Enterprise Database Technologies

INSTITUTE OF technology tallaght

CA1

x00111602

2018

Contents

[Part 1: Getting to know the Dataset using R. 2](#_Toc508710429)

[1. Information on each predictor variable 2](#_Toc508710430)

[a. Attribute Type 2](#_Toc508710431)

[b. % of missing data 2](#_Toc508710432)

[c. Max 2](#_Toc508710433)

[c. Min 2](#_Toc508710434)

[c. Mean 2](#_Toc508710435)

[c. Mode 2](#_Toc508710436)

[c. Median 3](#_Toc508710437)

[c. Standard Deviation 3](#_Toc508710438)

[d. The type of distribution (e.g normal) 3](#_Toc508710439)

[e. Is the numeric data skewed? Type 3](#_Toc508710440)

[f. The level of correlation (Highest) 3](#_Toc508710441)

[Overall findings/observations 4](#_Toc508710442)

[2. Construct a histogram for each numerical variable, with an overlay of the target variable. 5](#_Toc508710443)

[Construct a bar chart for each categorical variable, with an overlay of the target variable. 8](#_Toc508710444)

[Graphically and statistically detect outliers. 10](#_Toc508710445)

[Part 2: Cleaning and Transforming the Dataset (using Weka) 12](#_Toc508710446)

[6. Using Weka 12](#_Toc508710447)

[7. Using Weka 14](#_Toc508710448)

# Part 1: Getting to know the Dataset using R.

## Information on each predictor variable

### Attribute Type

Age: numeric Sex: binary Cp: nominal Trestbps: numeric Cholesterol: numeric

Fasting blood sugar < 120: binary Restecg: nominal Diastpexerc: numeric

Thalach: numeric Exang: binary Oldpeak: numeric Slope: nominal Ca: ordinal

Thal: nominal Class: binary

### % of missing data

Age: 0 Sex: 0 Cp: 0 Trestbps: 0 Cholesterol: 0.64%

Fasting blood sugar < 120: 0 Restecg: 0.97% Diastpexerc: 0

Thalach: 0 Exang: 0 Oldpeak: 0 Slope: 0

Ca: 0 Thal: 0 Class: 0.64%

### Max

Age: 87 Sex: Male Cp: NoTang Trestbps: 200 Cholesterol: 564

Fasting blood sugar < 120: 1 Restecg: Normal Diastpexerc: 128

Thalach: 202 Exang: 1 Oldpeak: 6.2 Slope: Up

Ca: 3 Thal: Rev Class: Sick

### Min

Age: 18 Sex: Female Cp: Asymptomatic Trestbps: 94 Cholesterol: 126

Fasting blood sugar < 120: 0 Restecg: Abnormal Diastpexerc: 60.16

Thalach: 71 Exang: 0 Oldpeak: 0 Slope: Down

Ca: 0 Thal: Fix Class: Healthy

### Mean

Age: 54.23701 Sex: NA Cp: NA Trestbps: 131.4318 Cholesterol: 246.2712

Fasting blood sugar < 120: 0.1461039 Restecg: NA Diastpexerc: 84.81338

Thalach: 149.7208 Exang: 0.324675 Oldpeak: 1.049351 Slope: NA

Ca: 0.6688312 Thal: NA Class: NA

### Mode

Age: 58 Sex: Male Cp: NoTang Trestbps: 120 Cholesterol: 204

Fasting blood sugar < 120: FALSE Restecg: Normal Diastpexerc: 76.8

Thalach: 162 Exang: FALSE Oldpeak: 0 Slope: Up

Ca: 0 Thal: Normal Class: Healthy

### Median

Age: 55 Sex: NA Cp: NA Trestbps: 130 Cholesterol: 240

Fasting blood sugar < 120: 0 Restecg: Normal Diastpexerc: 83.2

Thalach: 153 Exang: 0 Oldpeak: 0.8 Slope: NA

Ca: 0 Thal: NA Class: NA

### Standard Deviation

Age: 9.775942 Sex: NA Cp: NA Trestbps: 17.49002 Cholesterol: NA

Fasting blood sugar < 120: 0.3537851 Restecg: NA Diastpexerc: 10.71816

Thalach: 22.76805 Exang: 0.469015 Oldpeak: 1.166037 Slope: NA

Ca: 0.9312189 Thal: NA Class: NA

### The type of distribution (e.g normal)

Age: p-value = 0.0007914 -> Deviates from Normality

Trestbps: p-value = 8.431e-07 -> Normal Distribution

Cholesterol: p-value = 4.548e-09 -> Normal Distribution

Diastpexerc: p-value = 3.104e-07 -> Normal Distribution

Thalach: p-value = 5.608e-05 -> Normal Distribution

Oldpeak: p-value < 2.2e-16 -> Deviates From Normal Distribution

Ca: p-value < 2.2e-16 -> Deviates From Normality

### Is the numeric data skewed? Type

Age: -0.1612079, Negatively

Trestbps: 0.7219654, Positively

Cholesterol: 1.13237, Positively

Diastpexerc: 0.7654076, Positively

Thalach: -0.5391399, Negatively

Oldpeak: 1.229654, Positively

Ca: 1.186547, Positively

### The level of correlation (Highest)

Age: 0.2385335 diastbpexerc

Trestbps: 0.9505095 diastbpexerc

Cholesterol: NA

Diastpexerc: 0.9505095 trestbps

Thalach: -0.04608416 trestbps

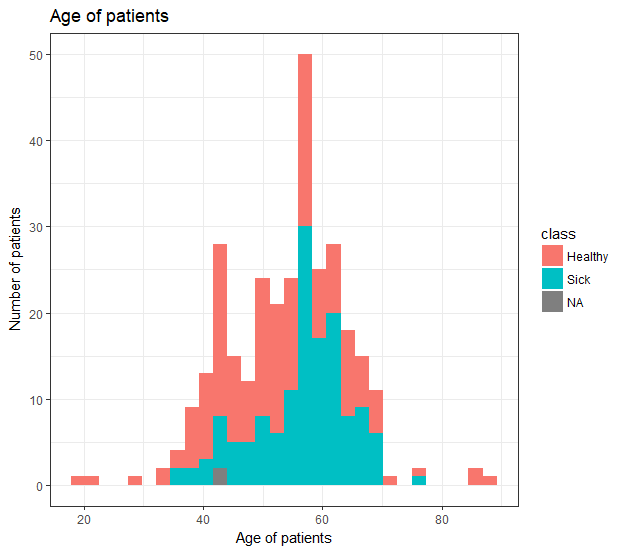
Oldpeak: 0.2164572 age

Ca: 0.3120721 age

### Overall findings/observations

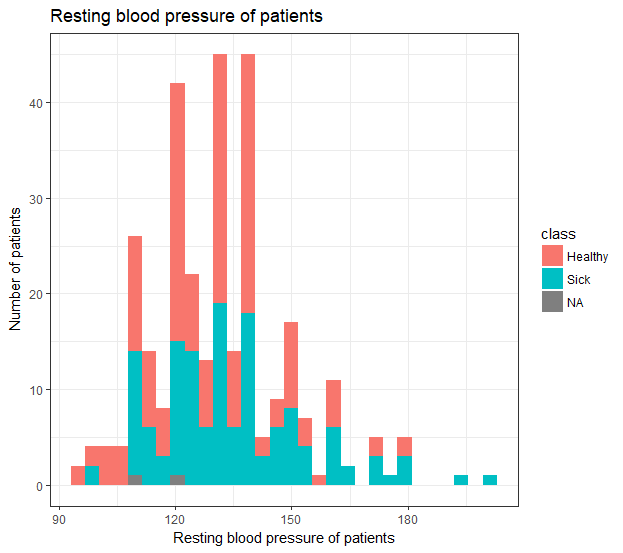
From

## Construct a histogram for each numerical variable, with an overlay of the target variable.

**Age Histogram:**

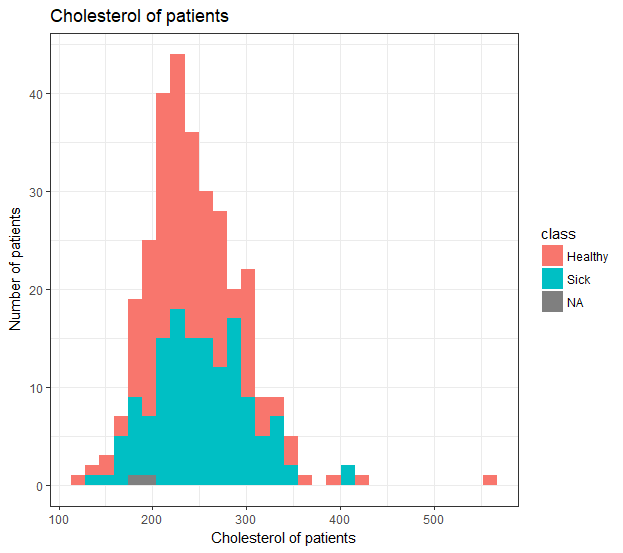
We could see from this histogram that the majority of sick patients are aged 35 to 70 years old. We could also deduct from the graph that thirty out of the fifty 55-57 year olds were sick. You can see from the histogram that the older the patients are (the further you go down the right of the X axis), the more sick the patients are. Also, if there were more older patients in the range 70-80 you could say from looking at the overlay of the class variable that they would also be sick. There seems to be more younger patients that are healthy than there are older patients.

From reading the age attribute histogram and also looking at the skewness of the attribute you could say that it would make an appearance in a machine learning algorithm. Due to the negative skewness and the occurrence of outliers using either IQR or (question 4) you could assume that the variable would make a significant appearance in a machine learning classification model. Using logistic regression where p is the probability of the outcome being a sick or healthy patient I think age would definitely be a variable to use in this type of scenario, using it over class would output the probability of patient being sick or healthy.

**Trestbps Histogram:**

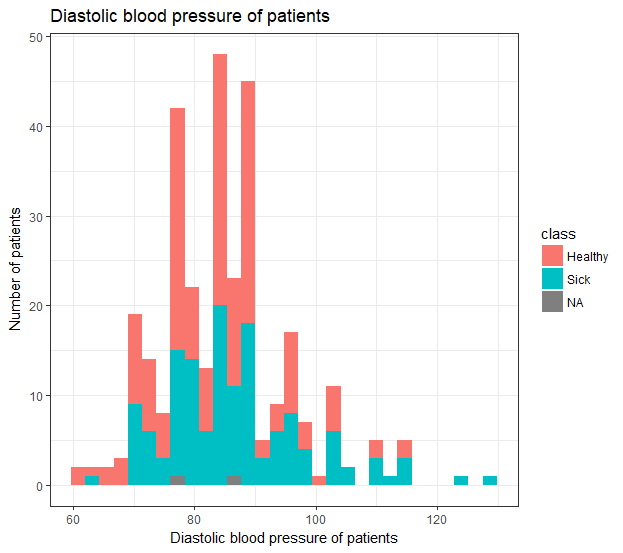
We can see that the majority of sick patients have a resting blood pressure of approximately 100-155. That being said it is clear that there are much more healthy patients within the same resting blood pressure bracket than there is sick patients. Nearly all patients with an RBP of over 160 are sick.

In terms of making an appearance in a machine learning classification model, the attributes distribution is normal distribution and positively skewed, skewness in a machine learning model is not ideal in regression modelling. Also the range of values in the data are close with min = 94 max = 200, putting the data into a logistic regression test it would be hard to say whether they would make a significant appearance in a machine learning model.

**Cholesterol Histogram:**

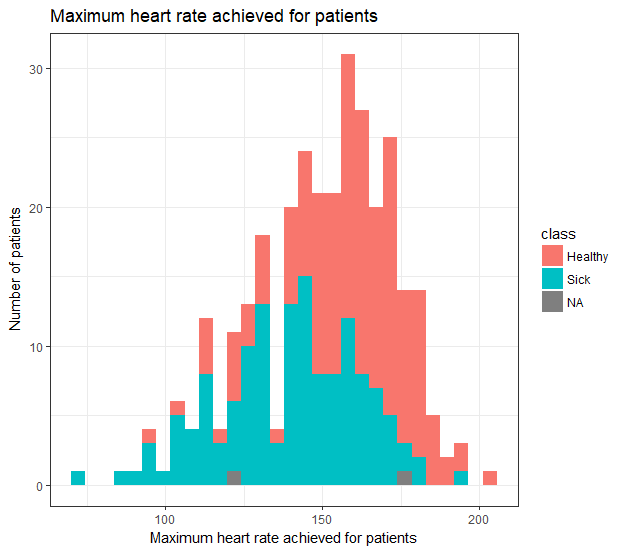
In this histogram the healthy patients greatly outweigh the sick patients. There are no more than 20 sick patients per each cholesterol reading, whereas the majority of the cholesterol readings have well over 20 healthy patients. The bulk of sick patients are in the 160-340 cholesterol reading range.

Because of the skewness and the look of normal distribution between the cholesterol of patients with the target class attribute I would say that this variable would make a significant appearance in a machine learning model. Although with some of the data missing in cholesterol it would be hard to perform to perform logistic regression testing on the attribute and to get the outcome if a patient is sick or healthy.

**Diastbpexerc Histogram:**

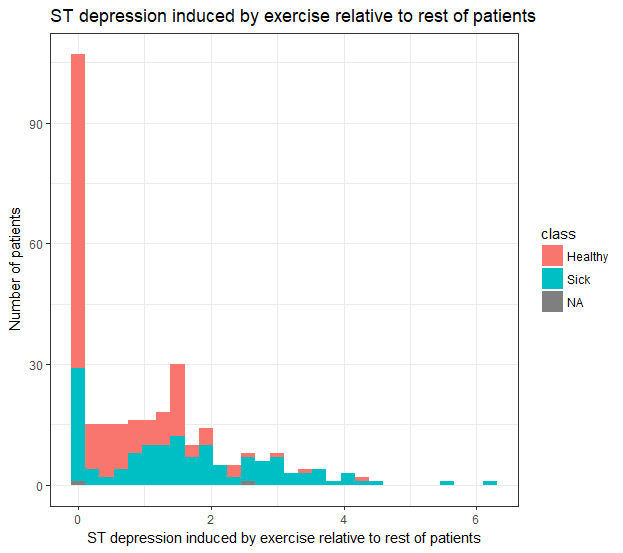
After viewing this histogram we established that it was extremely similar to the trestbps (Resting blood pressure) histogram. The attribute information tells us that trestbps is resting blood pressure and the diastbpexerc is also pressure in the arteries when the heart rests between beats. This would explain why the two graphs are so similar with a similar number of sick and healthy patients.

Most doctors define blood pressure too low if its below 90, causes can be strokes, heart attacks, and kidney failure, we would class anybody with these symptoms as ‘Sick.’ Looking at the histogram you can see that most of the sick patients are <= 90, taking this conclusion and analysis from the graph you could say that the Diastbpexerc attribute could be used in a machine learning classification model as the info the graph is giving us allows us to conclude it would give us an accurate probability of a patient being sick or healthy.

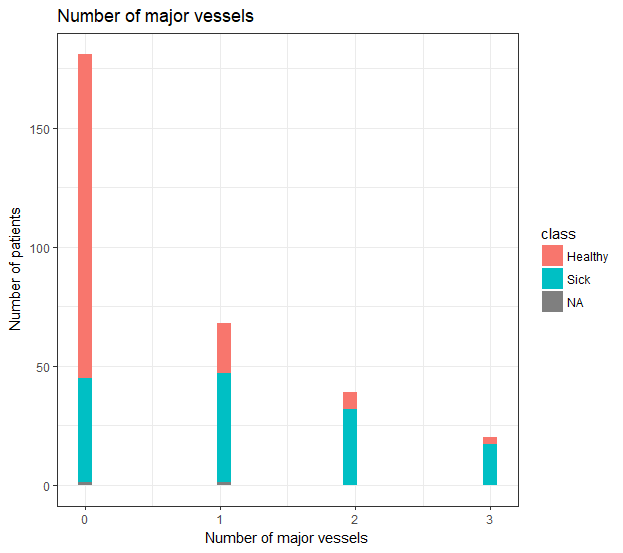
**Thalach Histogram:**

This histogram was the clearest example of a negatively distributed histogram. The majority of healthy patients had a high max heart rate while the majority of sick patients had a lower max heart rate. This might be because the healthier patients were more active and fit than the sick patients. This is a clear indication that keeping active and exercising means keeping healthy.

This attribute could be used in a machine learning model, taking the fact that a lower heart rate more or less means patient is sick and high heart rate means healthy you would expect this variable to make a significant appearance in a machine learning classification model.

**Oldpeak Histogram:**

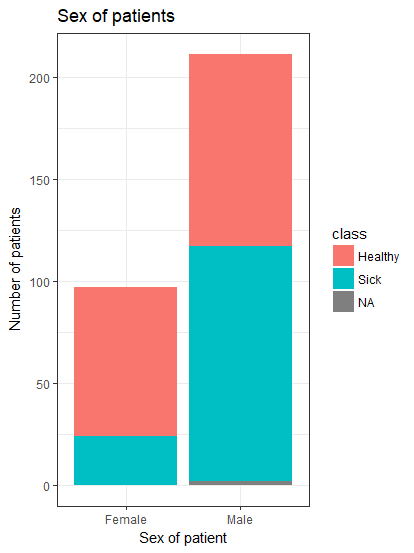
In this histogram the majority of patients are healthy with a large spike at 0 telling us that a lot of patients had 0 ST depression findings on the electrocardiogram. 30 of these patients however are still sick. Most patients with at least 2 ST depression readings where still healthy while most patients with over 3 readings where sick. This is a clear indication that the higher the ST depression finding then the more likely you are to be sick.



**Ca Histogram:**

This histogram shows us that if you have 1 or more major vessels you are nearly most certainly sick. If you have 0 then you are more than likely healthy with more than half of the patients with 0 readings being healthy. The readings for 1, 2 and 3 major vessels are nearly full of sick patients. Approx. 18 out of 20 patients with 3 major vessels are sick.

## Construct a bar chart for each categorical variable, with an overlay of the target variable.

**Sex Bar Chart:**

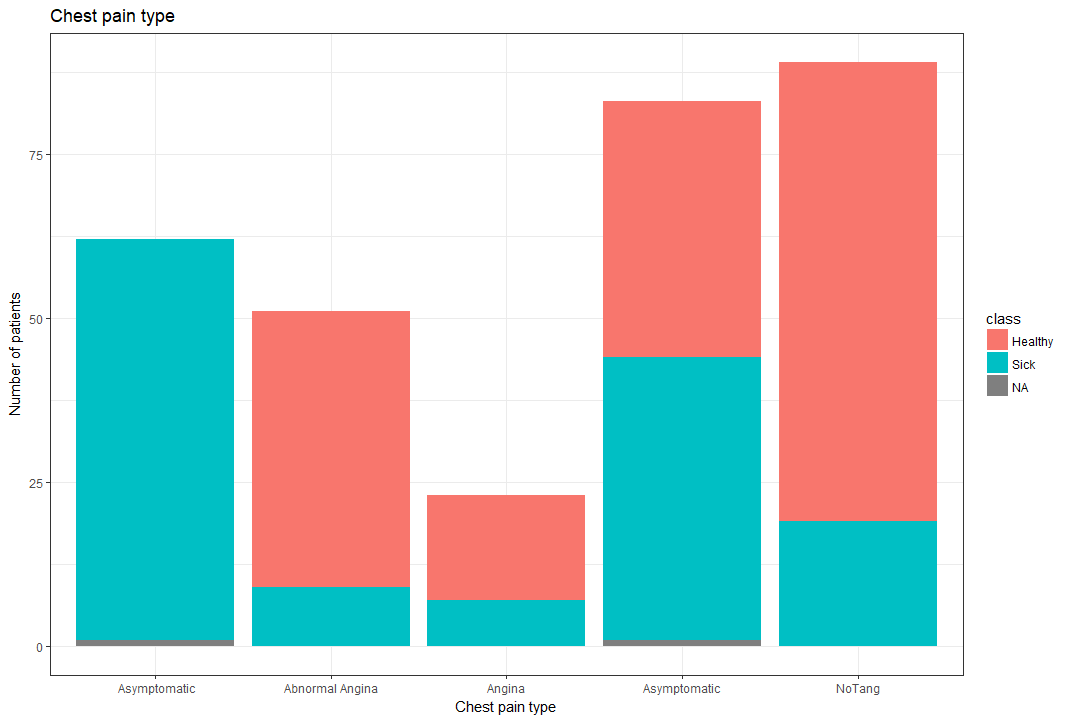
This bar chart showed us that approximately more than half of males were sick. This was not the same for females as approximately only one quarter of females were sick.We found from this bar chart that there were more sick males than there were females altogether in the data set.

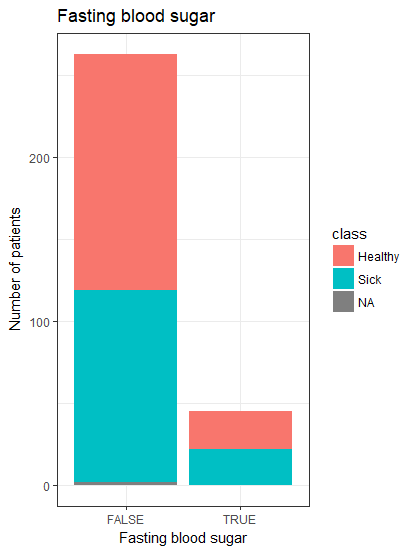
You would expect to see the Sex attribute used in a machine learning algorithm because it defines whether a patient is male or female, you can see from the histogram that there is an even correlation between the male sick and healthy patients which is useful when choosing attributes to use in a machine learning model.

**Chest Pain Bar Chart:**

From this bar chart we could see that all patients with Asymptomatic chest pains were sick. This means Asymptomatic is the most deadly chest pain of all. NoTang is the healthiest chest pain to have when it comes to scale as the most patients have this chest pain and it holds the healthiest amount of patients regarding chest pain.

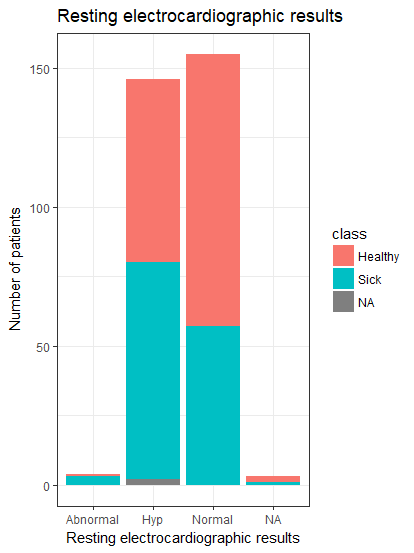
This attribute would be used in a machine learning model because you can conclude that patients in the Asymptomatic category are sick and most in NoTang are healthy, taking these conclusions would be useful in a machine learning model. In logistic regression this attribute would be helpful in defining the probability of a patient being sick or healthy.



**Fasting Blood Sugar Bar Chart:**

More than half of the patients that are not fasting are considered healthy. For patients that are fasting half are healthy and half are sick. If fasting is causing the sickness then going by this bar chart the healthier option would be to not fast.

With half the patients being sick or healthy in the False category it would be difficult for a machine learning model to decide whether a patient is sick or healthy based off Fasting blood sugar category.

****

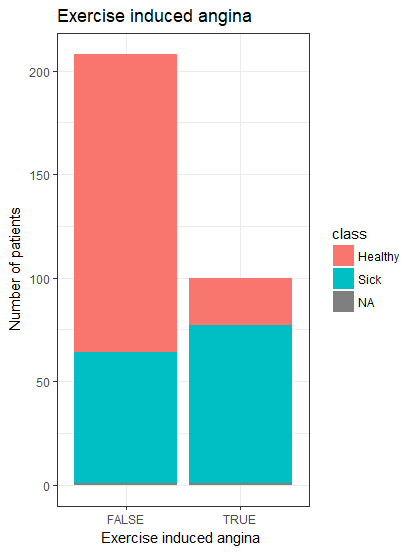
**Resting Electrocardiographic Bar Chart:**

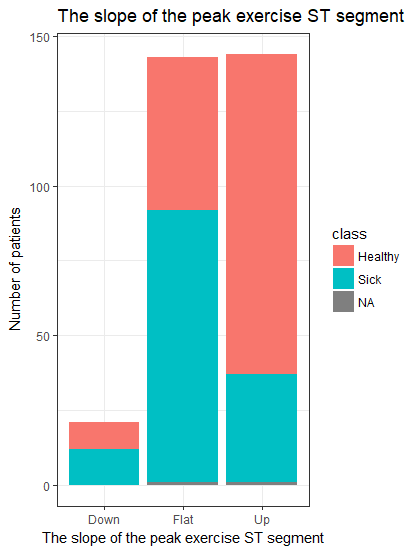
Most Abnormal results meant that the patients were sick. Most Hypertrophy results also meant that the patients were sick. More than half of patients with Normal results were considered healthy and were patients results were Non-Applicable there were slightly more healthier patients than sick.

Because the normal category has half sick and half healthy patients it would be difficult for a machine learning model to class whether a patient is sick or healthy based off this predictor attribute, therefore I would not expect to see this variable in a machine learning model.

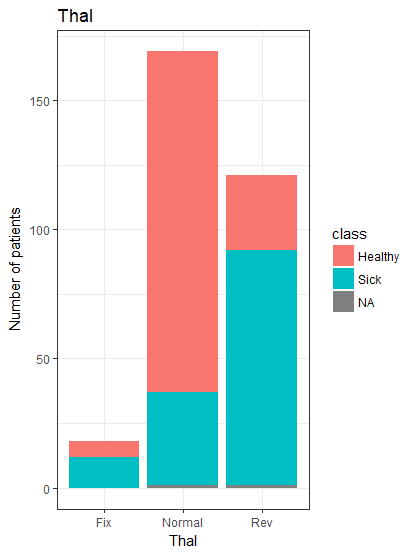
**Exercise Induced Angina Bar Chart:**

The majority of patients that did experience angina after exercise were considered sick. Patients that did not experience angina (200+ patients) were considered healthy with only 60 patients or so being considered sick without experiencing angina.

****

**Slope Bar Chart:**

There was slightly more sick patients than healthy where the slope of the peak exercise ST segment was down. There was much more healthy patients when the peak was up and when the peak was flat there was about two thirds more sick patients than healthy patients.

****

**Thalassemia Bar Chart:**

From looking at the bar chart it is clear to see that the majority of patients with a fixed defect or a reversible defect are considered sick. Patients with normal Thalassemia were mostly healthy with about one quarter of those being considered sick.

From this you could say that you would expect to see this attribute in a machine learning model because it almost definitively says that patients in Fix or Rev category are healthy and those with normal are sick, this data would be useful in a machine learning model, performing logistic regression on this attribute you would be able to calculate the probability of a patient being sick or healthy based off their thal category.

## Graphically and statistically detect outliers.

For this section we used ggplot’s boxplot to graphically determine the outliers in our numeric data.

To statistically detect outliers in the data we made a function that takes in the attribute and the column number of the attribute. We then found the interquartile range for the attributes data which gave us the upper and lower boundaries of the data. These boundaries were displayed on the console. If any number was outside of these boundaries it was considered an outlier. The function then displayed the outliers to the console on R.

## Investigate whether there are any correlated variables

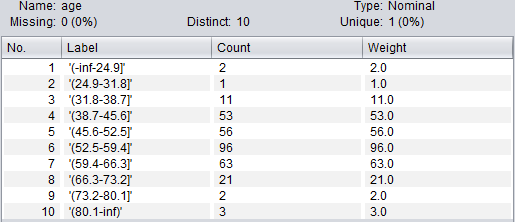
# Part 2: Cleaning and Transforming the Dataset (using Weka)

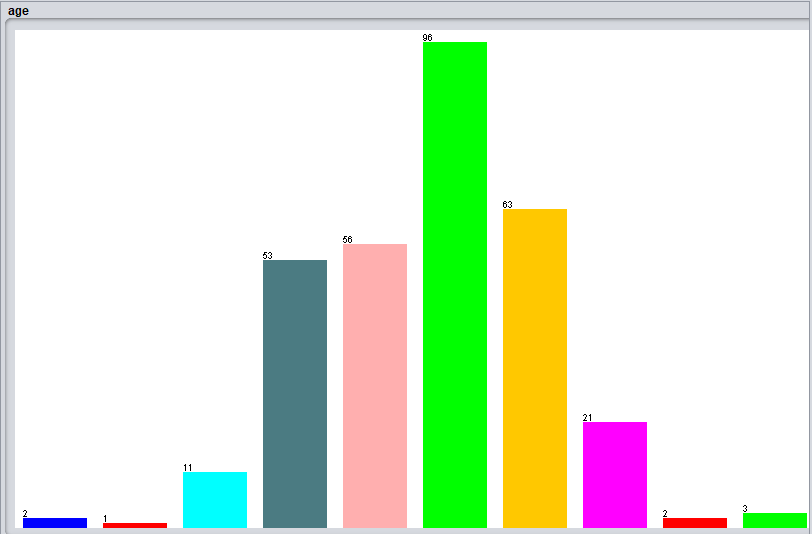
## 6. Using Weka

We chose age as our numeric variable to bin (discretise) the data using “equal width binning”. After loading the dataset into WEKA as an arff file we then began the cleaning and transforming steps. To perform bin (discretize) the data we set the number of bins to 10 and ran the algorithm. The algorithm output put the age variable into 10 bins with a range < 24.9 to > 80.1. It produced a bin interval of 6.9, nominal data with 0% missing and 1 unique value. The highest count with 96 values came from bin 6, the age range 52.5 – 59.4.

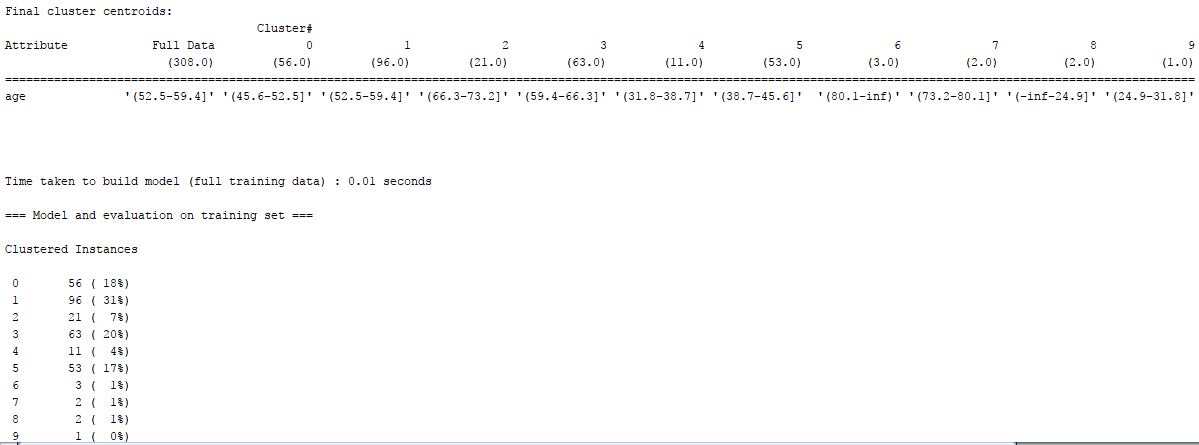
From looking at the histogram for age vs class you can see that most of the sick patients are also within this range.

Performing unsupervised discretization it broke the data into tasks that found the number of discrete values, the boundaries of the intervals, the range of the numeric age attribute.





Using K Means clustering algorithm on the numeric attribute we ran the simple k means function in WEKA. This produced a Clusterer output sheet with the run information, number of iterations: 2, number of clusters: 10, interval of 6.9. We chose 10 clusters to create a clustering algorithm similar to “equal width binning” algorithm. The output was similar to bin discretization but it grouped the clusters and produced a percentage of the count of patients in each cluster, the biggest cluster being the age group 52.5 – 59.4, 31% of the data and the smallest cluster being the age group 24.9 – 31.8, 0%.



For an optimal solution

## 7. Using Weka

## 8. Impute missing values

For this question we took the categorical variable restecg (resting electrocardiographic results) to develop an appropriate machine learning model to impute missing values for that attribute. There are a number of ‘NA’ rows in that column so imputing the missing values on this attribute makes sense. We used weka to perform the necessary actions. We started by making the restecg variable our target variable and deleting the current class target variable. The next step we took was to load the new impute file into weka explorer.