P8106 Midterm Project

Yunlin Zhou

Introdution

Motivation

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Using this data set, we would like to explore how those features related to the heart disease thus we can use them to predict a possible heart disease.

Data preparation and cleaning

As the table shows below, the data set has 11 predictor variables and 1 outcome variable(heart_disease), with 918 observations. When cleaning the data, we use factor() to change the type of character variable. There is no missing data in this data set. For better using this data set to train the models, we split the data set into two parts: training data (70%) and test data (30%).

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: heart disease or Normal

Table 1: Data summary

Name	Piped data
Number of rows	918
Number of columns	12
Column type frequency:	
character	6

numeric	6
Group variables	None

Variable type: character

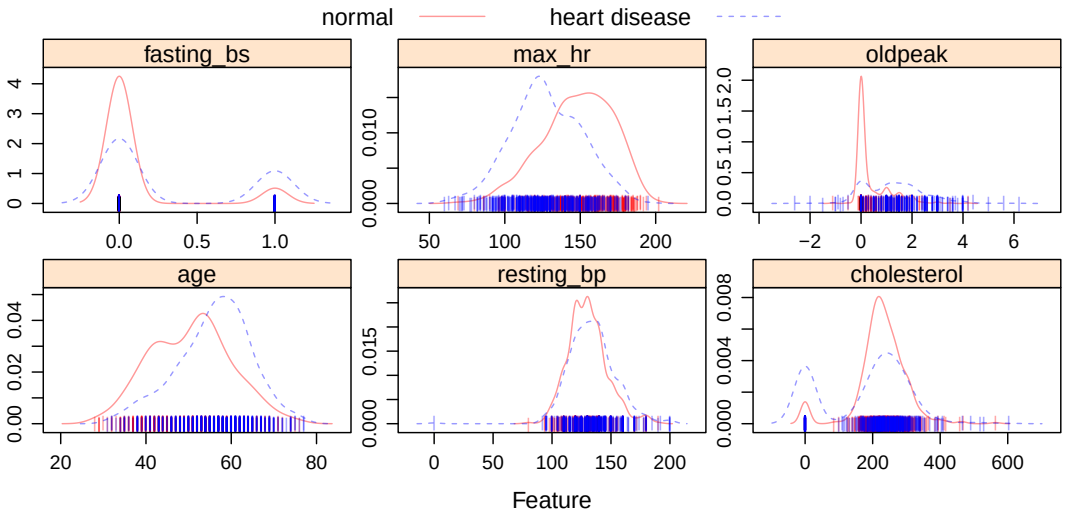
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
heart_disease	0	1	6	13	0	2	0
sex	0	1	1	1	0	2	0
chest_pain_type	0	1	2	3	0	4	0
resting_ecg	0	1	2	6	0	3	0
exercise_angina	0	1	1	1	0	2	0
st_slope	0	1	2	4	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	53.51	9.43	28.0	47.00	54.0	60.0	77.0	
resting_bp	0	1	132.40	18.51	0.0	120.00	130.0	140.0	200.0	
cholesterol	0	1	198.80	109.38	0.0	173.25	223.0	267.0	603.0	
fasting_bs	0	1	0.23	0.42	0.0	0.00	0.0	0.0	1.0	
max_hr	0	1	136.81	25.46	60.0	120.00	138.0	156.0	202.0	
oldpeak	0	1	0.89	1.07	-2.6	0.00	0.6	1.5	6.2	

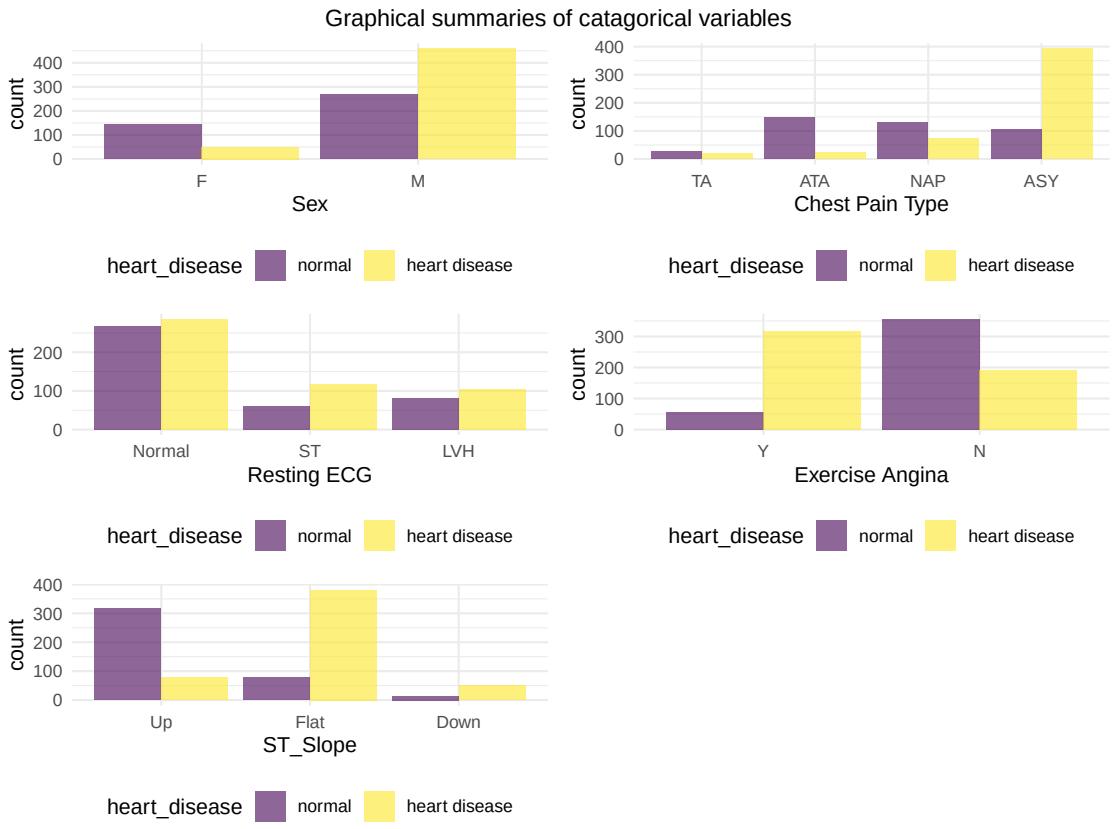
Exploratory analysis/visualization

Graphical summaries of continous variables



From the density plot of continuous variables above, we can see that the some features like oldpeak, have differences between the normal and heart-diseased people. For the some other features like resting_bp, the difference is not significant.

Graphical summaries of catagorical variables



As we can see from the plot above: male are tending to have the heart disease; Even if the patient has some normal features , they could still have heart disease.