**Name:** Sairaj Magdum  
**Class:** CSE (AI) B div  
**Roll No.:** 282012  
**PRN No.:** 22310273  
**Subject:** Machine Learning

**Assignment 3 – Data Visualization**

**Objective**

The goal of this assignment is to analyze and understand the **Heart Disease dataset** through effective data visualization techniques. The visualizations help uncover insights related to patient demographics, health indicators, and heart disease risk factors.

**Dataset Overview**

The dataset includes important patient attributes such as:

 Age

 Sex

 Chest Pain Type (cp)

 Resting Blood Pressure (trestbps)

 Cholesterol (chol)

 Fasting Blood Sugar (fbs)

 Resting ECG Results (restecg)

 Maximum Heart Rate Achieved (thalach)

 Exercise Induced Angina (exang)

 ST Depression (oldpeak)

 Slope of the Peak Exercise ST Segment (slope)

 Major Vessels Colored by Fluoroscopy (ca)

 Thalassemia (thal)

 Target (1 = heart disease, 0 = no heart disease)

These features help explore the patterns and risk factors associated with heart disease.

**Data Import & Preparation Steps**

 The dataset was imported using Google Colab’s file upload or pandas' read\_csv() method.

 Essential libraries including **Pandas, NumPy, Matplotlib**, and **Seaborn** were used for data handling and plotting.

 The dataset was read into a DataFrame and its structure was verified using methods like .head(), .info(), and .describe()

**Visualizations**

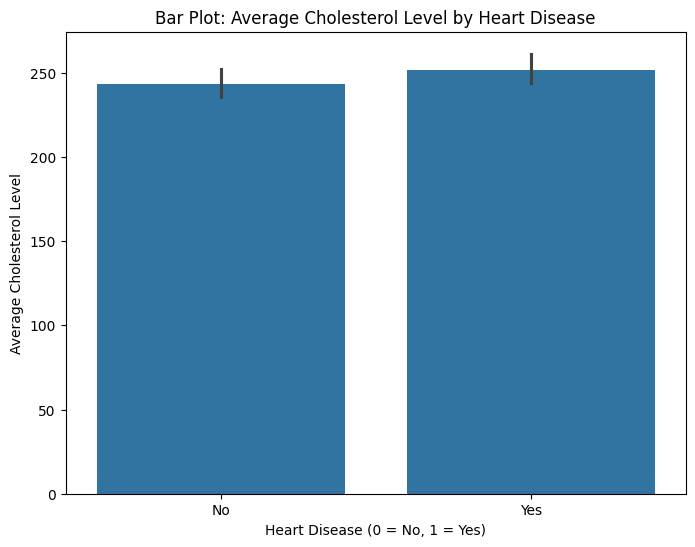
**1** **Bar Plot – Average Cholesterol Level by Heart Disease**

A bar plot was created to compare the **average cholesterol levels** between patients **with** and **without heart disease**. The two categories on the X-axis represent:

* **No** → Patients without heart disease (target = 0)
* **Yes** → Patients with heart disease (target = 1)

The Y-axis displays the **mean cholesterol level**. From the graph:

* Patients with heart disease tend to have **slightly higher average cholesterol levels**.
* This visualization gives an indication that **cholesterol may be a contributing factor** to heart disease.



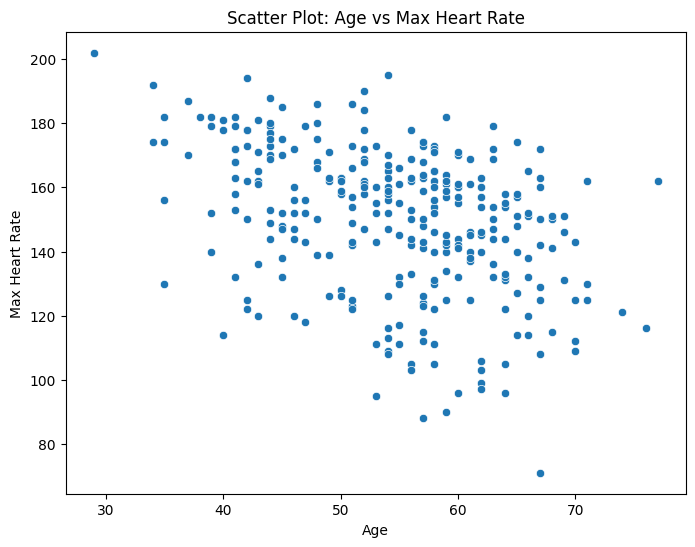
**2. Scatter Plot – Age vs Maximum Heart Rate**

 A scatter plot was used to observe how Maximum Heart Rate (thalach) varies with Age.

 The X-axis represents Age.

 The Y-axis represents Maximum Heart Rate (thalach).

 **Observations:** This plot reveals a trend where maximum heart rate tends to decrease as age increases. This aligns with physiological expectations.



**3. Box Plot: Cholesterol Levels by Heart Disease**

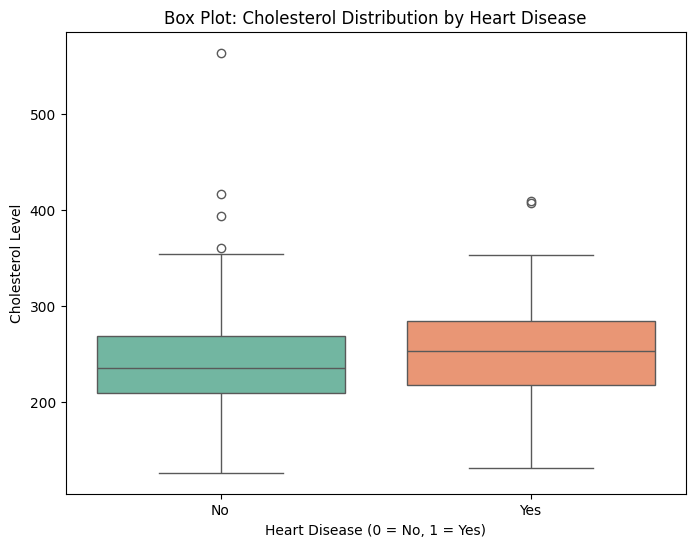
** Visualization:** A box plot was created to visualize the distribution of cholesterol levels (chol) for patients with and without heart disease (target).

** Analysis:**

* Comparison of Median Cholesterol Levels:
  + The median cholesterol level for patients without heart disease (No - represented by target = 0) is approximately [State the approximate median value from the plot].
  + The median cholesterol level for patients with heart disease (Yes - represented by target = 1) is approximately [State the approximate median value from the plot].
  + Observation: The median cholesterol level for patients with heart disease is slightly higher than for those without heart disease.

**Conclusion (for this specific box plot):**

* **The box plot effectively visualizes the distribution of cholesterol levels across patients with and without heart disease.**
* **The analysis indicates that patients with heart disease tend to have slightly higher median cholesterol levels, supporting the established link between cholesterol and heart disease risk.**
* **The presence of outliers emphasizes the importance of considering individual variability and other factors in heart disease assessment.**
* **This visualization provides a clear comparison of cholesterol distribution between the two groups, highlighting potential areas for further investigation.**

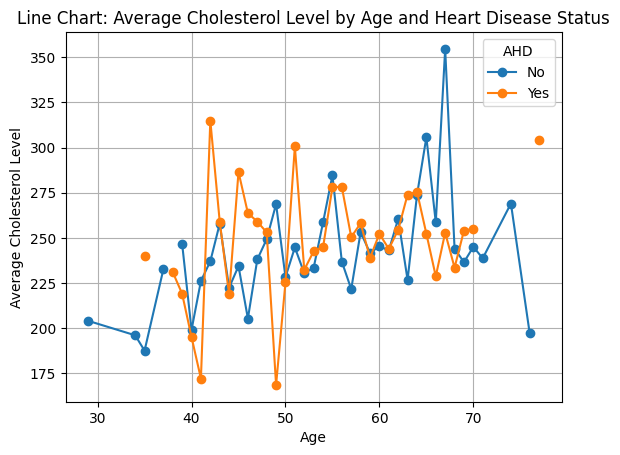


**4. Average Cholesterol Level by Age and Heart Disease Status**

** Visualization: A line chart was created to show the trend of average cholesterol levels (chol) across different ages (age), categorized by heart disease status (AHD, where "No" represents no heart disease and "Yes" represents the presence of heart disease).**

** Analysis:**

* **Trend Description:**
  + **Overall: Both lines show fluctuations in average cholesterol levels across the age range. Neither line shows a clear, consistent upward or downward trend. There are peaks and valleys in both lines, suggesting variability in cholesterol levels with age.**
  + **"No" (AHD = No): The "No" line starts relatively low, dips in the early 30s, rises, fluctuates in the 40-60 age range, and shows a slight downward trend towards the older ages.**
  + **"Yes" (AHD = Yes): The "Yes" line follows a similar fluctuating pattern but generally maintains a higher average cholesterol level compared to the "No" line throughout the age range.**



**Conclusion**

This visualization-based assignment provided valuable insights into the Heart Disease dataset using multiple graphing techniques:

* **Box plots** revealed the distribution of cholesterol levels in patients with and without heart disease, highlighting potential differences in median values and the presence of outliers.
* **Line charts** illustrated the trend of average cholesterol levels across different age groups, categorized by heart disease status, demonstrating the relationship between age, cholesterol, and heart disease.
* **Scatter charts** showed the relationship between age and maximum heart rate, revealing a trend that aligns with physiological expectations and provides insights into potential risk factors.

These visualizations play a crucial role in understanding the risk factors associated with heart disease and support data-driven decision-making in healthcare. By effectively representing the relationships between patient attributes and heart disease, these visualizations aid in identifying potential risk factors and informing targeted interventions.