STAT 650: Statistical Foundation for Data Science

Exploring Patient Profiles: Analyzing Health Indicators and Heart Disease Relationships Using Data Mining Techniques

Term Project

Submitted by:
Pranik Koirala

Submitted to:
Dr. Dixon Vimalajeewa

2022

Abstract

In an effort to improve patient health outcomes, the relationship between various health indicators and heart disease were measured via the exploration of a patient profile dataset. In order to properly analyze the dataset, a variety of data mining techniques were necessary. The data consisted of many outliers for different variables, and categorical data was represented in an unorthodox manner, in which higher values did not represent the significance of the variable. We used box plots, interquartile ranges, and column mean imputations to find and resolve the bias caused by outliers. To establish relational inferences between the variables, a correlation matrix was made using phi-k values. Also, bar plots and contingency tables were constructed to address the misleading correlations resulting from categorical data. Through this process, it was found that patients with high maximum heart rate, atypical angina or non-anginal chest pain, or with 0 blood vessels colored by fluoroscopy have higher incidences of heart disease.

Introduction

According to the Centers for Disease Control and Prevention (CDC), heart disease is the leading cause of death.[1] One person dies every 34 seconds in the United States from cardiovascular disease and one of five deaths is caused due to heart disease.[1] Identifying heart disease in its earliest stages could help millions of patients who don't realize that they are at risk from this potentially life-threatening condition. Through analysis of this dataset, we aim to get a better understanding of the relationship between a patient's medical profile and their risk of getting a heart condition. In order to successfully do this, it is important to have proper data collection, cleaning, and analysis skills. Developing a successful data mining approach would aid doctors in ascertaining a patient's risk for heart disease and allow them to prescribe preventative treatments for better patient health outcomes.

Data

The data was collected from Kaggle:
https://www.kaggle.com/code/balmeetkaur/heart-disease-data-insights-and-predictions/data
The dataset has the medical profiles of 1,025 patients that consist of 14 variables that relate to the health of each patient. The 'target' variable, which indicates whether a patient has heart disease or not, is the dependent variable in our analysis. The 13 other variables will be used as explanatory variables to see if a certain trait increases a patient's risk of heart disease. Information regarding each of these variables can be found in Table 1.

Variable	Variable Type	Explanation	Values
age	Numerical	The person's age in years	
sex	Categorical	The person's sex	1: male, 0: female
cp	Categorical	The chest pain experienced	0: typical angina, 1: atypical angina, 2: non anginal pain, 3: asymptotic[2,3]

[1] CDC (2022), Heart Disease Facts

[2] Heart Disease Prediction: https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

[3] Heart Disease Dataset (UCI): https://archive.ics.uci.edu/ml/datasets/heart+disease

trestbps	Numerical	Resting blood pressure on admission to the hospital in mm Hg	
chol	Numerical	The person's cholesterol measurement in mg/dl	
fbs	Categorical	The person's fasting blood sugar > 120 mg/dl	1: true, 0: false
restecg	Categorical	Resting electrocardiographic measuring results	0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Numerical	The person's maximum heart rate achieved	
exang	Categorical	Exercise induced angina	1: yes, 0: no
oldpeak	Numerical	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)	
slope	Categorical	The slope of the peak exercise ST segment	0: upsloping, 1: flat, 2: downsloping
ca	Categorical	Number of major vessels colored by fluoroscopy	0, 1, 2, 3, 4
thal	Categorical	A blood disorder - Thalassemia	0: NULL (dropped from the dataset previously), 1: fixed defect (no blood flow in some part of the heart), 2: normal blood flow, 3: reversible defect (a blood flow is observed but it is not normal)
target	Categorical	Heart disease	1: yes, 0:no

Table 1. Variable and values of each variable

The authors of the data requested to include the names of the investigators responsible for data collection: Andras Janosi, M.D. (Hungarian Institute of Cardiology. Budapest); William Steinbrunn, M.D. (University Hospital, Zurich, Switzerland); Matthias Pfisterer, M.D. (University Hospital, Basel, Switzerland); Robert Detrano, M.D., Ph.D. (V.A. Medical Center, Long Beach and Cleveland Clinic Foundation). The dataset used in this report has no privacy concern as it does not contain any personal information such as patients name, race, or address. Furthermore, this analysis is ethical as such analysis can help both patients and doctors in identifying individuals having high risk of heart disease and recommend suitable diagnosis afterwards.

Results

Out of 1,025 patients in the data, 69.5% are male and 30.4% are female. In addition, 51.3% of the patients have a heart disease (target = 1). We find that of the male patients 42.07% and of the female patients 72.43% have heart disease. Based on the summary statistics, we can see that the numerical variables for age and a person's maximum heart rate achieved (thalac) are left-skewed. The underlying reason behind why the variable thalac is left-skewed is that younger people in general exhibit higher heart rates (shown in figure 1). On the other hand, the variables for blood pressure, cholesterol, and ECG peaks induced by exercise are right-skewed.

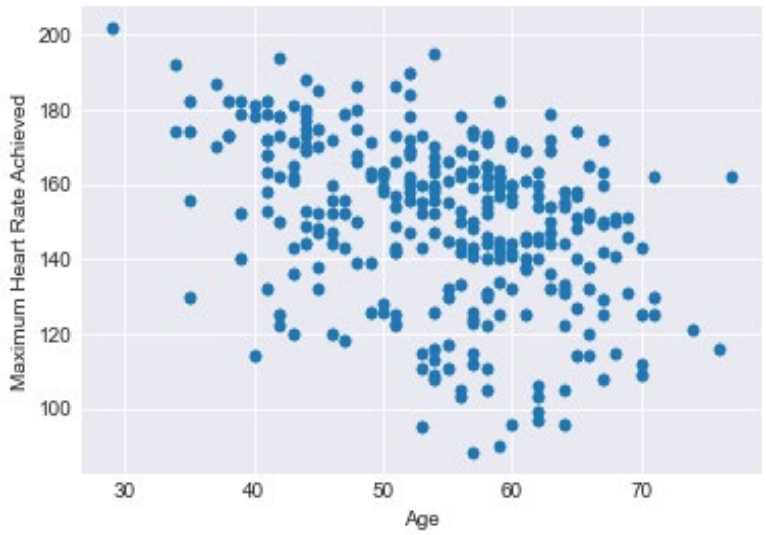


Figure 1: Scatter plot of maximum heart rate achieved and age

	age	trestbps	chol	thalac	oldpeak
Mean	54.4341	131.6117	246	149.1141	1.0715
Standard Deviation	9.07229	17.5167	51.5925	23.0057	1.1750
IQR	13	20	64	34	1.8
Range	48	106	438	131	6.2
Max	77	200	564	202	6.2
Min	29	94	126	71	0
Skewness	-0.2485	0.7386	1.0725	-0.5130	1.2091
Median	56	130	240	152	0.8

Table 2: Summary Statistics of Numerical Variables

We check for outliers[4] in the numerical variables visually with boxplots and quantitatively. There are in total 57 outlier observations for the numerical variables of trestbps, chol, thalach, and oldpeak. Thus, to handle these, we replace the outliers with NaNs and imputed NaNs with column means. After doing so, because of its heavy right tail, the variable oldpeak has new outliers towards the upper bound. Therefore,

[4] Outliers are defined as any value greater than 3rd Quantile + (1.5 * IQR) and lower than 1st Quantile - (1.5 * IQR)

we re-imputed the oldpeak variable with column mean once again. After this, we get the oldpeak variable without outliers, but with its distributional shape intact.

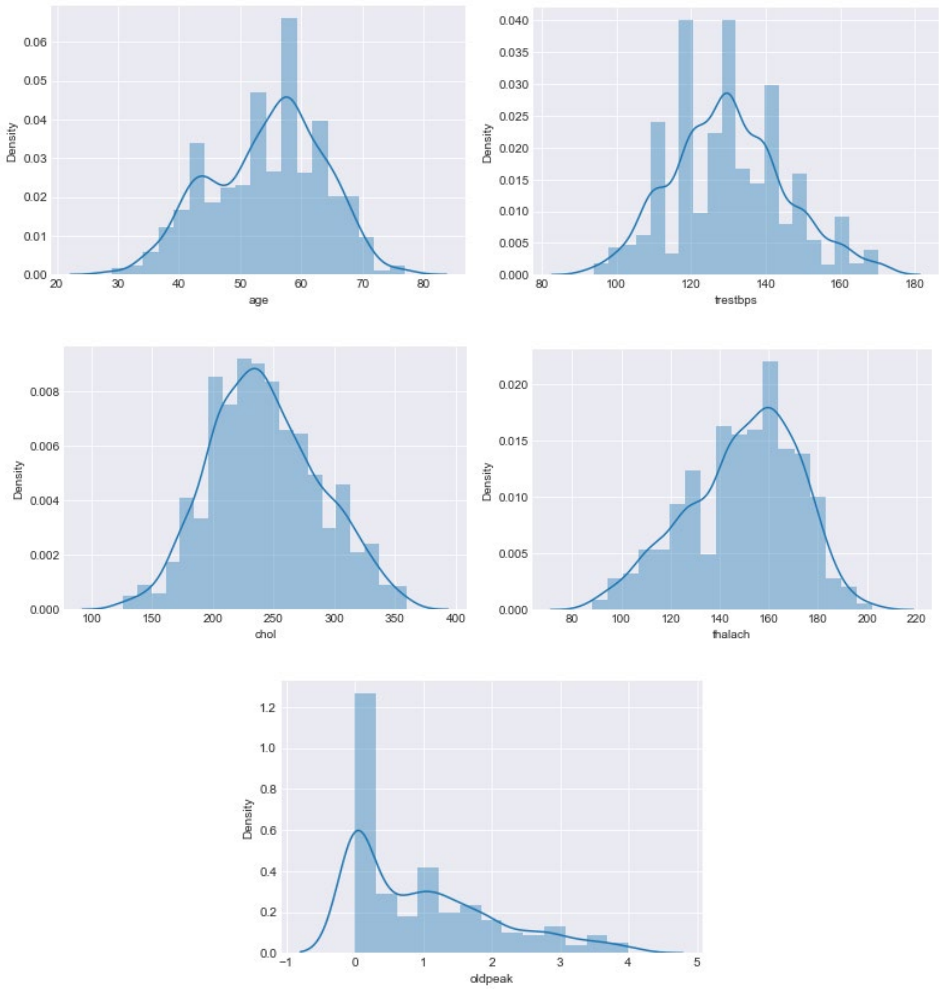


Figure 2: Histogram Plots of Numerical Variables

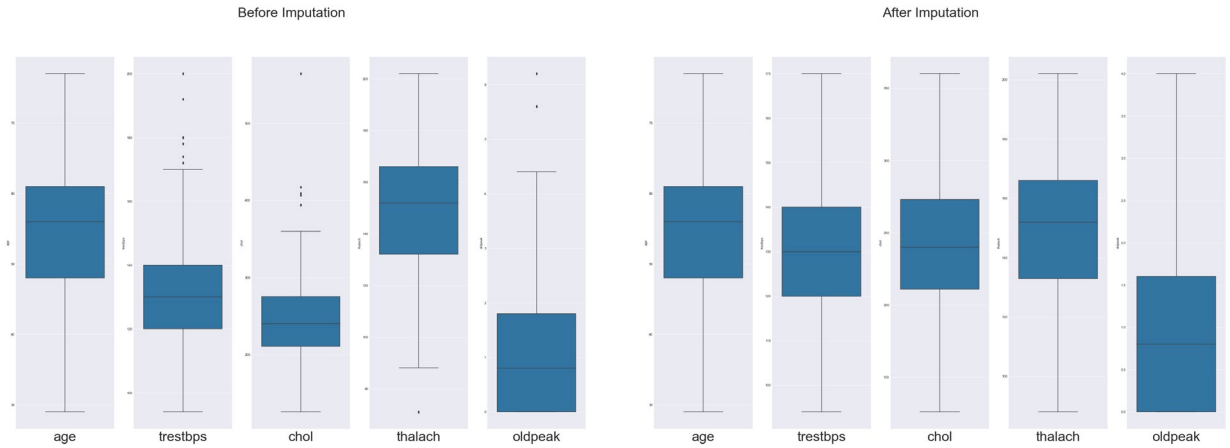


Figure 3: Boxplots of Numerical Variables (left: before imputation and right: after imputation)

Correlations among variables

Figure 4 shows correlations among all variables using Phi_k criteria. Unlike Pearson's and Spearman's correlations, Phi_k shows correlations for categorical variables while also capturing non-linear dependencies. This correlation matrix does not provide direction (positive/negative) of the relation as it is scaled between 0 and 1. Phi_k criteria has been utilized as the dataset contains many categorical variables. Categorical variables ‘cp’ (chest pain type) and ‘thal’ (presence of chest pain) have the highest correlation with ‘target’ (heart disease). However, due to the nature of the categorical expressions, a higher ‘category’ does not necessarily correlate to a higher risk of heart disease. For example, a ‘cp’ value of 3 indicates a patient that is asymptomatic for chest pain. This is not as indicative of heart disease as a cp value of 2, which highly correlates with heart disease, as displayed in figure 5 below. The same can be said of ‘thal’, which is further discussed in the contingency table section below. The variables ‘oldpeak’ (peaks in ECG) and ‘exang’ (exercise angina) have the highest correlation with heart disease among the continuous variables.

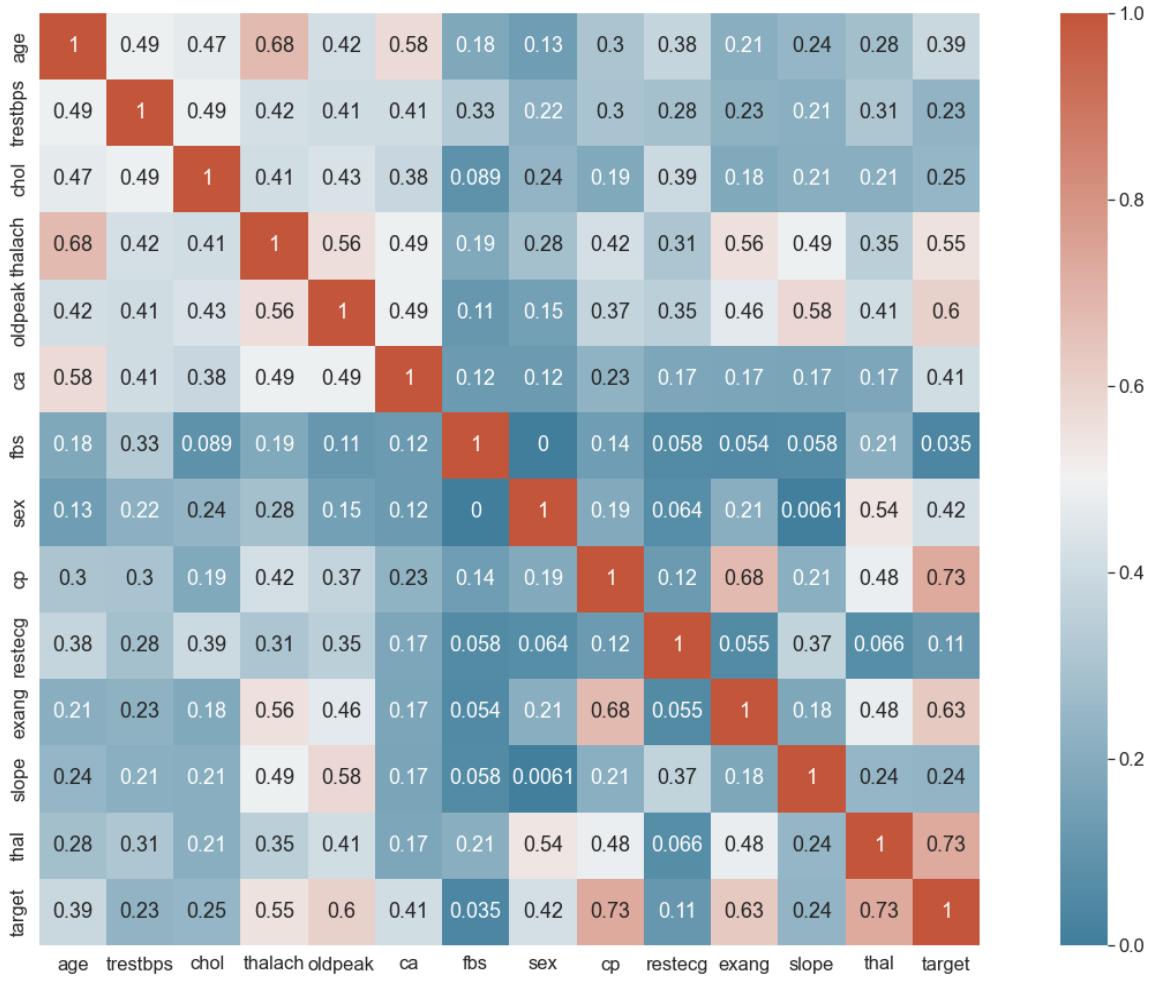


Figure 4: Heatmap to visualize relationship between all features utilizing Phi_k criteria

To better understand the correlation between heart disease and other categorical variables, we utilized bar plots as shown in Figure 5. In the dataset, the females are more prone to have heart disease. Opposite to common conviction, high fasting blood sugar has a very weak correlation with heart disease. Among patients with high fasting blood pressure, more than 50% are free of heart disease. The variable restecg = 1 representing presence of ST-T wave abnormality in the ECG report is more prevalent among patients having heart disease. It is also surprising that many patients have heart disease but they do not experience any exercise angina (exang = 0). So, people may keep developing heart disease without noticing visual effects such as exercise angina. The downsloping of the variable slope (the slope of the peak exercise ST segment) has a very high correlation with developed heart disease. Most heart patients have 0 major vessels colored by fluoroscopy (ca = 0). Patients having normal blood flow (thal = 2) have high association with heart disease which is counter intuitive. Chest pain (cp) is highly correlated with heart disease in general. Among types of heart pain, patients with non-anginal pain (cp = 2) are most likely to have heart disease.

Figure 6 shows relationships between heart disease and continuous variables utilizing box plots. The average age for patients with heart disease is lower than that of patients free of heart disease. Age has weak correlation with heart disease. Similarly, resting blood pressure (trestbps) and cholesterol (chol) have weak correlations with heart disease. Patients with high maximum heart rate (thalach) are more likely to have heart disease. Lower value of ST depression induced by exercise relative to rest (oldpeak) has a strong correlation with heart disease.

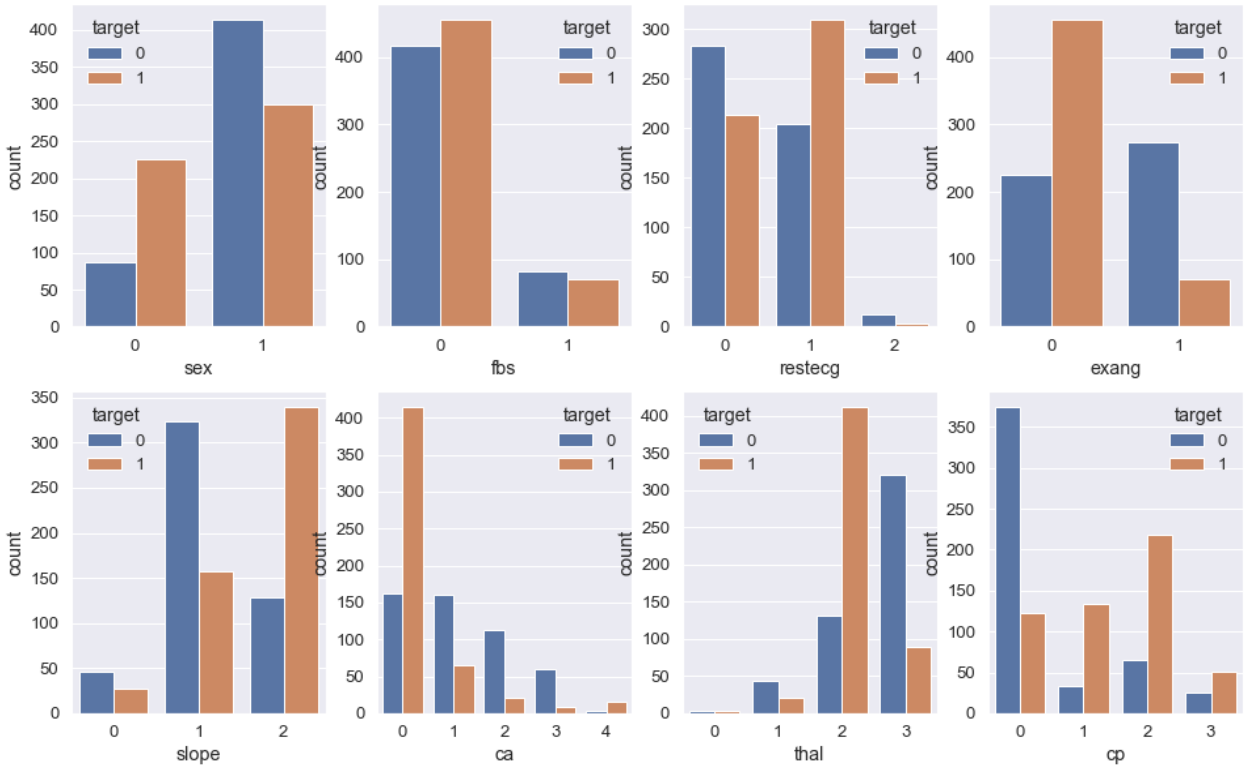


Figure 5: Correlation between heart disease and other categorical variables

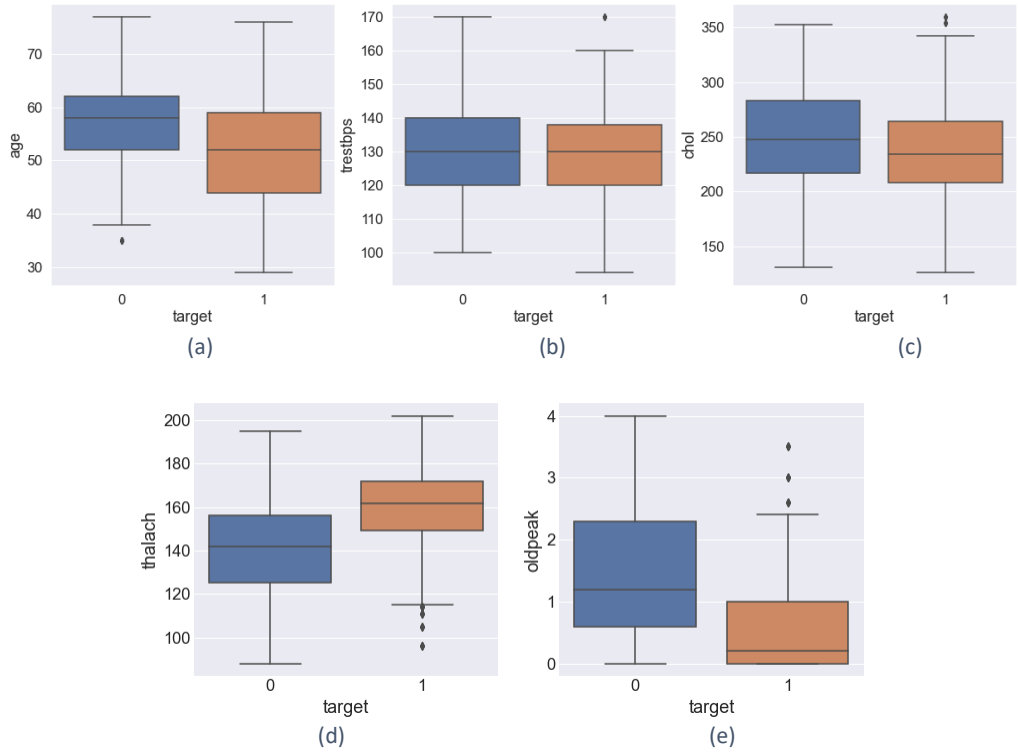
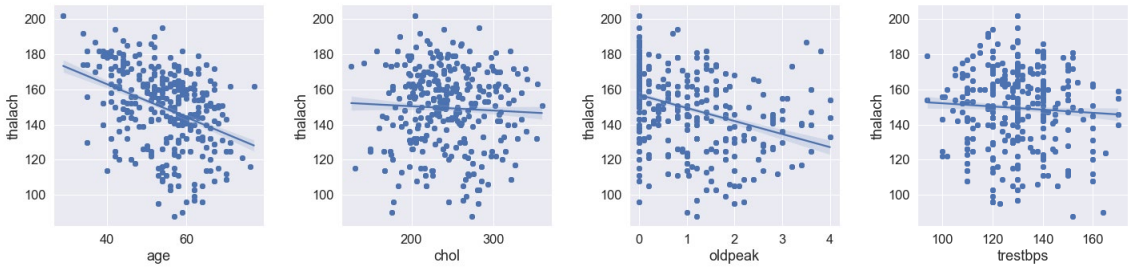


Figure 6: Correlation between heart disease and continuous variables. (a) age, (b) resting blood pressure, (c)cholesterol, (d) max heart rate, (e) ST depression induced by exercise relative to rest (ECG plot)



(a) *(b)* *(c)* *(d)*

Figure 7: Correlation between maximum heart rate and other continuous variables

As stated earlier, the maximum heart rate is highly correlated with heart disease and it is also one of the variables that patients can control using lifestyle change and/or medications. Due to this relative importance, the correlation of maximum heart rate with other continuous variables are shown in figure 7 utilizing lmplot. All of them have weak linear correlation with maximum heart rate. No nonlinear relationship was observed from the lmplot of continuous variables. Among the variables few of them are controllable by lifestyle changes and/or medicines. Patients should try to keep cholesterol and maximum blood pressure under control.

Contingency Tables and Probabilities

The contingency tables below show the percentage distribution of different variables within the target variable. Each variable is separated with respect to having heart disease or not having heart disease and the percentage of different classes within the variable are mentioned. Some variables such as Resting blood pressure (trestbps) showed weak correlation with heart disease in the box plot analysis and were excluded from the Probability/Contingency tables.

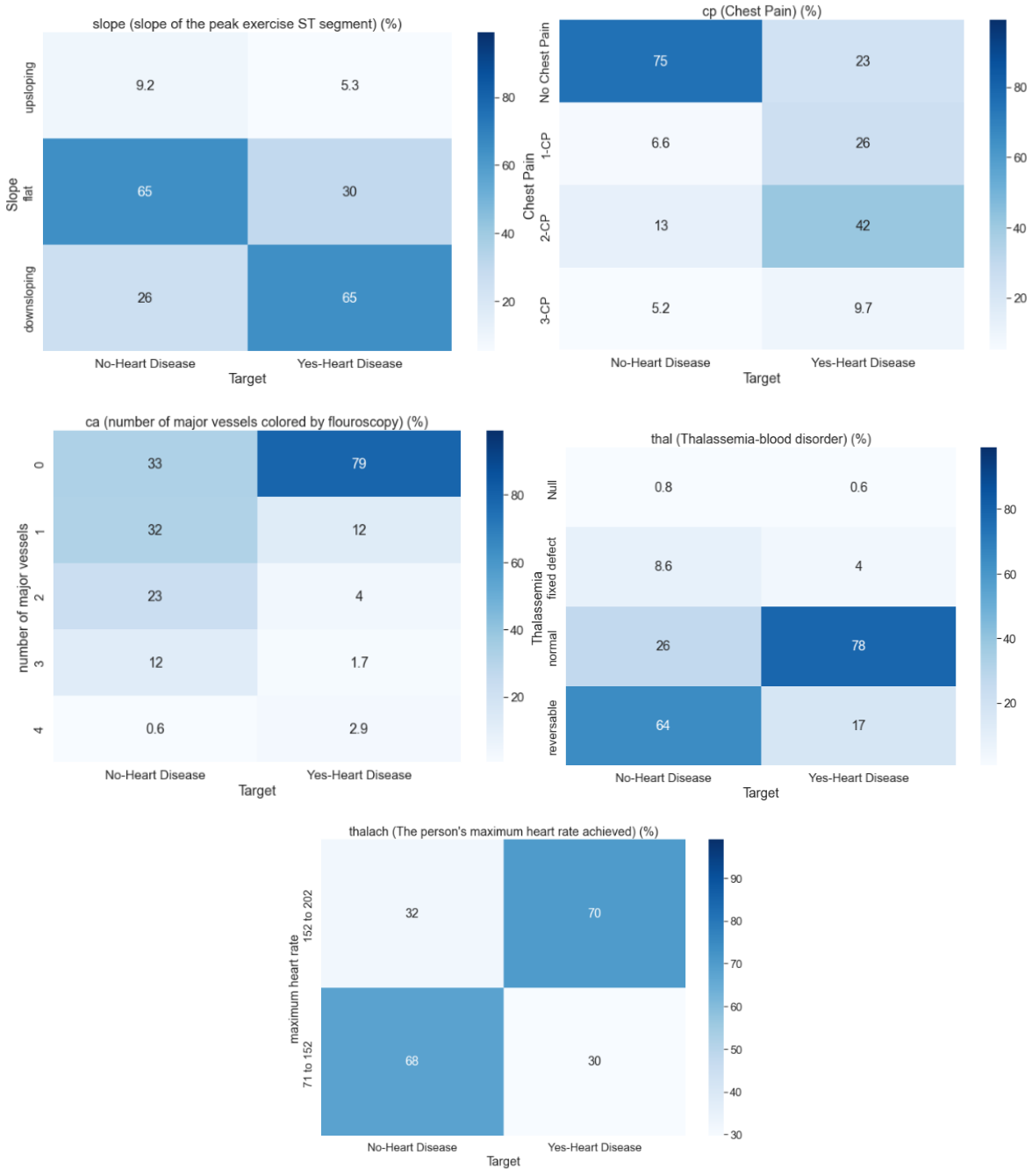


Figure 8: Contingency Tables for correlated variables on heart disease as percentages

Chest pain: From figure 8, people with one of the three types of chest pains combined to make only 24.8% of the sample size, however, alarmingly, patients with chest pains combined to a total of 77% of the total number of heart disease indicating a strong contingency between the two variables. In particular, patients with chest pain type 1 and chest pain type 2 (non-anginal pain) have a higher chance of having heart disease than patients with chest pain type 3. Thus, chest pain can be considered as a good precursor for determining if the patient has heart disease.

Slope of peak exercise ST Segment: The distribution of different slopes of the peak exercise ST segment also shows differences in patients with heart disease and no heart disease. The contingency table shows that patients with downslope of the peak exercise ST segment are more prone to have heart disease; 65% of the patients with heart disease have a downslope whereas only 26% of patients with no heart disease have a downslope. Also, patients with flat slopes are more likely to not have heart disease according to the contingency table.

No. of Major Vessels Colored by Fluoroscopy: The distribution of variable 'ca' (i.e., the number of major vessels colored by fluoroscopy) is found to be different in two target classes. 79% of patients with zero major vessels colored by fluoroscopy have heart disease.

Thalassemia-Blood Disorder: Looking at the contingency table/heatmap The distribution of A blood disorder - Thalassemia in 'Yes- Heart Disease' and 'No- Heart Disease' is clearly not similar. However, the investigation does not show patients with fixed defects having heart disease. In contrast, the contingency table shows normal patients in terms of Thalassemia having heart diseases. This gives us the insight that Thalassemia is not a good precursor of detecting heart disease in a patient.

Maximum heart rate achieved: The maximum heart rate data is converted into categorical data by dividing the whole dataset into two groups: having less than and more than 152 heart rate. The contingency table shows that patients with a maximum heart rate more than 152 constitute 70 % of the people with heart disease. Likewise, 68% of the people with no heart disease have a maximum heart rate less than 152, indicating maximum heart rate can indicate a person's chance of having heart disease.

From Figure 8, it is evident that some of the variables such as maximum heart rate and certain types of chest pains (1 and 2) have strong contingencies with the target variable and are therefore good predictors for heart disease. However, none of the variables have a strong enough contingency to predict heart disease on its own and thus contingencies from multiple variables should be considered when predicting heart disease for greater accuracy.

Conclusion

Relationships between all categorical and continuous variables are explored and illustrated as such data exploration can be useful in promoting patient awareness and assisting doctors in evaluating patients' risk of heart disease more accurately. Different data visualizations, such as count plots, boxplot and contingency table all showed the same trends and relationships among the variables. Doctors and patients should be extra vigilant observing highly correlated features, such as maximum heart rate, chest pain type and major vessels colored by fluoroscopy. In order to understand more on heart disease risk factors, such analysis should be repeated when more recent data is available. Since model building was out of scope, the study did not make use of any predictive models but more work can be done in this area.