Project Step 1

Minu Pabbathi, Ziqian Zhao, Paul Zhang

2024-04-24

About the data

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It originally contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. 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 `chest.pain.type` 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 `resting.bp.s` variable is a numeric variable that states the resting blood pressure in mmHg.

The `cholesterol` variable is a numeric data which give an overview of person's cholesterol levels. The `fasting blood sugar` 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 `resting electrocardigram` 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 `maximum heart rate` 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 binary data "yes" or "no". 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 `ST 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

Age

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   47.00   55.00   54.02   60.25   77.00
```

The ages of the patients range from 29 to 77. The average age is 54.02 and the median age is 55.

Resting Blood Pressure

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   120.0   130.0   132.5   140.0   200.0
```

The resting blood pressure ranges from 94 to 200. The average resting blood pressure is 132.5 mmHg and the median is 130 mmHg.

Cholesterol

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   211.0   243.0   248.1   282.0   564.0
```

Resting Electrocardiogram

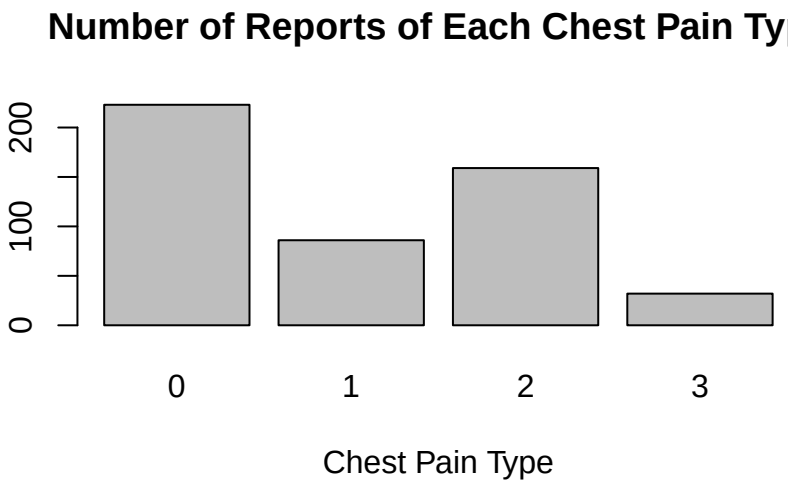
Nominal	0	1	2
Meaning	normal	abnomality in ST-T waves	Show left ventricular hypertrophy
No. of people	263	233	4

Max Heart Rate

This data is a numeric dataset which records the maximum heart rate achieved.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   137.0   152.0   150.2   168.0   202.0
```

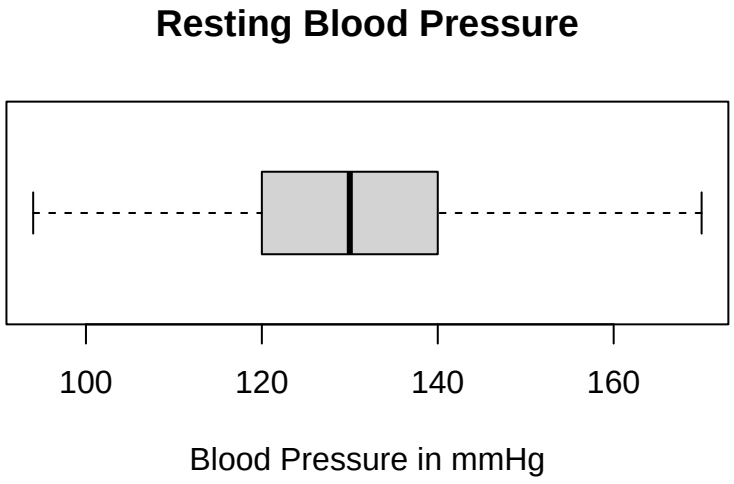
Chest Pain Type



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

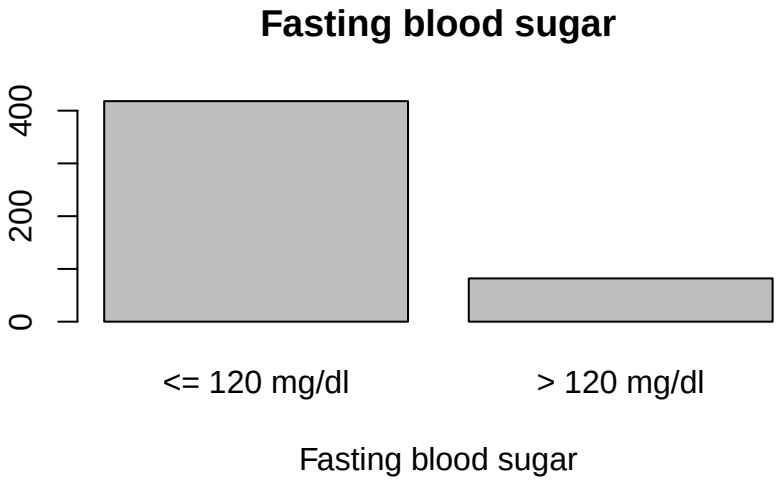
The barplot indicates that majority of patients experienced typical angina, followed by non-anginal pain, atypical angina, and asymptomatic patients.

Resting Blood Pressure

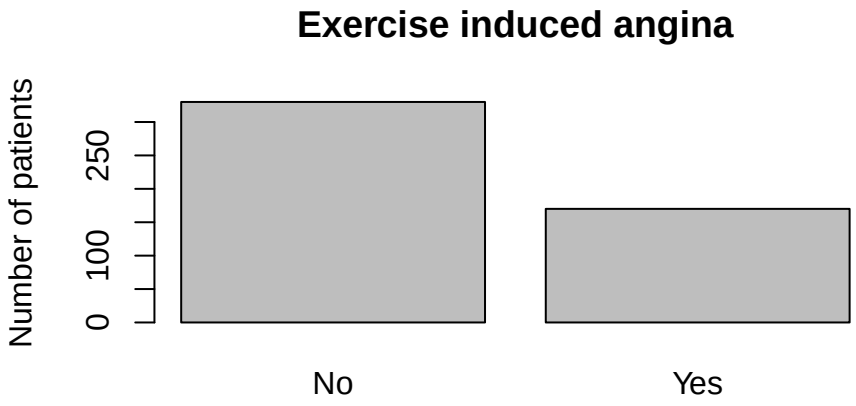


The boxplot (outliers excluded) indicates that median resting blood pressure is around 130 and the data is evenly spread.

Fasting Blood Sugar



Exercise induced angina



Target