

Datasheets for Dataset 1 & Dataset 5

Motivation for Dataset Creation

Why was the dataset created?

OEWS is a semiannual survey which is created for specific tasks: measuring occupational employment and wage rates for wage and salary workers in nonfarm establishments in the United States.

What (other) tasks could the dataset be used for?

Yes.

Has the dataset been used for any tasks already?

Yes.

[STEM occupations: past, present, and future: Stella Fayer, Alan Lacey and Audrey Watson, January 2017](#)

[Occupational employment and wages in metro and non-metro areas; Stella Fayer and Audrey Watson, November 2020](#)

[Employment and Wages in Healthcare Occupations: Stella Fayer and Audrey Watson, December 2015](#)

Dataset Composition

What are the instances?

Columns	Description
Area data	area, area_title, area_type Area code, area name, area type for U.S., state, Metropolitan Statistical Area (MSA) or New England City and Town Area (NECTA)
	prim_state The primary state for the given area. "US" is used for the national estimates.
	naics, naics_title North American Industry Classification System (NAICS) code and title for the given industry
	i_group Industry level. Indicates cross-industry or NAICS sector
Occupation data	own_code Ownership type
	occ_code, occ_title, o_group The 6-digit Standard Occupational Classification (SOC) code, the SOC title and SOC occupation level for the occupation
Employment data	tot_emp Estimated total employment number rounded to nearest 10 (self-employed included)
	emp_prse Percent relative standard error (PRSE) for the employment estimate
	pct_total Percent of industry employment in the given occupation.
	jobs_1000 The number of jobs (employment) in the given occupation per 1,000 jobs in the given area
	loc_quotient The location quotient represents the ratio of an occupation's share of employment in a given area to that occupation's share of employment in the U.S. as a whole.
Wage data	h_mean, a_mean Mean hourly/annual wage
	mean_prse Percent relative standard error (PRSE) for the mean wage estimate.
	h_pctl10, h_pctl25, h_median, h_pctl75, h_pctl90 Hourly wage in different percentile
	a_pctl10, a_pctl25, a_median, a_pctl75, a_pctl90 Annual wage in different percentile
	annual, hourly "TRUE" when only hourly/annual wages are given

How many instances of each type are there?

The dataset 1 consists of 390,706 entries including wage and employment information for different occupations in different areas. The dataset 5 consists of 1,448 entries containing the standard code and detailed classification. The standard occupational classification dataset 1 uses is based on dataset 5.

Is everything included or does the data rely on external resources?

Everything is included in the datasheet.

Data Collection Process

How was the data collected? Over what time-frame was the data collected?

OEWS are calculated with data collected from employers in all industry sectors in every state and the District of Columbia.

Each year, two semiannual panels of approximately 180,000 to 185,000 sampled establishments are contacted, one panel in May and the other in November. Responses are obtained by mail, Internet or other electronic means, email, telephone, or personal visit. The May 2020 estimates are based on responses

from six semiannual panels collected over a 3-year period: May 2020/2019/2018, November 2019/2018/2017. The unweighted sampled employment of 83 million across all six semiannual panels represents approximately 56 percent of total national employment. The overall national response rate for the six panels, based on the 50 states and the District of Columbia, is 69 percent based on establishments and 66 percent based on weighted sampled employment.

If the dataset is a sample, then what is the population?

OEWS estimates are constructed from a sample of about 1.1 million establishments.

Does the dataset contain all possible instances?

The dataset does not contain all possible instances

Is there information missing from the dataset and why?

Dataset 1:

- * = indicates that a wage estimate is not available
- ** = indicates that an employment estimate is not available
- # = indicates a wage equal to or greater than \$100.00 per hour or \$208,000 per year

Dataset 5: no missing information.

Dataset Preprocessing

What preprocessing/cleaning was done?

No preprocessing/cleaning was done.

Dataset Distribution

How is the dataset distributed?

Dataset 1 comes from [The Occupational Employment and Wage Statistics \(OEWS\) survey](#). Dataset 5 comes from the [Standard Occupational Classification](#).

When will the dataset be released/first distributed?

Dataset 1 is first released on March 31, 2021.

Dataset 5 is first published on November 28, 2017.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

Bureau of Labor.

Will the dataset be updated?

Yes. Estimates are generally released in late March or early April. Please check the [OEWS homepage](#) around that time for a scheduled release date.

Legal & Ethical Considerations

If the dataset relates to people, were they informed about the data collection? Were they told what the dataset would be used for and did they consent?

Yes. Nearly all of BLS's surveys are voluntary, which means the individuals, households, and organizations selected for their survey samples can choose whether to participate.