Racial and Gender Disparities in Wages and Employment: Analyzing Trends Across Education Levels in USA (1979–2022)

Muhammad Muaviya Ijaz
Friedrich-Alexander-Universität Erlangen-Nürnberg

November 28, 2024

1 Introduction

This project investigates the disparities in wages and employment-to-population ratios across different education levels, with a focus on Black and White populations of men and women in the USA from 1979 to 2022. By using two primary datasets, this analysis seeks to explore how education affects employment opportunities and wage inequalities among men and women, with a particular emphasis on the effects of race and gender. To achieve these objectives, the project aim is to integrate these data sources by using an automated data pipeline to further extract and uncover key insights and trends from the transformed data.

2 Question

2.1 Primary Question

The main question that this project answers is: What is the impact of education level on wages and employment ratio across the black and white male and female population in the USA (1979–2022)?

2.2 Secondary Question

The secondary question that this project also supports is: How do wages and employment ratios correlate with each other for white and black men and women based on their education levels?

3 Data Sources

3.1 Wages by Education in the USA (1973-2022)

This dataset provides a comprehensive overview of average hourly wage statistics in the USA, segmented by education level, from 1973 to 2022. It covers different age brackets and races within the USA for each gender.

- **Data Source:** Kaggle
- **Data Source URL:** https://www.kaggle.com/datasets/asaniczka/wages-by-education-in-the-usa-1973-2022
- **Reason for Selection:** This dataset was chosen due to its extensive coverage of wage data over a significant time period. It includes wage information for both men and women across various education levels, making it suitable for analyzing wage disparities based on education and gender.

- **Structure and Quality of Data:** The dataset is organized in a CSV-based tabular format with over 60 columns detailing average hourly wages across various education levels for both men and women. The data is accurate, with no missing or duplicate values. Although the column names are generally clear, some require renaming for better clarity. The dataset maintains consistency with appropriate data types and values assigned to each column. However, for memory optimization, the year column needs type conversion, and certain columns and year rows (1973–1978) should be eliminated to meet the specific requirements of this analysis. The dataset spans from 1973 to 2022, providing valuable historical data.
- **License:** CC0: Public Domain, allowing free use, modification, and distribution and waives copyright. Attribution is encouraged but not required. Source: Economic Policy Institute's State of Working America Data Library, available at https://www.epi.org/data/

3.2 Employment-to-Population Ratio for USA (1979-2022)

This dataset offers detailed information on the employment-to-population ratio and the total population in the United States, covering the period from 1979 to 2022.

- **Data Source:** Kaggle
- **Data Source URL:** https://www.kaggle.com/datasets/asaniczka/employment-to-population-ratio-for-usa-1979-2023
- **Reason for Selection:** This dataset was chosen as it gives an extensive overview of the employment to population ratio for different age groups for both genders, covering the same time period as the first dataset. Also, the data is consistent and relevant, making it compatible with the first dataset.
- **Structure and Quality of Data:** The dataset is organized in a tabular CSV format and contains over 120 columns, detailing the employment-to-population ratio for various age groups and education levels. It is free of missing or duplicate values and maintains consistency across all columns. However, some column names require renaming, and certain integer columns need type conversions to optimize memory usage. Additionally, many columns need to be removed for our narrowed analysis. The dataset covers the period from 1979 to 2022, providing a comprehensive historical view of employment trends, though the most recent data may not fully reflect current employment dynamics.
- **License:** CC0: Public Domain, allowing free use, modification, and distribution, and waives copyright. Attribution is encouraged but not required. Source: Economic Policy Institute's State of Working America Data Library, available at https://www.epi.org/data/

4 Data Pipeline

The data pipeline for this project was implemented in **Python**, with **VSCode** as the IDE. The **Kaggle API** was used for data extraction, while **Pandas** facilitated data manipulation and transformation steps. Finally, **SQLite database** was utilized for storing the merged dataset.

1. **Extraction:** A Kaggle token was set up to use the Kaggle API. The datasets were then extracted with a retry mechanism to handle bad URLs and network errors. The CSV files were generated through an automated process that unzipped the dataset ZIP files.
2. **Data Cleaning and Transformation:** Using Pandas, unnecessary columns and rows were removed, and column names were standardized. New columns were created using existing data from the wages dataset. The datasets were then merged using a join method on the year column, with type conversion applied to this column. Although the data was complete and consistent, additional checks for missing and duplicate values were implemented.
    - **Wages By Education Dataset (1973-2022):**

- Row Year entries from 1973-1978 were removed for compatibility with the second dataset
- Kept only the year column and columns related to the wages of white and black men and women.
- Renamed column names for different race groups for clarity.
- Added a new column to describe the mean hourly wage for different education levels.
- Implemented generic checks for handling empty dataframes, missing values, and duplicate entries.

- **Employment-to-Population Ratio Dataset (1979-2022):**
  - Only kept the year and columns related to black and white men and women population for different age groups and education levels.
  - Columns name are standardized for better clarity
  - Performed standard checks to handle empty dataframes, missing values, and duplicate entries.

3. **Merging and Loading Datasets:** Merged the datasets on a common key attribute and loaded them into a database sink.
   - Merged both the datasets using **pd.merge()**, using the year column as the key.
   - Year column datatype changed to int32 for better memory usage
   - Load the merged dataset into a sqlite database sink in the **'data/'** directory of the project.
4. **Meta Quality Measures and Error Handling:** The automated pipeline includes error handling mechanisms for network calls and data validation checks to address issues arising from invalid inputs and incorrect arguments in various functions.
5. **Problems Encountered and Solutions:** Despite the data being largely clean with minimal issues during the cleaning and transformation processes—aside from removing some irrelevant columns—the **primary challenge** was selecting the relevant columns from both datasets. This was **resolved** by focusing on columns specific to Black and White populations and adding additional columns necessary for a narrowed analysis across these two racial groups.

5 Result and Limitations

The automated pipeline results in well-structured relational data stored as a SQlite database file.

5.1 Structure and Quality of Data:

The transformed and merged dataset is stored as an SQLite database file. It is free from missing or duplicate values and maintains consistent data types and values across all columns. The dataset has been narrowed down to approximately 90 columns, with well-formed column names that align with the specific requirements of the analysis.

5.2 SQLite as Output Data Format:

The output format chosen is SQLite database file. This choice was made as it is lightweight and very efficient in storing and querying structured data. The other reason for choosing it correlates with the size of the dataset. As the dataset size for our project is not huge, this format would be an ideal choice as it deals well with small to medium-sized datasets. Finally, it is serverless and easy to use across different systems and environments with various analytical tools.

5.3 Limitations:

The merged dataset may introduce bias in the analysis of male and female populations, as the employment-to-population ratio dataset contains percentages, whereas the wages dataset consists of numeric wage values. This discrepancy complicates combined analysis.