Info 250: Final Project

Sabah Lala

Final Academic Report: Factors affecting Heart Disease

**Dataset:**

***Heart Disease UCI.***

It is found on Kaggle at <https://www.kaggle.com/ronitf/heart-disease-uci> however, the dataset comes from a subset of the Cleveland Database containing heart disease data.

This dataset has been generated specifically for Machine Learning researchers. The creator of this particular dataset has been accredited to V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. The date that this dataset was donated was 1988-07-01.

The **target** variable in this dataset classifies the patient as having (1,2,3,4) or not having (0) heart disease. The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

**Data Dictionary:**

A screenshot of a cell phone

Description automatically generated

\*Note: Thalach, in particular, is the metric for maximum heart rate achieved.

**Purpose:**

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. Exploring any possibility of how certain factors can lead to increased chances of getting heart disease is a worthwhile project to pursue for the general public of American (but more importantly for those that are prone to HD).

Statistical analysis has identified many risk factors associated with heart disease such as age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity, lack of physical exercise, etc. I am going to run statistical testings using the Cleveland heart disease dataset to assess factors like maximum heart rate one can achieve during exercise and how it is associated with a higher likelihood of getting heart disease.

**Audience:**

Are you or your loved one worried about heart disease? Have your cholesterol test results come higher than normal? Is there a genetic probability of heart disease that runs in your family? Then you should read this report to understand what different factors can affect the probability of you having heart disease!

Jenna just came back from her doctor’s office after experiencing chest pain. The doctor ran some tests and want to monitor her levels over a couple of days. Jenna wants to get a better idea on the probability that she might suffer from heart disease and get a better grasp on what factors seem most important in causing it. These are some of the questions on her mind:

1. As age increases, does that also increase the possibility of having heart disease (target variable = 1 or 2 or 3 or 4)?

2. Is there a particular gender that is more prone to getting heart disease?

3. Is there a correlation with the serum in cholesterol and blood pressure to the likelihood of getting heart disease?

4. If I am having chest pain, does that definitely indicate heart disease?

After searching on google, she stumbled on this article and it gave her an idea of which factors she should look out for the most.

The target audience that I am trying to pursue is people that would like to get a little more information on the variables that factor in to making them more probable of getting heart disease in the near future. This way, they can monitor their levels constantly and know when to reach their provider if they suspect that they will get it in the near future.

They may already know that heart disease is caused by a lot of different factors like the environment, what you eat, genetics, etc and also may have some familiarity with some of the variables in the dataset if someone in their family already has heart disease. The aspect of the data that they might be interested in is how the different variables each can factor into how probable they are of getting heart disease.

**Exploratory Data Analysis:**

* To get some statistical descriptions of the variables, I used the summary and str functions.

A screenshot of a cell phone

Description automatically generated

A close up of text on a black background

Description automatically generated

* Converting diagnosis class into outcome variable

I noticed that the outcome variable target has more than two levels. I will map any non-zero values that can be coded as an "event” to represent a binary 1/0 outcome.

* Identifying important clinical variables for visualizations

I can run statistical tests to see which ones are related to heart disease. I will explore the associations for each variable in the dataset. Depending on the type of the data (i.e., continuous or categorical), I will use t-test or chi-squared test to calculate the p-values.

**Visualizations**

The following visuals all will be static graphs created in R using the library ggplot.

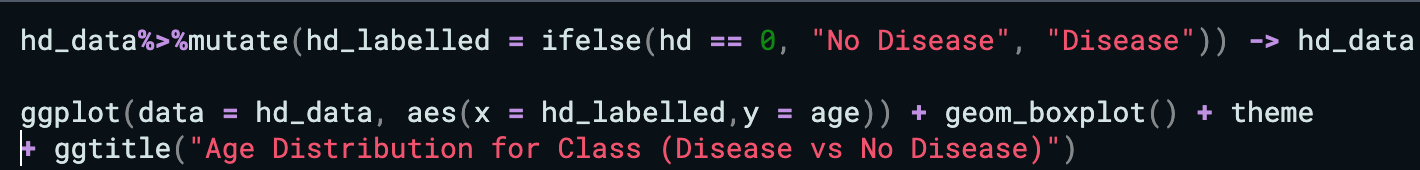
I will plot the age, sex, and maximum heart rate distributions with respect to the outcome variable. This will give a sense of both the direction and magnitude of the relationship.

**“Does age have an effect on our target variable?”**

Visual #1:

Age is continuous so boxplot makes sense. In this case, we see that the approximate median age that you are if you have HD is around 52 years old. The spread of the data in terms of age can clearly be seen through this boxplot.

Code:



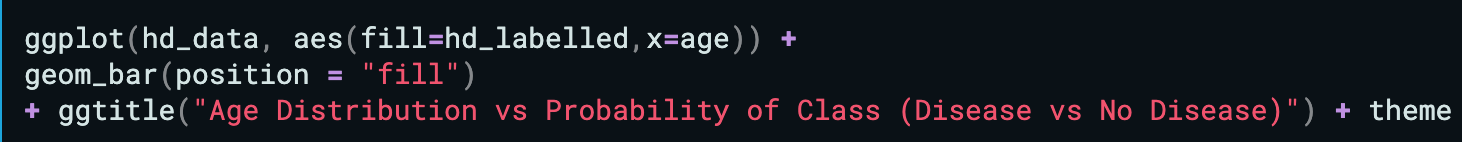
A screenshot of a social media post

Description automatically generated

Visual #2:

Since we determined that Age can have an effect on heart disease, this is another way to visualize age and its distribution over the probability of having heart disease.

Code:



A close up of a logo

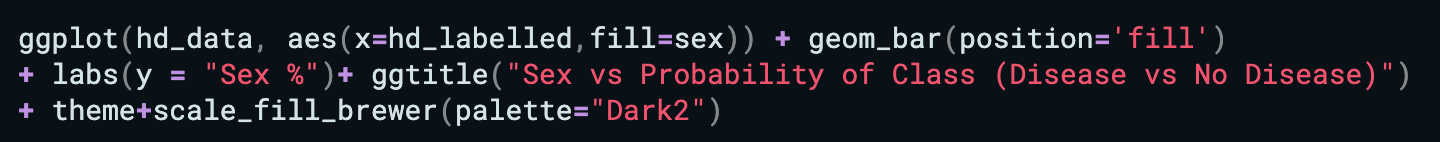
Description automatically generated

**“Does sex have an effect on our target variable?”**

Visual #3:

Graphing sex using a bar plot since it is a binary variable in this dataset. We can see through the graph that sex does not seem to have that much of an effect because approximately the same percentage of males and females have heart disease. However, we do see that males in this dataset tend to have higher chance of not having heart disease.

Code:



A screenshot of a cell phone

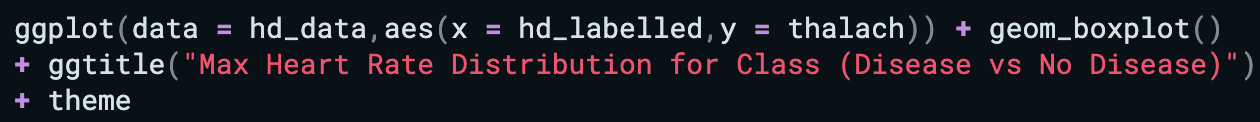
Description automatically generated

Visual #4:

**“Does thalach have an effect on our target variable?”**

Through the boxplot we can see that the median maximum heart rate that can indicate heart disease is somewhere around 160 bps. There are a considerable number of outliers as well. The spread for heart disease is much smaller than that of no heart disease.

Code:



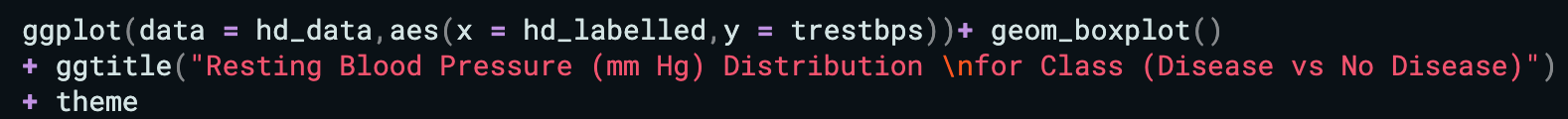
A screenshot of a social media post

Description automatically generated

Visual #5:

Even though blood pressure did not show up as a statistically significant variable, given my medical background, I would like to also explore the relationship between the target variable and blood pressure. Especially, looking at any outliers. This is best done by a box plot.

Code:



A screenshot of a social media post

Description automatically generated

Citations:

UCI Machine Learning Repository: Heart Disease Data Set, archive.ics.uci.edu/ml/datasets/heart+Disease.