Analysis on USA wage level by demographic

INTRODUCTION

This dataset provides the poverty-level wages and hourly wages by education level, gender and race. It covers the years from 1973 to 2022.
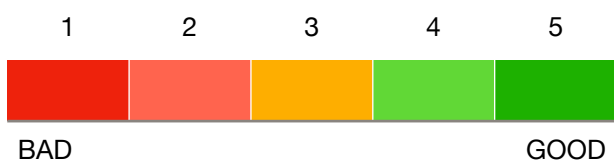
Created By: Asaniczka, Economic Policy Institute
Source: **1.** https://www.kaggle.com/datasets/asaniczka/poverty-level-wages-in-the-usa-dataset-1973-2022 **2.** https://www.kaggle.com/datasets/asaniczka/wages-by-education-in-the-usa-1973-2022
Time Range: 1973-2022

DATA QUALITY

	Mark
Completeness missing value ... no missing value	5
Data types	1
Data distribution	2
Consistency	5

1	2	3	4	5

BAD　　　　　　　　　　　　　　GOOD

USAGE RECOMMENDATION

- Explore the wage gap based on education level and identify trends over time.

- Analyze the impact of education on wages for different demographic groups (gender, race/ethnicity).

- Investigate changes in wage inequality between education groups over the decades.

- Examine the relationship between wages and economic indicators such as inflation and GDP.

BADGES

Alert Counts

4

	Num
Completness	0
Provenance	0
Collection	2
Repetitiveness	1
Variety	1
Description	1
value names	1
Compostion	1*
Racial/Gender bias	1*

* represents potential alerts

ABOUT

Objectives and alerts are corresponding to the USAGE RECOMMENDATION and ALERT COUNTS.

OBJECTIVES
- Analyze the trend of annual poverty-level wages over the years.

- Compare the hourly poverty-level wages between men and women.

- Explore the distribution of wages across different poverty wage ranges.

- Calculate the percentage change in poverty-level wages from one year to another.

- Analyze the wage gap trends over the years between different races.

- Analyze the wage gap trends over the years between different education levels.

- Analyze the wage gap trends over the years between different genders.

- Investigate the wage gap by two or three elements.

- Conduct regression analysis to analyze the wage gap while controlling for relevant factors such as education, age, and geographic location.

- Discover the potential effects of policy changes or economic transformations on the wage gap.

ALERTS

- **Repetitiveness.** There too many similar values, like high_school, men_high_school, white_high_school, black_high_school… The value is too detailed, which might influence the efficiency on finding and analyzing data.

- **Variety.** The variety is single, only wage. And data types are all numerical. You need to find more related datasets to research deeply.

- **Value names.** Some value names about poverty level is not clear, or hard to understand

- **Racial/Gender bias*.** The innate traits like race or gender might not be suitable for the cause of different wage and poverty levels.

Suggestions
Corresponding to the Alerts

- **Split.** Separate the data frame to two. One is for basic values and another one includes more detailed value.

- **Find more.** Find more related sources, such as employment rate, or do some background research.

- **Change.** Change the value names to more simple names.

- **No Racial/Gender analysis*.** Not recommend on spending too much time researching on this part.

DATASET INFORMATION

Description

1. **TELL US ABOUT THIS DATASET.**

   This dataset provides the poverty-level wages and hourly wages by education level, gender and race. It covers the years from 1973 to 2022.

2. **IS THERE AN INTENDED PURPOSE FOR THE DATASET? WHAT DOMAIN WAS IT DESIGNED FOR?**

   The purpose of this dataset is to assess what elements influence the wage.

3. **WHAT IS THE LICENSE UNDER WHICH THE DATASET IS MADE AVAILABLE?**

   Economic Policy Institute.

4. **ARE THERE ANY ADDITIONAL RESTRICTIONS UNDER WHICH THE DATASET IS MADE AVAILABLE? ARE THERE ANY LEGISLATION, CODES OF PRACTICE OR GUIDANCE FROM THE JURISDICTION OR DOMAIN IN WHICH THE DATA WAS COLLECTED ABOUT THE USE OF THIS DATA? ARE THERE TASKS FOR WHICH THE DATASET CANNOT BE USED? PLEASE PROVIDE A DESCRIPTION OR GUIDANCE.**

   No.

5. **ARE THERE TASKS FOR WHICH THE DATASET WOULD BE CAUTIONED AGAINST BEING USED? IF SO, PLEASE PROVIDE A DESCRIPTION.**

   No.

6. **HAS THE DATASET BEEN USED FOR ANY TASKS ALREADY? IF SO, PLEASE PROVIDE A DESCRIPTION AND LINKS TO PAPERS OR SYSTEMS USING THE DATASET?**

   No.

Composition

1. **DOES THE DATASET HAVE A METADATA REPOSITORY OR DATA DICTIONARY? IF YES, PLEASE PROVIDE THE LINK AND IF NOT, PLEASE EXPLAIN WHAT EACH FIELD MEANS.**

   No data dictionary exists for this dataset.

2. **IN THE CASE OF EXPLICITLY RELATED INSTANCES, ARE THESE RELATIONSHIPS INDICATED? (E.G., RELATING AN INSTANCE ABOUT A USER TO INSTANCES ABOUT THEIR POSTS)**

   No.

3. **IS ANY INFORMATION MISSING FROM INDIVIDUAL?**

   No missing values.

4. **IS YOUR DATASET A SAMPLE? IF SO, WHAT WAS THE SAMPLING STRATEGY USED (E.G. DETERMINISTIC, PROBABILISTIC WITH SPECIFIC SAMPLING PROBABILITIES), AND DOES IT ACCURATELY REPRESENTS THE INTENDED OUTPUT?**

   No

5. **WAS ANY PREPROCESSING/CLEANING/ LABELING OF THE DATA DONE? IF SO, PLEASE PROVIDE A DESCRIPTION.**

   Yes, the dataset comes from two datasets ("Poverty-Level Wages in the USA Dataset " and "Wages by Education in the USA") and has two new columns.
   ("college_poverty_level" and "diff")

6. **IF THE DATA WAS PROCESSED, WAS THE RAW DATA SAVED IN ADDITION TO THE PREPROCESSED/CLEANED/ LABELED DATA? IF YES, PLEASE PROVIDE A LINK TO THE RAW DATA.**

   1. https://www.kaggle.com/datasets/ asaniczka/poverty-level-wages-in-the-usa-dataset-1973-2022

   2. https://www.kaggle.com/datasets/ asaniczka/wages-by-education-in-the-usa-1973-2022

7. **ARE THERE RECOMMENDED DATA SPLITS? IF SO, PLEASE PROVIDE A DESCRIPTION OF THESE SPLITS, EXPLAINING THE RATIONALE BEHIND THEM.**

   Yes. The dataset should be more simple and clear. It could separate to the basic values and other more detailed values.

8. **ARE THERE ANY ERRORS, SOURCES OF NOISE, OR REDUNDANCIES IN THE DATASET? IF SO, PLEASE PROVIDE A DESCRIPTION.**

   No.

9. **DOES THE DATASET INCLUDE INFORMATION FROM UPSTREAM SOURCES? IF SO, NAME THESE SOURCES, THEIR STABILITY, AND ANY KNOWN USAGE LIMITATIONS.**

   No.

10. **IF THE DATASTET INCLUDES INFORMATION FROM UPSTREAM SOURCES, IS THIS DATASET TECHNICALLY DEPENDENT UPON THOSE SOURCES?**

    N/A

11. **IS THERE CONFIDENTIAL DATA INCLUDED IN THIS DATASET?**

    No.

12. **DOES THE DATASET IDENTIFY ANY SUBPOPULATIONS (E.G., BY AGE, GENDER)? IF YES, PLEASE INDICATE THEIR DISTRIBUTIONS WITHIN THE DATASET WHERE POSSIBLE / AVAILABLE.**

    Yes. By gender and race.

Provenance

1. **WHO CREATED THE DATASET AND ON BEHALF OF WHICH ENTITY?**

   Asaniczka, in Economic Policy Institute

2. **HOW IS / WAS THE COLLECTION AND MANAGEMENT OF THIS DATASET FUNDED?**

   Non-profit private funding.

Collection

1. **OVER WHAT TIMEFRAME WAS THE DATA COLLECTED? IF THE DATA CONTINUES TO BE UPDATED, PLEASE INDICATE.**

   It's from 1973 to 2022 and it's not updated so far.

2. **WHAT MECHANISMS OR PROCEDURES WERE USED TO COLLECT THE DATA?**

   I have no information on how it was collected.

3. **WHAT IS THE RELATIONSHIP BETWEEN THE DATASET COLLECTOR AND OWNERS OR MANAGERS OF THE DATASET?**

   Superior and subordinate.

4. **IF INDIVIDUALS' DATA IS INCLUDED IN THIS DATASET, DID THOSE INDIVIDUALS CONSENT TO THE COLLECTION AND USE OF THEIR DATA? IF SO, PLEASE DESCRIBE CONSENT PROCEDURE.**

   N/A

5. **IF INDIVIDUALS' DATA IS INCLUDED IN THIS DATASET, WAS THIS DATA ALTERED TO ENSURE HIGHER LEVELS OF PRIVACY? IF YES, PLEASE DESCRIBE ANY PRIVACY PROCEDURES FOLLOWED WITH REGARDS TO THIS DATASET (ANONYMIZATION EFFORTS, PRIVACY PROTOCOLS, SUPRESSION TECHNIQUES, ETC).**

   N/A

6. **HAS THE DATA BEEN REVIEWED FOR QUALITY?**

   Yes.

7. **WERE ANY ETHICAL REVIEW PROCESSES CONDUCTED (E.G., BY AN INSTITUTIONAL REVIEW BOARD)? IF SO, PLEASE PROVIDE A DESCRIPTION OF THESE REVIEW PROCESSES, INCLUDING THE OUTCOMES, AS WELL AS A LINK OR OTHER ACCESS POINT TO ANY SUPPORTING DOCUMENTATION.**

   No.

Management

1. **HOW IS THE DATA BEING MANANGED AT REST AND IN TRANSIT?**

   Internally.

2. **HOW CAN THE DATA BE ACCESSED? IF THIS IS GOING TO CHANGE OVER TIME, PLEASE INDICATE THIS.**

   It could be access online and it might not be changed.

3. **WILL THE DATASET BE DISTRIBUTED TO ANY OTHER INDIVIDUALS/THIRD PARTIES (E.G. COMPANIES, INSTITUTIONS, ORGANIZATIONS)? IF YES, WHEN WILL THE DATASET BE DISTRIBUTED AND HOW WILL THE DATASET BE DISTRIBUTED (E.G. GITHUB, TARBALL ON THE WEBSITE, ETC)?**

   No

4. ***WILL THE DATASET BE UPDATED (E.G., TO CORRECT LABELING ERRORS, ADD NEW INSTANCES, DELETE INSTANCES)? IF SO, PLEASE DESCRIBE HOW OFTEN, BY WHOM, AND HOW UPDATES WILL BE COMMUNICATED TO USERS (E.G., MAILING LIST, GITHUB)?***

   No

5. ***IS THERE A MECHANISM THROUGH WHICH AN INDIVIDUAL CAN REQUEST THEIR PERSONAL INFORMATION BE REMOVED?***

   N/A