Project Work 4 - Data Report

Question

How can educational attainment and income distribution data be transformed and processed to derive insights into demographic trends and socio-economic factors?

Data Sources

The datasets used are:

1. Education dataset: [ACSST1Y2023.S1501](https://data.census.gov/table/ACSST1Y2023.S1501)

2. Income dataset: [ACSST1Y2023.S1901](https://data.census.gov/table/ACSST1Y2023.S1901)

Both datasets are provided under the Census Bureau's Terms of Service, which can be reviewed at: [https://www.census.gov/data/developers/about/terms-of-service.html].

Compliance Note:

"This product uses the Census Bureau Data API but is not endorsed or certified by the Census Bureau."

Data Pipeline

1. Data Loading:
   - Raw data was imported from CSV files.
2. Cleaning:
   - Symbols like %, $, and commas within numbers were preserved.
   - Misformatted or missing data were identified and handled.
3. Verification:
   - Data cleaning steps were iteratively refined.

4. Finalization:

   - Prepared cleaned datasets for analysis.

Challenges and Solutions

Challenges:

- Maintaining number formatting (e.g., 2,051,545).

- Preserving symbols like % and $ during processing.

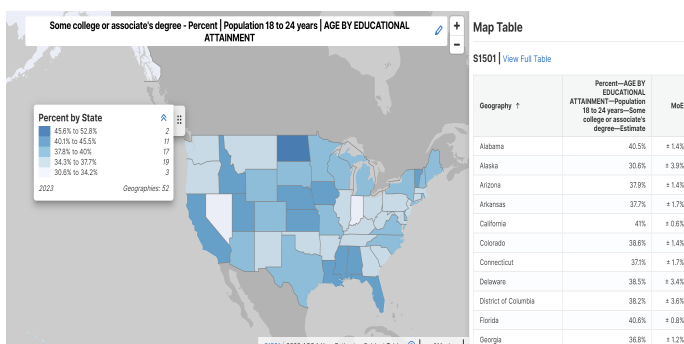- Managing inconsistencies in the source data.

Solutions:

- Applied regular expressions for targeted cleaning.

- Validated and previewed cleaned data iteratively.

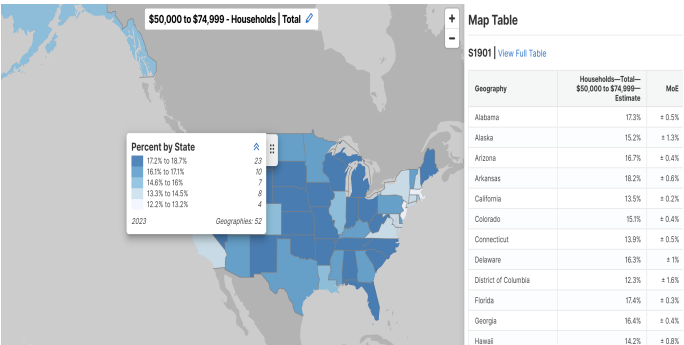- Built robust scripts to automate and streamline the pipeline.

Results and Limitations

Output Datasets:

1. Education Data: Cleaned and structured for demographic analysis.

2. Income Data: Ready for socio-economic insights.

Figures summarizing dataset structure:

Data Report

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%	± 0.8%
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%
Florida	17.4%	± 0.3%
Georgia	16.4%	± 0.4%
Hawaii	14.2%	± 0.8%

Critical Reflection

Strengths:

- High-quality cleaned datasets.

- Retention of essential symbols (% and $).

Limitations:

- Potential inaccuracies from missing data.

- Future data updates require pipeline adjustments.

Overall, the datasets are well-prepared for further analysis and integration with other studies.