Heart Disease Science

Finn B

1/17/2021

Contents

1 Introduction

The purpose of this report is the analysis and methodology of several health data of patients from 1988. The data shows if a patient has heart disease. It describes a range of conditions that affect the heart. The data set used is a data set provided by **Donald Bren School of Information and Computer Sciences** from the **University of California, Irvine** originally. This project will concentrate on a database from the **V.A. Medical Center, Long Beach and Cleveland Clinic Foundation** provided created by **Robert Detrano, M.D., Ph.D.**. The data was sourced from Kaggle, where the data was initially processed.
Origin of this database: Archive.ics.uci

2 Data exploration and cleaning

2.1 Data exploration

This report excludes 62 attributes from the original database to work with a subset of 14 attributes, containing **13 features** and **one outcome variable** to consider if a patient has heart disease. The database contains health data of **303 patients**.
On a first view you can see what features will accompany the final outcome variable in this project. Before heading into the analysis we need to understand what the different attributes tell us:

```
## # A tibble: 6 x 14
##     age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak slope
##   <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl> <dbl>
## 1    63     1     3      145   233     1       0     150     0     2.3     0
## 2    37     1     2      130   250     0       1     187     0     3.5     0
## 3    41     0     1      130   204     0       0     172     0     1.4     2
## 4    56     1     1      120   236     0       1     178     0     0.8     2
## 5    57     0     0      120   354     0       1     163     1     0.6     2
## 6    57     1     0      140   192     0       1     148     0     0.4     1
## # ... with 3 more variables: ca <dbl>, thal <dbl>, target <dbl>
```

Attribute	Meaning
age	Patients age (29-77 years)
sex	Female (0) and Male (1)
cp - **chest pain type**	asymptomatic (0); atypical angina (1); non-anginal pain (2); typical angina (3)
trestbps - **resting blood pressure**	in mm/Hg on admission to the hospital[1]
chol - **serum cholesterol**	in mg/dl
fbs - **fasting blood sugar**	> 120 mg/dl; no(0) yes(1)
restecg - **resting electrocardiographic results**	probable or definite left ventricular hypertrophy by Estes' criteria(0); normal(1); having ST-T wave abnormality(2)
thalach	maximum heart rate achieved
exang - **exercise induced angina**	no(0); yes(1)
oldpeak	ST depression induced by exercise relative to rest
slope - **slope of peak exercise ST segment**	downsloping(0); flat(1); upsloping(2)
ca - **number of major vessels colored by flourosopy**	vessels(0-3)
thal - **Thalium Stress Test Result**	null(0); fixed defect(1); normal(2); reversible defect(3)

[1] Judging from the values, the systolic pressure (the pressure when the heart pushes blood out) is given here.

2.2 Data cleaning

For data cleaning some data from the original data base will be changed. The levels of 'sex' will be changed to 'female' and 'male'. The levels of 'target' will be changed to 'disease' and 'no disease' to have a quiet better overview. Furthermore some of the attributes will be encoded as factors to enable a better work. These vectors are: **sex, cp, fbs, restecg, exang, slope, ca, thal, disease(target)**.

```
str(HeartData)
```

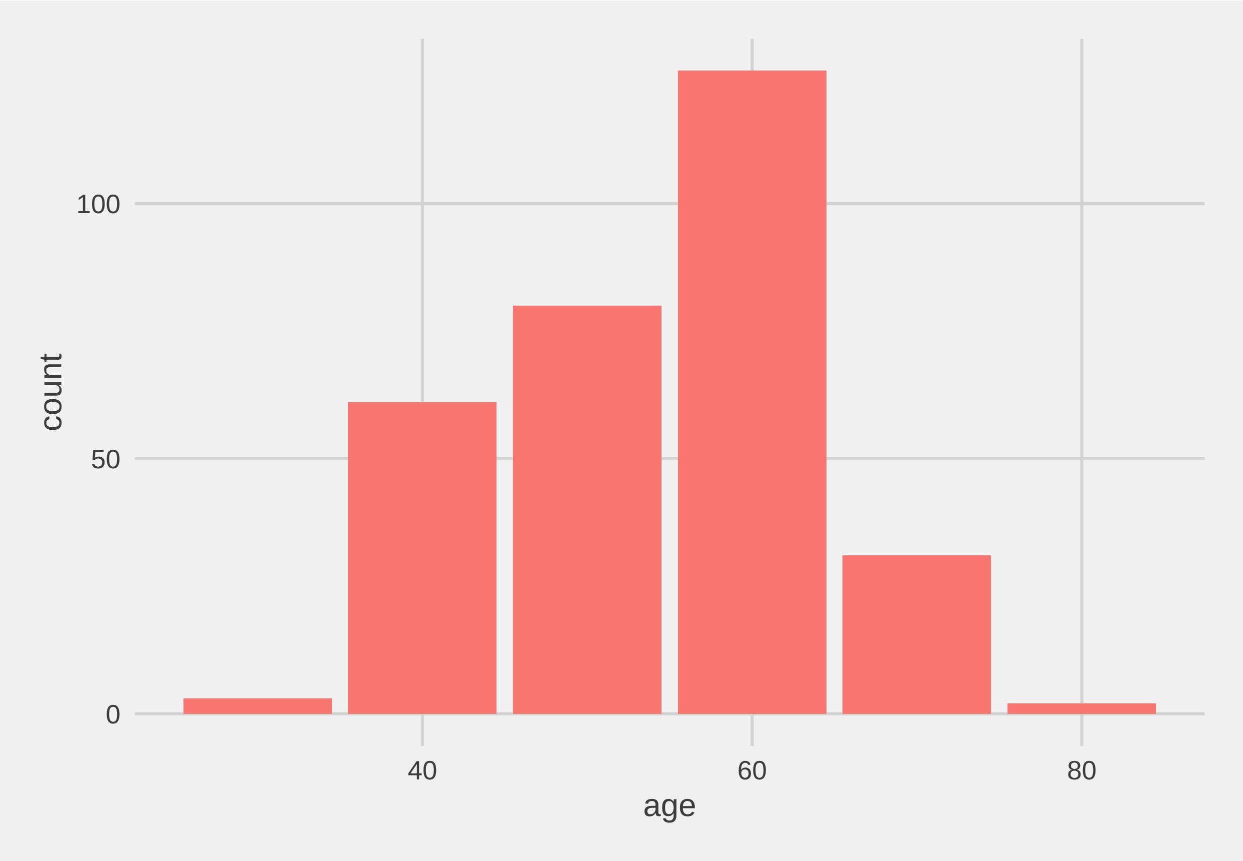
```
## tibble [303 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : Factor w/ 2 levels "female","male": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "asymptomatic",..: 4 3 2 2 1 1 2 2 3 3 ...
##  $ trestbps: num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : Factor w/ 2 levels "<=120",">120": 2 1 1 1 1 1 1 1 2 1 ...
##  $ restecg : Factor w/ 3 levels "left vetricular hypertrophy",..: 1 2 1 2 2 2 1 2 2 2 ...
##  $ thalach : num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
##  $ exang   : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope   : Factor w/ 3 levels "downsloping",..: 1 1 3 3 3 2 2 3 3 3 ...
##  $ ca      : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
##  $ thal    : Factor w/ 4 levels "null","fixed defect",..: 2 3 3 3 3 2 3 4 4 3 ...
##  $ disease : Factor w/ 2 levels "disease","no disease": 2 2 2 2 2 2 2 2 2 2 ...
```

3 Data analysis

In this part of the project we will dig deeper into the attributes and potential effects on the disease. But first we will have a look on the most obvious and superficial indicators: Age and Sex.

3.1 Age

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    29.0    47.5    55.0    54.4    61.0    77.0
```



The age range goes from 29 years to 77 year. The median age is at 55 years, while we can see that the most patients are between 55 and 65 years.

3.2 Sex

sex	count
female	96
male	207

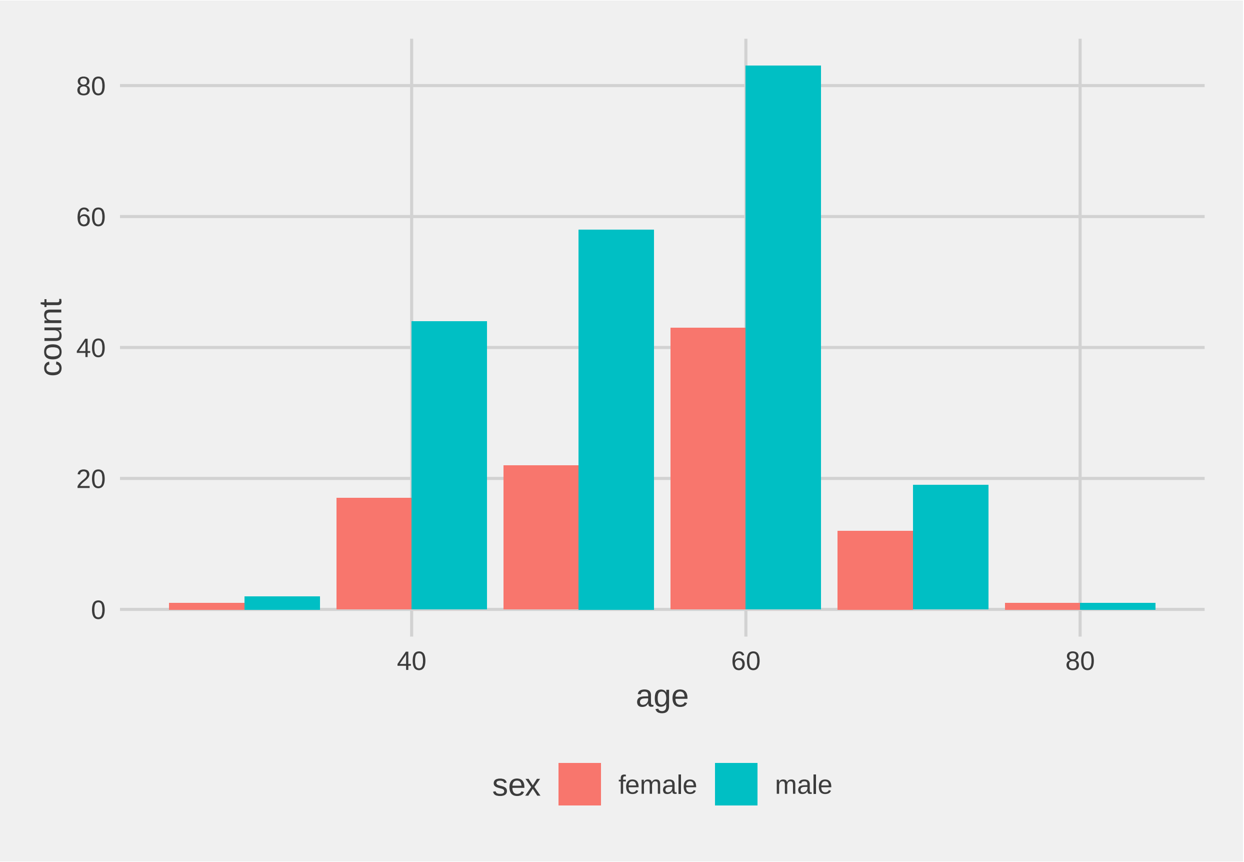
The distribution by sex is dominated by male patients, around 68% of the patients are male.
The mean age of female patients is quite lower than the mean age of male patients.

This can be seen in the mean of patients that have heart diseases as well:
female:

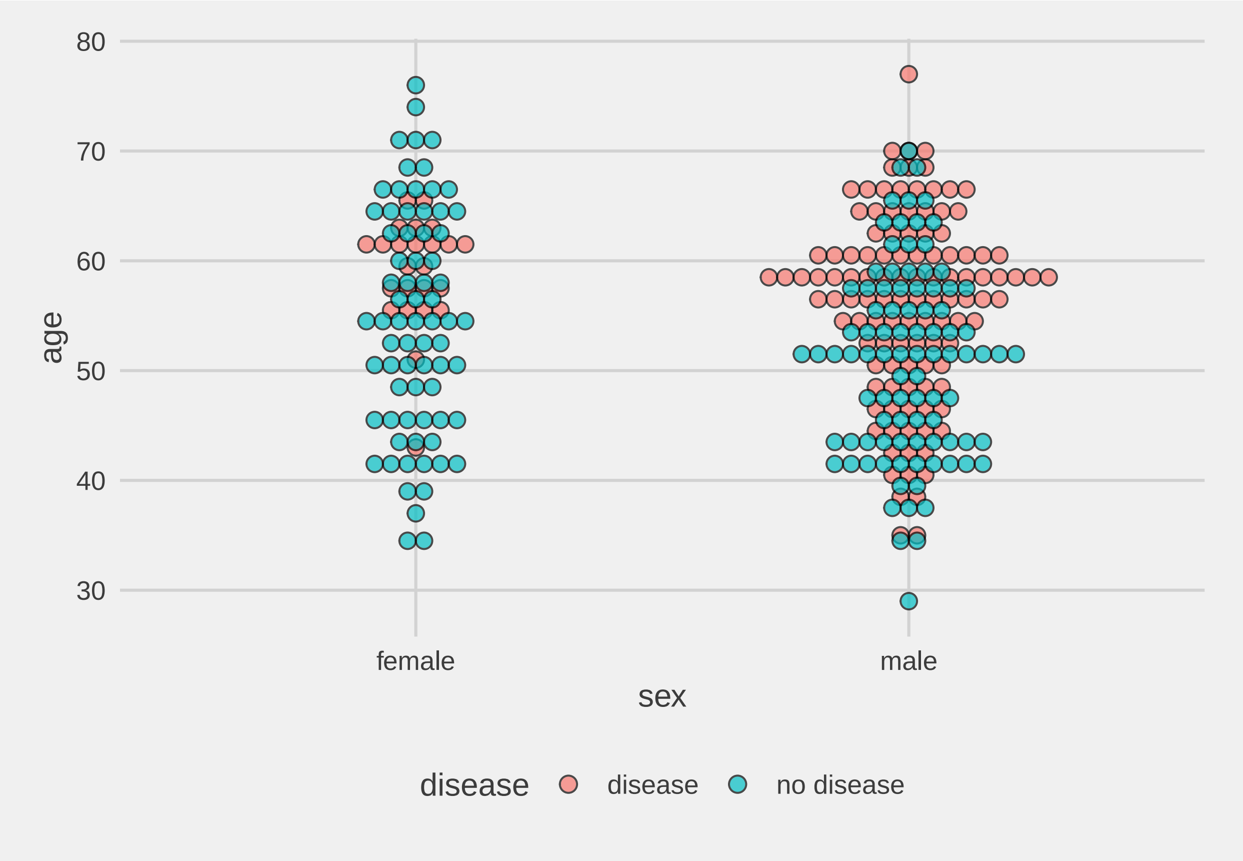
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    43.0    56.8    60.5    59.0    62.0    66.0
```

male:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    35.0    51.0    57.5    56.1    61.0    77.0
```

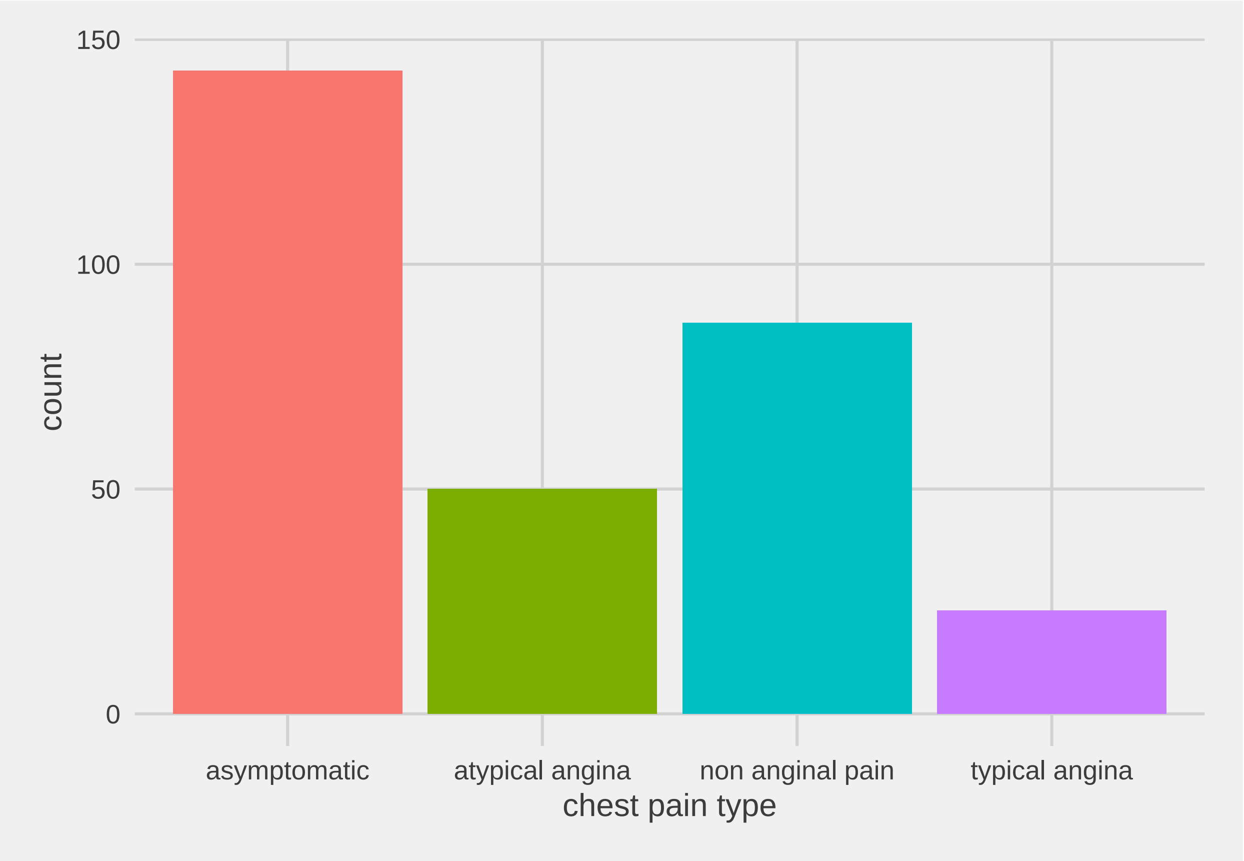


As we can see in the following plot there are much more female patients without heart disease than with heart disease. Furthermore the number of male patients with and without diseases seem to be similar while there are more male patients around 60 with heart disease.



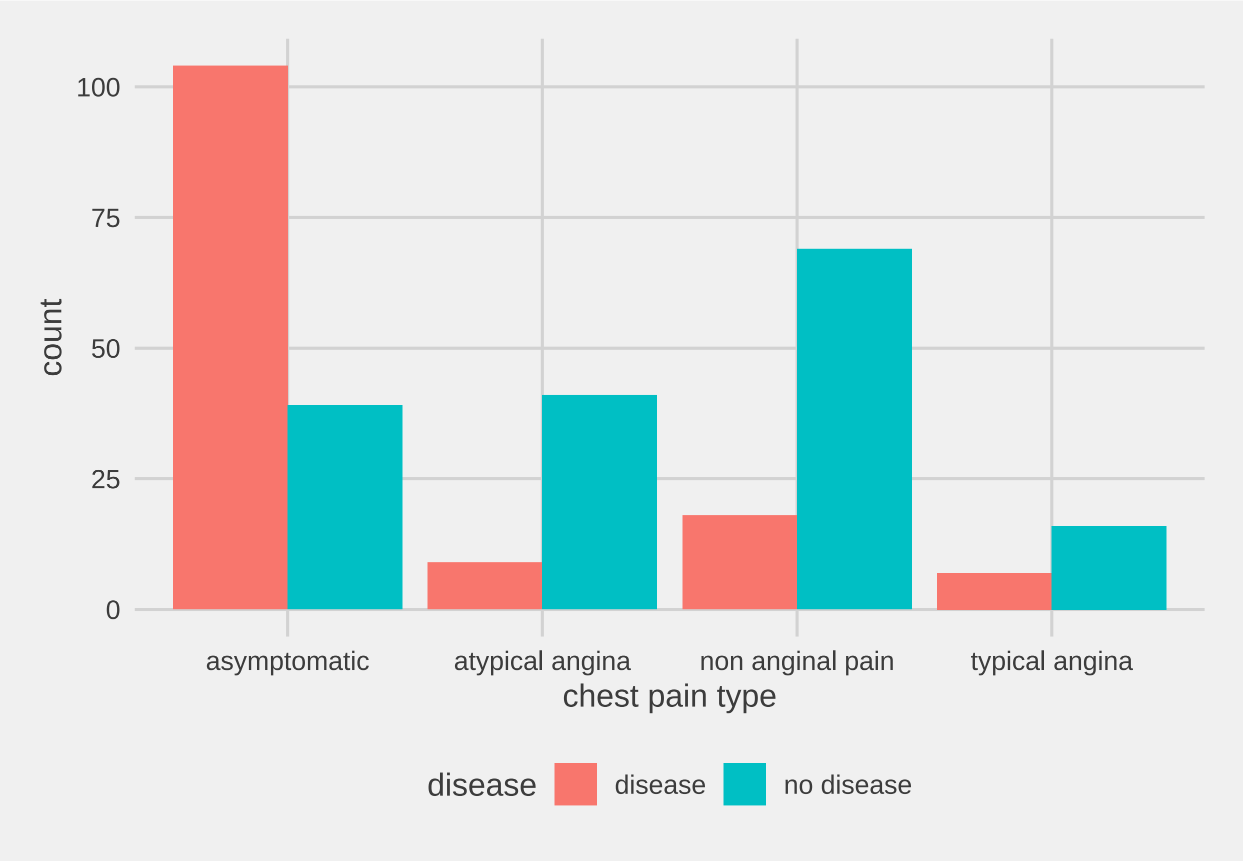
3.3 Chest pain type

As previously researched, there are four types of chest pain. **Asymptomatic** pain means that a patient has no symptoms/pain. Angina is a pain that the patient has near the heart, often described as chest tightness. Angina pain is categorized into **atypical** angina and **typical** angina.
Most patients data shows asymptomatic and non anginal pain. Only a small amount of patients had typical angina:



Something that may not have been expected by many is the following result. Except of asymptomatic pain the proportion of patients with disease is far below 50% for each category of pain.

chest pain	disease (prop)
asymptomatic	0.727
atypical angina	0.180
non anginal pain	0.207
typical angina	0.304

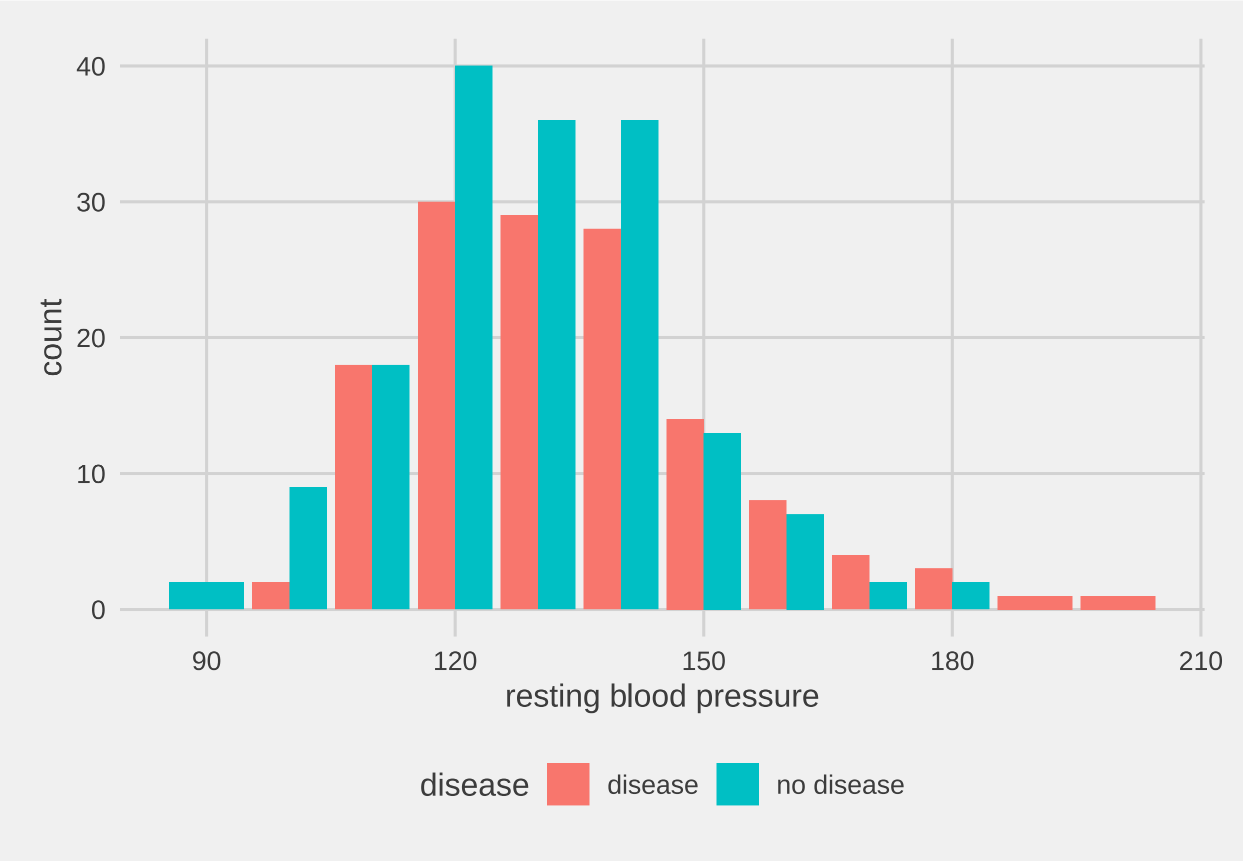


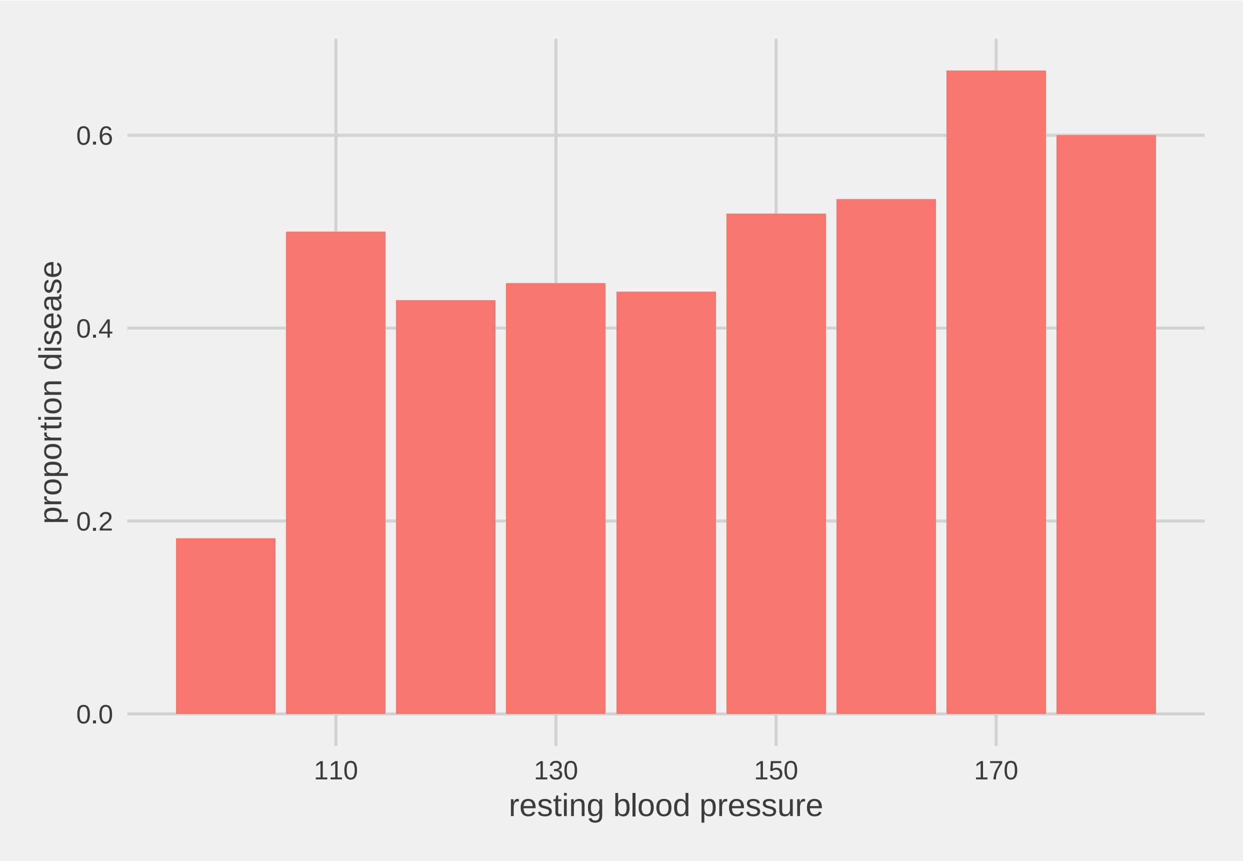
3.4 Resting blood pressure

For most patients (68%) the blood pressure is higher than the ideal systolic blood pressure of between 90 and 120mm/Hg. In the second plot we can observe an increasing proportion of disease with a higher systolic resting blood pressure.

```
HeartData %>%
  summarize(mean(trestbps>120))
```

```
## # A tibble: 1 x 1
##   ‘mean(trestbps > 120)‘
##                    <dbl>
## 1                  0.680
```

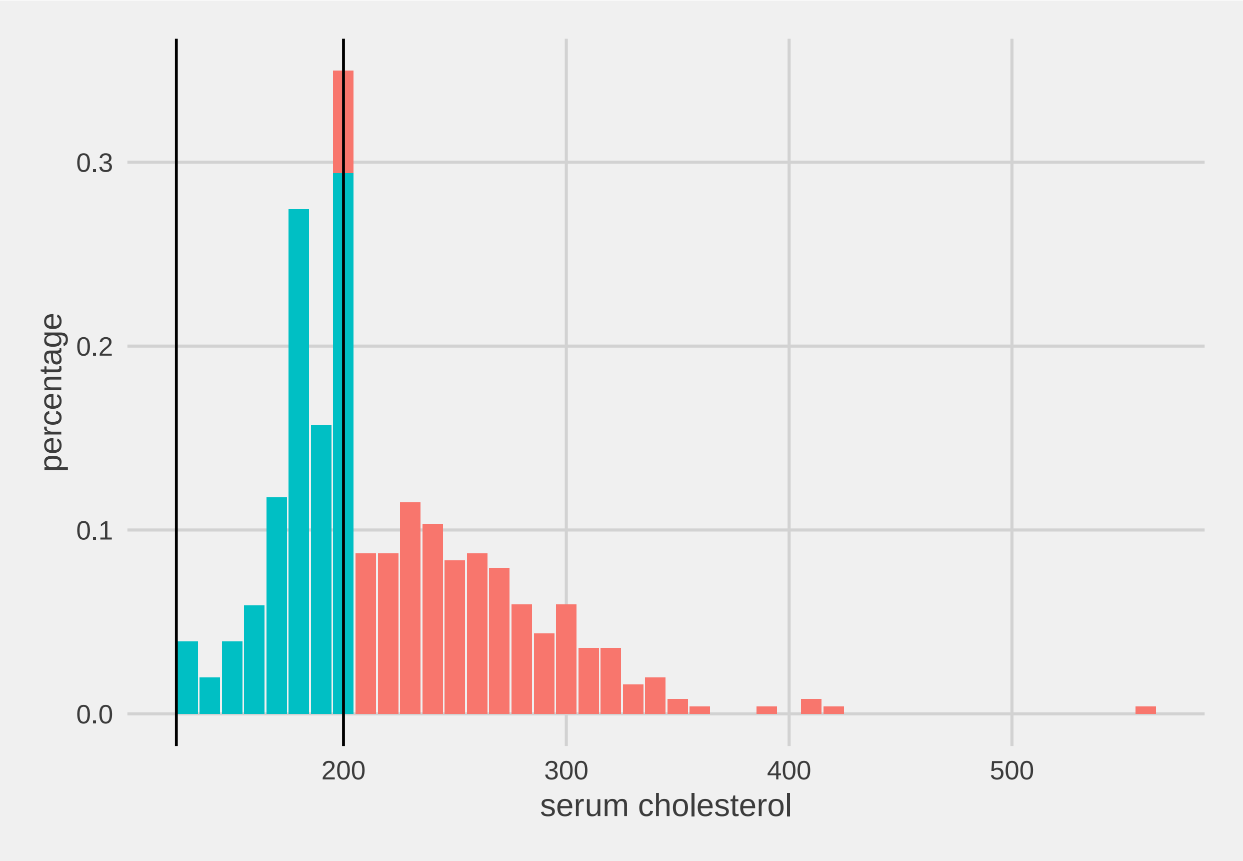

40
30
20
10
0
count
90
120
150
180
210
resting blood pressure
disease
disease
no disease



3.5 Serum cholesterol

Since there is too little information about what type of cholesterol level is given we assume total cholesterol. Healthy cholesterol level for adults is between 125mg/dL and 200mg/dL.

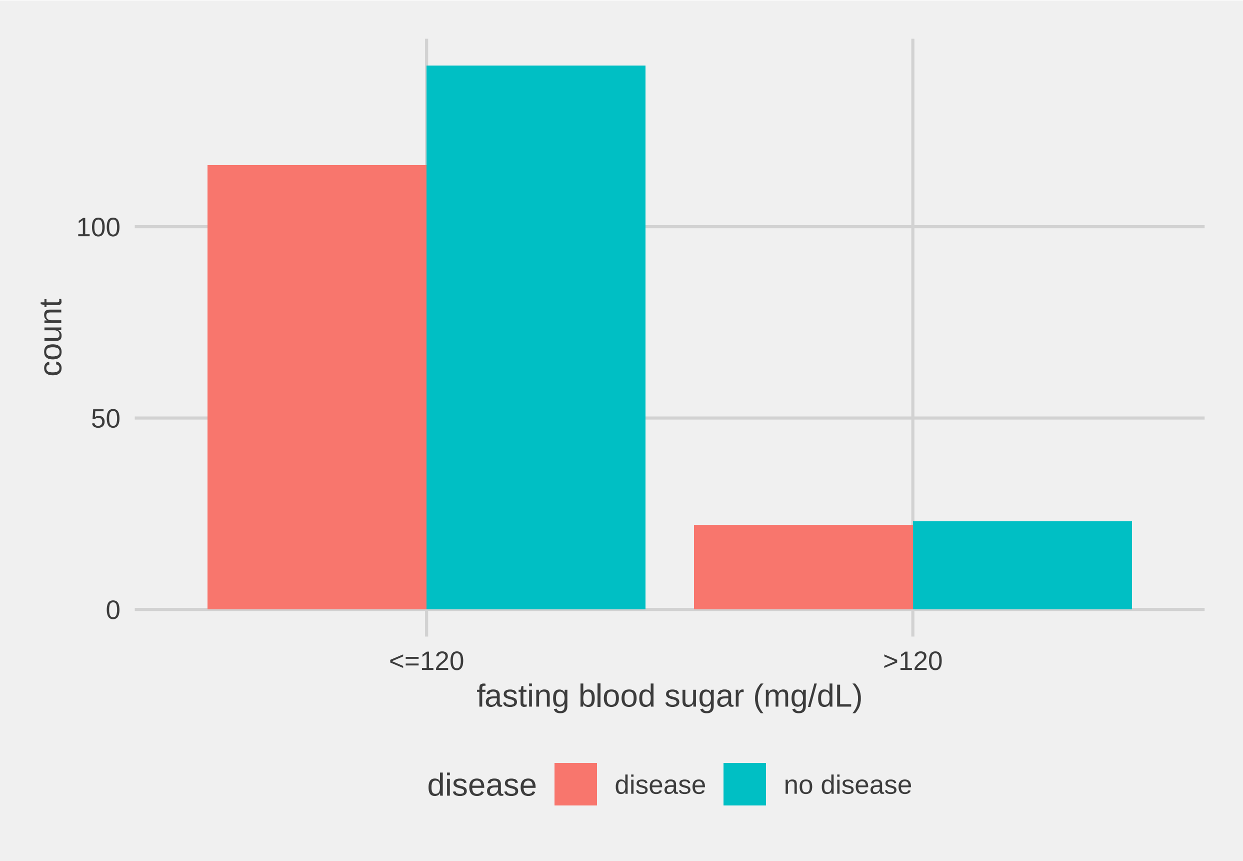
Most patients serum cholesterol is higher than the healthy range:



3.6 Fasting blood sugar

Normal blood sugar levels of **non-diabetic** people are between **72mg/dL** and **99mg/dL** when fasting. Fasting blood sugar levels of **100mg/dL** up to **125mg/dL** are already described as **prediabetic**, while **fbs** $>$ **125mg/dL** are diagnosed as **diabetic**.
The study shows two possible outcomes of **fbs**: **<= 120mg/dL** and **>120mg/dL**.It must be considered, that people with a **fbs >120mg/dL** are at greater risk of developing heart disease or cardiovascular disease, however the symptoms of the patient may be caused by diabetes and secondary diseases.

Only a few patients have blood sugar levels in the range where diabetes would be diagnosed. The number of patients with and without disease are similar. The most patients have fasting blood sugar levels of 120 and lower.



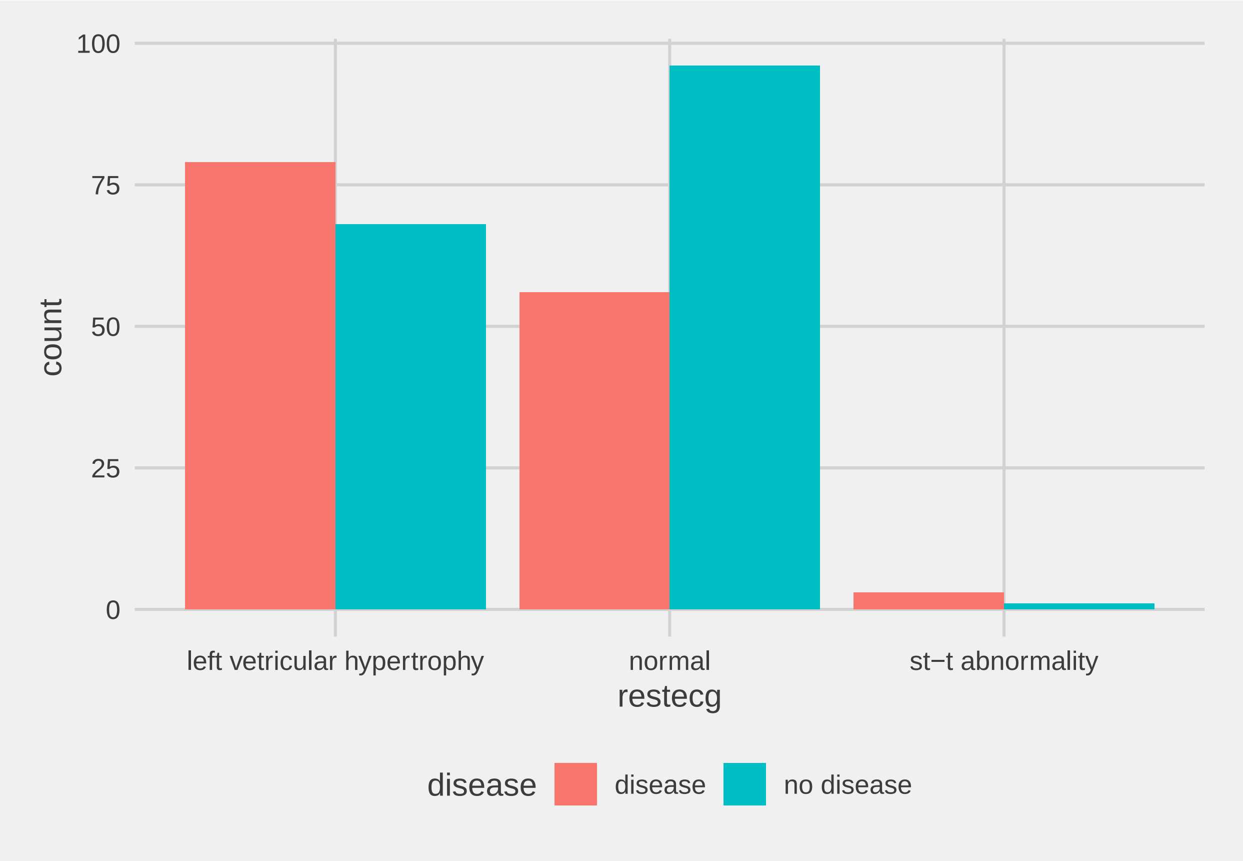
There were 14.90% of patients with a critical value of fasting blood sugar level in the database, while the prevalence of diabetes in the US was 4.90%[2] in the year 1990. So we can observe a much higher prevalence in the study from 1988.

[2] Diabetes trends in the U.S.: 1990-1998

3.7 Resting electrocardiographic results

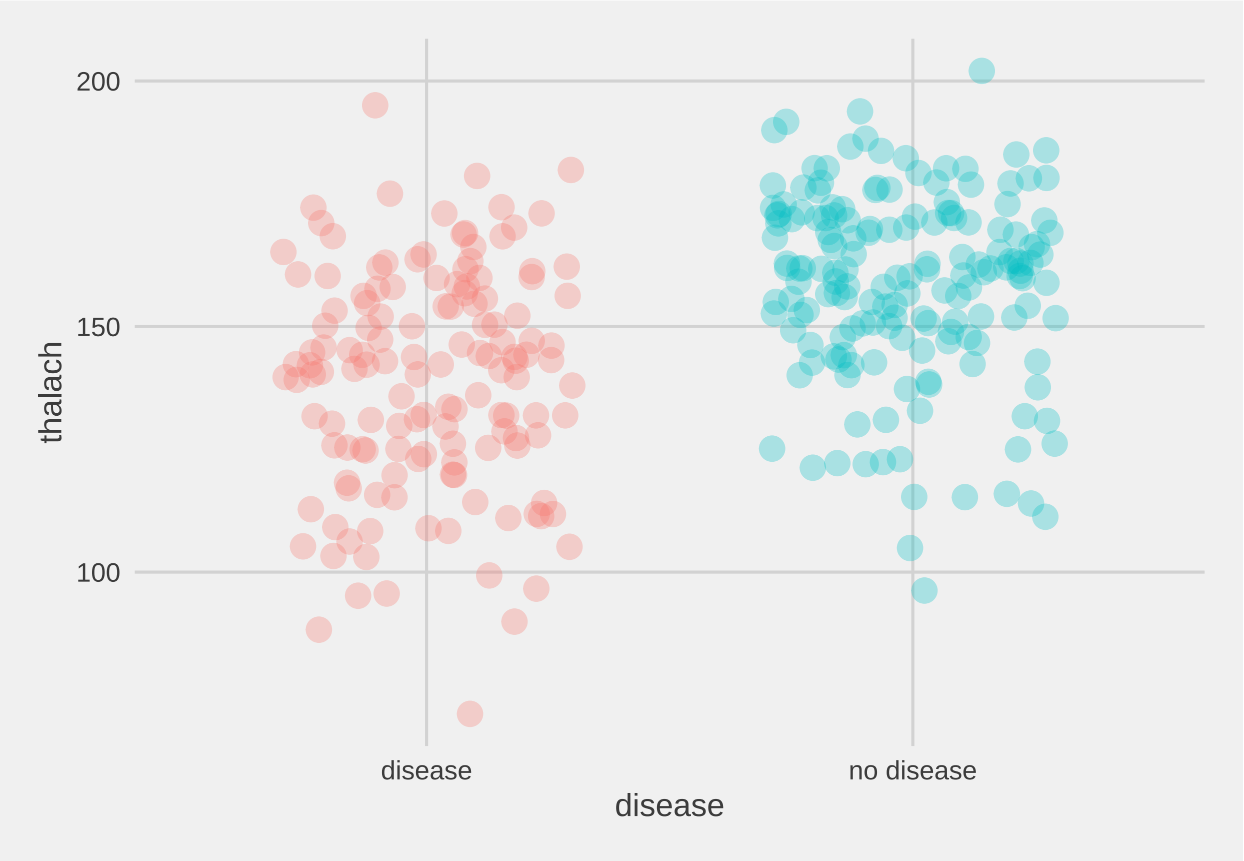
As results of the restecg there are three potential outcomes:

- **left ventricular hypertrophy:**
  Left ventricular hypertrophy is enlargement and thickening (hypertrophy) of the walls of the heart's main pumping chamber.
- **Normal:**
  No abnormalities or hypertrophies.
- **Having ST-T wave abnormality:**
  Abnormalities of **ST-** and/or **T wave** in the imaging procedures of the electrocardiogram.



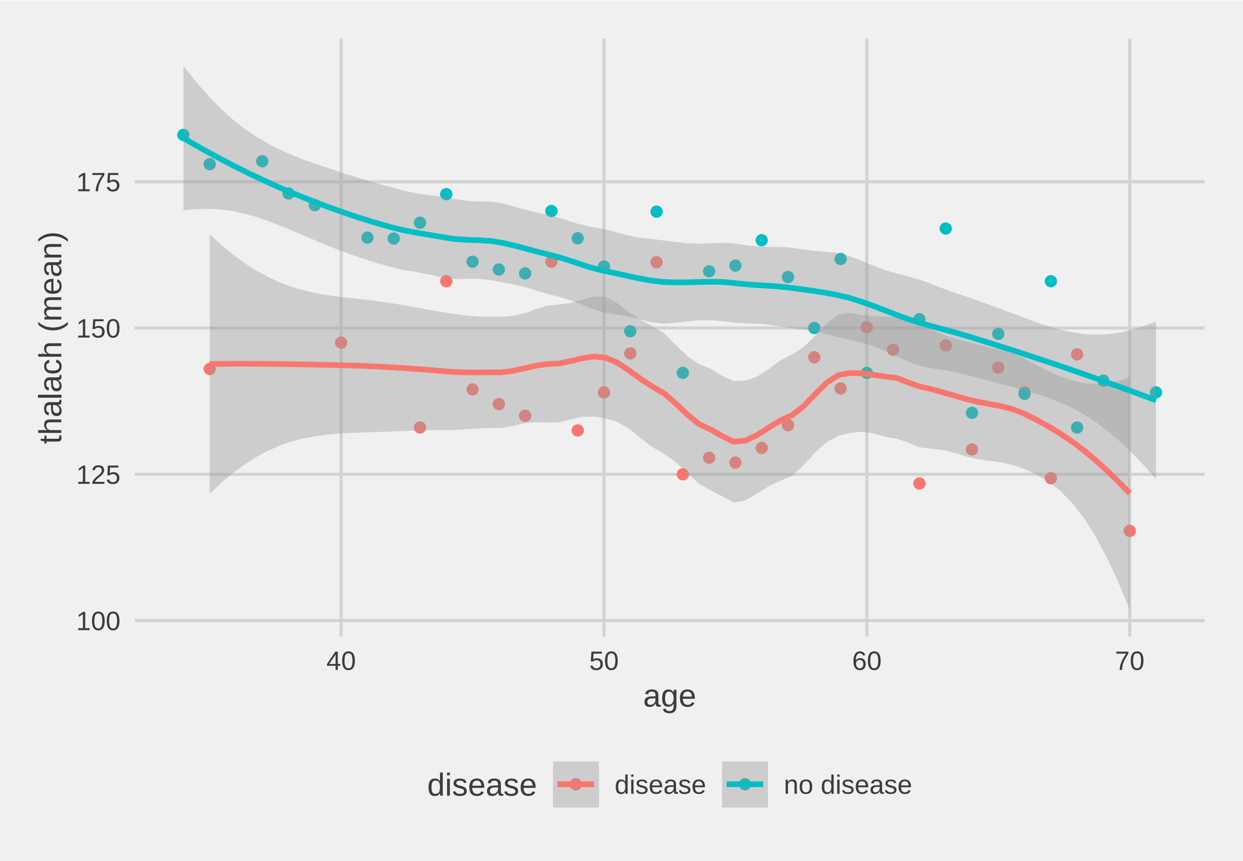
3.8 THALACH

THALACH is the **maximum heart rate** that has been achieved of each patient.
We observe a lower maximum heart rate for patients with disease than without disease:



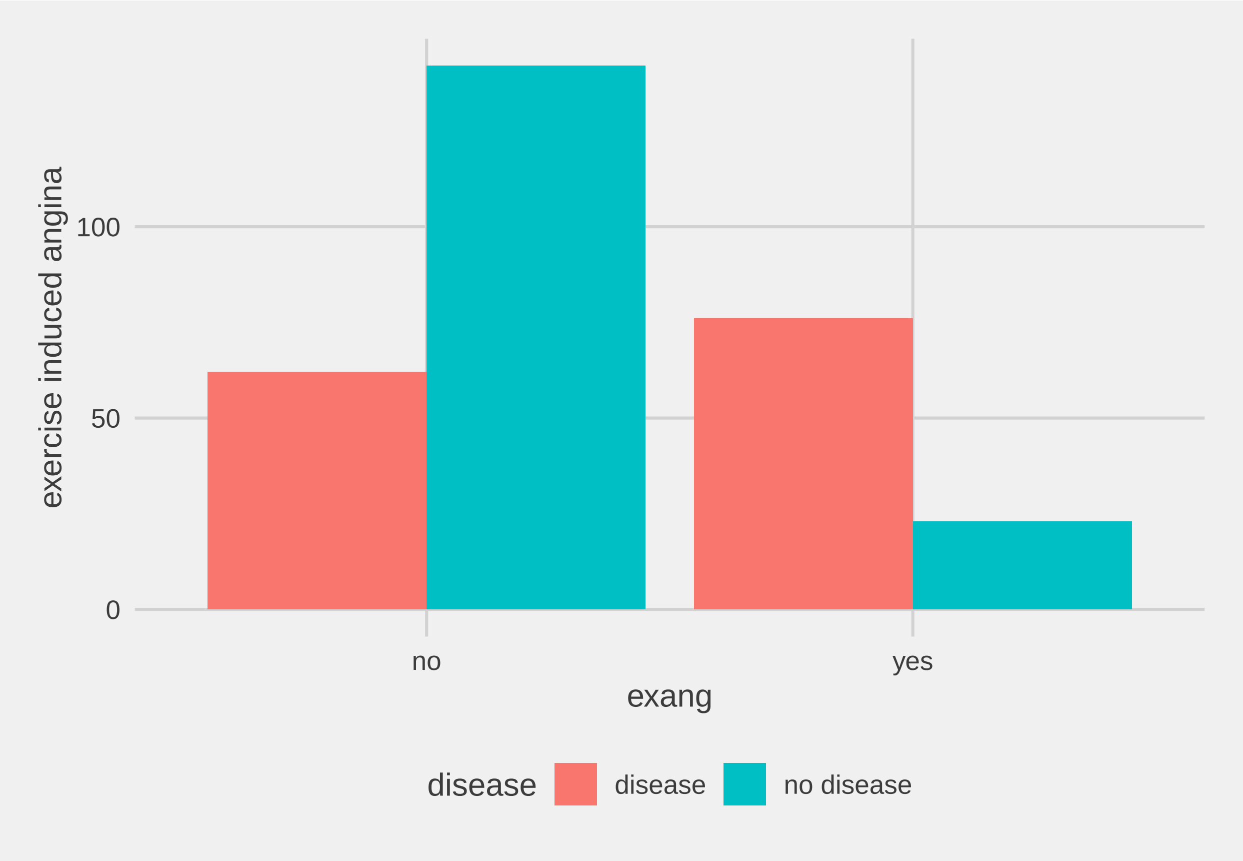
disease	mean	median
disease	139	142
no disease	158	161

As you can see in the next chart the average maximum heart rate decreases with age. An interesting abnormality is that patients with heart disease show a lower maximum heart rate at almost any age.



3.9 Exercise induced angina

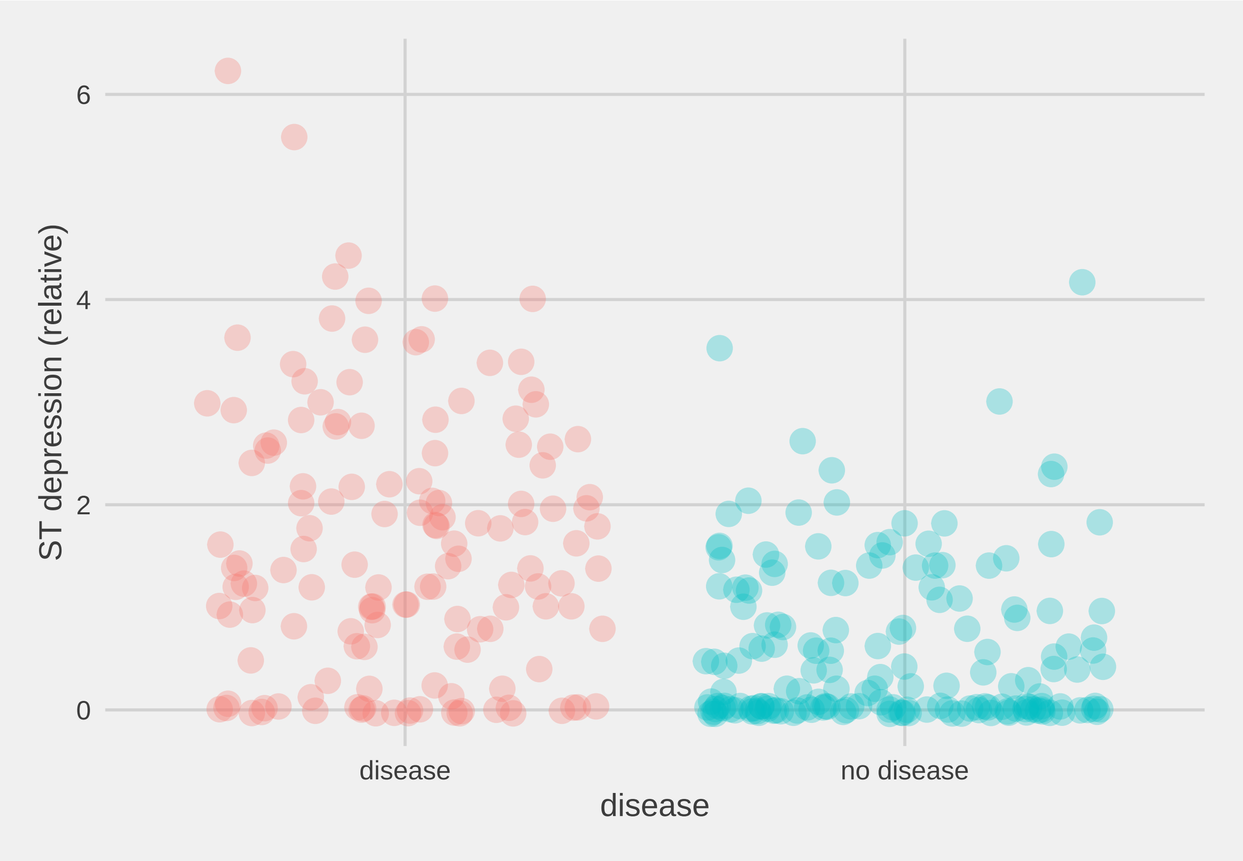
We can assume that angina indicates heart disease at exercise more often than without. We can tell by the fact that most patients with exercise induced angina had heart disease but only a minority of patients with heart disease had angina outside the exercises, most were asymptomatic[3].



[3] 3.3 chest pain type and disease

3.10 ST depression induced by exercise relative to rest

We can say that a greater ST depression is a sign of an increased probability of heart disease. The following findings from the database show, that the ST depression increase at exercise for patients with heart disease is greather than for patients without heart disease:



disease	mean	median
disease	1.586	1.4
no disease	0.583	0.2

Mean and median show increased values of ST depression relation in people with heart disease than in people without heart disease.

3.11 Slope of peak exercise ST segment

3.12 Major vessels colored by flourosopy

3.13 Thalium Stress Test Result

4 Methods

5 Results

6 Conclusion