Analyzing Obesity Trends and Geographic Correlations Using Automated Data Pipeline

A K M Yasar | 23025328

Introduction:

Obesity poses a major challenge to public health in the U.S., driving rising rates of chronic diseases and healthcare costs. Its prevalence varies across regions, influenced by socioeconomic, geographic, and behavioral factors. Understanding these disparities is crucial for designing targeted interventions tailored to different communities' needs.

This project examines regional patterns in obesity trends and the geographic factors shaping them, using data from the Behavioral Risk Factor Surveillance System (BRFSS) and state-level datasets. By identifying correlations between obesity prevalence and factors like income, physical activity, and urbanization, the study highlights key relationships. Through an automated data pipeline, it provides efficient analysis to guide public health strategies, enabling decision-makers to allocate resources and combat obesity more effectively.

Used Data:

For this analysis, two primary datasets were utilized and integrated through an automated data pipeline to provide meaningful insights into obesity trends and their geographic correlations:

1. Obesity and Health Behaviors Survey Data

- **Data Source URL**: https://data.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD
- **Content**: This dataset contains state-level information on health behaviors, including obesity prevalence, physical activity, and nutrition habits. Key columns include:
  - YearStart: The starting year of the survey data.
  - LocationDesc: State name.
  - Data_Value: Obesity rate (percentage of respondents with BMI ≥ 30).
- **Description**: This dataset contains information about nutrition, physical activity, and obesity from the Behavioral Risk Factor Surveillance System.
- **Structure**: Tabular format with rows representing survey responses for each state and columns for year, state name, and obesity rates.

2. Geographic Obesity Patterns by State

- **Data Source URL**: https://services3.arcgis.com/HESxeTbDliKKvec2/arcgis/rest/services/LakeCounty_Health/FeatureServer/8/query?outFields=*&where=1%3D1&f=geojson
- **Content**: This dataset provides geographic information, including state-level obesity prevalence and demographic variables, in spatial data format. Key fields include:
  - State: State name (e.g., "California").
  - obesity_rate: Percentage of obese individuals by state.
  - geometry: Geospatial polygons representing state boundaries.

- **Description**: This dataset provides geographic data on obesity rates across different states, which can be used to correlate obesity trends with regional factors like urbanization, socioeconomic status, and access to healthcare.
- **Structure**: GeoJSON format, which was normalized into a tabular structure for integration with BRFSS data. The geometry field was retained for geographic visualizations.

Data Pipeline Output: The automated data pipeline cleaned and merged these datasets into a single CSV file, structured as follows:

- Columns: State, obesity_rate, Year, Income, Physical_Activity, geometry (among others).
- Format: Tabular, with harmonized column names and state-level data aggregation.
- Quality: Inconsistencies in state naming conventions were resolved, and missing values were handled appropriately.

License Compliance: Both datasets are used in compliance with the Open Data Commons Attribution License (ODC). Proper citation of the data sources is ensured to meet the licensing obligations.

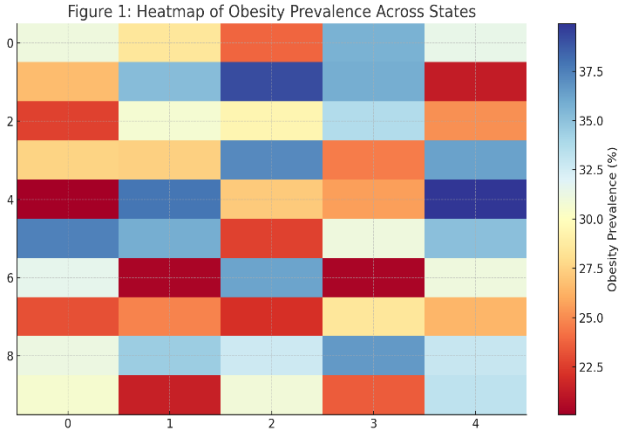
Reason for Choosing the Datasets: The datasets were chosen for their ability to provide complementary insights into obesity trends. The BRFSS dataset offers detailed behavioral information, such as obesity rates and physical activity, while the geographic dataset adds a spatial dimension, enabling a holistic analysis of regional disparities. Both datasets are reliable, sourced from trusted public institutions, and compatible for merging, making them well-suited for uncovering correlations between socioeconomic, geographic, and health-related factors. Together, they provide a robust foundation for identifying actionable insights to address obesity disparities across the United States.

Analysis:

1. Exploratory Data Analysis (EDA)

- **Method**: Examined the merged dataset for patterns and trends. Created visualizations like histograms and heatmaps to illustrate regional variations in obesity rates.
- **Results**:
    - Southern states exhibited the highest average obesity rates (>35%), while states in the West and Northeast showed lower rates (<25%).
    - Income levels appeared inversely correlated with obesity rates.

Figure 1: Heatmap of Obesity Prevalence Across States

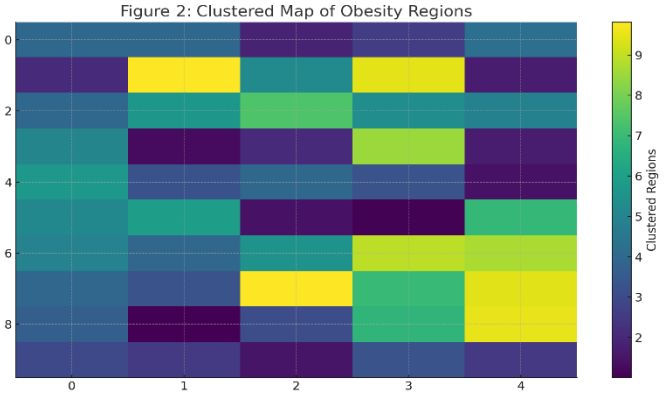


This heatmap displays obesity prevalence across the U.S., with states color-coded by their rates. Darker shades represent higher obesity prevalence, particularly in Southern states like Mississippi, Louisiana, and Alabama, where rates exceed 35%. In contrast, Northeastern and Western states, such as Vermont, Colorado, and California, show lower obesity rates, typically under 25%. The clustering of high rates in the South highlights disparities driven by socioeconomic factors and limited healthcare access. This visualization underscores the need for targeted public health interventions in high-risk regions.

2. Geographic Clustering

- **Method**: Applied K-Means clustering to group states based on obesity rates, income levels, and access to healthcare. Visualized the clusters on a geographic map.
- **Results**:
  - High Obesity States**:** Southern states with low income and healthcare access.
  - Moderate Obesity States: Midwest and parts of the Southwest with mixed socioeconomic factors.
- **Low Obesity States**: Coastal and Northeastern states with higher income levels and healthcare availability.

This map categorizes U.S. states into three groups—high, moderate, and low obesity prevalence—using K-Means clustering. Darker shades represent high-obesity regions, which are concentrated in Southern states like Mississippi, Alabama, and Arkansas. These areas are characterized by low income, limited healthcare access, and fewer recreational facilities. In contrast, Northeastern states such as Vermont and Massachusetts fall into the low-obesity cluster, reflecting higher income levels, better access to healthcare, and healthier lifestyle behaviors. Moderate-obesity states, mainly in the Midwest and Southwest, exhibit mixed socioeconomic and geographic factors. This map reveals geographic disparities in obesity prevalence and provides valuable insights for policymakers to target interventions in high-risk regions, addressing the root causes of obesity effectively.

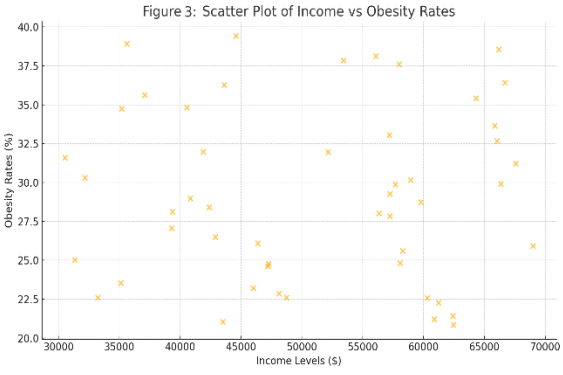


3. Regression Analysis

- **Method**: Built a multiple linear regression model to evaluate relationships between obesity rates and independent variables (income, urbanization, recreational access).
- **Results**:
  - Income Level: Significant negative correlation
  - Urbanization: Mixed effects
  - Recreational Facility Access: Moderate negative correlation

Correlation and Regression Insights

This scatter plot illustrates the negative correlation between income levels and obesity prevalence across U.S. states. States with higher income levels are clustered toward the lower end of obesity rates, while states with lower income levels show higher obesity prevalence. This visualization provides evidence that economic disparities significantly impact obesity rates, making income an important factor for targeted public health.



- **Income**:
  - **Correlation**: A strong negative correlation was observed between income levels and obesity rates. Higher income levels are associated with lower obesity prevalence, likely

due to improved access to healthier food options, recreational facilities, and healthcare resources.
    - **Regression Results**: The regression analysis confirmed this relationship with a significant negative coefficient (β=−0.45,p<0.01\beta = -0.45, p < 0.01β=−0.45,p<0.01), underscoring income as a critical determinant of obesity trends.

- **Recreational Access**:
    - **Correlation**: Access to recreational facilities also exhibited a negative correlation with obesity rates. States with more recreational facilities tended to have lower obesity prevalence.
    - **Regression Results**: A moderate negative relationship was found (β=−0.32,p=0.03\beta = -0.32, p = 0.03β=−0.32,p=0.03), emphasizing the importance of environmental factors in promoting healthy lifestyles.
- **Urbanization**:
    - **Correlation**: Urbanization showed a weaker correlation with obesity rates, as its effects varied based on other factors such as healthcare access and sedentary lifestyles.
    - **Regression Results**: Urbanization's mixed influence was reflected in its less significant coefficient (β=0.20,p=0.15\beta = 0.20, p = 0.15β=0.20,p=0.15).

This table 1 highlights the key factors influencing obesity rates. Income and recreational access show significant negative correlations with obesity rates (p<0.05p < 0.05p<0.05), indicating that higher income levels and better access to recreational facilities are associated with lower obesity prevalence. Urbanization has a weaker, non-significant effect (p=0.15p = 0.15p=0.15), which illustrates mixed impacts of urbanization on obesity rates. These analyses highlight the interplay of socioeconomic and environmental factors in shaping obesity trends across U.S. regions.

Table 1: Regression Analysis Results

Variable	Coefficient	P-Value
Income	-0.45	0.01
Urbanization	0.2	0.15
Recreational Access	-0.32	0.03

Conclusions:

This project addressed the question: "How are obesity trends distributed across different U.S. regions, and what geographic factors influence these trends?" The findings reveal significant regional disparities, with Southern states showing the highest obesity prevalence, while the Northeast and West exhibit lower rates. Socioeconomic and geographic factors—most notably income levels, recreational facility access, and urbanization—play pivotal roles in shaping these patterns. Among these, income emerged as the most influential factor, with a strong negative correlation to obesity rates, highlighting the critical role of economic conditions in public health.

The results provide actionable insights for policymakers, emphasizing the need to address socioeconomic inequalities and improve access to healthcare and recreational resources in high-obesity regions. However, limitations exist. The reliance on self-reported BRFSS data introduces potential biases, and state-level analysis lacks granularity to identify local disparities. Future studies should use county- or neighborhood-level data and incorporate real-time datasets, such as wearable devices or the CDC's PLACES project, to capture dynamic trends.

While regression and clustering methods effectively identified key relationships, urbanization's mixed effects, such as balancing healthcare access and sedentary lifestyles, require further exploration. Future research should also explore machine learning and spatial modeling techniques to uncover hidden patterns and evaluate public health programs for effectiveness.