VITAL SIGNS: UNRAVELING PATTERNS OF HEART HEALTH FROM CLINICAL DATA INSIGHTS

By: Yiran Hu and Koki James

Introduction

In the United States, heart disease has been the leading cause of death every year for over 100 years, claiming the lives of more people than strokes, cancer, and respiratory diseases. According to the National Safety Council, over 700,000 people died due to heart disease in 2022. Many factors can increase the chance of getting heart disease, such as physical inactivity, obesity, and drug & alcohol use. Heart disease develops over time, so being able to identify a high risk of developing it is vital as it allows for the patient to work on their health to potentially improve their condition. Even though heart disease itself is typically not reversible, knowledge of potential heart issues also allows doctors to provide more solutions to alleviate the problem as many of the symptoms can actually be reversed. We will analyze a dataset of patients from a Cleveland hospital that have come in with some level of chest discomfort to identify trends and irregularities in the data in order to get a better understanding of the complexities of heart disease research.

Dataset Description and Selection Justification

This dataset is sourced from Kaggle, the leading platform for data science and machine learning, which provides an extensive collection of datasets across various domains. Our group has selected the "Heart Attack Analysis & Prediction Dataset" for our study. Specifically designed for heart disease classification, this dataset aims to predict the risk of heart disease using a variety of medical indicators from patients.The dataset encompasses information on 303 individuals, covering crucial health indicators such as age, gender, exercise-induced angina, the number of major vessels, types of chest pain, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, and maximum heart rate achieved. Each feature provides valuable

insights into the patients' cardiac health status, making this dataset an invaluable resource for our heart disease analysis and prediction project.

Reasons for Choosing This Dataset

1. Comprehensiveness and Diversity:

The dataset offers a comprehensive coverage of factors related to heart disease. It spans ages from under thirty to over seventy and encompasses different genders. Analyzing such a rich and varied sample allows us to derive conclusions that are both universally applicable and relevant.

2. Well-Organized and Clean Data:

The dataset is well-organized, with each feature clearly delineated and independent from others. It consists of clean data, free from missing values, which facilitates straightforward analysis. Furthermore, the dataset provides detailed categorizations for certain features—such as dividing Chest Pain Type into four distinct categories—enabling more precise statistical analysis and better organization of the data.

3. Reliability and Insights for Heart Attack Risk Prediction:

This dataset provides critical and reliable information for analyzing and predicting the risk of heart attacks in patients. Through the data provided by these 300+ individuals, we aim to uncover underlying connections and potential undiscovered trends. We seek insights that could enlighten scientists, medical professionals, and the general public alike on the characteristics closely associated with heart disease. For instance, we are interested in exploring the relationship between the number of major vessels and the incidence of heart disease—whether a greater number of vessels correlates with a higher probability of heart disease. Through a comprehensive analysis of these features, we can gain a deeper understanding of the risk factors for heart disease, thereby predicting the likelihood of its occurrence. This type of analysis holds significant value for medical research and provides crucial information for patient health management.

Objectives and Key Questions

In our comprehensive analysis of the Heart Attack Analysis & Prediction Dataset, we aim to uncover patterns and correlations that can advance our understanding of heart disease risks. Our overarching question is: What are the critical factors related to heart disease in patients? To answer this, our analysis seeks answers to the following key questions:

1. What is the Correlation of Resting Blood Pressure and Age with Heart Disease?

We plan to find out how resting blood pressure and patient age interact in the context of heart disease within our dataset. Our objective is to map out any significant trends showing the relationship between these two factors and the likelihood of heart disease across different demographics.

2. What is the Distribution of Heart Disease Across Different Age Groups and Genders?

We want to examine the distribution of heart disease prevalence among varying age brackets and between genders. We aim to determine which age groups and gender categories exhibit a higher probability of heart disease.

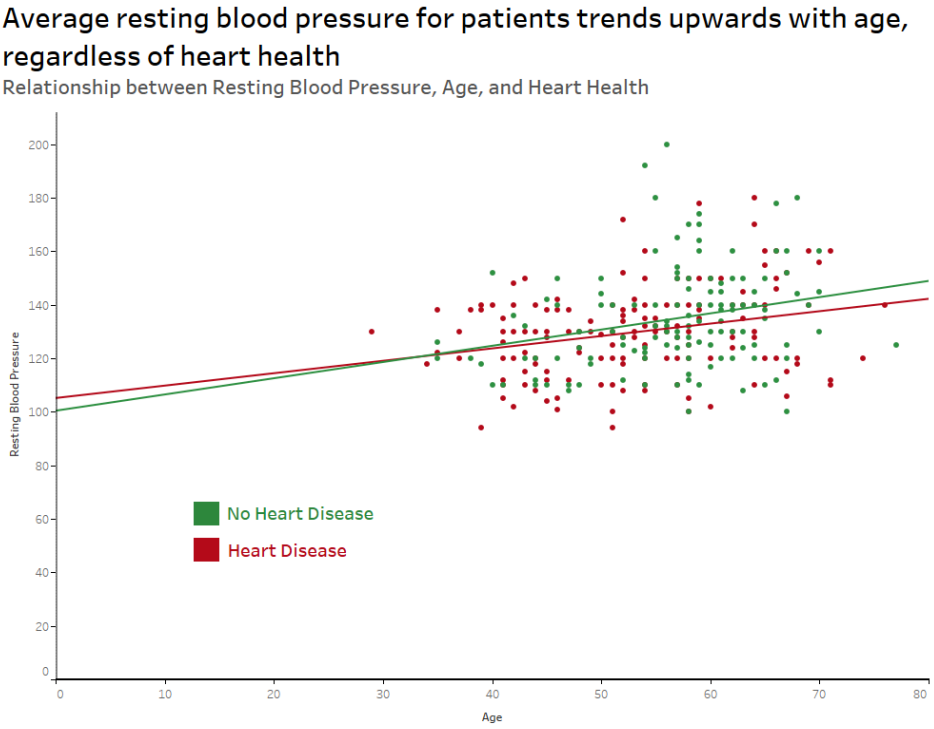
3. How Does the Distribution of Maximum Heart Rate Relate to Heart Disease?

We seek to analyze the correlation between the maximum heart rate observed in participants and the prevalence of heart disease. Does the data suggest a significant difference in the distribution of maximum heart rates between those with and without heart disease?

4. What is the Correlation of Chest Pain and Electrocardiographic Results with Heart Disease?

We aim to analyze how the distribution of chest pain types correlates with heart disease incidence. Simultaneously, we will assess how various electrocardiographic results align with these incidences. By examining these two indicators, we hope to uncover any patterns that may provide deeper insights into the diagnosis and classification of heart disease.

1. Understanding the Correlation of Resting Blood Pressure and Age with Heart Disease



A. Methodology

In order to analyze the correlation of resting blood pressure and age with heart disease in the patients, we needed to split the patients into categories: those diagnosed with heart disease and those without it. From there, we could plot the information on a chart with our independent variable (age) compared to our dependent variable (resting blood pressure) as a scatter plot. Each point would represent one patient. Finally, we added a trend line for patients with heart disease and one for those without heart disease to more easily compare the two categories.

B. Findings

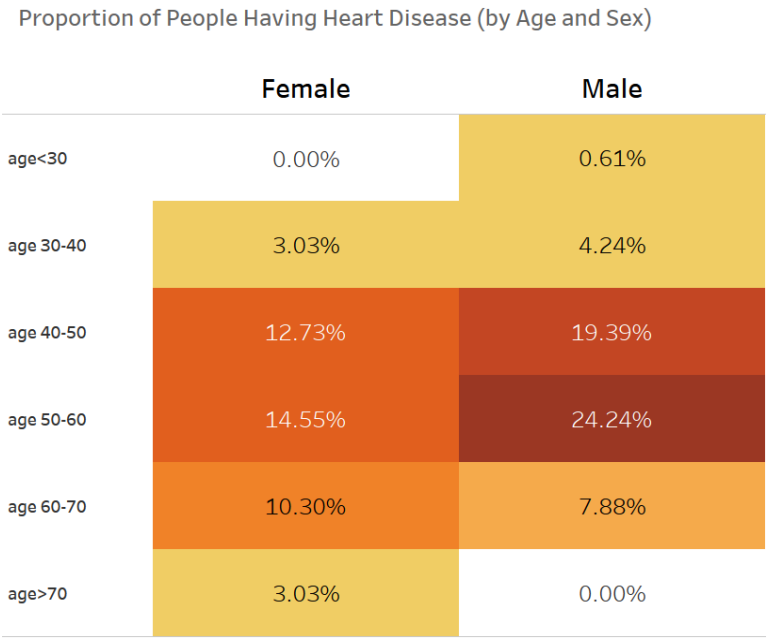
The scatter plot reveals that patients that came in with chest pain at the Cleveland hospital tended to have a higher blood pressure regardless of their heart disease diagnosis. In fact, those without heart disease are expected to have a higher blood pressure on average as the age increases compared to those with heart disease. However, the younger age demographic tended to have more people with heart disease

that had high blood pressure. This could indicate that high blood pressure is tied closely to chest pain in general, and that there is at least some level of correlation between high blood pressure and heart disease.

C. Limitations

Because this dataset only consists of individuals who have come into the Cleveland hospital with chest discomfort, it cannot be generalized that these findings are representative of the overall population. In fact, 206 of the 303 patients have a resting blood pressure higher than normal levels. While resting blood pressure doesn't seem to be a predictor of heart disease in these patients, it is correlated with having chest pain to some degree. High blood pressure can lead to damage of the heart, so many of the patients that do not have heart disease are still at a high risk for heart disease. This could be something to be researched further. Generalizations also cannot be made about those below 40 and above 70 as not as many people in these categories came into with chest pain. This could be due to a general lack of heart disease prevalence at lower ages and from people dying due to heart disease at the higher age levels.

2. Analysis of the Distribution of Heart Disease Across Different Age Groups and Gender



A. Methodology

To explore the distribution of heart disease by age and gender, we commenced by examining the age distribution of patients within the Heart Attack Analysis & Prediction Dataset, which ranges from under 30 to over 70 years old. We segmented the dataset into decade-based age groups and categorized the patients accordingly. We then calculated the proportion of individuals with heart disease within each age and gender category relative to the total number of individuals with heart disease in the dataset. These proportions were represented using a heat map, a graphical tool that effectively illustrates variance through color gradation based on the magnitude of the proportions.

B. Findings

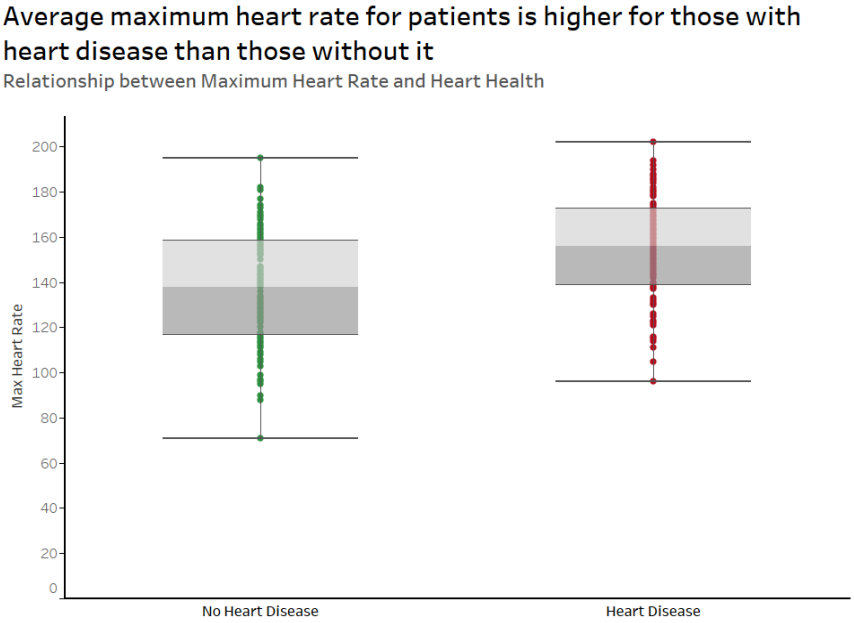
The heat map revealed that within our dataset, males aged 50-60 have the highest proportion of heart disease, accounting for 24.24% of all cases. This was followed by males in the 40-50 age bracket and females in the 50-60 age bracket, with proportions of 19.39% and 14.55%, respectively. Therefore, we hypothesize that the observed prevalence within our dataset may suggest a possible correlation between an increased risk of heart disease and individuals in the male group aged 40-60, as well as females aged 50-60.

C. Limitations

It is important to note that the number of individuals in each age and gender group within our dataset is not uniform, which may influence the results of the analysis to some extent. For example, the female group aged 40-50, representing 12.73%, does not appear as prominently as the other mentioned categories. However, this group had a small representation in the dataset, consisting of only 22 individuals out of a total of 303, with 165 individuals having heart disease. Within this specific subset, 21 individuals had heart disease, and only one did not. Consequently, the heat map may underestimate the risk of heart disease in the 40-50 age group for females. Whether this is an actual underestimation cannot be determined with our current sample size. A larger sample could

potentially clarify this ambiguity. Additionally, the proportion of individuals over 70 with heart disease is very low in our dataset. We caution against interpreting this to mean that the risk of heart disease is low in this age group, as the dataset contains only a few individuals over 70.

3. How Does the Distribution of Maximum Heart Rate Relate to Heart Disease?



A. Methodology

In order to explore how maximum heart rate relates to heart disease, we created a box plot to see how those without heart disease compared to those with it. We plotted each patient's maximum heart rate as a point on our chart. When the box and whiskers are added to the chart, it gives us a good idea of how the values compare to each other.

B. Findings

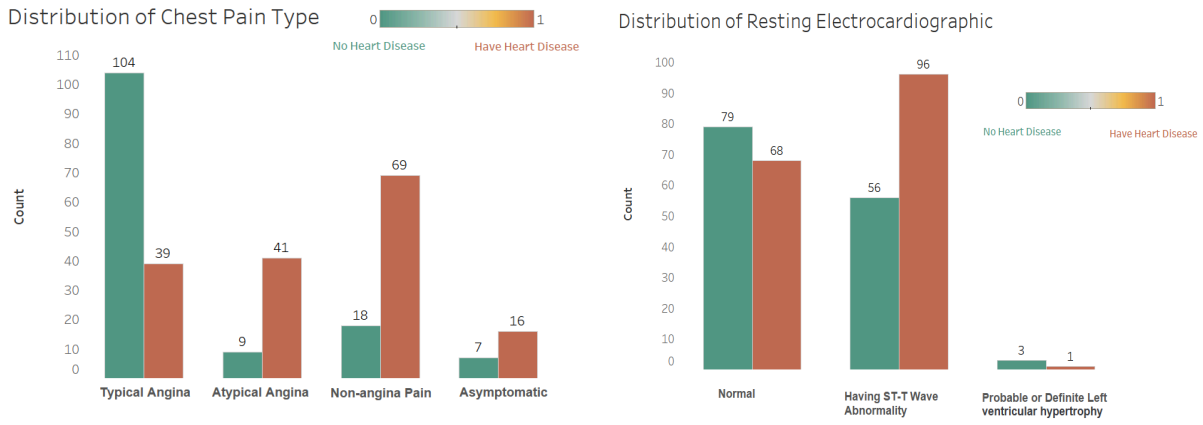
The box plot shows us that on average, those with heart disease have a higher heart rate than those without heart disease in the given dataset. The average maximum heart rate of those without heart disease is around 140 while the average for those with heart disease was near 150. The whiskers are

around the same length so it shows that the amount of variation in the maximum heart rates of the patients is around the same for both categories. This could imply that the heart must work harder for those with heart disease than without it, so it could potentially indicate a positive correlation between heart rate and heart disease.

C. Limitations

As mentioned numerous times above, these findings are only representative of the dataset, and not of the general population. It's also unknown how the maximum heart rate number was calculated; it could yield different results if the maximum heart rate is taken during exercise or after it. However, since this gives insight into a potential correlation between heart rate and heart disease, it still opens the door for more potential research on these two factors.

4. Exploring the Distribution of Chest Pain and Electrocardiographic Results in Relation to Heart Disease



A. Methodology

For chest pain types, we categorized individuals from the dataset into four distinct classifications—typical angina, atypical angina, non-anginal pain, and asymptomatic—alongside their heart disease status. We employed bar charts with contrasting colors to clearly delineate each category, thereby facilitating a straightforward statistical and visual examination. A similar approach

was applied to resting electrocardiographic results, dividing individuals based on their electrocardiogram (ECG) readings—normal, having ST-T wave abnormality, or showing probable or definite left ventricular hypertrophy—and their diagnosis of heart disease.

B. Findings

From the chest pain distribution analysis, it was observed that the majority of individuals with typical angina did not suffer from heart disease. In contrast, for the other three categories of chest pain, the number of individuals with heart disease constituted the majority, particularly in the atypical angina and non-anginal pain categories. Therefore, it appears that having atypical angina or non-anginal pain is a notable indicator of potential heart disease presence.

The distribution of resting electrocardiographic results revealed no substantial difference in the normal ECG category concerning heart disease prevalence. However, a significant majority of the individuals with ST-T wave abnormalities were found to have heart disease, suggesting that such an ECG finding may be indicative of underlying heart issues.

C. Limitations

Our analysis faced limitations, especially concerning the resting electrocardiographic results, where the group with probable or definite left ventricular hypertrophy was too small to deduce any substantial correlation between ECG results and heart disease presence. This limitation in sample size restricts the strength of any potential conclusions that can be drawn from this particular subset.

Last Remarks

Overall, the most insightful finding was that high blood pressure seems to be correlated with the chest pain that patients came into the hospital for. However, those with heart disease had a higher frequency of atypical angina and ST-T Wave abnormalities. This could imply that high blood pressure is a precursor of heart disease to some degree. We also found that a heightened heart rate seems to be

correlated with a higher chance of heart disease diagnosis. For people at home, we hope that this blog raises awareness around heart disease in general and the importance of going to the doctor for any level of chest discomfort or pain. Especially for those ages 40-60, it could be beneficial to keep track of your own blood pressure and heart rate. With more improved technology such as Apple Watches, this is easier than ever. Catching heart issues early will allow for potentially less severe symptoms.

For researchers, we hope that this blog encourages more research to be done regarding heart disease. There were certain findings that would need more research to conclude anything confidently. For example, is it that women are less likely to have heart disease or are they less likely to go to the hospital given symptoms of heart disease? This is one of many research questions that could be explored to further our understanding of heart disease. In searching for datasets to conduct our analysis, it was very difficult to find ones that would have enough instances that could be more representative of a larger population. Heart disease is a very hard topic to research due to the fact that it can be caused by so many different factors and can lead to sudden death from heart attack during the course of a study. Because of this, most data comes from hospitals, but even this is flawed because individuals have to choose to come into the hospital with no benefit of pay (unlike a formal study). However, if more hospitals could publicly publish the numbers that we discussed today, there could be more analysis done and a proper prediction model could be developed. Heart disease has been the leading cause of death in the United States for the past 100 years, but it doesn't have to be that way for the next 100 years.