

## Exploratory Data Analysis- Summary

Our data consists of 12 columns and 918 rows. There are 7 numeric categories and 5 categorical.

1. **Age** - age of the patient [year]
2. **Sex** - sex of the patient [M: Male, F: Female]
3. **ChestPainType** - a type of chest pain [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** - 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** - having a heart disease [1: heart disease, 0: Normal]

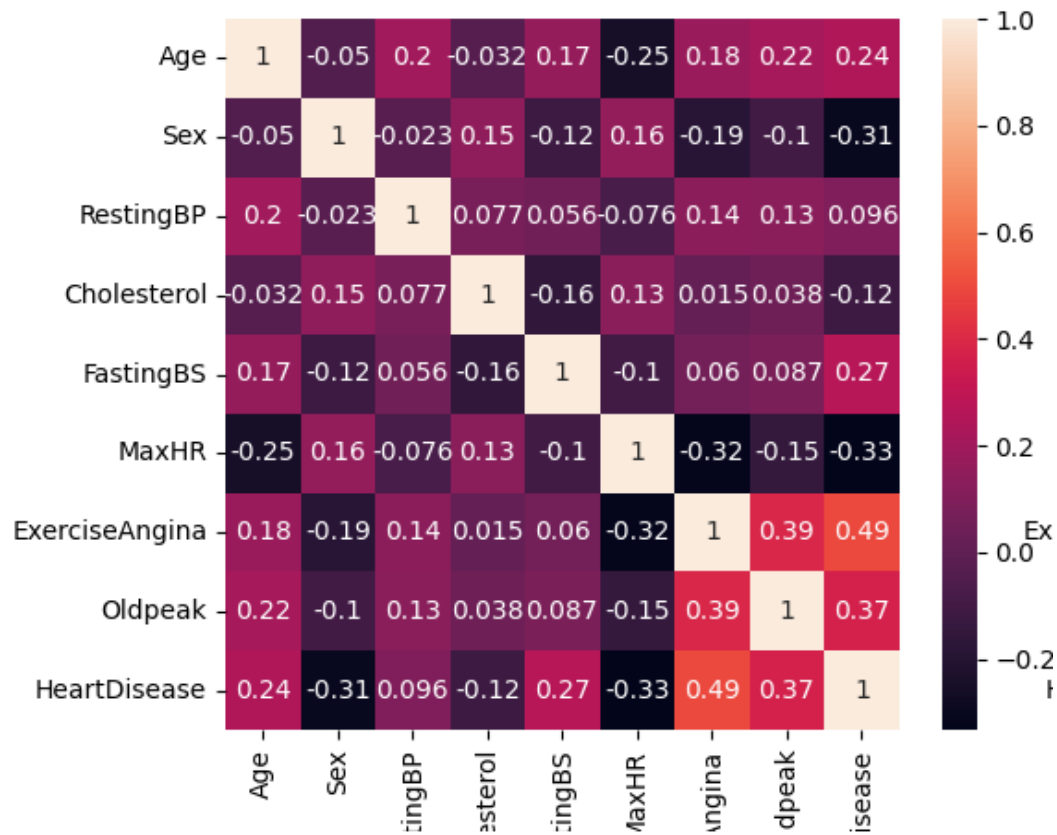
Summary of attributes- taken from our attribute description file

There are no missing values. Statistically, some outliers exist, but it seems these are not incorrect measurements or erroneous inputs, so we see no reason to omit them.

Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
39	120	339	0	170	0.0	0
53	113	468	0	127	0.0	0
53	145	518	0	130	0.0	1
36	112	340	0	184	1.0	0
54	150	365	0	134	1.0	0
52	112	342	0	96	1.0	1
44	150	412	0	170	0.0	0
32	118	529	0	130	0.0	1
40	150	392	0	130	2.0	1
40	120	466	1	152	1.0	1
59	130	338	1	130	1.5	1
58	180	393	0	110	1.0	1
56	170	388	0	122	2.0	1
50	140	341	0	125	2.5	1

Sample of outliers taken from our DataQualityAssessment.py file

There aren't many very strong correlations between the variables. Exercise Angina and Oldpeak both have relatively significant correlation between each other and heart disease. The correlations also tell us that more males had heart disease issues than women. Older age is also seen to correlate with positive diagnosis of heart disease.



Correlation matrix, created by our correlations.py file