Exploring Heart Disease

Data Science · November 20, 2025· Mr. Merrick

Overview

We will use the **Heart Disease (CDC BRFSS)** dataset (from Kaggle), which contains responses from over 400,000 U.S. adults. For this activity you will focus on categorical variables and two-variable relationships (e.g., HeartDisease vs AgeCategory).

Key Variables

Some of the main variables you will use:

- `HeartDisease`: Has the respondent ever been told they have coronary heart disease (CHD) or myocardial infarction (MI)? (Yes/No)
- `AgeCategory`: Categorical age groups (e.g., *18–24*, *25–29*, ...).
- `GenHealth`: Self-reported general health (e.g., *Excellent*, *Very good*, *Good*, *Fair*, *Poor*).
- `Smoking`: Has the person smoked at least 100 cigarettes in their life? (Yes/No)
- `SleepTime`: Numeric: hours of sleep in a typical 24-hour period.
- Additional categorical variables you might explore: `Diabetic`, `PhysicalActivity`, `Asthma`, `KidneyDisease`, `Gender`, `Race`.

Guiding Questions (Two-Variable Categorical Focus)

Answer each question using R and include both your **plots** and **written interpretations**. When you compare two variables, treat `HeartDisease` as the **outcome** and other variables as **explanatory**.

Q1. How common is heart disease in this sample?
Create a bar chart showing how many people in the dataset have `HeartDisease = "Yes"` and `"No"`.

(a) Use counts or a simple table to estimate the probability that a randomly chosen adult in this dataset has heart disease.

(b) Suppose someone is *18–24 years old*. Would you use the same overall probability from part (a) to estimate their chance of having heart disease? Why or why not?

(c) Estimate the probability that a person has heart disease **given** that they are in the *18–24* AgeCategory. Explain how this conditional probability is different from your answer in part (a).

Q2. Heart disease vs. general health (counts).
Make a bar chart with `GenHealth` on the x-axis and bars filled by `HeartDisease`. This shows the number of people in each general health category, split by whether they have heart disease.

- Why is it hard to compare the risk of heart disease across `GenHealth` categories using **counts** alone?
- Which categories have the largest sample sizes, and why does that matter?

Q3. Heart disease vs. general health (relative bar chart).
Re-make your plot from Q2 as a **relative bar chart** by using `position = "fill"` in `geom_bar`. This makes each bar show **proportions** instead of raw counts.

- Compare the proportion of people with heart disease in each `GenHealth` category.
- From this plot, would you say there is an **association** between general health and heart disease? Explain briefly.

Q4. Heart disease across age categories.
Use a relative bar chart to compare the proportion of heart disease across `AgeCategory`.

- Which age groups appear to have the **lowest** estimated risk of heart disease?
- Which age groups appear to have the **highest** estimated risk?
- How might this pattern connect to what you know about heart disease and age?

Q5. Heart disease and smoking.
Use a relative bar chart to compare the proportion of heart disease for people who **have** smoked at least 100 cigarettes (`Smoking = "Yes"`) versus those who have not.

- Does the proportion of heart disease look higher among smokers, non-smokers, or are they similar?
- What might this suggest about the relationship between smoking and heart disease in this dataset?

Q6. Sleep time: what is "normal"?
Make a histogram of `SleepTime` for all adults in the dataset.

- Based on the distribution, what range of sleep times would you classify as "normal" for most people (e.g., 6–9 hours)? Explain how you chose your range.
- Are there many people with very short or very long sleep times? How might those extreme values affect your conclusions?

Q7. Heart disease and sleep (after trimming).
First, **filter** the dataset to include only people with `SleepTime` between 2 and 15 hours. Then create either:

- a relative bar chart of `SleepTime` grouped into categories (e.g., `Low`, `Normal`, `High`) vs. `HeartDisease`, **or**
- a relative bar chart where you treat `SleepTime` as a discrete numeric x-axis and fill by `HeartDisease`.

- Do people who sleep much less or much more than your "normal" range appear to have a higher proportion of heart disease?

- How could confounding factors (like age or general health) influence this pattern?

Q8. Your choice: explore one more association.
Choose **one additional categorical variable** (for example, `Diabetic`, `PhysicalActivity`, `Asthma`, `KidneyDisease`, or `Gender`) and use a relative bar chart to compare heart disease across its categories.

- State clearly which variable you chose and why.
- Describe any association you see between that variable and heart disease.
- If you were a public health researcher, would this association make you curious to investigate further? Explain briefly.

Dataset Description and Citation

The dataset you are using is a processed subset of a larger CDC public health survey.

- The original data come from the **Behavioral Risk Factor Surveillance System (BRFSS)**, an annual telephone survey that collects data on health-related risk behaviors, chronic health conditions, and use of preventive services in all 50 states, the District of Columbia, and several U.S. territories.
- The version used here (often distributed as `heart.csv`) includes variables such as heart disease status, age category, smoking, BMI, diabetes, sleep time, and more.
- **Example citation:** Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Survey Data*. Processed heart disease subset accessed via Kaggle.

Tutorial Solutions (R)

Q1. Overall Heart Disease Prevalence

```r
# Bar chart of HeartDisease counts
heart %>%
  ggplot(aes(x = HeartDisease)) +
  geom_bar() +
  labs(
    title = "Counts of Heart Disease in the Sample",
    x = "Heart Disease Status",
    y = "Number of Adults"
  ) +
  theme_classic()
```

```r
# Table of counts and estimated probabilities
table(heart$HeartDisease)

# Probability distribution for HeartDisease
table(heart$HeartDisease) / length(heart$HeartDisease)
```

```r
# Conditional probability given AgeCategory = "18-24"
heart %>%
  filter(AgeCategory == "18-24") %>%
  select(HeartDisease) %>%
  table()

# Two-way cross-tab
table(heart$HeartDisease, heart$AgeCategory)
```

Q2. Heart Disease vs General Health (Counts)

```r
heart %>%
  ggplot(aes(x = GenHealth, fill = HeartDisease)) +
  geom_bar(position = 'dodge') +
  labs(
    title = "Heart Disease Counts by General Health",
    x = "General Health",
    y = "Count of Adults",
    fill = "Heart Disease"
  ) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Q3. Heart Disease vs General Health (Relative Bar)

```
p_q3 <- heart %>%
  ggplot(aes(x = GenHealth, fill = HeartDisease)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion with Heart Disease by General Health",
    x = "General Health",
    y = "Proportion (within each health category)",
    fill = "Heart Disease"
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(p_q3)
```

Q4. Heart Disease vs Age Category

```
p_q4 <- heart %>%
  ggplot(aes(x = AgeCategory, fill = HeartDisease)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion with Heart Disease by Age Category",
    x = "Age Category",
    y = "Proportion (within age group)",
    fill = "Heart Disease"
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(p_q4)
```

Q5. Heart Disease vs Smoking

```
p_q5 <- heart %>%
  ggplot(aes(x = Smoking, fill = HeartDisease)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion with Heart Disease by Smoking Status",
    x = "Smoking (100+ cigarettes in life)",
    y = "Proportion (within smoking group)",
    fill = "Heart Disease"
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_classic()

print(p_q5)
```

Q6. Distribution of Sleep Time

```
p_q6 <- heart %>%
  ggplot(aes(x = SleepTime)) +
  geom_histogram(bins = 30, color = "white") +
  labs(
    title = "Distribution of Sleep Time",
    x = "Hours of Sleep in 24 Hours",
    y = "Number of Adults"
  ) +
  theme_classic()

print(p_q6)
```

Q7. Heart Disease vs Sleep (Trimmed)

```
# Filter to sleep between 2 and 15 hours
heart_trim <- heart %>%
  filter(SleepTime >= 2, SleepTime <= 15)
```

Option A: Treat SleepTime as a discrete variable

```
p_q7a <- heart_trim %>%
  ggplot(aes(x = as.factor(SleepTime), fill = HeartDisease)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion with Heart Disease by Sleep Time (2-15 hrs)",
    x = "Sleep Time (hours)",
    y = "Proportion (within each sleep hour)",
    fill = "Heart Disease"
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))

print(p_q7a)
```

Option B: Use sleep categories

```
heart_trim <- heart_trim %>%
  mutate(
    SleepCategory = case_when(
      SleepTime < 6 ~ "Low",
      SleepTime <= 9 ~ "Normal",
      TRUE ~ "High"
    )
  )

p_q7b <- heart_trim %>%
  ggplot(aes(x = SleepCategory, fill = HeartDisease)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion with Heart Disease by Sleep Category",
    x = "Sleep Category",
    y = "Proportion (within sleep category)",
    fill = "Heart Disease"
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_classic()

print(p_q7b)
```

Q8. Explore One More Variable

```
p_q8 <- heart %>%
  ggplot(aes(x = Diabetic, fill = HeartDisease)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion with Heart Disease by Diabetic Status",
    x = "Diabetic Status",
    y = "Proportion (within diabetic group)",
    fill = "Heart Disease"
  ) +
  scale_y_continuous(labels = percent_format()) +
  theme_classic()

print(p_q8)
```