Obesity Prediction Data Analysis

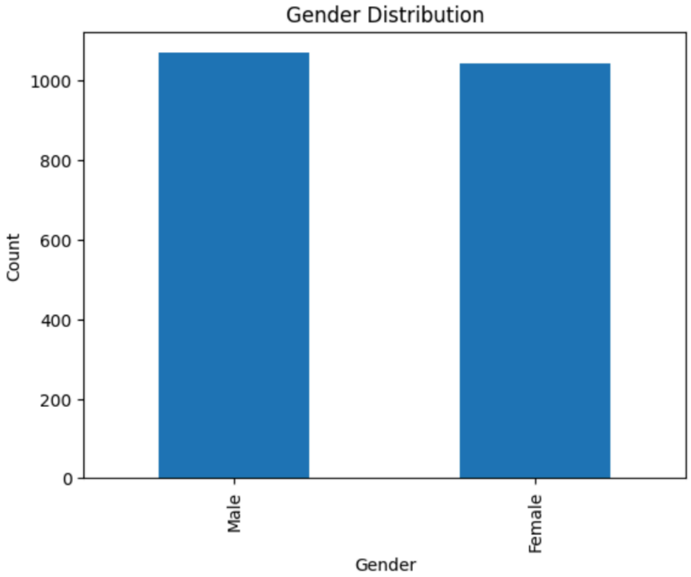
Nguyen Quoc Anh Cuong - 220214
Programming for Data Science and Visualization course

Story Telling

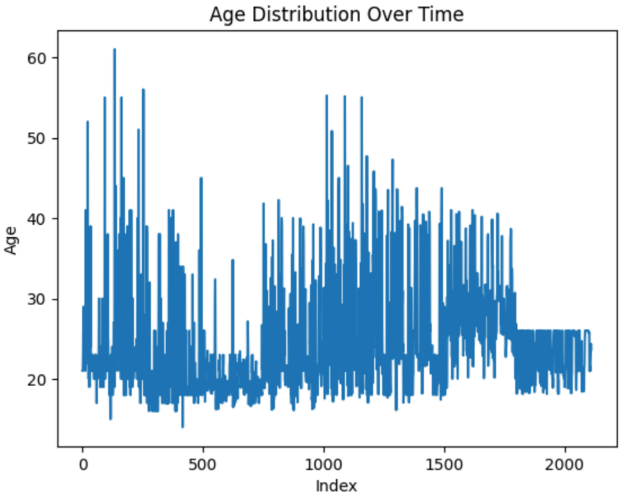
Obesity is a complex disease influenced by various factors, including an imbalance between nutrient intake and expenditure, a sedentary lifestyle, and environmental, genetic, and epigenetic elements. The presence of an obesogenic environment, characterized by easy access to food, urbanization, and the widespread consumption of processed foods, significantly contributes to obesity rates. This condition is also associated with several comorbidities, such as type 2 diabetes, cardiovascular disease, and certain cancers.

To understand the factors contributing to obesity, we conducted a comprehensive analysis using a dataset from individuals in Mexico, Peru, and Colombia. We began by importing necessary libraries like Pandas, NumPy, Matplotlib, and Seaborn, and loaded the dataset from a CSV file. We verified the data structure and summarized it by counting unique values for categorical variables, determining the shape of the data frame, and calculating descriptive statistics. Handling missing data was crucial, so we dropped rows with any null values and replaced null values in numerical columns with the mean. We then extracted rows meeting specific criteria, removed duplicates, and randomly selected a fraction of rows for analysis. Various visualizations, including bar charts, line graphs, histograms, scatter plots, box plots, and heatmaps, were created to explore data relationships.

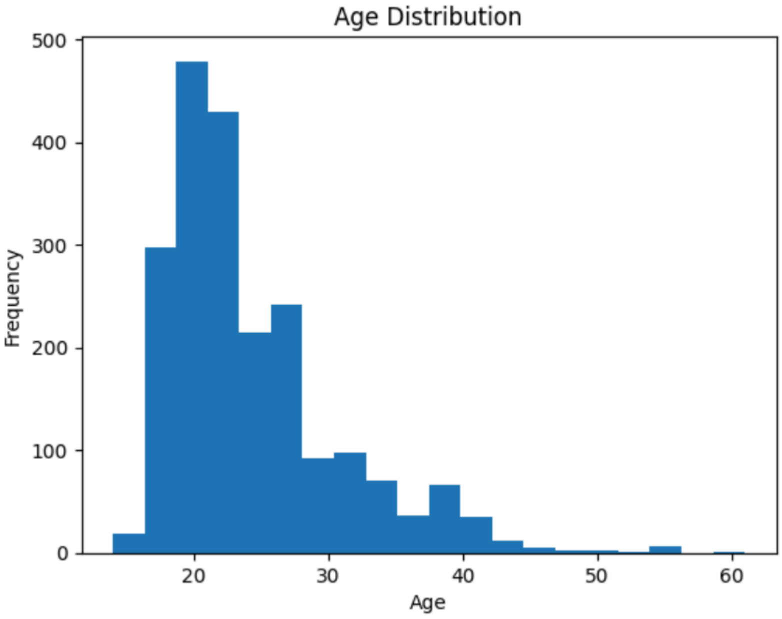
During data cleaning, we handled missing values using forward fill, detected and addressed outliers using z-score, and corrected inconsistencies by removing duplicate records. Our exploratory data analysis involved generating a correlation matrix and pairplot to identify key patterns and correlations. The results revealed several insights. A bar chart showed a balanced gender distribution:



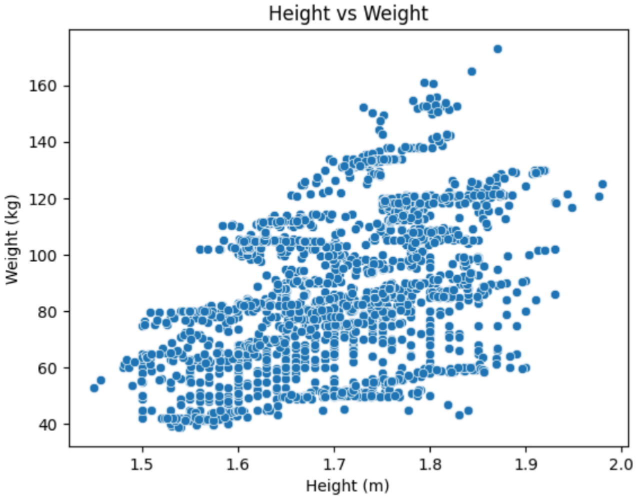
A line graph illustrated age distribution over time, with higher variability in earlier indices:



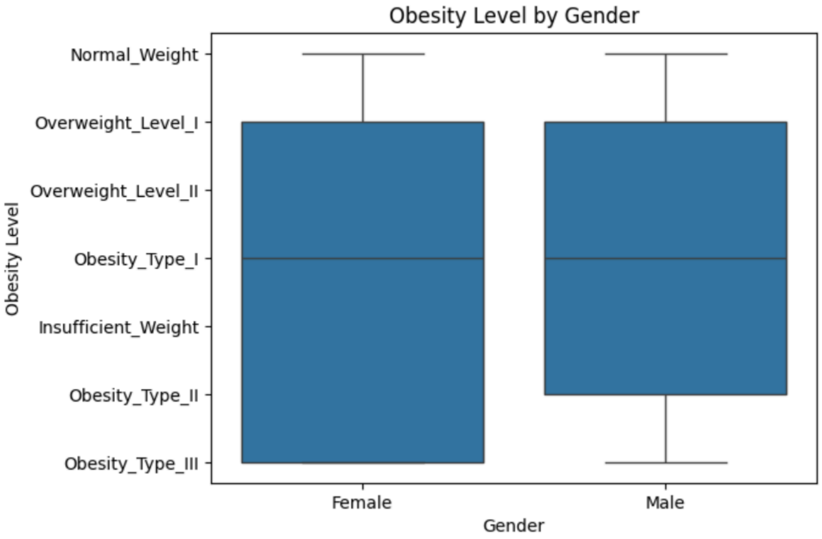
A histogram indicated that most participants were young adults, with fewer older individuals:



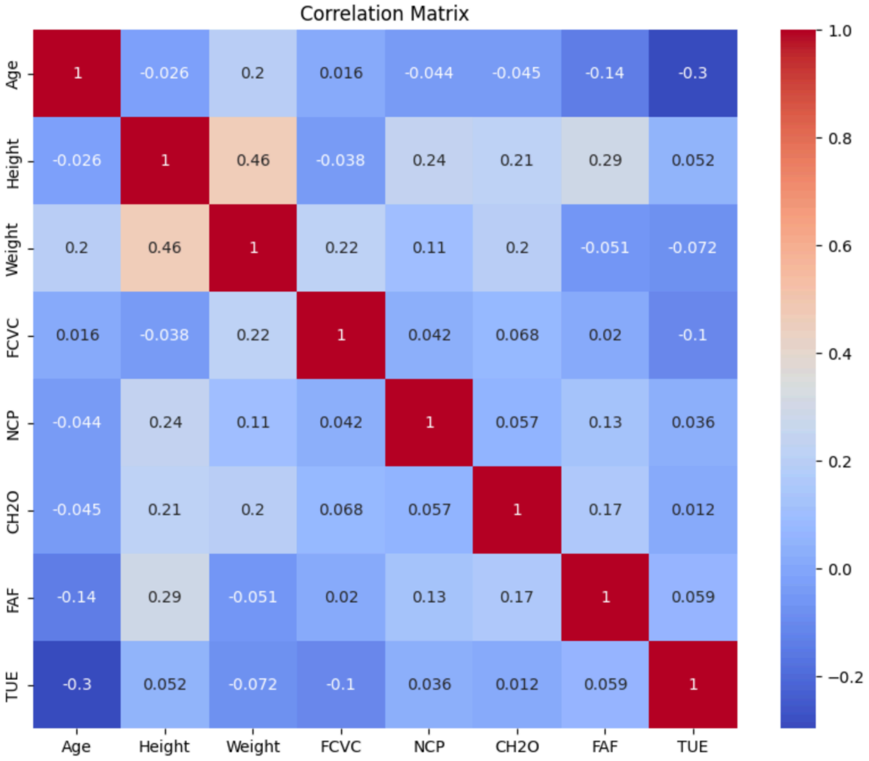
A scatter plot showed a positive correlation between height and weight:



A box plot compared obesity levels by gender, revealing similar median obesity levels for both males and females:



The heatmap highlighted significant correlations, such as the strong positive correlation between height and weight:



This analysis provided key insights into obesity levels among individuals from Mexico, Peru, and Colombia. The balanced gender distribution ensured unbiased comparisons, and the age distribution highlighted the critical age range for lifestyle changes. The scatter plot validated the dataset's accuracy, and the box plot suggested that factors other than gender, such as lifestyle and dietary habits, play a significant role in obesity levels. The heatmap emphasized the importance of dietary habits and physical activity in determining obesity levels.

My findings align with existing research that highlights the impact of dietary habits and physical activity on obesity. For instance, the study "Obesity: causes, consequences, treatments, and challenges" published in the Journal of Molecular Cell Biology discusses how imbalanced energy intake and expenditure, coupled with a sedentary lifestyle, contribute to obesity. The study also emphasizes the role of environmental, genetic, and epigenetic factors in obesity, which is consistent with our observations. Another study published in BMJ Public Health explores the public health implications of obesity and supports the need for comprehensive interventions to address this growing issue.

In conclusion, this analysis of the Obesity Prediction dataset from Mexico, Peru, and Colombia reveals significant insights into the factors contributing to obesity levels. The study highlights the critical role of dietary habits and physical activity in determining obesity levels, with frequent consumption of high-caloric food and low vegetable intake being strongly associated with higher obesity levels. Additionally, lower frequency of physical activity and higher daily time spent using technological devices are linked to increased obesity levels. These findings underscore the importance of promoting healthy eating and regular physical activity as part of public health interventions to combat obesity. However, the study's limitations, including its geographical focus and reliance on self-reported data, suggest the need for further research. Expanding the dataset to include participants from other regions and investigating additional lifestyle factors such as sleep patterns and stress levels could provide a more comprehensive understanding of obesity. Overall, this analysis provides valuable insights that can inform public health strategies aimed at reducing obesity rates and improving overall health outcomes.

Report

Executive Summary

Obesity is a significant public health issue globally, with the World Health Organization (WHO) estimating that as of 2022, 59% of adults are living with overweight or obesity. This report analyzes the Obesity Prediction dataset, which includes data from Mexico, Peru, and Colombia. Through data cleaning, exploratory data analysis (EDA), and visualization, we identified key patterns and correlations. The analysis revealed significant relationships between dietary habits, physical activity, and obesity levels. These findings can inform public health strategies and interventions to combat obesity.

Introduction

Obesity is a multifactorial disease primarily caused by an imbalance between nutrient intake and expenditure, along with a sedentary lifestyle. Environmental, genetic, and epigenetic factors also play crucial roles. The presence of an obesogenic environment, including aspects such as food availability, urbanization, and the widespread consumption of processed foods, significantly influences obesity rates. Additionally, obesity is associated with various comorbidities, including type 2 diabetes, cardiovascular disease, and certain cancers.

Methodology

1. Data Loading:

- Imported necessary libraries (Pandas, NumPy, Matplotlib, Seaborn).
- Loaded the dataset from a CSV file using Pandas.
- Verified the data structure using `head()`, `info()`, and `describe()` methods.

2. Data Summarization:

- Counted unique values for categorical variables.
- Determined the shape of the data frame.
- Calculated descriptive statistics (min, max, mean, standard deviation).

3. Handling Missing Data:

- Dropped rows with any null values.
- Replaced null values in numerical columns with the mean.

4. Subset Observation:

- Extracted rows meeting specific criteria (e.g., age > 30).
- Removed duplicate rows.
- Randomly selected a fraction of rows for analysis.

5. Data Visualization:

- Created various visualizations (bar charts, line graphs, histograms, scatter plots, box plots, heatmaps) to explore data relationships.

6. Data Cleaning:

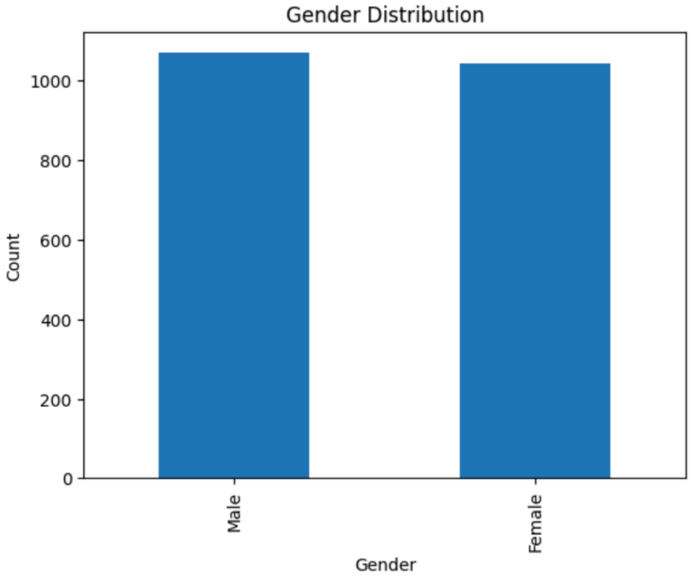
- Handled missing values using forward fill.
- Detected and addressed outliers using z-score.
- Corrected inconsistencies by removing duplicate records.

7. Exploratory Data Analysis (EDA):

- Conducted initial exploration and descriptive statistics.
- Generated correlation matrix and pairplot to identify key patterns and correlations.

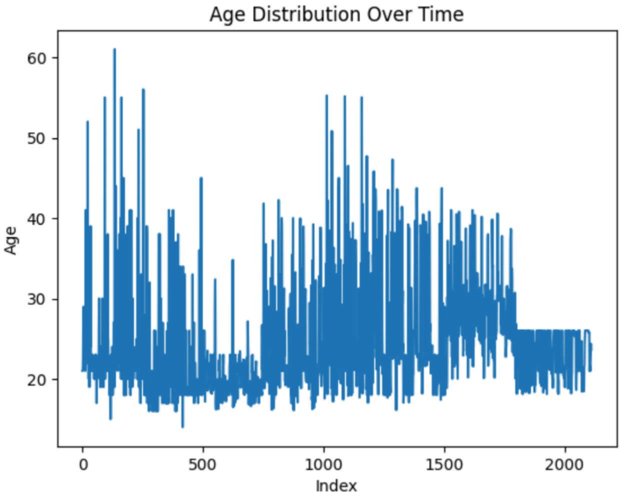
Results

1. Bar Chart for Gender Distribution



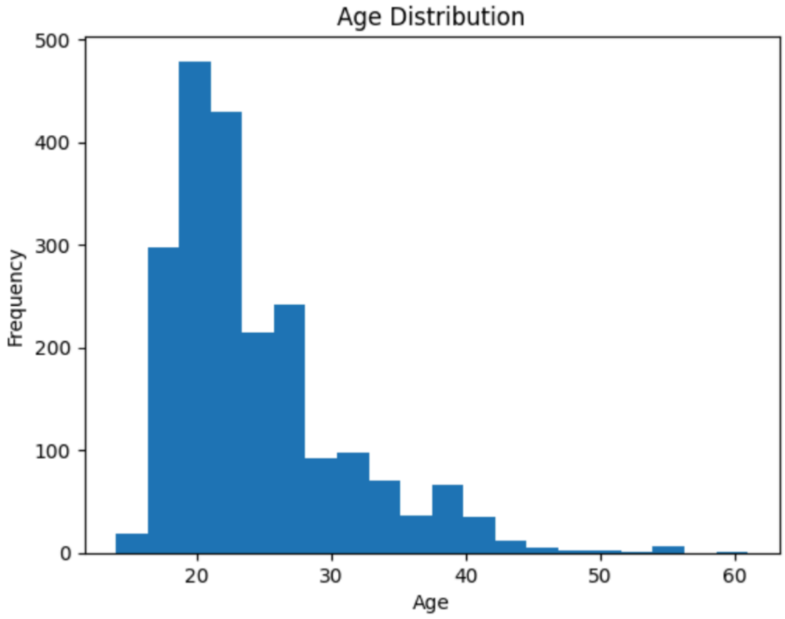
The bar chart shows the distribution of gender in the dataset. It indicates a balanced distribution of male and female participants, with both genders having an equal count of approximately 1000.

2. Line Graph for Age Distribution Over Time



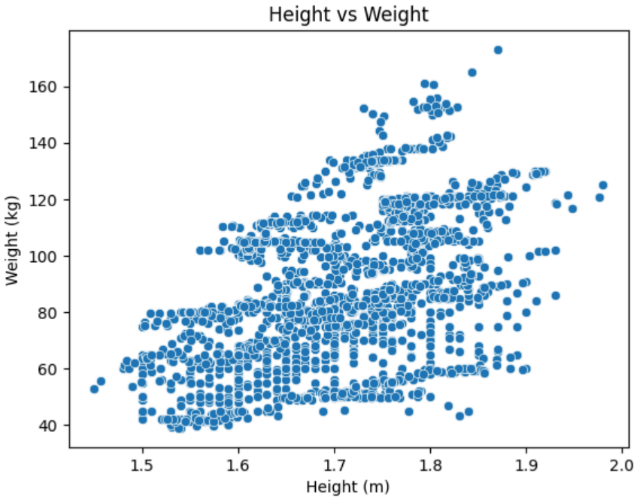
The line graph illustrates the age distribution of participants over time. The x-axis represents the index, ranging from 0 to 2000, and the y-axis represents age, ranging from 0 to 60. The graph shows fluctuations in age values, with higher variability in the earlier indices and a noticeable decrease in variability after approximately index 1500.

3. Histogram for Age Distribution



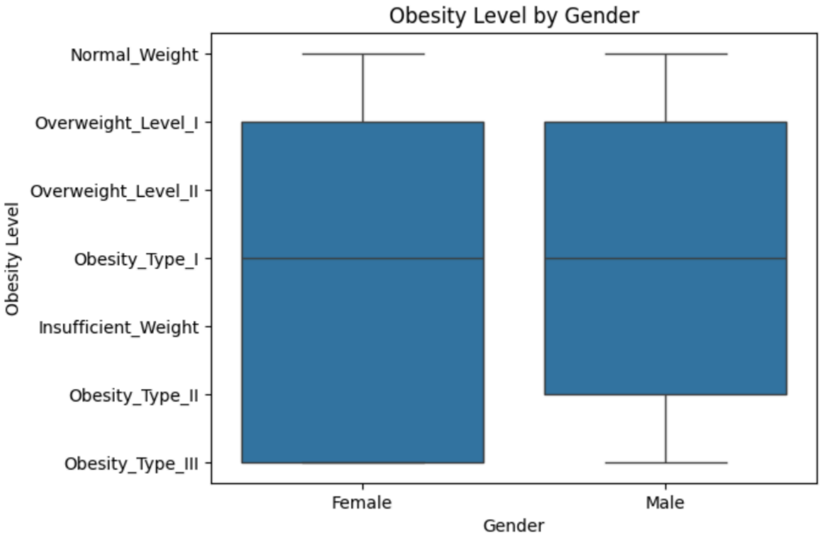
The histogram displays the frequency distribution of participants' ages. The x-axis represents age, ranging from 0 to 60 years, and the y-axis represents frequency, ranging from 0 to 500. The highest frequency of individuals falls within the age range of approximately 15 to 25 years, with the peak frequency just under 500. As age increases beyond this range, the frequency decreases significantly, with very few individuals above the age of 50.

4. Scatter Plot for Height vs Weight



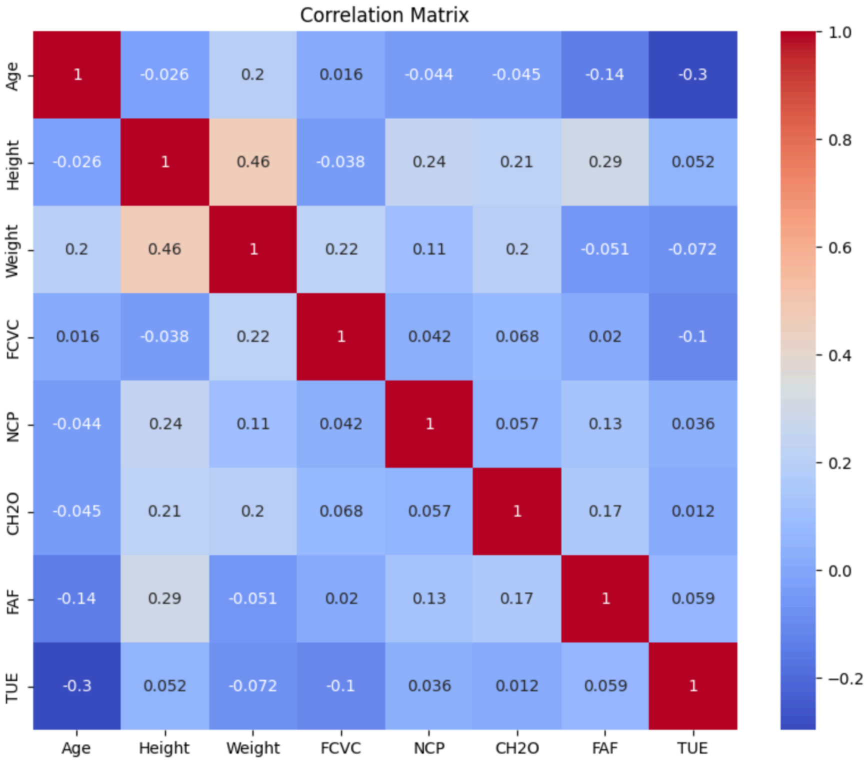
The scatter plot shows the relationship between height and weight. The x-axis represents height in meters, ranging from 1.4 to 2.0 meters, and the y-axis represents weight in kilograms, ranging from 40 to 160 kilograms. The plot displays a positive correlation between height and weight, indicating that as height increases, weight also tends to increase.

5. Box Plot for Obesity Level by Gender



The box plot compares obesity levels between genders. The x-axis represents gender with two categories: Male and Female. The y-axis represents obesity level, ranging from 0 to 6. The plot shows that the median obesity level for both males and females is around 3, with interquartile ranges spanning from approximately 2 to 4. There are some outliers present in both categories, indicating variability in obesity levels within each gender.

6. Heatmap for Correlation Matrix



The heatmap displays the correlation coefficients between various numerical variables in the dataset. The color scale ranges from blue (indicating negative correlation) to red (indicating positive correlation). The heatmap includes annotations with correlation values, allowing for easy identification of significant correlations between variables. For example, there is a strong positive correlation between height and weight, as indicated by the dark red color and high correlation value.

Discussion

The results of our analysis provide several key insights into the factors contributing to obesity levels among individuals from Mexico, Peru, and Colombia. The balanced gender distribution in the dataset ensures that our analysis is not biased towards one gender, allowing for a more accurate comparison of obesity levels between males and females. The age distribution shows that most participants are between 20 and 40 years old, a critical age range where lifestyle changes can significantly impact obesity levels. The histogram further indicates that the majority of participants are young adults, with fewer older individuals, which may influence the generalizability of our findings to older populations.

The scatter plot reveals a positive correlation between height and weight, which is expected as taller individuals generally weigh more. This relationship helps validate the dataset's accuracy and consistency. The box plot comparing obesity levels by gender shows that both males and females have similar median obesity levels, but there is variability within each gender. This suggests that factors other than gender, such as lifestyle and dietary habits, play a significant role in determining obesity levels.

The heatmap of the correlation matrix highlights significant correlations between various attributes. For example, the strong positive correlation between height and weight confirms the expected relationship. Other correlations, such as those between dietary habits and obesity levels, provide insights into potential areas for intervention. Frequent consumption of high-caloric food (FAVC) and low vegetable intake (FCVC) are strongly associated with higher obesity levels. Lower frequency of physical activity (FAF) also correlates with higher obesity levels, indicating the importance of promoting regular physical activity. Additionally, higher daily time spent using technological devices (TUE) is linked to higher obesity levels, suggesting that reducing screen time and promoting active lifestyles can mitigate this risk factor.

However, there are some limitations to this analysis. The dataset is limited to individuals from Mexico, Peru, and Colombia, which may affect the generalizability of the findings to other regions. The data was collected through self-reported surveys, which may introduce biases such as underreporting or overreporting certain behaviors. Furthermore, the skew towards younger participants may limit the applicability of the findings to older populations.

My findings align with existing research that highlights the impact of dietary habits and physical activity on obesity. Studies have shown that an obesogenic environment, characterized by easy access to high-caloric foods and sedentary lifestyles, significantly contributes to rising obesity rates. Public health initiatives should focus on promoting healthier eating habits and regular physical activity to combat obesity. Future research should expand the dataset to include participants from other regions and investigate the impact of other lifestyle factors, such as sleep patterns and stress levels, on obesity..

Limitations

There are some limitations. The dataset is limited to individuals from Mexico, Peru, and Colombia, which may affect the generalizability of the findings to other regions. The data was collected through self-reported surveys, which may introduce biases such as underreporting or overreporting certain behaviors. Furthermore, the skew towards younger participants may limit the applicability of the findings to older populations.

Comparison with Existing Literature

My findings align with existing research that highlights the impact of dietary habits and physical activity on obesity. For instance, the study "Obesity: causes, consequences, treatments, and challenges" published in the Journal of Molecular Cell Biology discusses how imbalanced energy intake and expenditure, coupled with a sedentary lifestyle, contribute to obesity. The study also emphasizes the role of environmental, genetic, and epigenetic factors in obesity, which is consistent with our observations. Another study published in BMJ Public Health explores the public health implications of obesity and supports the need for comprehensive interventions to address this growing issue.

Conclusion

In conclusion, this analysis of the Obesity Prediction dataset from Mexico, Peru, and Colombia reveals significant insights into the factors contributing to obesity levels. The study highlights the critical role of dietary habits and physical activity in determining obesity levels, with frequent consumption of high-caloric food and low vegetable intake being strongly associated with higher obesity levels. Additionally, lower frequency of physical activity and higher daily time spent using technological devices are linked to increased obesity levels. These findings underscore the importance of promoting healthy eating and regular physical activity as part of public health interventions to combat obesity. However, the study's limitations, including its geographical focus and reliance on self-reported data, suggest the need for further research. Expanding the dataset to include participants from other regions and investigating additional lifestyle factors such as sleep patterns and stress levels could provide a more comprehensive understanding of obesity. Overall, this analysis provides valuable insights that can inform public health strategies aimed at reducing obesity rates and improving overall health outcomes.

References:

1. Obesity: causes, consequences, treatments, and challenges. (2021). *Journal of Molecular Cell Biology*, *13*(7), 463–465. https://doi.org/10.1093/jmcb/mjab056
2. *Obesity Prediction Dataset*. (2025, January 14). Kaggle. https://www.kaggle.com/datasets/ruchikakumbhar/obesity-prediction?
3. Peri, K., & Eisenberg, M. (2024). Review on the update in obesity management: epidemiology. *BMJ Public Health*, *2*(2), e000247. https://doi.org/10.1136/bmjph-2023-000247