**MACHINE LEARNING LAB 2**

EDA ON CATEGORICAL VALUES

*Name: Vishal Sreekumar*

*Regnum: 2241163*

*Class: 5 BCA B*

**Introduction to EDA on Categorical Values**

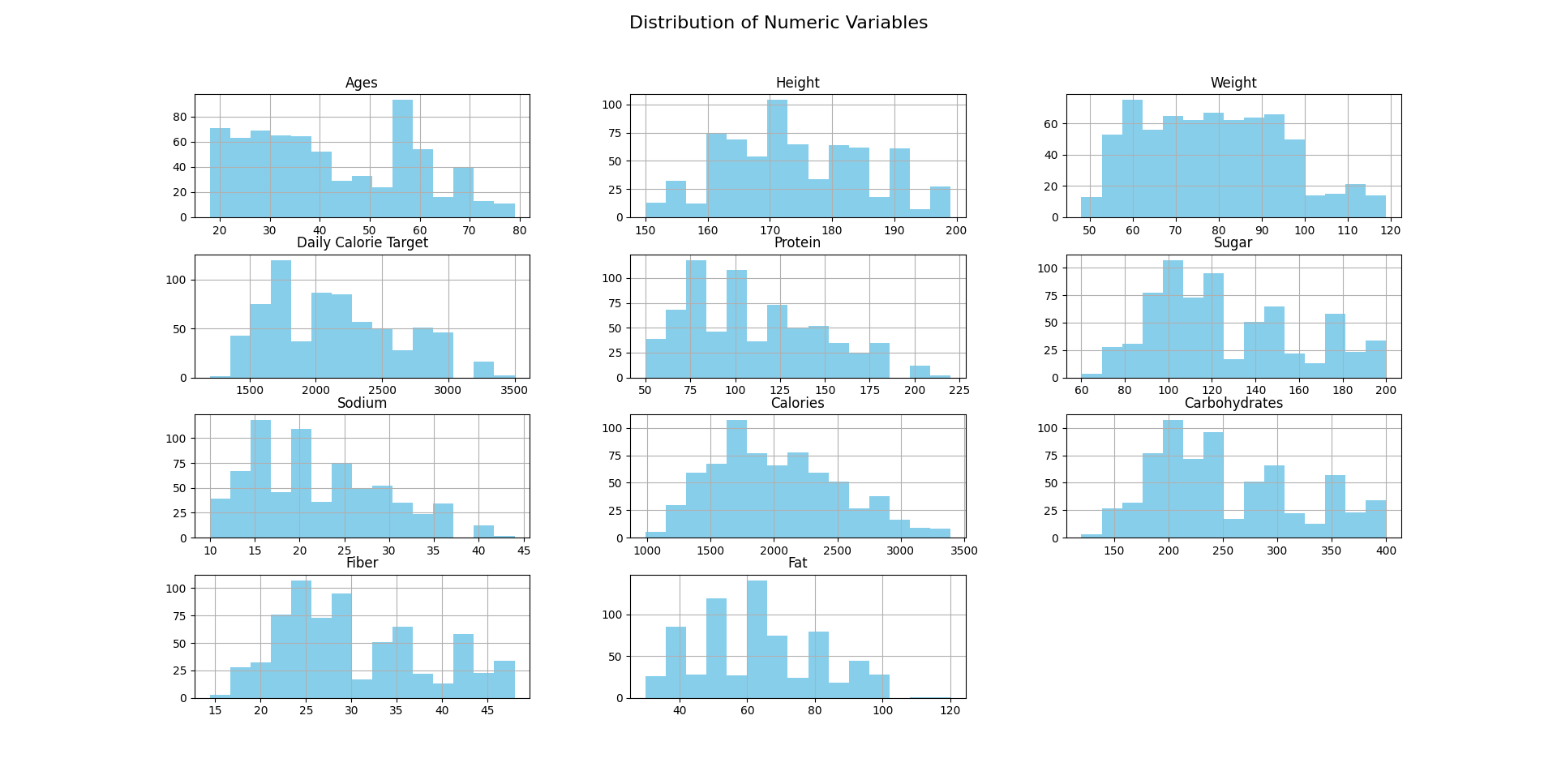
Exploratory Data Analysis (EDA) for categorical values focuses on understanding the distribution and relationships within categorical attributes in a dataset. These analyses help identify patterns, detect outliers, and establish insights for further modeling or decision-making.

EDA typically involves:

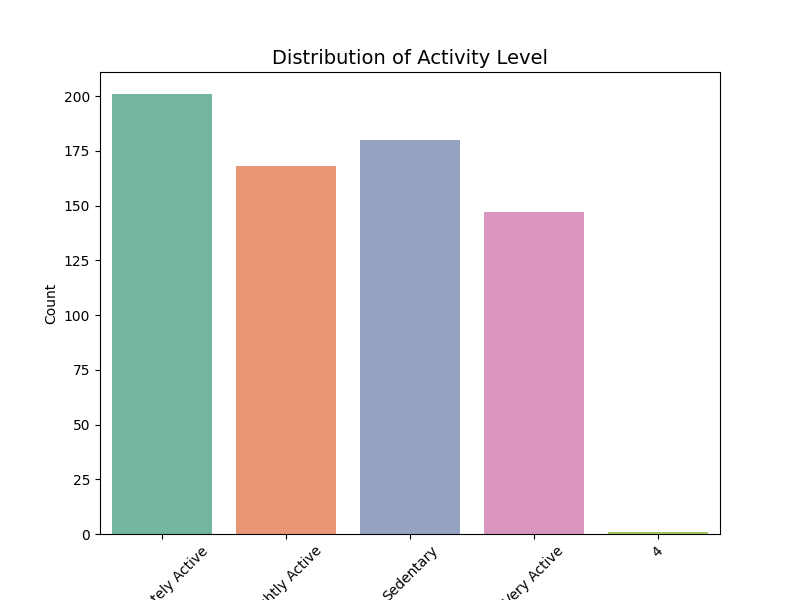
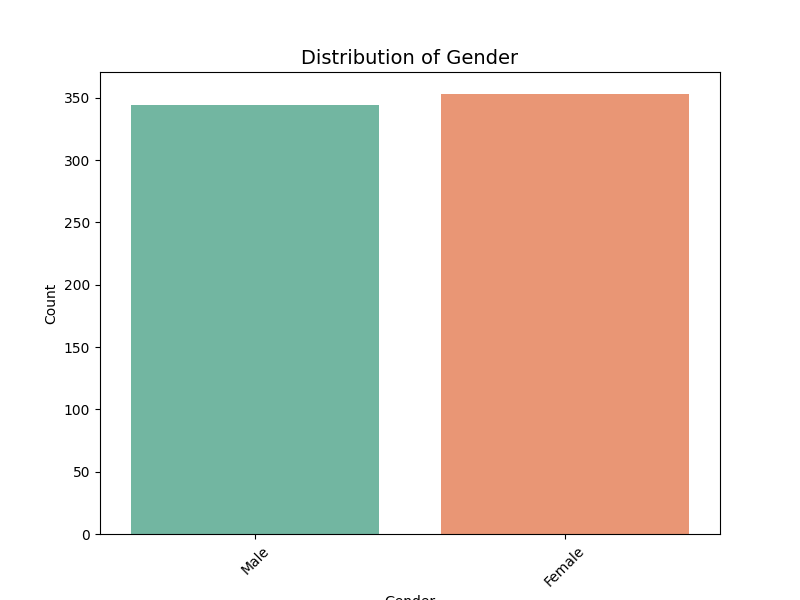
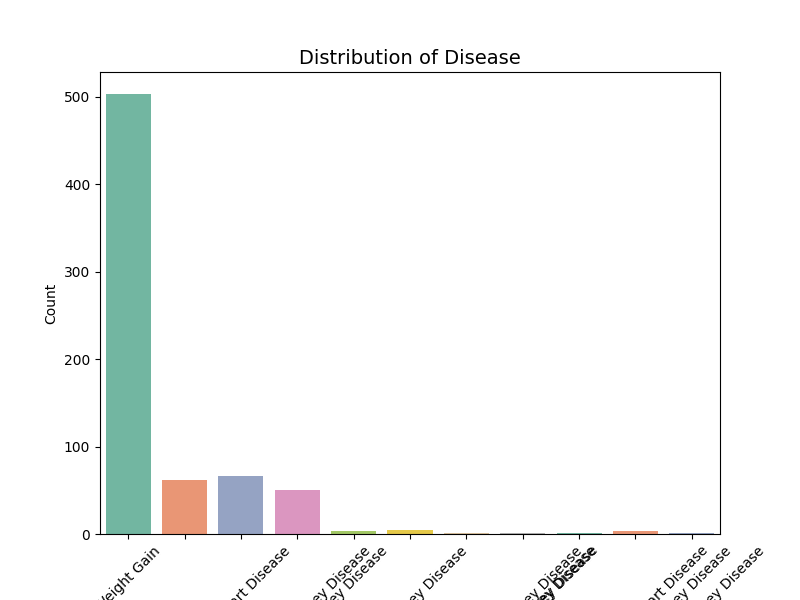
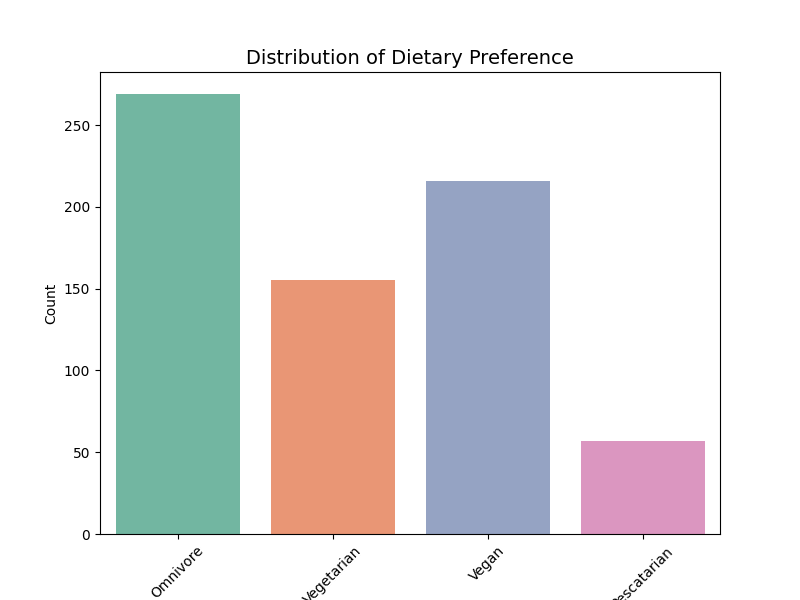
1. **Univariate Analysis:** Examining individual categorical variables to understand their frequency distribution.
2. **Bivariate Analysis:** Investigating relationships between a categorical variable and another (categorical or numerical) variable.
3. **Multivariate Analysis:** Analyzing interactions between multiple variables simultaneously.

**Description, Observation, and Implication for Each Figure**

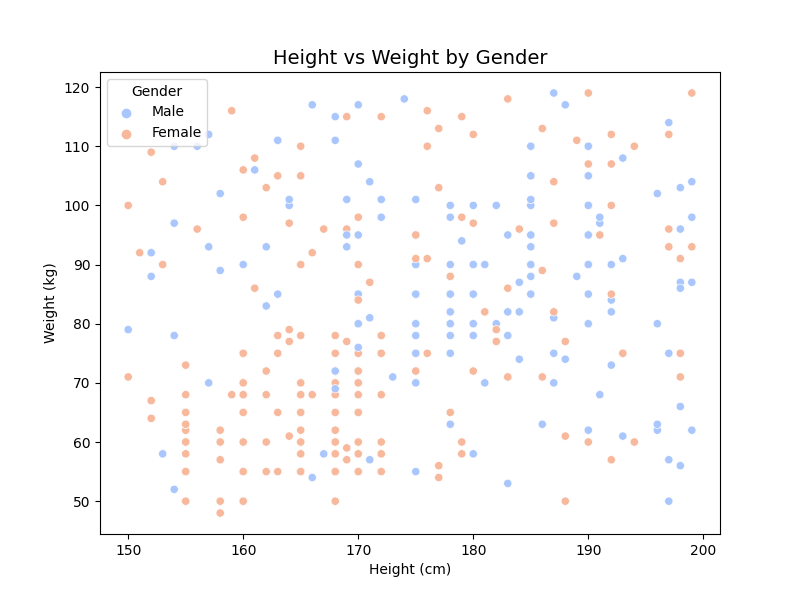
**1. Histograms for Numeric Variables (Cell 4)**

* **Description:** Histograms of numeric variables such as Ages, Height, and Calories show their frequency distribution.
* **Observation:** Certain variables, like Height, exhibit normal distribution, while others, like Calories, may have outliers or skewed data.
* **Implication:** Skewness or outliers in variables like Calories might require transformations or handling for statistical modeling.  
    
    
  

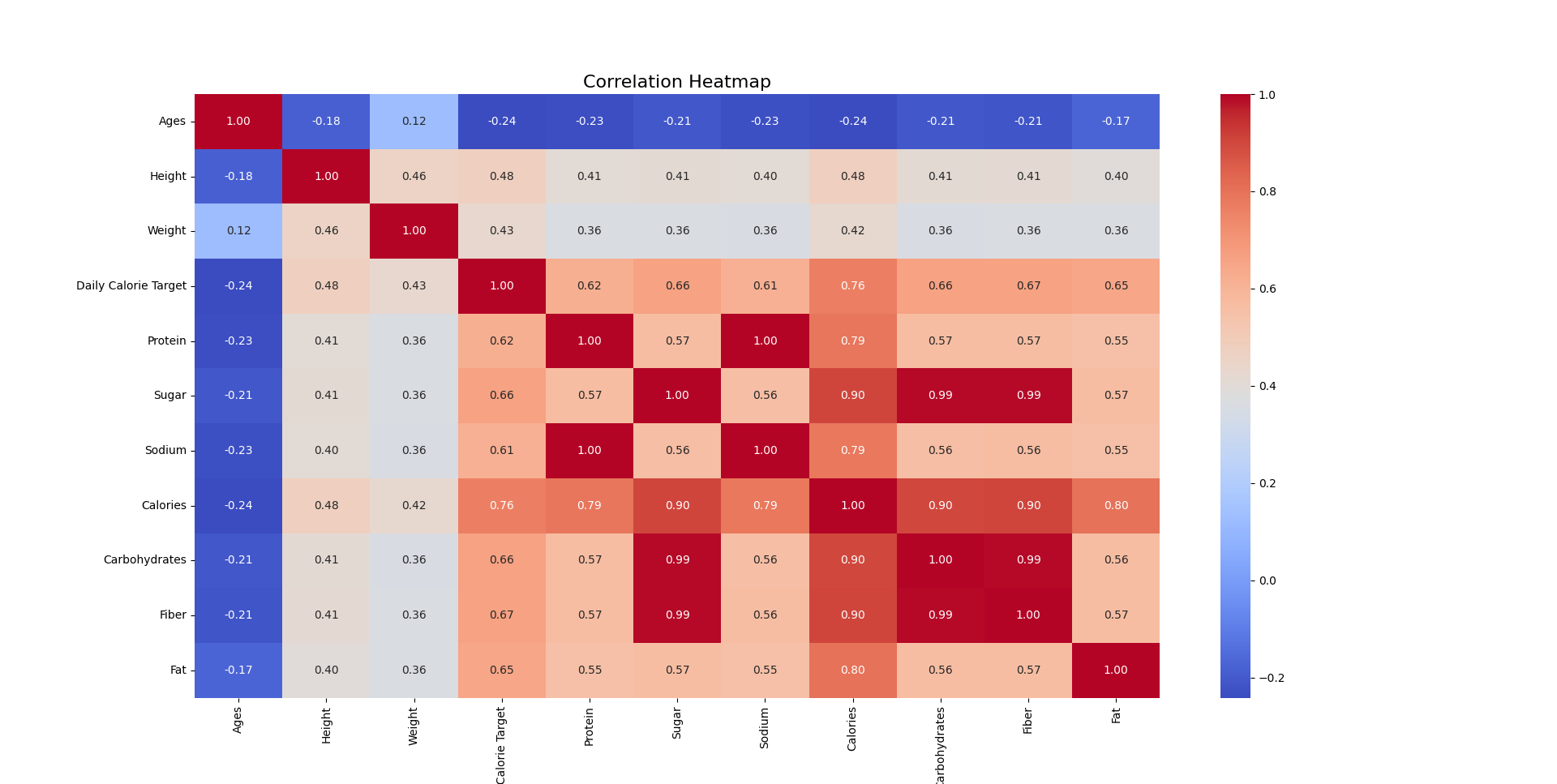
**2. Count Plots for Categorical Variables (Cell 5)**

* **Description:** Count plots visualize the distribution of categorical variables (Gender, Activity Level, Dietary Preference, Disease).
* **Observation:** For instance, Gender distribution might be nearly equal, while specific Activity Levels dominate.
* **Implication:** Skewed categories may require special handling during sampling or analysis to avoid bias.  
    
    
  

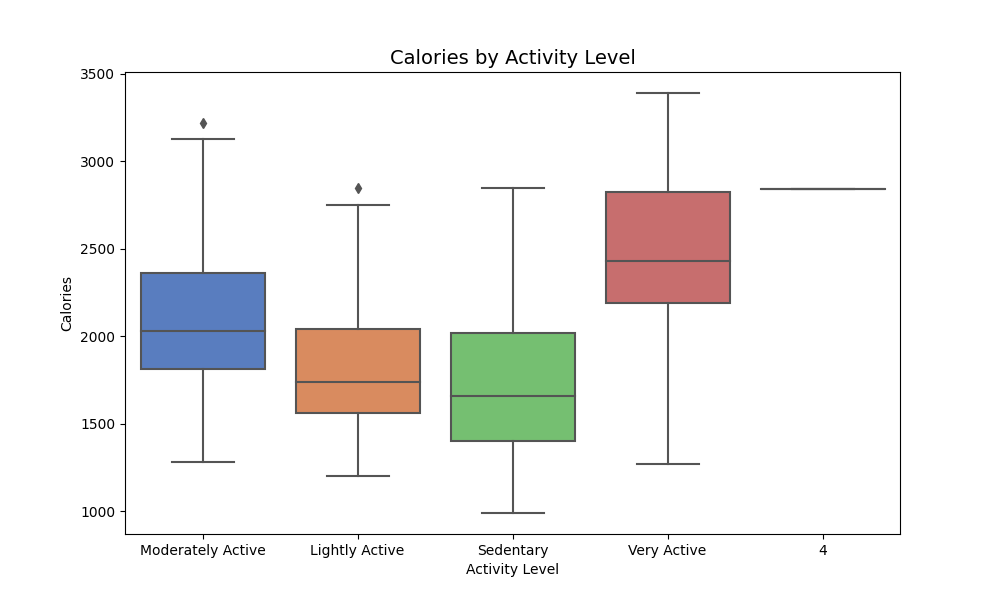
**3. Scatterplot for Height vs Weight by Gender (Cell 6)**

* **Description:** A scatterplot shows the relationship between Height and Weight, grouped by Gender.
* **Observation:** A positive correlation is evident; males and females may exhibit distinct clusters.
* **Implication:** Gender-specific analysis might be required to understand the height-weight relationship accurately.  
    
  

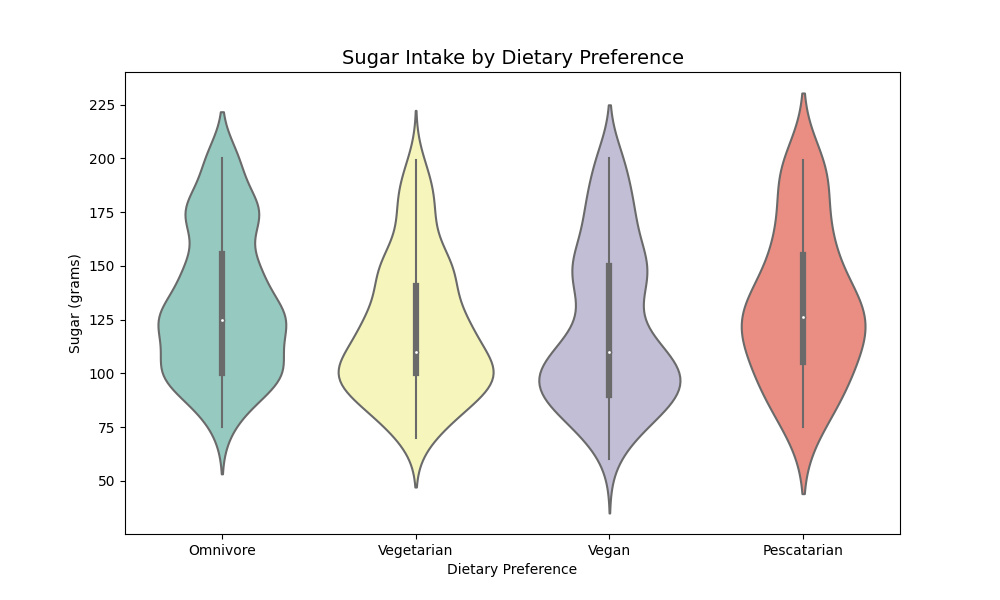
**4. Correlation Heatmap for Numeric Variables (Cell 7)**

* **Description:** Heatmap displays correlations among numeric variables, with darker colors indicating stronger correlations.
* **Observation:** For instance, Calories might correlate strongly with Fat and Carbohydrates.
* **Implication:** Highly correlated variables could lead to multicollinearity, impacting regression models.  
    
  

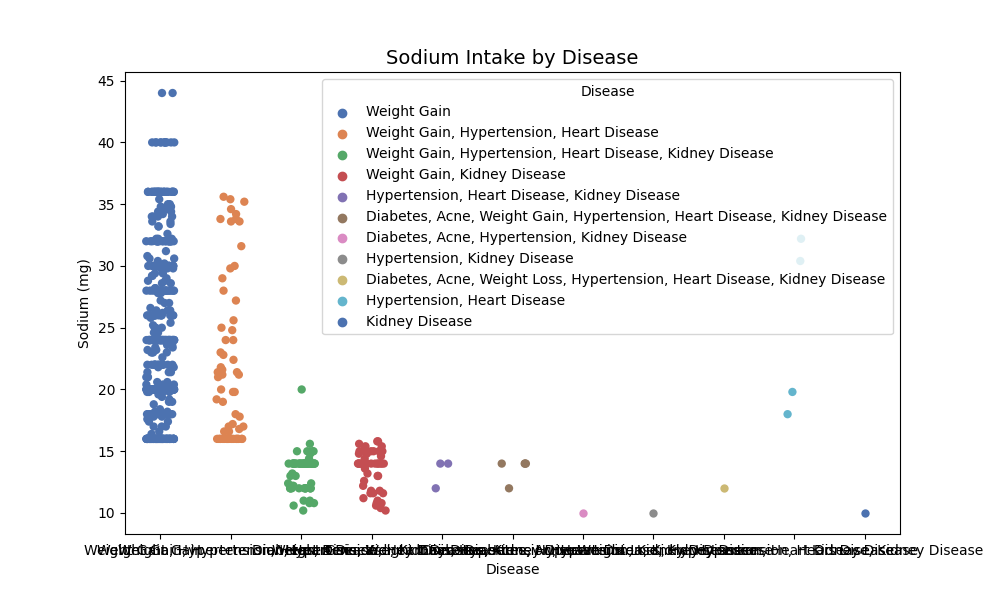
**5. Boxplot of Calories by Activity Level (Cell 8)**

* **Description:** Boxplot shows calorie distribution across different Activity Levels.
* **Observation:** Higher activity levels correspond to wider calorie ranges and higher median values.
* **Implication:** Suggests a positive association between physical activity and calorie intake.  
    
  

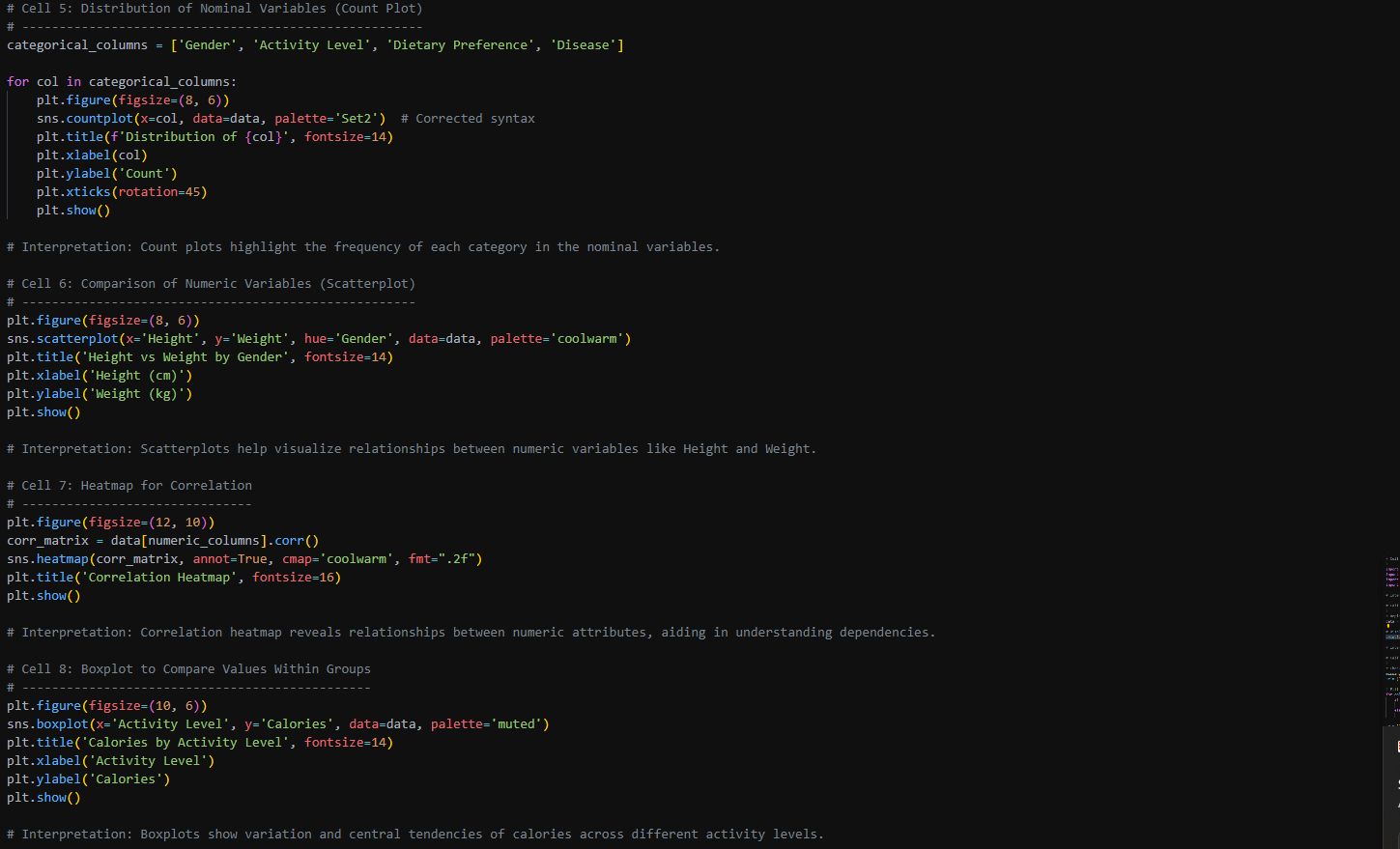
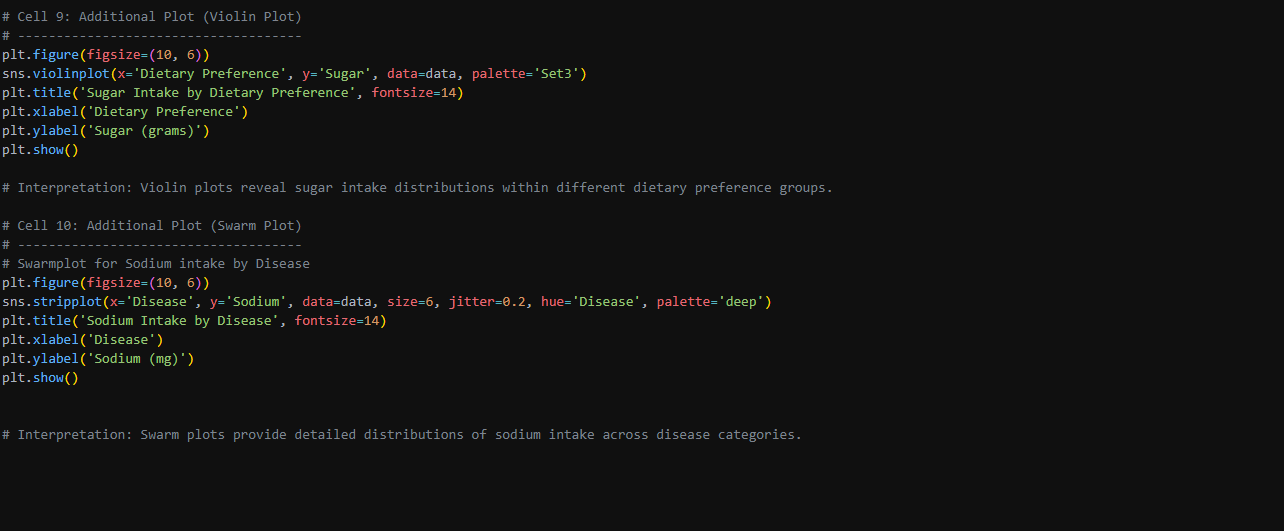
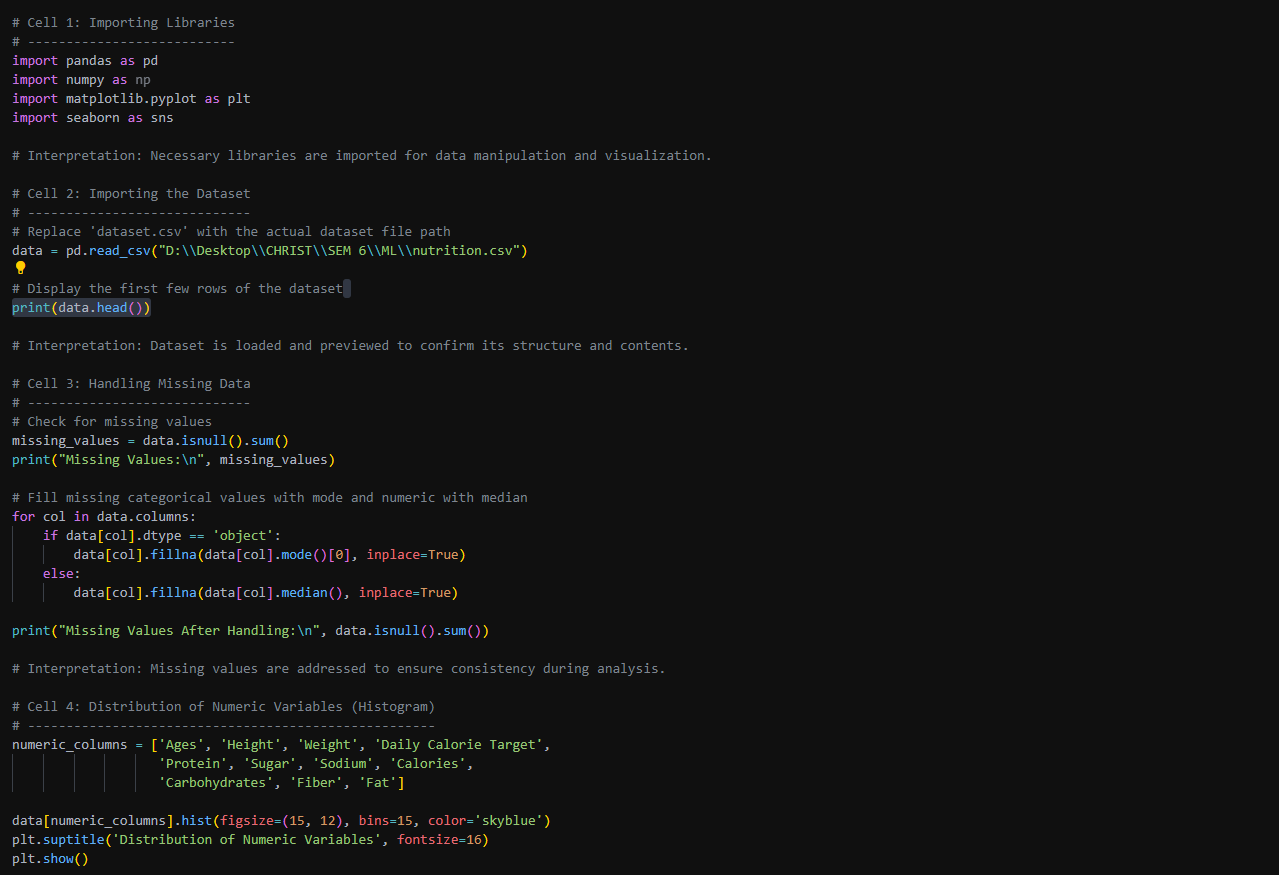
**6. Violin Plot for Sugar Intake by Dietary Preference (Cell 9)**

* **Description:** Violin plot reveals the distribution of Sugar for each Dietary Preference.
* **Observation:** Variations within dietary groups indicate significant differences in sugar consumption.
* **Implication:** Dietary preferences must be considered when analyzing nutrition-related outcomes.  
    
    
  

**7. Swarm Plot for Sodium Intake by Disease (Cell 10)**

* **Description:** Swarm plot details sodium intake for each disease category.
* **Observation:** Sodium levels vary significantly among diseases, with overlapping data points.
* **Implication:** Sodium intake could be a distinguishing factor for specific diseases.  
    
  

**Source Code:**

****

**Conclusion**

EDA on categorical values provides critical insights into the dataset’s structure and trends. It highlights variable distributions, interdependencies, and potential biases. This analysis revealed:

* Skewness in distributions and relationships (e.g., height vs weight, dietary preference vs sugar).
* Significant correlations, such as calories with macronutrients.
* Distinct patterns across categorical groups (e.g., disease and sodium).

The findings guide preprocessing steps like handling skewness, balancing categories, and preparing data for predictive modeling.