Labor Market Dynamics: Education and Unemployment Trends in U.S. Adults (2017–2022)

Maryam Homayoon - 23265350

Educational attainment and unemployment are key factors shaping the economic and social landscape of the United States. Higher education is often linked to greater employment stability, while events like the COVID-19 pandemic have exposed vulnerabilities in the labor market. This analysis examines the relationship between unemployment rates and educational attainment across age groups from 2017 to 2022. By spanning a time frame that captures both pre- and post-pandemic trends, this analysis provides a holistic view of how education influences employment stability over time and across life stages, aiming to uncover patterns that demonstrate the impact of educational attainment on economic resilience.

I. Question

How do unemployment rates correlate to educational attainment levels among adults aged 18 and older in the United States from 2017 to 2022?

Initially, the focus was on crime-related questions using FBI data but shifted to exploring the relationship between unemployment and educational attainment due to challenges in programmatically accessing and downloading the data.

II. Data Sources

Addressing this correlation requires two datasets. The first should provide unemployment rates segmented by age and educational degree for the selected years. The second should detail educational attainment levels by age group and year. Therefore the selected datasets are:

1. *Educational Attainment of the Population 18 Years and Over, by Age and Sex for All races[1]:* This dataset, provided by the U.S. Census Bureau, offers comprehensive annual statistics on educational attainment across various age and demographic groups in the United States from the Current Population Survey’s Annual Social and Economic supplement (ASEC).

2. *Unemployment Rates of Persons Aged 16 to 64 by Age Group and Highest Educational Attainment[2]:* Published by the National Center for Education Statistics (NCES), this dataset provides detailed unemployment statistics segmented by educational attainment levels and demographic groups for selected years, spanning 1975 to 2022. The data is sourced from U.S. Department of Commerce, Census Bureau, Current Population Survey (CPS).

A. Data Structure

1. *Educational Attainment of the Population 18 Years and Over, by Age and Sex for all races[1]:* This data is structured with categorical variables (e.g. educational levels such as "High school graduate" and "Bachelor's degree") and demographic categories (e.g. "Both sexes," "Age groups"). The rows and columns jointly define the structure, where rows represent specific characteristics and columns capture detailed metrics. While the data provides valuable quantitative insights, its structure is designed to be easily understood by humans, which makes automatic preprocessing challenging. Notably, the dataset contains no missing values, ensuring a high quality of data for analysis.

2. *Unemployment Rates of Persons Aged 16 to 64 by Age Group and Highest Educational Attainment[2]:* This dataset is structured with categorical variables (e.g., age groups such as "16 to 19 years old" and educational levels such as "Less than high school completion") and continuous variables representing unemployment rates across selected years (e.g., 1975, 1980, 2020). Rows detail educational categories, while

columns span years and related metrics. However, the dataset's irregular row formatting complicates direct analysis and preprocessing.

B. Data Quality

1. *The Educational Attainment* data[1] is considered Accurate as it is obtained from the official website of the United States government. In terms of Completeness, the data is annual which means that for the desired time span, these annual datasets must be combined. The data structure and data formats are Consistent through years. Timeliness is excellent since the data is separated by year. Relevance is not ideal due to the presence of metadata rows, which make the dataset less immediately usable without preprocessing.

2. *The Unemployment data*[2] has the same level of Accuracy as the other dataset. Completeness could be improved if the age ranges were more detailed, as the current ranges are quite broad. Consistency is excellent, with the official table updated each year by adding only one column for the most recent data. Timeliness is also very good, as the updated table is published annually. Relevance is not ideal due to the presence of numerous metadata rows, making the preprocessing harder.

C. Data Sources Licenses

1. Based on the census.gov policies on citation for *Educational attainment*[1], a simple citation is sufficient. As stated "Data users who create their own estimates using data from disseminated tables and other data should cite the Census Bureau as the source of the original data only. Conclusions drawn from any analysis of these data are the sole responsibility of the performing party." Therefore, their suggested citation format is used in this paper.

2. For *Unemployment dataset*[2], Permission to Replicate Information is granted, as stated: "all information on the U.S. Department of Education's NCES website at http://nces.ed.gov is in the public domain and may be reproduced, published, linked to, or otherwise used without NCES' permission. " Similarly, the citation rule is followed.

III. Data Pipeline

Due to the raw datasets structures including a lot of metadata such as table titles, footnotes, and empty rows for cleanness, preprocessing becomes challenging. To address this, the pipeline is coded with python to ease the process. Initially, I attempted to follow a traditional ETL pipeline architecture; however, I was unable to complete all the necessary transformations before loading the data. As a result, the current pipeline architecture is a mix of ETL and ELT.

- **Extraction:** Data is fetched from external URLs using the *requests* library. The fetched data is loaded into memory for initial processing using the *BytesIO* and *pandas.read_excel()* functions.
- **Transformation:** Each data undergoes extensive transformation and cleaning.
  - *Educational attainment data*[1] process includes: Extracting years from descriptive text to be added as a feature when combining all annual datasets. Age ranges are normalized by combining or excluding specific ranges to align with the more general age ranges in U*nemployment data*[2]. Educational levels are consolidated into broader categories to match those in the U*nemployment data*[2]. Column names are standardized, and unwanted rows are dropped to retain only data for both sexes.
  - The process for *Unemployment data*[2] highlights where the ETL architecture fails, requiring a mix of pre-load transformations and post-load transformations applied to the saved

SQLite table. The pre-load process includes: Removing unwanted rows such as foot notes, reformatting the data for each year into a single column, and cleaning row names. Post-load transformations involve: Appending the respective age range to the beginning of related educational level rows to handle formatting inconsistencies in the SQLite table; Additionally, the age ranges are adjusted by combining the '16 to 19' and '20 to 24' ranges into a new '18 to 24' age range.

- **Loading:** The final step involves saving the transformed and cleaned data into an SQLite database using the capabilities of sqlite3 library.

Challenges: During the development of the pipeline, one of the greatest challenges was the data structure which was easy to understand for human eyes but difficult to process programmatically. Another challenge was determining how to calculate and combine data for the age ranges of 16 to 19 and 20 to 24 to create the 18 to 24 age range.

Unfortunately, due to time constraints, no metadata has been implemented to handle errors or changes in the pipeline, leaving it unable to adapt to evolving data structures.

IV. Results ans Limitations

The output of the data pipeline is two separate SQLite databases with cleaned and transformed data from the datasets. The datasets are complete, with no missing values, a clear structure, and only the necessary data to provide insights into the question. The “educational_attainment” dataset organizes attainment data as numerical values for each category of education level, aggregated by age group and year; the “unemployed_data” stores unemployment data with rows representing different categories of age groups and educational attainment, while the year columns hold corresponding values. The output demonstrates high quality in terms of completeness, consistency, timeliness, and relevance.

The chosen final format is SQLite because of its robust data management, powerful querying capabilities, and scalability and flexibility.

V. Critical Reflection

While the pipeline effectively transforms and stores the data, several potential issues could impact the final analysis:

- Data Accuracy: the current datasets have not been checked for anomalies or duplicate values.
- Error Handling: the pipeline does not manage any errors or changes in the data and works exclusively with the current datasets.
- Limitations: The *unemployment dataset*[2], excludes individuals enrolled in school for the 16-19 and 20-24 age groups. This creates a mismatch with the *educational attainment data*[1], which includes the entire population. This limitation should be considered when interpreting the analysis.

References

[1] U.S. Census Bureau, “All Races,” Educational Attainment in the United States: 2022*,<https://www.census.gov/data/tables/2022/demo/educational-attainment/cps-detailed-tables.html>, accessed on November, 2024; *changing the year would result in the selected year, this analysis uses 2017-2022 data

[2] U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics