# Data Report: Analysis of Educational Attainment and Income Levels Across States

## I. Question

How do educational attainment levels and income distributions correlate across U.S. states, particularly focusing on populations aged 18-24 years and household income levels?

This inquiry aims to understand patterns in education and income data across states and identify potential trends, which can provide insights for policymakers and stakeholders in education and economic development.

## II. Data Sources

1. Educational Attainment Data:

- Source: American Community Survey (ACS) 2023 1-Year Estimates

- Dataset Link: https://data.census.gov/table/ACSST1Y2023.S1501

- Description: Contains educational attainment statistics across states, focusing on different age groups (18-24 years, 25+ years) and degree levels.

2. Income Data:

- Source: American Community Survey (ACS) 2023 1-Year Estimates

- Dataset Link: https://data.census.gov/table/ACSST1Y2023.S1901

- Description: Provides household income distribution data across states, grouped by income ranges such as $50,000-$74,999.

Licenses: Both datasets are provided under the Census Bureau's Terms of Service (https://www.census.gov/data/developers/about/terms-of-service.html). Compliance is ensured by proper attribution, non-commercial use for educational purposes, and no substantial modifications to the datasets.

## III. Data Pipeline

A. Overview:

- The data pipeline was built using Python with Pandas for data manipulation and cleaning.
- Workflow includes:
  - Data Cleaning: Removal of symbols and formatting inconsistencies, preservation of percentages and dollar signs.
  - Data Transformation: Filtered and structured the educational dataset, grouped income data by household income ranges.

B. Tools and Technologies:

- Python (Pandas, NumPy, Matplotlib)
- CSV for data storage, Matplotlib for visualizations.

C. Challenges:

1. Formatting Issues: The datasets contained a variety of special characters ($, %, commas in numbers) and inconsistencies. We implemented preprocessing steps to preserve meaningful symbols while standardizing the data. For instance:

   - Retained percentages and dollar signs by careful symbol-specific operations.
   - Reformatted commas in large numbers (e.g., 2,051,545) to ensure numerical integrity.

2. Column Alignment: Column misalignment in the raw CSVs caused difficulties in parsing. We manually inspected the files to understand the structure, splitting data accurately by headers.

3. Large Number Handling: Numbers with embedded commas (e.g., 2,051,545) and percentages with decimal points (e.g., 6.4%) required specialized handling. We implemented distinct rules to differentiate between commas in numeric

values and delimiters between columns.

4. Preservation of Data Meaning: It was critical to retain the meaning of values (e.g., percentages, dollar signs) without loss of accuracy during transformations. For this, we:

- Ensured symbols such as $ and % remained intact during data cleaning.
- Addressed edge cases where formats were mixed (e.g., some values had symbols, while others did not).

5. Dealing with Missing Data: Both datasets contained columns with missing values. We implemented a strategy to handle missing values by retaining rows for analysis but noting the gaps in the data report for transparency.

6. Dataset Size and Complexity: With over 150+ columns in the education dataset and 200+ columns in the income dataset, readability was a concern. We reduced the datasets by focusing on relevant rows and columns, such as specific age groups and income ranges, while preserving critical insights.

## IV. Results and Limitations

A. Results:

1. The cleaned datasets provide structured views of:

- Educational attainment by age groups and degree levels across states.
- Household income distributions grouped by ranges.

2. Visualizations include:

- Educational Attainment Map: Percentage of populations aged 18-24 with some college education across states.
- Income Distribution Map: Household income levels within the $50,000-$74,999 range across states.

# Data Report: Analysis of Educational Attainment and Income Levels Across States

B. Limitations:

- Data Gaps: Missing values in some columns may affect state comparisons.

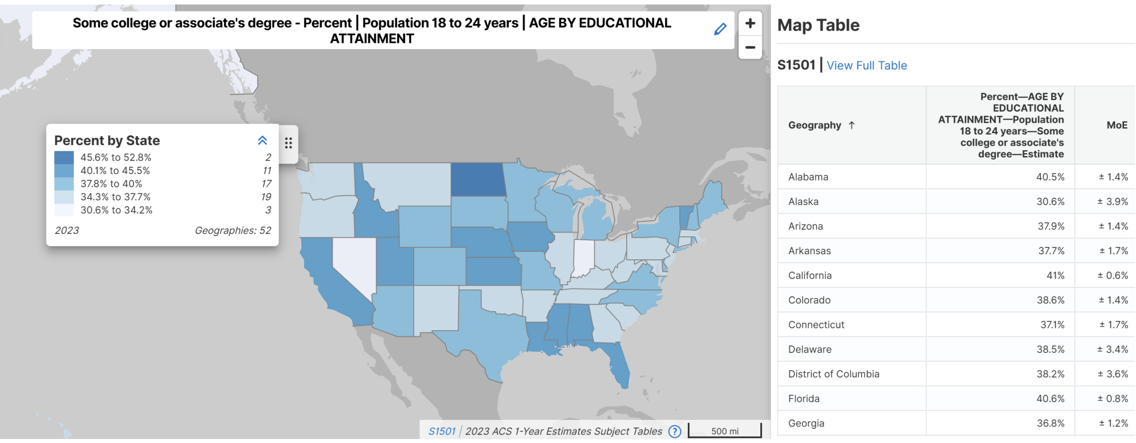
- Static Nature: The pipeline processes the 2023 ACS dataset; future datasets may require adjustments.

## V. Conclusion

The pipeline effectively cleans and processes educational and income data, providing valuable insights into state-level patterns. Future improvements could focus on enhancing automation and integrating external datasets for broader analysis.

## VI. Figures

Educational Attainment Data Sample:

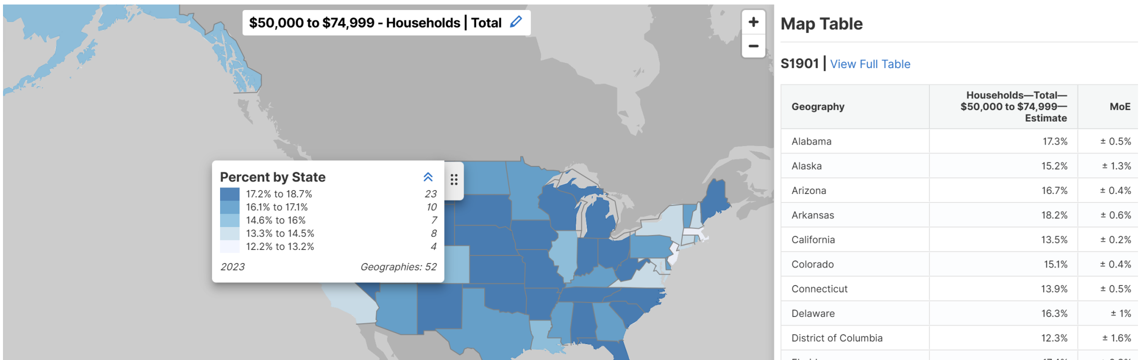

Some college or associate's degree - Percent | Population 18 to 24 years | AGE BY EDUCATIONAL ATTAINMENT
+
−
Percent by State
45.6% to 52.8% 2
40.1% to 45.5% 11
37.8% to 40% 17
34.3% to 37.7% 19
30.6% to 34.2% 3
2023
Geographies: 52
S1501 | 2023 ACS 1-Year Estimates Subject Tables
500 mi

Map Table

S1501 | View Full Table

| Geography ↑ | Percent—AGE BY EDUCATIONAL ATTAINMENT—Population 18 to 24 years—Some college or associate's degree—Estimate | MoE |
|---|---|---|
| Alabama | 40.5% | ± 1.4% |
| Alaska | 30.6% | ± 3.9% |
| Arizona | 37.9% | ± 1.4% |
| Arkansas | 37.7% | ± 1.7% |
| California | 41% | ± 0.6% |
| Colorado | 38.6% | ± 1.4% |
| Connecticut | 37.1% | ± 1.7% |
| Delaware | 38.5% | ± 3.4% |
| District of Columbia | 38.2% | ± 3.6% |
| Florida | 40.6% | ± 0.8% |
| Georgia | 36.8% | ± 1.2% |

Income Data Sample:



Map Table

S1901 | View Full Table

| Geography | Households—Total—$50,000 to $74,999—Estimate | MoE |
|---|---|---|
| Alabama | 17.3% | ± 0.5% |
| Alaska | 15.2% | ± 1.3% |
| Arizona | 16.7% | ± 0.4% |
| Arkansas | 18.2% | ± 0.6% |
| California | 13.5% | ± 0.2% |
| Colorado | 15.1% | ± 0.4% |
| Connecticut | 13.9% | ± 0.5% |
| Delaware | 16.3% | ± 1% |
| District of Columbia | 12.3% | ± 1.6% |
| [illegible] | [illegible] | [illegible] |