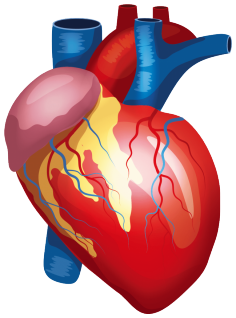

Exploring Heart Disease

March 2, 2022

The dataset used for this challenge is titled 'heart.csv'. Containing information if 400,000 adults from the annual CDC survey.

What topic does the dataset cover?

According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

Where did the dataset come from and what treatments did it undergo?

Originally, the dataset come from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. As the CDC describes: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.". The most recent dataset (as of February 15, 2022) includes data from 2020. It consists of 401,958 rows and 279 columns. Variables in the dataset are shown below:

- HeartDisease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
- BMI: Body Mass Index
- Smoking: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
- AlchoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
- Stroke: Has had a stroke or not.
- PhysicalHealth: Physical health describes illness and injury, for how many days during the past 30.
- MentalHealth: How many days during the past 30 days was your mental health not good?
- DiffWalking: Do you have serious difficulty walking or climbing stairs?
- Gender: Male or female.
- AgeCategory: Fourteen-level age category.

- Race: race/ethnicity value.
- Diabetic: Has Diabetes or not
- PhysicalActivity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
- GenHealth: What is your general health?
- SleepTime: On average, how many hours of sleep do you get in a 24-hour period?
- Asthma: Has asthma or not.
- KidneyDisease: Kidney disease or not.
- SkinCancer: Skin cancer or not.

Assignment

Complete each of the following questions using R Studio and submit your answers as a detailed report.

1. Create a barchart showing the number of people who have heart disease and who do not.
   (a) Using the chart you created estimate the probability someone has heart disease.
   (b) If you were told that a person was 18 years old and wanted to estimate the probability they have heart disease, would you use your estimated probability from part (a)? Why or why not?
   (c) Estimate the probability someone has Heart disease given they are 18 years old.

2. Create a barchart where general health is on the x-axis and each category is filled by whether or not a person has heart disease. Looking at the plot, why is it difficult to compare categories of counts?

3. Make the barchart you created in question 2 a *relative barchart* by specifying the argument position="fill" in the barchart geometry. From your plot would you infer that there is an association between general health and heart disease? Why?

4. Use a relative barchart to compare the proportion of heart disease across age categories. What can you infer from your plot?

5. Use a relative barchart to compare the proportion of heart disease across people who smoke. What can you infer from your plot?

6. Visualize the distribution for sleep times. What range of sleep times would you classify as 'normal'.

7. First filter the dataset to only include sleep times between 2 and 15 hours. Use a relative barchart to compare the proportion of heart disease across sleep times. What can you infer from your plot.

8. Use a relative barchart to explore the association between heart disease and one other variable in the dataset.