**MACHINE LEARNING LAB 1**

**EDA**

*Name: Vishal Sreekumar*

*Regnum: 2241163*

*Class: 5 BCA B*

**Introduction**

This dataset is designed for analyzing individual health and dietary patterns. It includes a comprehensive set of attributes such as demographic details, physical characteristics, activity levels, dietary preferences, and nutritional intake. The dataset is ideal for exploratory data analysis (EDA), allowing researchers to uncover patterns, correlations, and trends related to health, nutrition, and lifestyle. Insights from this dataset can guide personalized health recommendations, nutrition planning, and identification of risk factors for diseases.

**Dataset Description**

The dataset comprises 14 attributes:

* **Ages**: Represents the age of individuals in years.
* **Gender**: Categorical variable indicating gender (e.g., Male, Female).
* **Height**: Numeric attribute representing height in centimeters.
* **Weight**: Numeric attribute representing weight in kilograms.
* **Activity Level**: Categorical variable describing the physical activity level (e.g., Sedentary, Moderate, Active).
* **Dietary Preference**: Categorical variable indicating dietary habits (e.g., Vegetarian, Vegan, Non-Vegetarian).
* **Daily Calorie Target**: Numeric value specifying the target calorie intake per day.
* **Protein**: Numeric value representing the protein intake in grams.
* **Sugar**: Numeric value for daily sugar intake in grams.
* **Sodium**: Numeric value for sodium intake in milligrams.
* **Calories**: Total calories consumed per day, a numeric value.
* **Carbohydrates**: Numeric value for carbohydrate intake in grams.
* **Fiber**: Numeric value indicating fiber intake in grams.
* **Fat**: Numeric value representing fat consumption in grams.
* **Disease**: Categorical variable indicating any associated diseases or conditions (e.g., Diabetes, Hypertension, None).

**Significance**

* **Demographics**: Age and gender provide a basis for understanding patterns within population segments.
* **Nutrition**: Nutritional attributes such as protein, sugar, sodium, and fat intake give a snapshot of dietary habits.
* **Health Monitoring**: The Disease attribute enables linking dietary and lifestyle factors to health outcomes.
* **Activity Analysis**: The Activity Level attribute provides context for energy expenditure and calorie needs.

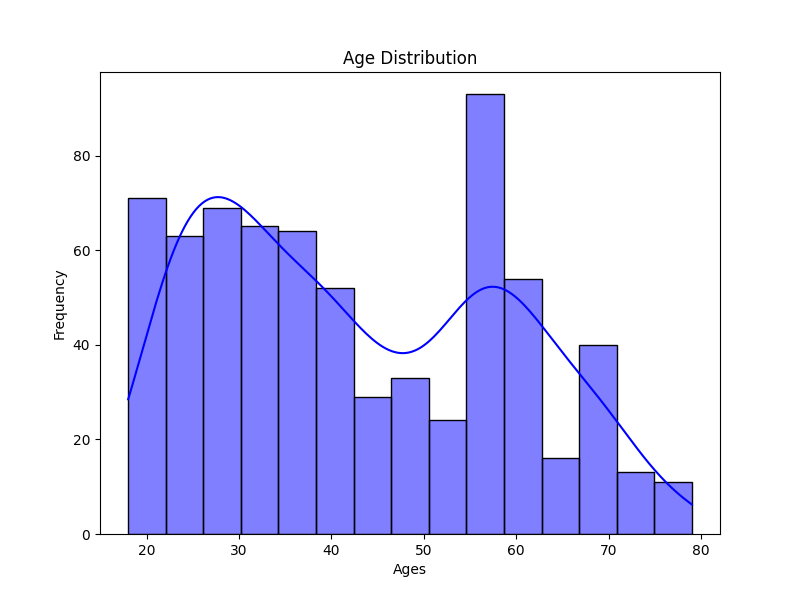
This dataset is suitable for tasks like nutritional planning, disease risk assessment, and developing health-focused machine learning models.

**Exploratory Data Analysis**

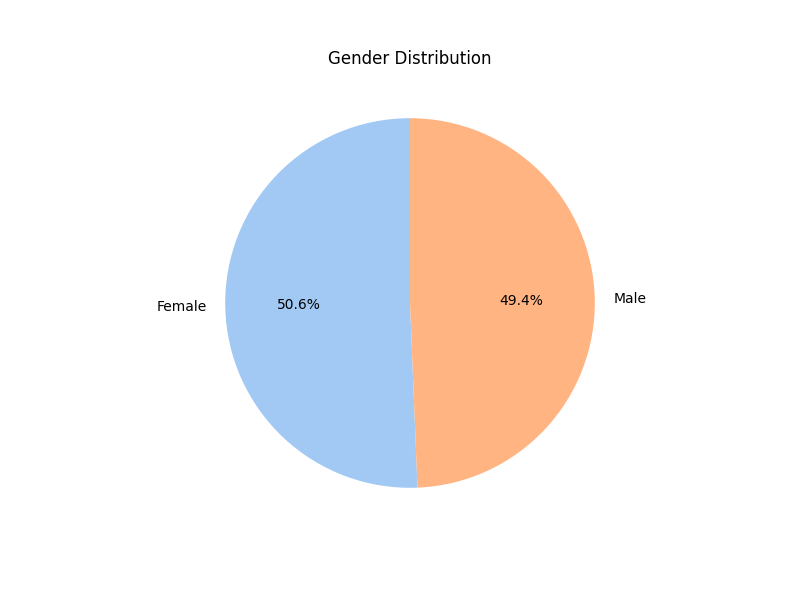
**Univariate Analysis:**

* **Description**: Examines one variable at a time to understand its distribution, central tendency, and spread.
* **Purpose**: Provides insights into individual variable characteristics such as mean, median, mode, range, or skewness.
* **Examples in Code**: Age distribution, Gender count, Activity Level count.

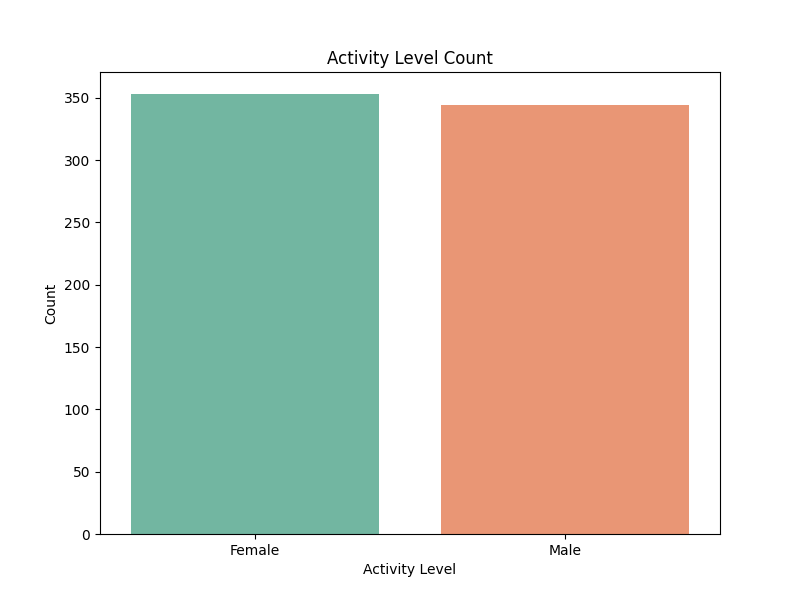
1. **Age Distribution (Histogram + KDE)**
   * **Description: A histogram with a KDE curve that visualizes the distribution of Ages in the dataset.**
   * **Observation: The dataset may show a peak in certain age ranges (e.g., young adults), indicating the dominant demographic group.**
   * **Implication: Target interventions or recommendations to the largest age group, such as diet plans or activity guidelines.**

****

1. **Gender Count (Pie Chart)**
   * **Description: A pie chart representing the proportion of males and females in the dataset.**
   * **Observation: Gender distribution may be balanced or skewed. For instance, females might represent 60% and males 40%.**
   * **Implication: An imbalanced dataset could bias health recommendations or analysis outcomes.**

****

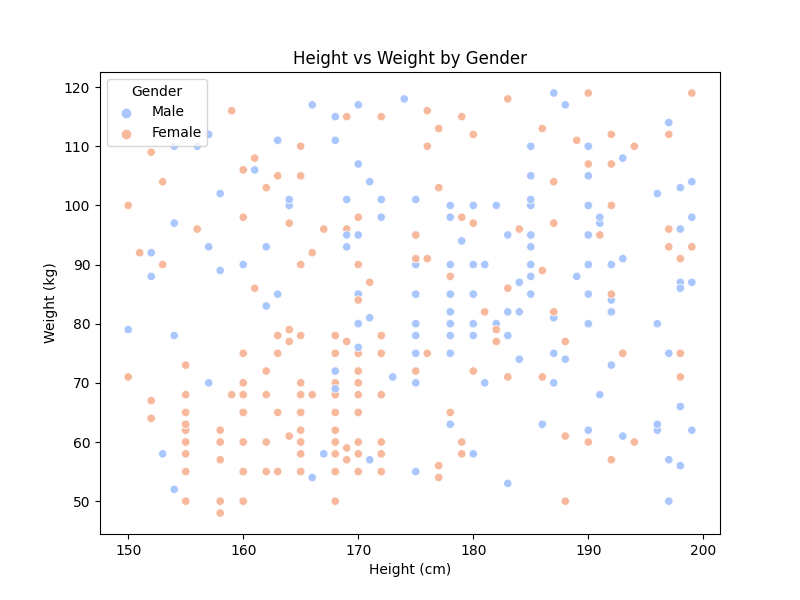
1. **Activity Level Count (Bar Chart)**
   * **Description: A bar chart showing the frequency of different activity levels (e.g., Sedentary, Moderate, Active).**
   * **Observation: Sedentary individuals might dominate, highlighting a need for promoting physical activity.**
   * **Implication: Reinforces the focus on fitness programs or tailored activity-specific dietary guidelines.**



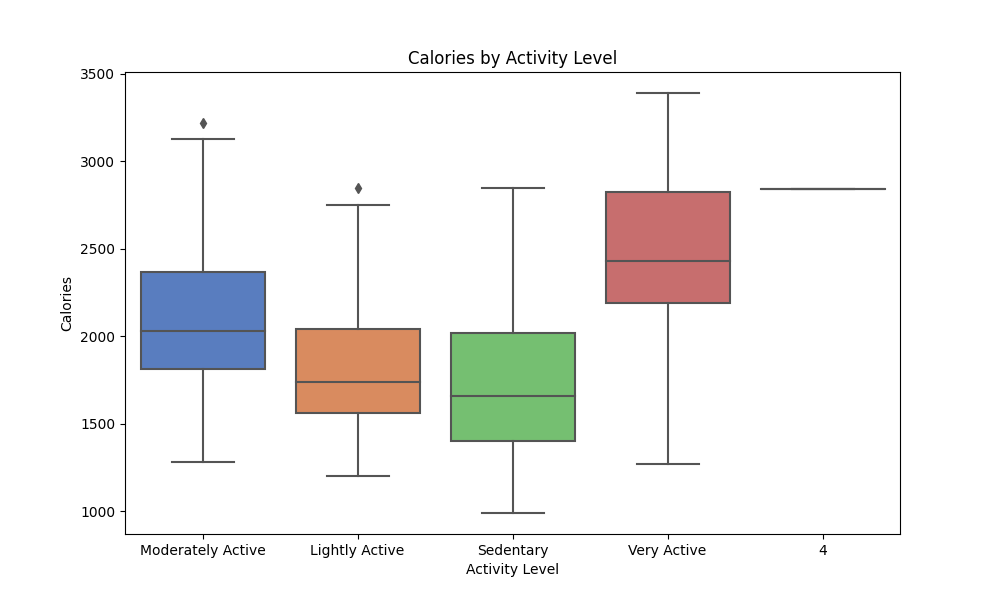
**Bivariate Analysis:**

* **Description: Explores the relationship between two variables and evaluates how one variable influences or is associated with another.**
* **Purpose: Identifies patterns, trends, or correlations between variables.**
* **Examples in Code: Height vs Weight by Gender, Calories by Activity Level, Sugar Intake by Dietary Preference.**

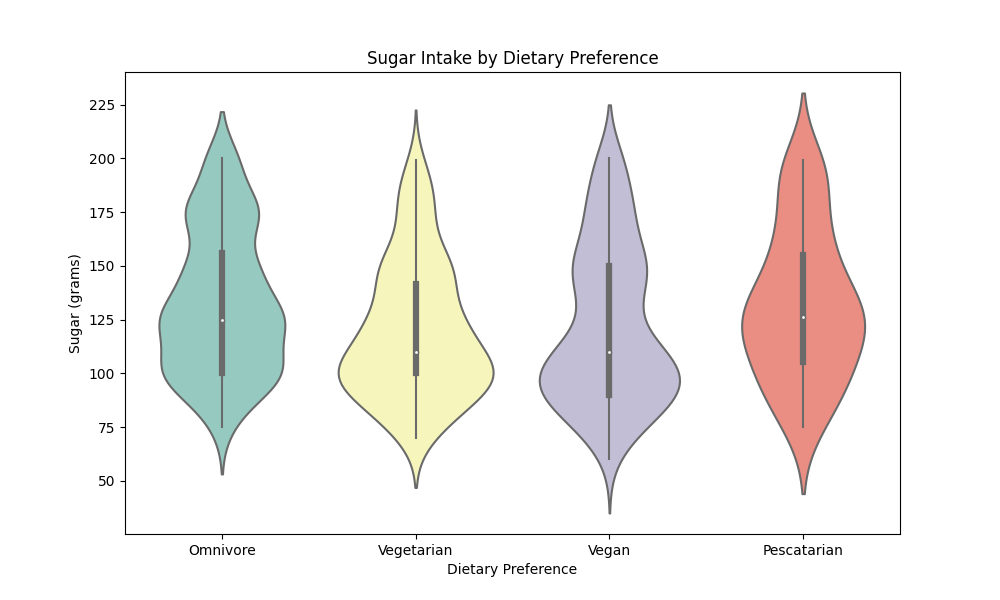
1. **Height vs Weight by Gender (Scatter Plot)**
   * **Description: Scatter plot depicting the relationship between Height and Weight, categorized by Gender.**
   * **Observation: Trends like taller individuals generally weighing more might emerge, with possible differences between genders.**
   * **Implication: Insight into height-weight ratios for gender-specific BMI calculations and health assessments.**



1. **Calories by Activity Level (Box Plot)**
   * **Description: Box plot showing calorie consumption across activity levels.**
   * **Observation: Active individuals likely have higher calorie intakes, with some variability within groups.**
   * **Implication: Verifies if calorie intake aligns with activity requirements, providing a benchmark for dietary adjustments.**



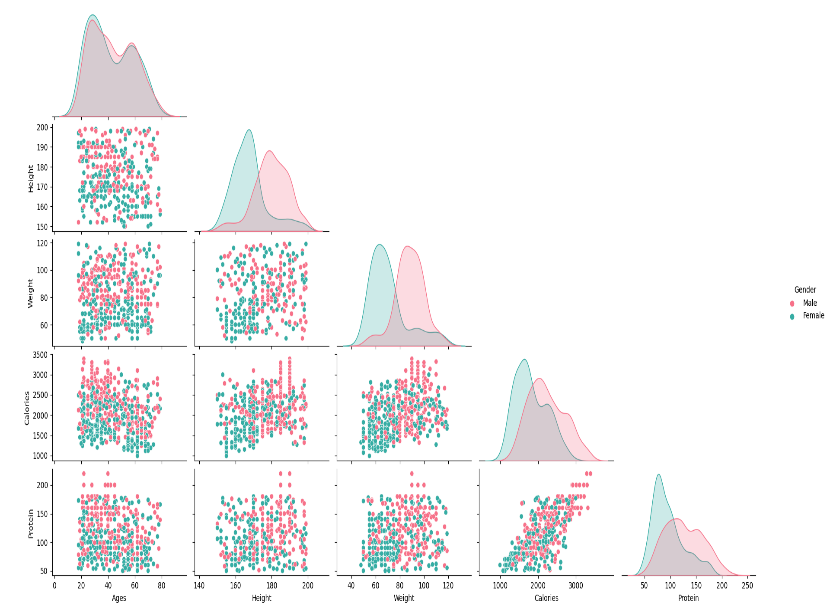
1. **Sugar Intake by Dietary Preference (Violin Plot)**
   * **Description: Violin plot visualizing sugar intake distribution for various dietary preferences.**
   * **Observation: High sugar intake may be more common among individuals with unrestricted diets compared to specific dietary plans.**
   * **Implication: Suggests prioritizing sugar reduction in broader dietary guidelines.**



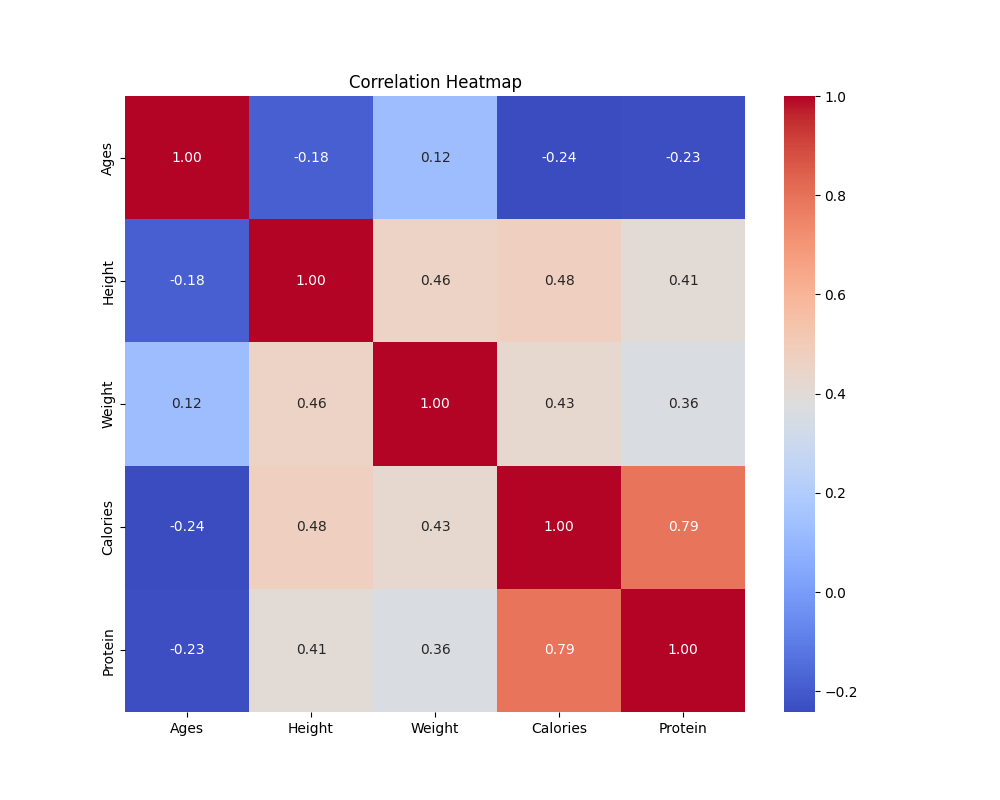
**Multivariate Analysis:**

* **Description: Analyzes interactions among three or more variables simultaneously to uncover complex relationships.**
* **Purpose: Useful for clustering, predicting trends, and understanding multifaceted dependencies.**
* **Examples in Code: Pairplot of numerical features, Correlation heatmap, Dietary Preference vs Disease, Sodium Intake by Disease.**

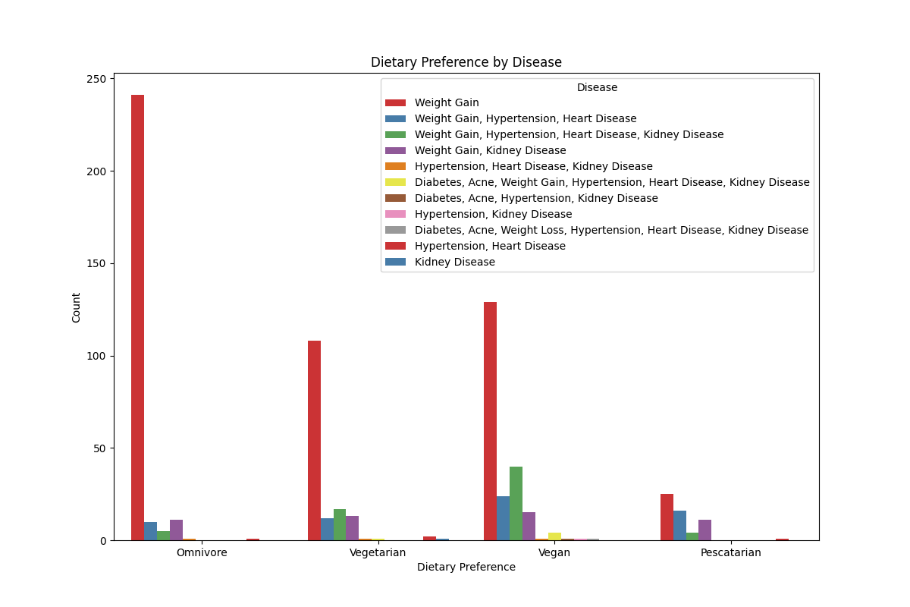
1. **Pairplot of Numerical Features by Gender**
   * **Description: Pairplot of Ages, Height, Weight, Calories, and Protein, categorized by Gender.**
   * **Observation: Clusters or relationships might emerge (e.g., males consuming more calories or having higher protein intake).**
   * **Implication: Guides the creation of gender-specific health or nutrition programs.**



1. **Correlation Heatmap**
   * **Description: Heatmap of correlation coefficients for numerical attributes.**
   * **Observation: Strong correlations (e.g., Calories and Protein) highlight interdependencies.**
   * **Implication: Supports feature selection for predictive modeling or targeted interventions.**



1. **Dietary Preference by Disease (Clustered Bar Chart)**
   * **Description: Clustered bar chart showing dietary preference distribution for different diseases.**
   * **Observation: Unhealthy diets might correlate with higher prevalence of lifestyle diseases.**
   * **Implication: Promotes awareness of dietary adjustments to reduce disease risk.**



1. **Sodium Intake by Disease (Swarmplot)**
   * **Description: Swarmplot visualizing sodium intake across diseases.**
   * **Observation: Diseases like hypertension might associate with higher sodium consumption.**
   * **Implication: Emphasizes sodium intake control in dietary plans for specific diseases.**

