## univariate analysis

### continuous (numeric) variables

**Age**: the age of the samples. [years]

图表, 直方图

AI 生成的内容可能不正确。

According to the plot:

The age distribution is approximately bell-shaped, centered around 55 years, with most samples between 40 and 65 years old. The sample includes few young (<35) or elderly (>70) individuals. There are no major outliers or strong skewness, so age can be used directly as a continuous input variable. However,it should be standardized or scaling before use.  
Operations needed: standardized or scaling.

**RestingBP**: resting blood pressure [mm Hg]

图表, 直方图

AI 生成的内容可能不正确。

According to the plot:

The resting blood pressure (RestingBP) distribution is obviously right-skewed, concentrated between 110–150 mmHg with a median around 125 mmHg. The extremely low value (near 0, not reasonable) should be treated as outliers. Other than that, most values fall in the normal range.

If we don’t consider the outliers, the remaining distribution appears approximately normal.

So before using it, there are 2 operations that should be applied:

1. Remove outliers.
2. Standardization or scaling.

**Cholesterol**: serum cholesterol [mm/dl]

图表, 直方图

AI 生成的内容可能不正确。

The value distribution centered around 200 mg/dL, typical for adult populations. However, there are a large number of zero values, which are likely data entry errors or missing-value codes. After removing these zeros and capping extreme values (>400 mg/dL), the remaining data appear approximately normal. We did some research on the value, >400 md/dL is possible in real-world situations. However, log-transformation should be applied to reduce the impact on the model.

So, the operations need to be applied before using:

1. Handle the zero-value samples.
2. Log-transformation.
3. Standardization or scaling.

**MaxHR**: maximum heart rate achieved

图表, 直方图

AI 生成的内容可能不正确。

According to the figure:

The variable is approximately normally distributed, ranging from 60 to 200 and centered near 140 bpm. No significant outliers appear, and the data seem symmetric. Therefore MaxHR can be used directly as a continuous feature after standardization. In terms of medical area, lower MaxHR may reflect poorer heart performance, which might correlate with a higher likelihood of heart disease.

The operations need to be applied before using:

1. Standarlization.

**Oldpeak**: oldpeak = ST [Numeric value measured in depression], 0 – normal. High value means high risk.

图表, 直方图

AI 生成的内容可能不正确。

Through the Figure:

The variable is highly right-skewed, with most values near 0 and a few extreme cases above 4. A small number of negative values likely to be measurement errors. If only consider the values in range 0–4, the distribution remains slightly skewed but realistic. Because higher Oldpeak values indicate greater ST depression, this feature is expected to be positively correlated with heart-disease risk. Transformation or scaling should be applied.

The operations need to be applied before using:

1. Handling negative values
2. Log-transformation.
3. Standardization or scaling.

### Categorical

**Sex**: sex of the patient [M: Male, F: Female]

图表, 条形图

AI 生成的内容可能不正确。

Through the data and figure:

The variable has of two categories: male (M) and female (F). In the dataset, the major samples are male (725, 79%). Through our research, this gender imbalance is typical in heart-disease studies but should be noted, as it may cause a slight bias during model training if not handled carefully.

Operations before using: One-Hot Encoding

**FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

图表, 条形图

AI 生成的内容可能不正确。

Through the figure:

The FastingBS variable is binary, showing if a patient’s fasting blood sugar is higher than 120 mg/dL. Most patients (about 75%) fall into the normal range (value = 0), while a smaller portion (around 25%) have higher level (value = 1). There are no outliers since the feature is categorical. The imbalance between the two categories should be further noticed in bivariate analysis to assess its relationship with the target.

**ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

图表, 条形图

AI 生成的内容可能不正确。

According to the figure and numbers:

The variable has four categories: ASY (496), NAP (203), ATA (173), and TA (46). ASY cases dominate the dataset (≈55%), while typical angina cases are rare (≈5%). This imbalance suggests that many patients exhibit no classical chest pain symptoms despite having heart disease. The categorical imbalance should be noted, as it may influence how models interpret this feature.

**RestingECG**: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

图表, 条形图

AI 生成的内容可能不正确。

The variable contains three categories: Normal (552), LVH (188), and ST (178). Most patients (≈60%) have normal resting ECG results, while the other groups show LVH or ST-T wave abnormalities. The distribution is moderately imbalanced but reasonable. This feature should be one-hot encoded for use in modeling.

Operations before using: One-Hot Encoding

**ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]

图表, 条形图

AI 生成的内容可能不正确。

The variable is binary. There are 547 patients (60%) reporting N and 371 (40%) reporting Y. This moderate imbalance is reasonable. The presence of exercise-induced angina is expected to be positively correlated with heart-disease risk.

Operations before using: One-Hot Encoding

**ST\_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

图表, 条形图

AI 生成的内容可能不正确。

The variable has three categories: Flat (460), Up (395), and Down (63). The distribution is moderately imbalanced: “Flat” is the most common and “Down” is the least. Since a flat or downsloping ST segment often indicates abnormal cardiac response, this feature is important and will be one-hot encoded. The rare “Down” category (indicate serious situation) should be treated carefully depending on model choice.

Operations before using: One-Hot Encoding

## Feature vs Feature Correlation Analysis

The Analysis will be applied on numerical features only, and the result of how strongly different features are linearly related to each other. Heatmap will be used as tool in the analysis.

The abnormal in ‘Cholesterol’ and ‘RestingBP’ should be handled before the analysis to avoid mistakes.

The two Feature-Feature heatmaps (Pearson & Spearman) are as follows:

图片包含 正方形

AI 生成的内容可能不正确。

图片包含 图表

AI 生成的内容可能不正确。

文本

AI 生成的内容可能不正确。

Both Pearson and Spearman correlation analyses showed similar patterns, so the relationships among numerical features are consistent and linear. So, we only discussed the Pearson correlation matrix for simplicity.

The heatmap showed that the variables (Age, RestingBP, Cholesterol, MaxHR, and Oldpeak) demonstrated no strong correlations (|r| > 0.8). The highest correlation was a moderate negative relationship between Age and MaxHR, meaning that older patients tend to achieve lower maximum heart rates. Mild positive correlation existed between Age and Oldpeak, while other feature pairs showed weak or negligible relationships. The numerical summary of Pearson correlations also confirms the analysis. These findings suggest that the numerical features are independent and suitable to use in the modeling stage without multicollinearity concerns.