Data Analysis Report

Project Title: DEPI Project – Exploratory Data Analysis on Cardiovascular Health

1. Objective

The purpose of this analysis is to perform exploratory data analysis (EDA) on a cleaned cardiovascular dataset. The goal is to identify patterns, evaluate relationships among variables, and derive insights that can support future predictive modeling tasks, particularly for cardiovascular disease (CVD) risk assessment.

2. Data Loading and Preparation

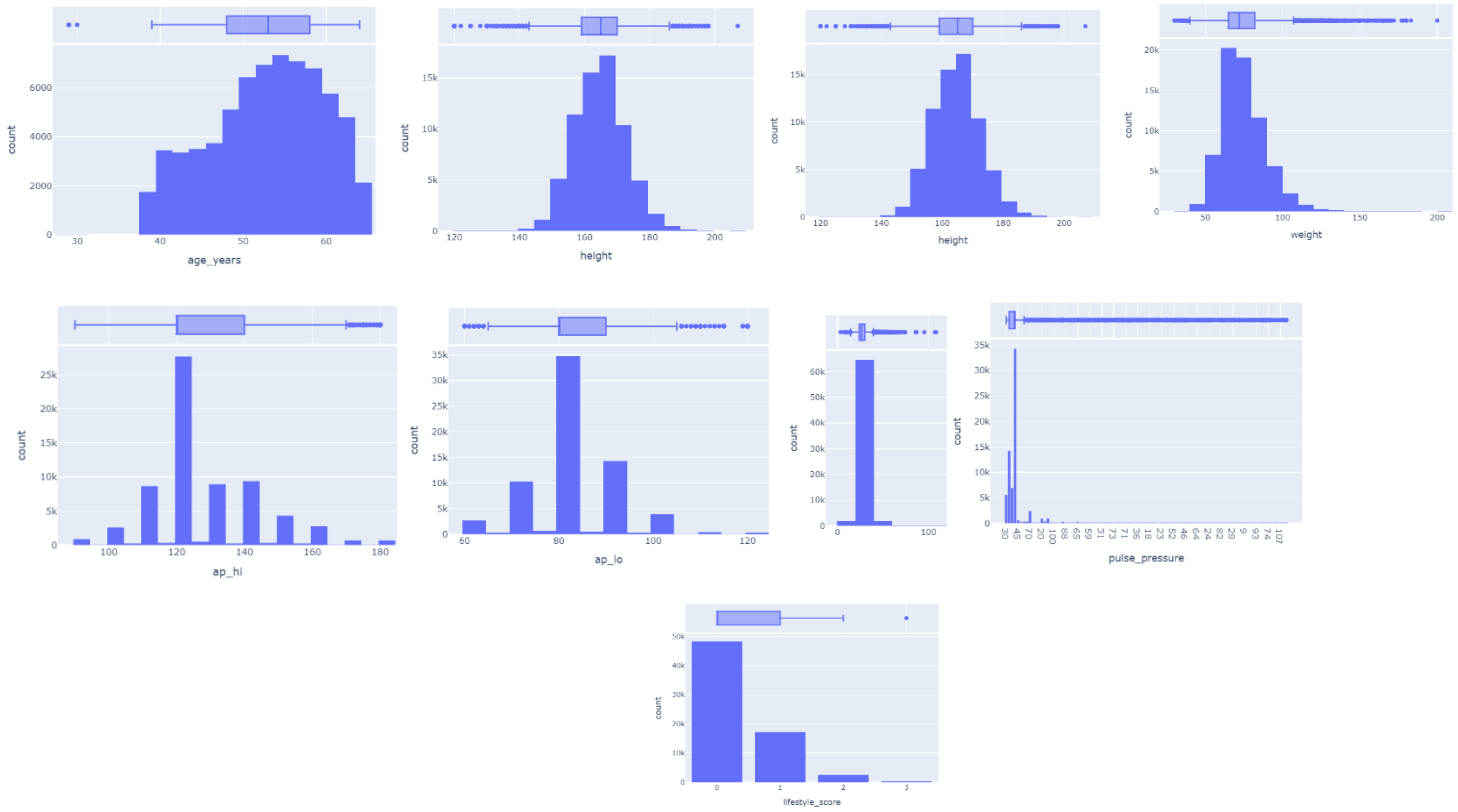
The dataset was imported from **cleaned_data.csv**, which uses tab-separated values. The preprocessing phase involved the following steps:

- **Feature Extraction and Renaming:**
  Variables were categorized into logical groups:
  - **Demographics & Lifestyle: gender**, **age_years**, **smoke**, **alco**, **active**
  - **Physical Measurements: height**, **weight**, **bmi**, **ap_hi** (systolic BP), **ap_lo** (diastolic BP), **pulse_pressure**
  - **Health Indicators: cholesterol**, **gluc**, **cardio**, **is_obese**, **lifestyle_score**, **bmi_category**, **bp_category**
- **Data Type Conversion:**
  Numerical values were cast to appropriate types (**int**, **float**), while categorical variables were converted to **category** data types.

3. Numerical Univariate Analysis

Univariate analysis was conducted to understand the distribution of individual features:

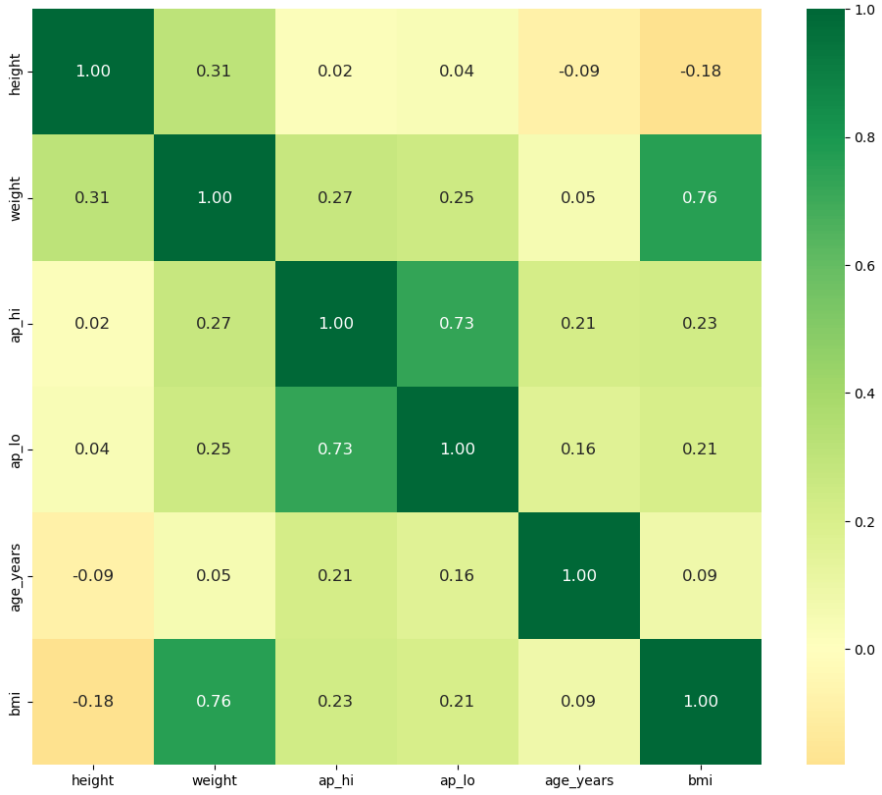
- **Visual Tools Used:** Histograms and box plots provided insight into the feature



- **Observations:**
  - Variables like **age_years**, **bmi**, **ap_hi**, and **ap_lo** exhibited skewness and potential outliers.
  - **lifestyle_score** also showed a wide range of values, suggesting variability in health-related behavior.

4. Correlation Analysis

A correlation matrix was visualized using a Seaborn heatmap to examine linear relationships among numerical features.



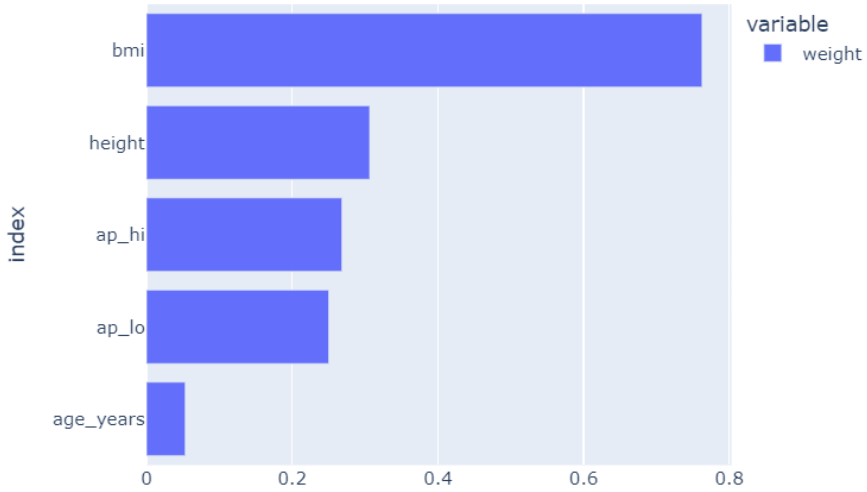
Key Findings:

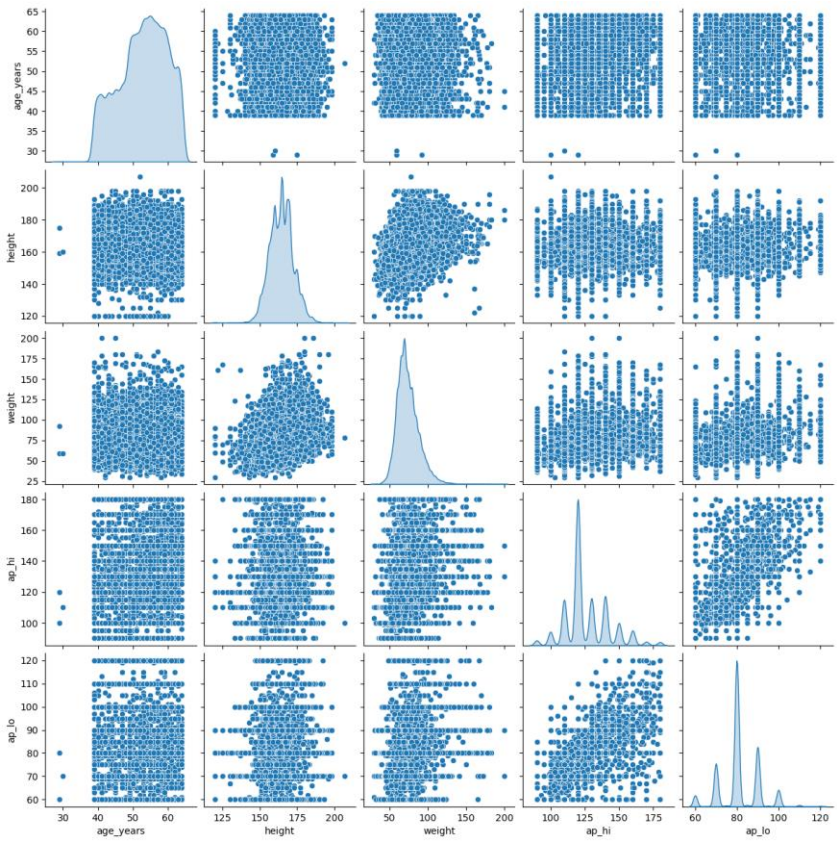
- Strong positive correlation:
    - Between **systolic** (**ap_hi**) and **diastolic** (**ap_lo**) blood pressure.
- Positive correlation:
    - Between **BMI** and **weight**, with a positive correlation between **BMI** and **height**.
- Moderate correlation:
    - Observed between the target variable **cardio** and variables like **age**, **blood pressure**, and **cholesterol levels**.

5. Feature Relationships

Multivariate analysis explored relationships among features:

- **Bar Plot:** Showed absolute correlations of features with **weight**.
- **Pair plot:** Revealed interactions between variables such as **cardio**, **age_years**, **ap_hi**, **ap_lo**, and lifestyle-related factors





6. Categorical Mapping

To improve readability and modeling interpretability, several binary categorical variables were relabeled and converted into **category** types:

Feature	Mapping
alco	0 → *Don't Consume*, 1 → *Consume Alcohol*
cardio	0 → *Absence*, 1 → *Presence*
active	0 → *Not physically active*, 1 → *Physically active*
smoke	0 → *Non-smoker*, 1 → *Smoker*

7. Categorical Univariate Analysis

Function uni_cat_analysis(col) generates both a pie chart and a bar chart for each categorical variable. This helps understand the distribution of categories within each feature.

Features Analyzed:

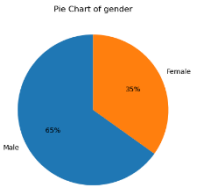
- Gender, cholesterol, gluc, smoke, active, cardio, alco, bmi_category, is_obese

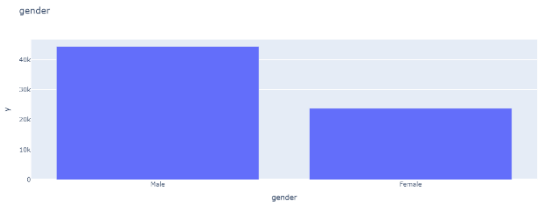
Each feature's frequency was visualized:

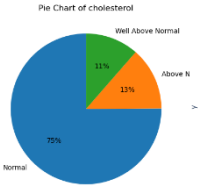
- Pie charts presented percentage distributions.
- Bar charts provided exact counts.

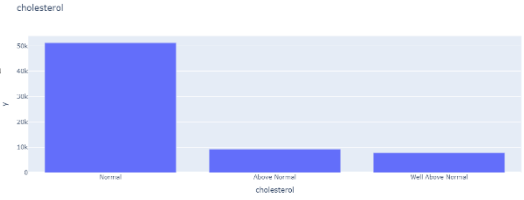
These visualizations highlighted class imbalances and revealed that:

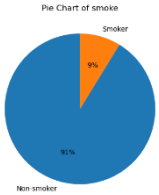
- Most individuals are **physically active** and **non-smokers**.
- There are more males than females.
- Most people fall into the **normal BMI** range, though a notable portion is **obese**.
- **Cholesterol and glucose levels** are mostly in the normal range, but elevated cases are still common.

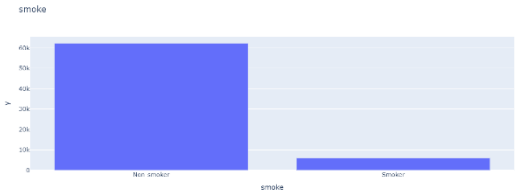


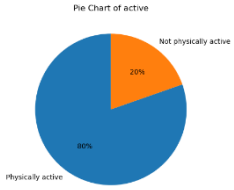


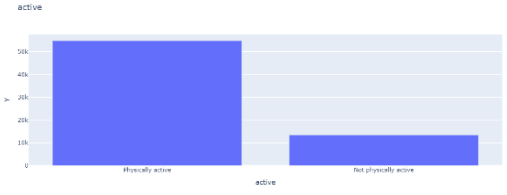


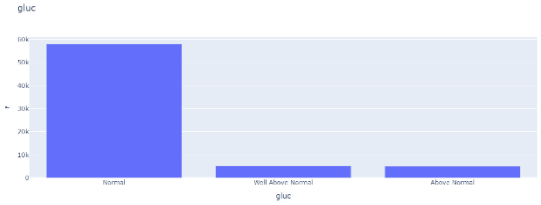


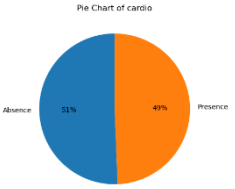


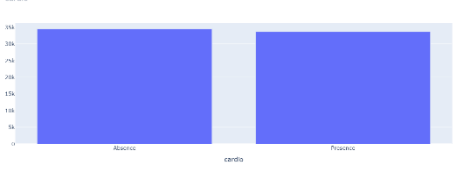


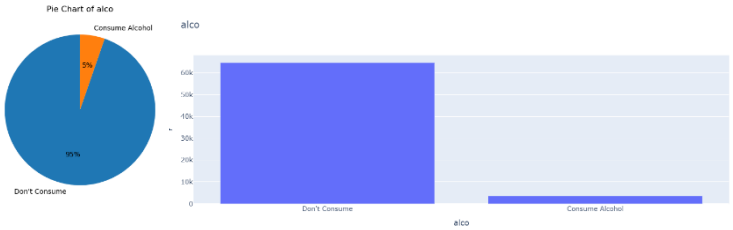


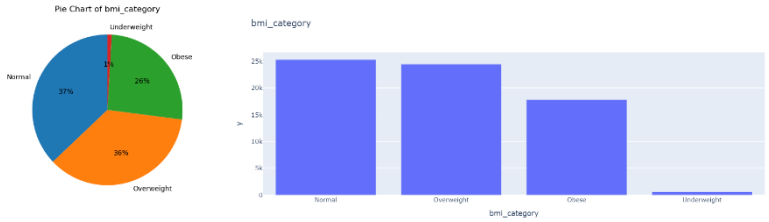


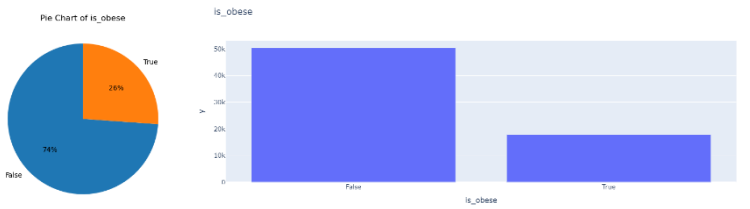







Pie Chart of alco
Consume Alcohol
5%
95%
Don't Consume
alco
Don't Consume
Consume Alcohol
alco


Pie Chart of bmi_category
Underweight
1%
Obese
26%
Normal
37%
36%
Overweight
bmi_category
Normal
Overweight
Obese
Underweight
bmi_category


Pie Chart of is_obese
True
26%
74%
False
is_obese
False
True
is_obese

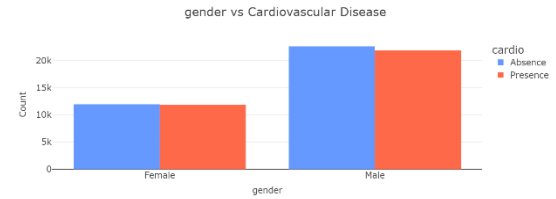
8. Categorical Features vs (Cardio)

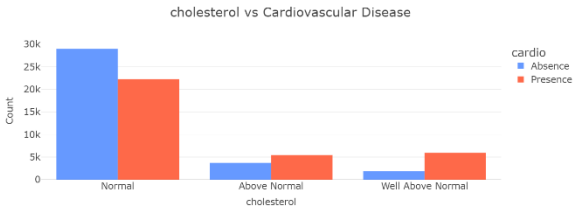
Visualized Features:

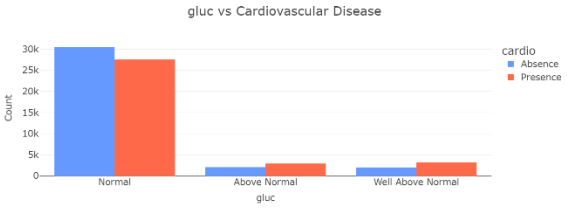
- gender, cholesterol, gluc, smoke, active, alco, bmi_category, is_obese, bp_category

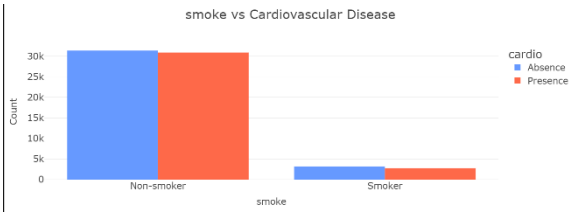
Observations:

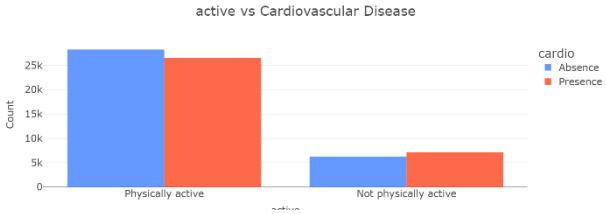
- Higher cholesterol and gluc levels are positively associated with cardiovascular disease.
- Physically inactive Individuals, **alcohol consumers**, and the **obese** have a higher incidence of cardio.
- **Hypertensive BP categories** (especially high systolic and diastolic) correlate strongly with cardio presence.

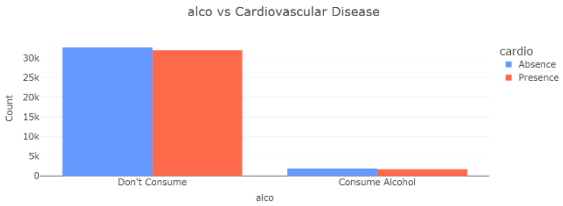


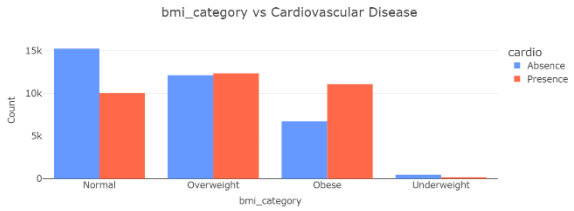


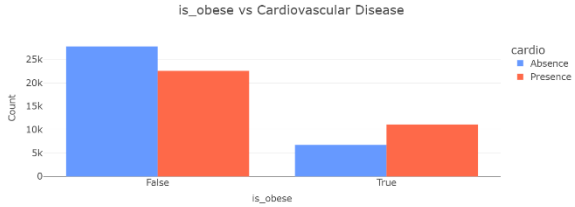


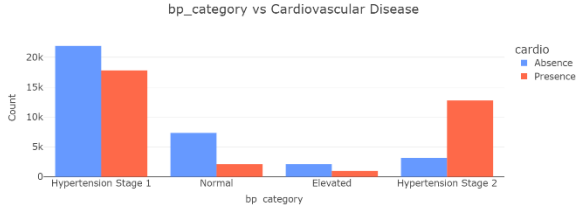


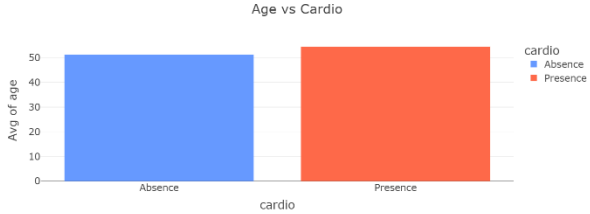






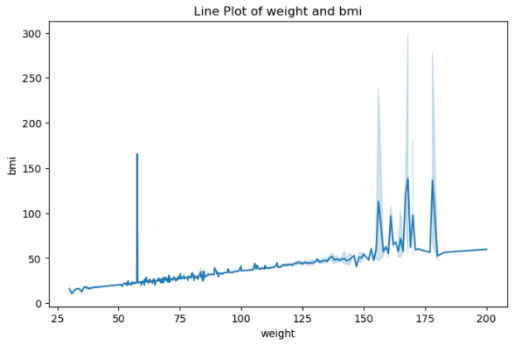






Line Plot: Weight and BMI

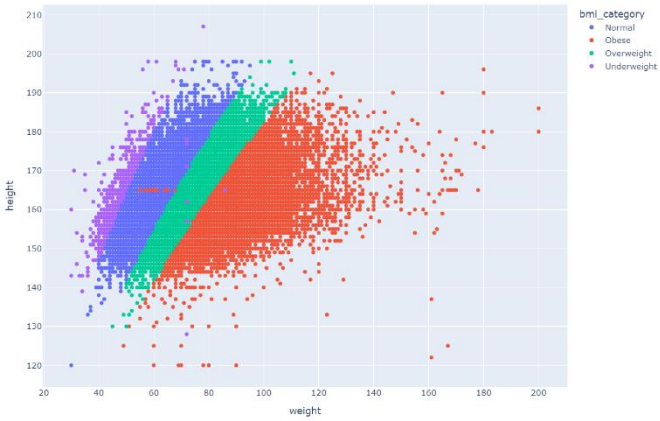
- A Seaborn line plot displayed the average BMI across increasing weight.
- Demonstrated a positive, near-linear relationship — validating BMI as a function of weight over height^2.
- Important for observing trends rather than raw distributions.

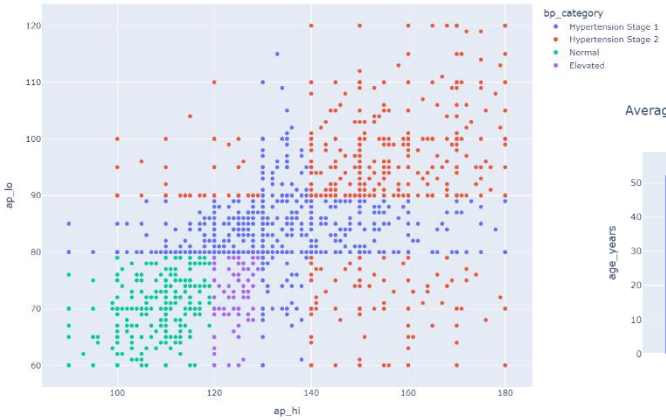


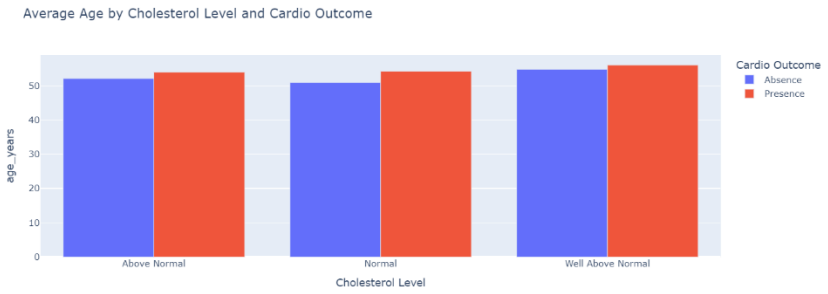
9. Multivariate Analysis

- Blood Pressure (ap_hi vs ap_lo): Higher systolic and diastolic values clustered more in individuals with cardiovascular disease, especially in hypertensive BP categories.
- Weight vs Height (by BMI Category): Obesity was more common among heavier and shorter individuals, confirming BMI category distributions.
- Line Plot (Weight vs BMI): A clear positive trend showed BMI increasing with weight, validating BMI's calculation.
- Average Age by Cholesterol and Cardio: At all cholesterol levels, individuals with cardiovascular disease were older on average, emphasizing the combined risk of age and high cholesterol.

.







10. Summary and Insights

This data analysis provided a structured overview of the cardiovascular dataset, highlighting key patterns and risk factors.

- **Data Preparation**: The dataset was cleaned and enriched with derived features such as BMI, age in years, obesity status, and blood pressure categories.
- **Risk Factors**: Age, high BMI, elevated blood pressure, cholesterol, and glucose levels were strongly linked to cardiovascular disease.
- **Lifestyle Indicators**: Smoking, alcohol use, and lack of physical activity showed notable associations with disease presence.
- **Multivariate Patterns**: Combined effects of age, cholesterol, and blood pressure revealed higher risk profiles, especially in older, obese individuals with poor health metrics.

Overall, the analysis emphasized that cardiovascular disease risk arises from both medical indicators and lifestyle behaviors — useful for further predictive modeling.