Report on Heart Attack Dataset

Team: Byte Brigade // Members: Nevin, Archana, Ashish, Elvin, Vaishnav

Objective

The goal of this project is to create an AI-driven solution that can efficiently process and analyze structured heart attack datasets to represent knowledge and generate valuable insights. This solution should identify patterns within the data and produce meaningful information to support decision-making processes.

Problem Description

In today's era of big data, organizations across various sectors generate enormous amounts of data daily. Properly processing and analyzing this data can uncover valuable insights that greatly enhance decision-making processes. The challenge lies in effectively representing this knowledge and extracting useful insights. Your task is to develop an AI-based solution capable of addressing this challenge by processing a provided structured dataset, representing the knowledge within it, and generating meaningful insights.

Dataset Source

- **Dataset Source:** `www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-`
- **Key Features:** The dataset contains reports from 304 patients on the likelihood of experiencing a heart attack, based on the following factors:
  - Age (`age`)
  - Gender (`sex`)
  - Chest pain type (`cp`)
  - Resting blood pressure (`trtbps`)
  - Serum cholesterol (`chol`)
  - Fasting blood sugar (`fbs`)
  - Resting electrocardiographic results (`restecg`)
  - Maximum heart rate achieved (`thalachh`)
  - Exercise induced angina (`exng`)
  - ST depression induced by exercise relative to rest (`oldpeak`)
  - Slope of the peak exercise ST segment (`slp`)
  - Number of major vessels colored by fluoroscopy (`caa`)
  - Thalassemia (`thall`)

Methods Used

- **Clustering Algorithm:** K-Means
- **Classification Algorithms:** Decision Tree Classifier, Logistic Regression
- **Regression Algorithms:** Linear Regression, Decision Tree Regressor

Tools

- **Python Libraries:**
  - Pandas: Data manipulation and analysis
  - NumPy: Numerical operations
  - Matplotlib and Seaborn: Data visualization
  - Scikit-learn: Machine learning and model evaluation

Results & Findings : Detailed Insight Generation Stages

1 Data Preprocessing Stage

1.1 Dataset Overview

- Initial Data Shape: The dataset contains 303 entries with 14 columns.
- Missing Values: There are no missing values in any column.
- Data Types: The dataset primarily consists of integer data types (13 columns) and one float data type.

1.2 Key Insights

- Data Completeness: The absence of missing values suggests that the dataset is well-maintained and does not require any imputation strategies.
- Dataset Structure: The dataset contains 14 columns, including the target variable (output).
- Memory Usage: The dataset's memory usage is relatively low (33.3 KB).
- Feature Types: The dataset includes both categorical and continuous features.
- Feature Relevance: The features selected are clinically relevant for predicting heart attacks.

1.3 Outlier Analysis

- Outliers were detected in several features:
  - trtbps: 2 outliers
  - chol: 4 outliers
  - thalachh: 1 outlier
  - oldpeak: 2 outliers
  - caa: 5 outliers
  - thall: 2 outliers
- Impact on Data Shape: After removing the detected outliers, the dataset's shape was reduced from 303 entries to 287 entries.
- Feature-Specific Insights:
  - trtbps (Resting Blood Pressure) and chol (Cholesterol Level) had relatively few outliers.
  - caa (Number of Major Vessels Colored by Fluoroscopy) had the highest number of outliers (5).

  - oldpeak (ST Depression Induced by Exercise) and thall (Thalassemia) had 2 outliers each.

- Data Integrity: Removing outliers ensures that the dataset is cleaner and more representative of the typical patient profile.

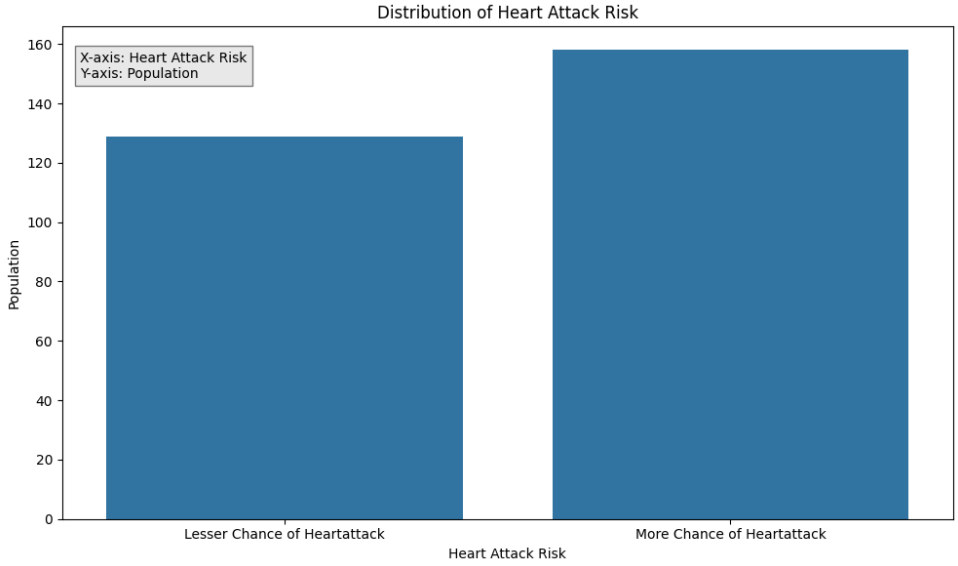
1.4 Dataset Composition

- The dataset contains 287 entries and 14 columns, including the target variable output.
- All columns are complete with no missing values.
- Feature Normalization: The features have been normalized to ensure a consistent scale.
- Initial Data Preview: The first few rows of the dataset provide a snapshot of various attributes.

2 Knowledge Representation Stage

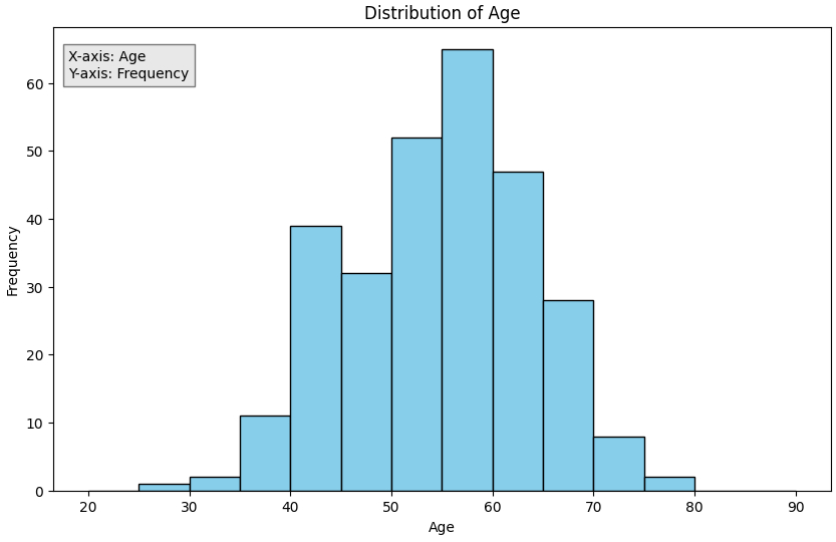
2.1 Distribution of Heart Attack Risk (Histogram)

- Population Split:
  - Lesser Chance of Heart Attack (0): 129 patients
  - More Chance of Heart Attack (1): 158 patients
- Implications: The dataset appears to have a relatively balanced representation between the two categories.



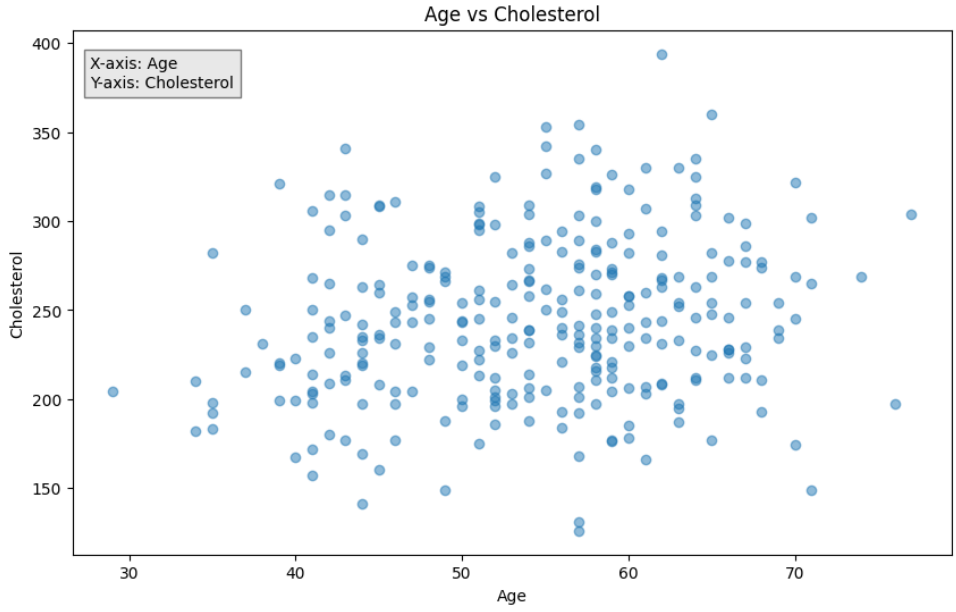
2.2 Distribution of Age (Histogram)

- The highest frequency appears to be in the range of 55-60 years.
- There is a generally increasing trend in frequency up to around 65 years, after which it declines gradually.



2.3 Age vs Cholesterol (Scatterplot)

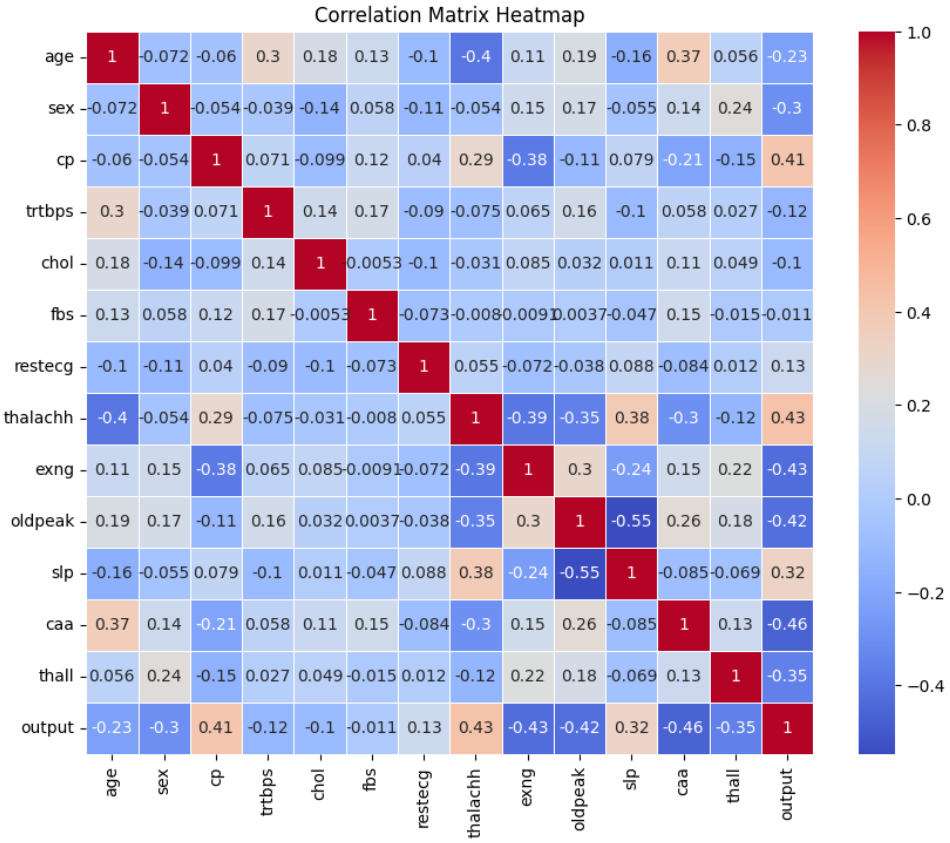
- Correlation Analysis: The positive correlation coefficient of 0.21 indicates a weak relationship between age and cholesterol levels.
- Age Extremes: The dataset includes individuals ranging from 29 to 77 years old.
- Cholesterol Extremes: The dataset shows a wide range of cholesterol levels, from a minimum of 126 mg/dl to a maximum of 564 mg/dl.
- Individual Cases:
  - The oldest person (77 years old) has a cholesterol level of 304 mg/dl.
  - The youngest person (29 years old) has a cholesterol level of 204 mg/dl.
  - The person with the highest recorded cholesterol (564 mg/dl) is aged 67 years.
  - The person with the lowest recorded cholesterol (126 mg/dl) is aged 57 years.
- Health Considerations: Monitoring cholesterol levels across different age groups is crucial for assessing cardiovascular health risks.



2.4 Correlation Matrix Heatmap

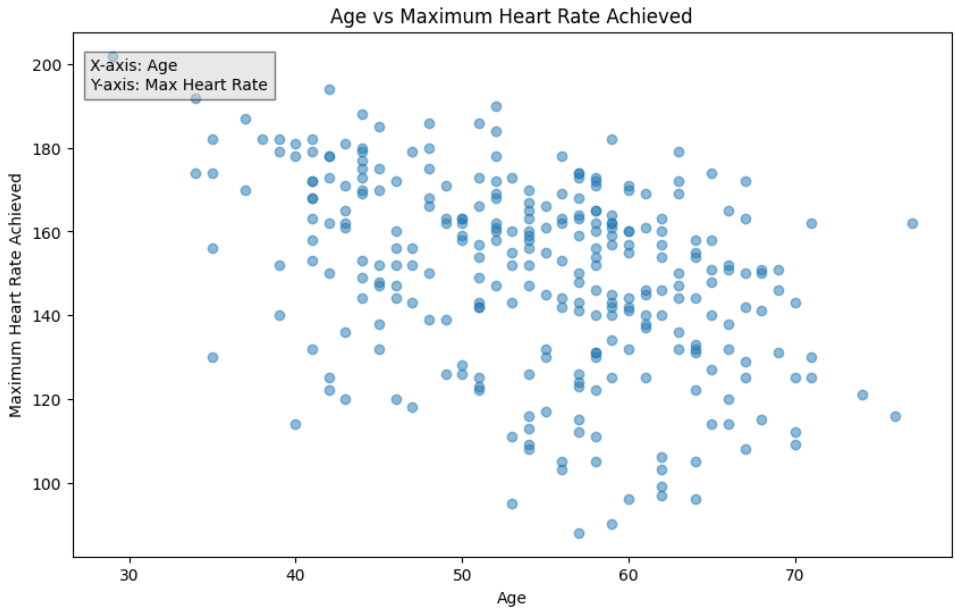
- Age and Health Factors:
  - Resting Blood Pressure (trtbps): Moderate positive correlation with age.
  - Maximum Heart Rate Achieved (thalachh): Moderate negative correlation with age.
  - Number of Major Vessels (caa): Moderate positive correlation with age.
- Sex and Heart Health:
  - Heart Attack Risk (output): Moderate negative correlation with sex.

- Chest Pain and Heart Health:
  - Exercise Induced Angina (exng): Moderate negative correlation with chest pain type (cp).
  - Heart Attack Risk (output): Moderate positive correlation with chest pain type (cp).
- Heart Rate and Exercise:
  - Exercise Induced Angina (exng): Moderate negative correlation with maximum heart rate achieved (thalachh).
  - ST Depression Induced by Exercise (oldpeak): Moderate negative correlation with maximum heart rate achieved (thalachh).
  - Slope of the Peak Exercise ST Segment (slp): Moderate positive correlation with maximum heart rate achieved (thalachh).
- Heart Health Indicators:
  - Heart Attack Risk (output): Moderate correlations with several factors including maximum heart rate achieved (thalachh), chest pain type (cp), exercise-induced angina (exng), ST depression induced by exercise (oldpeak), number of major vessels (caa), and thallium stress test result (thall).



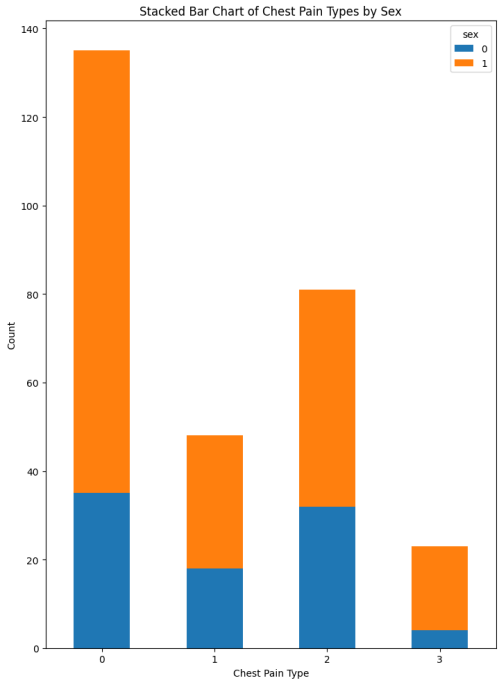
2.5 Age vs Maximum Heart Rate Achieved

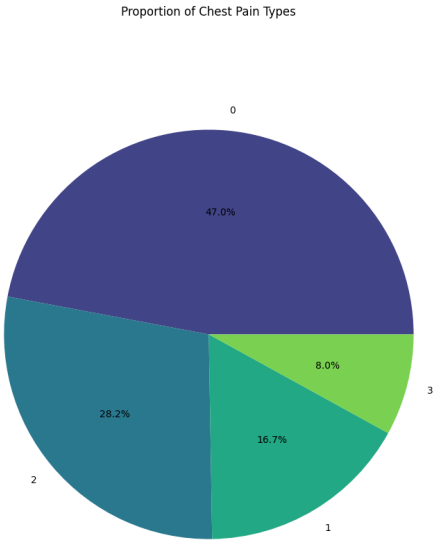
- Mean Age and Maximum Heart Rate Achieved: The average age is approximately 54.37 years, while the average maximum heart rate achieved is around 149.65 bpm.
- Correlation between Age and Maximum Heart Rate Achieved: The correlation coefficient of -0.40 indicates a moderate negative correlation.
- Standard Deviation: Age (9.08), Maximum Heart Rate Achieved (22.91).
- Outliers: No outliers detected in Age, one outlier detected in Maximum Heart Rate Achieved.



2.6 Stacked Bar and Pie Chart of Chest Pain Types by Sex

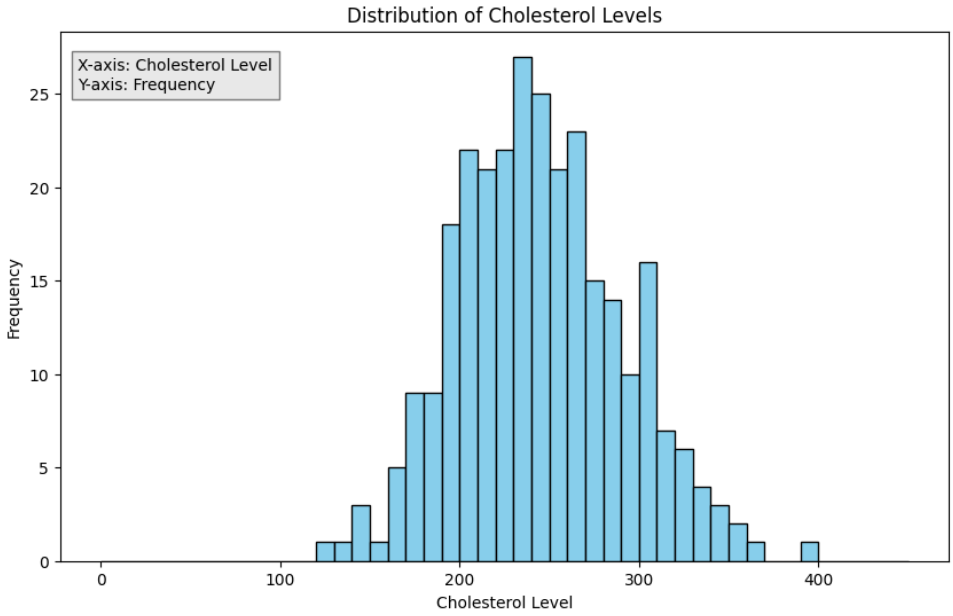
- Stacked Bar Chart of Chest Pain Types by Sex:
  - Type 0 (Typical Angina): Higher prevalence among males.
  - Type 1 (Atypical Angina): Similar distribution between males and females.
  - Type 2 (Non-anginal Pain): Slightly higher prevalence among females.
  - Type 3 (Asymptomatic): Higher prevalence among males.
- Proportion of Chest Pain Types (Pie Chart):
  - Type 0 (Typical Angina): 47.2%
  - Type 1 (Atypical Angina): 28.7%
  - Type 2 (Non-anginal Pain): 16.5%
  - Type 3 (Asymptomatic): 7.6%





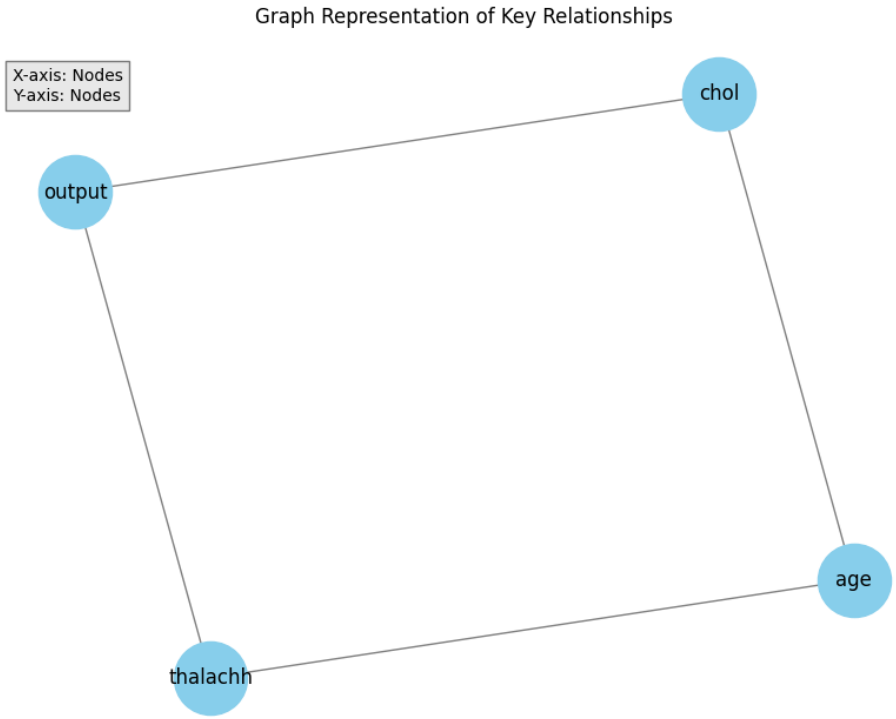
2.7 Cholesterol Levels Histogram

- Peak Frequencies: Highest frequency observed in the ranges 230-240 (28 occurrences) and 240-250 (26 occurrences).
- Overall Distribution: Clustering of cholesterol levels between 170 and 310.
- Low and High Extremes: Few patients with cholesterol levels in the extreme ranges (below 120 and above 370).
- Health Implications: Clustering of cholesterol levels in the 170-310 range suggests most patients are within the range where medical intervention might be recommended.



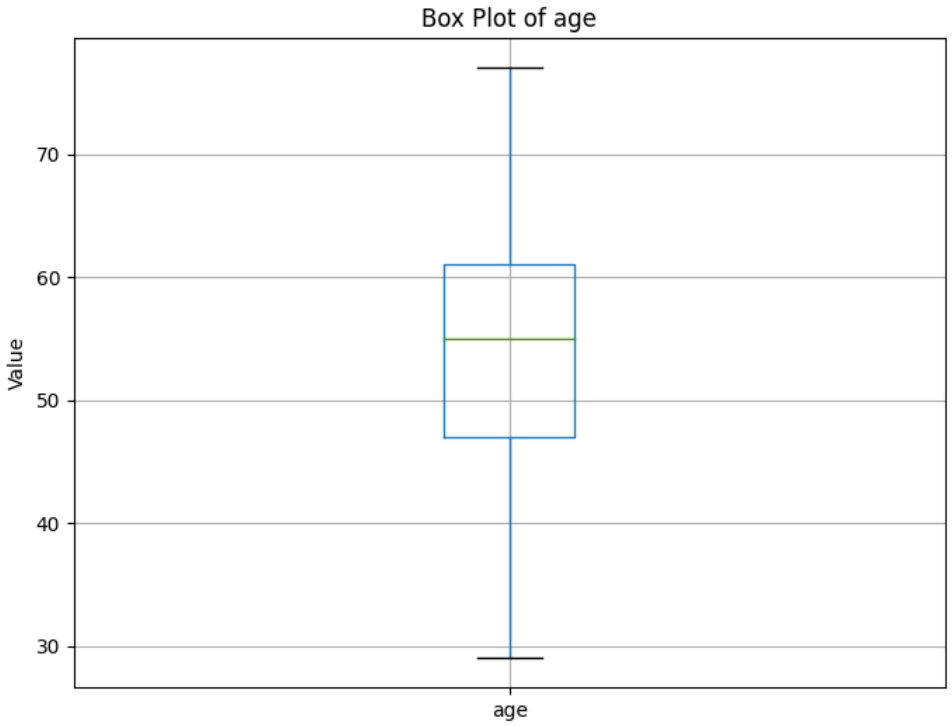
2.8 Graph Representation of Key Relationships

- Key Relationships Identified:
  - Age and Maximum Heart Rate Achieved (thalachh)
  - Age and Cholesterol Level (chol)
  - Maximum Heart Rate Achieved (thalachh) and Heart Attack Risk (output)
  - Cholesterol Level (chol) and Heart Attack Risk (output)
- Strength of Relationships: The presence of edges indicates a correlation or relationship between the connected nodes.
- Analysis Considerations: These relationships are valuable for understanding key risk factors for heart attacks and can guide further analysis or model development.

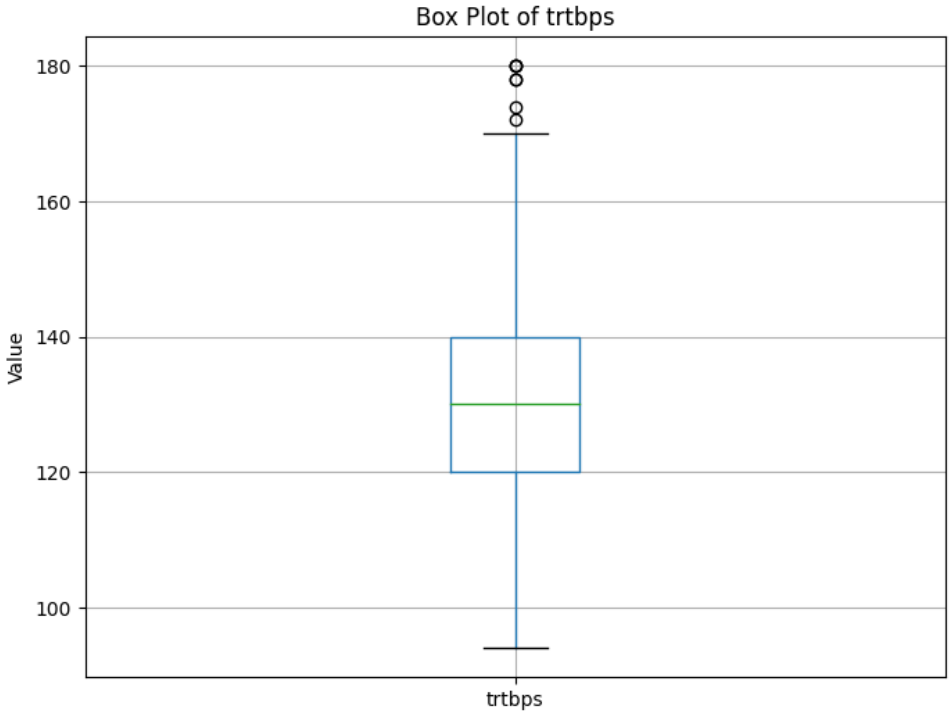


2.9 Box Plot of Age, Trtbps, Chol, Thalachh

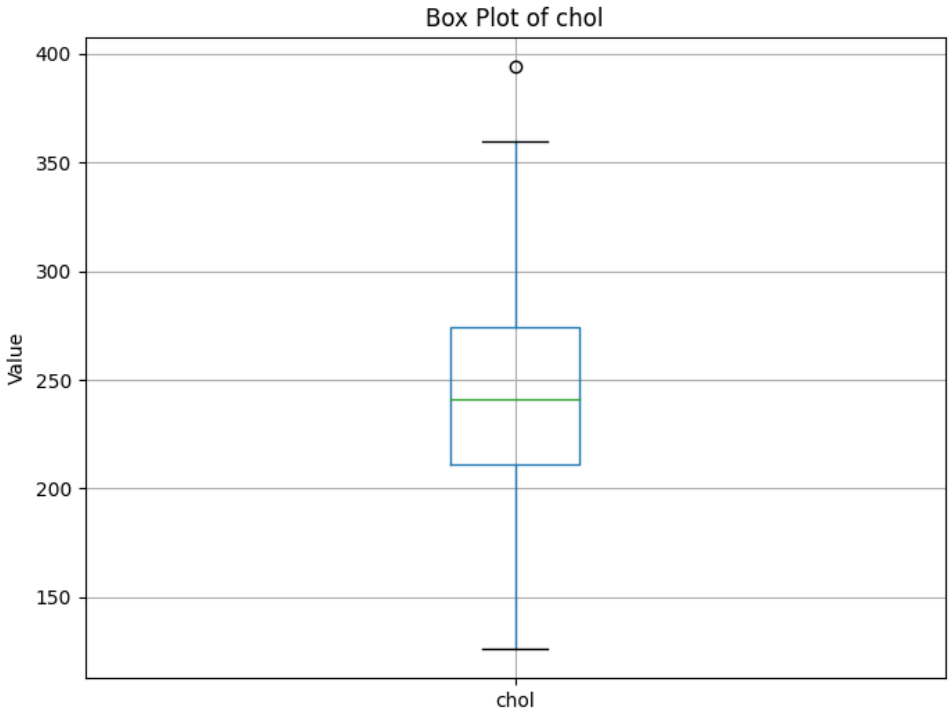
- **Age**: The distribution appears relatively balanced around the median without significant outliers, indicating a more uniform age distribution.



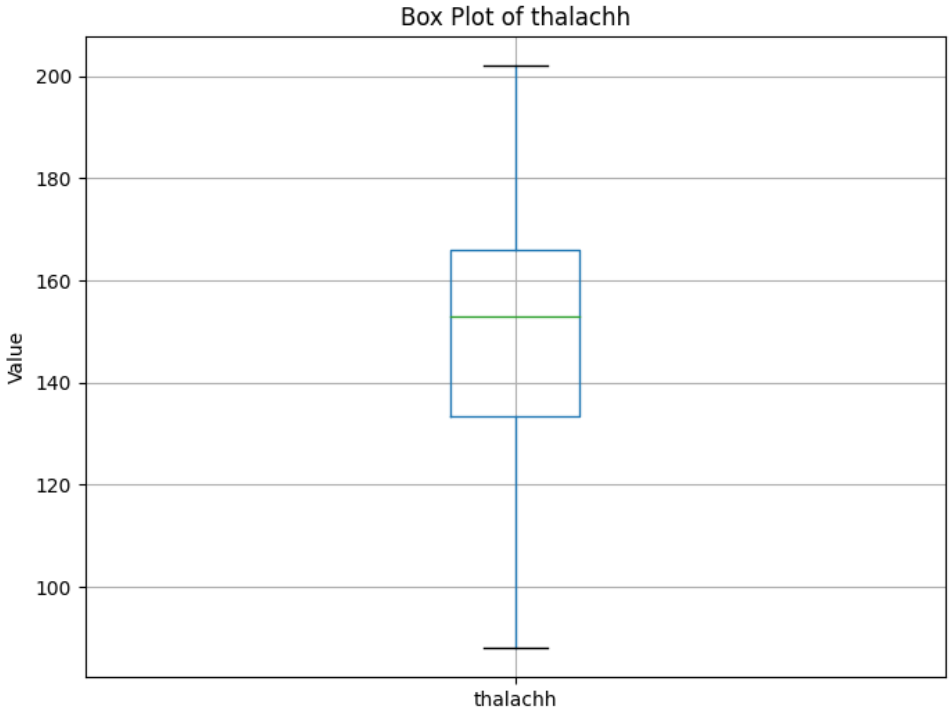
- **Resting Blood Pressure**: There are multiple outliers observed above the upper bound, suggesting some instances of unusually high resting blood pressure readings.



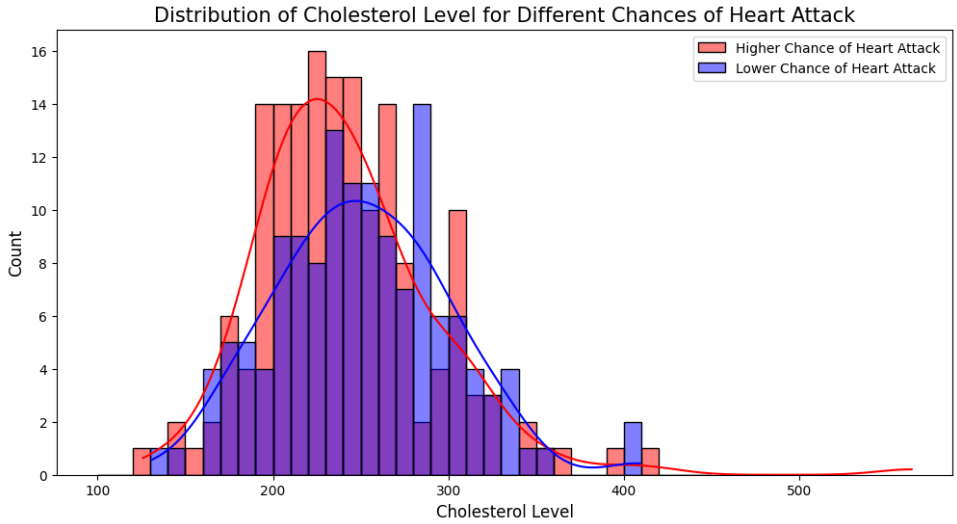
- **Cholesterol**: Several outliers are noted at higher cholesterol levels, which may indicate some individuals with significantly elevated cholesterol compared to the majority.



- **Maximum Heart Rate Achieved**: There is one outlier noted at a lower maximum heart rate, potentially indicating a distinct subset of individuals with lower heart rates compared to the median.



2.10 Distribution of Cholesterol Level for Different Chances of Heart Attack



Higher Chance vs. Lower Chance of Heart Attack:

- **Cholesterol Levels 120-260**: There seems to be a higher count of individuals with a higher chance of heart attack compared to those with a lower chance in these intervals.
- **Cholesterol Levels 280-300**: The count shifts, showing more individuals with a lower chance of heart attack compared to those with a higher chance.
- **Cholesterol Levels 300-320**: There is again a higher count of individuals with a higher chance of heart attack compared to those with a lower chance.

Specific Intervals:

- **Cholesterol Levels 180-200**: Shows a relatively high count of individuals with a higher chance of heart attack compared to those with a lower chance.
- **Cholesterol Levels 220-240**: Also shows a noticeable difference between higher and lower chances, with more individuals in the higher chance category.

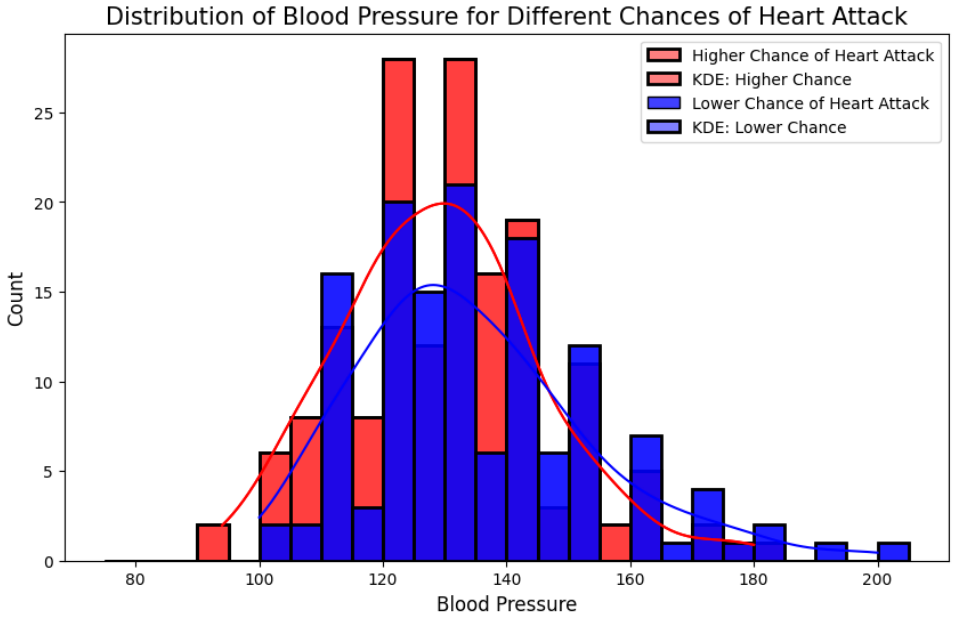
Outlier Detection:

The presence of outliers, especially at the extreme ends of the cholesterol spectrum (e.g., levels above 400), indicates cases that might warrant special attention due to their potential impact on heart health.

Distribution and Risk Assessment:

These insights suggest that cholesterol levels within certain ranges might correlate with a higher likelihood of heart attack. This information could be valuable for risk assessment and preventive measures in clinical settings.

2.11 Distribution of Blood Pressure for Different Chances of Heart Attack



Higher Chance of Heart Attack vs. Lower Chance:

Intervals where the count of individuals with a higher chance of heart attack is noticeably higher than those with a lower chance can indicate potential risk factors associated with higher blood pressure.

Distribution Patterns:

The distribution shows a varying pattern across different intervals. For instance, intervals around 120-125 and 130-135 show a significant number of individuals with both higher and lower chances of heart attack, indicating that blood pressure alone may not be a definitive indicator without considering other factors.

Risk Assessment:

Higher counts in certain intervals, such as 120-125 and 130-135, suggest these ranges might be critical for assessing cardiovascular risk. This aligns with medical understanding that blood pressure within specific ranges can impact heart health differently for different individuals.

Potential Outliers:

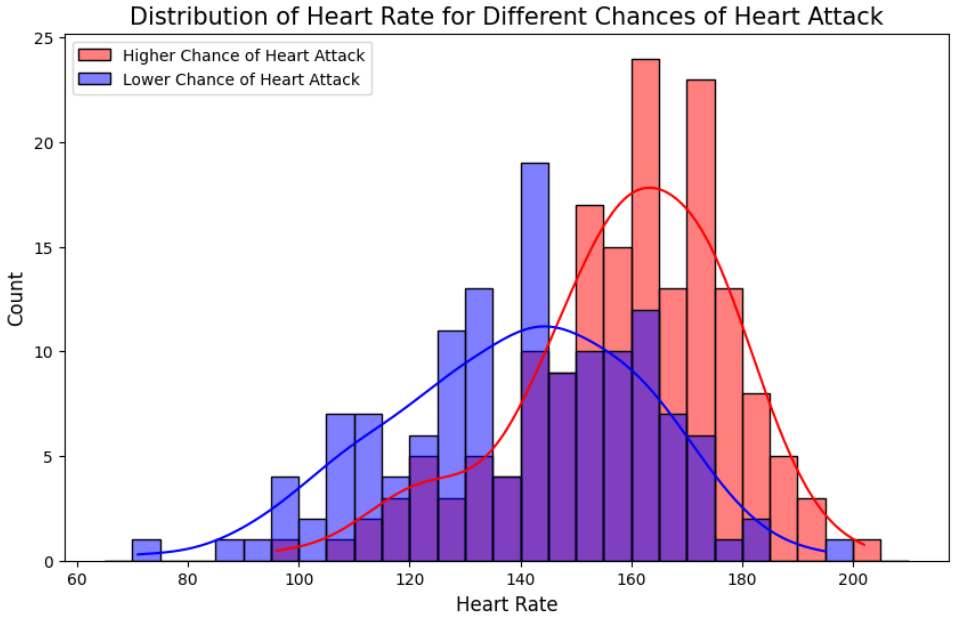
Although not explicitly noted here, outliers (such as extremely high or low values) in blood pressure could be crucial in understanding extreme risk cases that deviate significantly from typical patterns observed in the dataset.

Clinical Considerations:

Clinically, these insights can guide healthcare professionals in identifying potential risk factors associated with blood pressure levels and their correlation with the likelihood of a

heart attack. This understanding can lead to better preventive measures and personalized healthcare interventions.

2.12 Distribution of Heart Rate for Different Chances of Heart Attack



Lower Heart Rates (80-100 bpm):

- Individuals with lower heart rates (80-100 bpm) generally have a mix of lower and higher chances of heart attack.
- Lower chance intervals (e.g., 85-90 bpm, 90-95 bpm) tend to have fewer occurrences compared to higher chance intervals (e.g., 95-100 bpm).

Moderate Heart Rates (100-130 bpm):

Heart rates in the range of 100-130 bpm show a broader distribution. There's a notable shift towards higher chances of heart attack as heart rate increases within this range (e.g., 110-115 bpm, 120-125 bpm).

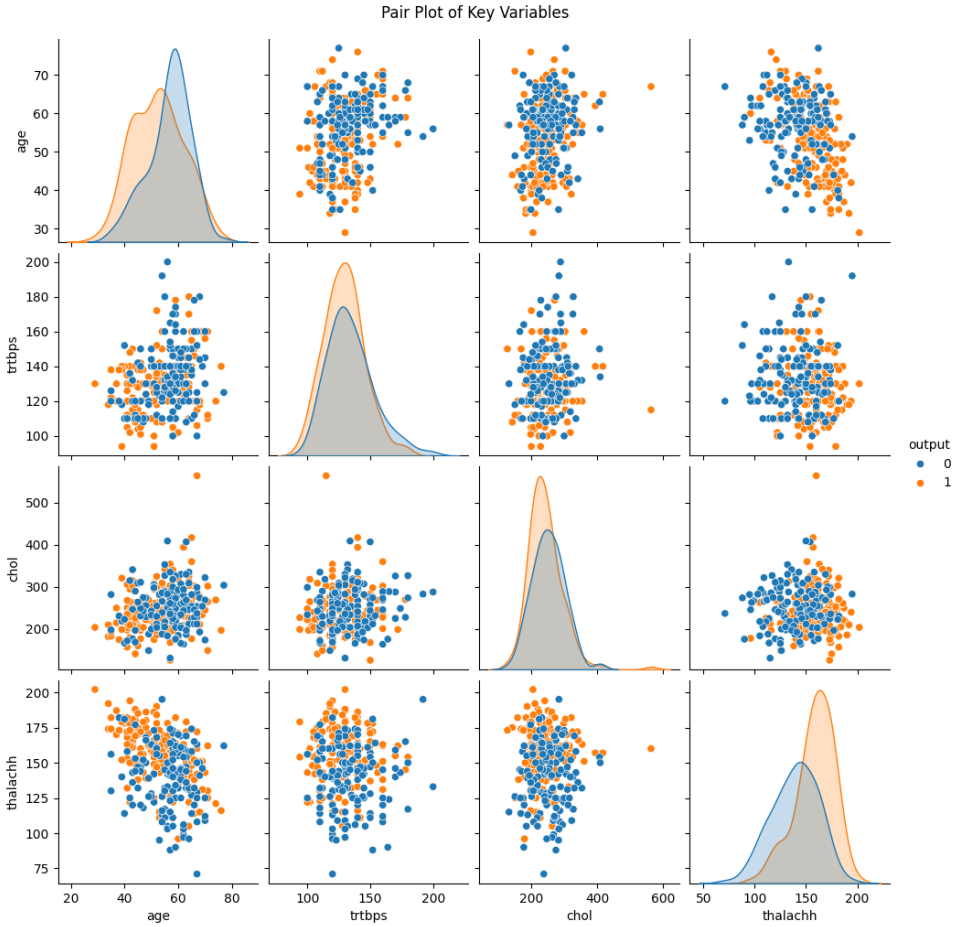
Higher Heart Rates (130-180 bpm):

Heart rates above 130 bpm generally show a higher incidence of heart attacks. Intervals like 130-135 bpm, 140-145 bpm, and 160-165 bpm consistently show higher chances of heart attack compared to lower chance intervals.

Very High Heart Rates (180-205 bpm):

Very high heart rates (180-205 bpm) are less common but still indicate a higher chance of heart attack, albeit with fewer data points.

2.13 Pair Plot



Age vs. Other Variables:

- **Age vs. Resting Blood Pressure (trtbps)**: There appears to be a slight positive correlation, suggesting that older individuals tend to have slightly higher resting blood pressure.
- **Age vs. Cholesterol Level (chol)**: There doesn't seem to be a strong linear trend, indicating that age alone may not significantly influence cholesterol levels.
- **Age vs. Maximum Heart Rate Achieved (thalachh)**: There's a noticeable downward trend as age increases, suggesting that older individuals generally achieve lower maximum heart rates.

Resting Blood Pressure (trtbps) vs. Other Variables:

- **Resting Blood Pressure vs. Cholesterol Level (chol)**: There doesn't appear to be a strong linear relationship between these variables, indicating that they may be independent factors.
- **Resting Blood Pressure vs. Maximum Heart Rate Achieved (thalachh)**: Similarly, there doesn't seem to be a clear linear relationship, suggesting independence or non-linear associations.

Cholesterol Level (chol) vs. Maximum Heart Rate Achieved (thalachh):

There appears to be no strong linear correlation between cholesterol levels and maximum heart rates achieved. This suggests that these variables may not directly influence each other linearly in the dataset.

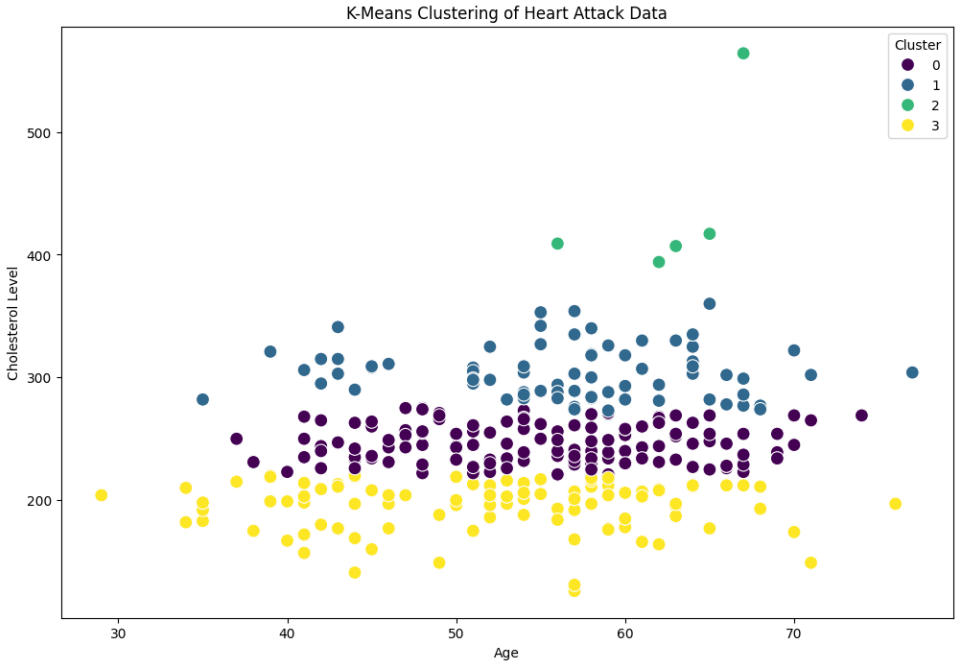
Effect of Output (Chance of Heart Attack):

The hue (output) shows how the distribution of each variable differs between individuals with a lower chance (0) and a higher chance (1) of heart attack. For instance, in the age vs. thalachh plot, there's a noticeable concentration of higher chances of heart attack (orange) in older individuals with lower maximum heart rates.

3. Pattern Identification

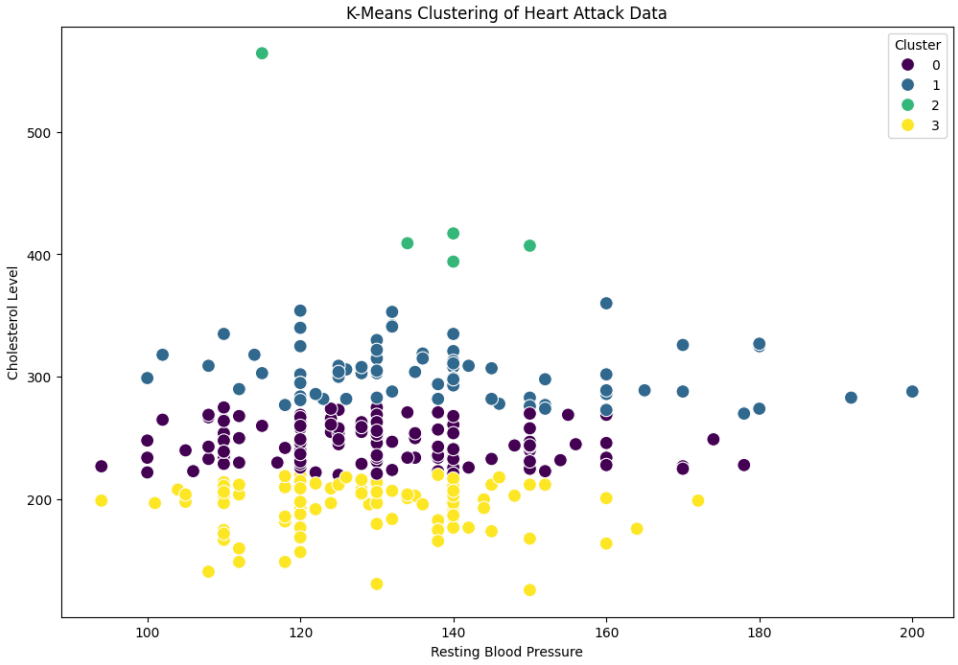
3.1 K-Means: Clustering Algorithm (2D)

(i) Age and Cholesterol:



Older individuals tend to cluster together with higher cholesterol levels, suggesting a potential correlation between age and cholesterol. Younger individuals are more spread out across the clusters, indicating diverse cholesterol levels.

(ii) Blood Pressure and Cholesterol:



Higher blood pressure readings tend to correlate with higher cholesterol levels in some clusters, particularly Cluster 1 (blue). Lower blood pressure readings are associated with varying cholesterol levels across different clusters.

Cluster Interpretation:

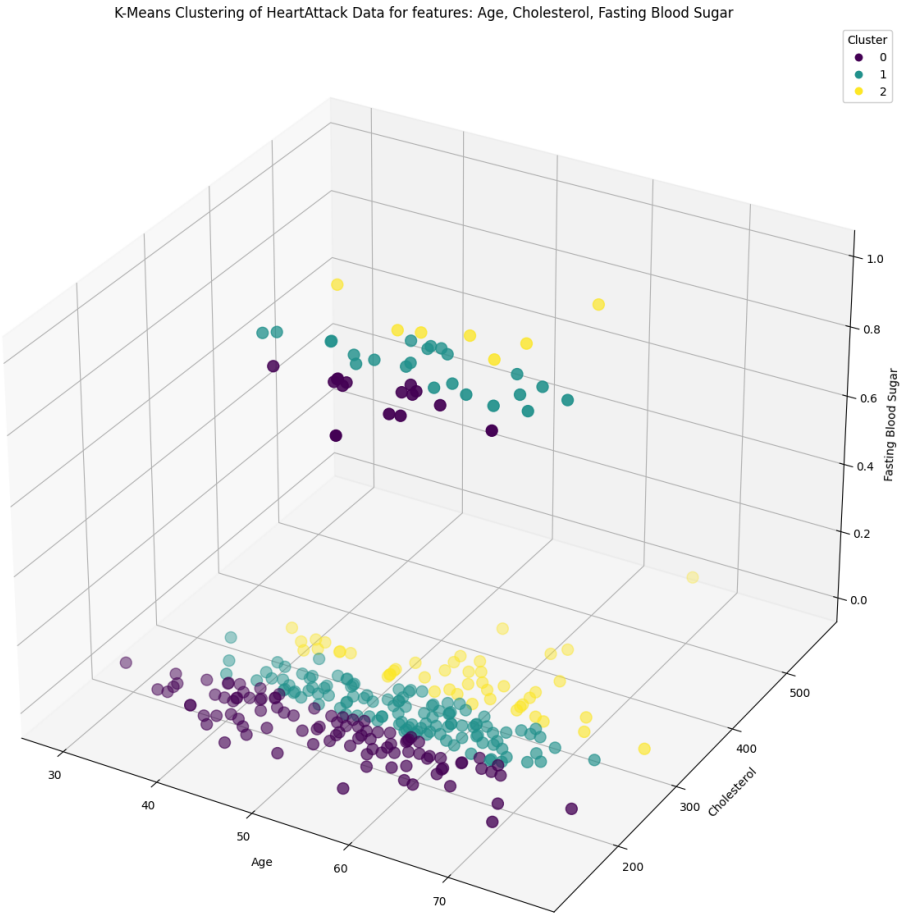
Cluster 1 (blue) may represent individuals at higher risk due to higher cholesterol levels and moderate blood pressure. Cluster 3 (yellow) suggests a younger group with higher blood pressure readings, indicating potential risk factors in younger demographics.

Practical Implications:

These clusters can help identify groups at higher risk of heart issues based on age, blood pressure, and cholesterol levels. Healthcare interventions and preventive measures could be targeted based on these identified clusters to mitigate potential risks associated with heart attacks.

3.2 K-Means: Clustering Algorithm (3D)

(i) Insights from the Sorted DataFrame



Based on the sorted DataFrame for features ['age', 'chol', 'fbs'], we can draw the following insights:

1. **Age and Cholesterol Relationship:**
    - Older Age and Lower Cholesterol:
        - The oldest individual in the dataset is 77 years old with a cholesterol level of 304.
        - Individuals like those aged 71 have a cholesterol level of 149, which is relatively low for their age.
    - Middle-Aged Individuals:
        - A significant number of middle-aged individuals (ages 41 to 59) exhibit varied cholesterol levels ranging from 177 to 327.
2. **Fasting Blood Sugar (fbs):**
    - The fbs value is 0 across all sorted entries, indicating that fasting blood sugar levels above 120 mg/dl are not prevalent in this subset.
3. **K-Means Clustering:**
    - **Cluster 0 (Blue):**
        - Appears to have a mix of ages with cholesterol levels ranging from 149 to 214.
        - Individuals in this cluster have varied conditions like restecg and thalachh, but all have fbs as 0.
    - **Cluster 2 (Green):**
        - Consists of individuals with very high cholesterol levels (295 to 327).
        - This cluster also includes individuals with severe heart conditions (ca=3 and thall=3).
4. **Heart Attack Risk (output):**
    - Individuals with higher cholesterol levels (above 295) mostly belong to cluster 2 and show varied output values (indicating both higher and lower chances of heart attack).
    - Middle-aged individuals with moderate cholesterol levels (ranging from 204 to 214) and varying restecg values mostly belong to cluster 0 and have output values indicating higher chances of heart attack.

Detailed Analysis:

- **Cluster Analysis:**
    - **Cluster 0:**
        - Predominantly middle-aged individuals (41 to 56) with moderate cholesterol levels.
        - These individuals generally show a higher chance of heart attack (output=1).
        - They have various restecg values, indicating different ECG results.

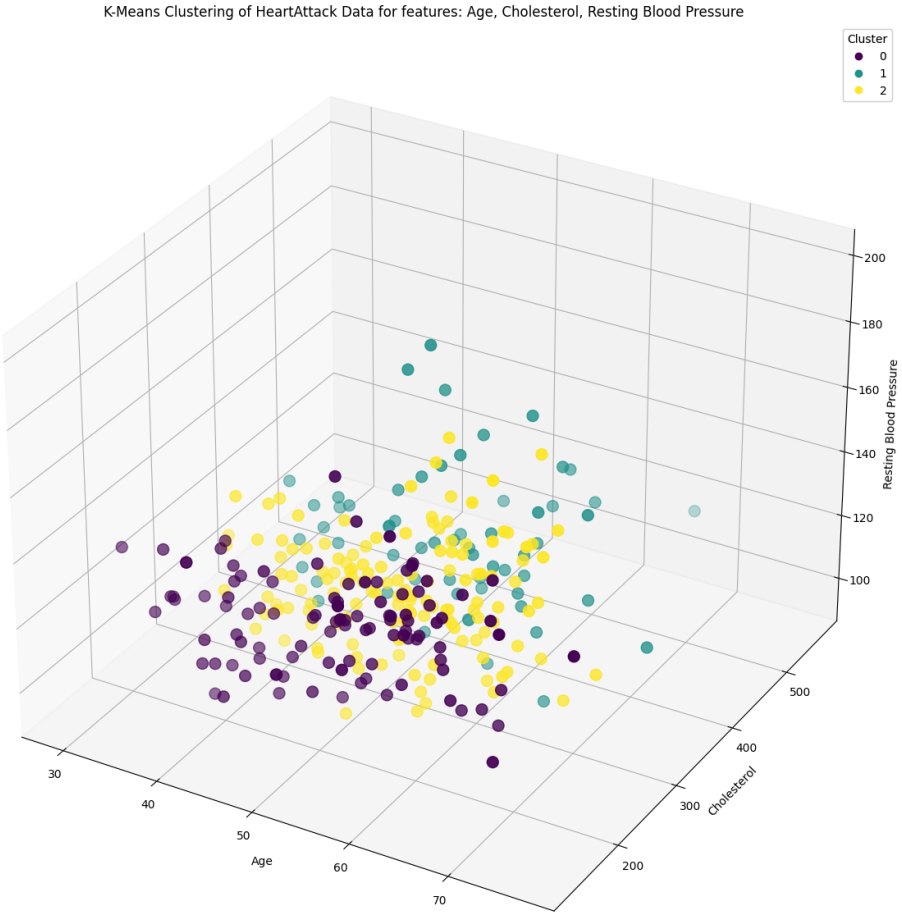
  - **Cluster 2:**
    * Contains individuals with very high cholesterol levels, including both younger and older individuals.
    * This cluster shows a mix of heart attack risks, but many are in higher risk categories.

- **Age and Risk Factors:**
  - Older individuals with lower cholesterol levels may still be at risk due to other factors like thalachh and trtbps.
  - Younger individuals with very high cholesterol levels are at significant risk of heart issues, as indicated by the clustering results.

- **Health Interventions:**
  - Targeted interventions can be designed for clusters. For instance, cluster 2 individuals (with high cholesterol) might benefit from aggressive cholesterol management and regular monitoring.
  - Middle-aged individuals in cluster 0 with moderate cholesterol levels but higher heart attack risk (output=1) may need comprehensive cardiovascular assessments.

- **Preventive Measures:**
  - Preventive measures should focus on managing cholesterol levels across all age groups.
  - Regular screenings and lifestyle modifications are essential for individuals in higher risk clusters.

(ii) Insights from the Sorted DataFrame for Features: ['age', 'chol', 'trtbps']



Observations and Patterns:

1. **Age and Cholesterol:**
   - Older Age with Lower Cholesterol:
     - The oldest individual in the dataset is 71 years old with a cholesterol level of 149 and resting blood pressure of 112, showing low cholesterol and blood pressure.
     - Individuals around the age of 64 and 66 have cholesterol levels ranging from 227 to 278 with varying blood pressure levels (170 and 146 respectively).
   - **Middle Age and Cholesterol:**
     - Individuals aged between 52 and 59 show cholesterol levels from 196 to 239, with varying blood pressure levels.
     - Cholesterol levels for individuals in their 50s are relatively higher (196 to 239) compared to those in their 40s.
   - **Young Adults:**

- The youngest individual in this subset is 34 years old with a cholesterol level of 210 and resting blood pressure of 118.

2. **Resting Blood Pressure (trtbps):**

   - **Normal Blood Pressure Range:**
     - Individuals with normal or slightly lower resting blood pressure (e.g., 112, 118) have cholesterol levels ranging from 149 to 210.
   - **Elevated Blood Pressure:**
     - Higher cholesterol levels (e.g., 227, 278) are associated with higher resting blood pressure (e.g., 170, 146).

3. **K-Means Clustering:**

   - **Cluster 0:**
     - Contains individuals with moderate cholesterol levels (149 to 210) and resting blood pressure levels (112 to 136).
     - This cluster consists of both younger (34 to 56 years) and middle-aged individuals (52 to 56 years).
   - **Cluster 2:**
     - Contains individuals with higher cholesterol levels (227 to 278) and resting blood pressure levels (130 to 170).
     - This cluster includes both older (57 to 66 years) and middle-aged individuals (47 to 59 years).

4. **Heart Attack Risk (output):**

   - **Higher Risk:**
     - Individuals in Cluster 2 with higher cholesterol and blood pressure levels show a higher risk of heart attack (output=1).
     - Middle-aged individuals (47 to 64 years) in Cluster 0 also show a significant risk of heart attack.
   - **Lower Risk:**
     - The youngest individual (34 years) in the dataset shows a lower cholesterol level (210) and normal resting blood pressure (118), indicating a lower risk of heart attack (output=0).

Detailed Analysis:

- **Cluster Analysis:**
  - **Cluster 0:**
    - Predominantly includes individuals with moderate cholesterol levels (149 to 210) and normal to slightly elevated resting blood pressure levels (112 to 136).
    - This cluster consists of a mix of younger (34 to 56 years) and middle-aged individuals (52 to 56 years).

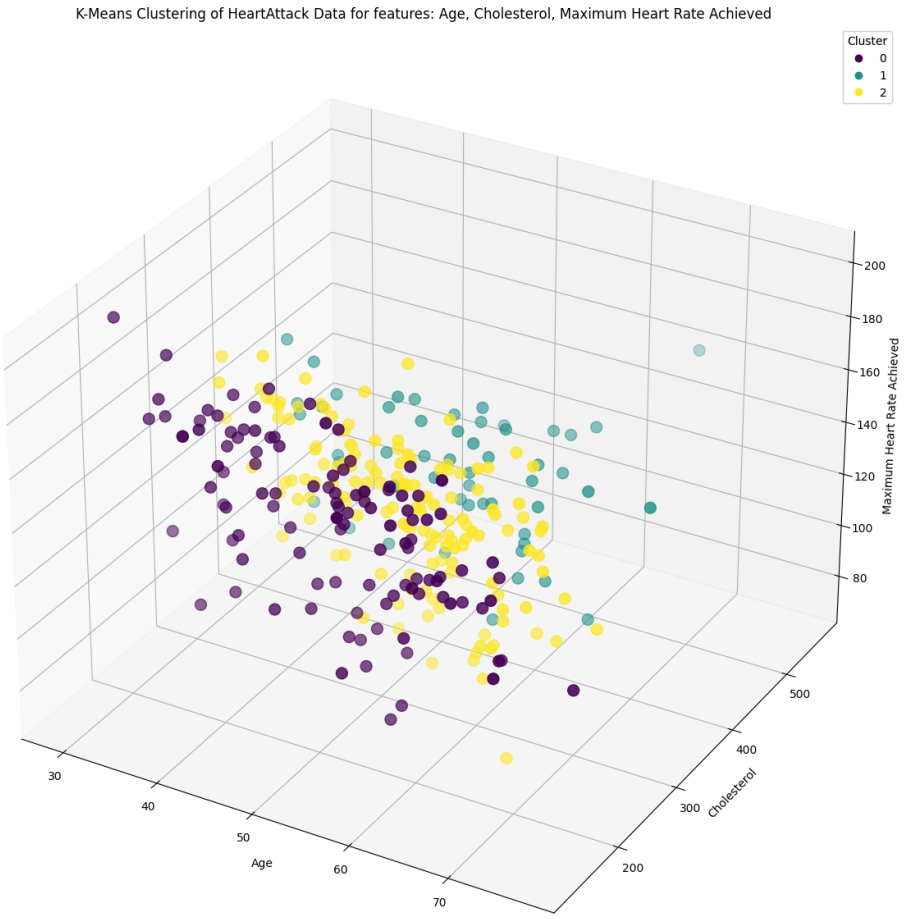
  - **Cluster 2:**
    * Contains individuals with higher cholesterol levels (227 to 278) and elevated resting blood pressure levels (130 to 170).
    * This cluster includes both older (57 to 66 years) and middle-aged individuals (47 to 59 years).

- **Age and Risk Factors:**
  - Older individuals (57 to 66 years) with higher cholesterol levels (227 to 278) and elevated blood pressure are at significant risk of heart issues, as indicated by Cluster 2.
  - Middle-aged individuals (47 to 64 years) with moderate cholesterol levels (149 to 239) also show a higher risk of heart attack, especially those in Cluster 0.

- **Health Interventions:**
  - Targeted interventions should focus on managing cholesterol and blood pressure levels across all age groups, particularly for individuals in Cluster 2.
  - Regular cardiovascular assessments and lifestyle modifications are crucial for individuals in Cluster 0 with moderate cholesterol and varying blood pressure levels.

- **Preventive Measures:**
  - Preventive measures should emphasize cholesterol management and regular blood pressure monitoring.
  - Lifestyle changes, including diet and exercise, are essential for reducing heart attack risks in both middle-aged and older individuals identified in Clusters 0 and 2.

(iii) Insights from the Sorted DataFrame for Features: ['age', 'chol', 'thalachh']



Observations and Patterns:

1. **Age, Cholesterol, and Maximum Heart Rate:**
   - **Older Age with Varying Cholesterol and Heart Rates:**
     - The oldest individual in the dataset is 71 years old with a cholesterol level of 149 and a maximum heart rate (thalachh) of 125.
     - Another older individual, aged 68, has a higher cholesterol level of 211 but a lower maximum heart rate of 115.
   - **Younger Age and High Cholesterol with High Heart Rates:**
     - A younger individual aged 37 has a cholesterol level of 215 and a high maximum heart rate of 170.
     - Another individual aged 43 has a cholesterol level of 211 and a high maximum heart rate of 161.
2. **Middle Age Individuals:**
   - **Higher Cholesterol with High Heart Rates:**

- Individuals aged between 47 and 57 show a wide range of cholesterol levels (126 to 240) and high maximum heart rates (161 to 179).
- For instance, a 57-year-old individual has a cholesterol level of 126 and a high maximum heart rate of 173.

3. **Resting Blood Pressure (trtbps):**

   - The resting blood pressure values do not show a consistent pattern across different ages, cholesterol levels, and maximum heart rates.

4. **K-Means Clustering:**

   - **Cluster 0:**
     - Contains individuals with moderate cholesterol levels (149 to 215) and varying maximum heart rates (115 to 170).
     - This cluster includes younger (37 to 43 years) and middle-aged individuals (57 to 68 years).
   - **Cluster 2:**
     - Contains individuals with higher cholesterol levels (236 to 278) and higher maximum heart rates (152 to 179).
     - This cluster includes middle-aged individuals (47 to 57 years).

5. **Heart Attack Risk (output):**

   - **Higher Risk:**
     - Individuals in Cluster 2 with higher cholesterol levels and higher maximum heart rates show a higher risk of heart attack (output=1).
     - Middle-aged individuals (47 to 57 years) in this cluster have high cholesterol and heart rates, indicating significant heart attack risks.
   - **Lower Risk:**
     - Younger individuals (37 to 43 years) in Cluster 0 with moderate cholesterol and varying heart rates show varied risks but tend to have higher heart attack risks (output=1).

Detailed Analysis:

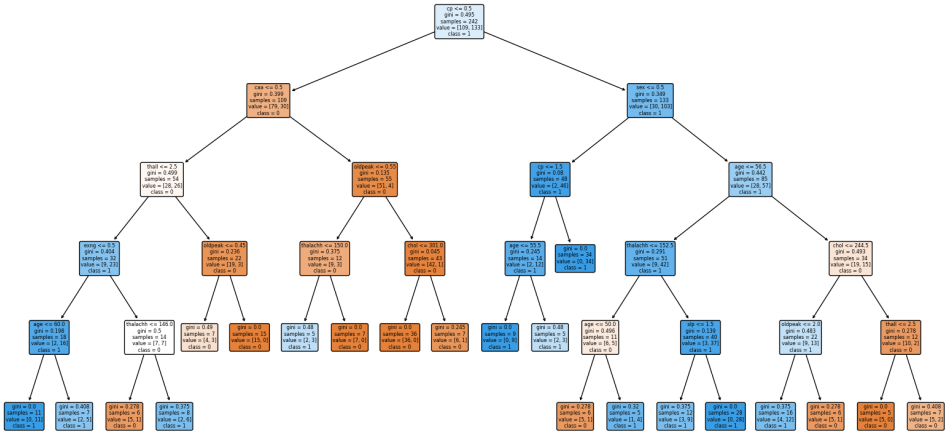
- **Cluster Analysis:**
  - **Cluster 0 (Moderate Cholesterol and Varying Heart Rates):**
    - Individuals are generally younger or middle-aged with moderate cholesterol levels and varying maximum heart rates.
    - This cluster has varied heart attack risks but shows a tendency for higher risks among younger individuals.
  - **Cluster 2 (High Cholesterol and High Heart Rates):**
    - Individuals are generally middle-aged with higher cholesterol levels and higher maximum heart rates.

    - This cluster shows a higher risk of heart attack, indicating the need for targeted interventions.

- **Health Interventions:**
  - **Cluster 0:**
    - Preventive measures can focus on maintaining cholesterol and managing heart rates, especially for younger individuals.
    - Regular monitoring and lifestyle modifications are essential to mitigate heart attack risks.
  - **Cluster 2:**
    - Aggressive management of cholesterol and heart rates is crucial for individuals in this cluster.
    - Comprehensive cardiovascular assessments and interventions are necessary to reduce heart attack risks.

- **Preventive Measures:**
  - **Diet and Exercise:**
    - Encouraging a healthy diet and regular exercise can help maintain cholesterol and heart rates within normal ranges.
  - **Regular Check-ups:**
    - Regular health check-ups and screenings are vital, especially for individuals in higher risk clusters.
  - **Medication Management:**
    - Appropriate medication management for cholesterol and heart rates can significantly reduce heart attack risks.

3.3 Decision Tree Classification Algorithm



Cross-Validation Scores:

- The model's cross-validation accuracy scores vary from 0.70 to 0.82, with a mean CV accuracy of 0.7621. This indicates that the model generalizes reasonably well across different subsets of the data.

Training Accuracy:

- The training accuracy is 0.8845, which is significantly higher than the mean CV accuracy. This suggests that the model fits well to the training data but might be slightly overfitting.

Test Accuracy:

- The test accuracy is 0.8361, which is closer to the cross-validation scores. This indicates good performance on unseen data and suggests the model is reasonably well-calibrated.

Precision, Recall, and F1-Score:

- Both precision and recall for classes 0 and 1 are high, around 0.83-0.84. The F1-scores are also consistent, indicating balanced performance across both classes.
- The weighted averages for precision, recall, and F1-score all align with the overall accuracy, reinforcing that the model performs consistently across classes.

Feature Importance:

- **Most Important Features:**
  - **Chest Pain Type (cp):** The most significant feature with an importance score of 0.3600. This implies that the type of chest pain is a critical factor in predicting heart attack risk.
  - **Age (age):** The second most important feature with an importance score of 0.1175. Age is a well-known risk factor for heart disease.

  - **Thalassemia (thall) and Number of Major Vessels (caa):** Both features have similar importance scores (around 0.11), indicating that they significantly contribute to the model's predictions.

- **Moderately Important Features:**
  - **Maximum Heart Rate Achieved (thalachh):** This feature has an importance score of 0.0982, suggesting that the maximum heart rate during exercise is an important predictor.
  - **ST Depression Induced by Exercise (oldpeak):** With an importance score of 0.0677, this feature also contributes significantly to the model.

- **Least Important Features:**
  - **Sex (sex):** Although important with a score of 0.0604, it is less influential compared to the top features.
  - **Cholesterol (chol), Exercise Induced Angina (exng), and Slope of the Peak Exercise ST Segment (slp):** These features have lower importance scores (ranging from 0.0125 to 0.0361).

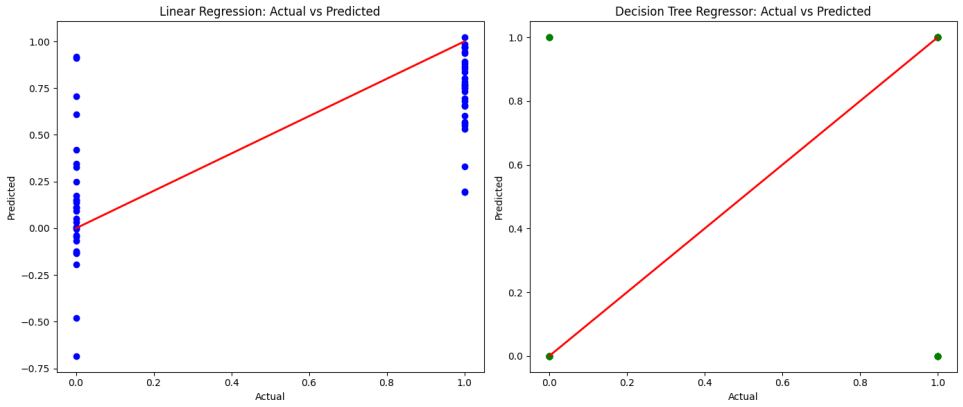
- **Zero Importance Features:**
  - **Resting Blood Pressure (trtbps), Fasting Blood Sugar (fbs), and Resting Electrocardiographic Results (restecg):** These features have zero importance, indicating that they do not contribute to the model's predictive power.

3.4 Logistic Regression Model

Model Performance Metrics:

- **Accuracy:** 0.8681
  - This indicates that the model correctly predicts 86.81
- **Precision:**
  - **Class 0 (Less chance of heart attack):** 0.90
  - **Class 1 (More chance of heart attack):** 0.84
  - Precision measures the proportion of true positives (correctly predicted cases) among all positive predictions. For class 0, it correctly identifies 90
- **Recall (Sensitivity):**
  - **Class 0:** 0.82
  - **Class 1:** 0.91
  - Recall measures the proportion of true positives that were correctly identified out of all actual positives. For class 0, 82
- **F1 Score:**
  - **Weighted Average:** 0.8776
  - The F1 score is the harmonic mean of precision and recall, providing a single metric to evaluate a model's performance. A higher F1 score indicates better overall performance in terms of both precision and recall.
- **ROC-AUC Score:** 0.8665
  - The ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) score is a measure of the model's ability to distinguish between classes. An AUC score close to 1 suggests a good separation between classes, with higher values indicating better performance.
- **Overall Model Performance:**
  - The logistic regression model achieves a high accuracy and balanced precision and recall scores for both classes, indicating robust predictive capability across the dataset.
- **Class-Specific Insights:**
  - Class 1 (more chance of heart attack) shows slightly higher recall than precision, indicating that the model is better at correctly identifying individuals at risk of a heart attack compared to those who are not at risk.
  - Class 0 (less chance of heart attack) shows higher precision, meaning that when the model predicts someone as having less chance of a heart attack, it is usually correct.

3.5 Regression Algorithms: Linear Regression and Decision Tree Regressor



Performance Insights:

- **Linear Regression**
  - **Mean Squared Error (MSE):** 0.1248
    - MSE measures the average squared difference between the actual and predicted values. A lower MSE indicates better model performance. Here, the MSE of 0.1248 suggests that the model's predictions are reasonably close to the actual values.
  - **R-squared (R^2):** 0.4994
    - R^2 represents the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 value of 0.4994 indicates that approximately 49.94
- **Decision Tree Regressor**
  - **Mean Squared Error (MSE):** 0.2131
    - The MSE for the Decision Tree Regressor is higher at 0.2131, indicating that its predictions are less accurate compared to the Linear Regression model.
  - **R-squared (R^2):** 0.1455
    - The R^2 value of 0.1455 indicates that only 14.55

Insights and Comparisons:

- **Prediction Accuracy:**
  - The Linear Regression model has a lower MSE (0.1248) compared to the Decision Tree Regressor (0.2131). This means that Linear Regression provides more accurate predictions in terms of mean squared error.
- **Explanatory Power:**

- The R^2 value for Linear Regression (0.4994) is significantly higher than that of the Decision Tree Regressor (0.1455). This indicates that Linear Regression explains a larger portion of the variance in the target variable compared to the Decision Tree Regressor.

- **Model Complexity:**
  - Linear Regression is a simple and interpretable model, which often works well for linear relationships between the features and the target variable. The lower MSE and higher R^2 suggest that in this case, a linear relationship might be more appropriate for your data.
  - Decision Trees can capture non-linear relationships and interactions between features. However, in this case, it appears that the Decision Tree model is overfitting or not capturing the underlying pattern as effectively as the Linear Regression model.

- **Choosing the Right Model:**
  - Based on the MSE and R^2 values, Linear Regression seems to be the better model for this specific dataset and problem. It provides more accurate predictions and explains a greater portion of the variance in the target variable.
  - However, if the data has complex non-linear relationships, it might be worth exploring other non-linear models or tuning the hyperparameters of the Decision Tree Regressor to improve its performance.

Conclusion:

- The Linear Regression model outperforms the Decision Tree Regressor in terms of both accuracy (lower MSE) and explanatory power (higher R^2) for this dataset. It is likely a more suitable model for predicting the target variable given the provided metrics. However, it's important to consider the nature of the data and potentially explore other models or techniques to ensure the best possible predictive performance.

FUTURE WORK:

- Requires the Linear Regression and Decision tree regressor to be more accurate
- Requires a massive Dataset for better accuracy of machine learning algorithum