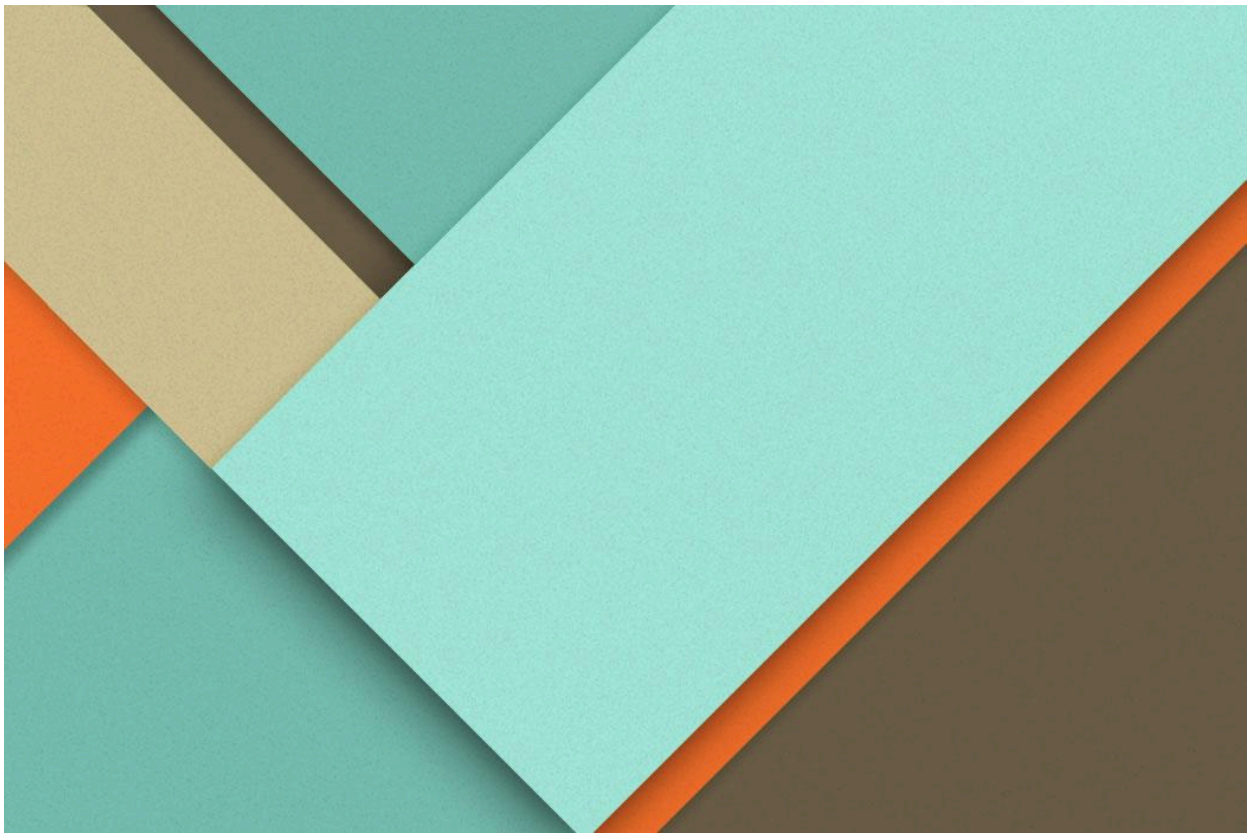

Analyzing U.S. Unemployment Through Education and Income Data

CS-GY 6513 (Big Data) - Fall 2025
NYU Tandon School of Engineering

—

Mariia Onokhina (mo2851) Yanka Sikder (ys4780) Yuqi Wang (yw4338)

Project Abstract

The relationship between education, income, and employment opportunities plays a major role in shaping the labor market in the United States. Unemployment rates and wage levels still vary widely across states and industries, while the national economy continues to evolve. This project aims to use large-scale data from the U.S. Census Bureau and the Bureau of Labor Statistics (BLS) to explore how education and income patterns relate to employment outcomes. We will analyze millions of records on unemployment rates, wage distributions, and educational attainment using Apache Spark, Hive, and Dask, to uncover trends that help explain differences in labor market stability across regions.

Problem Statement

The difference between unemployment and wages across the United States remains a challenge, even as the national economy grows and evolves. States with similar economic situations often experience different employment outcomes. In this project, we aim to show how the disparities between unemployment and wages are linked to socio-economic factors like education levels, income, and access to job opportunities.

Analyzing these factors with traditional data processing software is difficult because the data involved are massive, complex, and spread across multiple sources such as the U.S. Census Bureau and the Bureau of Labor Statistics. Therefore, we will use Big Data technologies such as Apache Spark, Hive, and Dask to integrate and analyze millions of socioeconomic records, aiming to identify meaningful patterns and correlations between education, income, and employment. The goal is to better understand how these factors shape labor market stability across states and industries.

Objectives

The primary goal of this project is to analyze and model the relationships between education, income, and employment outcomes across the United States using Big Data technologies. To achieve this, we will perform:

1. **Data Integration and Preprocessing:**

   Combine and clean large-scale datasets from the U.S. Census Bureau and the Bureau of Labor Statistics (BLS), including unemployment rates, wage distributions, and educational levels, to create a unified framework for analysis. The final integrated dataset will also contain and be structured at the county-by-year level using FIPS codes as primary geographic identifiers. Occupational and wage data

from BLS will be joined via SOC (Standard Occupational Classification) codes, while industry-level data will be linked using NAICS (North American Industry Classification System) codes. These methods ensure consistent merging of demographic, wage, and unemployment data.

2. **Correlation Analysis:**

   Identify relationships between educational quality, income inequality, and employment stability across different states and industries.

3. **Predictive Modeling:**

   Develop and fine-tune predictive models using PySpark MLlib and Dask ML to estimate unemployment trends, targeting a Mean Absolute Error (MAE) ≤ 0.5% and an $R^2 \geq 0.85$ on state-level unemployment predictions. Quantify correlations between education level, median wage, and unemployment using Pearson and Spearman correlation coefficients, with 95% confidence intervals (CI) reported for all metrics.

4. **Statistical Evaluation:**

   Evaluation of the impact of education and income distribution on unemployment and workforce participation using regression and statistical analysis, and models are expected to achieve a minimum absolute correlation coefficient $|r| \geq 0.30$ among core variables.

5. **Data Visualization:**

   Design interactive dashboards and plots using Apache Superset to communicate findings of any patterns on how education and income factors contribute to labor market disparities.

6. **Recommendations:**

   Provide data-driven recommendations to policymakers and researchers aimed at reducing unemployment inequality and improving workforce development strategies across the states.

Data Sources

I. 2015 - 2023 Labor force data by county, annual averages

Link to Dataset: Tables and Maps : U.S. Bureau of Labor Statistics

File format: Excel

Approximate File size: 233KB (each file)

Approximate Number of records: 3,223 (each file)

Purpose: to track local employment and unemployment trends.

II. 2015 - 2023 Occupational Employment and Wage Statistics (All Data):

Link to Dataset: Occupational Employment and Wage Statistics (OEWS) Tables

File format: Excel

Approximate File size: 7.7MB

Approximate Number of records: 414,438

Purpose: to analyze wages by occupation.

III. 2015- 2023 The Employment Cost Index (ECI):

Link to Dataset: Tables : U.S. Bureau of Labor Statistics

File format: Excel

Approximate File size: 2.4MB

Approximate Number of records: 35,622

Purpose: to measure the change in the hourly labor cost to employers over time.

IV. 2015 - 2023 Consumer Price Index:

Link to Dataset: Archived Consumer Price Index Supplemental Files : U.S. Bureau of Labor Statistics

File format: Excel

Approximate File size: 12.8KB

Approximate Number of records: 45

Purpose: to monitor inflation over time.

V. 2015 - 2023 American Community Survey Public Use Microdata Sample (ACS PUMS)

Link to Dataset: PUMS Data

File format: CSV

Approximate File size: 6-8GB

Approximate Number of records: 3-5 million records

Purpose: to analyze demographic, educational, and income characteristics of individuals and households across the U.S.

Total:

After integrating datasets from the Bureau of Labor Statistics (BLS) and the U.S. Census Bureau (ACS PUMS) covering the years 2015–2023, the combined dataset is expected to contain approximately 1 million records, totaling around 500–700 MB in Parquet format (compressed). Each record represents a unique state–year–education–income combination, incorporating key indicators such as unemployment rates, wage distributions, income levels, and educational attainment across all U.S. states.

Proposed Technologies & Programming Language

Data Storage: Hadoop Distributed File System (HDFS), Hive

We will clean and transform datasets that will be stored in HDFS as Parquet files with Snappy compression for efficient querying. The data lake will be partitioned by state and year, enabling parallel reads for analytical queries in Hive and Spark. ETL processes will include extraction via BLS APIs, transformation in Spark, and load into Hive tables for downstream ML tasks.

Machine Learning and Modeling: PySpark MLlib, Dask-ML

Data Visualization: Apache Superset, Matplotlib, Seaborn, Plotly

Data Analysis Libraries: Pandas, NumPy

Distributed Computing: Apache Spark, Dask

Programming Language: Python 3