Validating data analysis through benchmark

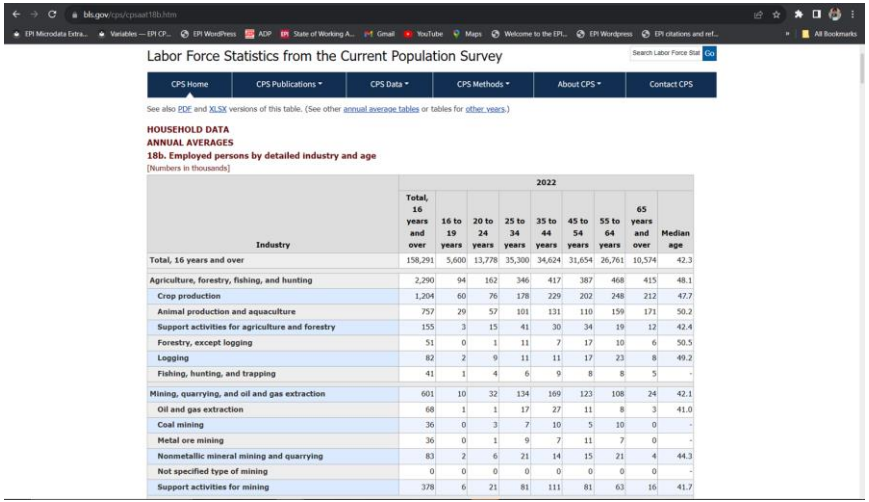
A hands-on programming workshop for analyzing public source data

About Me

- Jori Kandra
  - Reproducibility pipelines
  - Increasing replicability
  - Building out data programming

What is benchmarking & data validation?

- Comparing internal data to externally (or sometimes internally) published sources
  - Census & Bureau of Labor Statistics



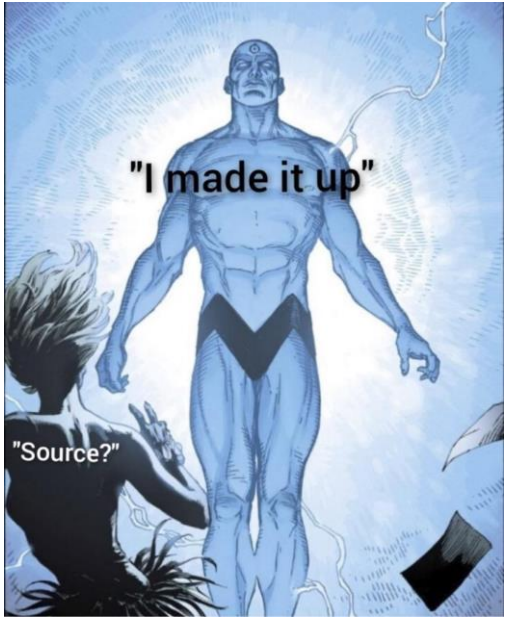
Labor Force Statistics from the Current Population Survey

CPS Home | CPS Publications | CPS Data | CPS Methods | About CPS | Contact CPS

See also PDF and XLSX versions of this table. (See other annual average tables or tables for other years.)

HOUSEHOLD DATA
ANNUAL AVERAGES
18b. Employed persons by detailed industry and age
[Numbers in thousands]

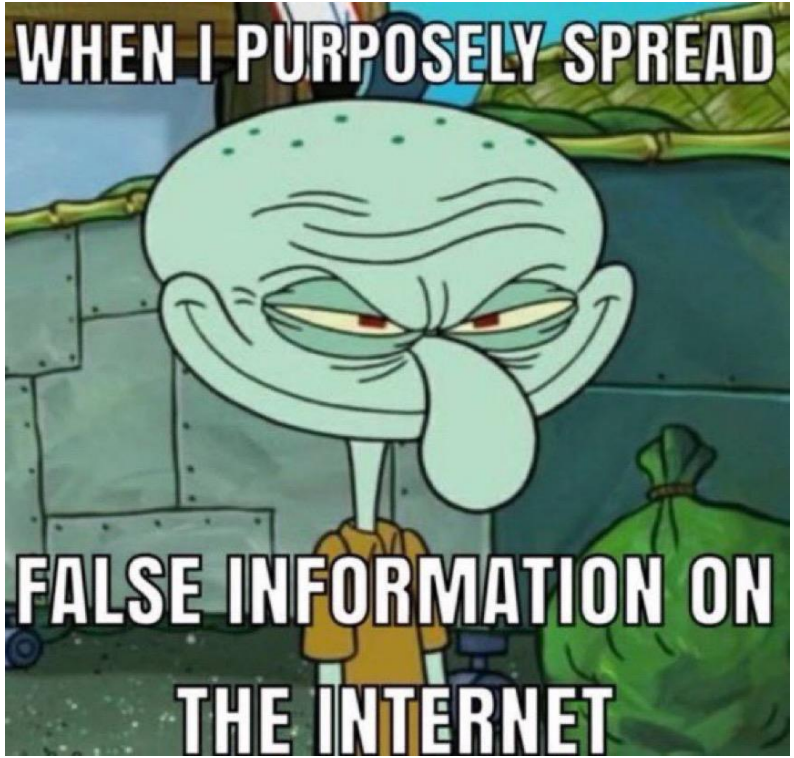
Industry	2022								
	Total, 16 years and over	16 to 19 years	20 to 24 years	25 to 34 years	35 to 44 years	45 to 54 years	55 to 64 years	65 years and over	Median age
Total, 16 years and over	158,291	5,600	13,778	35,300	34,624	31,654	26,761	10,574	42.3
Agriculture, forestry, fishing, and hunting	2,290	94	162	346	417	387	468	415	48.1
Crop production	1,204	60	76	178	229	202	248	212	47.7
Animal production and aquaculture	757	29	57	101	131	110	159	171	50.2
Support activities for agriculture and forestry	155	3	15	41	30	34	19	12	42.4
Forestry, except logging	51	0	1	11	7	17	10	6	50.5
Logging	82	2	9	11	11	17	23	8	49.2
Fishing, hunting, and trapping	41	1	4	6	9	8	8	5	-
Mining, quarrying, and oil and gas extraction	601	10	32	134	169	123	108	24	42.1
Oil and gas extraction	68	1	1	17	27	11	8	3	41.0
Coal mining	36	0	3	7	10	5	10	0	-
Metal ore mining	36	0	1	9	7	11	7	0	-
Nonmetallic mineral mining and quarrying	83	2	6	21	14	15	21	4	44.3
Not specified type of mining	0	0	0	0	0	0	0	0	-
Support activities for mining	378	6	21	81	111	81	63	16	41.7



Source(s):
Dude trust me

Importance of benchmarking

- Accuracy of research
  - Incorrect data can be costly & unethical
- Starting point for analysis
  - Verifying universe
  - Testing definitions
- Extend capabilities of organization’s research



Hands-on example

Older workers were devastated by the pandemic downturn and continue to face adverse employment outcomes

EPI testimony for the Senate Special Committee on Aging

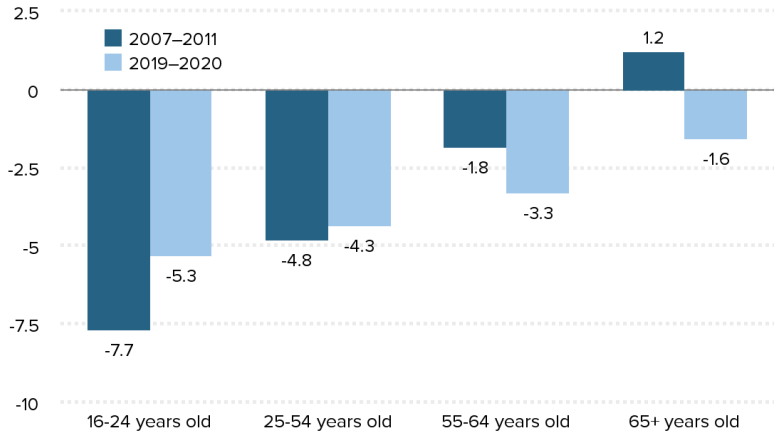
Testimony • By **Elise Gould** • April 29, 2021

Testimony prepared for the Senate Special Committee on Aging hearing on "A Changing Workforce: Supporting Older Workers Amid the COVID-19 Pandemic and Beyond"

Elise Gould
Senior Economist
Economic Policy Institute

Older workers face a more difficult recovery from the COVID-19 recession

Change in employment-to-population ratios during the Great Recession and the COVID-19 recession, 2007–2011 and 2019–2020, by age group



Source: Authors' analysis of EPI Current Population Survey Extracts, Version 1.0.14 (2021), https://microdata.epi.org.

Economic Policy Institute

EARN code library

economic.github.io/earn_code_library/code/prev_bootcamps.html

EPI Microdata Extra... | Variables — EPI CP... | EPI WordPress | ADP | State of Working A... | Gmail | YouTube | Maps | Welcome to the EPI... | EPI Wordpress | EPI citations and ref... | Digest of Education...

EARN Code Library

- Home
  - About
- EPI-built resources
  - R libraries
  - Stata packages
- R code
  - Historical state union density
  - Annotated examples
- Stata code
  - Wage deciles by state
- EARNCon Resources
  - Visualization workshop 2023
  - Benchmarking workshop
  - 2021 and 2022 data bootcamps

2021 and 2022 data bootcamps

EPI Bootcamp Website

2022 Data analysis in R (Video)

Data analysis in R slides
Historical minimum wage data
CPI-U-RS data
EPI CPS ORG data

2022 Data bootcamp: Code Library Presentation (Video)

Code library powerpoint

2022 Data analysis in Stata (Video)

On this page

- 2022 Data analysis in R (Video)
- 2022 Data bootcamp: Code Library Presentation (Video)
- 2022 Data analysis in Stata (Video)
- 2021 First steps in Stata (Video)
- 2021 Using Stata effectively (Video)
- 2021 First steps in R (Video)
- 2021 Using R effectively (Video)

EPI Microdata Extracts

EPI Microdata Extracts

Home | How to use the EPI extracts | Citations & sources | Variables | Methodology | Changes

Home

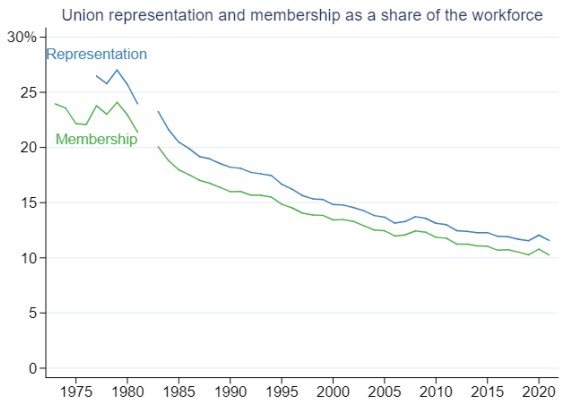
EPI Microdata Extracts

Download the data

The latest EPI CPS extracts are available in Stata format:

Extract	Years	File	Version
CPS Basic monthly	1976 - 2023m8	epi_cpsbasic_2000_2023.zip epi_cpsbasic_1976_1999.zip	1.0.43
CPS May	1973 - 1981	epi_cpsmay_1973_1981.zip	1.0.43
CPS ORG	1979 - 2023m8	epi_cpsorg_1979_2023.zip	1.0.43

union: Represented by a union as a member or covered by contract



Description

`union` indicates whether the respondent is represented by a union as a member or covered by contract. See also `uncov` and `unmem`.

EPI microdata extracts

Availability

Sample	Years
Basic	1984 - present
May	1977 - 1981
ORG	1983 - present

Note

Union status based on both membership and coverage is not available prior to 1977 or in 1982. On the Basic files, union is available 1984 - present.

Values

Value	Label
0	Not union represented
1	Union represented

Code

Variable creation

```
*****************************************************************************
* Union represented: member or covered
*****************************************************************************
gen byte union = .
replace union = 0 if unmem ~= . | uncov ~= .
replace union = 1 if unmem == 1
replace union = 1 if uncov == 1

* recode as missing in may before 1977 because no union coverage available
if $maycps == 1 {
    if $date <= tm(1976m12) {
        replace union = .
    }
}

* recode as missing in basic before 1984 because no union coverage available
if $monthlycps == 1 & $earnerinfo == 1 & $basicfile == 1 {
    if $date < tm(1984m1) {
        replace union = .
    }
}

* recode as missing in org before 1983 because no union coverage available
if $monthlycps == 1 & $earnerinfo == 1 & $basicfile == 0 {
    if $date < tm(1983m1) {
        replace union = .
    }
}

lab var union "Represented by a union as a member or covered by contract"
lab def union 1 "Union represented" 0 "Not union represented"
lab val union union
notes union: Not defined when union coverage is missing
notes union: Only available in 1977-1981 May, 1983-present ORG
notes union: Not available on 1976-1983 basic monthly files
notes union: Not available prior to 1977 or in 1982
notes union: Derived from unmem and uncov
```

Bureau of Labor Statistics: bls.gov

An official website of the United States government Here is how you know

U.S. BUREAU OF LABOR STATISTICS

HOME | SUBJECTS | DATA TOOLS | PUBLICATIONS | ECONOMIC RELEASES | CLASSROOM | BETA

Announcement

» BLS is seeking your comments on our plan to improve the Import and Export Price Indexes. Read More

OCT 03

Job openings increase to 9.6 million in August; hires and total separations change little

The number of job openings increased to 9.6 million on the last business day of August. The number of hires and total separations changed little at 5.9 million and 5.7 million, respectively.

HTML | PDF | RSS | Charts

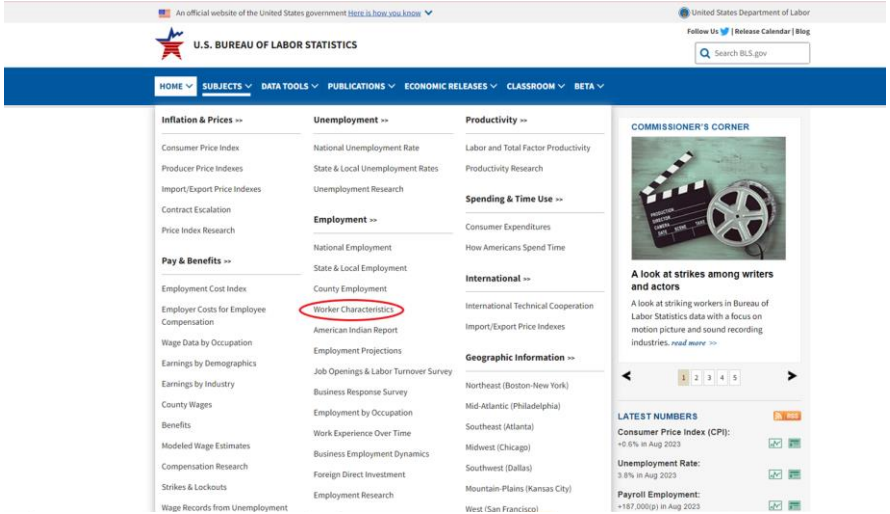
09/27/2023 **August jobless rates up over the year in 189 of 389 metro areas; payroll jobs up in 66**

09/21/2023 **Retirement benefits are available to 94% of private industry union workers in March 2023**

09/21/2023 **In 2021-22, 14% of the U.S. population, or 37.1 million people, provide unpaid eldercare**

09/19/2023 **July job openings rates down in 9 states, up in 3; quits down in 7**

All Releases >>



BLS CPS Table 3

HOUSEHOLD DATA
ANNUAL AVERAGES
3. Employment status of the civilian noninstitutional population by age, sex, and race
[Numbers in thousands]

Age, sex, and race	2020							
	Civilian noninstitutional population	Civilian labor force						Not in labor force
		Total	Percent of population	Employed		Unemployed		
				Total	Percent of population	Number	Percent of labor force	
TOTAL								
16 years and over	260,329	160,742	61.7	147,795	56.8	12,947	8.1	99,587
16 to 19 years	16,567	5,720	34.5	4,695	28.3	1,025	17.9	10,846
16 to 17 years	8,764	2,082	23.8	1,718	19.6	364	17.5	6,682
18 to 19 years	7,803	3,638	46.6	2,977	38.2	661	18.2	4,164
20 to 24 years	20,918	14,487	69.3	12,497	59.7	1,990	13.7	6,431
25 to 54 years	126,045	102,630	81.4	95,331	75.6	7,298	7.1	23,416
25 to 34 years	44,844	36,506	81.4	33,426	74.5	3,080	8.4	8,339
25 to 29 years	22,583	18,332	81.2	16,596	73.5	1,736	9.5	4,252
30 to 34 years	22,261	18,174	81.6	16,831	75.6	1,344	7.4	4,087
35 to 44 years	41,330	33,983	82.2	31,807	77.0	2,177	6.4	7,347
35 to 39 years	21,375	17,608	82.4	16,472	77.1	1,136	6.5	3,767
40 to 44 years	19,955	16,376	82.1	15,335	76.8	1,041	6.4	3,580
45 to 54 years	39,871	32,141	80.6	30,099	75.5	2,042	6.4	7,730
45 to 49 years	19,676	16,171	82.2	15,149	77.0	1,022	6.3	3,505
50 to 54 years	20,195	15,969	79.1	14,949	74.0	1,020	6.4	4,226
55 to 64 years	42,191	27,293	64.7	25,454	60.3	1,839	6.7	14,898
55 to 59 years	21,348	15,399	72.1	14,375	67.3	1,025	6.7	5,949
60 to 64 years	20,843	11,894	57.1	11,079	53.2	815	6.8	8,950
65 years and over	54,608	10,613	19.4	9,818	18.0	795	7.5	43,995

BLS CPS Table 18

HOUSEHOLD DATA
ANNUAL AVERAGES
18b. Employed persons by detailed industry and age
[Numbers in thousands]

Industry	2020								
	Total, 16 years and over	16 to 19 years	20 to 24 years	25 to 34 years	35 to 44 years	45 to 54 years	55 to 64 years	65 years and over	Median age
Total, 16 years and over	147,795	4,695	12,497	33,426	31,807	30,099	25,454	9,818	42.5