Simple Exploration of Heart Disease Data

Miftaj Chowdhury, Abhi Adusumilli, Brady Seyler

Wednesday, December 17, 2025

Introduction

The goal of this project is to explore patterns related to heart disease using patient health data. We focus on understanding what factors are associated with heart disease. The data set is formatted similarly to the UCI Heart Disease data set. Each row represents a patient, and variables include age, sex, cholesterol level, resting blood pressure, maximum heart rate, and a binary heart disease indicator.

Inserting the Data

```
# A tibble: 6 x 14
     X1    X2    X3    X4    X5    X6    X7    X8    X9   X10   X11 X12   X13
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr>
1    63     1     1   145   233     1     2   150     0   2.3     3 0.0   6.0
2    67     1     4   160   286     0     2   108     1   1.5     2 3.0   3.0
3    67     1     4   120   229     0     2   129     1   2.6     2 2.0   7.0
4    37     1     3   130   250     0     0   187     0   3.5     3 0.0   3.0
5    41     0     2   130   204     0     2   172     0   1.4     1 0.0   3.0
6    56     1     2   120   236     0     0   178     0   0.8     1 0.0   3.0
# i 1 more variable: X14 <dbl>
```

We first started by inserting the data into R. This is what the data frame first looked liked before we tidied it.

```
Rows: 219
Columns: 14
$ age      <dbl> 63, 67, 37, 41, 56, 57, 53, 57, 56, 44, 52, 57, 48, 54, 48, 4~
$ sex      <fct> Male, Male, Male, Female, Male, Female, Male, Male, Female, M~
```

```
$ cp       <fct> 1, 4, 3, 2, 2, 4, 4, 4, 2, 2, 3, 3, 2, 4, 3, 2, 1, 1, 2, 3, 3~
$ trestbps <dbl> 145, 120, 130, 130, 120, 120, 140, 140, 140, 120, 172, 150, 1~
$ chol     <dbl> 233, 229, 250, 204, 236, 354, 203, 192, 294, 263, 199, 168, 2~
$ fbs      <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0~
$ restecg  <dbl> 2, 2, 0, 2, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 0, 0~
$ thalach  <dbl> 150, 129, 187, 172, 178, 163, 155, 148, 153, 173, 162, 174, 1~
$ exang    <fct> 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
$ oldpeak  <dbl> 2.3, 2.6, 3.5, 1.4, 0.8, 0.6, 3.1, 0.4, 1.3, 0.0, 0.5, 1.6, 1~
$ slope    <fct> 3, 2, 3, 1, 1, 1, 3, 2, 2, 1, 1, 1, 3, 1, 1, 1, 2, 1, 2, 2, 1~
$ ca       <dbl> 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ thal     <dbl> 6, 7, 3, 3, 3, 3, 7, 6, 3, 7, 7, 3, 7, 3, 3, 3, 3, 3, 3, 3, 3~
$ target   <fct> No Disease, Disease, No Disease, No Disease, No Disease, No D~
```

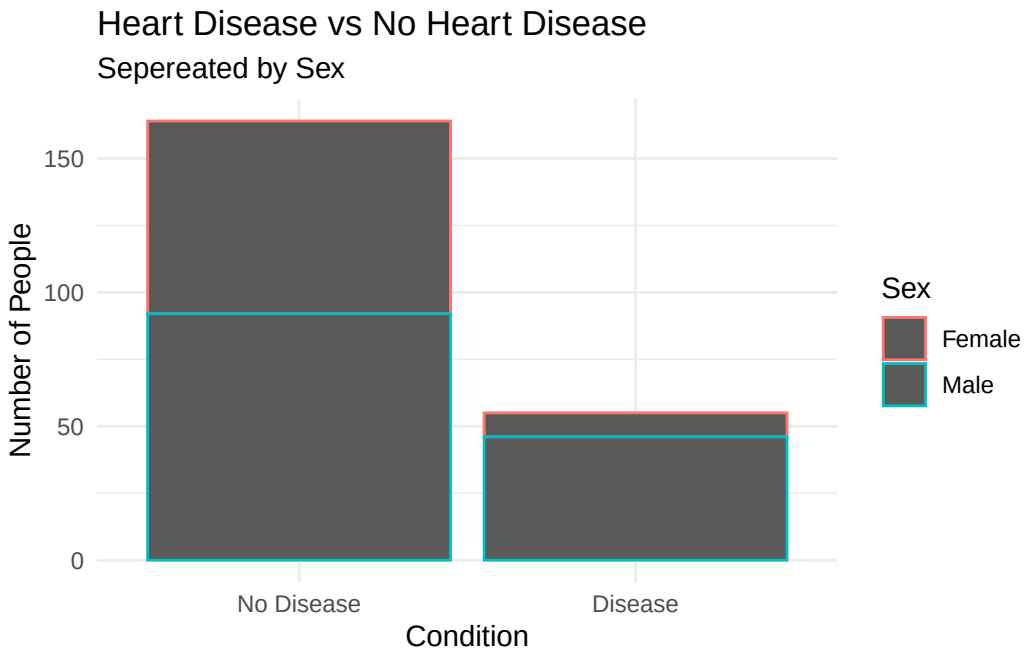
We then tidied the data, which included removing all cases where the patient wasn't labeled as either having or not having heart disease. Additionally, we changed the output of sex and target from 0 or 1 to Female or Male and No Disease or Disease respectively. This is a look at the tidied data.

```
# A tibble: 1 x 3
  average_age average_cholesterol average_resting_bp
        <dbl>               <dbl>              <dbl>
1        53.3                244.               130.
```

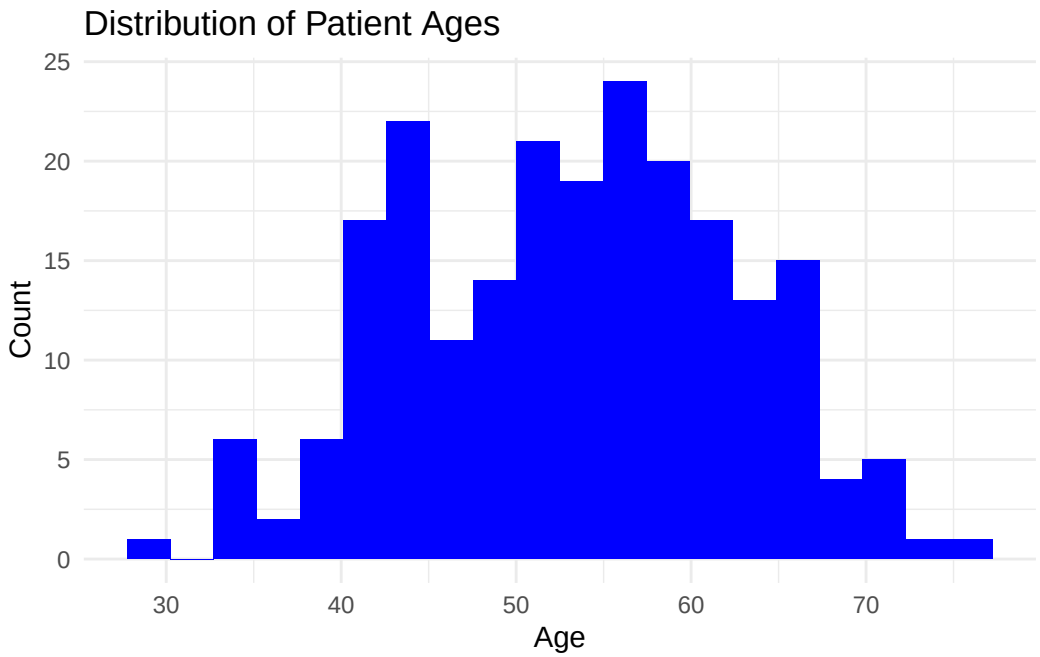
```
# A tibble: 4 x 3
  sex    target         n
  <fct>  <fct>      <int>
1 Female No Disease    72
2 Female Disease        9
3 Male   No Disease    92
4 Male   Disease       46
```

These two tables show summary statistics of the data, including average age, cholesterol, and resting blood pressure. It also shows how many men and women in the data did and did not have heart disease. These tables will become more relevant with the visualizations below.

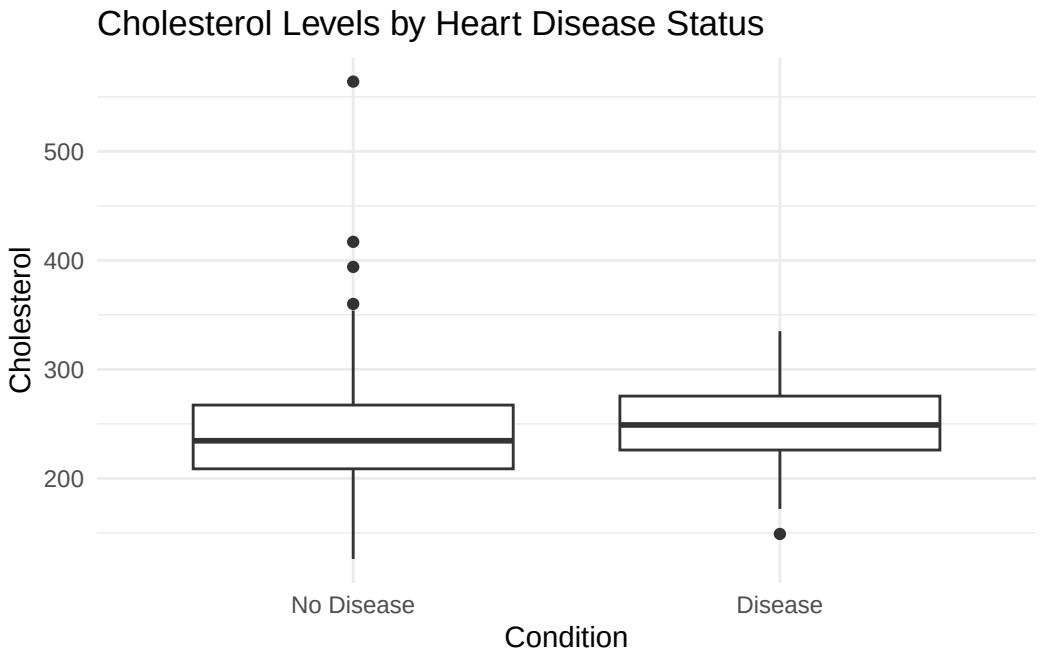
Visualizations



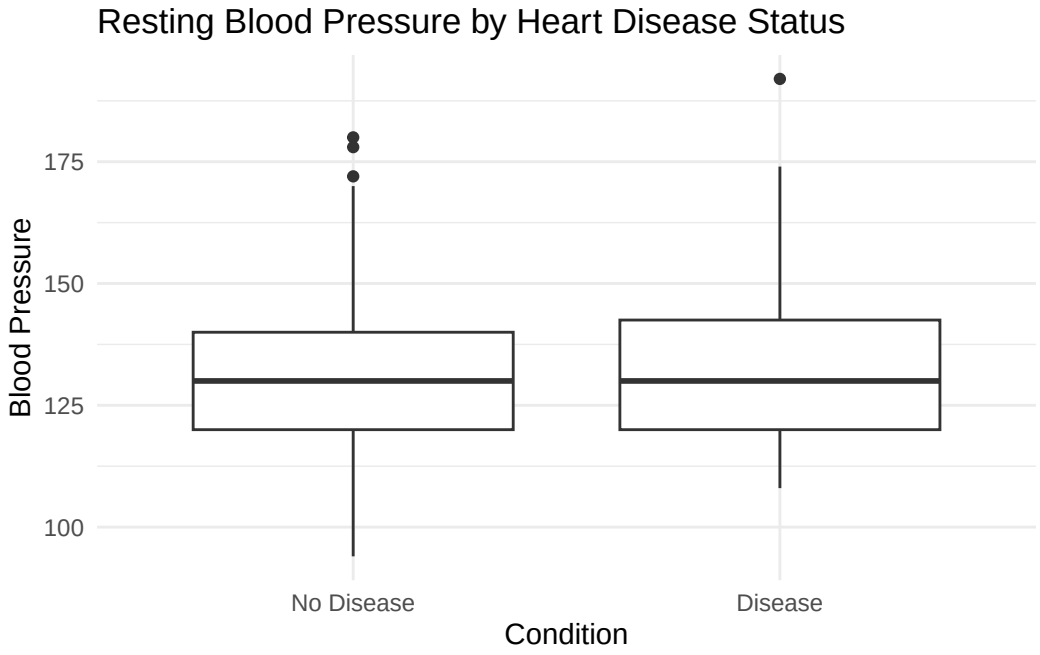
This plot shows the number of people in the data set that had and did not have heart disease. It additionally shows the sex proportions between who did and did not have heart disease. From the graph you can see that there are significantly more men that had heart disease than women in the data, which coincides with the table that was shown earlier.



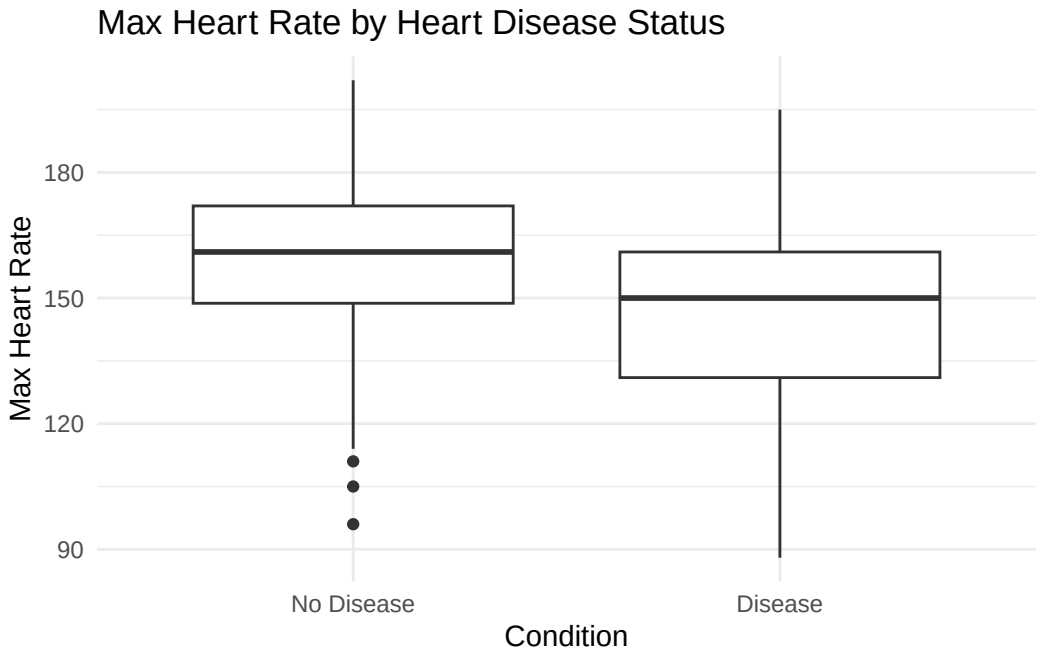
Knowing the average age of the people in the data set is important when interpreting the data, which is what this histogram shows us. The range of ages spans from about 30 to 75. The distribution of the ages is bimodal, with the two peaks being around 45 and 55. This information will help bring context to future visualizations by knowing the ages of the people in the data set.



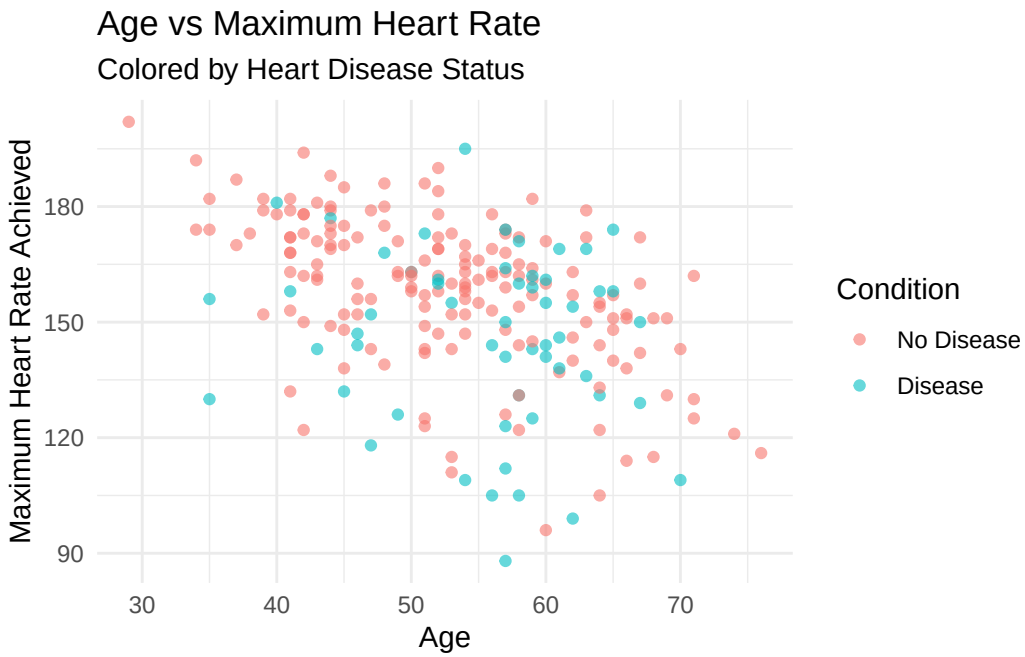
This side by side box plot shows the difference in cholesterol levels for people with and without heart disease. The average cholesterol levels for people with heart disease seems to be higher than the levels for people without. While we can't confirm exact causation with just this plot, there is certainly a possibility that higher cholesterol levels could be a factor in heart disease.



This side by side box plot shows the difference in resting blood pressure for people with and without blood pressure. The average blood pressure is almost identical for people with and without blood pressure, so this can not be used as an indicator for heart disease from this data set alone.



This side by side box plot shows the difference in max heart rate for people with and without heart disease. The average max heart rate for people with heart disease seems to be less than the average heart rate for people without heart disease. This should be considered when looking into factors for heart disease.



This scatter plot explores the relationship between age and maximum heart rate achieved. As age increases, maximum heart rate is expected to decrease. So when combining this with the previous plot, an increase in age could be a factor in getting heart disease, since people with heart disease had lower max heart rates than people without heart disease.

Conclusion

After creating multiple visualizations we can claim that older men with low max heart rates and high cholesterol levels are at most risk to get heart disease.

Code Appendix

```
# Loading useful packages
library(tidyverse)
library(ggplot2)
# Reading the heart data into R
heart <- read_csv("heart.csv .txt", col_names = FALSE)
# Viewing the first lines of the data set
head(heart)
```

```r
# These names come from the original UCI Heart Disease dataset.
colnames(heart) <- c(
  "age", "sex", "cp", "trestbps", "chol", "fbs",
  "restecg", "thalach", "exang", "oldpeak",
  "slope", "ca", "thal", "target"
)
#Converting the data into all numeric values
heart_clean <- heart %>%
  mutate(across(everything(), as.numeric))
#Removing unneeded rows
heart_clean <- heart_clean %>%
  filter(target <= 1)
#Fixing the names for the factor variables
heart_clean <- heart_clean %>%
  mutate(
    sex = factor(sex, labels = c("Female", "Male")),
    target = factor(target, labels = c("No Disease", "Disease")),
    cp = factor(cp),
    exang = factor(exang),
    slope = factor(slope)
  )
# Viewing the tidied data
glimpse(heart_clean)
#These give us a quick overview of the data.
heart_clean %>%
  summarize(
    average_age = mean(age),
    average_cholesterol = mean(chol),
    average_resting_bp = mean(trestbps)
  )
# Count how many males/females have disease/no disease
heart_clean %>% count(sex, target)
# Bar Graph shows information on who does and doesn't have heart disease.
ggplot(heart_clean,
       aes(
         x = target,
         colour = sex)
       ) +
  geom_bar() +
  labs(
    title = "Heart Disease vs No Heart Disease",
    subtitle = "Sepereated by Sex",
```

```
    x = "Condition",
    y = "Number of People",
    colour = "Sex",
    alt = "Bar Chart showing how many men and women have heart disease within the data."
  ) +
  theme_minimal()
# Histogram shows information on the ages of the people in the data set.
ggplot(heart_clean,
       aes(age)
       ) +
  geom_histogram(bins = 20, fill = "blue") +
  labs(
    title = "Distribution of Patient Ages",
    x = "Age",
    y = "Count",
    alt = "Histogtram showing the distribution of the ages of the people in the dataset."
    ) +
  theme_minimal()
# Box plot shows difference in Cholesterol levels for people with and without heart disease.
ggplot(heart_clean,
       aes(target, chol)
       ) +
  geom_boxplot() +
  labs(
    title = "Cholesterol Levels by Heart Disease Status",
    x = "Condition",
    y = "Cholesterol",
    alt = "Box plots showing the cholesterol levels for people with and without heart disease
    ) +
  theme_minimal()
# Box plot shows difference in resting blood pressure for people with and without heart disea
ggplot(heart_clean,
       aes(target, trestbps)
       ) +
  geom_boxplot() +
  labs(
    title = "Resting Blood Pressure by Heart Disease Status",
    x = "Condition",
    y = "Blood Pressure",
    alt = "Box plots showing the resting blood pressure for people with and without heart dis
    ) +
  theme_minimal()
```

```r
# Box plot shows difference in max heart rate for people with and without heart disease.
ggplot(heart_clean,
       aes(target, thalach)
       ) +
  geom_boxplot() +
  labs(
    title = "Max Heart Rate by Heart Disease Status",
    x = "Condition",
    y = "Max Heart Rate",
    alt = "Box plots showing the max heart rate for people with and without heart disease."
    ) +
  theme_minimal()
# Scatterplot shows relationship between age and max heart rate.
ggplot(
heart_clean,
aes(
x = age,
y = thalach,
color = target
)
) +
geom_point(alpha = 0.6) +
labs(
title = "Age vs Maximum Heart Rate",
subtitle = "Colored by Heart Disease Status",
x = "Age",
y = "Maximum Heart Rate Achieved",
color = "Condition",
alt = "Scatter plot showing age on the x-axis and maximum heart rate on the y-axis, colored b
) +
theme_minimal()
```