Project Step 1

Minu Pabbathi, Ziqian Zhao, Paul Zhang

2024-04-26

About the data

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains information about patients and risks for heart disease as well as whether or not they eventually end up developing heart disease. The data set originally contains 1025 observations. For the sake of clearness of the scatter plot, we randomly choose 500 from the original dataset.

Data Description

The `age` variable is a numeric variable that lists the age of the patient in years. The `sex` variable is a binary variable that states the sex of the patient, with 0 representing female and 1 representing male. The `cp` variable is a categorical variable that classifies the type of chest pain the patient is experiencing; 0 represents typical angina, 1 represents atypical angina, 2 represents non-anginal pain, and 3 represents asymptomatic. The `trestbps` variable is a numeric variable that states the resting blood pressure in mmHg.

The `chol` variable is a numeric data which give an overview of person's cholesterol levels. The `fbs` variable is a binary data with 1 means the fast blood sugar is greater than 120mg/dl, and 0 means the fast blood sugar is not. The `restecg` variable is a nominal data with scale of 0 (normal), 1 (abnormality in ST-T waves), and 2 (show left ventricular hypertrophy), which record the resting electrocardiogram results. The `thalach` variable is a numeric data which records the maximum heart rate achieved.

The variable `exang` indicates that if the patient have exercise induced angina (chest pain), which is shown with 0 indicating "no" and 1 indicating "yes". The variable `oldpeak` is a float type data records the ST depression, which is a measure of abnormality of an electrocardiogram, and the measurement is in unit depression. The variable `slope` is the slope of the peak exercise ST segment, which is an electrocardiography read out indicating quality of blood flow to the heart, and the data are in nominal type including 0 (upsloping), 1 (flat) and 2 (downsloping). Finally, the variable `target` is a binary data showing whether the patient is having heart disease (1 = having heart disease and 0 = normal).

Summary Statistics and Graphs

Table 1: Data summary

Name	heart_disease_data_1
Number of rows	500
Number of columns	14

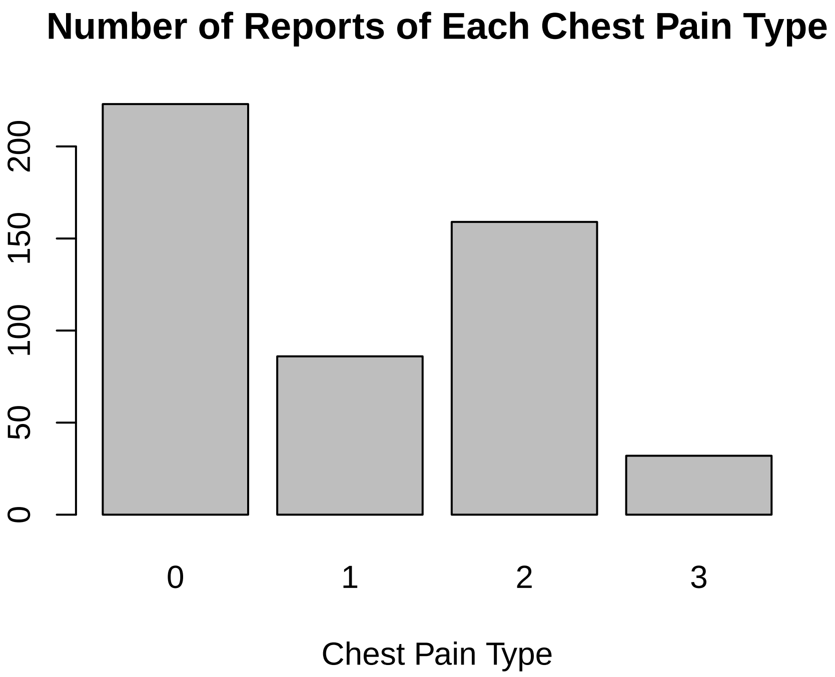
Column type frequency:

numeric	14
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	54.02	9.22	29	47	55.0	60.25	77.0	
sex	0	1	0.71	0.45	0	0	1.0	1.00	1.0	
cp	0	1	1.00	1.01	0	0	1.0	2.00	3.0	
trestbps	0	1	132.46	18.06	94	120	130.0	140.00	200.0	
chol	0	1	248.13	55.44	126	211	243.0	282.00	564.0	
fbs	0	1	0.16	0.37	0	0	0.0	0.00	1.0	
restecg	0	1	0.48	0.52	0	0	0.0	1.00	2.0	
thalach	0	1	150.19	23.31	71	137	152.0	168.00	202.0	
exang	0	1	0.34	0.47	0	0	0.0	1.00	1.0	
oldpeak	0	1	1.03	1.17	0	0	0.8	1.60	6.2	
slope	0	1	1.41	0.62	0	1	1.0	2.00	2.0	
ca	0	1	0.76	1.07	0	0	0.0	1.00	4.0	
thal	0	1	2.29	0.63	0	2	2.0	3.00	3.0	
target	0	1	0.53	0.50	0	0	1.0	1.00	1.0	

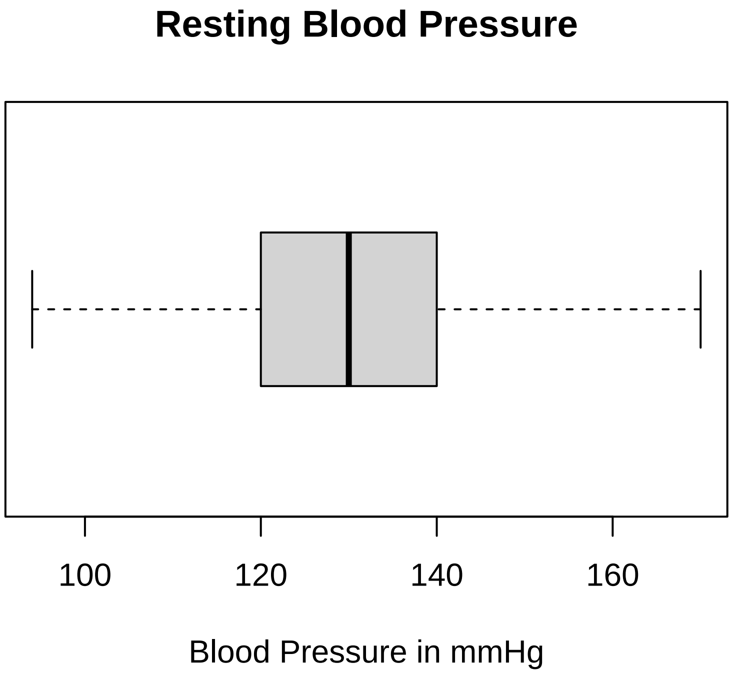
Chest Pain Type



0	1	2	3
typical angina	atypical angina	non-anginal pain	asymptomatic

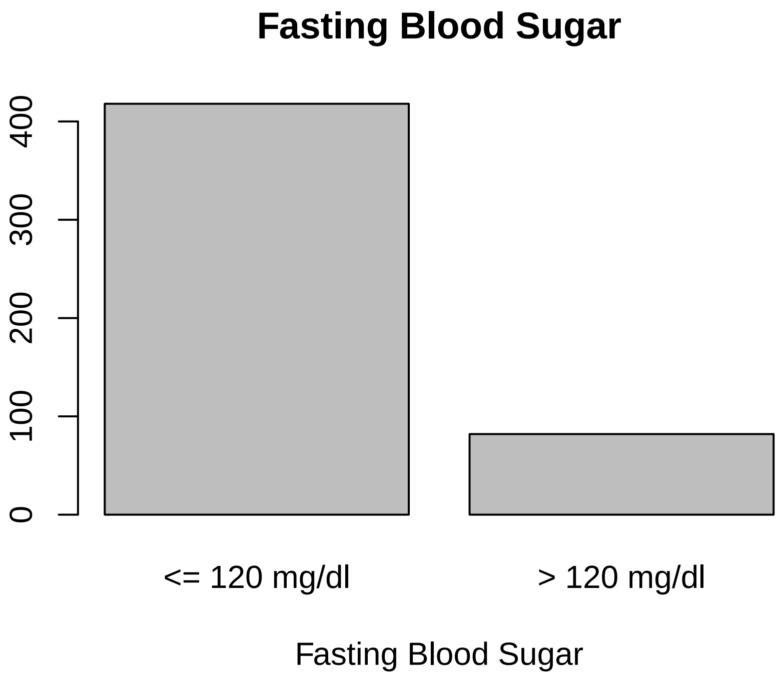
The barplot indicates that majority of patients experienced typical angina, followed by non-anginal pain, atypical angina, and asymptomatic patients.

Resting Blood Pressure



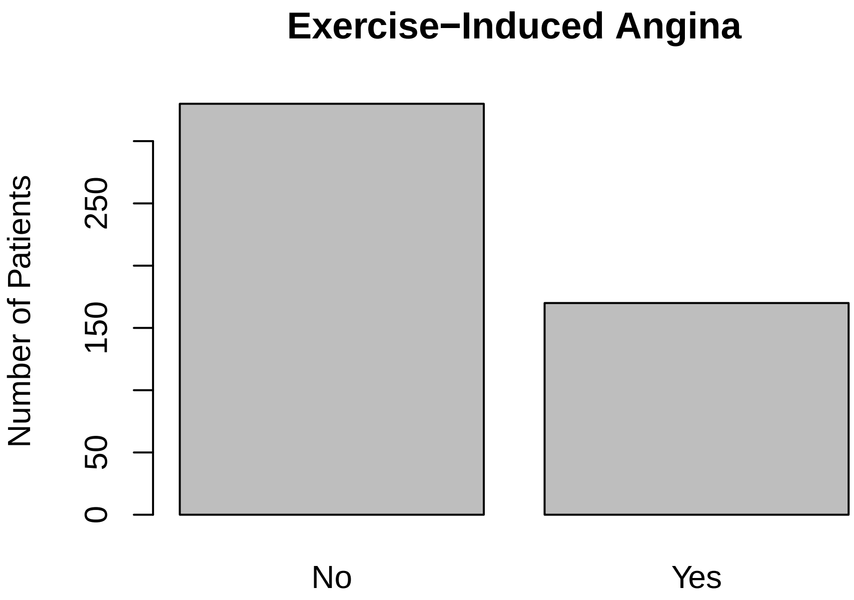
The boxplot (outliers excluded) indicates that median resting blood pressure is around 130 and the data is evenly spread.

Fasting Blood Sugar



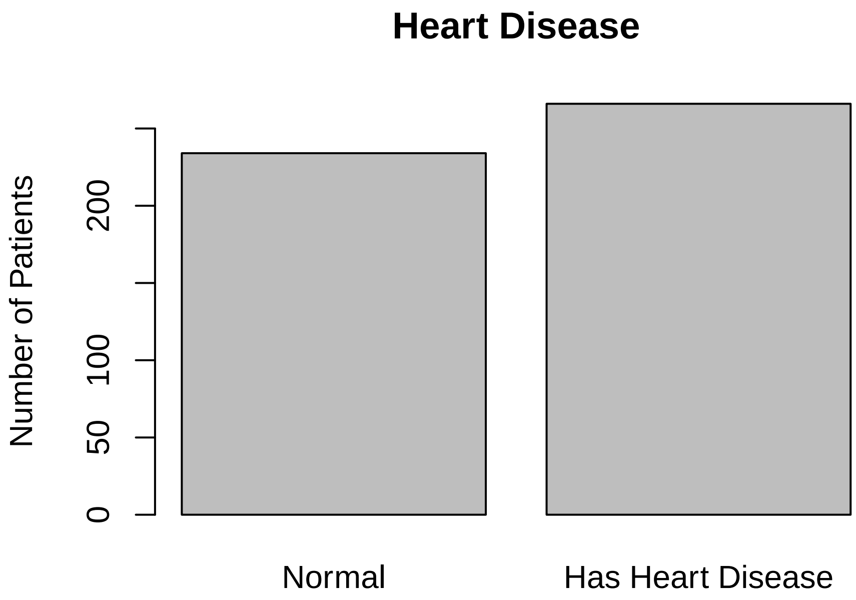
The majority of patients had a fasting blood sugar of less than 120 mg/dl.

Exercise-Induced Angina



Fewer patients experienced exercise-induced angina than not.

Target



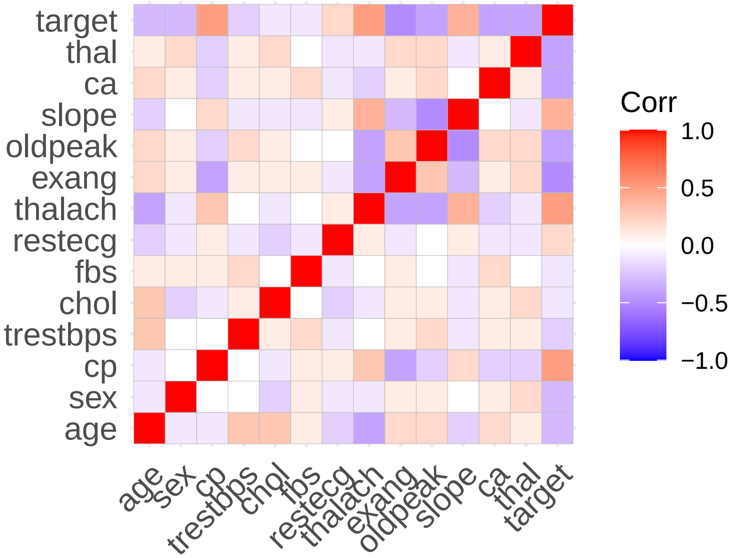
Roughly half of the patients were found to have heart disease and half were not.

Relationships

Correlation

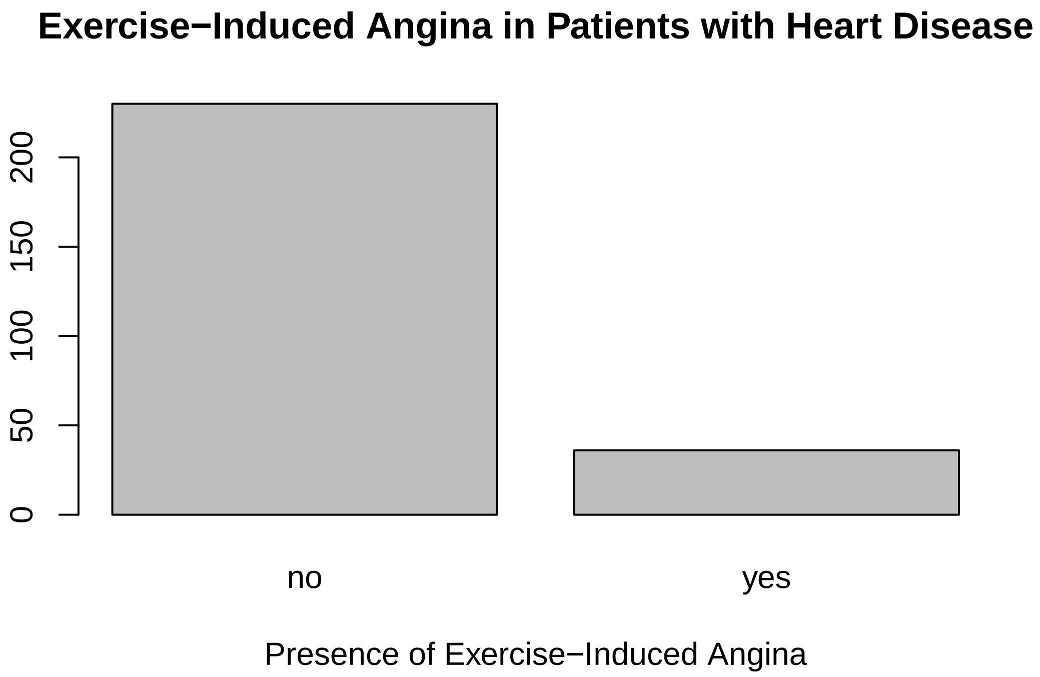
The following correlation matrix represent the relationship between different variables and the interaction between each other.

Each cell represents a correlation coefficient, the red color, which has a coefficient of 1, represents the strong positive correlation between two variables; whereas, the white color with coefficient of 0 indicates little (no) correlation between two variables, and the purple color with coefficient of -1 indicates a strong negative correlation.



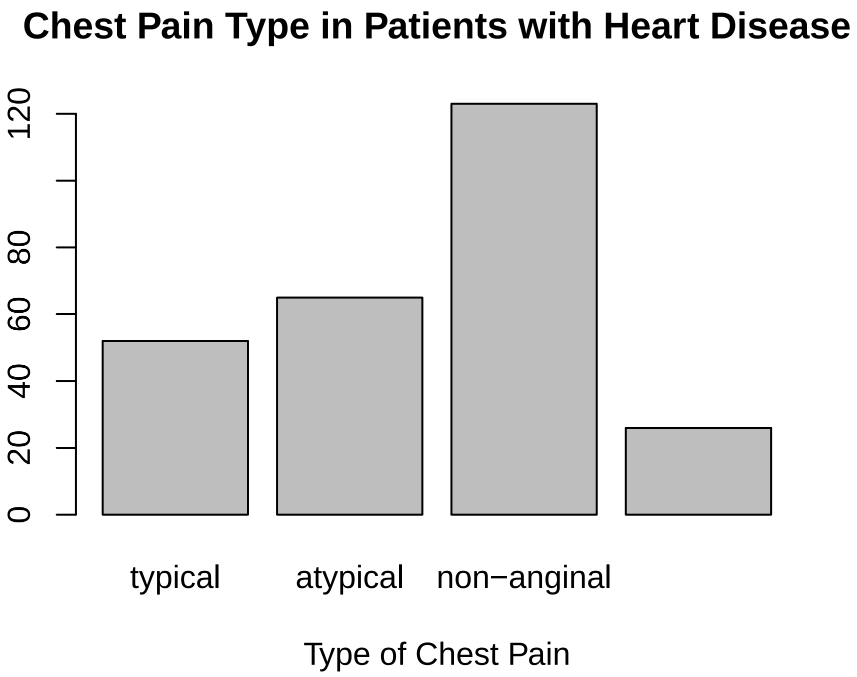
The correlation matrix of numeric data shows that there isn't a strong positive correlation between any two variables, and between maximum heart rate achieved and target; a relatively strong negative correlation between ST slope and oldpeak; and little correlation between sex and cheast pain type, cholesterol and fasting blood sugar.

Exercise-Induced Angina vs. Target



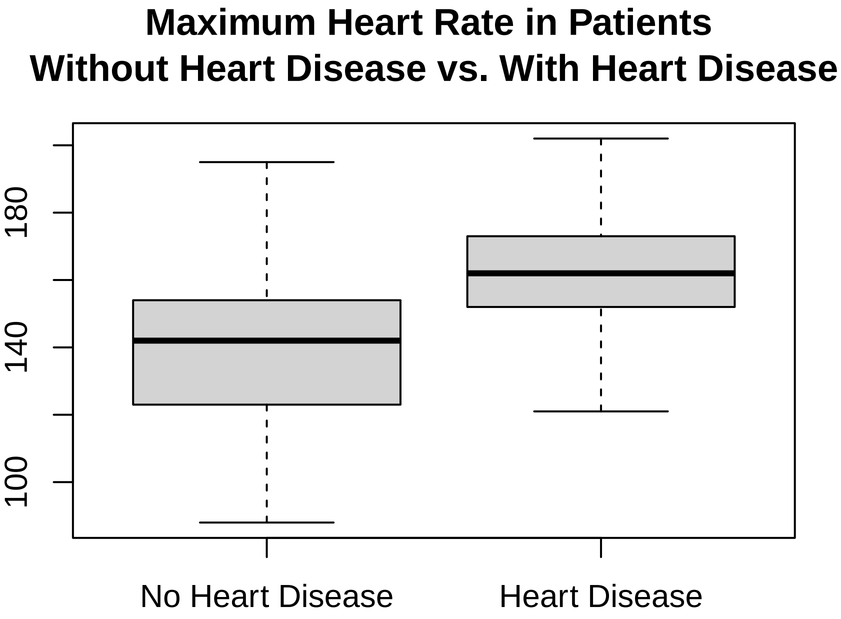
Of patients with heart disease, the majority did not experience exercise-induced angina. This could imply that exercise-induced angina is not a major factor in predicting heart disease.

Chest Pain Type vs. Target



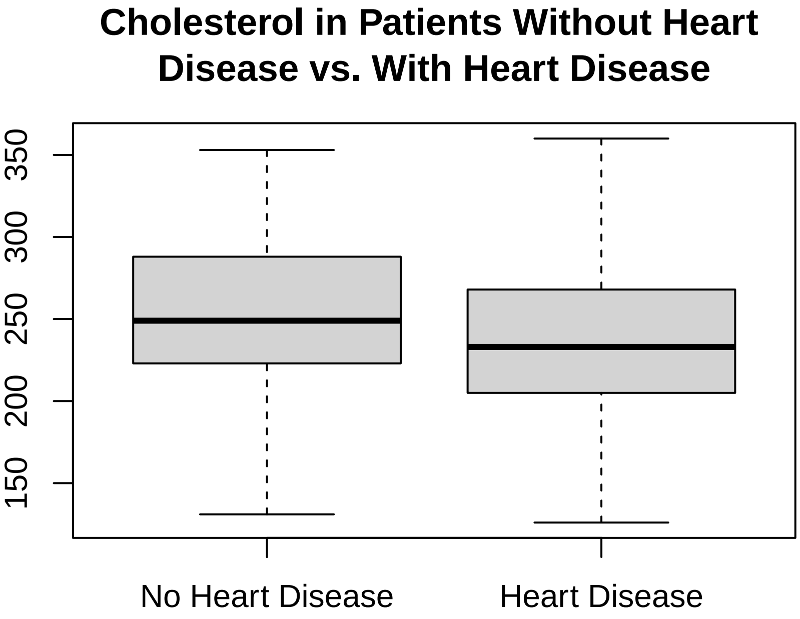
Most patients with heart disease experienced non-anginal pain followed by atypical angina, typical angina, and no symptoms.

Maximum Heart Rate vs. Target



When controlled for outliers, patients with heart disease tended to have higher maximum heart rates than patients without, indicating that maximum heart rate could be important in detecting heart disease.

Cholesterol vs. Target



When controlled for outliers, cholesterol levels appear similar in patients with and without heart disease. In those with heart disease, cholesterol levels appear more evenly spread than those without heart disease.