# Project Challenge: Heart Disease Analysis

**Applied Data Science with Python for Beginners Bootcamp Contest #1**

Tina Munda

January 2026

# Introduction

- Objective: Which factors are associated with heart disease in a clinical sample (N=303)?
- Dataset: Cleveland Clinic evaluation sample (not general population!)
- Sample size: 303 patients
- Variables:
  - age,
  - sex,
  - resting blood pressure in mm Hg (trestbps),
  - serum cholestorol in mg/dl (chol),
  - chest pain type (cp),
  - exercise-induced angina (exang),
  - fasting blood sugar (fbs),
  - maximum heart rate achieved in exercise test (thalach),
  - heart_disease (yes or no)

# Benchmarks / thresholds – cholesterol vs 240 mg/dl

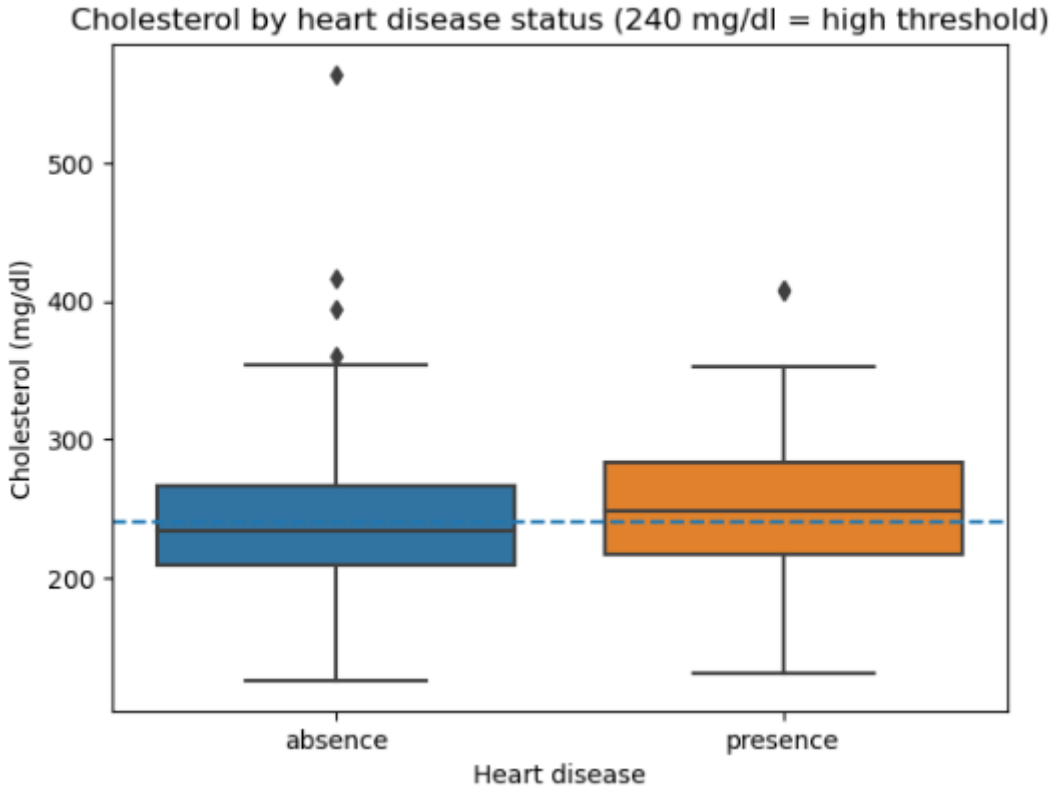
- Question: For each subgroup (yes_hd, no_hd), is mean cholesterol significantly greater than 240 mg/dl (the "high cholesterol" threshold)?
- Results:
  - **Heart disease (presence):** Mean cholesterol **251.47 mg/dl** (**+11.47** above 240);

    one-sample t-test vs 240

    **p(one-sided)=0.0035 → significantly higher!**
  - **No heart disease (absence):** Mean cholesterol **242.64 mg/dl** (**+2.64** above 240);

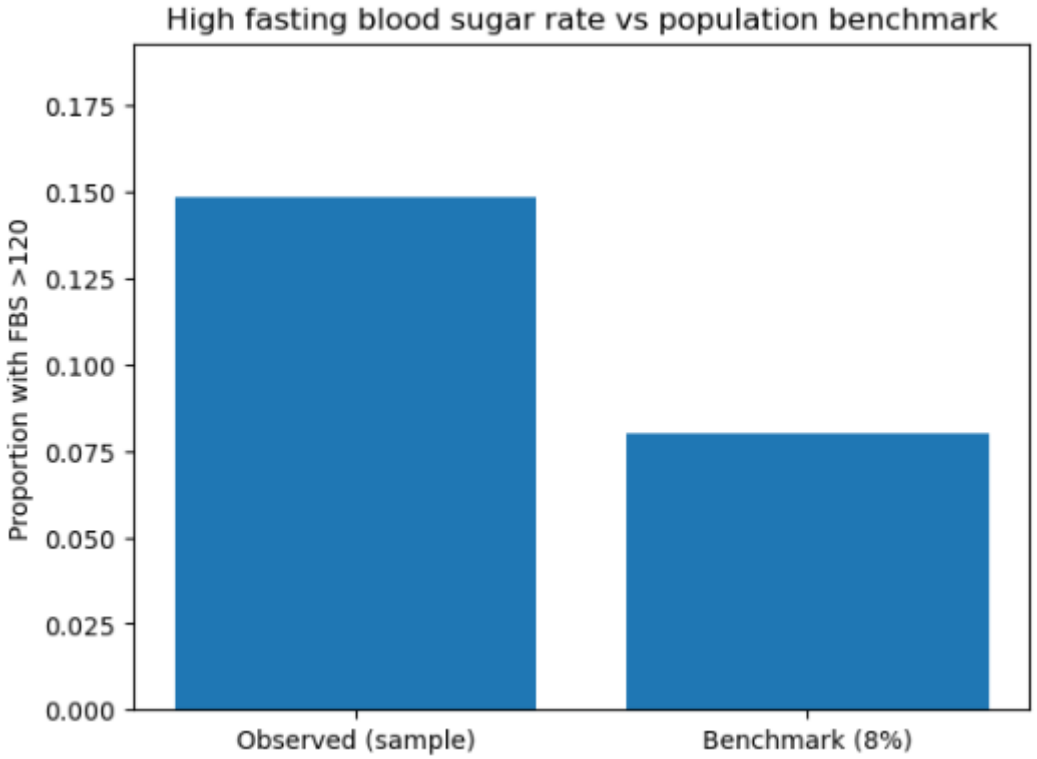
    one-sample t-test vs 240

    **p(one-sided)=0.264 → not significantly higher**.
- Conclusion: In this sample, **only the heart disease group shows evidence that average cholesterol exceeds the "high cholesterol" threshold**.



# Benchmarks / thresholds – fasting blood sugar (fbs) vs 8% population benchmark

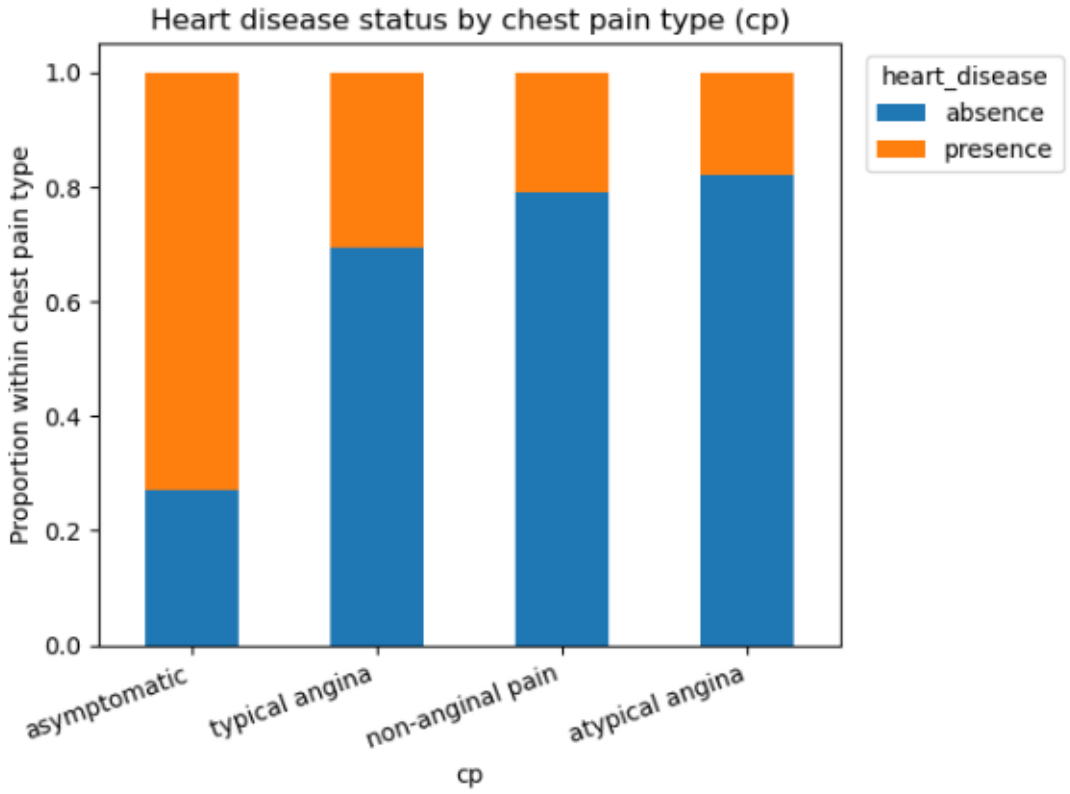
- Question: Is the rate of **FBS >120 mg/dl** in this sample consistent with an **8%** population baseline (1988), or is it **higher**?
- Results:
  - **High FBS (>120): 45** patients → **14.85%** (45/303)
  - **Expected at 8%: ~24** patients (0.08 × 303)
  - **Binomial test (H1: rate > 8%): p = 4.69e–05 → significantly higher!**
- Conclusion: This clinical sample shows an **elevated high-FBS rate** compared to the 8% benchmark (expected ~24, observed 45), suggesting this clinical sample is **not representative** of the general population (from 1988) on this metric.




**observed 45 vs expected 24**

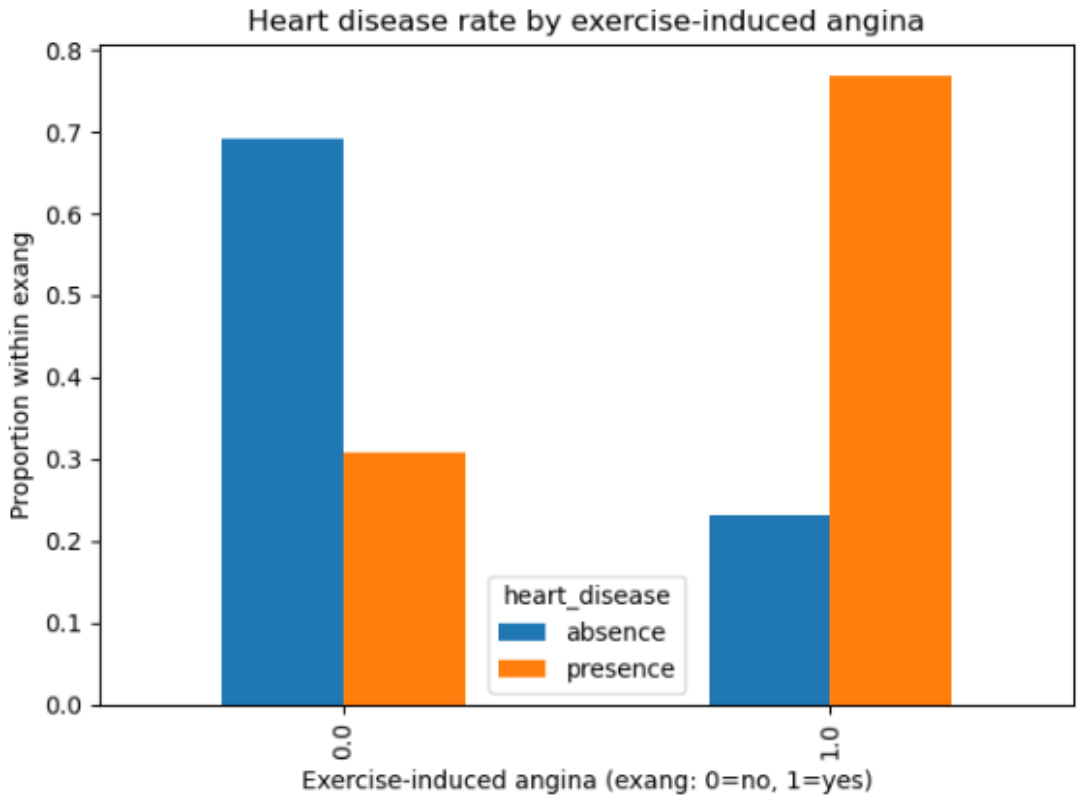
# Symptom + exercise signals – kind of chest pain vs heart disease

- Question: Is **chest pain type** (cp) associated with whether a patient will ultimately be diagnosed with **heart disease**?
- Results (contingency table + $\chi^2$ test):
  - **Yes**, chest pain type and heart disease diagnosis are significantly associated in this sample
    - Chi-square test: **p = 1.25e–17 → statistically significant association!**
  - Among all cp types, **asymptomatic** patients show the highest heart disease count: **105 presence vs 39 absence** (105/144 = **72.9% presence**). Other pain types skew toward **absence** (e.g., atypical angina: **9/50 = 18% presence**).
- Conclusion: Chest pain type is **strongly associated** with heart disease diagnosis; in this sample, **asymptomatic** patients are much more likely to be diagnosed with heart disease.



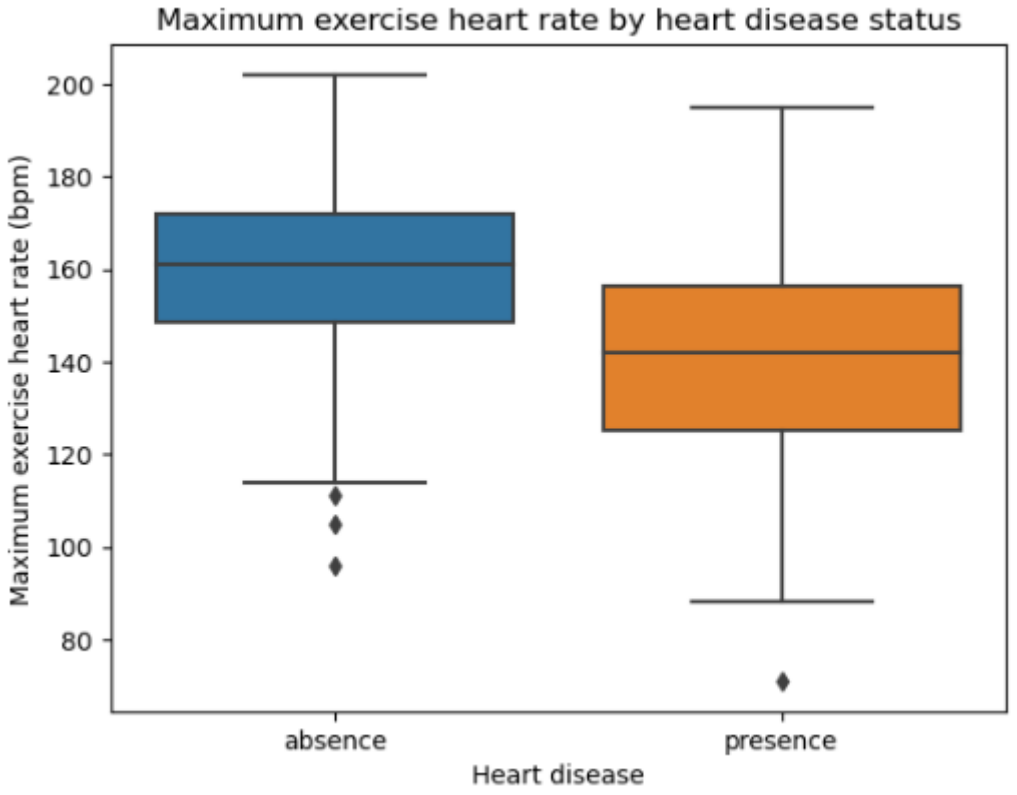
# Symptom + exercise signals – exercise-induced angina vs heart disease

- Question: Is **exercise-induced angina** (exang) associated with whether a patient will ultimately be diagnosed with **heart disease**?
- Results (2×2 contingency table + $\chi^2$ test):
  - **No exercise-induced angina** (exang = 0): **30.9%** (**63/204**) diagnosed with heart disease
  - **Exercise-induced angina present** (exang = 1): **76.8%** (**76/99**) diagnosed with heart disease
  - Chi-square test of independence: **p = 1.41e–13** → **significant association**
- Conclusion: **Exercise-induced angina** is a **strong indicator** in this sample: patients experiencing exercise-induced angina (exang=1) are **much more likely** to have heart disease.



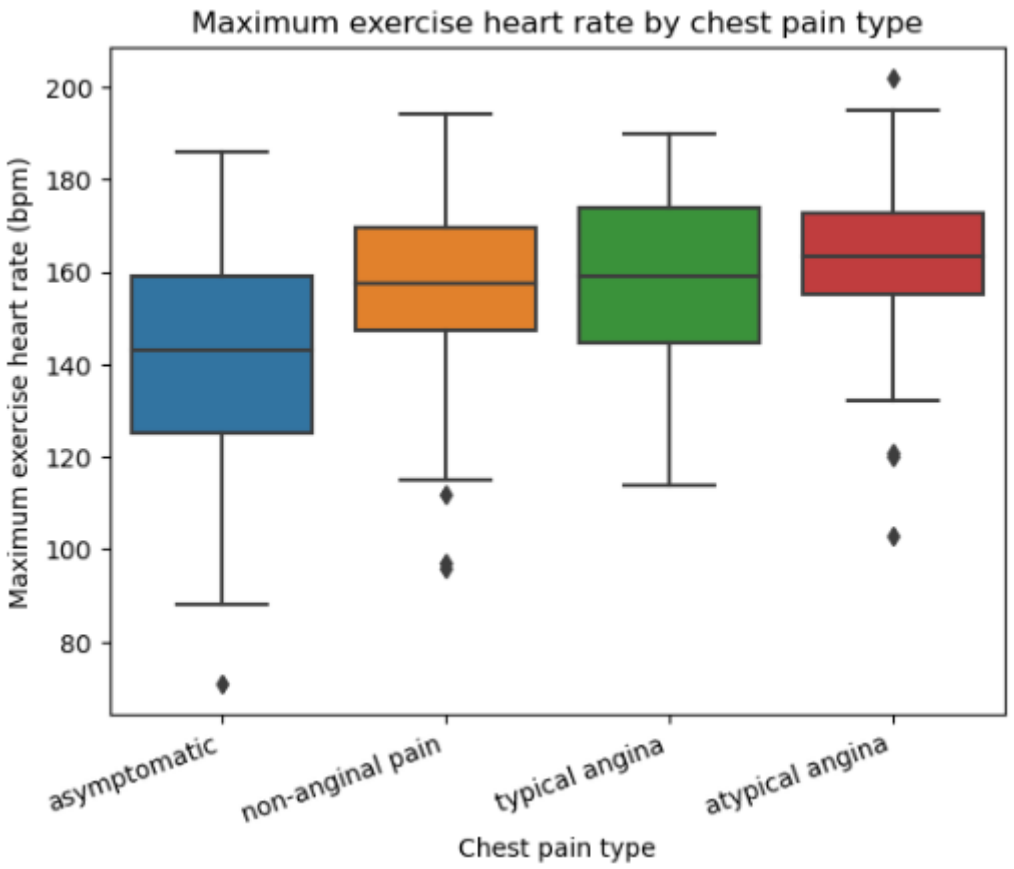
# Symptom + exercise signals – maximum exercise heart rate vs heart disease

- Question: Is **maximum exercise heart rate** (thalach) associated with whether or not a patient will ultimately be diagnosed with **heart disease**?
- Results:
  - Patients with heart disease reached a **lower** max heart rate, **~19 bpm lower** on average :
    - Mean difference: **–19.12 bpm**; median difference: **–19.0 bpm**
    - Two-sample t-test: **p = 3.46e–14** → **statistically significant!**
- Conclusion: Patients diagnosed with heart disease achieve a **substantially lower** maximum heart rate during the exercise test (**≈ 19 bpm lower**).



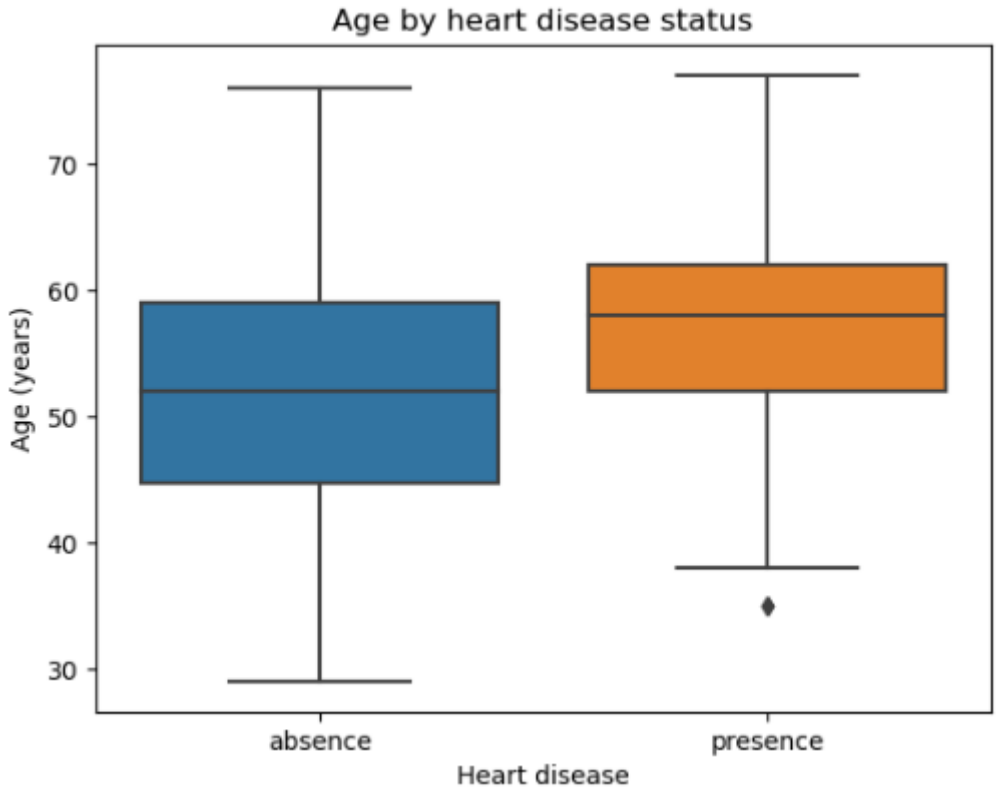
# Symptom + exercise signals – maximum exercise heart rate by chest pain type

- Question: a) Do patients **with different chest pain** have the **same average maximum exercise heart rate** (thalach), or does **at least one** chest pain group have a different average thalach? b) **If so, which of those pairs are significantly different?**
- Results a) (4-group comparison):
  - Side-by-side- boxplots show **clear differences in typical thalach across cp categories** (asymptomatic appears lowest).
    - One-way ANOVA: **p = 1.91e–10 → at least one pair of chest pain types has a different mean thalach.**
- Results b) (Tukey HSD, FWER=0.05):
  - **Asymptomatic** vs atypical angina: **+21.74 bpm (p-adj < 0.001) → significant**
  - **Asymptomatic** vs non-anginal pain: **+14.73 bpm (p-adj < 0.001) → significant**
  - **Asymptomatic** vs typical angina: **+15.28 bpm (p-adj = 0.0081) → significant**
  - All other pairwise comparisons: **not significant** (p-adj ≥ 0.248)
- Conclusion: a) People with typical angina, non-anginal pain, atypical angina, and asymptomatic people **do not all have the same average thalach**. b) The overall difference is driven by the **asymptomatic group**, which has a **significantly lower mean thalach** than each of the other chest pain types; the other three types have **similar** average thalach.



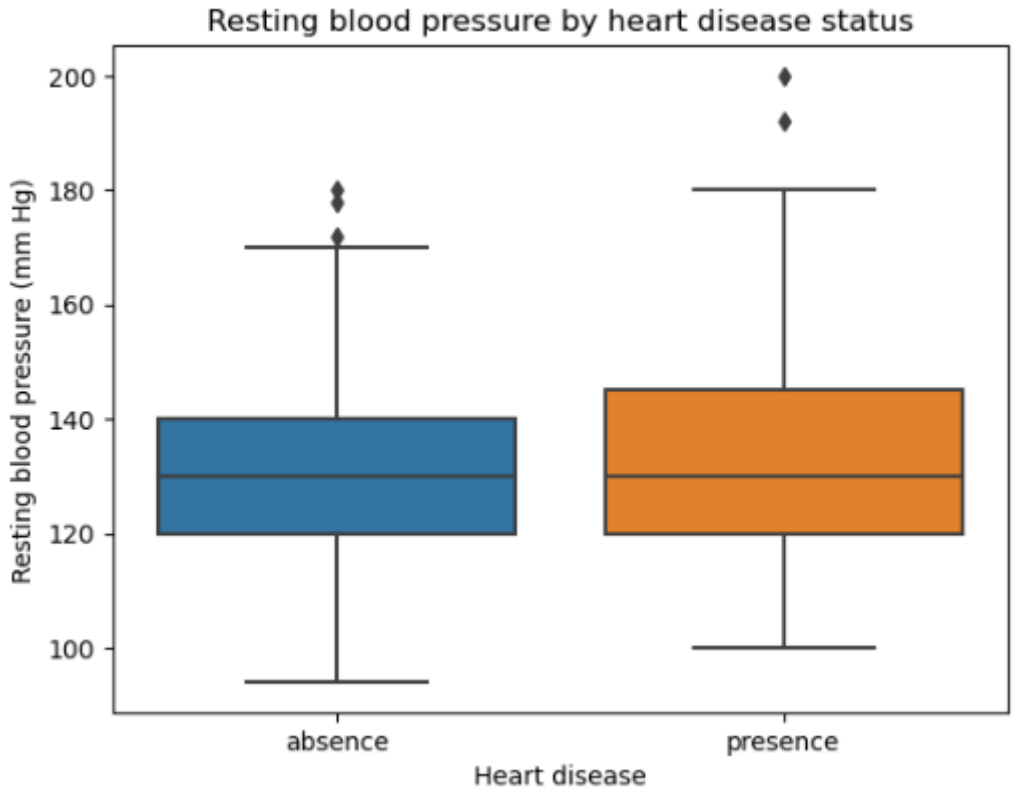
# Demographics/vitals – age vs heart disease

- Question: Is **age** associated with whether a patient will ultimately be diagnosed with **heart disease**?
- Results (group comparison + Welch t-test):
  - **With heart disease (presence):** mean age **56.6** years (median **58**)
  - **Without heart disease (absence):** mean age **52.6** years (median **52**)
    - Difference: **+4.0 years** on average (**+6** years at the median)
  - Welch t-test: **p = 7.06e–05** → **statistically significant**
- Conclusion: In this sample, heart disease is significantly associated with **older age**; the diagnosed group is **several years older** on average.



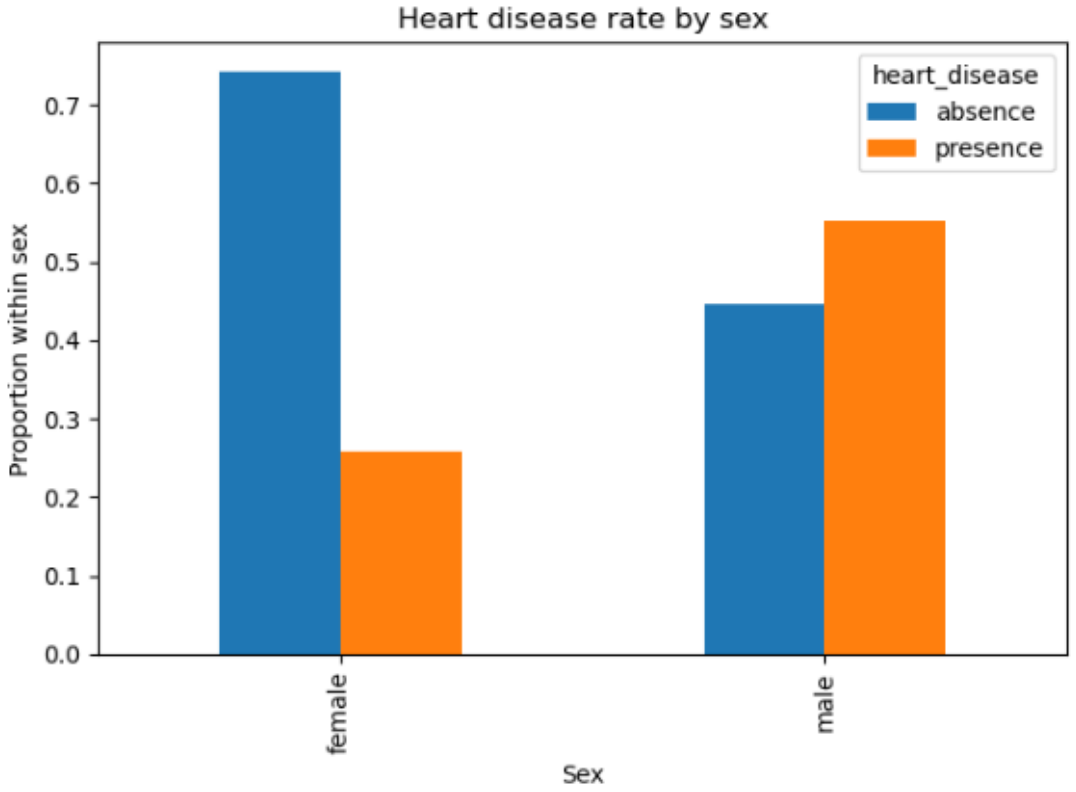
# Demographics/vitals – resting blood pressure vs heart disease

- Question: Is **resting blood pressure** (trestbps) associated with whether a patient will ultimately be diagnosed with **heart disease**?
- Results:
  - Resting BP is **higher** in the heart disease group by about **+5.32 mm Hg** on average
    - Mean difference **+5.32**; median difference **0.0**
    - Two-sample t-test: **p = 0.00855** → **statistically significant!**
- Conclusion: Resting blood pressure shows a **small but significant** association with heart disease; the average is higher in the heart disease group, though the **median is unchanged**.



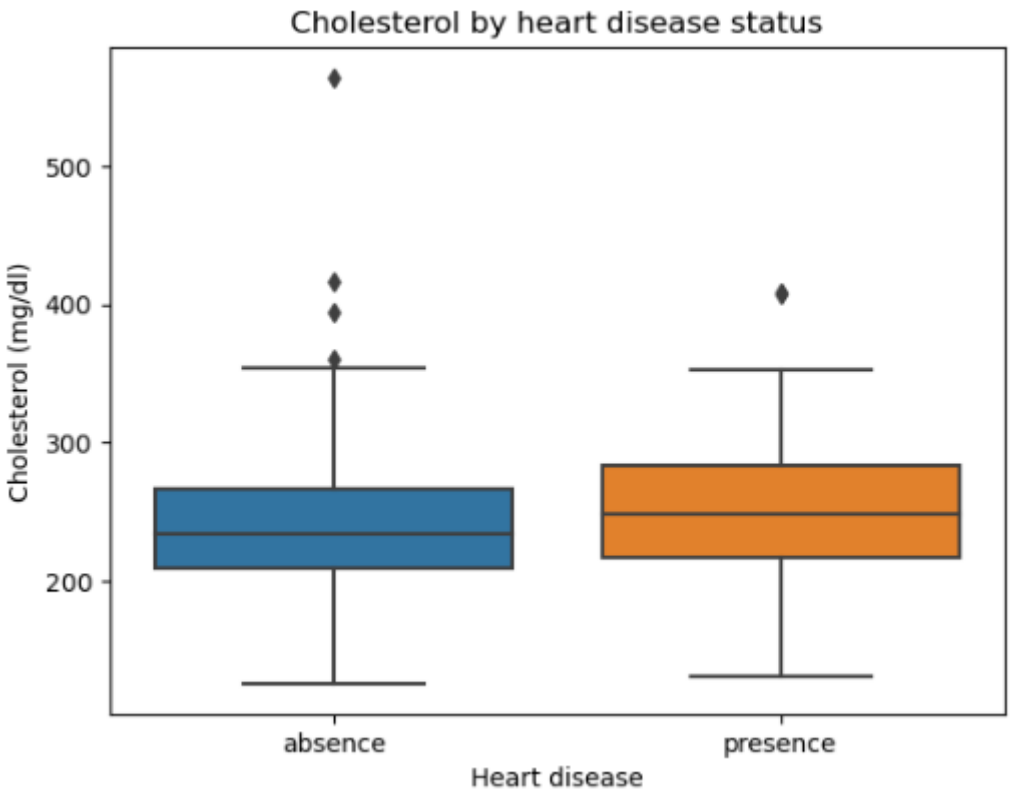
# Demographics/vitals – sex vs heart disease

- Question: Is **sex** associated with whether a patient will ultimately be diagnosed with **heart disease**?
- Results (2×2 table + $\chi^2$ test):
  - **Female: 25.8%** (**25/97**) diagnosed with heart disease
  - **Male: 55.3%** (**114/206**) diagnosed with heart disease
  - Chi-square test: **p = 2.67e−06** → **statistically significant association**
- Conclusion: In this clinical sample, **males are more likely to be diagnosed** with heart disease than females.



# Demographics/vitals – cholesterol vs heart disease

- Question: Is **cholesterol** (chol) associated with whether a patient will ultimately be diagnosed with **heart disease**?
- Results:
  - Cholesterol is **higher** in the heart disease group, but the difference is modest
    - Mean difference **+8.83 mg/dl**; median difference **+14.5 mg/dl**.
    - Two-sample t-test: **p = 0.137** → **not statistically significant**
- Conclusion: In this sample, cholesterol is **not a strong discriminator** between heart disease vs no heart disease (despite being above the 240 mg/dl threshold within the heart disease group(see slide 3)).



# Key takeaways

- **Symptom + exercise signals are the strongest markers:** chest pain type (**$\chi^2$ p = 1.25e–17**) and exercise-induced angina (**HD rate 76.8% vs 30.9%, $\chi^2$ p = 1.41e–13**) show large separation.
- **Exercise capacity differs strongly by diagnosis:** patients with heart disease reached **~19 bpm lower** max exercise heart rate (**p = 3.46e–14**).
- **Demographics/vitals add signal:** heart disease patients are **older** (**+4 years mean; p ≈ 7.06e–05**) and have **slightly higher resting BP** (**+5.3 mmHg; p = 0.0086**).
- **Lab tests tell a nuanced story:** heart-disease group's mean cholesterol is **>240** (threshold test), but cholesterol is **not a strong between-group discriminator** (two-sample p ≈ 0.137). High FBS is **elevated vs 8% benchmark** (14.85% vs 8%, **p = 4.69e–05**).

## Recommendations

- For quick screening in similar clinical settings, prioritize **symptom/exercise indicators**: **cp**, **exang**, and **thalach** (largest separations).
- Use **age + resting BP** as supportive risk context.
- Continue monitoring **metabolic risk (chol, FBS)** even if they're weaker discriminators here – they matter for prevention and overall cardiovascular risk.

## Limits

- **Not a general-population sample** (clinic evaluation cohort): rates (e.g., FBS) won't match population baselines.
- **Associations ≠ causation**; unmeasured confounders (meds, comorbidities, lifestyle) may drive patterns.
- Some subgroups are small (e.g., typical angina), and multiple comparisons can inflate false positives (Tukey helps for cp–thalach pairs).