EDA for Heart disease dataset

DataFrame fragment demonstration

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
0	69.0	1.0	0.0	160.0	234.0	1.0	2.0	131.0	0.0	0.1	1.0	1.0	0.0	0.0
1	69.0	0.0	0.0	140.0	239.0	0.0	0.0	151.0	0.0	1.8	0.0	2.0	0.0	0.0
2	66.0	0.0	0.0	150.0	226.0	0.0	0.0	114.0	0.0	2.6	2.0	0.0	0.0	0.0
3	65.0	1.0	0.0	138.0	282.0	1.0	2.0	174.0	0.0	1.4	1.0	1.0	0.0	1.0
4	64.0	1.0	0.0	110.0	211.0	0.0	2.0	144.0	1.0	1.8	1.0	0.0	0.0	0.0

DataFrame columns information

RangeIndex: 297 entries, 0 to 296

Data columns (total 14 columns):

Column Non-Null Count Dtype

--- ------ -------------- -----

0 age 297 non-null int64

1 sex 297 non-null int64

2 cp 297 non-null int64

3 trestbps 297 non-null int64

4 chol 297 non-null int64

5 fbs 297 non-null int64

6 restecg 297 non-null int64

7 thalach 297 non-null int64

8 exang 297 non-null int64

9 oldpeak 297 non-null float64

10 slope 297 non-null int64

11 ca 297 non-null int64

12 thal 297 non-null int64

13 condition 297 non-null int64

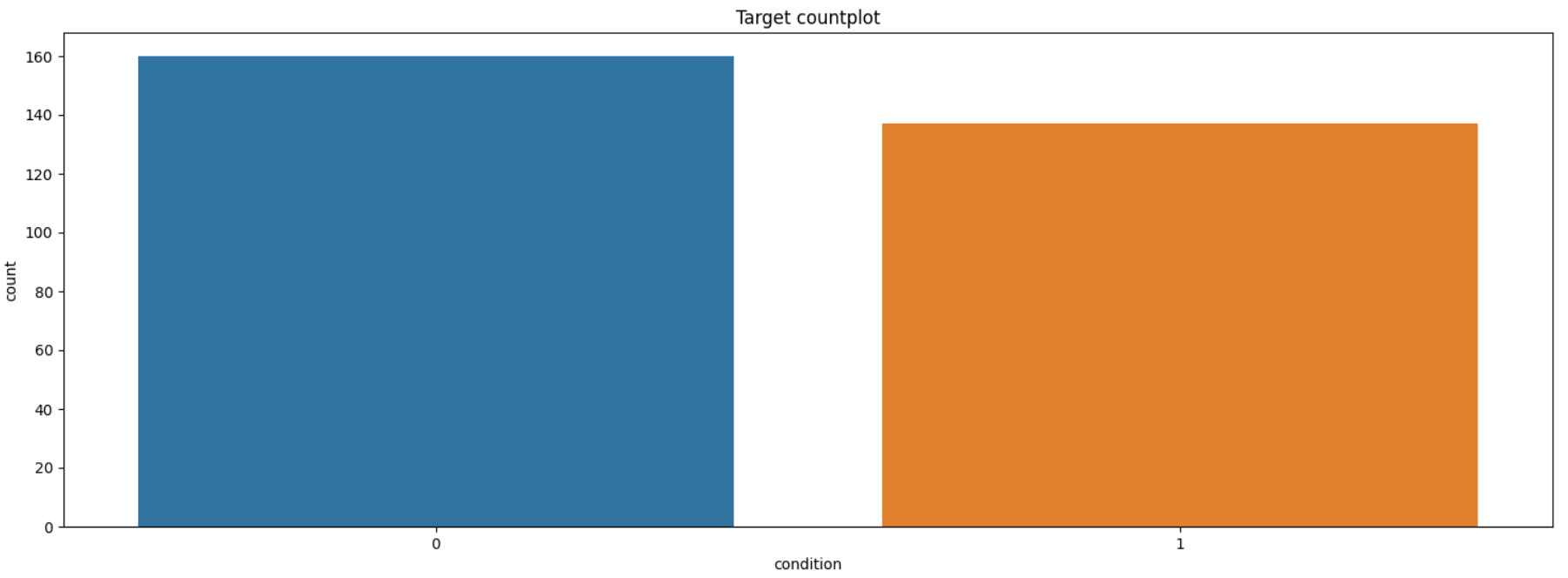
dtypes: float64(1), int64(13)

memory usage: 32.6 KB

DataFrame columns statistics

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
count	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0	297.0
mean	54.54	0.68	2.16	131.69	247.35	0.14	1.0	149.6	0.33	1.06	0.6	0.68	0.84	0.46
std	9.05	0.47	0.96	17.76	52.0	0.35	0.99	22.94	0.47	1.17	0.62	0.94	0.96	0.5
min	29.0	0.0	0.0	94.0	126.0	0.0	0.0	71.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	48.0	0.0	2.0	120.0	211.0	0.0	0.0	133.0	0.0	0.0	0.0	0.0	0.0	0.0
50%	56.0	1.0	2.0	130.0	243.0	0.0	1.0	153.0	0.0	0.8	1.0	0.0	0.0	0.0
75%	61.0	1.0	3.0	140.0	276.0	0.0	2.0	166.0	1.0	1.6	1.0	1.0	2.0	1.0
max	77.0	1.0	3.0	200.0	564.0	1.0	2.0	202.0	1.0	6.2	2.0	3.0	2.0	1.0

Checking the target balance



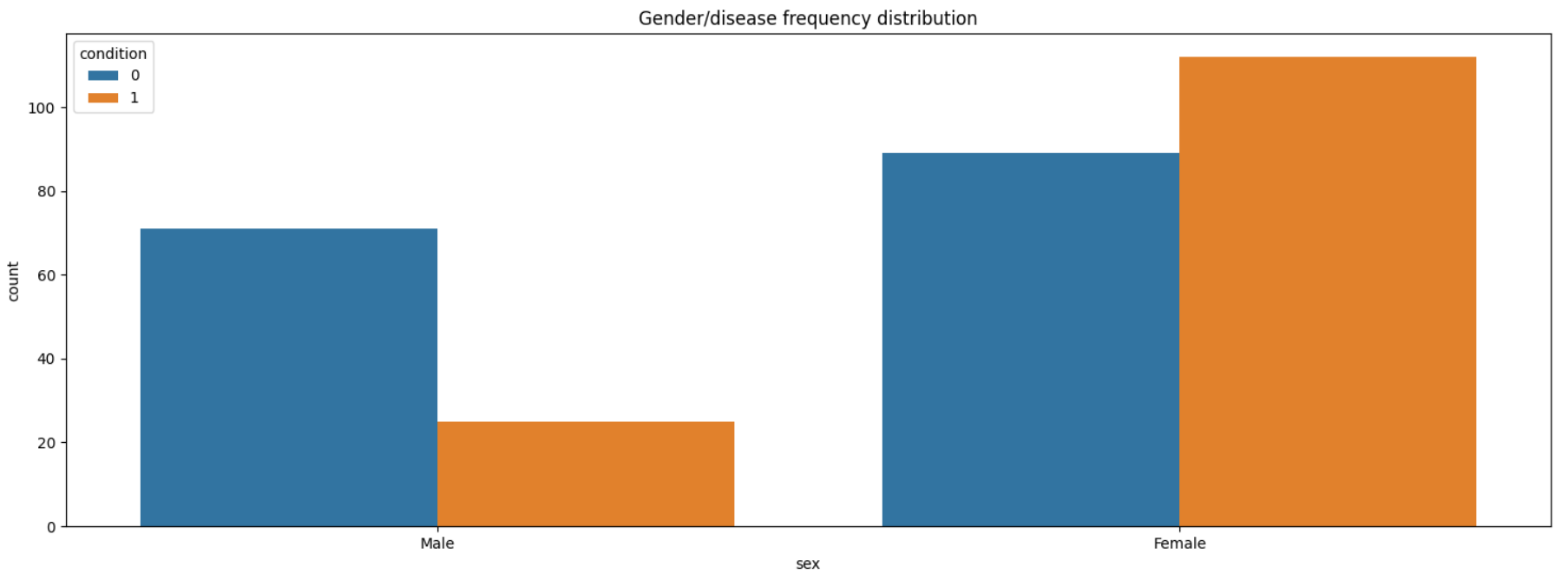
	condition
0	160
1	137

targer 1's count in percentage is 0.46 %

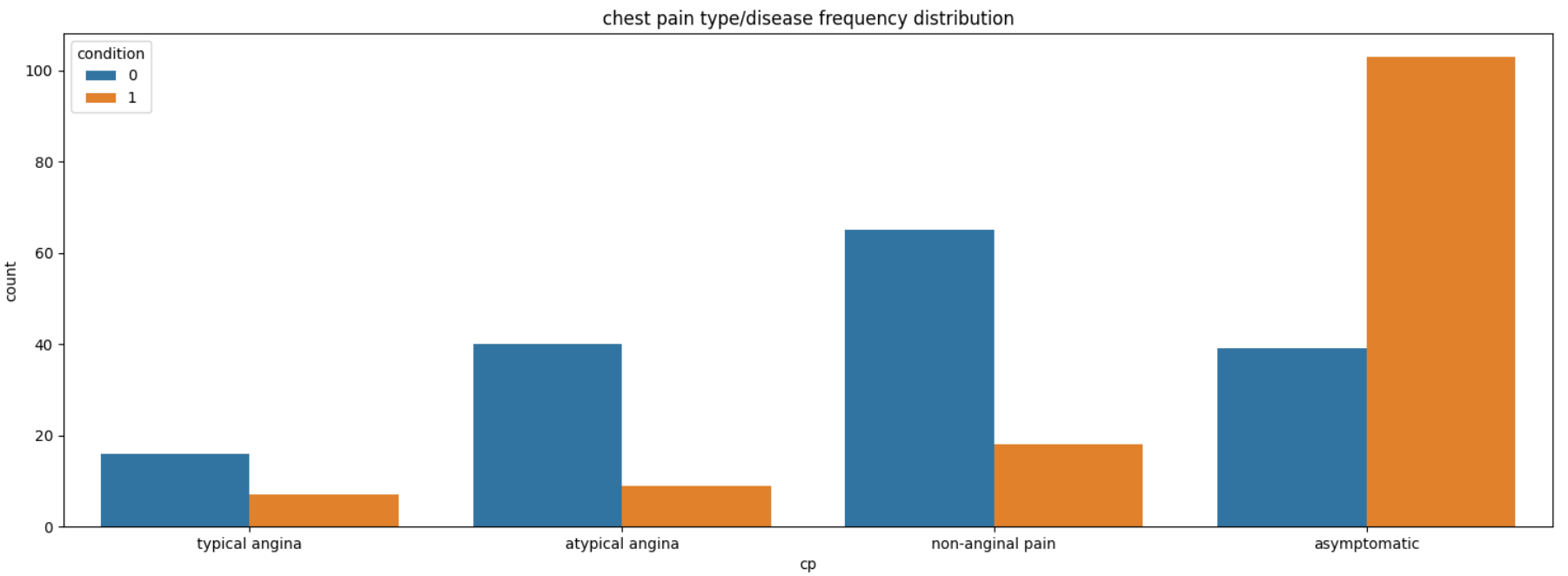
target 0's count in percentage is 0.54 %

Conclusion: dataset is unbalanced

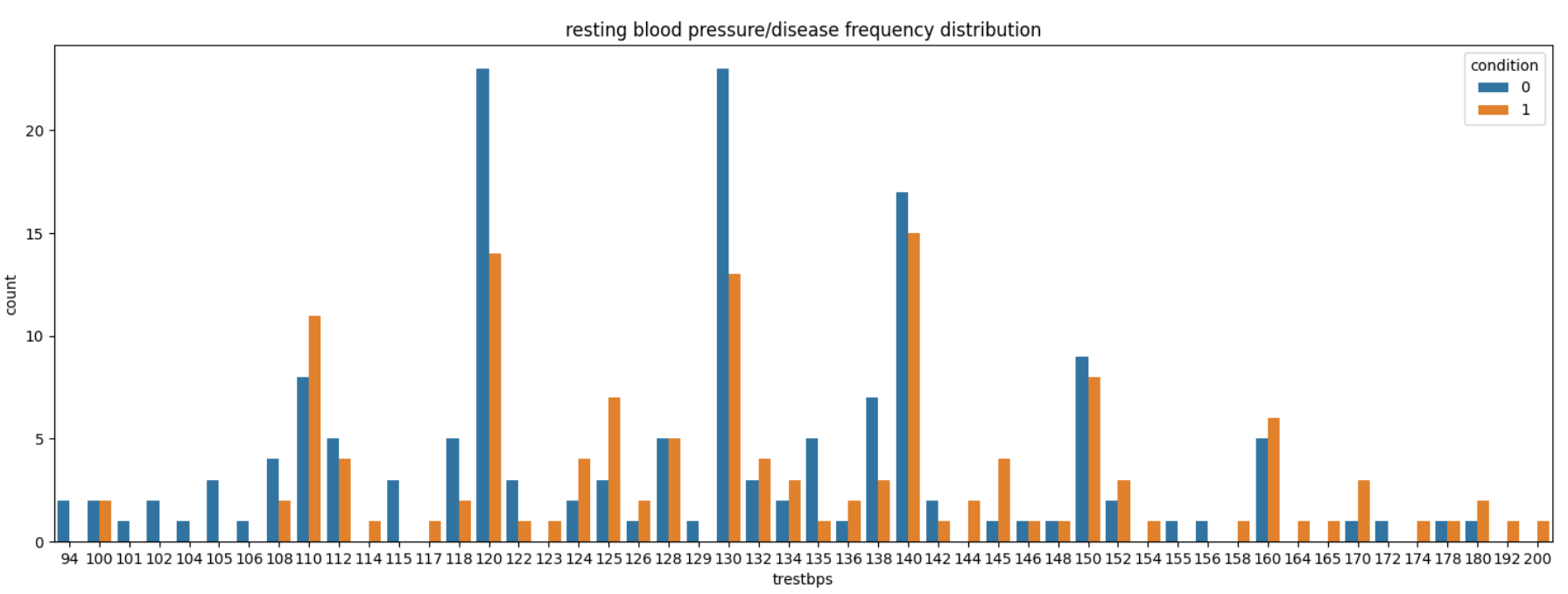
Parameters distribution plots

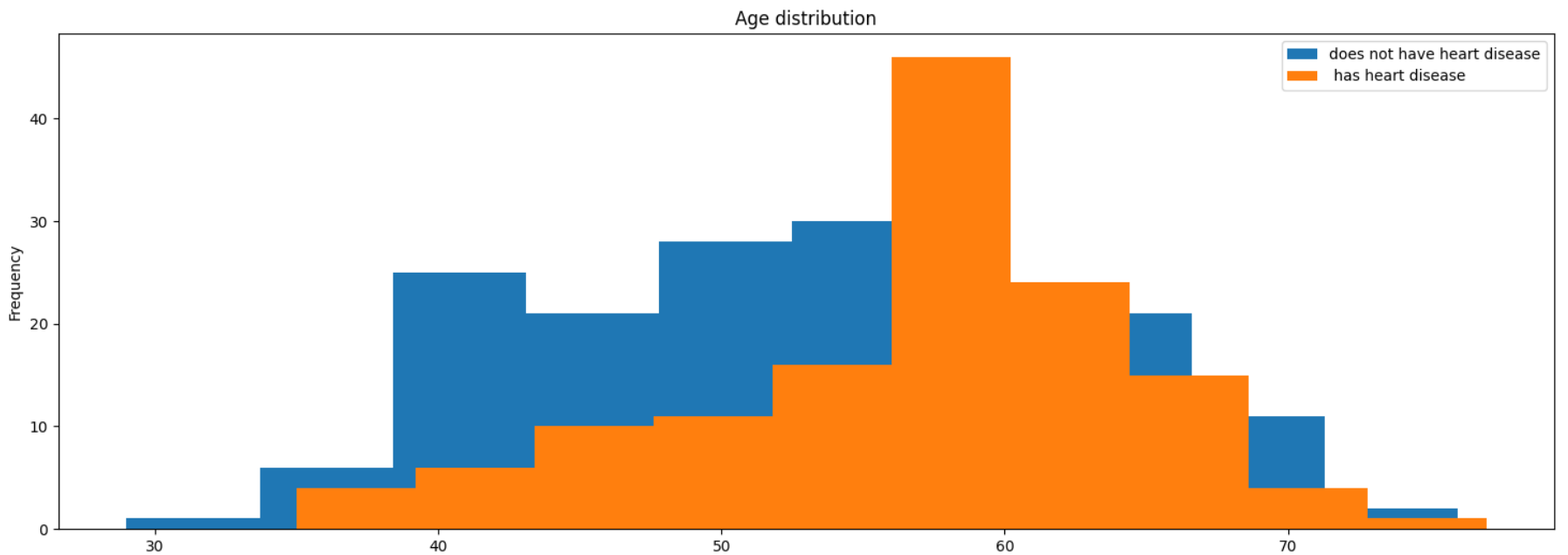


Conclusion: Females has heart diesease more frequent than males

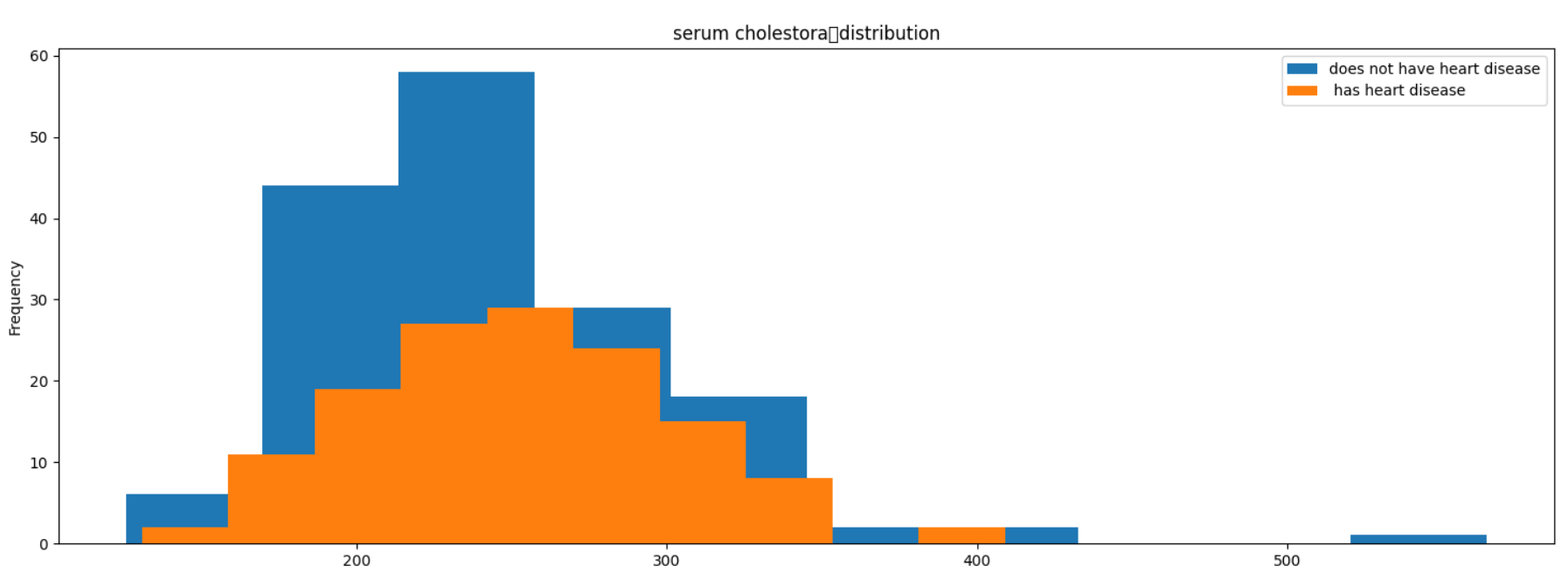


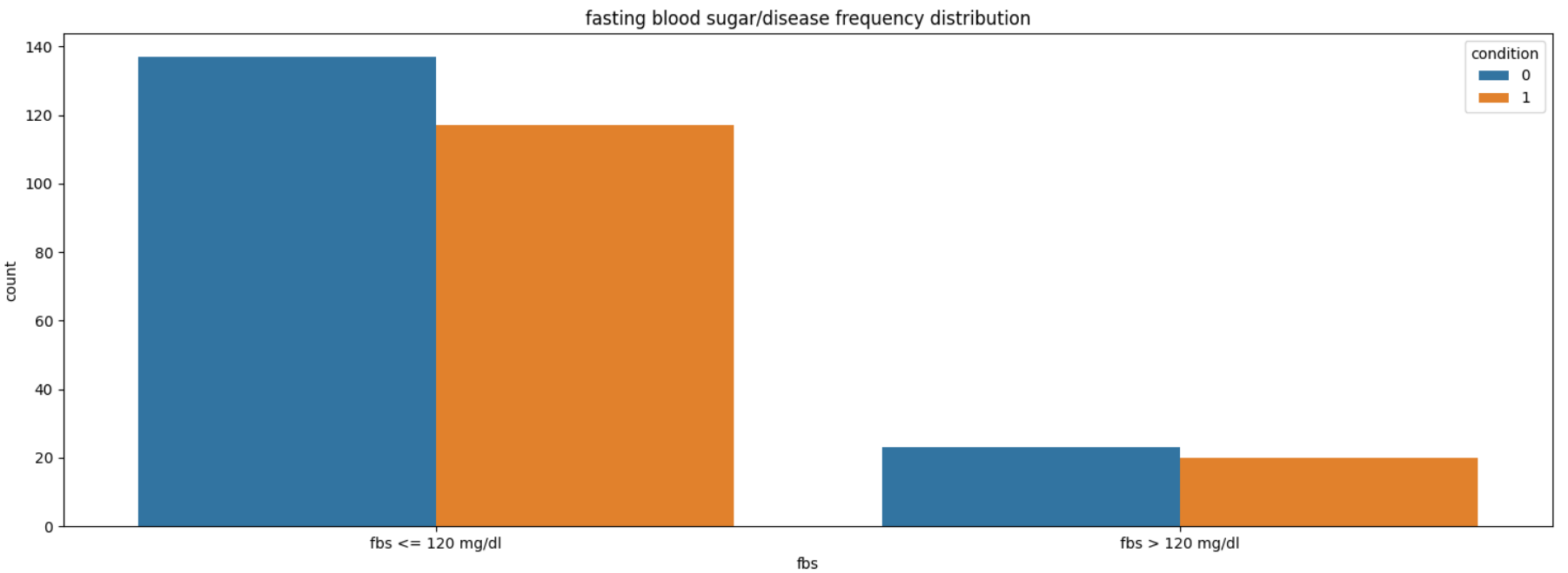
conclusion most of patients with heart disease were asymptomatic



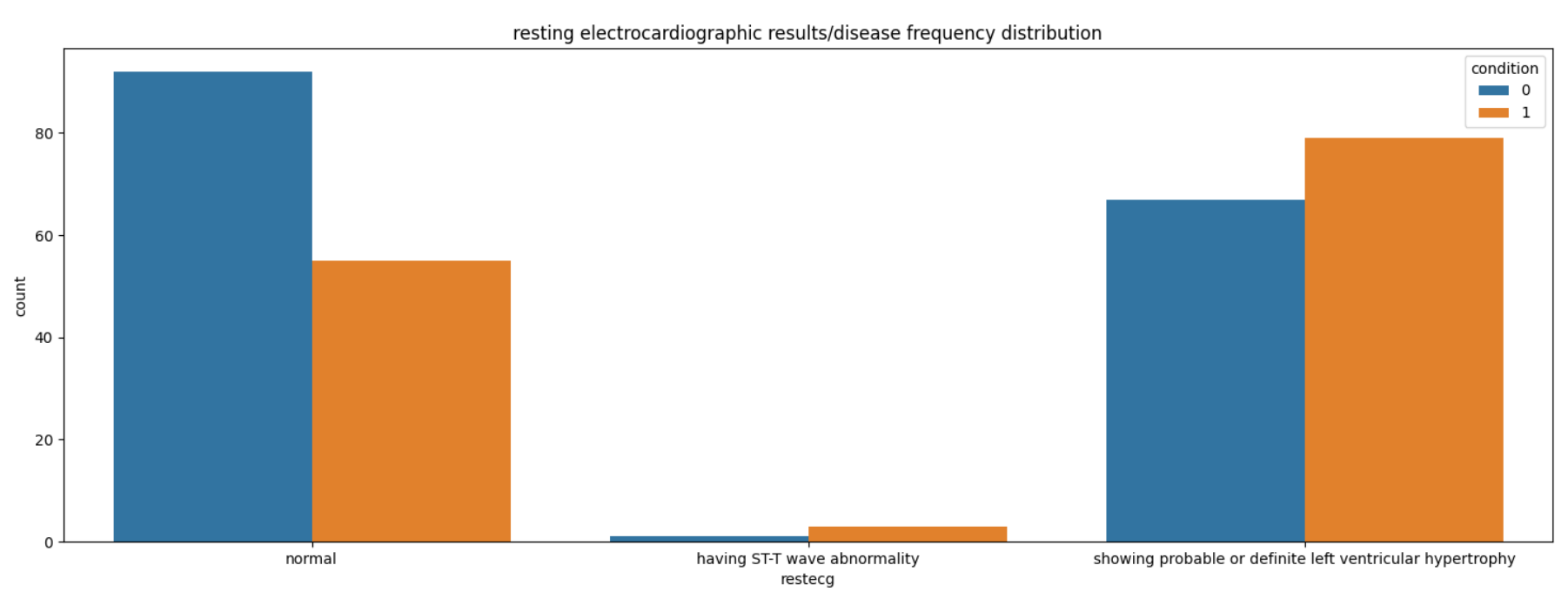


Conclusion: Eldery patient more frequent have heart disease

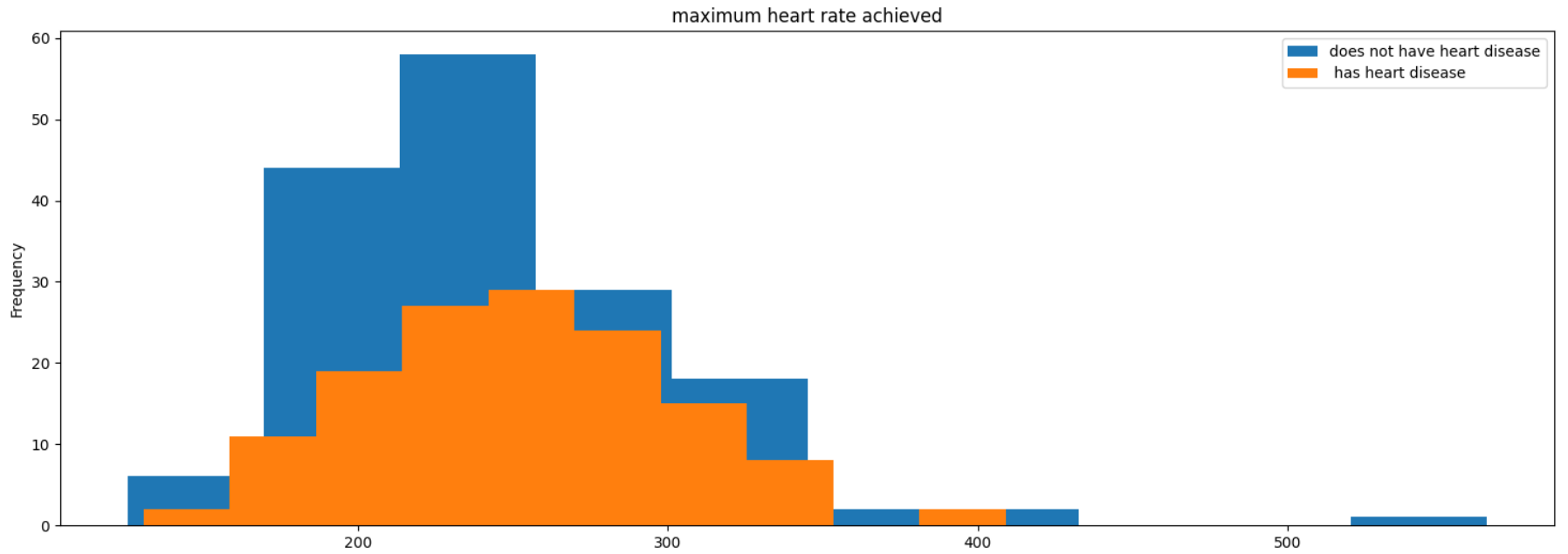




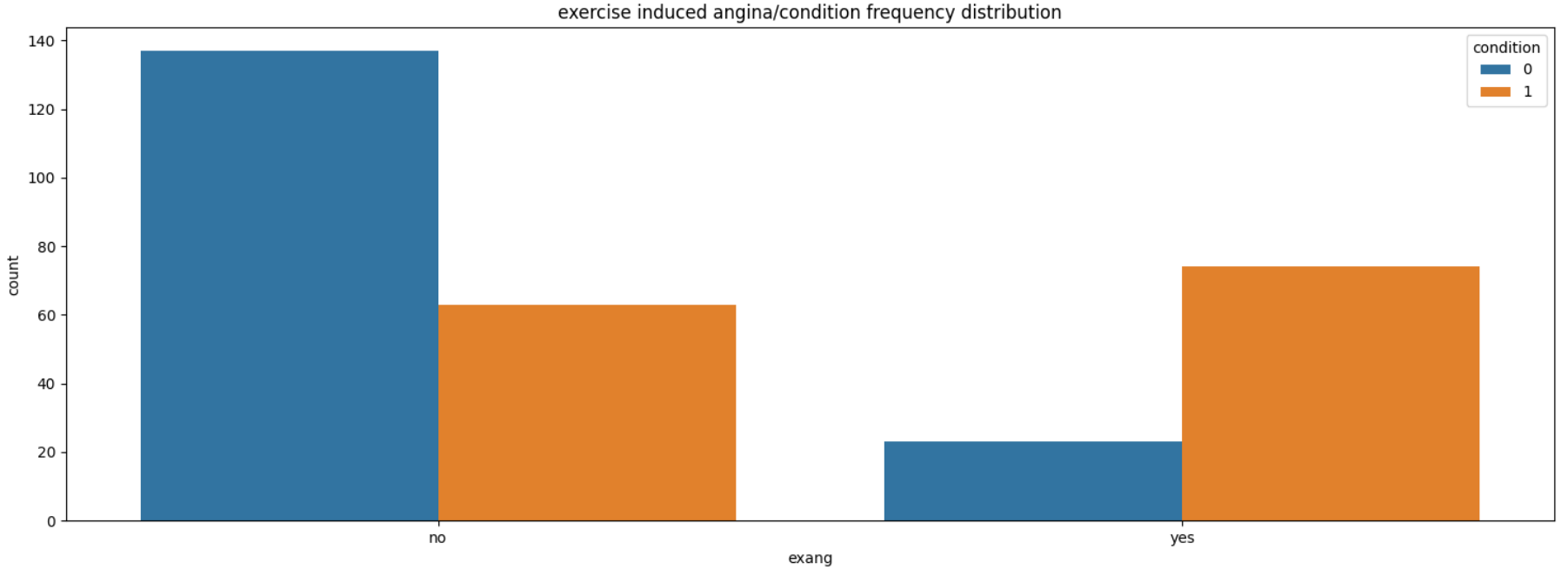
Conclusion: people with low fbs more frequent have heart disease



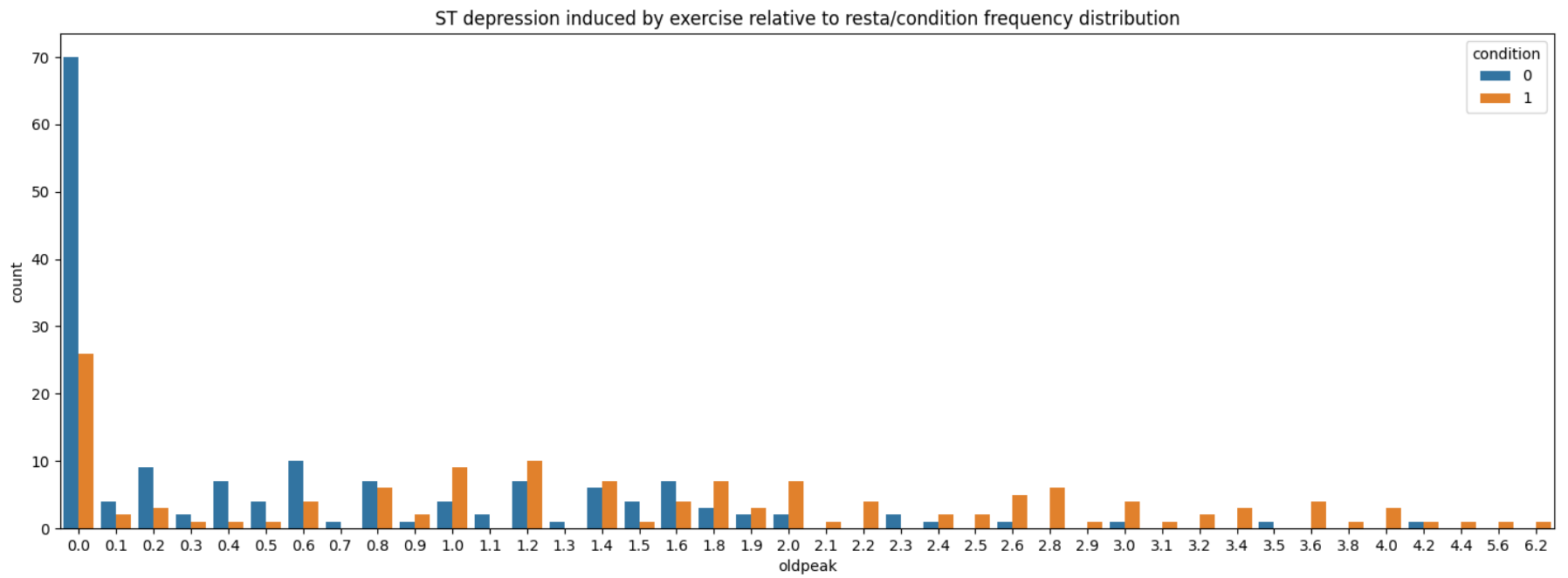
Conclusion: Patients showing left ventriculuar hupertrophy more frequent has heart disease



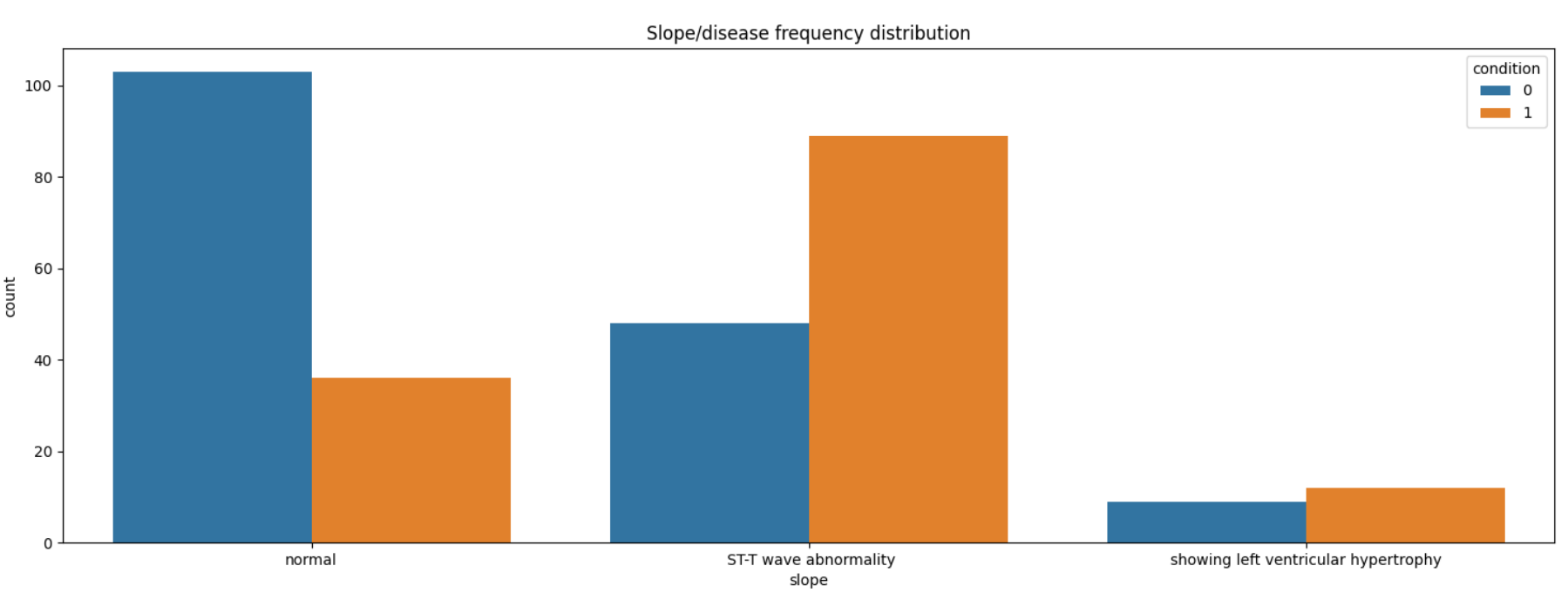
exercise induced angina/condition frequency distribution



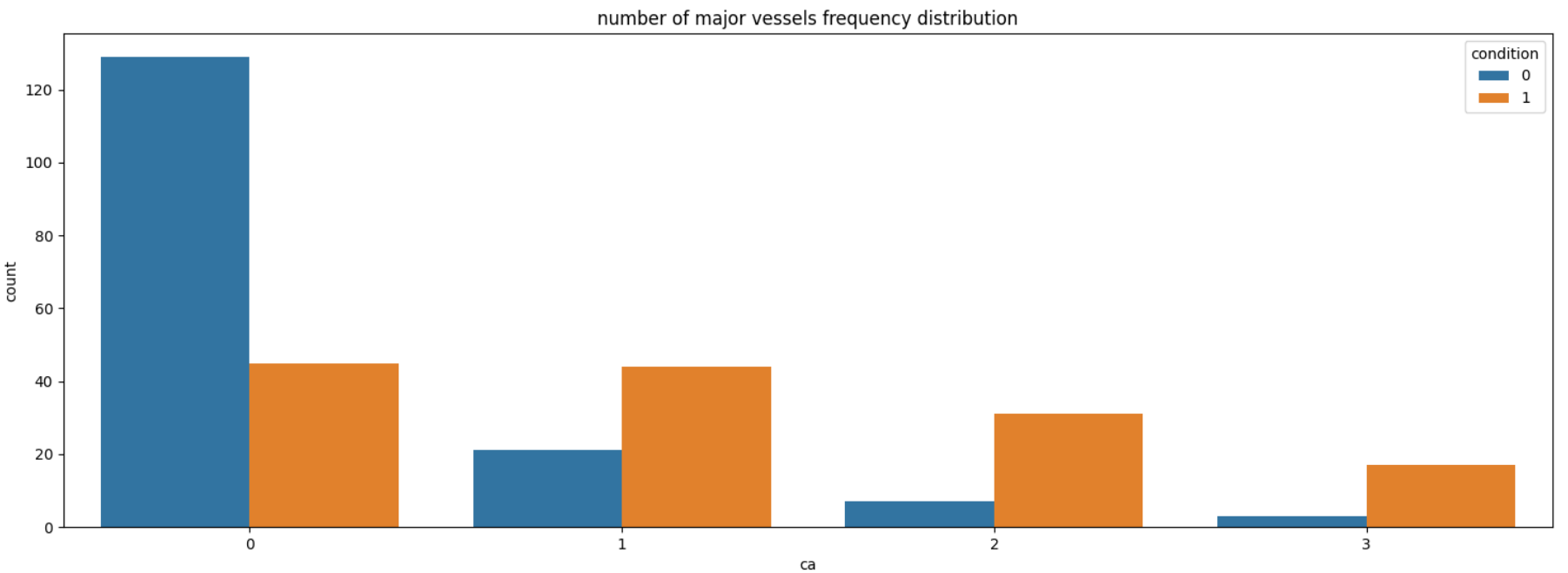
Conclusion: patients who had excercise included angine more frequent had heart disease



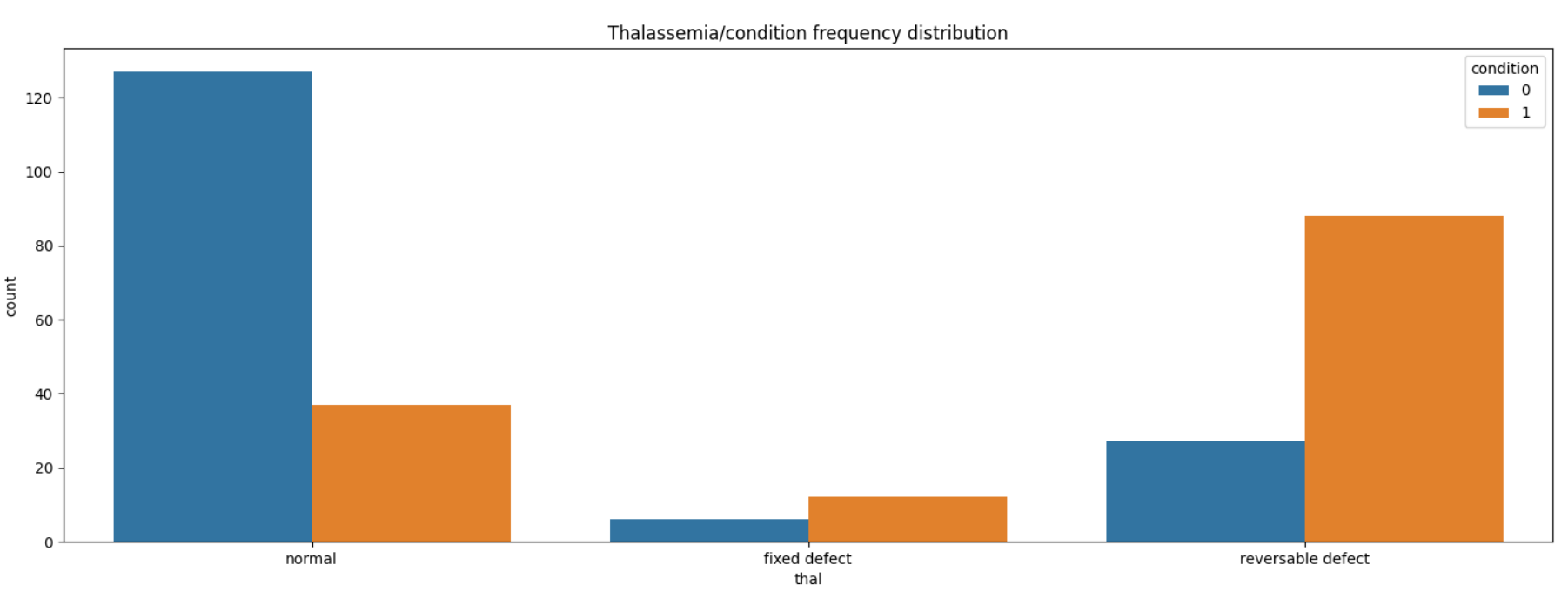
Conclusion: patients with high ST depression have more frequent heart disease



Conclusion: patients with flat slope more frequent have the heart disease

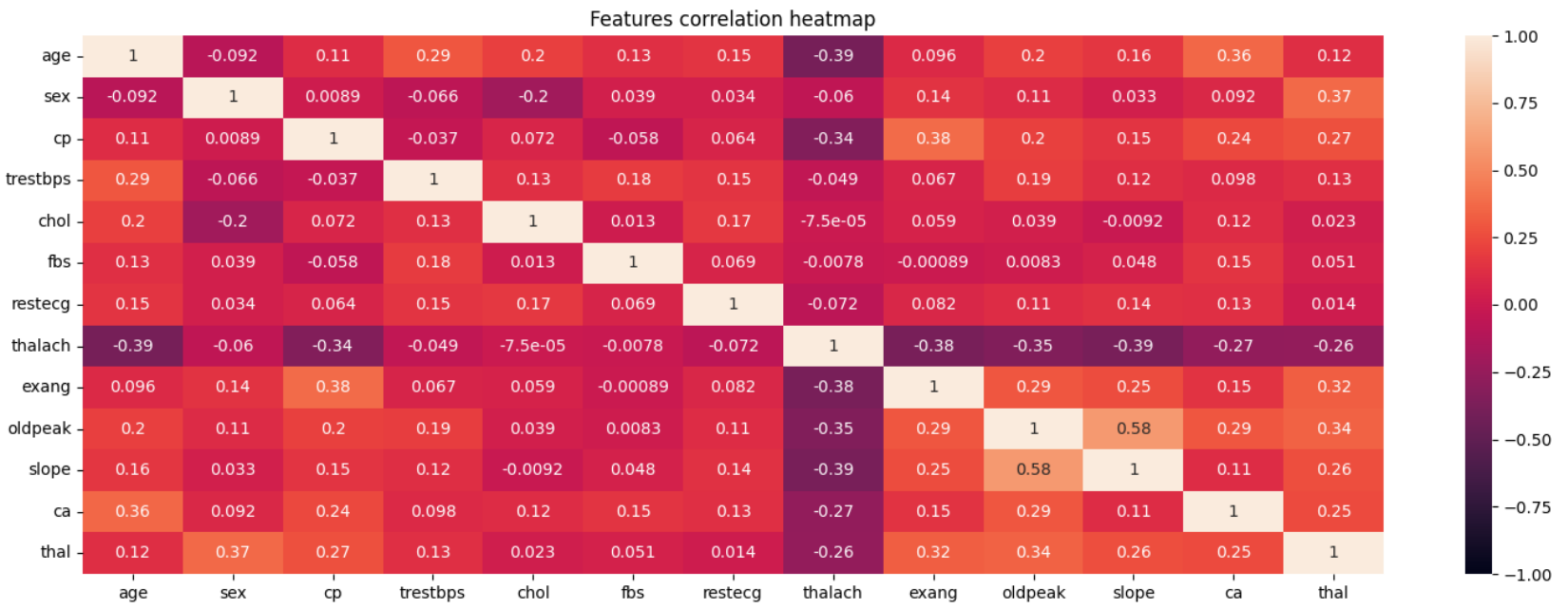


Conclusion: patients with more number of major vessels have more frequent heart disease



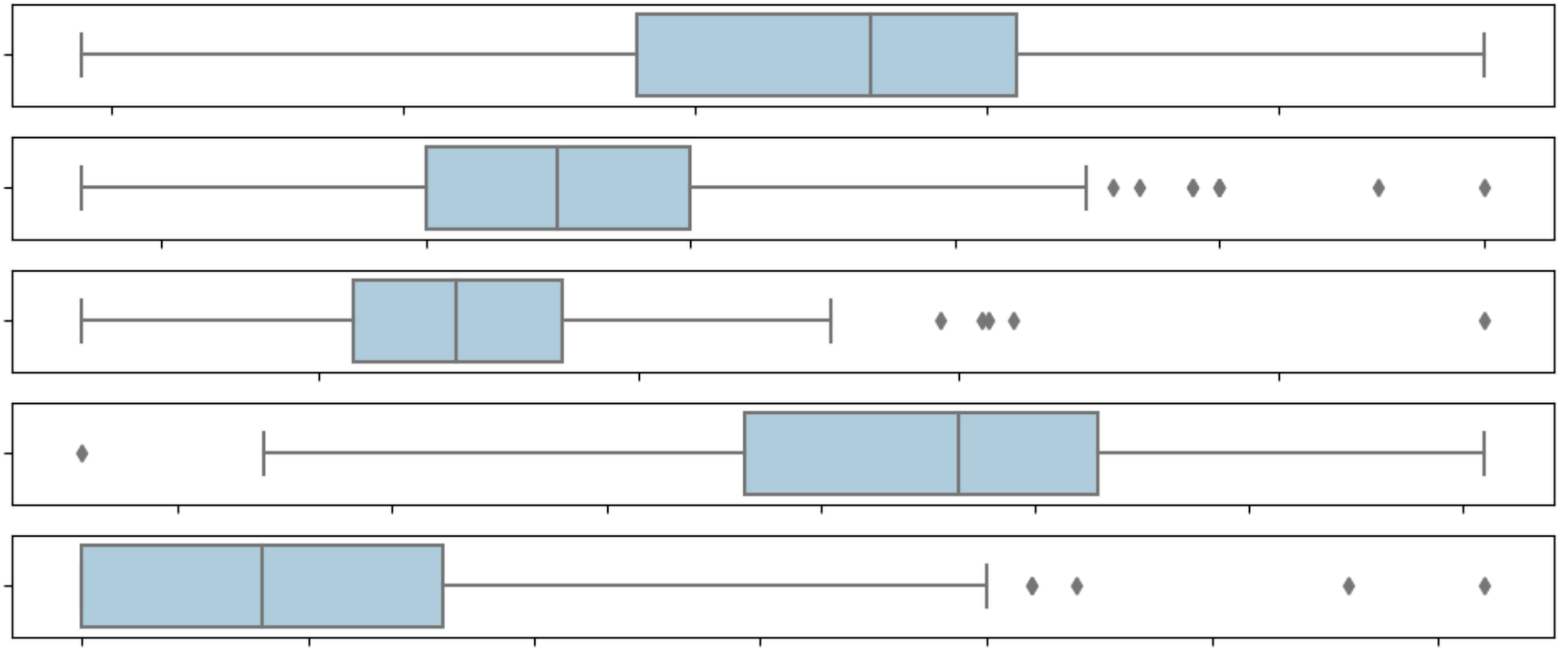
Conclusion: patient with fixed or reversable defect more frequent have heart disease

Correlation heatmap



Conclusion: No really strong correlation between any pair of variables

Outlier search



Conclusion: this dataset contains outliers