Simple EDA Analysis in R

2024-10-05

Scenario: Community Health Initiative

Imagine you're a health researcher working for a local government. Your city has noticed an increase in heart disease cases over the past few years and wants to implement targeted health programs to address this issue. You've been given a dataset of 30 residents who participated in a recent health screening event. This dataset includes various health metrics and whether or not each person has been diagnosed with heart disease.

Your task is to analyze this data to help inform the city's health initiatives.

Loading Necessary Libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(effsize)
library(readxl)
```

Loading the data

```
health_data<-read_xlsx(path="health_data.xlsx")
```

Summary and structure of the data

```
summary(health_data)
```

```
##    patient_id         age             bmi        blood_pressure
##  Min.   : 1.00   Min.   :33.00   Min.   :21.80   Length:30
##  1st Qu.: 8.25   1st Qu.:41.25   1st Qu.:24.55   Class :character
##  Median :15.50   Median :50.50   Median :27.00   Mode  :character
##  Mean   :15.50   Mean   :51.03   Mean   :27.08
##  3rd Qu.:22.75   3rd Qu.:60.50   3rd Qu.:29.02
##  Max.   :30.00   Max.   :70.00   Max.   :33.20
##   cholesterol       glucose       smoking_status exercise_frequency
##  Min.   :160.0   Min.   : 85.00   Min.   :0.0    Min.   :0.0
##  1st Qu.:186.2   1st Qu.: 95.75   1st Qu.:0.0    1st Qu.:1.0
##  Median :207.5   Median :109.00   Median :0.5    Median :1.5
##  Mean   :208.7   Mean   :111.40   Mean   :0.5    Mean   :1.9
##  3rd Qu.:228.8   3rd Qu.:119.50   3rd Qu.:1.0    3rd Qu.:3.0
```

```
##  Max.   :270.0   Max.   :160.00   Max.   :1.0    Max.   :5.0
##   sleep_hours    heart_disease
##  Min.   :4.000   Min.   :0.0
##  1st Qu.:6.000   1st Qu.:0.0
##  Median :6.500   Median :0.5
##  Mean   :6.433   Mean   :0.5
##  3rd Qu.:7.000   3rd Qu.:1.0
##  Max.   :8.000   Max.   :1.0
```

```
str(health_data)
```

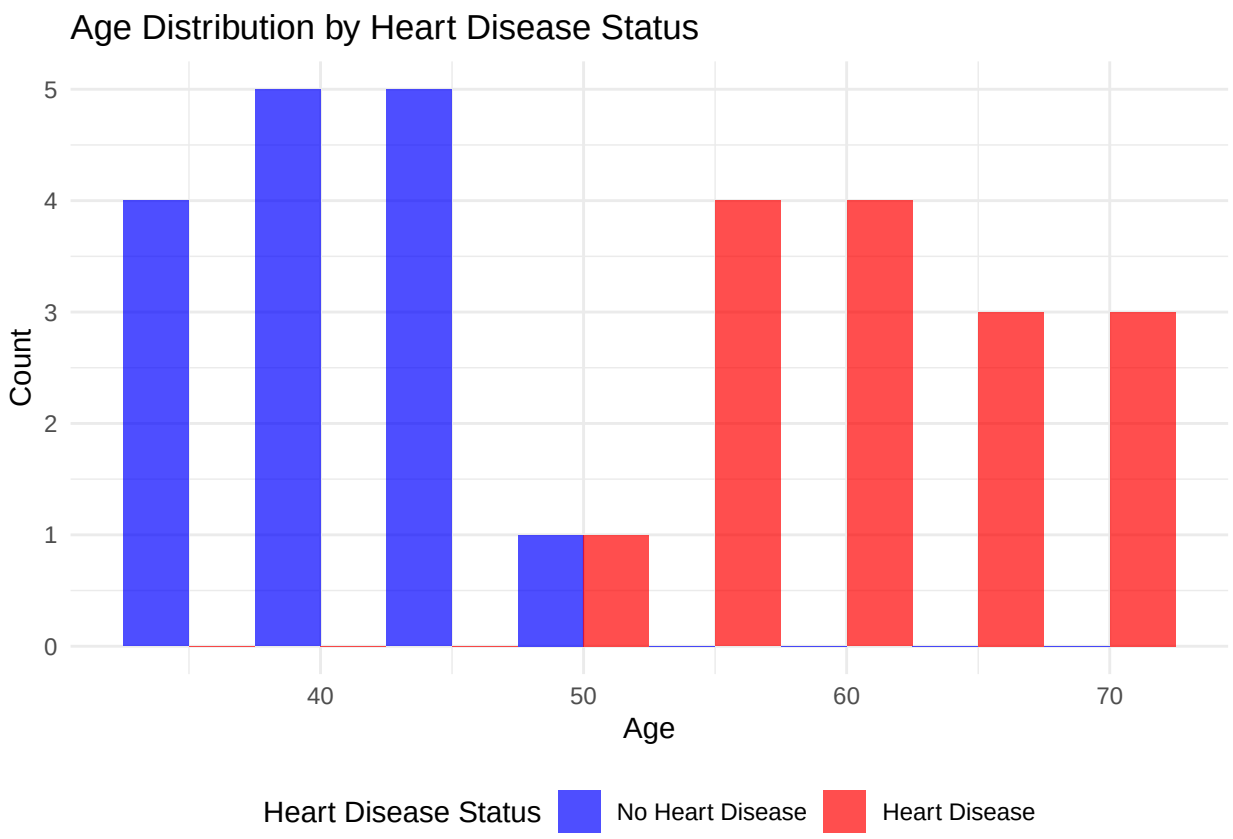
```
## tibble [30 x 10] (S3: tbl_df/tbl/data.frame)
##  $ patient_id        : num [1:30] 1 2 3 4 5 6 7 8 9 10 ...
##  $ age               : num [1:30] 45 62 38 55 41 70 33 58 49 67 ...
##  $ bmi               : num [1:30] 24.2 28.5 22.1 30.8 25.6 27.3 21.8 31.2 26.7 29.1 ...
##  $ blood_pressure    : chr [1:30] "120/80" "138/88" "118/75" "142/92" ...
##  $ cholesterol       : num [1:30] 185 210 165 240 195 220 160 250 200 230 ...
##  $ glucose           : num [1:30] 95 110 88 130 98 115 85 140 105 120 ...
##  $ smoking_status    : num [1:30] 0 1 0 1 0 1 0 1 0 1 ...
##  $ exercise_frequency: num [1:30] 3 1 4 0 2 1 5 0 2 1 ...
##  $ sleep_hours       : num [1:30] 7 6 8 5 7 6 8 5 7 6 ...
##  $ heart_disease     : num [1:30] 0 1 0 1 0 1 0 1 0 1 ...
```

1. Distribution of Age

We need this in order to:

- To understand the age distribution of our sample.
- To see if heart disease is more prevalent in certain age groups.
- To help the city decide which age groups to target for health programs.

```
ggplot(health_data, aes(x = age, fill = factor(heart_disease,
                                               levels = c(0, 1),
                                               labels = c("No Heart Disease", "Heart Disease")))) +
  geom_histogram(binwidth = 5, position = "dodge", alpha = 0.7) +
  scale_fill_manual(values = c("No Heart Disease" = "blue", "Heart Disease" = "red")) +
  labs(title = "Age Distribution by Heart Disease Status",
       x = "Age",
       y = "Count",
       fill = "Heart Disease Status") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



The histogram shows that a higher proportion of older individuals have heart disease compared to younger ones. As age increases, the red bars (representing those with heart disease) become more prominent, indicating that heart disease is more common in older age groups. This suggests a strong correlation between aging and the likelihood of heart disease.

Therefore, we should implement targeted health initiatives focusing on older adults. This could include regular health screenings, educational programs on heart disease prevention, specifically designed for older populations.

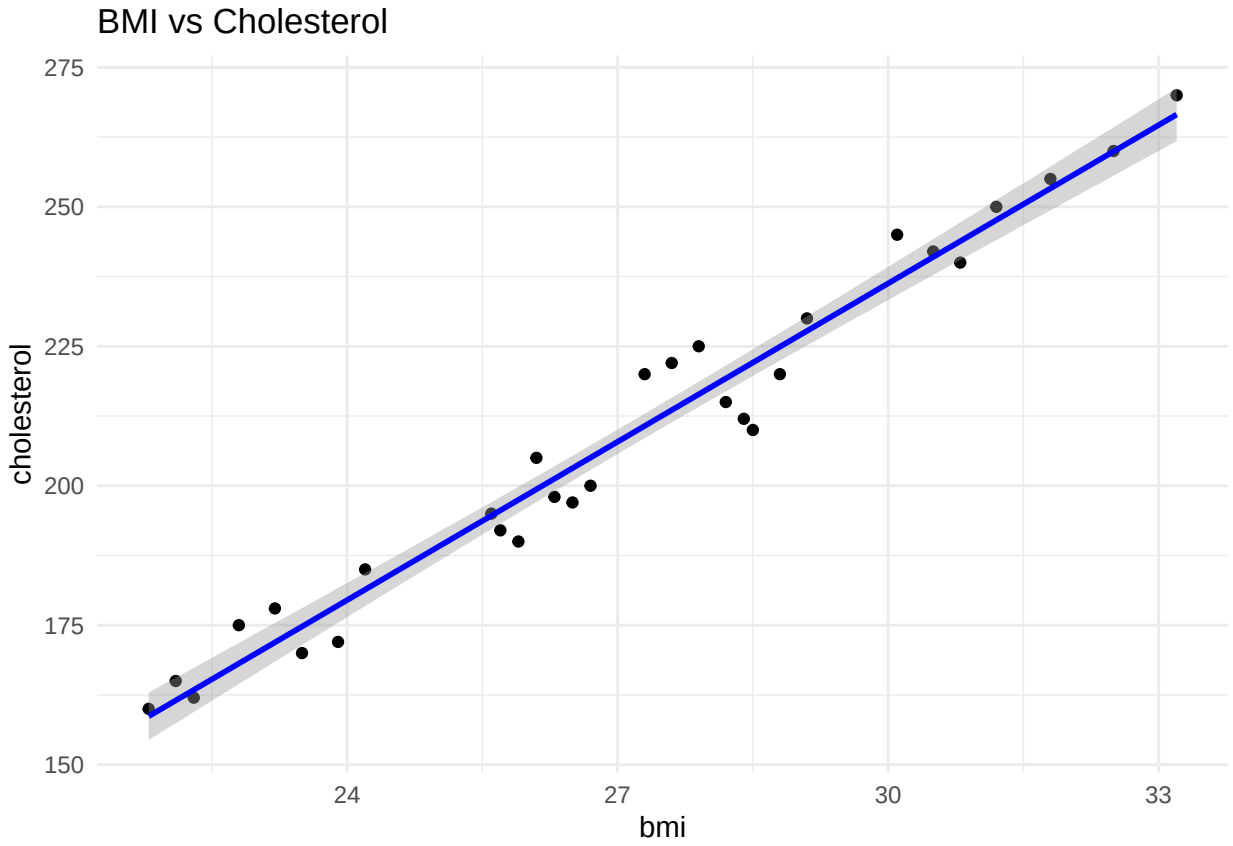
2. Scatter Plot of BMI vs. Cholesterol

We need this to:

- To visualize the relationship between BMI and cholesterol levels.

```
ggplot(health_data, aes(x = bmi, y = cholesterol)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  theme_minimal() +
  ggtitle("BMI vs Cholesterol")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



There is a clear positive relationship between BMI and cholesterol. As BMI increases, cholesterol levels also tend to increase.

The points are distributed in a pattern that suggests a nearly linear relationship, meaning that a higher BMI is associated with higher cholesterol.

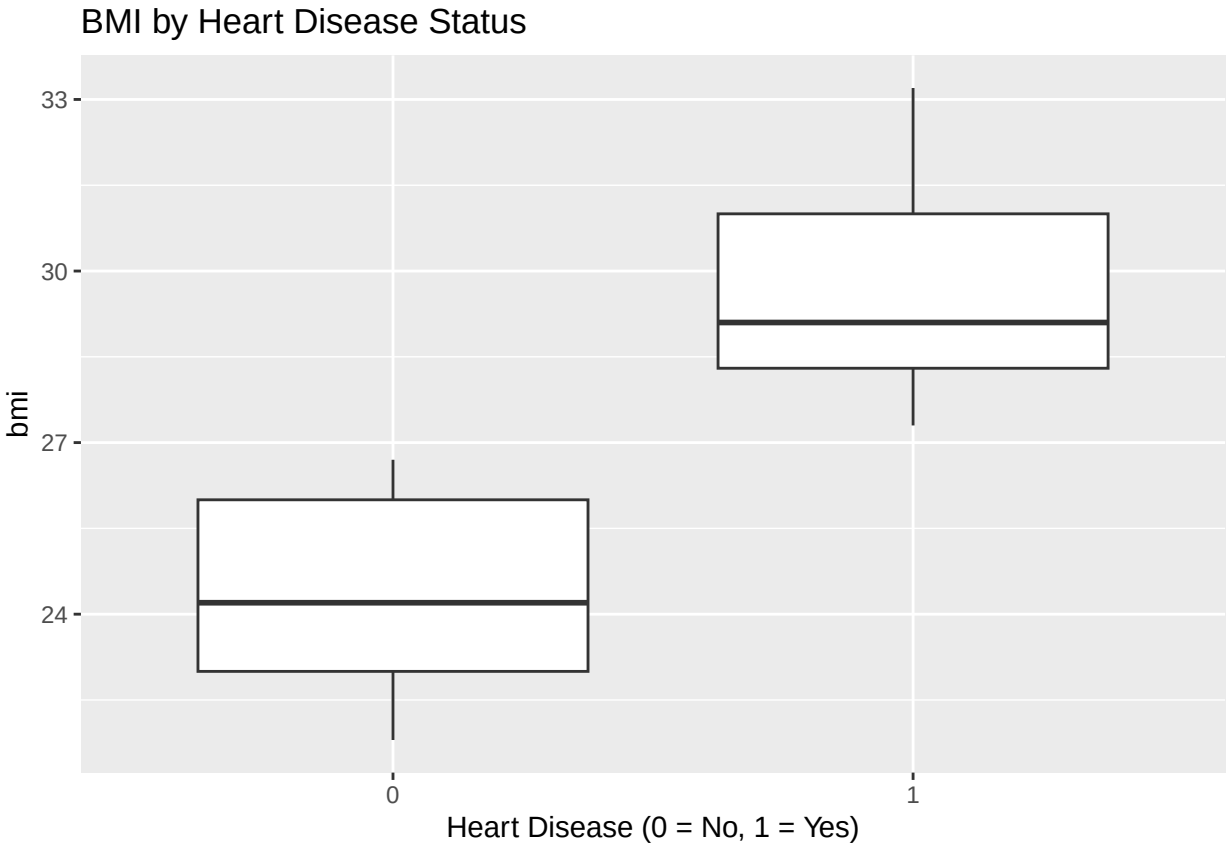
The city should consider a combined weight management and cholesterol education programs and initiatives.

3. Boxplot of BMI by Heart Disease Status

We need this to:

- To compare BMI between those with and without heart disease.
- To see if there's a clear BMI threshold associated with higher heart disease risk.

```
ggplot(health_data, aes(x = factor(heart_disease), y = bmi)) +
  geom_boxplot() +
  ggtitle("BMI by Heart Disease Status") +
  xlab("Heart Disease (0 = No, 1 = Yes)")
```



Individuals with heart disease generally have higher BMI values compared to those without heart disease.

There is more variability in BMI among individuals with heart disease.

This plot suggests a potential link between higher BMI and heart disease presence.

Based on the box plot, implement weight management programs to reduce high BMI, as it is linked to a higher risk of heart disease. Launch education campaigns on the link between high BMI and heart disease, promoting healthy lifestyle choices.

4. Correlation Analysis

We need to understand how different health metrics relate to each other and identify which factors might be most important in predicting heart disease risk.

```
round(cor(health_data[, c("age", "bmi", "cholesterol", "glucose")]),3)
```

```
##                age   bmi cholesterol glucose
## age          1.000 0.711       0.729   0.630
## bmi          0.711 1.000       0.983   0.964
## cholesterol 0.729 0.983       1.000   0.984
## glucose     0.630 0.964       0.984   1.000
```

Based on the correlation matrix, BMI, cholesterol, and glucose levels are all highly correlated with each other, and moderately correlated with age. This suggests that as age increases, BMI, cholesterol, and glucose levels also tend to rise, contributing to heart disease risk.

We should promote weight control, balanced nutrition, and regular screenings to manage these interconnected risks and reduce heart disease in the population.

5. T-test for BMI between Heart Disease Groups

We need this to:

- To statistically confirm if the BMI difference between groups is significant.
- To provide concrete evidence for policy decisions.

```
t.test(bmi ~ heart_disease, data = health_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  bmi by heart_disease
## t = -7.9948, df = 27.896, p-value = 1.075e-08
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -6.641421 -3.931913
## sample estimates:
## mean in group 0 mean in group 1
##        24.44000        29.72667
```

Since the p-value is extremely small, we reject the null hypothesis and conclude that there is a statistically significant difference in the BMI means between individuals with and without heart disease.

6. Cohen's d Calculation

We need this to:

- To understand the practical significance of the BMI difference, not just statistical significance.

```
cohens_d_bmi <- cohen.d(health_data$bmi ~ health_data$heart_disease)
```

```
## Warning in cohen.d.formula(health_data$bmi ~ health_data$heart_disease):
## Cohercing rhs of formula to factor
```

```
cohens_d_cholesterol <- cohen.d(health_data$cholesterol ~ health_data$heart_disease)
```

```
## Warning in cohen.d.formula(health_data$cholesterol ~
## health_data$heart_disease): Cohercing rhs of formula to factor
```

```
print(cohens_d_bmi)
```

```
##
## Cohen's d
##
## d estimate: -2.919306 (large)
## 95 percent confidence interval:
##     lower     upper
## -3.994228 -1.844385
```

```
print(cohens_d_cholesterol)
```

```
##
## Cohen's d
##
## d estimate: -3.028474 (large)
## 95 percent confidence interval:
##     lower     upper
## -4.124313 -1.932635
```

Unlike p-values, which only indicate whether an effect exists, Cohen's d quantifies the size of the difference, providing practical implications for health interventions.

A large effect size suggests that addressing BMI through health programs could be crucial in reducing the risk of heart disease in the community.