Factors that Could Lead to Heart Disease

By:

LaShae Clarke

Felipe Flores

Bryan McElvy

Dataset Description:

The Heart Disease dataset originates from the UC Irvine Machine Learning Repository. The kaggle dataset used the same amount of features that was used in the UC Irvine dataset. However, it is important to note that the kaggle dataset is possibly a cleaner version because it has fewer rows, 270, than the UC Irvine dataset which had 303 rows. There were no null values displayed in the kaggle dataset. We chose this dataset because it was something new to learn about and we could explore the factors that possibly lead to heart disease. Out of the other datasets that we explored, this dataset provided concise information and interpretable data. Here is the link of the dataset that we will be using: https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var

When predicting heart disease, it is important to research the features and check their importance. Age is an important factor when considering heart disease. This dataset contains the ages of 29 to 77. Heart disease generally affects older people more than younger people [1]. More specifically, in a study of people that had heart disease, it affected 82% of people who were older than 65 [1]. Age and gender are also closely related to each other with respect to heart disease. Unfortunately, after menopause, from past data, “women’s death rate from heart disease increases” [1]. Even though it largely affects older individuals, it can still affect younger people too [1].

Gender is another factor associated with heart disease. In this dataset a value of 1 indicates that the person is a male and a value of 0 indicates that the person is a female. As women age, it is more common for them to have heart disease as opposed to men, who start to experience it at an earlier age than women [2]. Science Direct states that “As for common factors, age, hypertension, total cholesterol and low-density lipoprotein (LDL)-cholesterol have a great influence in men. In contrast, smoking, diabetes, triglyceride and high-density lipoprotein (HDL)-cholesterol levels mainly have effect on women.” [3]. The difference in what correlates more to heart disease with respect to gender is very important to take into account.

Chest pain is another feature explored in this dataset. In this dataset, patients experienced 4 different types of chest pains. A value of 1 indicates that they experienced typical angina, a value of 2 indicates that they experienced atypical angina, a value of 3 indicates that they experienced non-anginal pain, and a value of 4 indicates that they were asymptomatic. People who experience an angina chest pain normally have little amounts of blood traveling to their heart [4]. Atypical angina “can also be caused by non-cardiac causes, such as musculoskeletal issues or because of a psychiatric condition” [4]. Non-anginal pain is “recurring pain in your chest — typically, behind your breast bone and near your heart — that is not related

to your heart" [5]. Asymptomatic chest pain patients tend to experience "three or four times that of anginal attacks" [6]. This relates to the target variable because if a patient has barely any blood traveling to the heart it could lead to heart disease.

Another feature researched is blood pressure. Blood pressure is defined as "the pressure of blood pushing against the walls of your arteries" [7]. In this dataset we will be specifically looking at the systolic blood pressure of the patient. This blood pressure specifically "measures the pressure in your arteries when your heart beats." [7] A normal systolic blood pressure is normally defined as less than 120 mm Hg. An at risk for high blood pressure patient usually has a blood pressure around the range of 120 – 139 mm Hg. A high blood pressure would be indicated through a systolic blood pressure reading of 140 mm Hg or higher. When a patient has a higher blood pressure than the normal reading it could lead to future heart disease [7]. This is a really important statistic for us to know because looking at the dataset, we could determine if that blood pressure is considered high or low.

The next feature explored is cholesterol. This column represents the total amount of cholesterol in the patients blood. Too much cholesterol can clog the arteries, making it harder for blood to travel throughout the cardiovascular system, which could lead to a higher risk for heart disease. In this dataset the patients cholesterol levels range from 126 to 564. A total cholesterol of less than 200 is considered to be a good cholesterol level. A cholesterol level between 200 – 239 is close to being classified as a high cholesterol. And a cholesterol level that is above 240 is considered high [8]. As stated in the gender and heart disease research, cholesterol can be a factor in what causes heart disease.

Fasting blood sugar (FBS) is also a feature present in the database. FBS is a diagnostic test in which the patient abstains from eating or drinking anything but water for 8-12 hours before having their blood glucose content tested. Although it's generally used to screen for diabetes, it stands to reason that heart disease might affect the blood glucose content of an afflicted patient, hence why it was included as a feature in this database [https://my.clevelandclinic.org/health/diagnostics/21952-fasting-blood-sugar]. For this database, FBS is a categorical variable, where a value of 1 refers to FBS over 120 mg/dL – which would be considered high blood sugar – and a value of 0 refers to FBS below that value, which is normal, non-pathological behavior.

EKG results are also included as a feature in this database. EKG (or ECG) refers to electrocardiography, which is a recording of the electrical activity of the heart. In terms of this database, EKG results are a categorical variable with three possible values: 0 for patients with normal EKG, 1 for patients with an abnormal ST segment and/or T wave, and 2 for patients

who, by Estes' criteria, have probably or definite left ventricular hypertrophy [https://www.mayoclinic.org/tests-procedures/ekg/about/pac-20384983][https://archive.ics.uci.edu/ml/datasets/Heart+Disease].

Heart rate is another feature within this dataset. Heart rate refers to the amount of cardiac cycles per unit time. The normal range of human heartbeats is 60-100 beats per minute (bpm). A heart rate above (tachycardia) or below (bradycardia) this range is considered pathological [https://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979][https://archive.ics.uci.edu/ml/datasets/Heart+Disease]. In this dataset, heart rate is a numerical variable.

Exercise angina is a feature of the dataset that corresponds to whether or not the patient suffered chest pain as a result of reduced blood flow to the heart during exercise. Exercise angina is a known biomarker for heart disease, which is likely why it was included in this dataset. Exercise angina is a categorical variable, where a value of 1 refers to patients that suffered from it and a value of 0 refers to patients that did not [https://www.mayoclinic.org/diseases-conditions/angina/symptoms-causes/syc-20369373][https://archive.ics.uci.edu/ml/datasets/Heart+Disease].

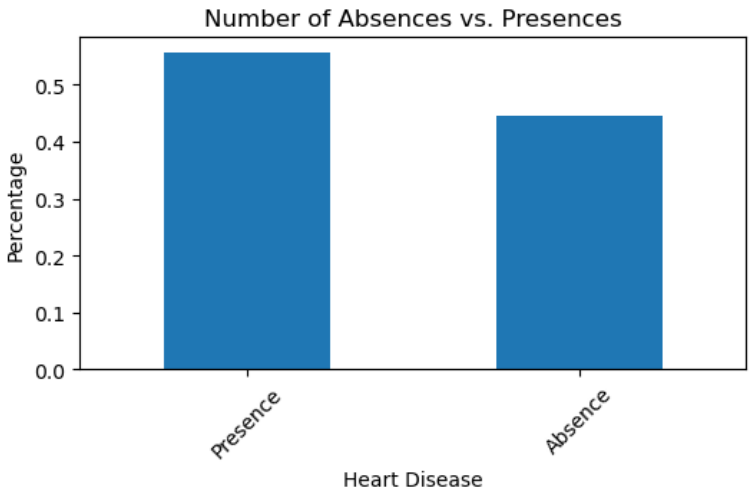
ST depression along with ST slope are other features in the data set that are similar. For depression and slope, these features look at the EKG readings and are looking for segments in the test that appear abnormally low and below the patient's baseline. From here it would categorize the occurrence numerically with a 1 mean upsloping depression, a 2 for a flat depression, or 3 for a downward sloping in ST slope. The ST depression values are measures of how much these slopes deviate from the baseline. ST depression is thought to be a more accurate EKG in regards to diagnosing heart disease[9]. However, in a heatmap that we created it seems to not be the case. While the two features have high correlations with one another they do not have a high correlation with the heart disease classification.

Thallium is an attribute that measures the patient's blood flow during exercise or at rest[10]. The attribute is categorized into 3 categories which are normal, fixed defect, and reversible defect. Fixed defect and reversible defect are used to signify that there are problems with the patient's blood flow.Leading to the patient being diagnosed with a heart disease. This along with the high accuracy of Thallium scans being true leads to a high correlation of the patients having heart disease.

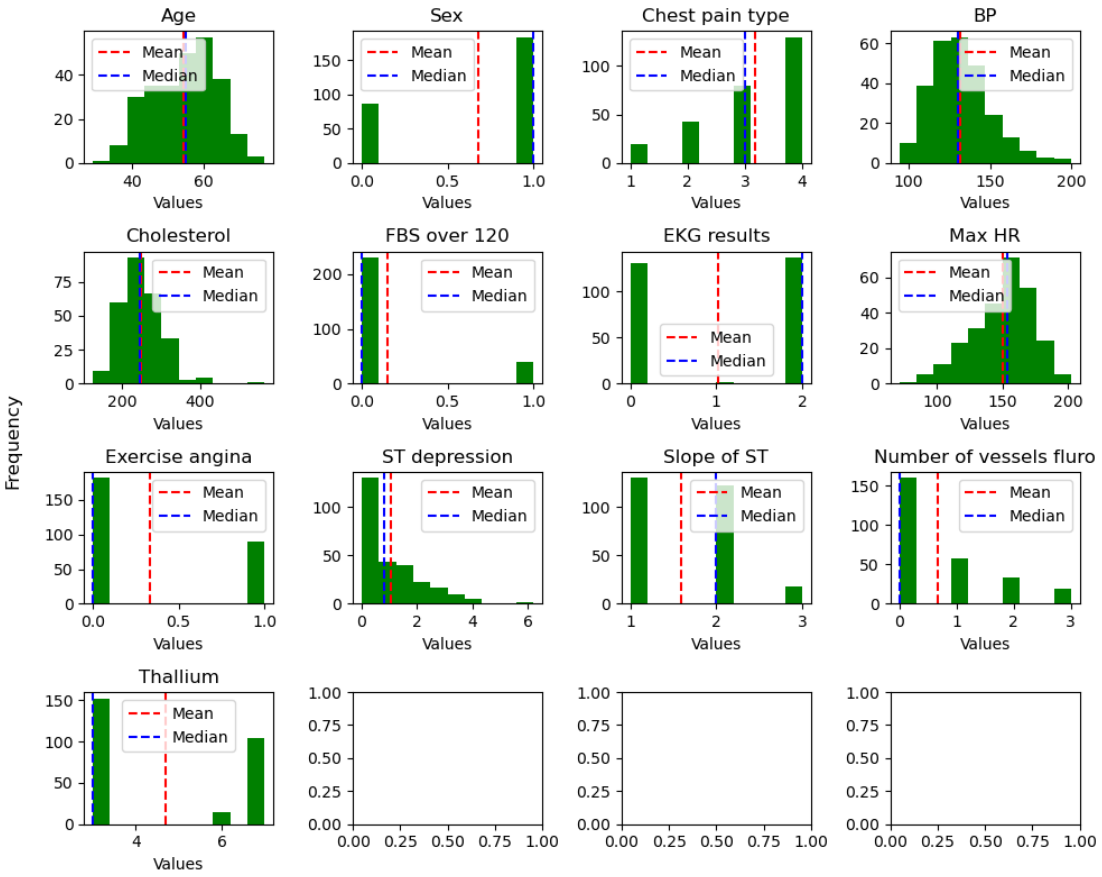
The last feature is the number of vessels Fluro which is the amount of vessels that are seen in fluoroscopy images.[12] This attribute has values that range from 0-3 with the more

vessels seen increasing the likelihood that a patient has a heart disease. However, this does not seem to be something that is medically proven as of yet but still can hold some correlation. But from correlations that have been made this does seem to be a possible conclusion.

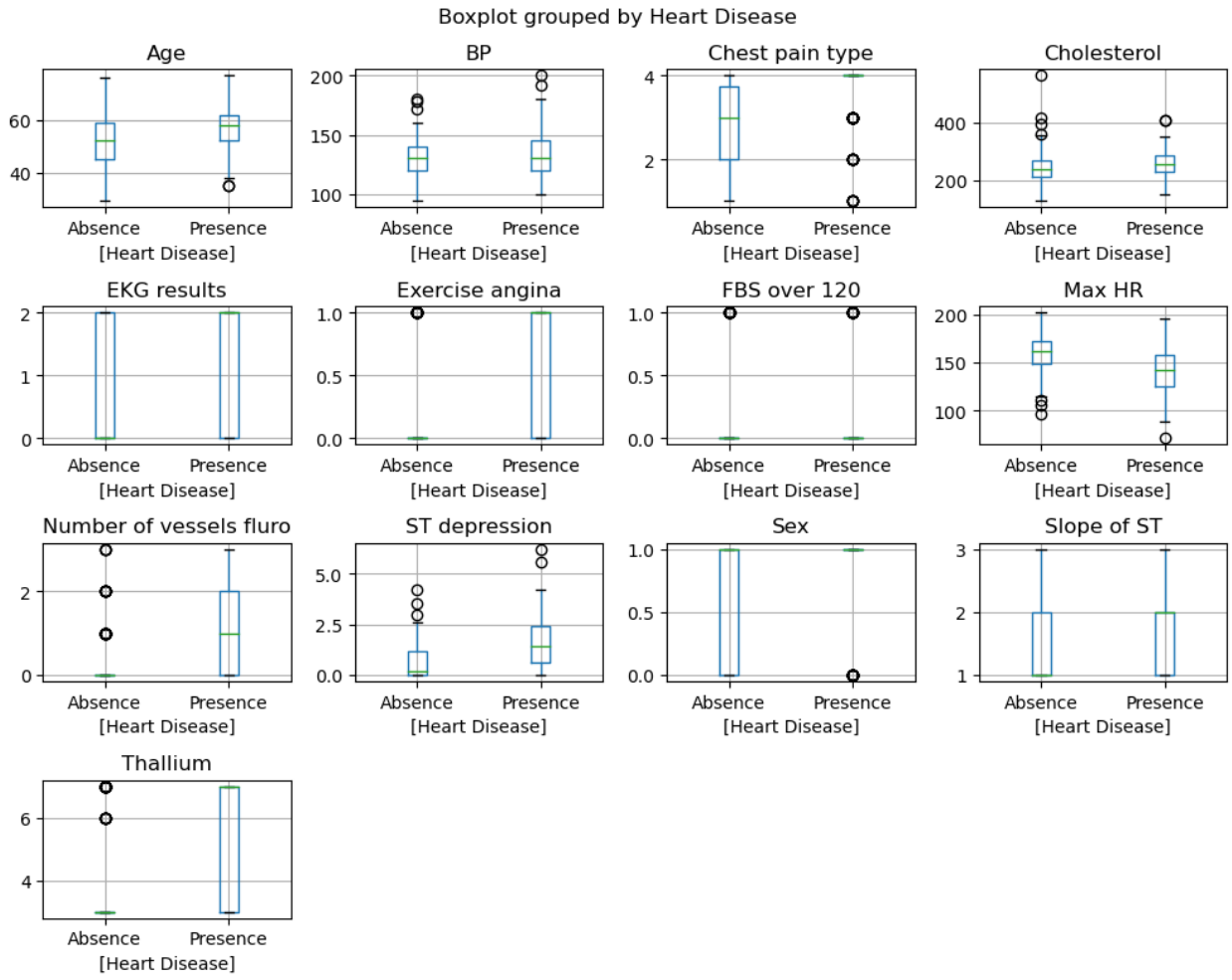
Exploratory Data Analysis



In this dataset more than 50% of people had heart disease. So we can infer that the majority of the patients exhibit multiple factors that could lead up to heart disease. This also shows that this dataset was able to identify people with heart disease pretty well, with the features given.



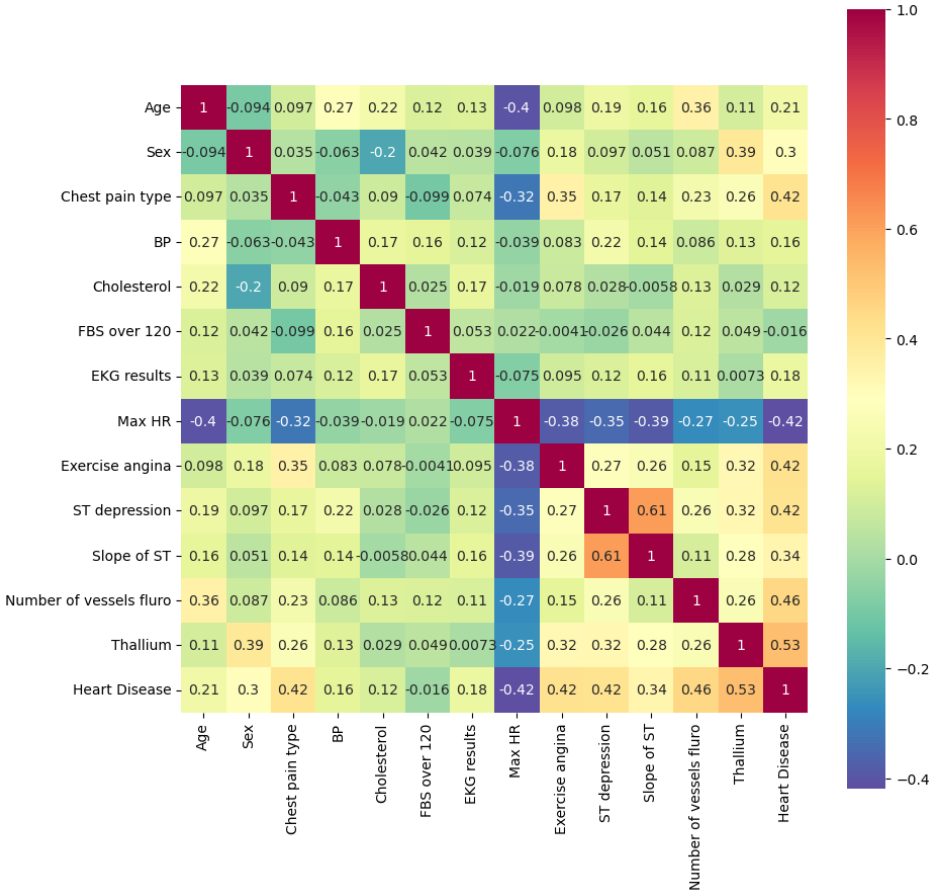
As seen by the previous image it is important to note that more than 50% of the patients had heart disease. Also, the research above indicates that older individuals are more likely to have heart disease than younger individuals. It's no surprise that the meridian of the data is a bit lower than 60. For gender, we see mostly men were patients in this data collection. More people experienced asymptomatic angina, where they did not experience chest pain, however, they still had angina. Again, this aligns with knowing that more than 50% had heart disease. A high blood pressure is considered to be above 140 mm Hg. The median of this dataset is 130 mm Hg, indicating that a lot of the patients could be at risk for heart disease, as indicated by past research. High cholesterol can be another factor that could lead to heart disease, which would be defined as over 240. In the dataset, the median is 245. This also indicates that a person could be at risk for heart disease.



From the age boxplot, it displays the median at a higher level of the patients who had heart disease versus the patients that did not. This shows that patients who tended to be older had heart disease on average. Shockingly the median blood pressure was the same for patients with heart disease and without. Also, the higher blood pressure on both box plots tended to be outliers. For chest pain a range from 1 - 4 did not have heart disease. However, most people who had heart disease experienced a chest pain of 4. The median for cholesterol is higher for people with heart disease. It is important to note that some people who did not have heart disease were outliers, indicating that a person could have high cholesterol but still not have heart disease. As for the ST depression we can see that most of the patients with heart disease tend to have an upsloping depression or flat depression, but this trait also seems to be shared with patients without a heart disease as patients without heart disease tend to have a upsloping depression. Slope of Thallium stress test shows that patients with either fixed or reversible defects tend to have a heart disease. While the ones that have a normal reading then do not be

diagnosed with heart disease. As for Fluoroscopy it can be seen that the majority of patients with 1 vessel seen have heart disease and the majority of patients with none seen are clear. There are also some outliers that have 1 or more vessels seen that don't have heart disease.

Feature Engineering:



Above is a heatmap that directly describes the feature relationships with each other and the target variable. FBS over 120 should be removed because it does not have a strong relationship with heart disease as it is - 0.016. We believe due to moving this one feature, it will make for a future stronger model. 3 more features that we are considering removing are cholesterol, EKG results, and blood pressure. The correlation between these features and heart disease are all below 0.2. We will test our model with and without these features to see if they improve or skew its results. Another noteworthy feature is Thallium as it has a high correlation with heart disease. Which is something that we want to look into in the future to determine why this is. As of the current moment we believe that it is due to Thalliums high accuracy in

determining heart disease in the medical field. As it gives a look into someone's vesicular system.

Sources

[1] https://memorialhermann.org/services/specialties/heart-and-vascular/healthy-living/education/heart-disease-and-age#:~:text=Your%20risk%20for%20heart%20disease,under%20the%20age%20of%2065.

[2] https://www.sciencedirect.com/science/article/pii/S2590093519300256

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605/

[4] https://www.ppschicago.com/pain-management/chest-pain/atypical-chest-pain/

[5] https://my.clevelandclinic.org/health/diseases/15851-gerd-non-cardiac-chest-pain

[6]https://pubmed.ncbi.nlm.nih.gov/2238747/#:~:text=Abstract,or%20the%20usual%20anginal%20equivalents.

[7] https://www.cdc.gov/bloodpressure/about.htm

[8] https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk

[9]https://pubmed.ncbi.nlm.nih.gov/3739881/#:~:text=The%20ST%20segment%20shift%20relative,coronary%20artery%20disease%20(CAD).

[10]https://centennialheart.com/service/thallium-stress-test

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221136/#:~:text=Non%2Dreversible%20defect%20(fixed),myocardium%20or%20prior%20nontransmural%20MI.

[12]https://pubmed.ncbi.nlm.nih.gov/7005930/#:~:text=Cardiac%20fluoroscopy%20is%20an%20inexpensive,diagnosis%20of%20valvular%20heart%20disease