States	Female	15 to 17 years	6,321,420
United States	Female	18 and 19 years	4,296,716
United States	Female	20 years	2,175,299

Census data in open-source packages

tidycensus (R) and censusdis (Python)

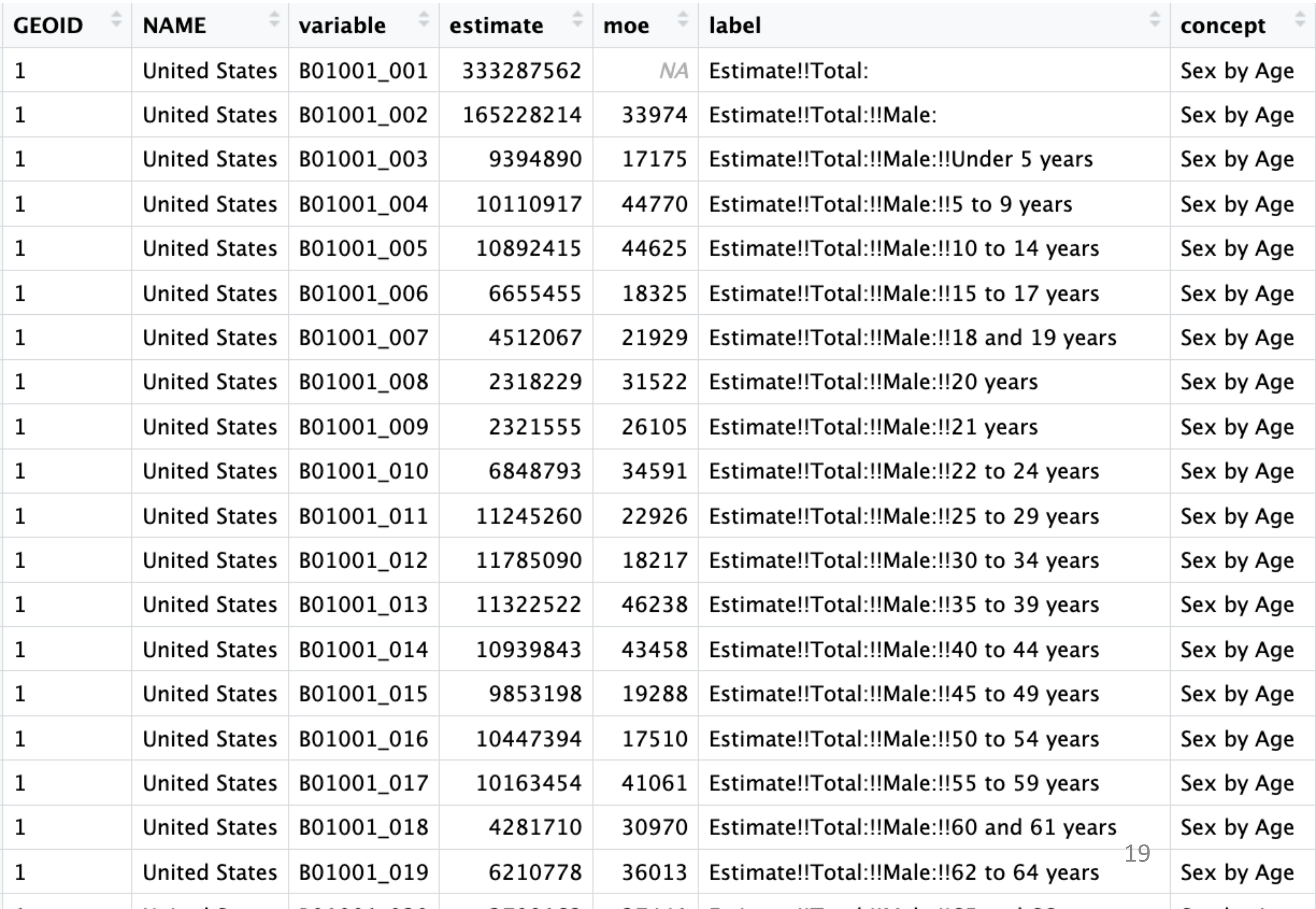
Same Table returned using **tidycensus**

Long format



GEOID	NAME	variable	estimate	moe
1	United States	B01001_001	333287562	NA
1	United States	B01001_002	165228214	33974
1	United States	B01001_003	9394890	17175
1	United States	B01001_004	10110917	44770
1	United States	B01001_005	10892415	44625
1	United States	B01001_006	6655455	18325
1	United States	B01001_007	4512067	21929
1	United States	B01001_008	2318229	31522
1	United States	B01001_009	2321555	26105
1	United States	B01001_010	6848793	34591
1	United States	B01001_011	11245260	22926
1	United States	B01001_012	11785090	18217
1	United States	B01001_013	11322522	46238
1	United States	B01001_014	10939843	43458
1	United States	B01001_015	9853198	19288
1	United States	B01001_016	10447394	17510
1	United States	B01001_017	10163454	41061
1	United States	B01001_018	4281710	30970
1	United States	B01001_019	6210778	36013

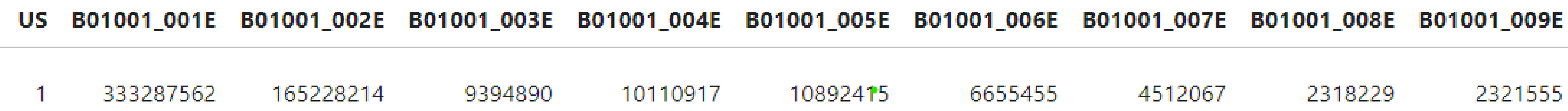
After joining variable labels:



GEOID	NAME	variable	estimate	moe	label	concept
1	United States	B01001_001	333287562	NA	Estimate!!Total:	Sex by Age
1	United States	B01001_002	165228214	33974	Estimate!!Total:!!Male:	Sex by Age
1	United States	B01001_003	9394890	17175	Estimate!!Total:!!Male:!!Under 5 years	Sex by Age
1	United States	B01001_004	10110917	44770	Estimate!!Total:!!Male:!!5 to 9 years	Sex by Age
1	United States	B01001_005	10892415	44625	Estimate!!Total:!!Male:!!10 to 14 years	Sex by Age
1	United States	B01001_006	6655455	18325	Estimate!!Total:!!Male:!!15 to 17 years	Sex by Age
1	United States	B01001_007	4512067	21929	Estimate!!Total:!!Male:!!18 and 19 years	Sex by Age
1	United States	B01001_008	2318229	31522	Estimate!!Total:!!Male:!!20 years	Sex by Age
1	United States	B01001_009	2321555	26105	Estimate!!Total:!!Male:!!21 years	Sex by Age
1	United States	B01001_010	6848793	34591	Estimate!!Total:!!Male:!!22 to 24 years	Sex by Age
1	United States	B01001_011	11245260	22926	Estimate!!Total:!!Male:!!25 to 29 years	Sex by Age
1	United States	B01001_012	11785090	18217	Estimate!!Total:!!Male:!!30 to 34 years	Sex by Age
1	United States	B01001_013	11322522	46238	Estimate!!Total:!!Male:!!35 to 39 years	Sex by Age
1	United States	B01001_014	10939843	43458	Estimate!!Total:!!Male:!!40 to 44 years	Sex by Age
1	United States	B01001_015	9853198	19288	Estimate!!Total:!!Male:!!45 to 49 years	Sex by Age
1	United States	B01001_016	10447394	17510	Estimate!!Total:!!Male:!!50 to 54 years	Sex by Age
1	United States	B01001_017	10163454	41061	Estimate!!Total:!!Male:!!55 to 59 years	Sex by Age
1	United States	B01001_018	4281710	30970	Estimate!!Total:!!Male:!!60 and 61 years	Sex by Age
1	United States	B01001_019	6210778	36013	Estimate!!Total:!!Male:!!62 to 64 years	Sex by Age

Same Table returned using **censusdis**

Wide format

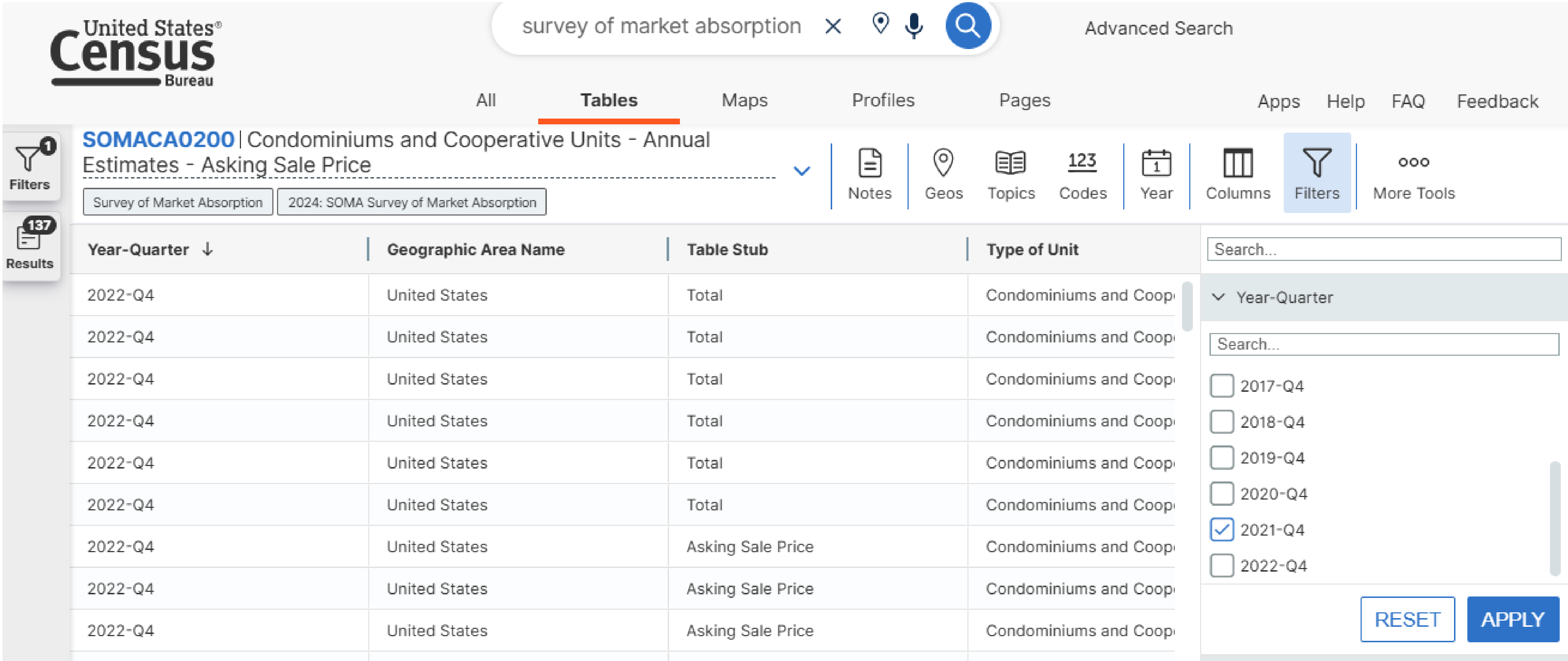


US	B01001_001E	B01001_002E	B01001_003E	B01001_004E	B01001_005E	B01001_006E	B01001_007E	B01001_008E	B01001_009E
1	333287562	165228214	9394890	10110917	10892415	6655455	4512067	2318229	2321555

What could tidy data enable?

Makes data easier to use for people and machines!

Tidy data enables filtering and sorting



SOMACA0200 | Condominiums and Cooperative Units - Annual Estimates - Asking Sale Price

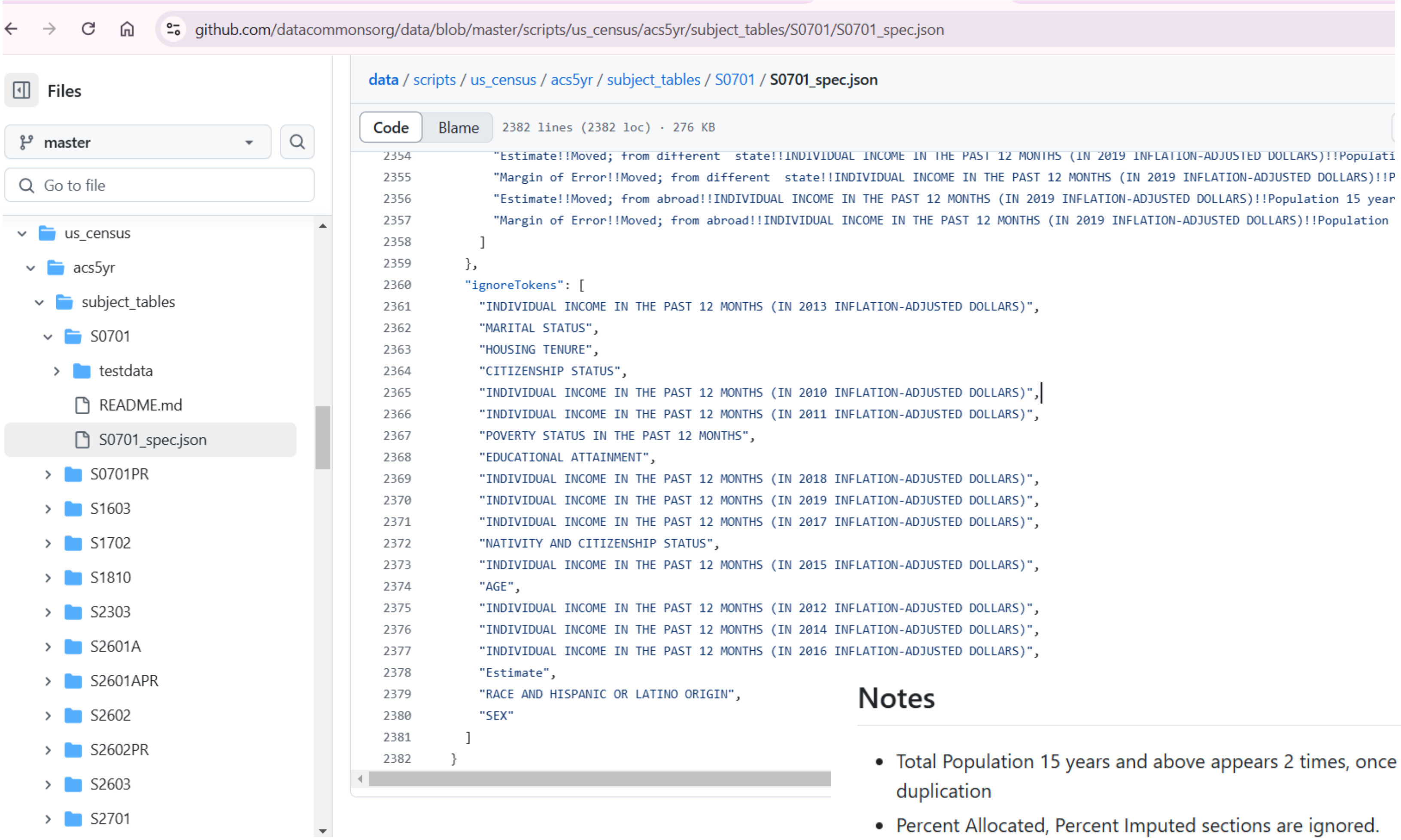
Survey of Market Absorption | 2024: SOMA Survey of Market Absorption

Year-Quarter ↓	Geographic Area Name	Table Stub	Type of Unit
2022-Q4	United States	Total	Condominiums and Coop
2022-Q4	United States	Total	Condominiums and Coop
2022-Q4	United States	Total	Condominiums and Coop
2022-Q4	United States	Total	Condominiums and Coop
2022-Q4	United States	Total	Condominiums and Coop
2022-Q4	United States	Total	Condominiums and Coop
2022-Q4	United States	Asking Sale Price	Condominiums and Coop
2022-Q4	United States	Asking Sale Price	Condominiums and Coop
2022-Q4	United States	Asking Sale Price	Condominiums and Coop

Year-Quarter

- □ 2017-Q4
- □ 2018-Q4
- □ 2019-Q4
- □ 2020-Q4
- ☑ 2021-Q4
- □ 2022-Q4

Tidy data reduces data transformation time and duplication of effort for data users


github.com/datacommonsorg/data/blob/master/scripts/us_census/acs5yr/subject_tables/S0701/S0701_spec.json
Files
master
Go to file
us_census
acs5yr
subject_tables
S0701
testdata
README.md
S0701_spec.json
S0701PR
S1603
S1702
S1810
S2303
S2601A
S2601APR
S2602
S2602PR
S2603
S2701
data / scripts / us_census / acs5yr / subject_tables / S0701 / S0701_spec.json
Code
Blame
2382 lines (2382 loc) · 276 KB

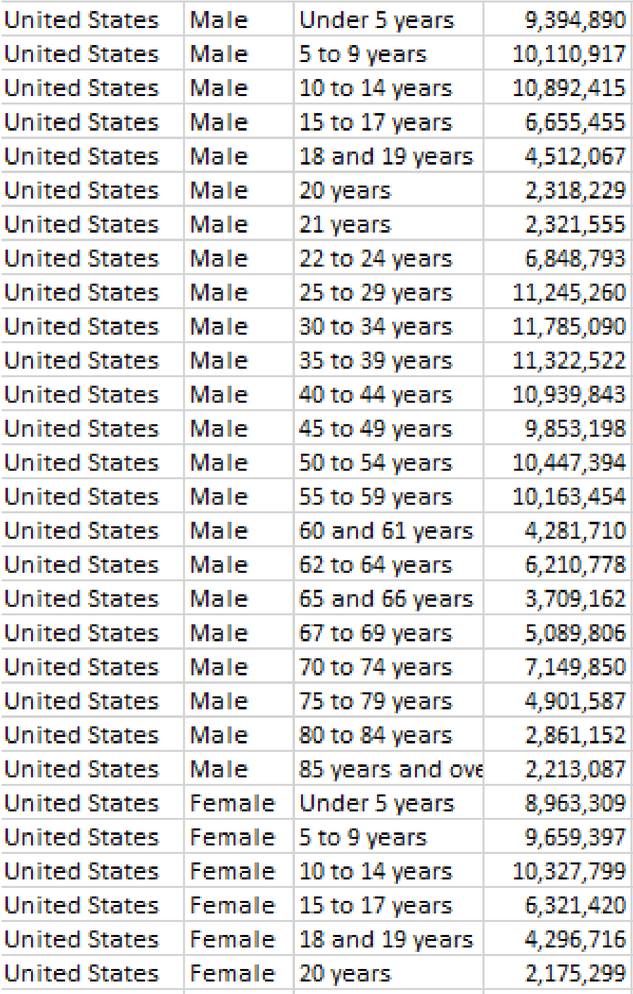
```
      "Estimate!!Moved; from different  state!!INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)!!Populati
      "Margin of Error!!Moved; from different  state!!INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)!!P
      "Estimate!!Moved; from abroad!!INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)!!Population 15 year
      "Margin of Error!!Moved; from abroad!!INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)!!Population
    ]
  },
  "ignoreTokens": [
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2013 INFLATION-ADJUSTED DOLLARS)",
    "MARITAL STATUS",
    "HOUSING TENURE",
    "CITIZENSHIP STATUS",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2010 INFLATION-ADJUSTED DOLLARS)",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2011 INFLATION-ADJUSTED DOLLARS)",
    "POVERTY STATUS IN THE PAST 12 MONTHS",
    "EDUCATIONAL ATTAINMENT",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2018 INFLATION-ADJUSTED DOLLARS)",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2017 INFLATION-ADJUSTED DOLLARS)",
    "NATIVITY AND CITIZENSHIP STATUS",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2015 INFLATION-ADJUSTED DOLLARS)",
    "AGE",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2012 INFLATION-ADJUSTED DOLLARS)",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2014 INFLATION-ADJUSTED DOLLARS)",
    "INDIVIDUAL INCOME IN THE PAST 12 MONTHS (IN 2016 INFLATION-ADJUSTED DOLLARS)",
    "Estimate",
    "RACE AND HISPANIC OR LATINO ORIGIN",
    "SEX"
  ]
}
```

Notes

- Total Population 15 years and above appears 2 times, once with marital status, once with income, one of it is ignored to prevent duplication
- Percent Allocated, Percent Imputed sections are ignored.
- Some Median Values appear as a Range rather than single value

Tidy data enables easier data joins for data users

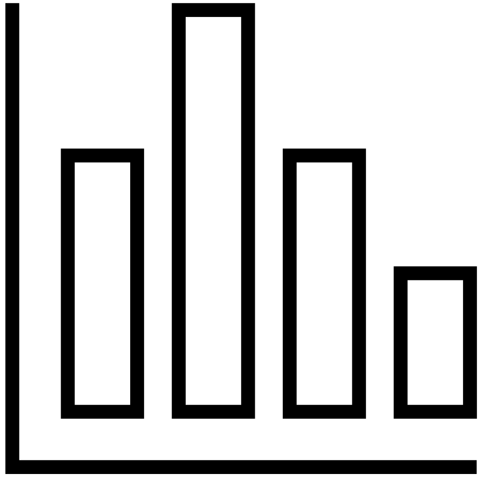
Tidy



United States	Male	Under 5 years	9,394,890
United States	Male	5 to 9 years	10,110,917
United States	Male	10 to 14 years	10,892,415
United States	Male	15 to 17 years	6,655,455
United States	Male	18 and 19 years	4,512,067
United States	Male	20 years	2,318,229
United States	Male	21 years	2,321,555
United States	Male	22 to 24 years	6,848,793
United States	Male	25 to 29 years	11,245,260
United States	Male	30 to 34 years	11,785,090
United States	Male	35 to 39 years	11,322,522
United States	Male	40 to 44 years	10,939,843
United States	Male	45 to 49 years	9,853,198
United States	Male	50 to 54 years	10,447,394
United States	Male	55 to 59 years	10,163,454
United States	Male	60 and 61 years	4,281,710
United States	Male	62 to 64 years	6,210,778
United States	Male	65 and 66 years	3,709,162
United States	Male	67 to 69 years	5,089,806
United States	Male	70 to 74 years	7,149,850
United States	Male	75 to 79 years	4,901,587
United States	Male	80 to 84 years	2,861,152
United States	Male	85 years and ove	2,213,087
United States	Female	Under 5 years	8,963,309
United States	Female	5 to 9 years	9,659,397
United States	Female	10 to 14 years	10,327,799
United States	Female	15 to 17 years	6,321,420
United States	Female	18 and 19 years	4,296,716
United States	Female	20 years	2,175,299

Can join this table on three variables

Reduced transformation burden makes visualizations and mapping quicker for data users



Benefits of tidy data for Census

The tidy data standard will:
make Census Data easier to use it for a **varied** set of users.

- Census developers developing public facing applications
- Folks using excel to analyze/present the data
- Journalists visualizing the data
- Data scientists creating ML models
- Package developers

Thank you

anna.vasylytsya@census.gov

inquiries@xd.gov