Analysis of Heart Disease Dataset

Omar Moustafa

900222400

January 22, 2023

DSCI 1411 (Winter 2023)

Project R

Table of Contents:

Introduction:

Heart disease is defined as a disease that negatively affects one's blood vessels or heart. Smoking, high blood pressure or cholesterol, having a poor diet, insufficient exercise, and obesity are the most common risk factors for various cardiac problems. One way to truly understand the severity and horrific effects of heart disease is by looking at statistics on its incidence within the United States. In the United States, heart disease is the leading cause of death for both sexes, and in the year 2020, it took the lives of approximately 697,000 people, accounting for one in every five fatalities.

"Heart Attack Analysis & Prediction" is a dataset that works with thirteen different variables of different types to examine multiple different trends and relationships to be able to make affirmative conclusions about which physical conditions are more or less susceptible to heart diseases. This dataset includes four databases regarding the diagnosis of heart disease, which were obtained from the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, the V.A. Medical Center in Long Beach, and also the University Hospital in Zurich.

Link To Source:

https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

What Question(s) Can Be Answered By Analyzing This Dataset?

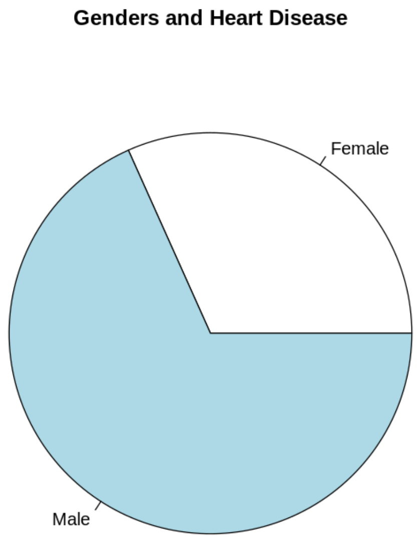
1. Which gender is more prone to heart disease than the other? Men or women?
2. During what age ranges are heart diseases most and least common to occur?
3. How does having heart disease affect your maximum achievable heart rate?
4. How does having heart disease affect one's condition of ST depression induced by exercise relative to rest?
5. What is the relationship between blood pressure and the chance of getting heart disease?
6. What is the relationship between levels of cholesterol and the chance of getting heart disease?

Each Variable Type & its Unit of Measurement:

Variable Name:	**Variable Type:**	**Unit of Measurement:**
Age	Quantitative	Age of the patient
Sex	Non-Ordinal Categorical, Character	Sex of the patient
CP [Chest Pain Type]	Ordinal Categorical	Level of chest pain
TRTBPS [resting blood pressure]	Quantitative	mmHg
CHOL [cholesterol]	Quantitative	mg/dl
FBS [fasting blood sugar >120]	Logical	mg/dl
RESTECG [resting electrocardiographic results]	Ordinal Categorical	N/A
THALACHH [maximum heart rate achieved]	Quantitative	bpm
EXNG	Logical	N/A
OLDPEAK	Quantitative	ST depression induced by exercise relative to rest
SLP	Ordinal Categorical	ST Segment
CAA	Ordinal Categorical	Number of major vessels
THALL	Ordinal Categorical	N/A
Diagnosis	Logical	N/A

Graphical Displays & Analysis:

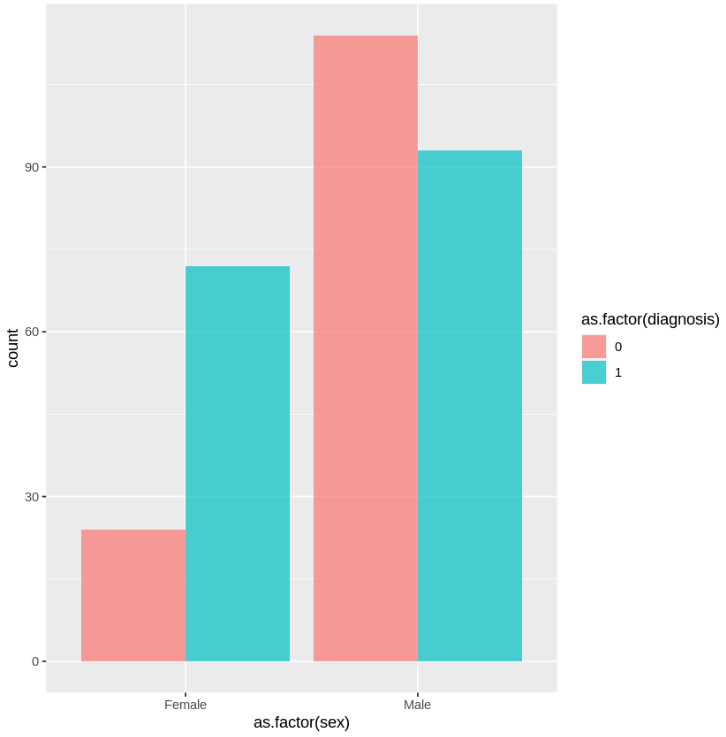
1. Pie Chart



Analysis of the Pie Chart:

The dataset's observations included 303 people, 207 of whom were male and 96 of whom were female.

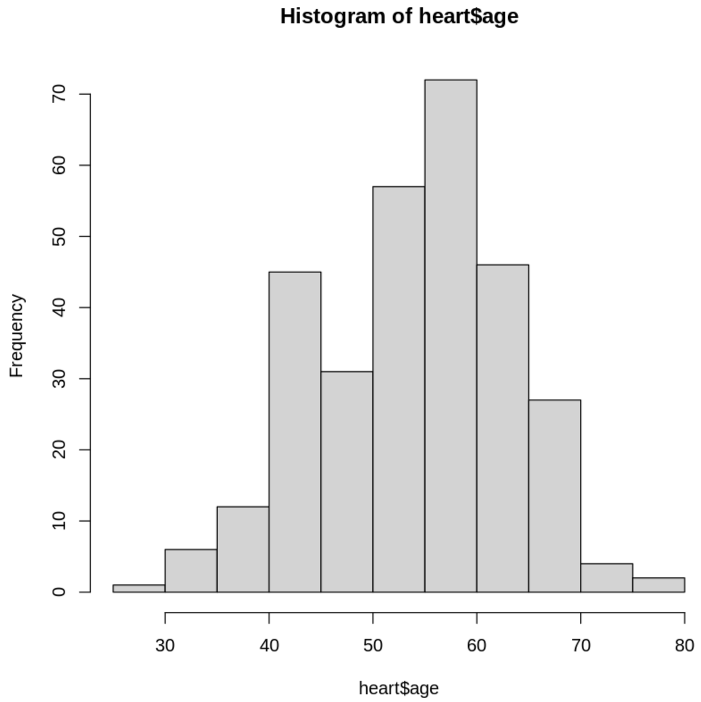
2. Bar Chart



Analysis of the Bar Chart:

The x-axis of the bar chart above is sex, and the y-axis is their count. Each of the two parts of the axis is separated by a diagnosis factor, where the participants that have not been diagnosed are labeled with the color blue, and those who have been diagnosed are labeled with pink. So, this bar chart shows that among the participants in this dataset, the ratio of diagnosed males to non-diagnosed males is a lot closer than the ratio of diagnosed females to non-diagnosed females.

3. Histogram



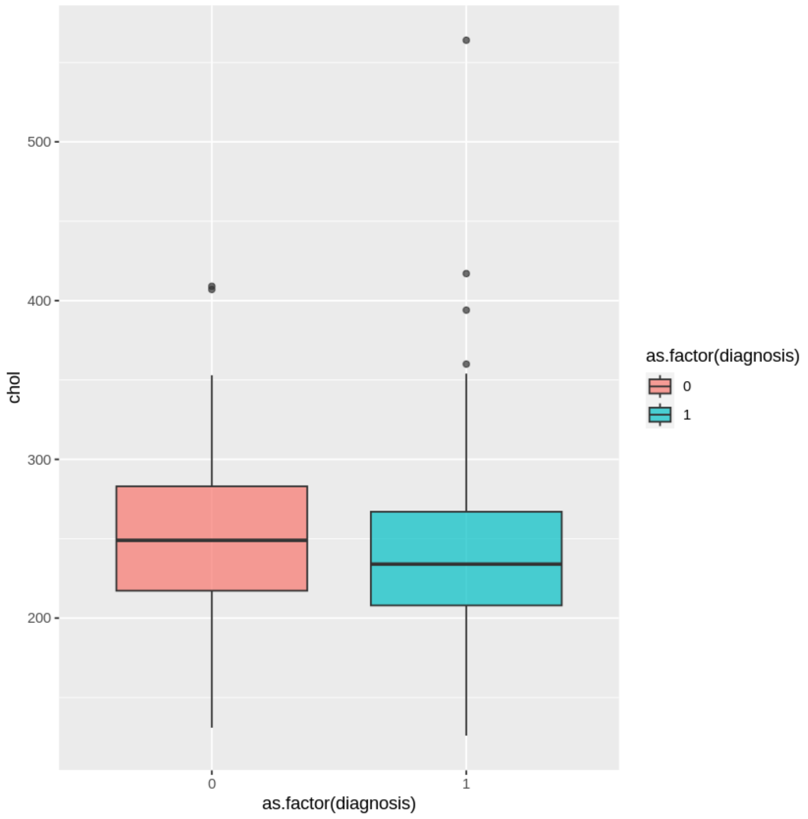
Analysis of the Histogram:

Above is a histogram where the x-axis represents the age range of the study’s 303 participants and the y-axis represents the frequency of age ranges of 5 years starting with the data set’s minimum age and ending with its maximum range.

It can be seen that the most common age range of the participants is 55 to 60 years old, and the least common is 25 to 30 years old. Because this is a data set that focuses on one’s chances of getting heart disease, it would be reasonable to infer that 55 to 60 is by far the most common age range for when people do get heart disease, and it is very uncommon for someone to get heart disease before the age of 30 or after the age of 75.

4. Box Plots

Box Plot #1:



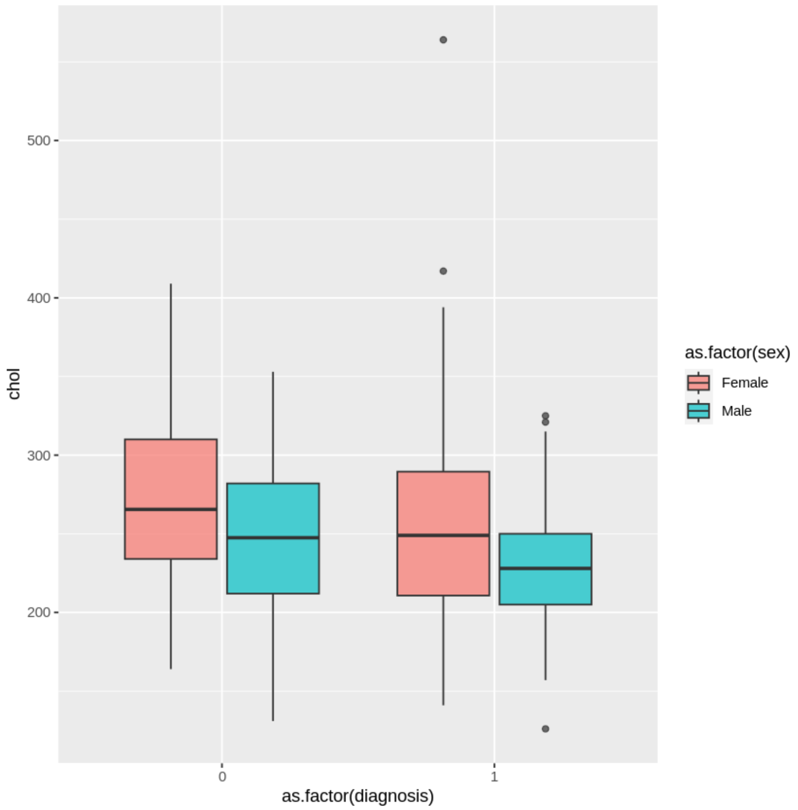
Hypothesis: the higher the level of cholesterol, the higher the chance of getting heart disease

Analysis of Box Plot #1:

A box plot with the x-axis being the heart disease diagnosis is shown above, where 0 represents being diagnosed with heart disease (shown in pink) and 1 represents not being diagnosed (shown in blue), and the y-axis shows the level of cholesterol in both sets of participants.

From this particular box plot, it can be seen that the participants who have been diagnosed with heart disease have higher levels of cholesterol than those who have not been diagnosed. Specifically analyzing the box plots, those who have been diagnosed with heart disease have higher minimum, maximum, median, and quartile values for their levels of cholesterol than the participants who have not been diagnosed. Therefore, it can be concluded that having higher levels of cholesterol will lead to a higher chance of getting heart disease, representing a directly proportional relationship between these two variables.

Box Plot #2:



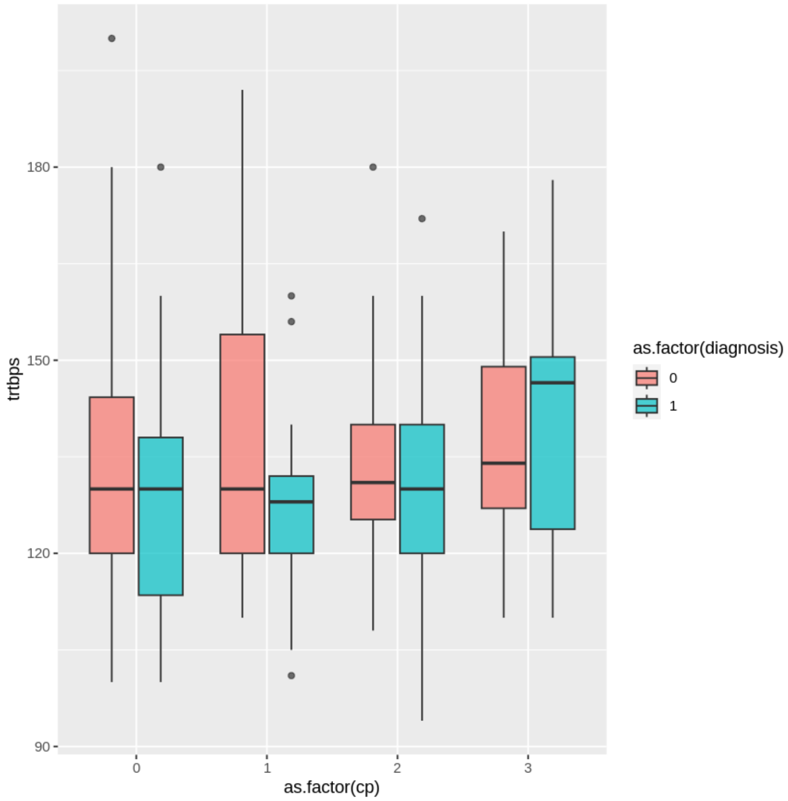
Analysis of Box Plot #2:

Above is another box plot with the x-axis being the heart disease diagnosis and the y-axis showing the level of cholesterol in both sets of participants; however, a separation between the male and female participants was added to see which gender generally has higher cholesterol levels. Knowing which gender has higher cholesterol levels can lead to the conclusion that the higher-cholesterol gender is more vulnerable to heart disease than the other.

In the graph above, where the males are represented by the color blue and the females are represented by pink, it can be observed that, regardless of whether they have been diagnosed with heart disease or not, females generally have higher cholesterol levels than males. This observation would lead to the valid conclusion that women are more susceptible or prone to heart disease than males because of the directly proportional relationship between cholesterol levels and diagnosis, which was confirmed by the previous box plot.

To make sure that the box plot above provides valid information, it can be scientifically supported by the article "Why Cholesterol Matters For Women," published by John Hopkins Medicine, which states that "women have higher levels of HDL cholesterol than men because the female sex hormone estrogen seems to boost this good cholesterol."

Box Plot #3:



Hypothesis: having high blood pressure leads to a higher chance of getting heart disease

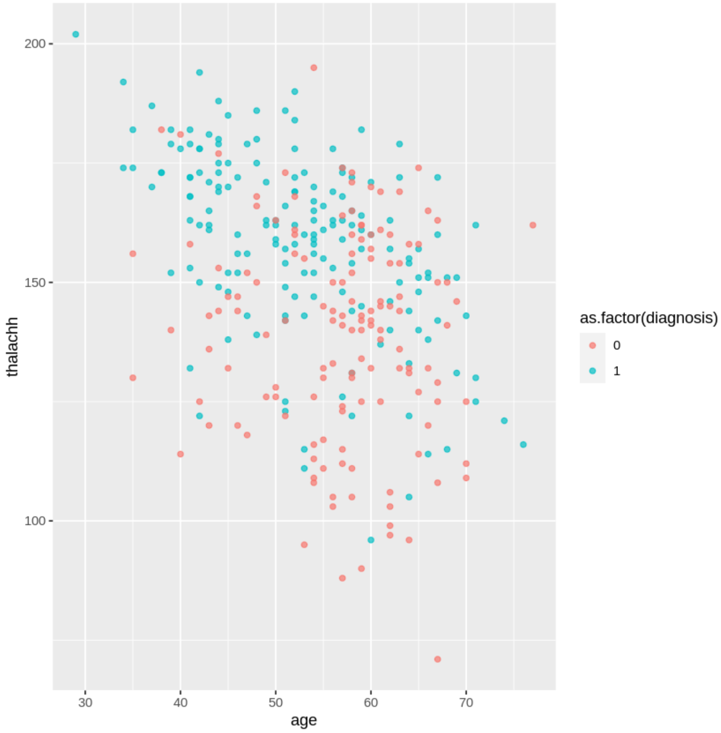
Analysis of Box Plot #3:

Box Plot #3 has the type of chest pain as its x-axis, where 0 represents the worst type of chest pain, "typical angina," 1 represents "atypical angina," 2 represents "non-anginal pain," and 3 is the least painful, "asymptomatic." As for the y-axis, Box Plot #3 uses the resting blood pressure of the participants, which is measured in mmHg, as its y-axis in order to explore the effect of having heart disease on one's blood pressure. Additionally, Box Plot #3 also uses the color pink to represent the participants who have been diagnosed with heart disease and the color blue to represent the participants who have not been diagnosed.

From this specific graph, we can see that those who have been diagnosed with heart disease have higher blood pressure at every level of chest pain, with the exception of asymptomatic. For typical and atypical angina and also non-anginal pain, it is clear that those with heart disease have higher minimum, maximum, and quartile values than those without heart disease. Accordingly, it is reasonable to say that the effect of having heart disease on one's blood pressure is that it badly increases it.

5. Scatter Plots

Scatter Plot #1:



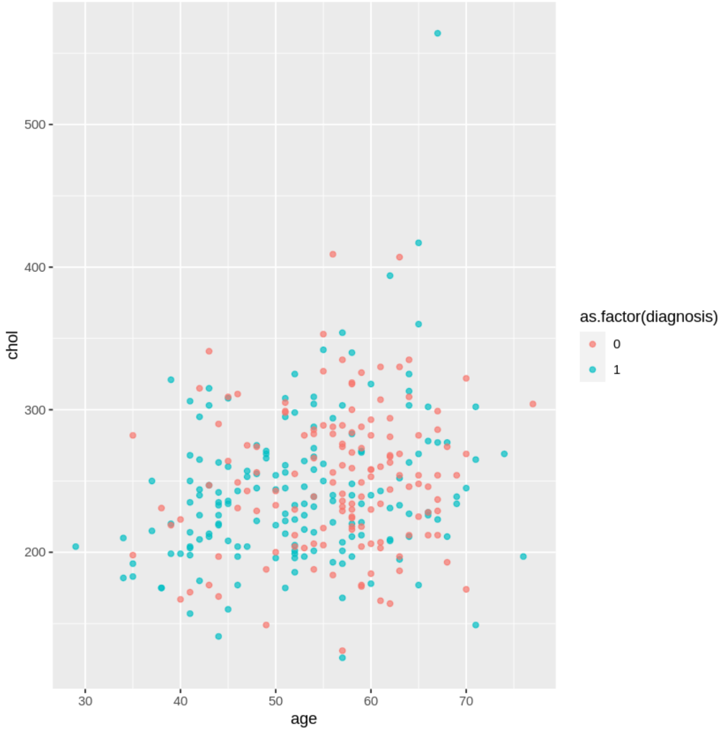
<u>Hypothesis:</u> Having heart disease would lower the maximum heart rate that one could reach

Analysis of the Dot Plot:

The figure above is a dot plot that has the ages of the participants as the x-axis and their maximum achievable heart rate as the y-axis, with the pink dots representing those who have heart disease and the blue dots representing those who have not been diagnosed.

In this dot plot, it can be clearly seen that the vast majority of heart disease patients have a maximum heart rate that is less than 150 bpm, and the vast majority of the participants without heart disease have a maximum heart rate that is between 175 and 190 bpm. Mathematically, a person's maximum achievable heart rate can be calculated by subtracting their age from 220, so if we take two 55-year-old participants from the data set, one of whom is a heart disease patient and the other who is not, the figure above shows the heart disease patient's maximum achievable heart rate is around 140 bpm and the non-diagnosed participant's maximum achievable heart rate is around 160 bpm. Therefore, the relationship that can be deciphered from this particular dot plot is that having heart disease affects one’s maximum achievable heart rate by lowering it from what it should naturally be.

Scatter Plot #2:

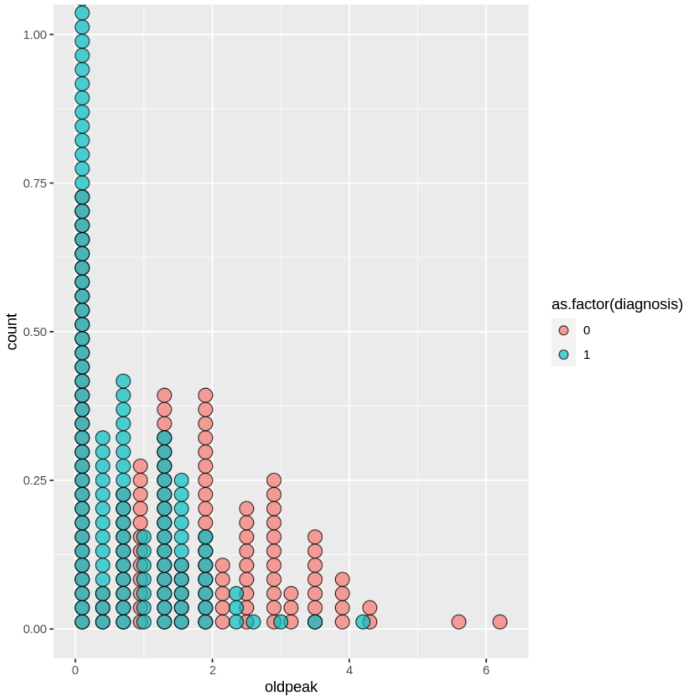


Analysis of the Scatter Plot:

The scatter plot above works to observe at what ages it is most common to be diagnosed with heart disease, with the x-axis representing the ages of the participants and the y-axis being the participants' levels of cholesterol. Additionally, the color pink represents the participants who have been diagnosed with heart disease, while the color blue represents the participants who have not been diagnosed. So, it can be observed that the majority of the participants who have been diagnosed with heart disease are over the age of 55.

Another trend that can be seen in this scatter plot, although it is not a very strong trend, is that the older the participant is, the higher their level of cholesterol is, representing a direct relationship between the variables age and cholesterol. This trend can also be used to explain why the majority of diagnosed participants were 55 or older: before middle age, one does not have a relatively high level of cholesterol, but as one gets older and closer to middle age, their cholesterol levels rise, making them more susceptible to heart disease due to the direct relationship, i.e., the higher the cholesterol levels, the greater the chance of getting heart disease.

6. Dot Plot



Hypothesis: Having heart disease would make one’s condition of ST depression induced by exercise relative to rest more critical

Analysis of the Dot Plot:

The dot plot above uses its x-axis to represent the variable “oldpeak,” which is the ST depression induced by exercise relative to rest; the farther down the x-axis, the worse the case of oldpeak is; and the y-axis is used to count them. Adding to this, whether the participant was or was not diagnosed with heart disease was used as a factor separator to observe the trend and relationship between the ST depression induced by exercise relative to rest and having heart disease. So, examining this dot plot would lead to the observation and conclusion that those with heart disease tend to experience a more critical condition of oldpeak. This is because it can be seen that most of the heart disease patients are located after x = 2 on the graph, while the vast majority of participants without heart disease are seen before x = 2. Therefore, it is safe to say that heart disease leads to a worse case of ST depression that is induced by exercise and relative to rest.

Summary & Conclusion:

Prior to actually graphing and analyzing the data, a hypothesis was made for most of these questions to have an uncertain answer to the research questions that are about to be confirmed or denied. The hypothesis for the relationship between levels of cholesterol and the chance of getting heart disease was that the higher the level of cholesterol, the higher the chance of getting heart disease. Another hypothesis that was made was for the relationship between heart disease and one’s maximum achievable heart rate, which was that having heart disease would lower the maximum heart rate that one could reach. A third hypothesis was made for the relationship between heart disease and one’s blood pressure, which is that having high blood pressure can lead to getting heart disease. As well as the previous three hypotheses, a fourth hypothesis was made for the relationship between having heart disease and one’s condition or severity level of ST depression induced by exercise relative to rest, which was that having heart disease would make the condition even more critical. All of these hypotheses were indeed proven correct after graphing and analyzing the data and observable trends. In summary, after revisiting the various questions that can be answered by analyzing the data, it was concluded that females are more prone to heart disease than males; higher levels of cholesterol mean a higher chance of getting heart disease; heart disease increases one's blood pressure, decreases one's maximum achievable heart rate, and worsens one’s condition of ST depression induced by exercise relative to rest; and that the most and least common age ranges for when people get diagnosed with heart disease, respectively, are 55 to 60 and 25 to 30.

Since the serious danger and severity of heart disease are widely known and acknowledged, this dataset can be utilized to make it possible to predict heart disease from its parameters and variables. Using the questions that show a direct relationship between a certain variable and the possibility of getting heart disease and their detailed examinations, medical precautions can be taken to prevent that variable and its direct relationship with causing heart disease from having a serious effect on people. This particular trait has priceless value as it would serve to monitor people’s health and status and has the potential of reducing the huge number of deaths caused by heart disease. Even if it is a relatively small reduction, it would still definitely make a difference.

Works Cited:

Rahman, Rashik. "Heart Attack Analysis & Prediction Dataset." *Kaggle.com*. 2019. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset. Accessed January 18, 2023.

"Heart Disease Facts." *Centers for Disease Control and Prevention.* October 14, 2022 https://www.cdc.gov/heartdisease/facts. Accessed January 18, 2023.

"Other Conditions Related to Heart Disease." Centers for Disease Control and Prevention. January 20, 2022. https://www.cdc.gov/heartdisease/other_conditions. Accessed January 19, 2023.

Michos, Erin Donnelly. "Why Cholesterol Matters for Women." *John Hopkins Medicine.* https://www.hopkinsmedicine.org. Accessed January 19, 2023.