Datasheets for Dataset 2,3,4: Age, Gender and Race Dataset

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. The collected data can be used for analysis of national labor force statistics with demographic characteristics.

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

Yes. The dataset, age, race and gender could be used as a way to identify distribution of occupation and potential discrimination in U.S.

Has the dataset been used for any tasks already?

Yes.

The early 2000s: a period of declining teen summer employment rates; Teresa L. Morisi, May 2010

American Indians and Alaska Natives in the U.S. labor force; Mary Dorinda Allard and Vernon Brundage Jr., Nov 2019

Women Still Underrepresented Among Highest Earners; U.S. Department of Labor, Mar 2006

Dataset Composition

The dataset contains employment status of different occupations by age, race and gender group.

Data Collection Process

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

Occupational Employment and Wage Statistics (OEWS) are calculated with data collected from employers in all industry sectors in metropolitan and nonmetropolitan areas in every state and the District of Columbia. OEWS estimates are constructed from a sample of about 1.1 million establishments. 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.

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

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

Dataset Preprocessing

What preprocessing/cleaning was done?

In the dataset, for privacy issues, if the wage is equal to or greater than $100.00 per hour or $208,000 per year, then the number is replaced with "#". If the number of employees is lower than 50,000, the value will be missing ('-') as it will be personal-identifiable in that case.

Dataset Distribution

How is the dataset distributed?

The dataset distributed through website. Dataset 2,3,4 can be downloaded from Age, Gender, Race.

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

U.S. Bureau of Labor Statistics.

Will the dataset be updated?

Yes. The dataset will be updated each year by BLS. 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 (e.g., their attributes) or was generated by people, were they informed about the data collection?

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.

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?

The data were collected from the Current Population Survey (CPS). Employees who filled out the survey are fully aware of the

If it relates to people, could this dataset expose people to harm or legal action?

No. The dataset is not personal identifiable so it is unlikely that it will put people to harm or legal action. But the dataset might illustrate some inequity or discrimination in the U.S., which might cause potential misunderstanding between people and different social groups.

If it relates to people, were they provided with privacy guarantees?

The dataset provides privacy guarantees by assigning a value of NaN when the number is under 50, considering that the respondents would be identifiable when an occupation has less than 50 people.

Does the dataset comply with the EU General Data Protection Regulation (GDPR)?

Yes. The dataset meets the requirement of broad applicability, informed consent, withdraw consent, transparency, and obligations to maintain data security.

Does the dataset contain information that might be considered inappropriate or offensive?

Yes. The dataset shows potential inequity in the wages and employment situation, which might be inappropriate.