# **Project Report: Exploratory Data Analysis of BMI and Weight-Related Health Trends in the U.S. Population**

## **Introduction**

The prevalence of obesity and related health conditions has become a significant public health challenge in the United States, with major implications for healthcare costs, life expectancy, and overall population health. This project explores trends in body mass index (BMI) categories—specifically, normal weight, overweight, and obesity—among adults in the United States. Using data from the National Health and Nutrition Examination Survey (NHANES), the analysis aims to uncover patterns in weight status across different demographic segments and over time.

By understanding how obesity and overweight prevalence vary by age, sex, race, and survey period, this study highlights key areas for public health intervention and provides insights to support policy decisions. The findings contribute to a data-driven approach to mitigate obesity-related risks and improve population health outcomes.

### **Data Source**

The data utilized in this project is sourced from NHANES and accessed through the U.S. government’s Data Catalog. The dataset focuses on adults aged 20 and over and includes multiple demographic and temporal variables that enable comprehensive analysis. The dataset can be accessed on https://catalog.data.gov/dataset/normal-weight-overweight-and-obesity-among-adults-aged-20-and-over-by-selected-characteris-8e2b1.

# Data Description

The dataset includes a range of variables designed to support an in-depth examination of BMI categories within the U.S. adult population. Key variables are outlined in the data dictionary below:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **Health\_Status** | General weight status category (e.g., normal weight, overweight, obesity). |
| **BMI\_Category** | Specific BMI range classification (e.g., Grade 1 Obesity). |
| **Population\_Percentage\_Type** | Type of percentage metric (age-adjusted or crude percentage). |
| **Demographic\_Category** | Data segmentation category, such as sex, race, or age. |
| **Age\_Range** | Specific age group range (e.g., "20 years and over", "75 years and over"). |
| **Survey\_Period** | Survey years (e.g., 2005-2008), indicating the data collection period. |
| **ESTIMATE** | Estimated percentage of the population in a BMI category. |
| **SE** | Standard error of the estimate, indicating precision. |
| **FLAG** | Data quality flag for reliability considerations. |
| **Age\_Group** | Aggregated age group for broader analysis (e.g., "Young Adults"). |
| **Survey\_Midpoint** | Calculated midpoint of each survey period to facilitate time-series analysis. |
| **BMI\_Category\_Code** | Encoded BMI values for ordered visualizations. |

# Methodology

### **Step 1: Environment Setup**

The project was developed using Python, employing libraries like pandas for data manipulation, matplotlib and seaborn for visualization, and setting display options for better data readability. This initial setup ensures efficient data handling and visually clear outputs.

### **Step 2: Data Loading**

The dataset was loaded from Google Drive for analysis in Google Colab. After mounting Google Drive, the CSV file containing BMI-related data was read into a panda DataFrame.

### **Step 3: Initial Exploration**

To gain an understanding of the dataset, the initial exploration included displaying the first few rows, examining data types, checking for missing values, and reviewing descriptive statistics. This step revealed essential details on variable formats, missing data, and column distributions, setting the foundation for data cleaning and preparation.

### **Step 4: Data Cleaning**

Data cleaning involved several key steps:

1. **Handling Missing Values**: The SE (standard error) column had missing values, which were filled with the median value to ensure continuity without skewing the data. Missing values in the FLAG column were replaced with "No Flag".
2. **Data Type Optimization**: Categorical variables were converted to the category type to optimize memory usage and improve processing speed.
3. **Dropping Redundant Columns**: Unnecessary columns, such as PANEL\_NUM and AGE\_NUM, were dropped as they were not essential for analysis.
4. **Renaming Columns**: Columns were renamed for consistency with the data dictionary, enhancing readability and interpretation.

### **Step 5: Data Transformation and Feature Engineering**

Several new features were created to facilitate analysis:

* **Age Grouping**: Broader age categories were generated to aggregate data into “Young Adults,” “Middle-aged Adults,” “Older Adults,” and “Senior Adults.”
* **Survey Midpoint Calculation**: For easier time-series analysis, a midpoint value was calculated for each survey period.
* **BMI Category Encoding**: BMI categories were assigned numerical codes to support sorted visualizations.

# Exploratory Data Analysis (EDA)

The exploratory analysis included several visualizations to understand trends in BMI and health status across the population.

### 1. **BMI Distribution Across Age Groups**

A box plot was created to display the distribution of BMI categories (normal weight, overweight, obesity) across age groups. The findings revealed that BMI categories tend to shift upwards with age, with a higher prevalence of obesity among middle-aged and older adults. This trend suggests that age-specific interventions may be needed to manage weight more effectively across different life stages.

### 2. **Trends in BMI Categories Over Time**

Using a line plot, the analysis tracked the trend of BMI categories over survey years, revealing an increasing prevalence of obesity. While the percentage of individuals with normal weight has declined over the years, the overweight and obesity categories have shown a steady increase. This temporal trend highlights the growing public health challenge posed by obesity, calling for targeted actions to reverse this trend.

### 3. **BMI Comparison by Demographic Segments (Sex, Race)**

To understand BMI variations across demographic factors, a bar plot was generated focusing on sex and race. The results indicate disparities in BMI prevalence, with certain racial and ethnic groups exhibiting higher obesity rates. This demographic insight underscores the need to consider cultural, socioeconomic, and environmental factors when developing public health policies related to weight management.

### 4. **Standard Error Distribution by BMI Category**

A box plot of standard error by BMI category was created to assess data reliability. Standard errors were relatively consistent across categories, with slightly higher variability observed in extreme BMI ranges, such as severe obesity. This consistency suggests that BMI estimates are generally reliable for most categories, although caution may be warranted with estimates for higher BMI groups.

### 5. **Heatmap of BMI Estimates by Survey Period and Age Group**

A heatmap was created to visualize BMI category estimates by survey period and age group. The heatmap highlighted an upward trend in obesity estimates across all age groups over time, particularly among middle-aged and older adults. This visualization underscores the increasing obesity rates among U.S. adults, affirming the urgency of addressing this issue across different age demographics.

# Insights and Recommendations

The analysis provides several valuable insights into BMI and weight trends among U.S. adults:

1. **Age-Related Obesity Trends**: The findings confirm that older adults tend to exhibit higher obesity rates. Targeted public health interventions should consider age-related factors and promote weight management programs tailored to middle-aged and senior adults.
2. **Rising Obesity Over Time**: There is a clear upward trend in obesity rates across survey periods, emphasizing a worsening obesity crisis. Policies focused on dietary education, physical activity promotion, and access to healthy food options could help mitigate this trend.
3. **Demographic Disparities**: Differences in obesity prevalence across demographic segments, particularly among racial and ethnic groups, point to a need for culturally specific health interventions and community outreach efforts.
4. **Reliability of Estimates**: The analysis of standard error indicates that the data is generally reliable across BMI categories, though estimates for extreme obesity categories may require further validation.

# Conclusion

This exploratory data analysis sheds light on critical trends in BMI and weight status among U.S. adults, revealing a complex picture influenced by age, time, and demographic factors. As obesity rates continue to rise, these findings support a multi-faceted approach to public health interventions, including age-specific programs, culturally tailored outreach, and policies promoting healthy lifestyles. Addressing the obesity epidemic requires collaboration across sectors to ensure a healthier future for all demographic groups in the United States.