HEART ATTACK DATASET ANALYSIS

DATASET SUMMARY

The Heart Attack dataset contains 303 patient records and includes 14 clinical variables associated with cardiovascular health and the likelihood of heart disease. The dataset features both numerical variables (e.g., age, cholesterol, resting blood pressure, maximum heart rate) and categorical clinical indicators (such as chest pain type, ECG results, thalassemia, exercise-induced angina, etc).

The target variable (target) indicates the presence (1) or absence (0) of heart disease, with a slight class imbalance: 165 positive cases and 138 negative cases. No severe missing-value issues are observed, making the dataset suitable for direct EDA and predictive modeling.

Key variables show medically meaningful patterns:

- Age, cholesterol, oldpeak, and resting blood pressure show natural variability useful for risk profiling.
- Chest pain type, thalassemia, number of vessels shown via fluoroscopy, and exercise-induced angina demonstrate strong categorical separation between patients with and without heart disease.
- Important numerical predictors like thalach (maximum heart rate achieved) and oldpeak (ST depression) show clear distribution differences between classes.

VARIABLE BY VARIABLE INSIGHTS

1. age — Age of the patient

Type: Numerical

Insight:

Heart disease risk increases with age. Patients above **45–50 years** tend to show higher incidence of heart attack.

Why it matters:
Age is one of the strongest predictors of cardiovascular events.

2. sex — Gender (1 = male, 0 = female)

Type: Categorical
Insight:

- Traditionally, men have a higher risk of early-onset heart disease.
- EDA usually shows **more males** in the dataset.

Why it matters:
Gender influences cholesterol levels, blood pressure, and heart-disease patterns.

3. cp — Chest Pain Type (0–3)

Type: Categorical
Insight:
Chest pain types correlate strongly with heart disease:

- **0 = Typical angina** (most dangerous)
- **1 = Atypical angina**
- **2 = Non-anginal pain**
- **3 = Asymptomatic** (highly predictive of disease)

In most datasets:

- cp = 0 and cp = 3 tend to have **high target positivity**.

Why it matters:
One of the top categorical predictors.

4. trestbps — Resting Blood Pressure (mm Hg)

Type: Numerical
Insight:

- High resting BP (>130–140 mmHg) increases heart attack risk.

Why it matters:
Hypertension is a major cardiovascular risk factor.

5. chol — Serum Cholesterol (mg/dl)

Type: Numerical
Insight:

- Higher cholesterol means higher coronary risk.
- Often right-skewed with outliers.

Why it matters:
High cholesterol contributes to arterial blockage.

6. fbs — Fasting Blood Sugar (1 = >120 mg/dl)

Type: Binary Categorical
Insight:

- Elevated fasting blood sugar suggests possible diabetes, a major heart-disease risk factor.
- Many datasets have **imbalanced values** (mostly 0).
- Check how much target=1 differs by category.

Why it matters:
Diabetic patients tend to develop heart complications earlier.

7. restecg — Resting ECG Results

Type: Categorical
Meaning:

- **0** = Normal
- **1** = ST-T wave abnormality
- **2** = Left ventricular hypertrophy

Insight:

- Abnormal ECG (1 and 2) often show higher heart-disease frequency.
- Useful predictor but less impactful than cp or thalach.

8. thalach — Maximum Heart Rate Achieved

Type: Numerical
Insight:

- LOWER max heart rate is typically associated with higher disease risk.
- Patients with heart disease often reach **lower thalach values** during exercise tests.
- Strong negative correlation with target.

Why it matters:
Very strong numerical predictor of heart disease.

9. exang — Exercise-Induced Angina (1 = yes)

Type: Binary Categorical
Insight:

- If exercise triggers chest pain, risk is higher.
- Usually shows clear separation: exang = 1 correlates with target = 1.

10. oldpeak — ST Depression (exercise vs rest)

Type: Numerical
Insight:

- Measures heart stress response.
- Higher oldpeak = more serious heart abnormality.
- Often right-skewed.
- Strong positive correlation with disease.

Why it matters:
One of the best predictive features.

11. slope — Slope of the Peak Exercise ST Segment

Type: Categorical
Meaning:

- **0** = Upsloping
- **1** = Flat
- **2** = Downsloping

Insight:

- Flat or downsloping slopes often indicate heart disease.

12. ca — Number of Major Vessels Colored by Fluoroscopy (0–3)

Type: Numerical / Categorical
Insight:

- More blocked vessels → higher risk.
- Strong predictive power.

Why it matters:
Directly measures coronary blockage.

13. thal — Thalassemia (3 categories)

Common Coding:

- **1 = Fixed defect**
- **2 = Normal**
- **3 = Reversible defect**

Insight:

- Reversible defect (3) is strongly associated with heart disease.
- Usually one of the best categorical predictors.

14. target — Heart Disease (1 = yes, 0 = no)

Type: Binary
Insight:

- The variable being predicted.

SUMMARY OF THE MOST IMPORTANT PREDICTORS

Based on typical medical and statistical patterns:

Top Numerical Predictors

1. thalach (max heart rate)
2. oldpeak (ST depression)
3. age
4. trestbps
5. chol

Top Categorical Predictors

1. cp (chest pain type)
2. thal
3. ca
4. exang
5. sex

CONCLUSION

The exploratory analysis reveals strong patterns linking clinical features to heart disease presence:

1. Key Clinical Indicators Strongly Predict Heart Disease

Features such as chest pain type (cp), thalassemia (thal), number of affected vessels (ca), and exercise-induced angina (exang) demonstrate significant differences between patients with and

without heart disease. These categorical variables show clear class separations, making them valuable diagnostic indicators.

2. Numerical Variables Exhibit Meaningful Risk Trends

Variables including age, resting blood pressure, serum cholesterol, and ST depression (oldpeak**)** follow expected medical trends. Positive cases generally exhibit:

- Higher age
- Higher oldpeak
- Lower maximum heart rate achieved (thalach)

Such patterns align with known cardiovascular risk factors.

3. Correlation Analysis Highlights Critical Predictors

The correlation matrix indicates:

- **Oldpeak** and **thalach** have some of the strongest relationships with the target.
- Mild correlations exist among certain numeric variables (e.g., age and resting BP), but not enough to cause severe multicollinearity.
- Categorical features (cp, thal, ca) complement numerical trends and enhance overall predictive strength.

4. The Dataset Is Suitable for Predictive Modeling

The combination of clean distributions, minimal missing data, medically relevant variables, and acceptable class balance makes this dataset ideal for:

- Logistic regression
- Random forests
- XGBoost
- SVM
- Clinical risk-profiling models

5. Clinical & Data-Driven Insight Align Perfectly

The EDA confirms well-known medical knowledge: individuals with high blood pressure, abnormal ECGs, lower exercise capacity, abnormal ST depression, and specific chest pain types show significantly higher heart-disease risk. These insights reinforce the dataset's reliability and support its use in machine-learning–based health-risk prediction.