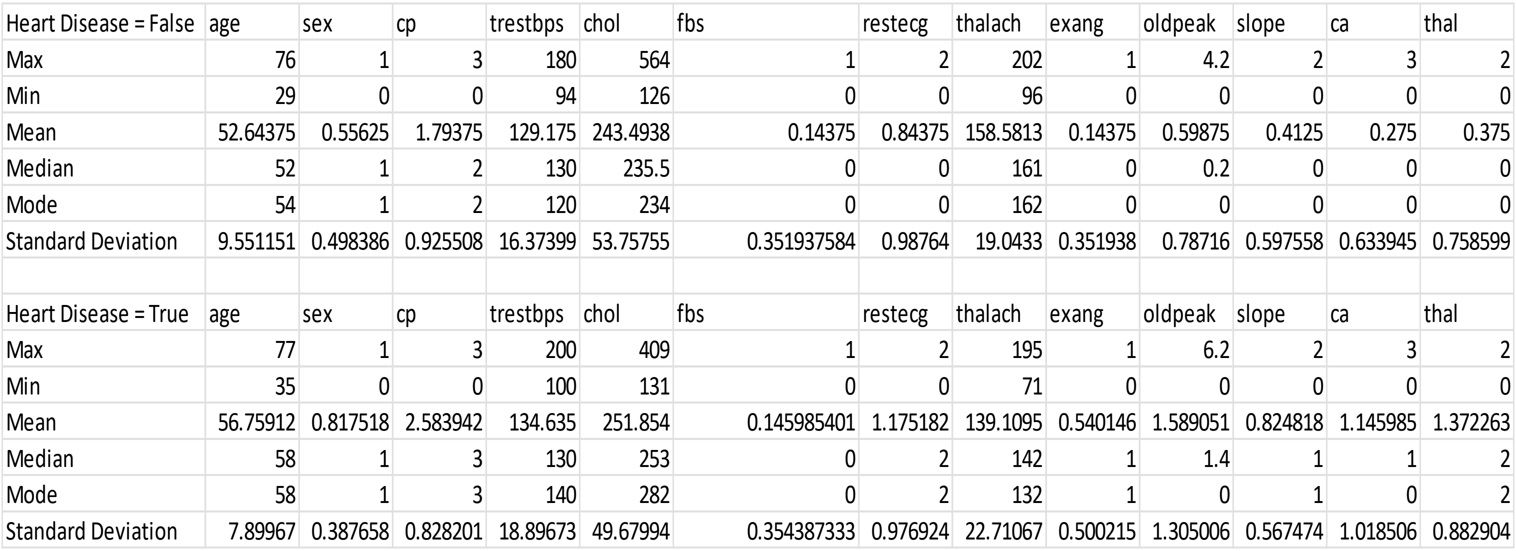
Alex Tuttle
Applied Machine Learning
Project Two

Our problem is to determine is patients do or do not have heart disease. Our data set is pulled from the UCI Machine Learning Repository. It can be found under the Heart Disease Data Set. This data set include 297 instances with 75 attributes. However, the data is limited to 14 total attributes including the predicted attribute. This attribute is the condition of the patient related to having heart disease. There are two classes available, which include true if the patient has heart disease and false if the patient does not. There are four databases within this data set; however, we are using the Cleveland data set. This is due to the proximity of Cleveland to Akron compared to the other data sets. The Cleveland data set is created by Robert Detrano, M.D., Ph.D., and the Cleveland Clinic Foundation. This data set is appropriate because it is data pulled from actual patients, which may or may not have heart disease. It is a randomized data set; thus, the accuracy of the prediction model can be better evaluated. The results will not be extremely high or low because of the randomization. The approach to solve this problem is to input the data set, use each attribute listed in the data set and create the training set with roughly 75% of the data. Then, create the testing set with the remaining 25%. While this is the main approach, the training and testing split may change to better suit the model.

Statistical Summary:



Heart Disease = False	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Max	76	1	3	180	564	1	2	202	1	4.2	2	3	2
Min	29	0	0	94	126	0	0	96	0	0	0	0	0
Mean	52.64375	0.55625	1.79375	129.175	243.4938	0.14375	0.84375	158.5813	0.14375	0.59875	0.4125	0.275	0.375
Median	52	1	2	130	235.5	0	0	161	0	0.2	0	0	0
Mode	54	1	2	120	234	0	0	162	0	0	0	0	0
Standard Deviation	9.551151	0.498386	0.925508	16.37399	53.75755	0.351937584	0.98764	19.0433	0.351938	0.78716	0.597558	0.633945	0.758599
Heart Disease = True	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
Max	77	1	3	200	409	1	2	195	1	6.2	2	3	2
Min	35	0	0	100	131	0	0	71	0	0	0	0	0
Mean	56.75912	0.817518	2.583942	134.635	251.854	0.145985401	1.175182	139.1095	0.540146	1.589051	0.824818	1.145985	1.372263
Median	58	1	3	130	253	0	2	142	1	1.4	1	1	2
Mode	58	1	3	140	282	0	2	132	1	0	1	0	2
Standard Deviation	7.89967	0.387658	0.828201	18.89673	49.67994	0.354387333	0.976924	22.71067	0.500215	1.305006	0.567474	1.018506	0.882904

Alex Tuttle
Applied Machine Learning
Project Two

All the attributes in the data set were used in the statistical summary for each class. The statistics show a generalization of the difference in each attribute when classifying if a patient has heart disease. One notable difference is the average age of people with heart disease. By looking at the mean of each class, those with heart disease tend to be in their upper-fifties, while those without tend to be in their lower-fifties. Another notable difference is those with heart disease have a generally higher cholesterol level. By looking at the data, each attribute is higher if a patient has heart disease than if they do not, excluding max heart rate in the thalach column. Overall, the model can determine whether a patient has heart disease based on the higher or lower levels for each attribute.