Data Analysis Report: CardioDetect

Executive Summary

This report analyzes the combined heart disease dataset used for the CardioDetect project. The dataset contains **5160 samples** and **13 features**.

Key Findings: 1. **Critical Data Quality Issues**: Several key features (`cp`, `restecg`, `exang`, `slope`, `ca`, `thal`) are missing >80% of their values. This severely limits the model's ability to learn from these medically significant indicators. 2. **Class Imbalance**: The dataset is imbalanced, with **77.7% Healthy** and **22.3% Disease** cases. This explains the model's tendency towards high accuracy but low recall (bias towards predicting "Healthy"). 3. **Data Anomalies**: There are **172 samples with 0 cholesterol** and **1 sample with 0 blood pressure**, which are physiologically impossible and likely represent missing data encoded as zeros. 4. **Risk Factors**: `thal` (Thalassemia), `cp` (Chest Pain), and `exang` (Exercise Induced Angina) show the strongest correlation with heart disease, despite the high missingness.

---

1. Dataset Overview

- **Total Samples**: 5160
- **Total Features**: 13
- **Target Variable**: `target` (0 = Healthy, 1 = Disease)

Target Distribution

The dataset is significantly imbalanced. - **Healthy (0)**: 4007 (77.66%) - **Disease (1)**: 1153 (22.34%)

> [!WARNING] **Imbalance Impact**: A "dumb" model predicting "Healthy" for everyone would achieve **77.7% accuracy**. This makes Accuracy a misleading metric. We must prioritize **Recall** (catching disease cases) and **F1-Score**.

---

2. Data Quality & Missing Values

The dataset suffers from extreme missingness in several features.

Feature	Missing Count	Percentage	Status
ca	4851	**94.01%**	■ Critical
thal	4726	**91.59%**	■ Critical
slope	4549	**88.16%**	■ Critical
oldpeak	4302	**83.37%**	■ Critical
exang	4295	**83.24%**	■ Critical
restecg	4242	**82.21%**	■ Critical
cp	4240	**82.17%**	■ Critical
fbs	90	1.74%	■ Good
chol	80	1.55%	■ Good
trestbps	59	1.14%	■ Good
thalach	56	1.09%	■ Good

> [!NOTE] The high missingness suggests that this "combined" dataset likely merges a smaller, high-quality dataset (with all fields) with a larger dataset that lacks these specific cardiac measures.

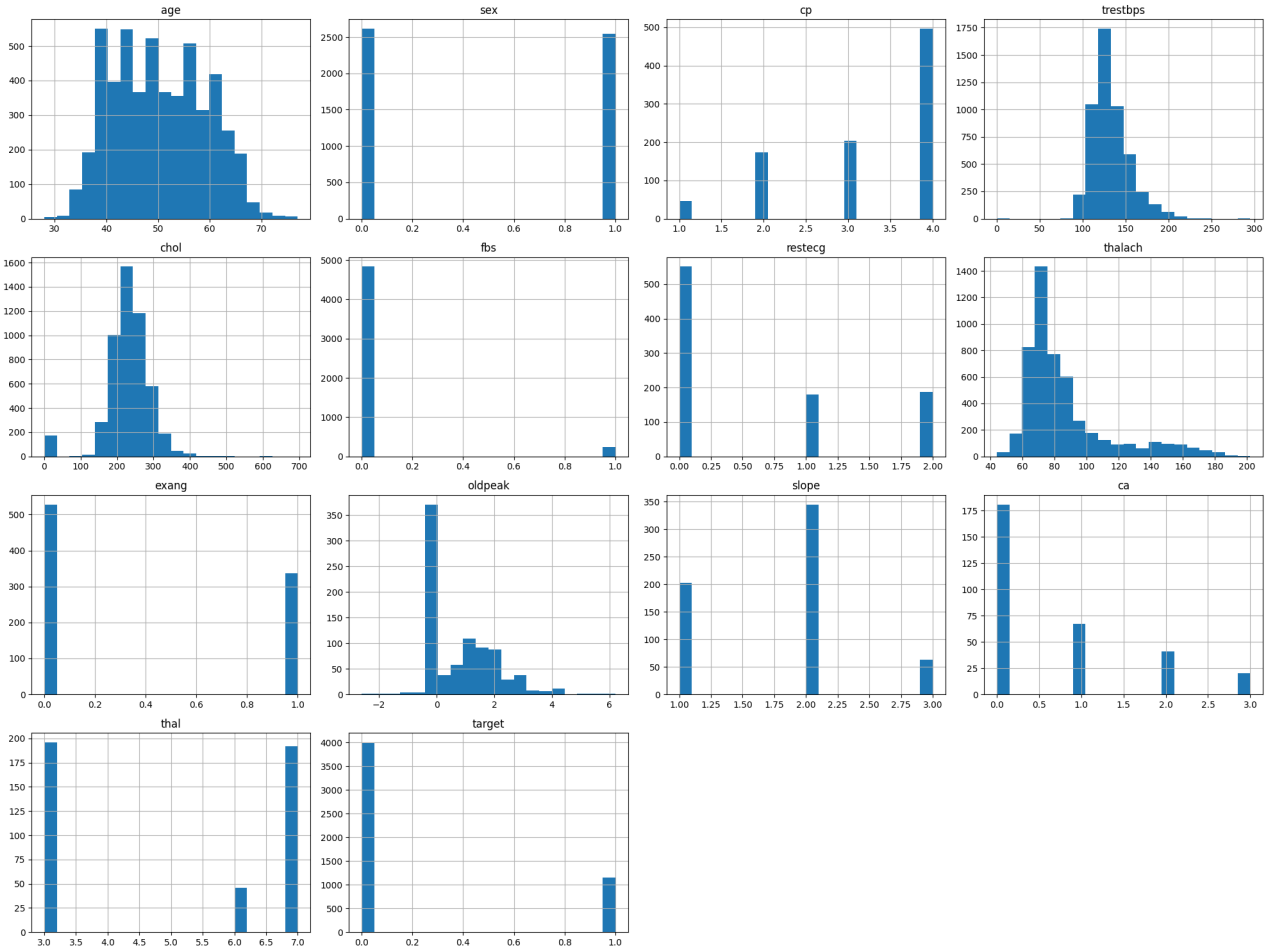
---

3. Numerical Features Analysis

Feature	Mean	Std	Min	Median	Max	Anomalies

age	50.3	8.9	28.0	50.0	77.0	None
trestbps	132.3	21.6	**0.0**	129.0	295.0	1 zero value
chol	230.1	63.2	**0.0**	232.0	696.0	**172 zero values**
thalach	86.3	27.7	44.0	77.0	202.0	None
oldpeak	0.88	1.09	-2.6	0.5	6.2	Negative values?

Visualizations



Distribution of numerical features. Note the normal distribution of Age and Trestbps.

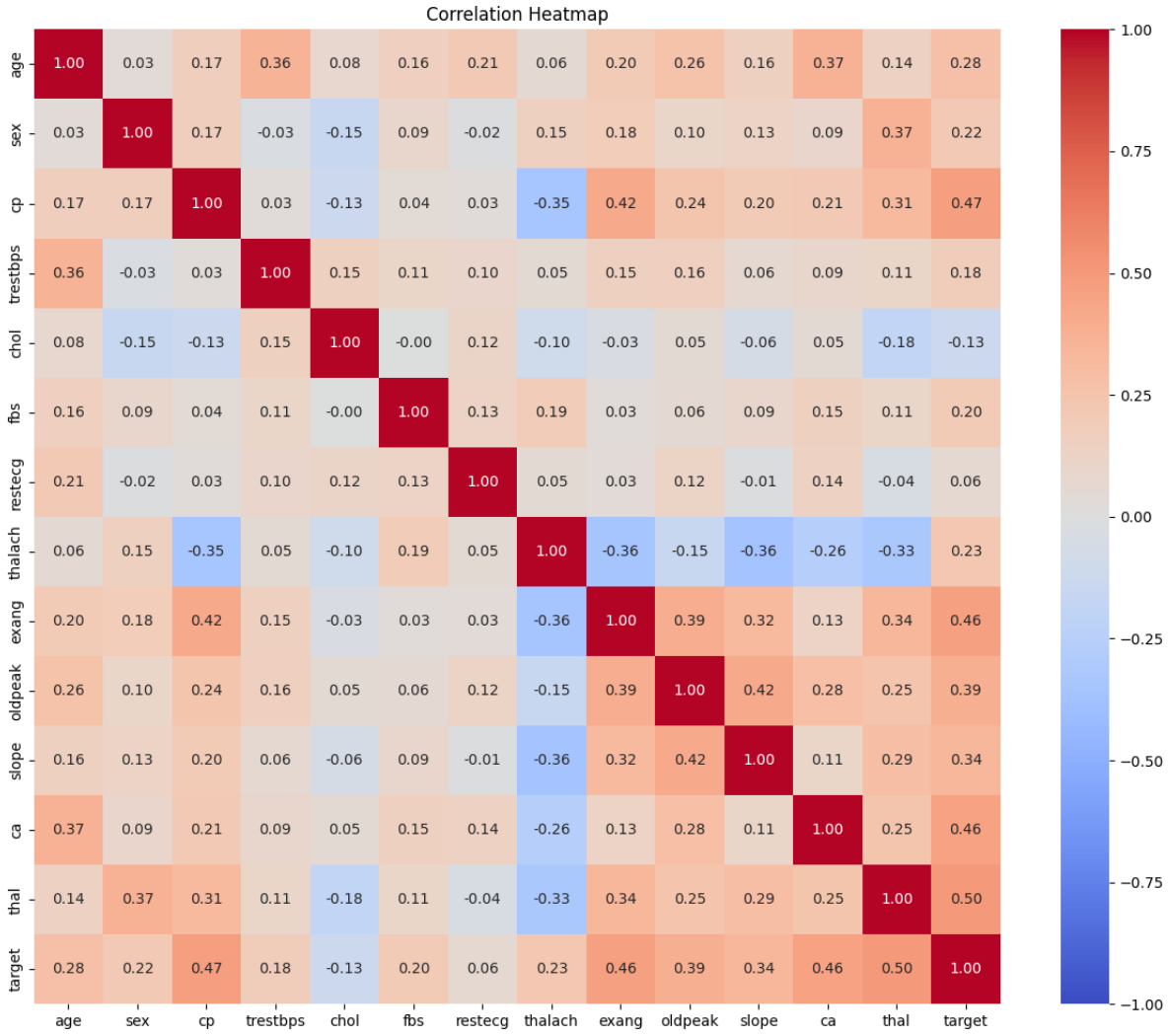
---

4. Correlation Analysis

Which features are most predictive of heart disease?

Feature	Correlation	Strength	Interpretation
thal	0.499	Strong	Thalassemia type is highly predictive.
cp	0.472	Moderate	Chest pain type is a key indicator.
exang	0.464	Moderate	Angina during exercise is a strong warning sign.
ca	0.456	Moderate	Number of major vessels colored by fluoroscopy.
oldpeak	0.386	Moderate	ST depression induced by exercise.
slope	0.337	Moderate	Slope of the peak exercise ST segment.
age	0.281	Low	Risk increases with age.
thalach	0.232	Low	Max heart rate achieved.

sex	0.217	Low	Gender plays a role (likely higher risk for males in this dataset).
chol	-0.127	Low	**Unexpected**: Negative correlation. Likely due to noise or "0" values.



Heatmap showing correlations between features and target.