Heart Disease Prediction: Exploratory Data Analysis Report

Randeep Sidhu

May 28, 2025

Table of Contents

1. Introduction

This is an exploratory data analysis (EDA) of the UCI Heart Disease data to find patterns, evaluate data quality, and prepare for machine learning in predicting heart disease. The goal is to determine significant risk factors and create a clean, encoded dataset ready for supervised modeling, like logistic regression. The data set consists of 920 patient records with 16 attributes, both numerical attributes (e.g., age, cholesterol, blood pressure) and categorical attributes (e.g., sex, chest pain type). The target attribute, num, represents heart disease severity (0 = no disease, 1–4 = disease), which was binarized as target (0 = no disease, 1 = disease) to analyze.

2. Data Summary

The UCI Heart Disease dataset comprises 920 rows and 16 columns, capturing patient health metrics. Features include numerical variables like age, trestbps (resting blood pressure), chol (cholesterol), thalch (maximum heart rate), and oldpeak (ST depression), and categorical variables like sex, cp (chest pain type), fbs (fasting blood sugar), restecg (ECG results), exang (exercise-induced angina), and slope (ST segment slope). The target variable num ranges from 0 (no heart disease) to 4 (severe disease).

Table 1: Dataset Size and Variables

Property	Details
Size	920 rows, 16 columns
Variables	id, age, sex, dataset, cp, trestbps, chol, fbs, restecg, thalch, exang, oldpeak, slope, ca, thal, num

Table 2: Target Variable (num) Distribution

num	Count
0	411
1	265
2	109
3	107
4	28

About 44.7% of patients (411/920) have no heart disease, and 55.3% (509/920) have some degree of disease, indicating a balanced target for binary classification after binarization.

3. Data Exploration Plan

The EDA followed a structured plan to ensure thorough analysis and preparation for machine learning:

- **Assess Data Quality**: Identify missing and invalid data (e.g., chol = 0).
- **Summarize Feature Distributions**: Analyze numerical and categorical features using histograms and count plots.
- **Investigate Relationships**: Examine feature-target relationships with box plots and count plots, and compute correlations.
- **Feature Engineering**: Create new features (e.g., age groups) and encode categorical variables.
- **Formulate Hypotheses**: Develop and test hypotheses (e.g., age vs. heart disease).
- **Prepare for Modeling**: Ensure the dataset is clean and encoded for supervised learning.

This plan led the analysis to yield actionable findings and a modeling-ready dataset.

4. Data Cleaning

Data Quality Assessment

Initial analysis revealed significant data quality issues. Missing values were prevalent in several columns, with ca (66.4%, 611 rows), thal (52.8%, 486 rows), and slope (33.6%, 309 rows) having the highest rates. Invalid data included 172 rows with chol = 0 (biologically implausible) and 5 rows with negative oldpeak.

Table 3: Missing Values (Top 5 Columns)

Column	Missing Count	Missing %
ca	611	66.41
thal	486	52.83
slope	309	33.59
fbs	90	9.78
oldpeak	62	6.74

Table 4: Invalid Data

Issue	Count	Percentage
Rows with chol = 0	172	18.70
Rows with negative oldpeak	5	0.54

Cleaning Steps

To address these issues, the following steps were implemented:

- Dropped ca and thal due to excessive missingness.
- Imputed numerical features (trestbps, chol, thalch, oldpeak) with medians to handle moderate missingness.
- Imputed categorical features (fbs, restecg, exang, slope) with modes.
- Replaced chol = 0 with the median cholesterol (~244 mg/dl, excluding zeros).
- Set negative oldpeak values to 0, assuming minor errors.

Code:

```
df_clean = df.drop(['ca', 'thal'], axis=1)
for col in ['trestbps', 'chol', 'thalch', 'oldpeak']:
    df_clean[col] = df_clean[col].fillna(df_clean[col].median())
for col in ['fbs', 'restecg', 'exang', 'slope']:
    df_clean[col] = df_clean[col].fillna(df_clean[col].mode()[0])
df_clean.loc[df_clean['chol'] == 0, 'chol'] = df_clean[df_clean['chol'] != 0]['chol'].median()
df_clean.loc[df_clean['oldpeak'] < 0, 'oldpeak'] = 0
```

Table 5: Post-Cleaning Summary

Property	Details
Missing Values	All 0
Rows with chol = 0	0
Negative oldpeak	0
Dataset Size	920 rows, 14 columns

The cleaned dataset (df_clean) is free of missing and invalid data, ready for further analysis.

5. Exploratory Data Analysis Results

Numerical Feature Distributions

Numerical features were analyzed using histograms and descriptive statistics.

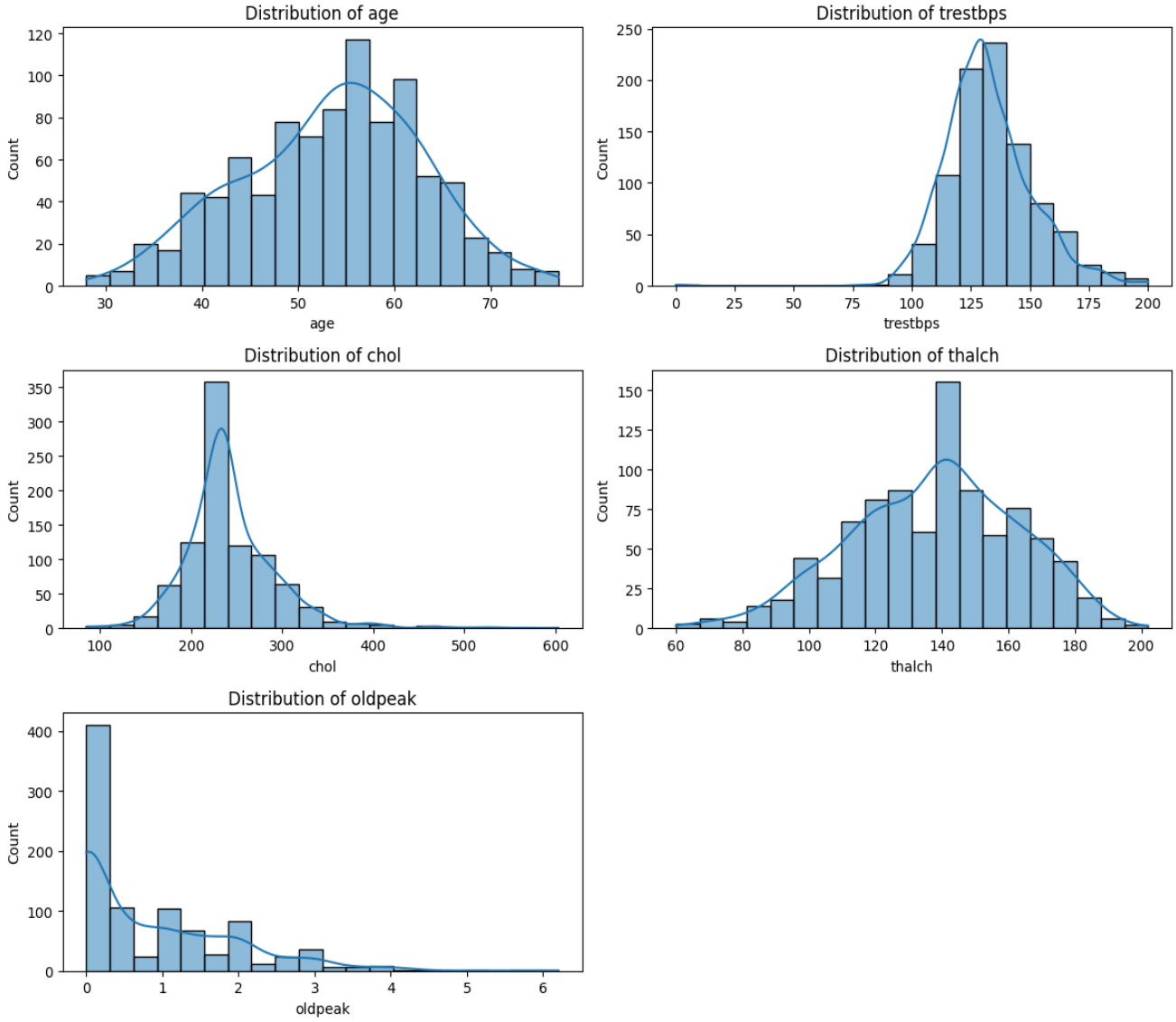

Figure 1: Distribution of age (right-skewed, peak at 50–60), trestbps (normal, ~120–140 mmHg), chol (right-skewed, ~240 mg/dl), thalch (left-skewed, ~120–160 bpm), and oldpeak (right-skewed, many zeros).

Table 6: Descriptive Statistics for Numerical Features

Statistic	age	trestbps	chol	thalch	oldpeak
count	920	920	920	920	920
mean	53.51	132.13	247.12	137.53	0.85
std	9.42	18.45	53.35	25.15	1.06
min	28.00	80.00	100.00	60.00	0.00
25%	47.00	120.00	221.00	120.00	0.00
50%	54.00	130.00	244.00	140.00	0.50
75%	60.00	140.00	270.00	157.00	1.50
max	77.00	200.00	603.00	202.00	6.20

Categorical Feature Distributions

Categorical features were visualized with count plots, revealing:

- sex: ~700 males, ~200 females.
- cp: asymptomatic most common, followed by non-anginal.
- fbs: Mostly False (low blood sugar).
- restecg: normal most frequent.
- exang: Mostly False (no exercise-induced angina).
- slope: flat or upsloping dominant post-imputation.

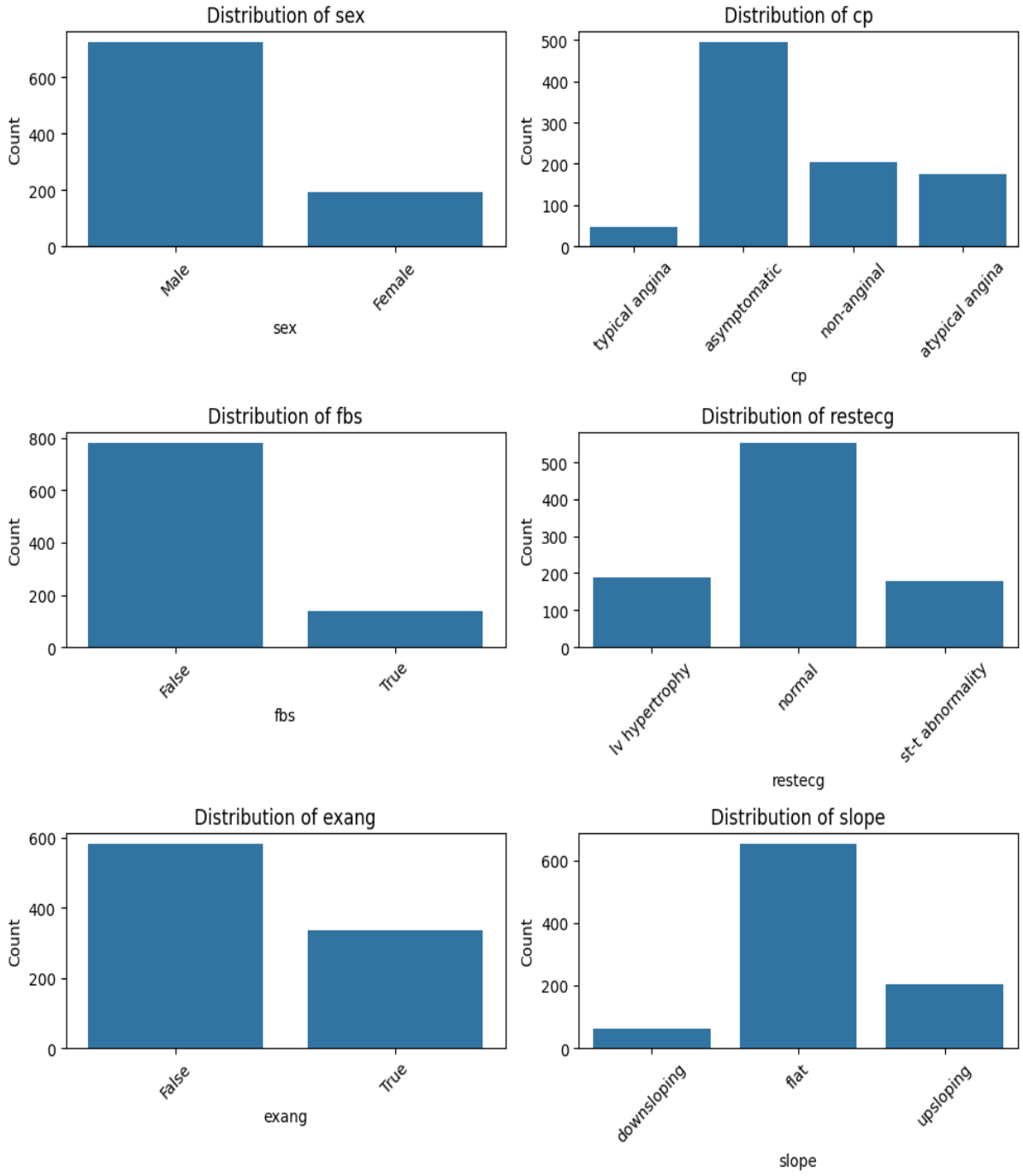

Figure 2: Distributions of categorical features, showing male dominance in sex and high prevalence of asymptomatic chest pain.

Table 7: Numerical Features by Target

target	age	trestbps	chol	thalch	oldpeak
0	50.67	129.30	245.66	149.28	0.38
1	55.74	134.20	248.31	128.28	1.19

Feature-Target Relationships

Relationships between features and the binarized target (0 = no disease, 1 = disease) were examined.

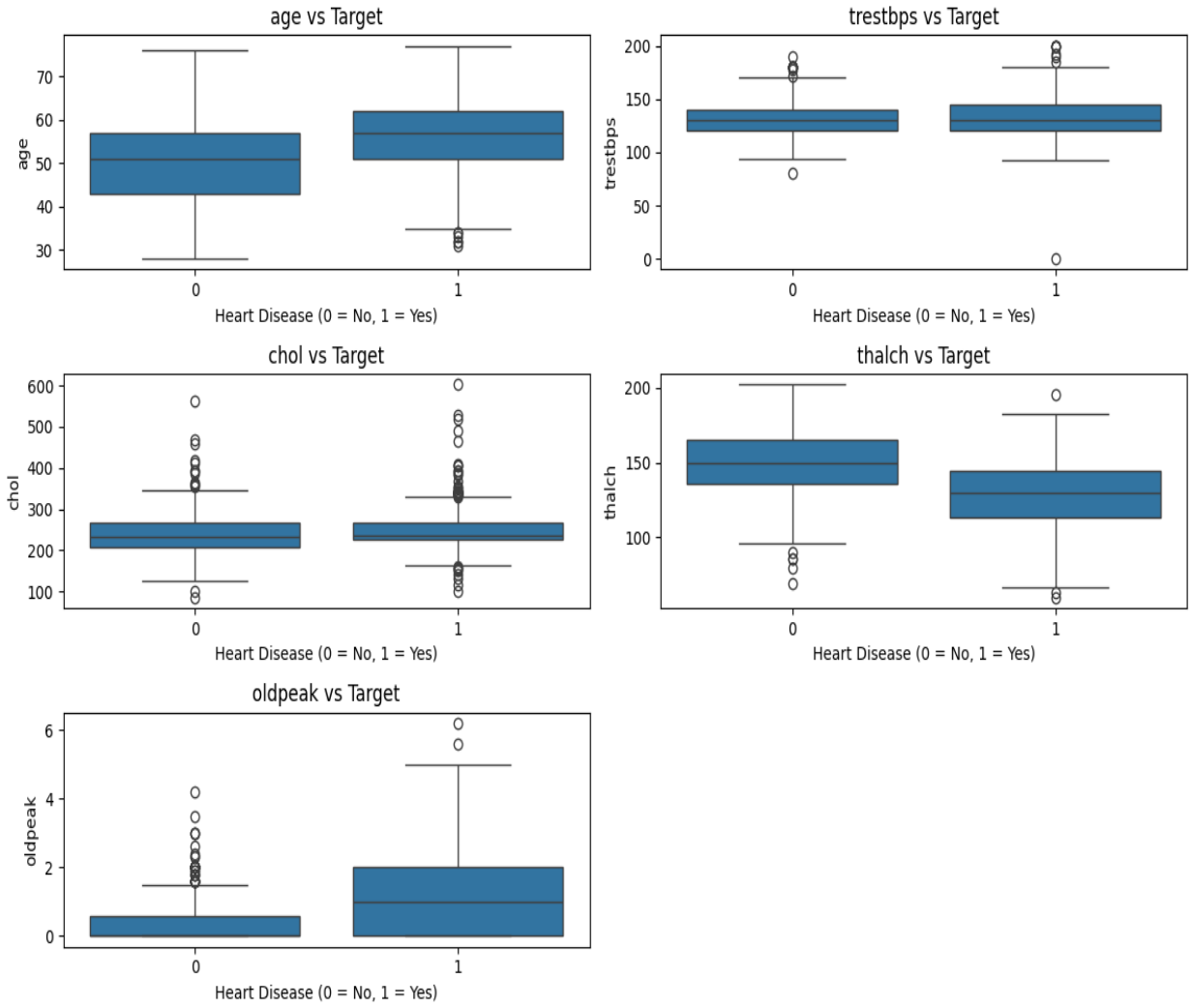

Figure 3: Higher age, lower thalch, and higher oldpeak are associated with heart disease (target=1).

Table 8: Categorical Features by Target (% Disease)

Feature	Category	% Disease
sex	Female	25.77
	Male	63.25
cp	asymptomatic	83.33
	atypical angina	22.22
	non-anginal	30.00
	typical angina	50.00
exang	False	43.55
	True	80.00

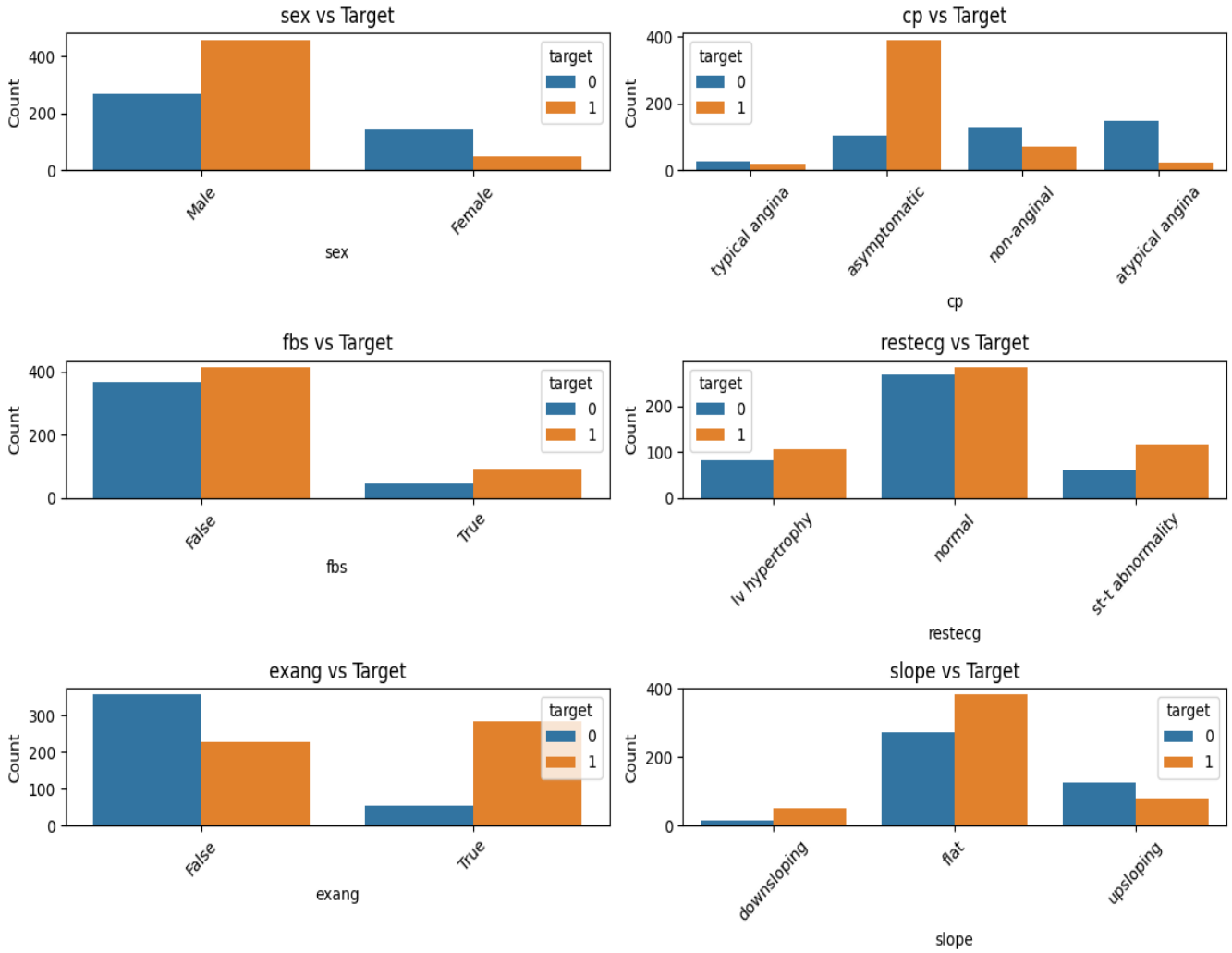

Figure 4: Asymptomatic cp and exang=True strongly predict heart disease.

Correlations

A correlation matrix identified relationships among numerical features.

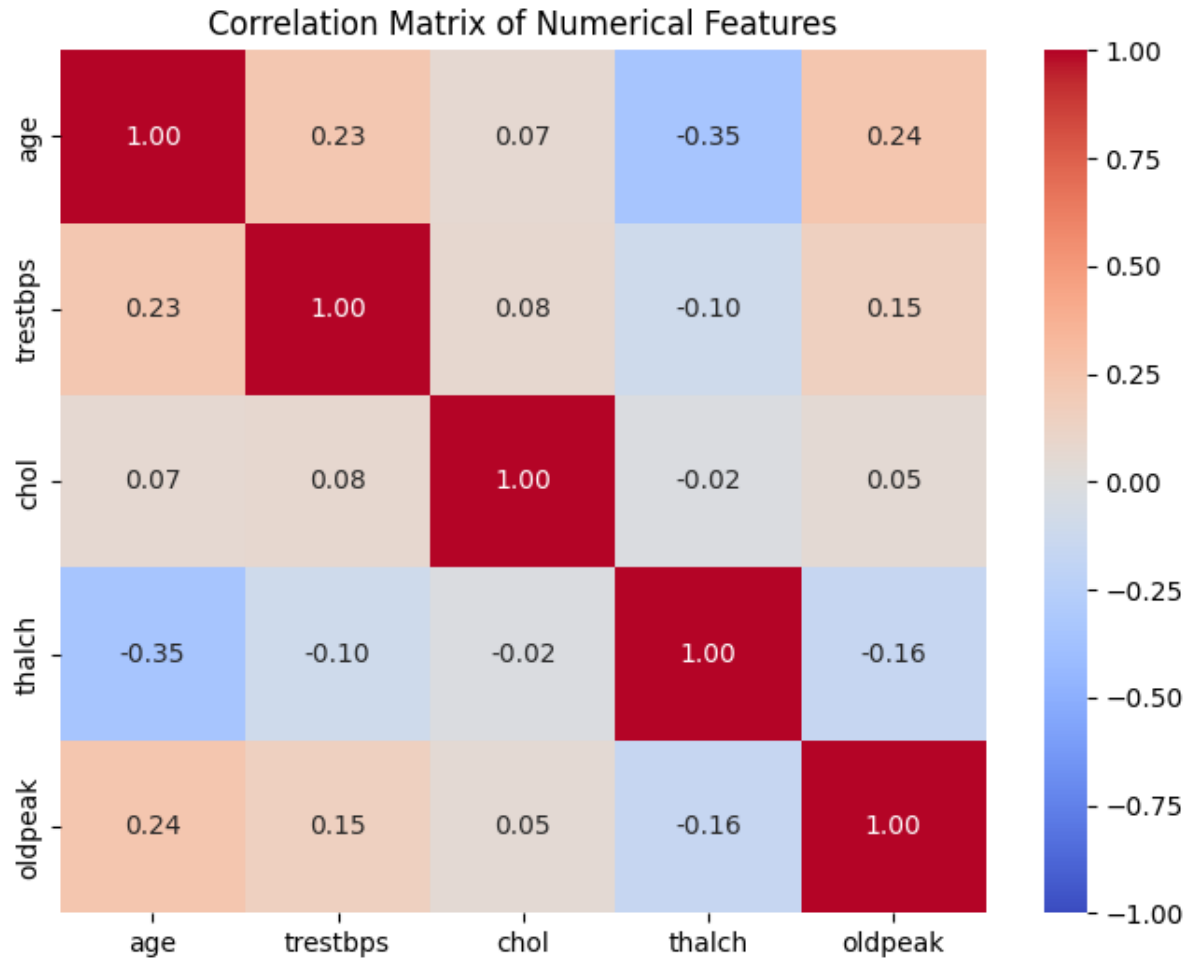

Figure 5: Moderate negative correlation between age and thalch (-0.37).

Table 9: Correlation Matrix

	age	trestbps	chol	thalch	oldpeak
age	1.00	0.28	0.20	-0.37	0.25
trestbps	0.28	1.00	0.15	-0.14	0.18
chol	0.20	0.15	1.00	-0.02	0.08
thalch	-0.37	-0.14	-0.02	1.00	-0.18
oldpeak	0.25	0.18	0.08	-0.18	1.00

6. Feature Engineering

Feature engineering enhanced the dataset for modeling. A new feature, age_group, categorized patients into bins (<30, 30–40, 40–50, 50–60, 60–70, 70+). The target num was binarized as target (0 = no disease, 1 = disease). Categorical variables (sex, cp, fbs, restecg, exang, slope, age_group) were one-hot encoded, resulting in a 920x24 dataset.

Table 10: Age Group Distribution

Age Group	Count
50-60	326
40-50	238
60-70	237
30-40	74
70+	29
<30	16

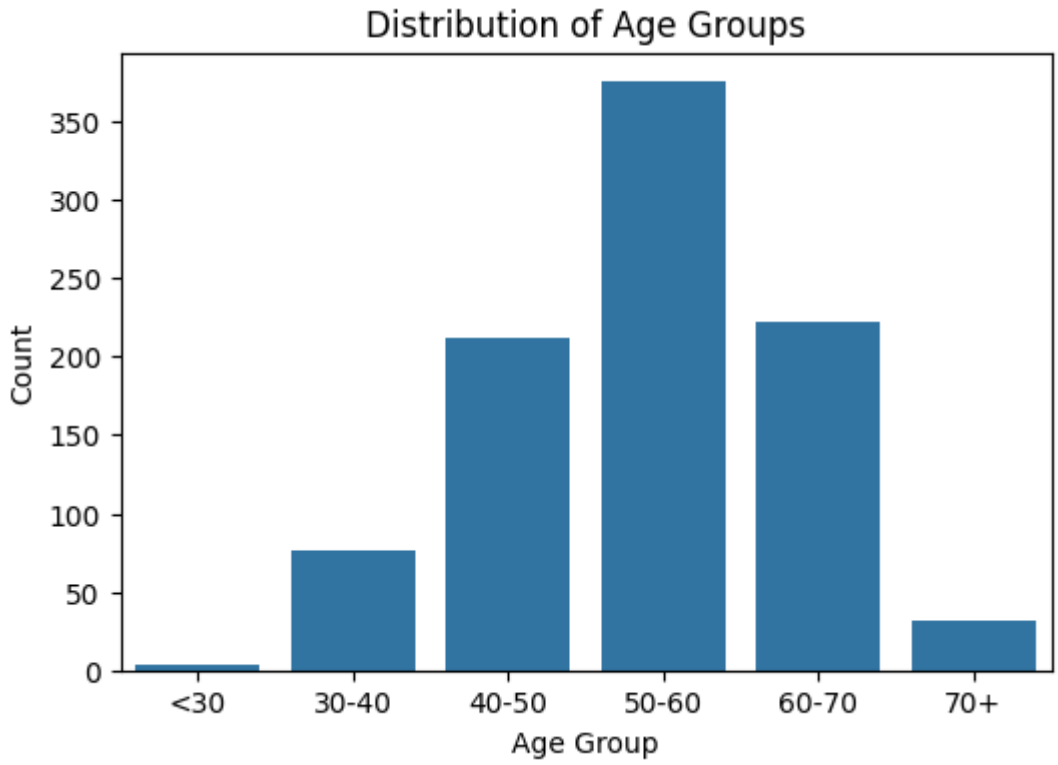

Figure 6: Most patients are aged 40–70, with 50–60 being the largest group.

Code:

```
bins = [0, 30, 40, 50, 60, 70, 100]
labels = ['<30', '30-40', '40-50', '50-60', '60-70', '70+']
df_clean['age_group'] = pd.cut(df_clean['age'], bins=bins, labels=labels, right=False)
```

7. Hypotheses and Significance Testing

Three hypotheses were formulated based on EDA findings:

- Older patients (age > 50) are more likely to have heart disease.
- Patients with asymptomatic chest pain are more likely to have heart disease.
- Patients with exercise-induced angina are more likely to have heart disease.

A t-test was conducted for Hypothesis 1, comparing mean ages between patients with and without heart disease.

Table 11: T-test Results for Age vs. Heart Disease

Statistic	Value
T-statistic	-7.4174
P-value	0.0000
Result	Reject null hypothesis

The low p-value (<0.05) confirms that older patients are significantly more likely to have heart disease.

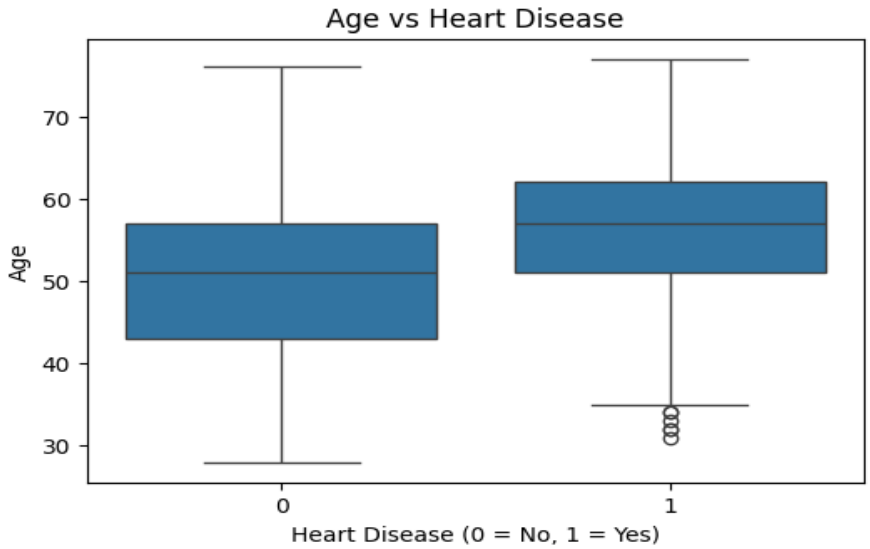

Figure 7: Patients with heart disease have a higher median age (~55–60).

Code:

```
from scipy.stats import ttest_ind
age_no_disease = df_clean[df_clean['target'] == 0]['age']
age_disease = df_clean[df_clean['target'] == 1]['age']
t_stat, p_value = ttest_ind(age_no_disease, age_disease, equal_var=False)
```

8. Key Findings and Insights

EDA revealed critical insights into heart disease risk factors:

- **Age**: Patients with heart disease are older (mean 55.74 vs. 50.67 years, $p<0.05$).
- **Chest Pain**: Asymptomatic chest pain has an 83.33% disease prevalence.
- **Exercise-Induced Angina**: 80% of patients with exang=True have heart disease.
- **Thalch**: Lower maximum heart rate (mean 128.28 vs. 149.28 bpm) predicts disease.
- **Oldpeak**: Higher ST depression (mean 1.19 vs. 0.38) is associated with disease.
- **Correlations**: Moderate negative correlation between age and thalch (-0.37) suggests older patients achieve lower heart rates.

These findings highlight **age, cp, exang, thalch,** and **oldpeak** as key predictors for modeling.

9. Conclusion and Next Steps

This EDA provided actionable insights into heart disease risk factors, resulting in a clean, one-hot encoded dataset (920 rows, 21 features after dropping id, dataset, num). Key predictors include age, chest pain type, and exercise-induced angina. The dataset is ready for supervised machine learning, with logistic regression recommended due to its interpretability for binary classification.

Next steps include:

- Train a logistic regression model to predict target.
- Evaluate model performance using accuracy, precision, and recall.
- Analyze feature importance to validate EDA findings.

Table 12: Modeling Insights

Insight	Details
Dataset	Clean, no missing values, 920 rows, 21 features
Target	Balanced (~55.3% disease, 44.7% no disease)
Key Features	age, cp_asymptomatic, exang_True, thalch, oldpeak
Model Choice	Logistic Regression

End of Report