Uncovering U.S. Obesity Trends: The Socioeconomic Impact of Income and Education Levels

Introduction

This project investigates the relationship between obesity rates and socioeconomic factors, such as income and education levels, across the United States. By analyzing data from two reliable sources, the Behavioral Risk Factor Surveillance System (BRFSS) and the Geographic Obesity Dataset, this study aims to understand regional disparities and their links to socioeconomic inequalities. The insights derived from this study can help guide public health interventions and policies to address obesity-related challenges effectively.

Question

Is there a correlation between obesity rates and socioeconomic factors, such as income or education levels?

Data sources

1. Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System

- Metadata URL: https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system/resource/0280bb9c-4de8-4b95-9642-93f727c4d305
- Data URL: https://data.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD
- Data Type: CSV
- Description: This data set provides state-specific data on various health metrics, including obesity rates, physical activity, and nutritional habits, with additional socioeconomic variables like income and education levels.

2. Obesity Rates and Geographic Information by State

- Metadata URL: https://data-lakecountyil.opendata.arcgis.com/datasets/lakecountyil::national-obesity-by-state/explore
- Data URL: https://services3.arcgis.com/HESxeTbDliKKvec2/arcgis/rest/services/LakeCounty_Health/FeatureServer/8/query?outFields=*&where=1%3D1&f=geojson
- Data Type: GeoJSON
- Description: This data set provides state-level obesity rates and other demographic information, allowing for spatial analysis. It includes variables that may enable a geographic analysis of obesity and its potential links with socioeconomic factors.

Reasons for Choosing These Data Sources

- **Relevance**: Focus on the U.S. and provide obesity-related metrics alongside socioeconomic indicators.
- **Coverage Period**: Recent data spanning multiple states, supporting robust correlation analysis.
- **Accessibility**: Publicly available datasets under open-data licenses, ensuring transparency.

Licenses and Permissions

Both datasets are publicly accessible under open-data licenses. Attribution is required, and modifications have been documented to ensure compliance with license terms. Links to the license information are provided in the metadata URLs above.

Data Pipeline

The data pipeline is implemented using Python, leveraging the Pandas and GeoPandas libraries for data manipulation and SQLite for data storage. The pipeline consists of the following steps:

Load Datasets

- **Socioeconomic Data**: Data was retrieved from the Behavioral Risk Factor Surveillance System (BRFSS) dataset, providing information on income, education, and other relevant variables.
- **Obesity Data**: Geographic obesity rates by state were extracted from a GeoJSON file, allowing spatial analysis and correlation with socioeconomic factors.

Preprocess & Cleaning the Data:

- **Socioeconomic Data**:
  - Selected relevant columns such as *YearStart*, *YearEnd*, *Location*, *Income*, and *Education*.
  - Standardized location names to lowercase and removed whitespace for consistency.
  - Removed rows with missing values to ensure data quality and accuracy.
- **Obesity Data:**
  - Retained essential columns, specifically *Location* and *Obesity*.
  - Standardized the Location column to match the format of the socioeconomic data for easier merging.
  - Removed rows with missing values to maintain data integrity.

Merge Datasets

The cleaned socioeconomic and obesity datasets were merged on the Location column to create a unified dataset. This merge enables the analysis of the relationship between socioeconomic factors and obesity rates across U.S. states.

Store Data

The merged dataset was stored in an SQLite database (*obesity_socioeconomic.db*) for efficient querying and exported as a CSV file (*merged_obesity_socioeconomic_data.csv*) for broader accessibility and integration.

Export Data

The processed dataset is available in two formats: SQLite for complex queries and analysis, and CSV for easy access and visualization in various tools.

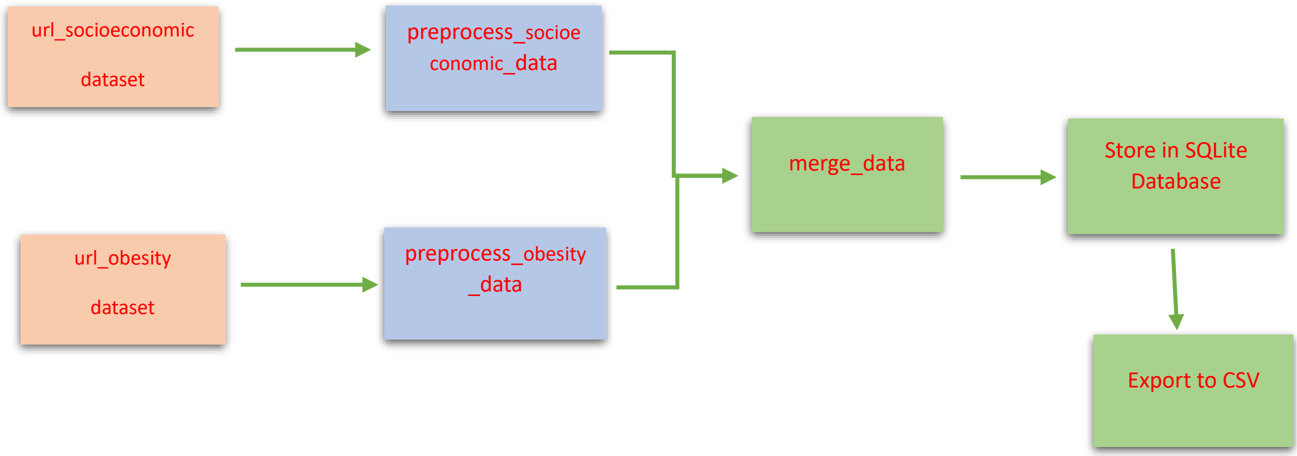


Fig 01: Automated Data Pipeline

Problems Encountered and Solved

Data Cleaning:

Handling datasets with varying structures and missing values posed a significant challenge. The socioeconomic dataset contained inconsistencies, such as missing values in key columns and extra spaces in column names. The obesity dataset had a similar issue with location mismatches and incomplete data. These issues were addressed through:

- Renaming and standardizing column names for consistency.
- Dropping rows with missing values in critical columns, such as "*Location*".
- Standardizing "*Location*" values to a lowercase, whitespace-trimmed format to ensure uniformity.

Data Merging:

Merging the two datasets required a common column, "*Location*", which had inconsistencies due to formatting differences between datasets. To overcome this:

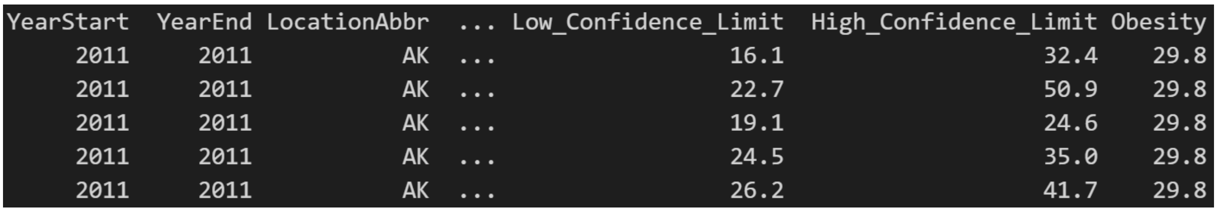
- Preprocessed both datasets to align their "*Location*" columns.
- Used an inner merge to retain only the relevant rows common across both datasets.
- Debugged mismatched "*Location*" entries by inspecting unique values and ensuring they were corrected.

Data Quality:

The processed datasets ensure accuracy with verified "*Location*" alignment, completeness by retaining only complete rows, and consistency through uniform formatting. The timely data spans recent years, providing relevant insights into obesity rates and socioeconomic factors.

Data Storage:

The final cleaned and merged dataset is stored in both SQLite and CSV formats for efficient querying and portability.



```
YearStart  YearEnd LocationAbbr  ... Low_Confidence_Limit  High_Confidence_Limit Obesity
     2011     2011           AK  ...                 16.1                   32.4    29.8
     2011     2011           AK  ...                 22.7                   50.9    29.8
     2011     2011           AK  ...                 19.1                   24.6    29.8
     2011     2011           AK  ...                 24.5                   35.0    29.8
     2011     2011           AK  ...                 26.2                   41.7    29.8
```

Fig 02: Merged Data Sample

Finally, the pipeline ensures high-quality, clean, and structured output datasets for further analysis. Challenges like missing values and inconsistent formatting were addressed using systematic preprocessing steps, enhancing data reliability and usability. Future work could focus on enhancing the granularity of the analysis by integrating additional data sources, refining the data pipeline for real-time processing, and applying advanced analytical techniques to uncover deeper patterns. Overall, this project highlights the potential of automated data pipelines to transform raw data into valuable insights, driving data-driven decision-making across domains.