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

Maryam Homayoon

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 highlighted 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 or 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 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 or 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 understandable by humans which makes it hard to do automatice preprocessing. Notably, the dataset doesn't contain any 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 irregular row formatting, complicate direct analysis and preprocessing.

B. Data Quality

1. *The Educational Attainment* data[1] Accuracy is considered good since it is gained through official website of the United States government. In terms of Completeness, the data is annual which means that for the desired time span we should combine each annual dataset. Regarding Consistency, the structure and data formats is consistent through years. Timeliness is at the best since the data is separated by year and it also results in great Relevancy since only the needed time span is fetched to the pipeline.

2. *The Unemployment data*[2] Accuracy is the same as the other dataset. Completeness could be better if the age ranges were more separated since the current age ranges are more general. Consistency is so good that each year the official table is updated just by one column regarding the most recent year data. Timeliness is also very good and every year, the updated table is published. Relevancy as defined in the course would be poor since all the data from 1975 till now is presented but it is not a problem in this case since we will simply drop those columns.

C. Data Sources Licenses

1. Based on: census.gov policies on citation for *Educational attainment*[1], simple citation is enough; as quoted "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, you can see their citation in their suggested format in this paper.

2. For *Unemployment dataset*[2], we have: Permission to Replicate Information as quoted: "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. " and likewise the citation rule is followed.

III. Data Pipeline

Due to the raw datasets structures including a lot of meta data such as table titles, foot notes, and empty rows for cleanness, which makes the preprocessing hard, the pipeline is coded with python to ease things. At first I tried to follow ETL pipeline architecture but, I could not do all the necessary transformations before loading the data so, 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 using the *BytesIO* and *pandas.read_excel()* functions for initial processing.
- **Transformation:** Each data undergoes extensive transformation and cleaning.
  - *Educational attainment data*[1] process includes: Extracting years from descriptive text since we will need to add it later as a feature when we combine all annual datasets; Normalizing age group ranges by combining or excluding specific ranges since the age ranges in U*nemployment data*[2] is more general; Consolidating educational levels into broader categories to match the U*nemployment data*[2]; Standardize column names; drop the unwanted rows to only have the data for both sexes.
  - The process for *Unemployment data*[2] is where the ETL architecture fails so we have some pre-load transformation and some transformation functions on the SQLite table we saved. The pre-

load process includes: removing unwanted rows like foot notes; correcting the data format for each year into one column; Row name cleaning. The transformations we do after are: Append the respective age range to the beginning of related educational level rows to handle formatting inconsistencies in the SQLite table; Adjusts the age ranges in the *Unemployment dataset*[2] by combining '16 to 19' and '20 to 24' into a new '18 to 24' age range.

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

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 hard to process via machine codes. Another challenge was to find out how to calculate and combine the data for the age ranges of 16 to 19 and 20 to 24 to reach the age rang of 18 to 24.

Unfortunately, due to time limitations there are no meta-data implemented to handle errors or changes in this the pipeline and it can’t adapt to changing data.

IV. Results ans Limitations

The output of the data pipeline is two separate SQLite databases with cleaned and transformed data from the datasets. The dataset is complete with no significant missing values, clear structure, and only the data that are needed to give an insight to the question. The “educational_attainment” dataset organizes educational attainment data as numerical values for each category of education level, aggregated by age group and year; the “unemployed_data” store unemployment data with rows representing different categories of age groups and educational attainment, while the year columns hold corresponding values. The output quality has good Completeness, Consistency, Timeliness, and Relevancy since we have cleaned and transformed it to our desired data.

The chosen final format is SQLite due to more robust data management, querying capabilities, and scalability and flexibility.

V. Critical Reflection

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

- Data Accuracy: the current datasets have not been checked for anomalies or duplicate values.
- Error Handling: this pipeline does not manage any errors or changes in the data and only works with the current tested 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