MVA Individual Project
Name - Sarjak Maniar
Email - sm2732@scarletmail.rutgers.edu

Heart Disease Dataset

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them.

The "target" field refers to the presence of heart disease in the patient.
It is integer-valued 0 = no disease and 1 = disease.

This is a multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis.
It is composed of 14 attributes.

One of the major tasks on this dataset is to predict, based on the given attributes of a patient that whether that particular person has heart disease or not, and the other is the experimental task to diagnose and find out various insights from this dataset, which could help in understanding the problem more.

There are various factors associated with the process of determining whether a person will have heart disease or not. In this project, we will do an analysis of some hypotheses and will come up with some conclusions for the same.

Dataset columns:

1) age: The person’s age in years

2) sex: The person’s sex (1 = male, 0 = female)

3) cp: chest pain type
- — Value 0: asymptomatic
- — Value 1: atypical angina
- — Value 2: non-anginal pain
- — Value 3: typical angina

4) trestbps: The person’s resting blood pressure (mm Hg on admission to the hospital)

5) chol: The person’s cholesterol measurement in mg/dl

6) fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

7) restecg: resting electrocardiographic results
Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
Value 1: normal
Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

8) thalach: The person's maximum heart rate achieved

9) exang: Exercise induced angina (1 = yes; 0 = no)

10) oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)

11) slope: the slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping
0: downsloping; 1: flat; 2: upsloping

12) ca: The number of major vessels (0–3)

13) thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
Value 1: fixed defect (no blood flow in some part of the heart)
Value 2: normal blood flow
Value 3: reversible defect (a blood flow is observed but it is not normal)

14) target: Heart disease (1 = no, 0= yes)

Note: I have stated the hypothesis, conclusion, and all the supporting evidence that led to the conclusion in the R file.
Here I am giving an overview of the analysis which I have performed

1) Hypothesis 1

Sex: Studies have shown that men are more likely to develop heart disease than women.
In the Heart Disease dataset, we can use a contingency table and chi-square test to investigate the association between sex and the presence of heart disease.

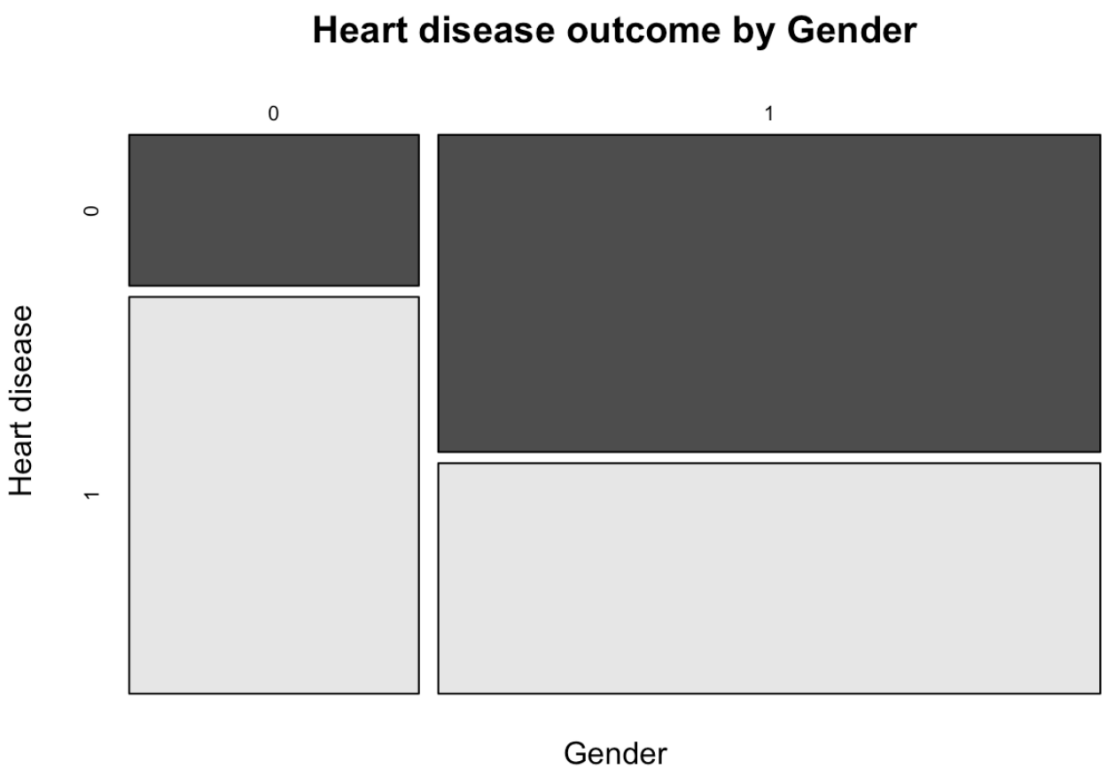
Which population is diagnosed more with heart disease? Male Population or Female Population?

The signs of a woman having a heart attack are much less noticeable than the signs of a male. In women, heart attacks may feel uncomfortable squeezing, pressure, fullness, or pain in the center of the chest. It may also cause pain in one or both arms, the back, neck, jaw, or stomach, shortness of breath, nausea, and other symptoms. Men experience typical symptoms of heart attack, such as chest pain, discomfort, and stress. They may also experience pain in other areas, such as arms, neck, back, and jaw, and shortness of breath, sweating, and discomfort that mimics heartburn.

There are 30.4 % females and 69.6% males in the dataset

	Has heart disease	Does not have heart disease
Females	86	226
Males	413	300

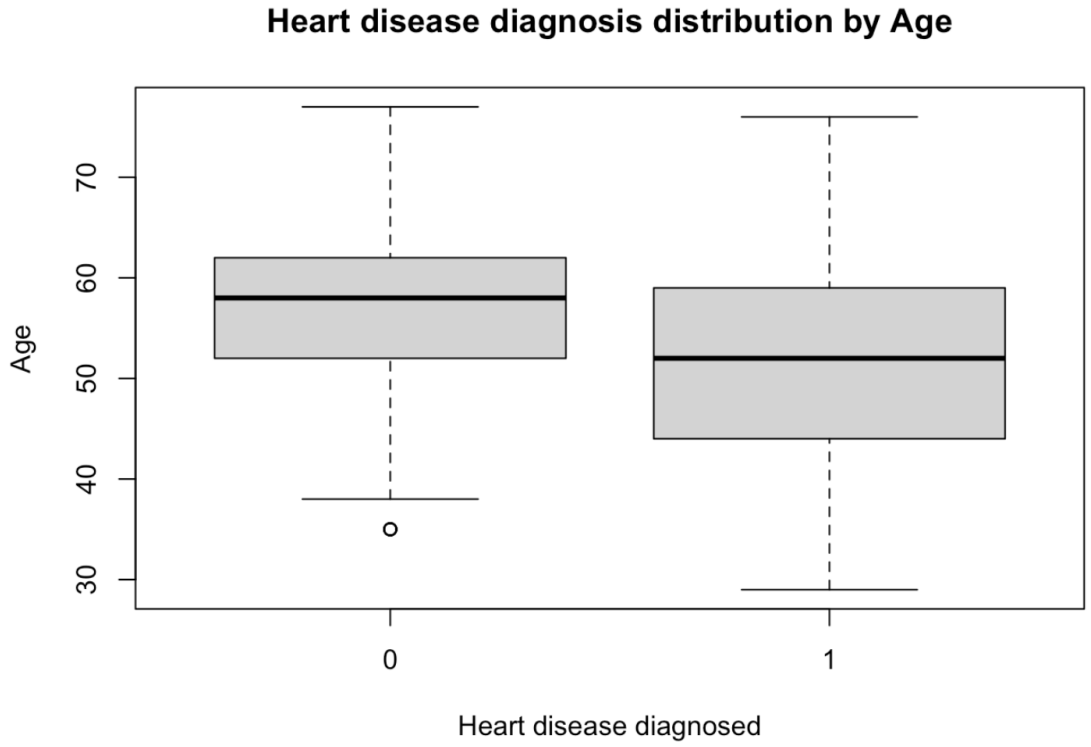
There are 86 females out of 312 who have been diagnosed with heart disease, and 413 males out of 713 were diagnosed with heart disease.
This indicates that 58% of males in this dataset are diagnosed with heart disease, whereas is only 27% of females are diagnosed with heart disease.



2) Hypothesis 2

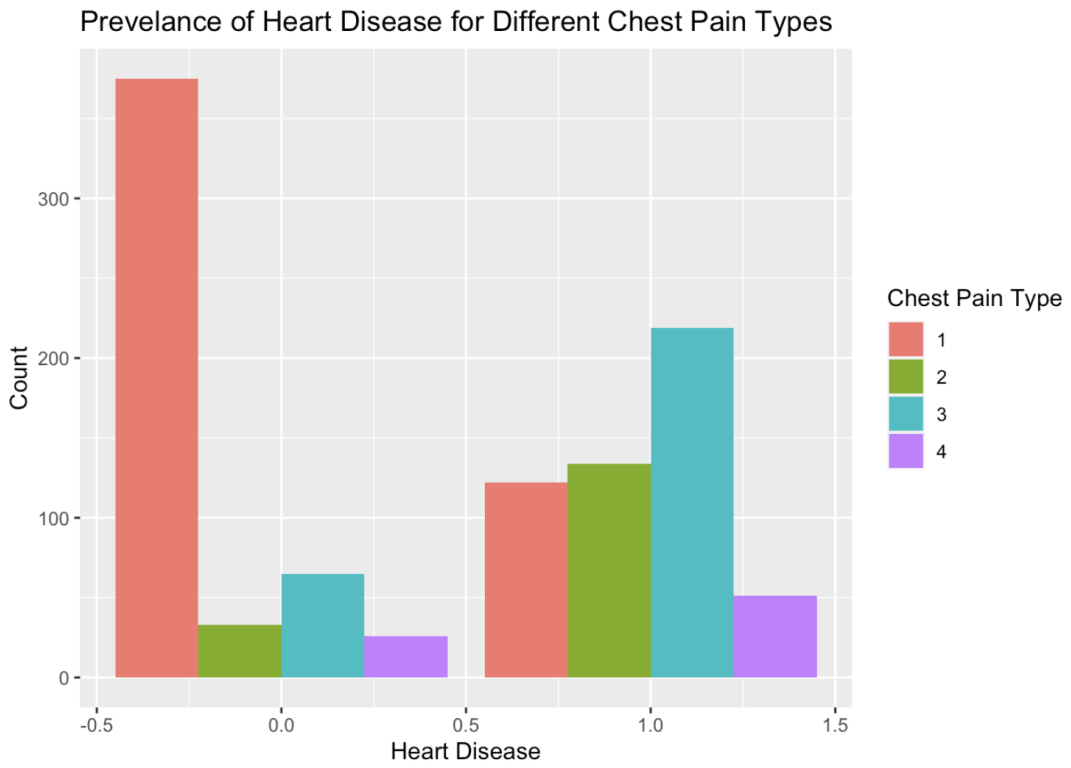
Age: As age increases, the likelihood of developing heart disease also increases. This is supported by previous studies on cardiovascular disease. In the Heart Disease dataset, we can use a scatter plot to visualize the relationship between age and the presence of heart disease and a correlation test to measure the strength of the relationship.

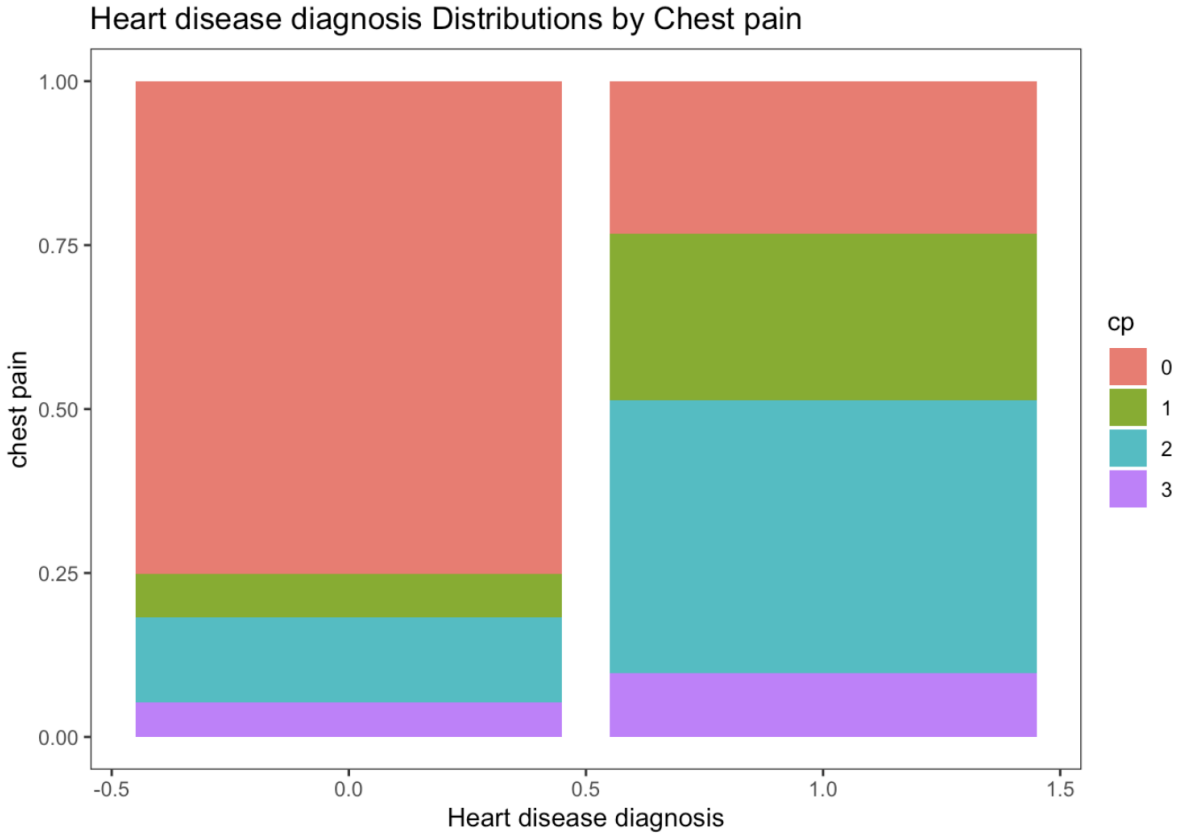

Prevelance of Heart Disease Across Age
0
1
60
40
20
0
Count
30
40
50
60
70
80
Age (years)
Heart Disease
1.00
0.75
0.50
0.25
0.00



Ans: We can conclude that people with a higher age are more likely diagnosed with a heart disease

3) Hypothesis 3
Chest pain type: The type of chest pain a patient experiences can provide importantly diagnostic information about heart disease.
cp: chest pain type
- — Value 0: asymptomatic
- — Value 1: atypical angina
- — Value 2: non-anginal pain
- — Value 3: typical angina


Prevelance of Heart Disease for Different Chest Pain Types
Count
300
200
100
0
-0.5
0.0
0.5
1.0
1.5
Heart Disease
Chest Pain Type
1
2
3
4


Heart disease diagnosis Distributions by Chest pain
chest pain
1.00
0.75
0.50
0.25
0.00
-0.5
0.0
0.5
1.0
1.5
Heart disease diagnosis
cp
0
1
2
3

There does appear to be a relationship between the type of chest pain and heart disease.
Interestingly, the asymptomatic chest pain type (Value 0) has the highest count for the presence of heart disease, while typical angina pain has the lowest count. There is a higher count of people without heart disease that have atypical or typical angina chest pain compared to people with heart disease. Angina is listed as one of the most common symptoms of heart attack, so this result is skeptical and needs further investigation, but we will assume it is correct for the current analysis.

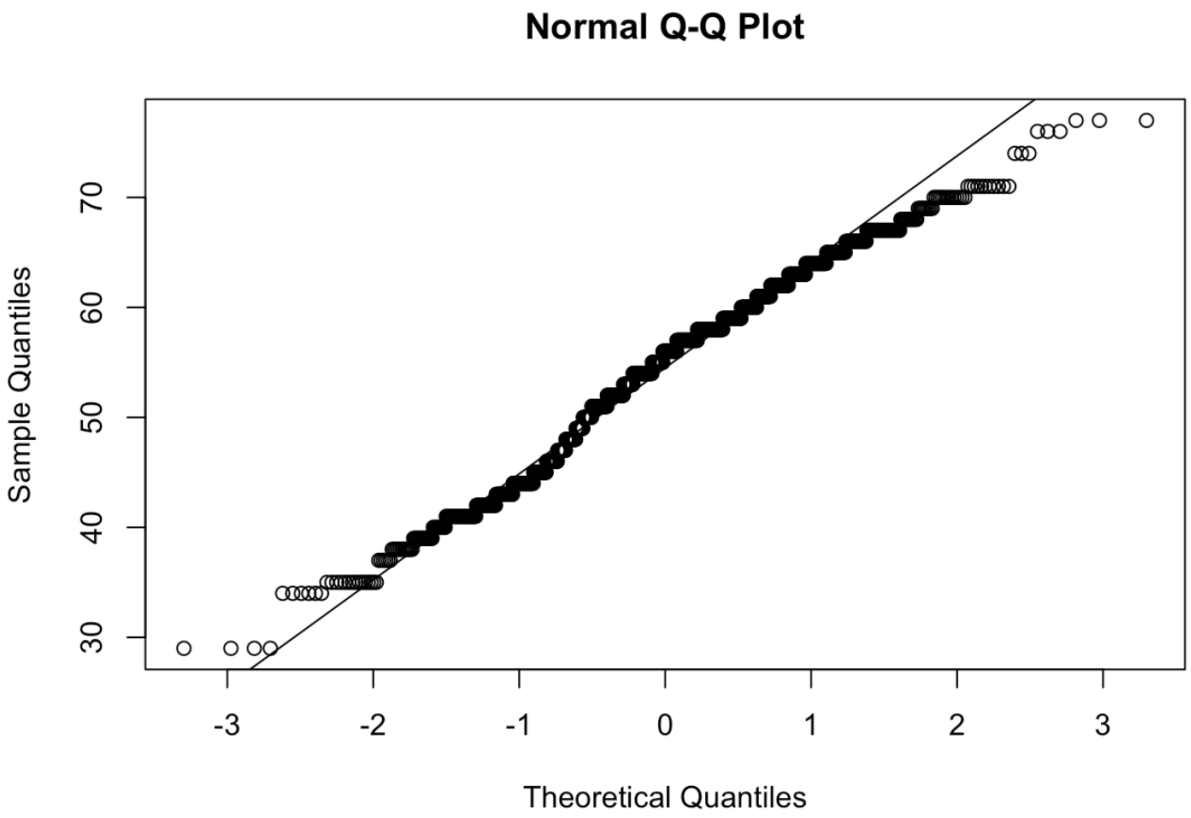
Ans: Therefore, chest pain type is a significant predictor of heart disease in patients.

Further, with more visualizations, we got the significant features:
"sex", "cp", "restecg", "exang", "slope", "ca", "thal"

4) Hypothesis 4

H0 = There is no association between chest pain and heart disease diagnosis

HA = There is an association between chest pain and heart disease diagnosis


Normal Q-Q Plot
Sample Quantiles
30
40
50
60
70
-3
-2
-1
0
1
2
3
Theoretical Quantiles

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = Train_Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5206  -0.3903   0.1197   0.5902   2.7044
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.970265   1.891980   3.156 0.001602 **
## age         -0.006612   0.014684  -0.450 0.652512
## sex         -1.760092   0.292218  -6.023 1.71e-09 ***
## cp           0.839629   0.115373   7.278 3.40e-13 ***
## trestbps    -0.020515   0.006730  -3.048 0.002303 **
## chol        -0.005038   0.002348  -2.145 0.031937 *
## fbs         -0.392153   0.331628  -1.183 0.237004
## restecg      0.442214   0.217176   2.036 0.041730 *
## thalach      0.026390   0.006503   4.058 4.95e-05 ***
## exang       -0.985515   0.260928  -3.777 0.000159 ***
## oldpeak     -0.591226   0.136340  -4.336 1.45e-05 ***
## slope        0.378445   0.223093   1.696 0.089819 .
## ca          -0.741602   0.118590  -6.254 4.01e-10 ***
## thal        -0.998061   0.183487  -5.439 5.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1064.15  on 767  degrees of freedom
## Residual deviance:  546.73  on 754  degrees of freedom
## AIC: 574.73
##
## Number of Fisher Scoring iterations: 6
```

5) Hypothesis 5

Is there a significant difference in the mean cholesterol levels between individuals with and without heart disease?
chol: The person’s cholesterol measurement in mg/dl

```
##
##  Two Sample t-test
##
## data:  heart$chol by heart$target
## t = 3.2134, df = 1023, p-value = 0.001353
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   4.015552 16.611444
## sample estimates:
## mean in group 0 mean in group 1
##        251.2926        240.9791
```

p-value = 0.001353

Since this value is less than 0.05, we can reject the null hypothesis and conclude that there is a significant difference in the mean cholesterol levels between individuals with and without heart disease.

Conclusion

From the above observations, we got the significant features which is correct:
"sex", "cp", "restecg", "exang", "slope", "ca", "thal"

1- Males are more vulnerable to be diagnosed with heart disease than females.

2- Chest Pain is most common factor that leads to heart disease for males and females.

3- Maximum heart rate achieved is the highest cause factor to cause heart disease for females where is Thalassemia is the highest to cause heart disease for males.

4- There is a high association between chest pain and heart disease diagnosis.

Limitation
The dataset is missing some useful information such as smoking, obesity or family history that can help in predicting.

The following conditions are associated with the increased prevalence of heart disease:

- Asymptomatic angina chest pain (relative to typical angina chest pain, atypical angina pain, or non-angina pain)
- Presence of exercise induced angina
- Lower fasting blood sugar
- Flat or down-sloaping peak exercise ST segment
- Presence of left ventricle hypertrophy
- Male
- Higher thelassemia score
- Higher age
- Lower max heart rate achieved
- Higher resting blood pressure
- Higher cholesterol
- Higher ST depression induced by exercise relative to rest