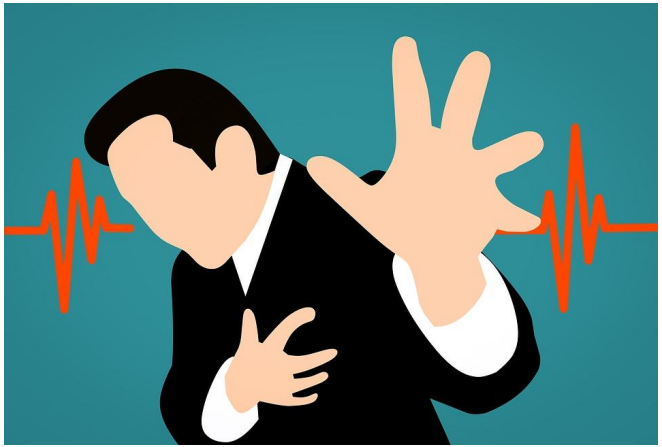
Predicting Heart Disease

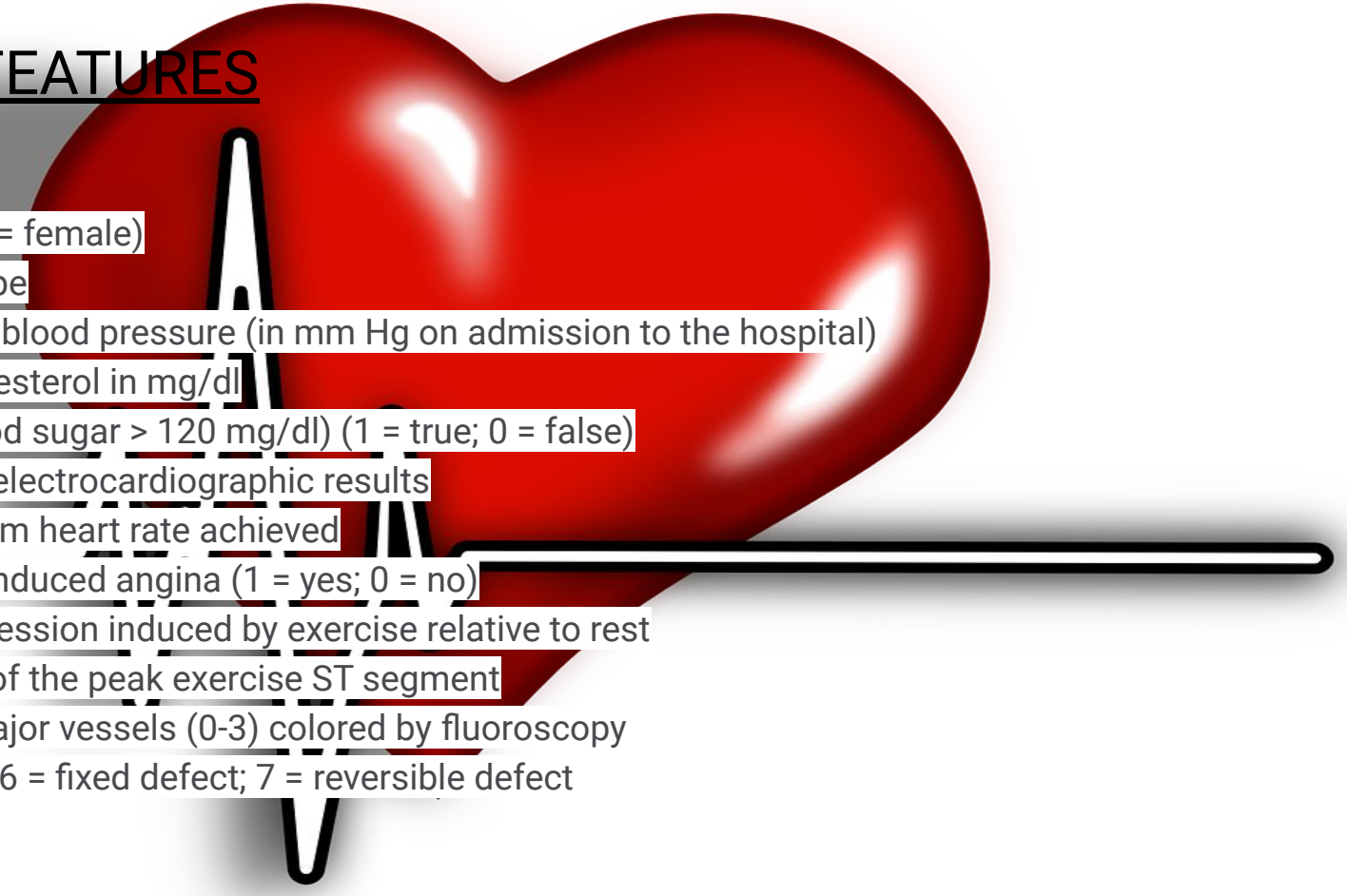
A presentation by Asher Palmer

The Dataset

The dataset I chose is from the University of California, Irvine housed at Kaggle. The dataset is on heart disease. There are 1025 rows with 13 columns. The target variable is described as whether the patient has heart disease.



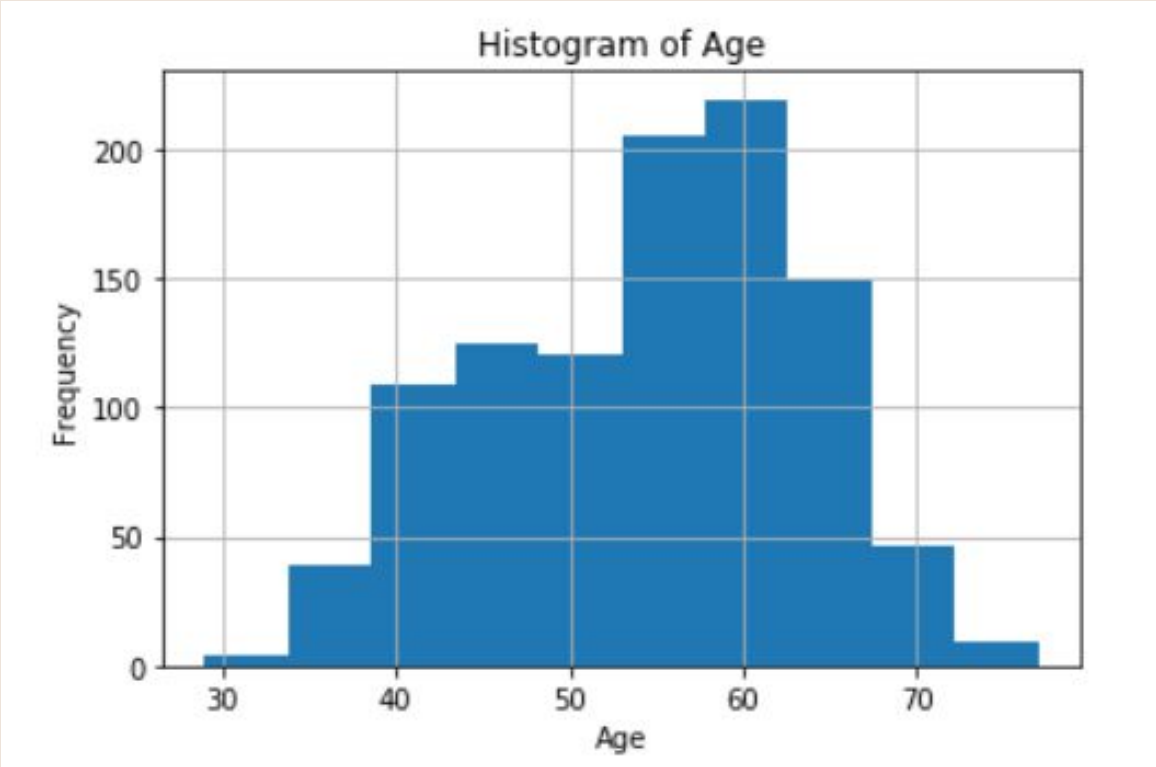
LIST OF FEATURES

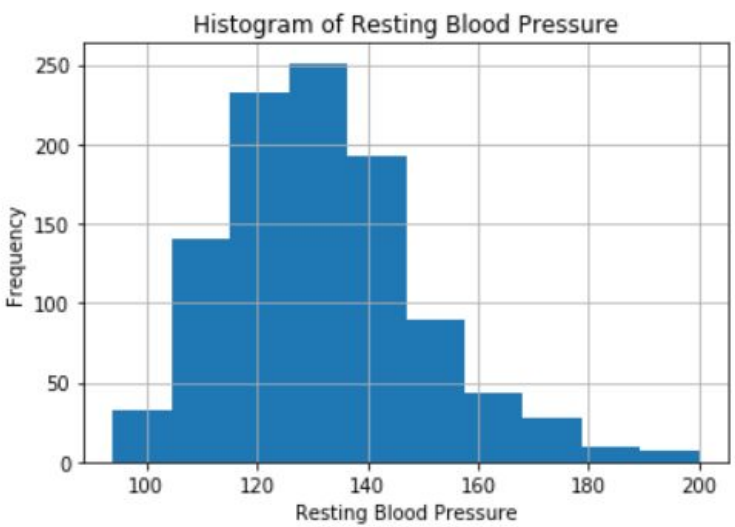


Age - age in years
Sex - (1 = male; 0 = female)
Cp - chest pain type
Trestbps - resting blood pressure (in mm Hg on admission to the hospital)
Chol - serum cholesterol in mg/dl
Fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg - resting electrocardiographic results
Thalach - maximum heart rate achieved
Exang - exercise induced angina (1 = yes; 0 = no)
Oldpeak - ST depression induced by exercise relative to rest
Slope - the slope of the peak exercise ST segment
Ca - number of major vessels (0-3) colored by fluoroscopy
Thal - 3 = normal; 6 = fixed defect; 7 = reversible defect
Target - 1 or 0

Histogram of Age

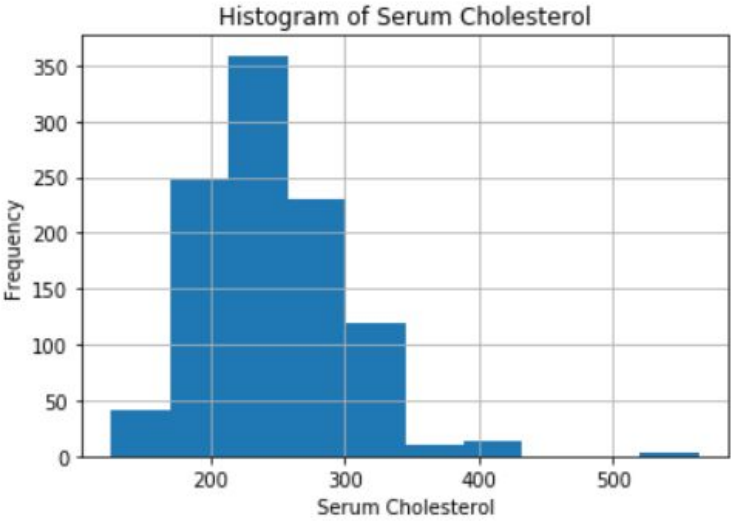
- Regular
- Ranges from 29 to 77 years
- Average Age is 54.43 years





Histogram of Resting Blood Pressure

These histograms are fairly regular but skewed to the left but not to the point of being too biased.

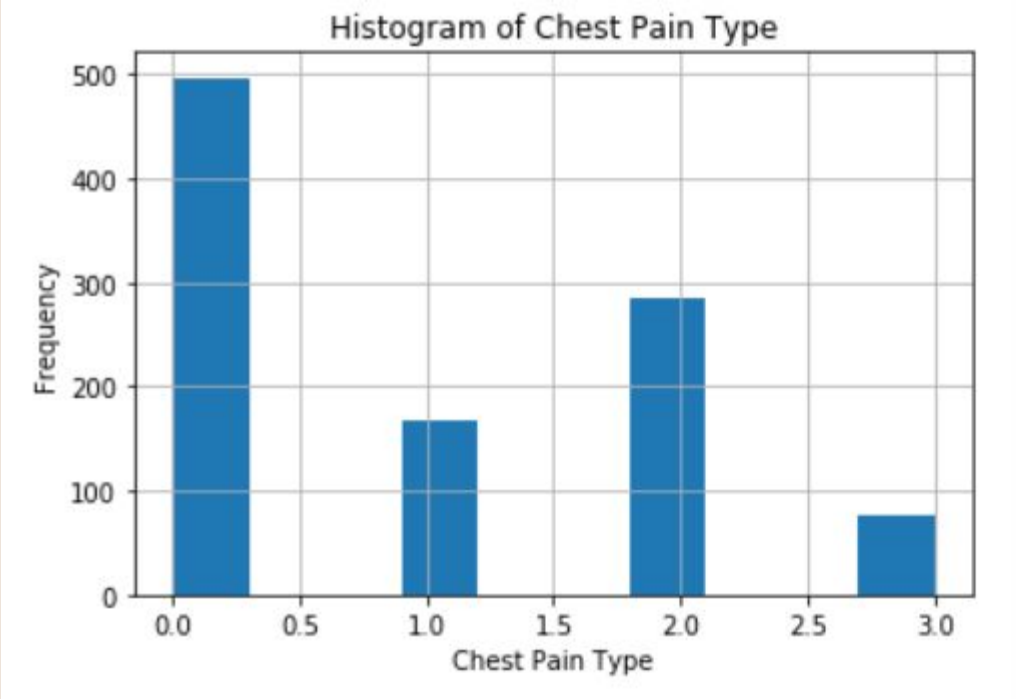


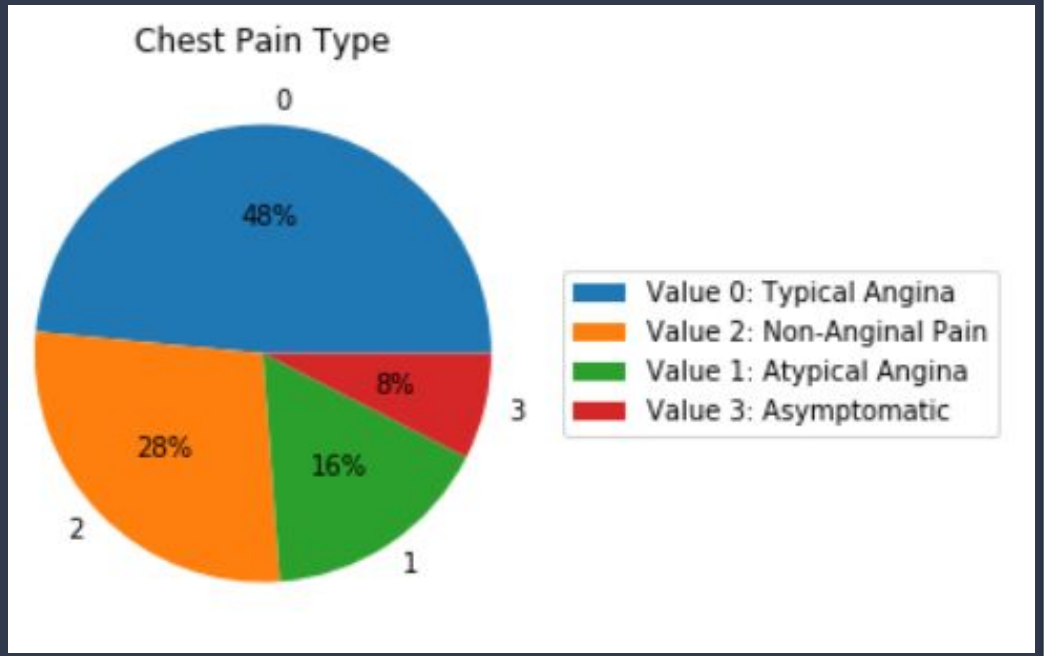
Histogram of Cholesterol

Bar Chart of Chest Pain Type

4 options

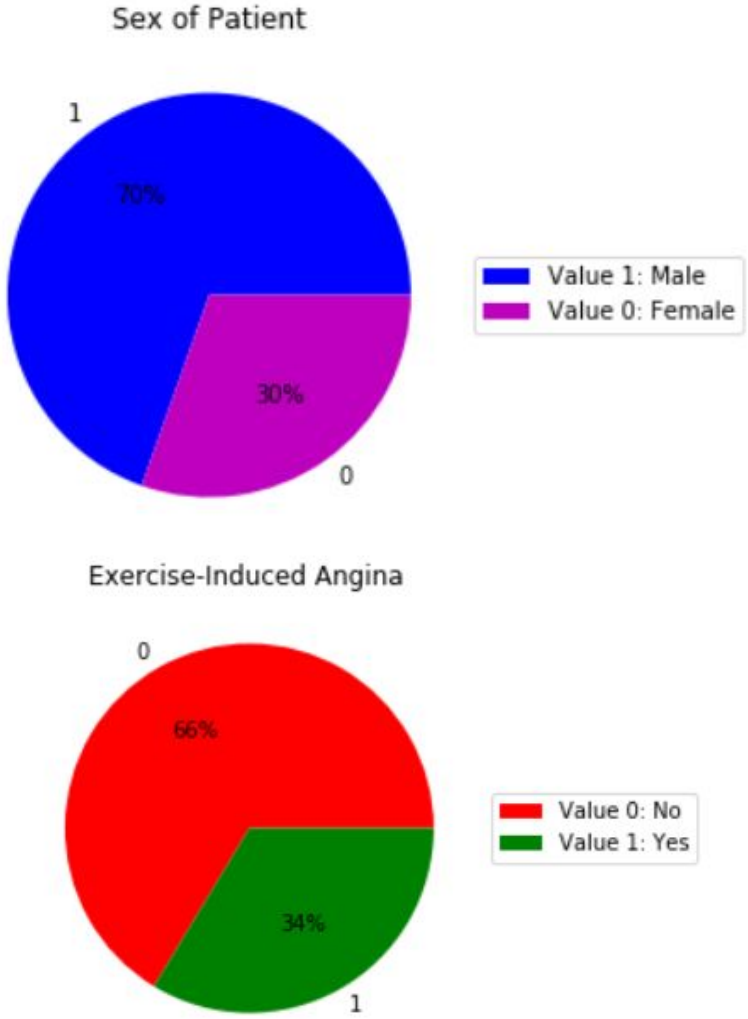
- Value 1: typical angina
- Value 2: atypical angina
- Value 3: non-anginal pain
- Value 4: asymptomatic

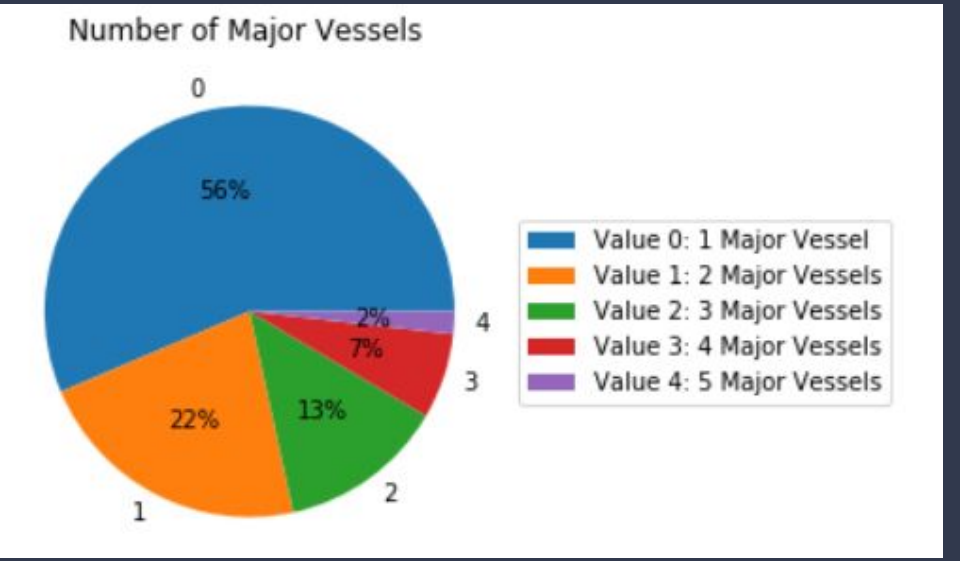


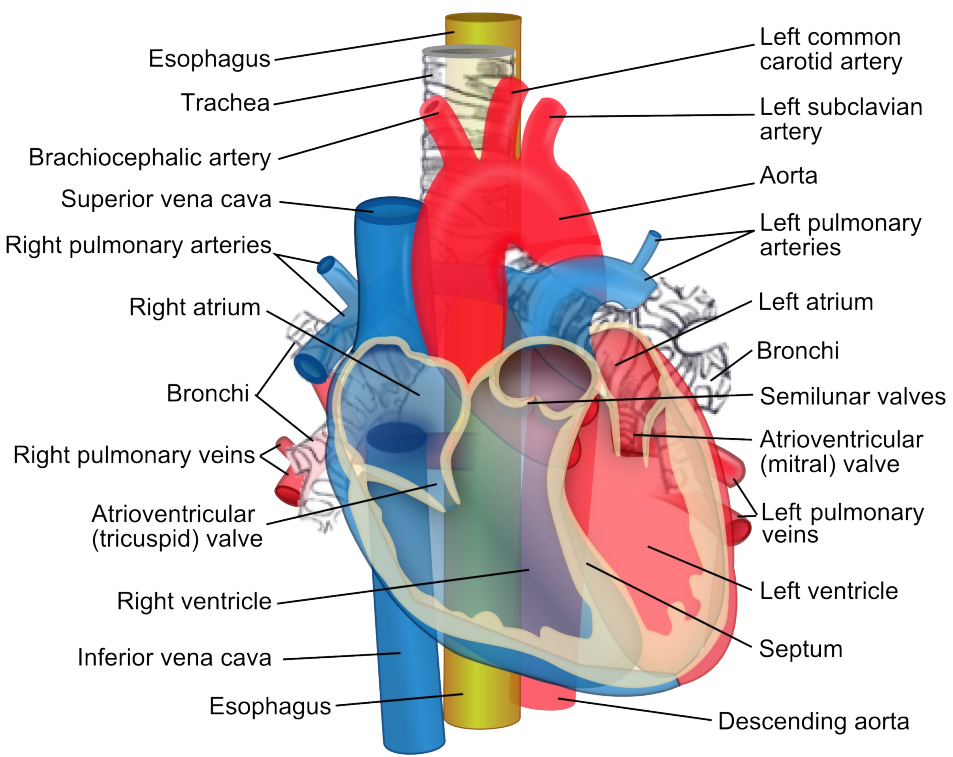


Chest Pain Pie Chart

- **Typical Angina**
  Chest pain or pressure, usually due to not enough blood flow to the heart muscle
- **Non-Anginal**
  Non-cardiac chest pain such as acid reflux
- **Atypical Angina**
  Possibly brought on by respiratory, musculoskeletal, or gastrointestinal diseases
- **Asymptomatic**
  Shows no symptoms


Sex of Patient
1
70%
30%
0
Value 1: Male
Value 0: Female
Exercise-Induced Angina
0
66%
34%
1
Value 0: No
Value 1: Yes



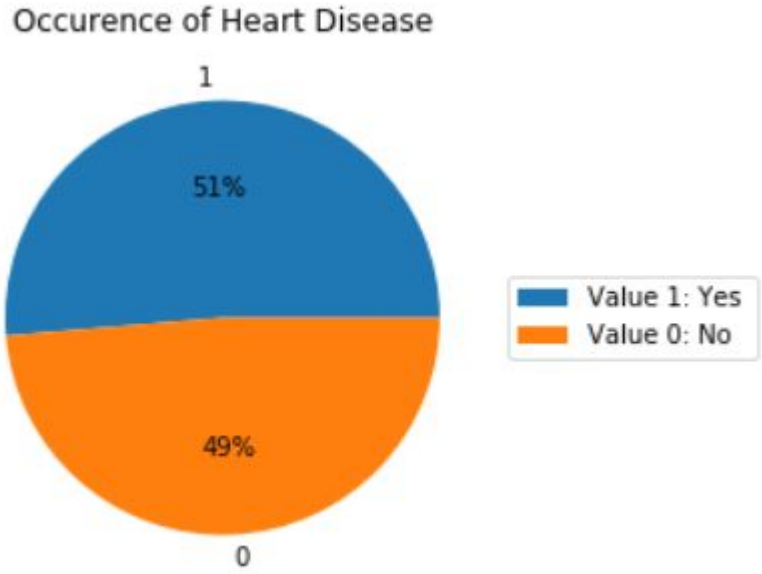


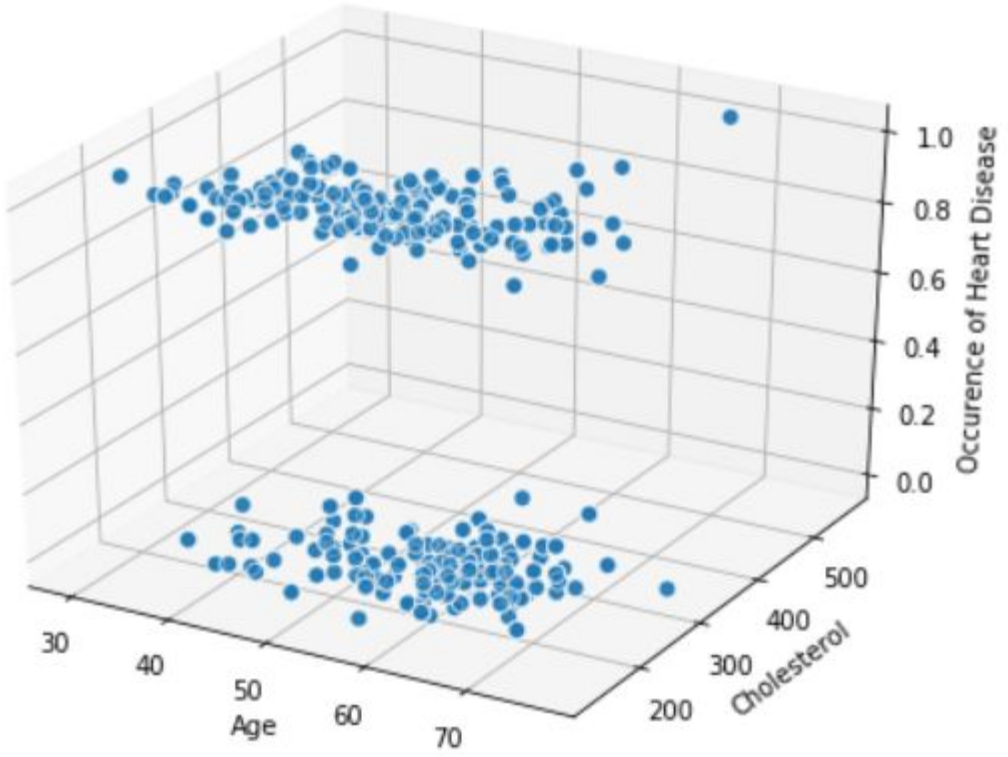
Major Blood Vessels

There are several major blood vessels and this pie chart describes the number of vessels involved in an incident.

The Important Info

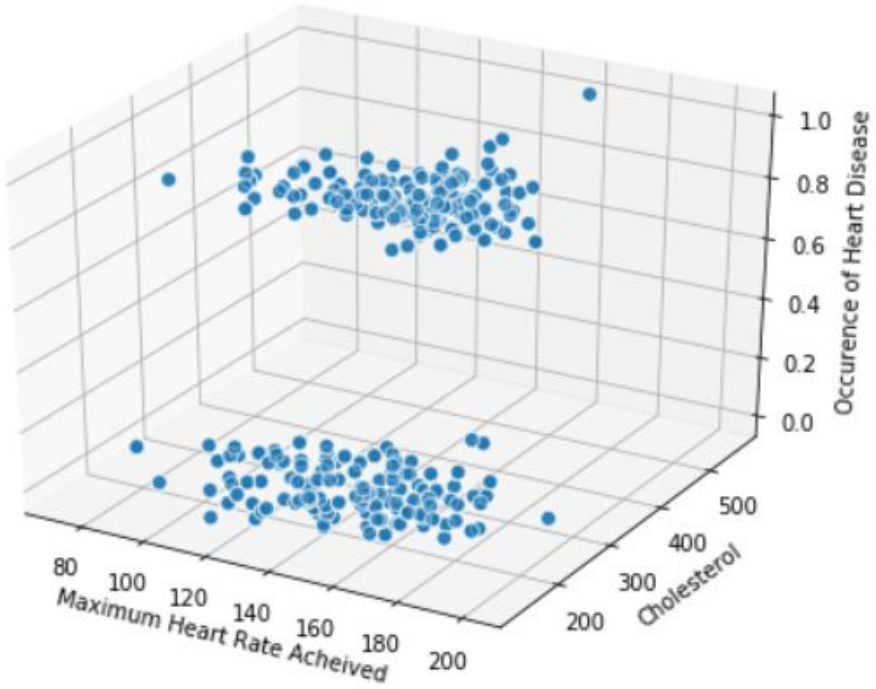
How many patients are getting heart disease? 51% of patients get heart disease in this sample.





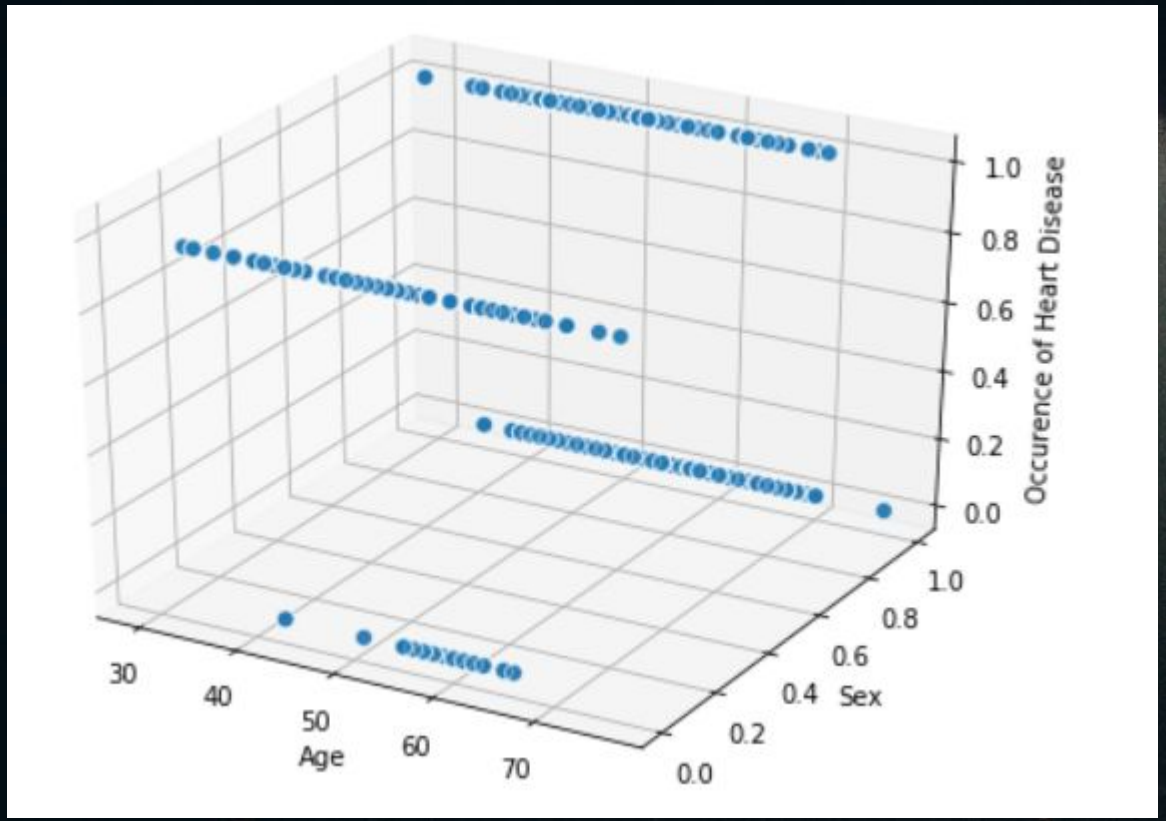
3D Scatterplot

This describes the relationship between age, cholesterol, and heart disease. As you can see, having low cholesterol may be ideal but does not contribute to heart disease.



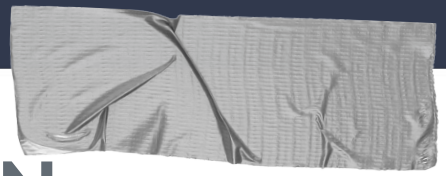
3D Scatterplot

The relationship between the maximum heart rate achieved during testing, cholesterol, and the occurrence of heart disease, Both clusters are spread fairly evenly between patients that get heart disease and those that don't.

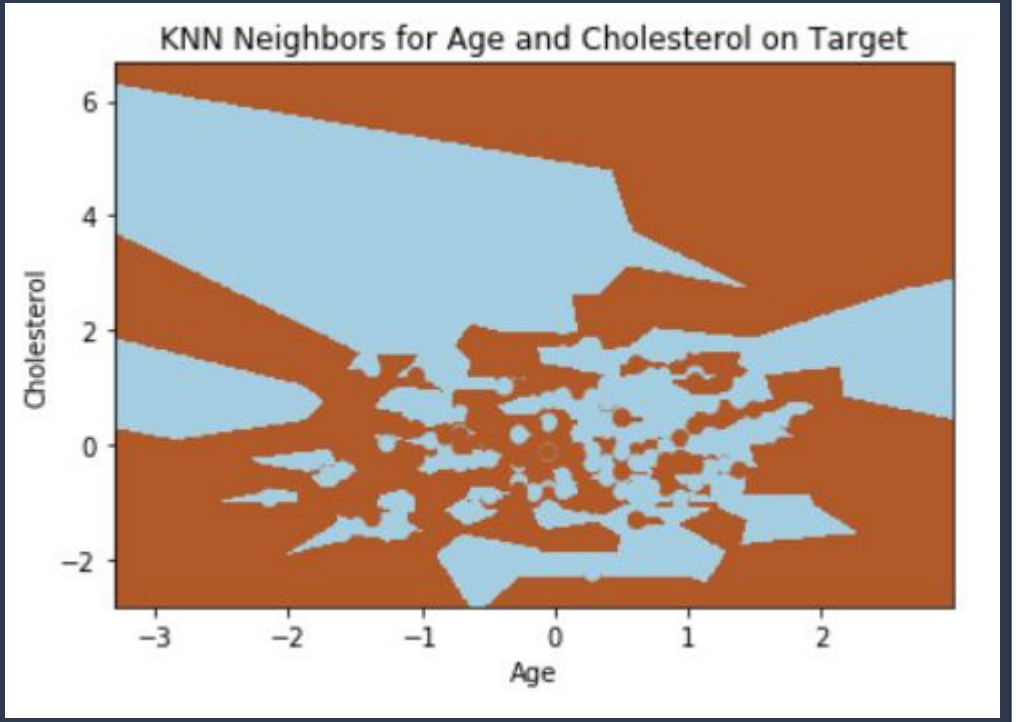


Age / Sex / Hearts

More people have heart disease but a larger portion of women have heart disease. Men and women regardless of age had heart disease while women in their mid 50s to 60s did not.



KNN



- To the left is the graphical representation of whether a patient will get heart disease based on age and cholesterol levels. As you can see, there is no clear relationship as where the brown space that indicates an area that heart disease has occurred.

```
array([1.        , 1.        , 1.        , 1.        , 1.        ,
       1.        , 1.        , 1.        , 0.97058824, 0.98019802])
```

```
Decision Tree Depth: 1
Model Performance: 75.99%.

Decision Tree Depth: 2
Model Performance: 73.07%.

Decision Tree Depth: 3
Model Performance: 83.02%.

Decision Tree Depth: 4
Model Performance: 84.29%.

Decision Tree Depth: 5
Model Performance: 89.75%.

Decision Tree Depth: 6
Model Performance: 94.73%.

Decision Tree Depth: 7
Model Performance: 97.17%.

Decision Tree Depth: 8
Model Performance: 98.44%.

Decision Tree Depth: 9
Model Performance: 99.80%.

Decision Tree Depth: 10
Model Performance: 99.90%.

--- Runtime: 24.84065190000001 seconds. ---
```

Random Forest / Decision Tree Models

The random forest model shows an accuracy of 100% as does the decision tree model.. The original data had well over 50 features and cut it down to 13. I believe that is why these accuracy numbers are so high.

```
R-squared simple Ridge Regression model:
0.5407524529252543
```

RIDGE

```
R² for the Lasso Regression model:
0.20928389033040784
```

LASSO

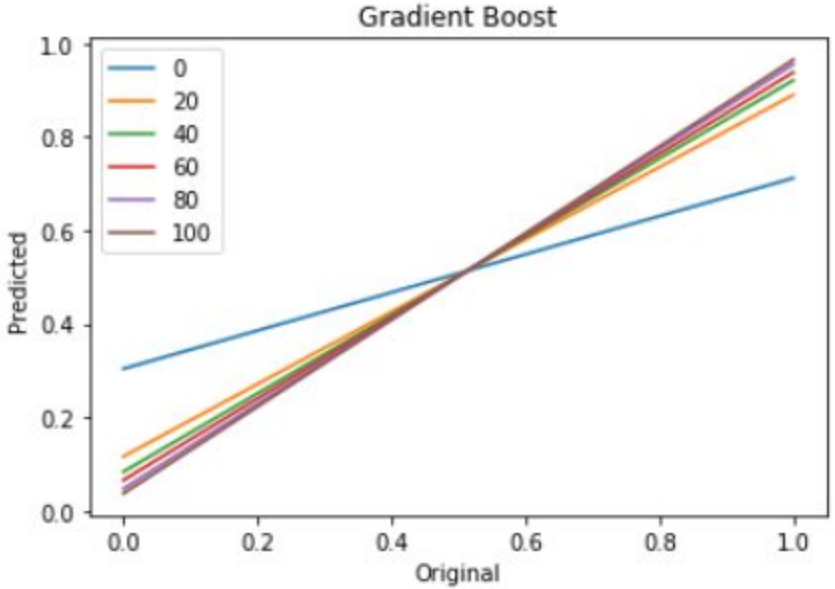
```
ANOVA Cross-Validation Scores: [0.69047619 0.73170732 0.70731707 0.68292683 0.775     ]

ANOVA Cross-Validation Score Averaged Across Folds: 71.75%.

Selected Features: ['cp', 'exang', 'oldpeak']
```

Neither Ridge nor Lasso Regressions have enough R squared to show the model fits the data well enough. The ANOVA shows that there is a significant difference between the means of the target and the means of the other features. There are 3 features that are highly correlated to the target variable: “cp” which is chest pain, “exang” which is an exercise induced pain, and “oldpeak” which is ST depression induced by exercise relative to rest.

```
Weak learner 0 R^2: -2.468939877125331
Weak learner 20 R^2: -0.05397616746412126
Weak learner 40 R^2: -0.017998362907154286
Weak learner 60 R^2: -0.018056631124267142
Weak learner 80 R^2: -0.027935908754259442
Weak learner 100 R^2: -0.023169191628083 62
```

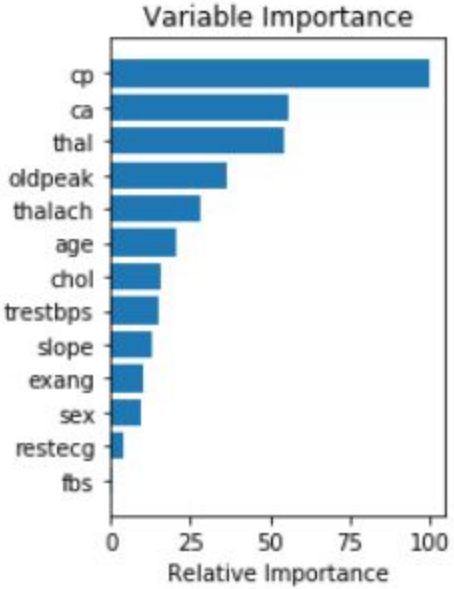



```
Overall R^2:
0.9792370863222565
```

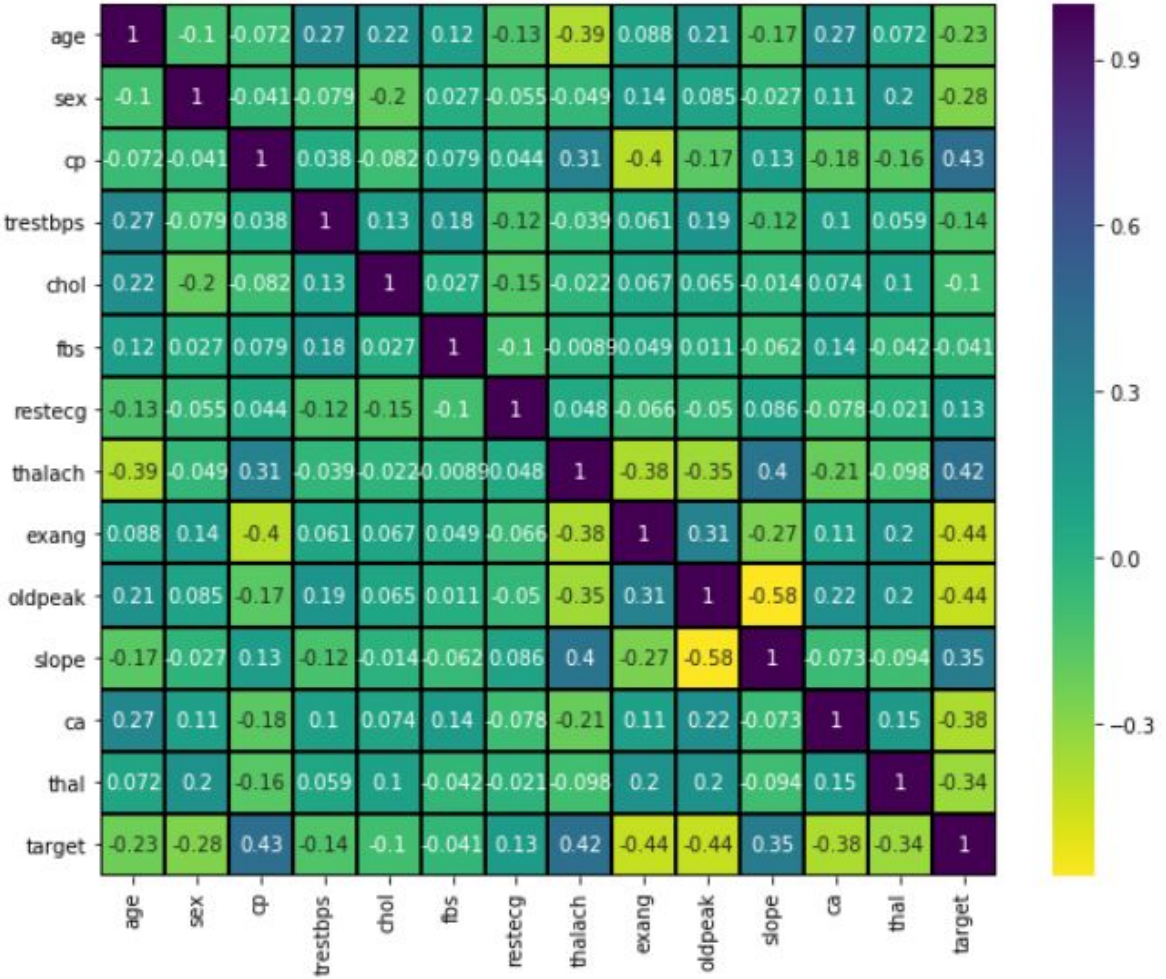
```
Training set accuracy:
Percent Type I errors: 0.0
Percent Type II errors: 0.0024390243902439024

Test set accuracy:
Percent Type I errors: 0.014634146341463415
Percent Type II errors: 0.01951219512195122
```

The gradient boost model has a R-squared of approximately 98% of the observed variation can be explained by the model's inputs. So far, this model and the random forest model are proving the best at fitting the data.



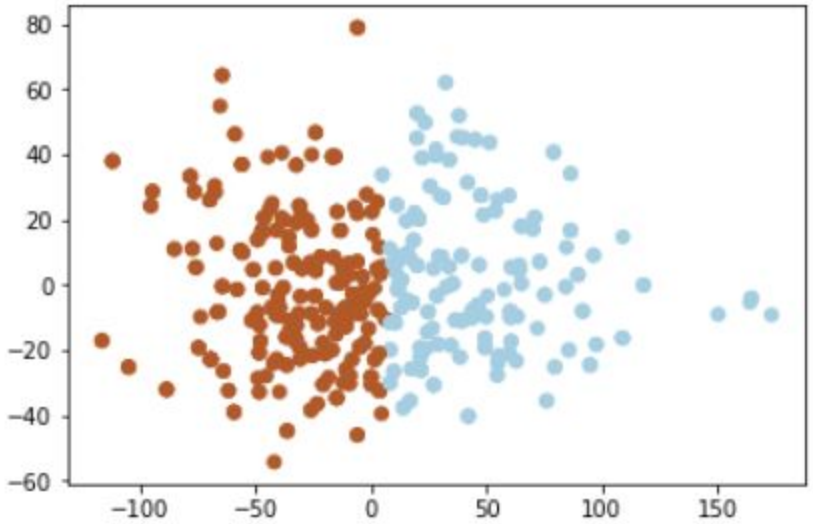
Both of these agree that cp is the most important feature. However, after that they differ as to which variables better correlate to the target variable.



```
 32/686 [>.............................] - ETA: 0s - loss: 0.2661 - accuracy: 0.9062WARNING:tensorflow:Early stopping cond
itioned on metric `val_loss` which is not available. Available metrics are: loss,accuracy
686/686 [==============================] - 0s 57us/sample - loss: 0.3344 - accuracy: 0.8455
Epoch 98/100
 32/686 [>.............................] - ETA: 0s - loss: 0.3646 - accuracy: 0.7812WARNING:tensorflow:Early stopping cond
itioned on metric `val_loss` which is not available. Available metrics are: loss,accuracy
686/686 [==============================] - 0s 60us/sample - loss: 0.3638 - accuracy: 0.8353
Epoch 99/100
 32/686 [>.............................] - ETA: 0s - loss: 0.3418 - accuracy: 0.8438WARNING:tensorflow:Early stopping cond
itioned on metric `val_loss` which is not available. Available metrics are: loss,accuracy
686/686 [==============================] - 0s 54us/sample - loss: 0.3345 - accuracy: 0.8382
Epoch 100/100
 32/686 [>.............................] - ETA: 0s - loss: 0.3609 - accuracy: 0.8438WARNING:tensorflow:Early stopping cond
itioned on metric `val_loss` which is not available. Available metrics are: loss,accuracy
686/686 [==============================] - 0s 58us/sample - loss: 0.3457 - accuracy: 0.8411
339/1 - 0s - loss: 0.3035 - accuracy: 0.8112

[0.3746840359978268, 0.81120944]
```

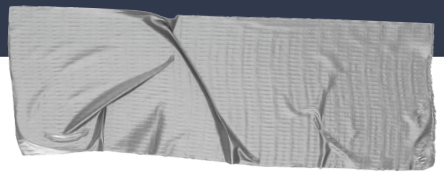
This is a snapshot of the bottom of a Keras epoch run. As you can see the evaluations of x_test and y_test at the bottom. The average model evaluation for y_test is 81% so this model is about 81% accurate. I did the Keras run to provide more variety to the models run on the data.



col_0	0	1
row_0		
0	170	117
1	158	241

accuracy score 0.5991253644314869.

An example of unsupervised learning run on the data. I split the data into 2 clusters because my target variable has only 2 options: diagnosed with heart disease and diagnosed with not having heart disease. Again, done to provide another explanation of the data, I don't believe that this accurately represents the data as shown by the accuracy score of approximately 60%.



Conclusion

Chest Pain is the most indicative of having heart disease. Each of these features are statistically significant and help in the explanation of heart disease.

➔ **Milestones**

We have found the most indicative factor of heart disease.

➔ **What's next?**

Find other studies that confirm or deny this conclusion. Find other datasets and see if those conclusions are consistent with this one.