Ramaiah Institute of Technology

(An Autonomous Institute, Affiliated to VTU)

MSR Nagar, MSRIT post, Bangalore-54

A Dissertation Report on

CLUSTERING BY K-MEDOIDS

Submitted by

Suhas Goutham	1MS15CS127
Smaranita Vasudev	1MS15CS124
Shravan A R	1MS15CS115
Rakshith R	1MS15CS101

Bachelor of Engineering in Computer Science & Engineering

Under the guidance of

Name of the guide	Sowmya B.J
Designation	Assistant Professor
Dept name	Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

M.S.RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, Affiliated to VTU)

BANGALORE-560054

www.msrit.edu, **2017**

INTRODUCTION:

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. This includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, heart arrhythmia, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis. Heart disease is the No. 1 cause of death in the world and the leading cause of death in the United States.

Symptoms:

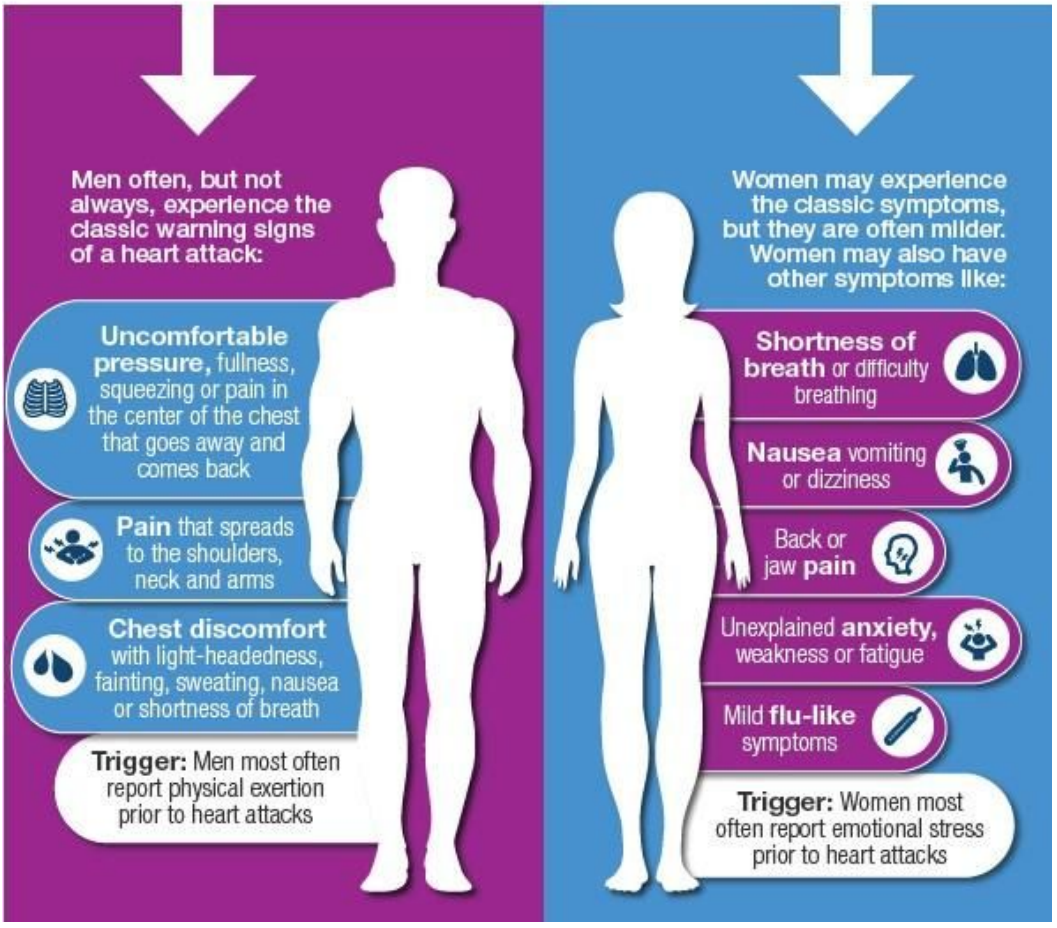
Some symptoms of heart disease are:

- **Pain in the chest**--the heart muscle is not getting enough flow to keep it going.
- **Trouble breathing**--blood may back up into the lungs.
- **Palpitations** (a feeling that the heart is beating too fast, too hard, or not regularly).
- **Swelling of feet or legs**--blood is backing up from the heart into the lower body.
- **Feeling weak** because the body and brain are not getting enough blood to supply them with oxygen.
- **Cyanosis** (skin turning a blue colour) means that too little oxygen is in the bloodstream to supply the cells in the body.

Treatment and Cure:

- Stop smoking
- Maintain a healthy lifestyle by consuming a healthy diet.
- Engage yourself in physical activities.
- Always maintain weight
- Avoid Stress and Depression
- Menopausal Hormone Therapy
- Procedures and Surgery like bypass and stents in the heart.
- Cardiac Rehabilitation

Comparison between men and women in getting a heart disease:



DATASET DESCRIPTION:

The dataset is: Heart disease in males.

SOURCE OF DATASET:

The link for the dataset:

eric.univ-lyon2.fr/~ricco/dataset/heart_disease_male.xls

ATTRIBUTES DESCRIPTION:

There are 8 attributes in the dataset.

1. **age –** It gives the age of the patient.

Age considered for analysis:

Minimum age=28 Maximum age=66

2. **chest_pain –** It tells the type of chest pain.

   chest_pain attribute in our dataset has the following values:

   - asympt: A condition might be **asymptomatic** if it fails to show the noticeable symptoms with which it is usually associated.
   - atyp_angina: Pleuritic,sharp,pricking,knife-like,pulsating,lancinating,choking pain.
   - non_anginal: A chest pain is very likely nonanginal if its duration is over 30 minutes or less than 5 seconds, it increases with inspiration, can be brought on with one movement of the trunk or arm, can be brought on by local fingers pressure, or bending forward, or it can be relieved immediately on lying down.
   - typ_angina: Sensations in chest of squeezing ,heaviness,pressure,vise-like aching.

3. **rest_bpress –** Indicates the resting blood pressure in mm hg on the admission to the hospital.

   rest_bpress attribute:

   - Min to Max in our dataset=92 to 200 mm hg
   - Generally,value for resting blood pressure in a healthy human being=140 mm hg

4. **blood_sugar –** It tells the fasting blood sugar. It's true if >120 mg/dl false otherwise.
5. **rest_electro –** It tells the type of resting electrocardiographic results.

   rest_electro attribute:

   - normal
   - left_vent_hyper: High blood pressure makes it difficult for your heart to pump blood. Like other muscles in your body, regular hard work causes your heart muscles to thicken and grow. This alters the way the heart functions. These changes usually happen in the main pumping chamber of the heart, the left ventricle. The condition is known as left ventricular hypertrophy (LVH)

- st_t_wave_abnormality:ST-T wave changes that are independent of changes in ventricular activation and that may be the result of global or segmental pathologic processes that affect ventricular repolarization:

  Drug effects (e.g., digoxin, quinidine, etc)

6. **max_heart_rate –** It tells the maximum heart rate achieved.
   - dataset maximum=188
   - dataset minimum=82
   - max heart rate of a healthy individual of age 47(mean age of our dataset) = 163

7. **exercice_angina –** It tells whether the patient has exercise induced angina.
   Angina may feel like pressure in the chest, jaw or arm due to exercise or stress.
8. **disease –** It tells whether the diagnosis was positive or negative i.e if the male has a heart disease or not .

DATA SET SIZE IN TERMS OF BYTES AND NUMBER OF TUPLES:

Size – 43.5 KB (44,544 bytes)

Number of Tuples – 209 Tuples.

INFERENCES WHICH CAN BE DRAWN WITH RESPECT TO EACH ATTRIBUTE:

1)We can infer if age is a factor of heart disease

2)We can also see how resting blood pressure and getting a disease are related

3)We can infer if blood sugar and heart disease are related

4)To see which kind of chest pain is prominent in different age groups

5)We can infer if pain while exercising, I. E exercise_angina and heart disease are related

6)Relation between resting blood pressure and blood sugar can be inferred

7)Inference as to how disease is determined by max_heart_rate and chest_pain can be found .

8)Is ecg a good indicator of heart disease? This can be inferred using the rest_electro attribute

ALGORITHM DESCRIPTION:

The k-medoids algorithm is a clustering algorithm which clusters the dataset by maximizing intra-group similarity and inter-group dissimilarity. It is an enhanced version of the k-means algorithm as it is more robust to noise and outliers.

This algorithm tries to minimize the distance between points in a cluster and the center point (medoid) by calculating the Manhattan distances from the point to the center. It groups the dataset of n-objects into k-clusters (value of k is determined by using Silhouette technique).

We can implement k-medoid clustering by using the Partitioning Around Medoids (PAM) algorithm. The algorithm works as follows:-

1) Initially we select the value of k from a dataset of n-objects and then select k objects from the dataset randomly as medoids (cluster centers).
2) We calculate the Manhattan distances from each object to all the k-centers and associate the data object to the nearest center (least distance).
3) We then find the total cost of this clustering by summing all the distances.
4) We repeat the above 3 steps by swapping the medoid to some other points in the dataset and find the total cost in each case.
5) Finally, we select the clustering with least total cost and the algorithm terminates.

R CODE SNIPPETS EXPLAINING ALGORITHM:

heart.R | abc | heart_disease_male

Source on Save | Run | Source

```r
library(xlsx)
library(cluster)
library(ggplot2)
abc <- read_excel("~/5TH SEM/DM/heart_disease_male.xls")
View(abc)
head(abc) #gives the first 6 values of a dataset
tail(abc)
str(abc)

summary(abc)

#Remove the outlier "?" in rest_electro
abc<-abc[!(abc$rest_electro=='?'),]

summary(abc)
str(abc)

#dotplot
ggplot(abc,aes(age,fill=disease))+geom_dotplot(binwidth = 0.80)
ggplot(abc,aes(rest_bpress,fill=disease))+geom_dotplot(binwidth = 2.00)
ggplot(abc,aes(max_heart_rate,fill=disease))+geom_dotplot(binwidth=1.50)
ggplot(abc,aes(blood_sugar,fill=disease))+geom_dotplot(binwidth=0.01)
ggplot(abc,aes(age,fill=chest_pain))+geom_dotplot(binwidth=0.90)
ggplot(abc,aes(chest_pain,fill=disease))+geom_dotplot(binwidth=0.050)
ggplot(abc,aes(rest_electro,fill=disease))+geom_dotplot(binwidth=0.0250)
```

1:1 (Top Level) R Script

heart.R | abc | heart_disease_male

Source on Save | Run | Source

```r
#plot using points
ggplot(abc,aes(age,max_heart_rate,color=disease))+geom_point()
ggplot(abc,aes(age,exercice_angina,color=disease))+geom_point()
ggplot(abc,aes(age,max_heart_rate,color=chest_pain))+geom_point()
ggplot(abc,aes(max_heart_rate,chest_pain,color=disease))+geom_point()


#change categorical values into numerical
abc$chest_pain<-sapply(abc$chest_pain,function(x) ifelse(x=="asympt",1,x))
abc$chest_pain<-sapply(abc$chest_pain,function(x) ifelse(x=="atyp_angina",2,x))
abc$chest_pain<-sapply(abc$chest_pain,function(x) ifelse(x=="non_anginal",3,x))
abc$chest_pain<-sapply(abc$chest_pain,function(x) ifelse(x=="typ_angina",4,x))


abc$disease<-sapply(abc$disease,function(x) ifelse(x=="positive",2,1))


abc$blood_sugar<-sapply(abc$blood_sugar,function(x) ifelse(x=="t",2,1))

abc$rest_electro<-sapply(abc$rest_electro,function(x) ifelse(x=="left_vent_hyper",1,x))
abc$rest_electro<-sapply(abc$rest_electro,function(x) ifelse(x=="st_t_wave_abnormality",2,x))
abc$rest_electro<-sapply(abc$rest_electro,function(x) ifelse(x=="normal",3,x))


abc$exercice_angina<-sapply(abc$exercice_angina,function(x) ifelse(x=="yes",2,x))
```

1:1 (Top Level) R Script

heart.R | abc | heart_disease_male

Source on Save | Run | Source

```
abc$exercice_angina<-sapply(abc$exercice_angina,function(x) ifelse(x=="yes",2,x))
abc$exercice_angina<-sapply(abc$exercice_angina,function(x) ifelse(x=="no",1,x))


cor(abc$age,abc$max_heart_rate)
#cor(abc[c("rest_bpress","disease")])
cor(abc[c("age","disease")])
#cor(abc[c("rest_bpress","blood_sugar")])
cor(abc[c("max_heart_rate","disease")])
cor(abc[c("max_heart_rate","blood_sugar")])
cor(abc[c("max_heart_rate","rest_electro")])

#to scale values
#abc<-scale(abc)
dataFrame<-as.data.frame(abc)
dataFrame
#for age and chestpain
pamx<-pam(dataFrame[c(2,8)],2)
pamx
pamx$medoids
inf1_a_c<-pamx$clustering


abc<-cbind(abc,inf1_a_c)
```

1:1 (Top Level) R Script

heart.R | abc | heart_disease_male

Source on Save | Run | Source

```
View(abc)

clusplot(pamx)

#age and heart disease
pamx1<-pam(dataFrame[c(1,8)],2)
pamx1
pamx1$medoids
inf2_a_c<-pamx1$clustering


abc<-cbind(abc,inf2_a_dis)
clusplot(pamx1)

#rest_b press and disease
rest_dis<-pam(dataFrame[c(3,8)],2)
rest_dis
rest_dis$medoids
inf3_r_dise<-rest_dis$clustering


abc<-cbind(abc,inf3_r_dise)
clusplot(rest_dis)

#rest_b press ,blood sugar and disease
```

1:1 (Top Level) R Script

heart.R | abc | heart_disease_male

Source on Save | Run | Source

```
#rest_b press ,blood sugar and disease
r_sug_dis<-pam(dataFrame[c(3,8)],2)
r_sug_dis
r_sug_dis$medoids
inf4_r_sug_dis<-r_sug_dis$clustering


abc<-cbind(abc,inf4_r_sug_dis)
clusplot(r_sug_dis)

#getting heart attack and max_heart rate are related
heart<-pam(dataFrame[c(6,8)],2)
heart
heart$medoids
inf<-heart$clustering


abc<-cbind(abc,inf)
clusplot(heart)

#chest pain in dif age groups
chest<-pam(dataFrame[c(1,2)],4)
chest
chest$medoids
inf1<-chest$clustering
```

1:1 (Top Level) R Script

heart.R | abc | heart_disease_male

Source on Save | Run | Source

```
clusplot(r_sug_dis)

#getting heart attack and max_heart rate are related
heart<-pam(dataFrame[c(6,8)],2)
heart
heart$medoids
inf<-heart$clustering


abc<-cbind(abc,inf)
clusplot(heart)

#chest pain in dif age groups
chest<-pam(dataFrame[c(1,2)],4)
chest
chest$medoids
inf1<-chest$clustering


abc<-cbind(abc,inf1)
clusplot(chest)
#plot(abc,pamx$clustering)
#ggplot(pamx,aes(age,chest_pain,color=disease))+geom_point()
#clusplot(pam(abc[c("age","disease")],2),xlab="age",ylab="disease",main="Graph")

```

1:1 (Top Level) R Script

INFERENCES WITH SNAPSHOTS OF THE GRAPHS:

NOTATIONS:

In our dataset we have categorical values. We converted them to numerical values to make it easier to work with:

1.Chest_pain has the following attribute values which have been converted into their respective numerical values using “sapply”:-

- asympt->1
- atyp_angina->2
- non_anginal->3
- typ_angina->4

2. Disease has positive and negative as its categorical values which are converted as follows:-

- positive->2
- negative->1

3. Blood_sugar attribute has two categorical values:-

- t->2
- f->1

4.rest_electro attribute has the following categorical values:-

- left_vent_hyper->1
- st_t_wave_abnormality->2
- normal->3

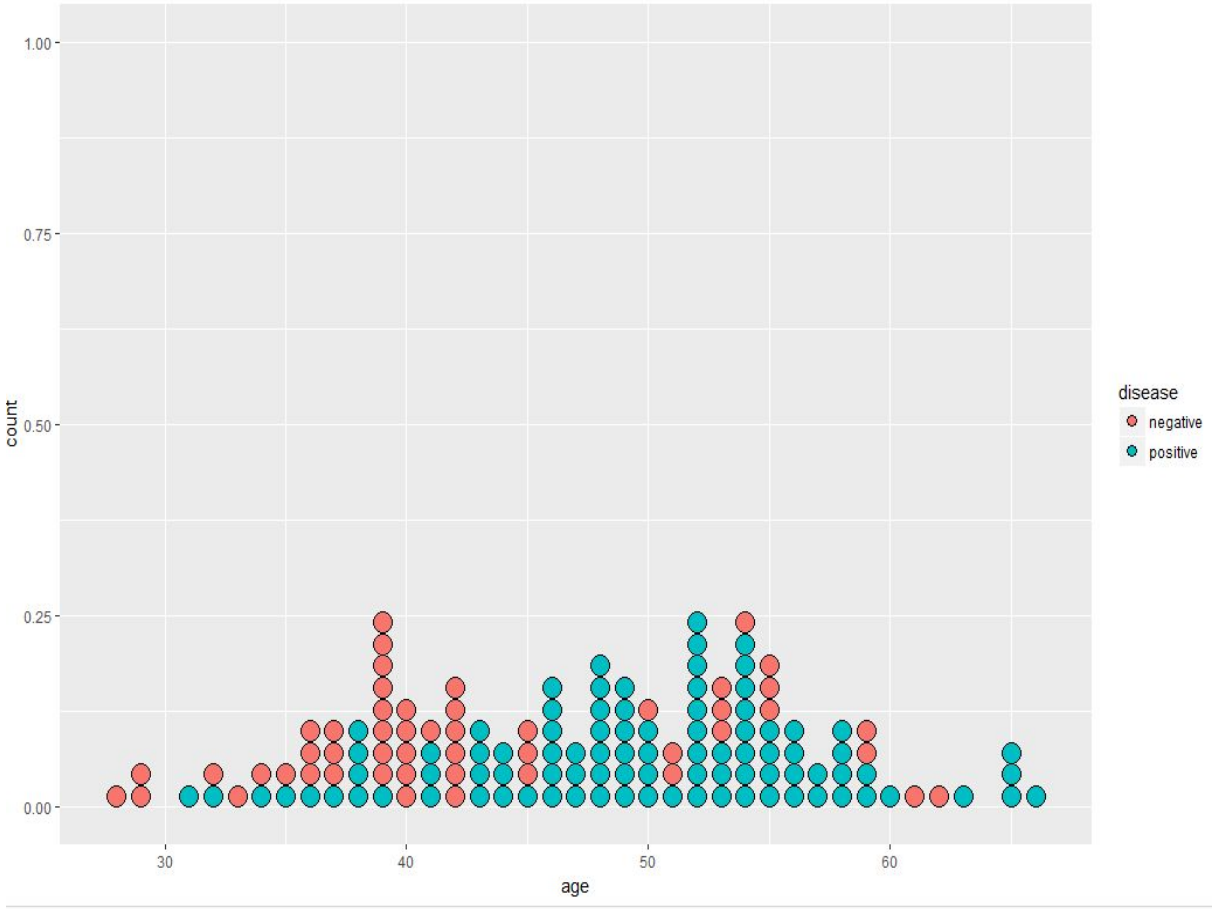
5. Excercise_angina attribute has the following categorical values:-

- yes->2
- no->1

.

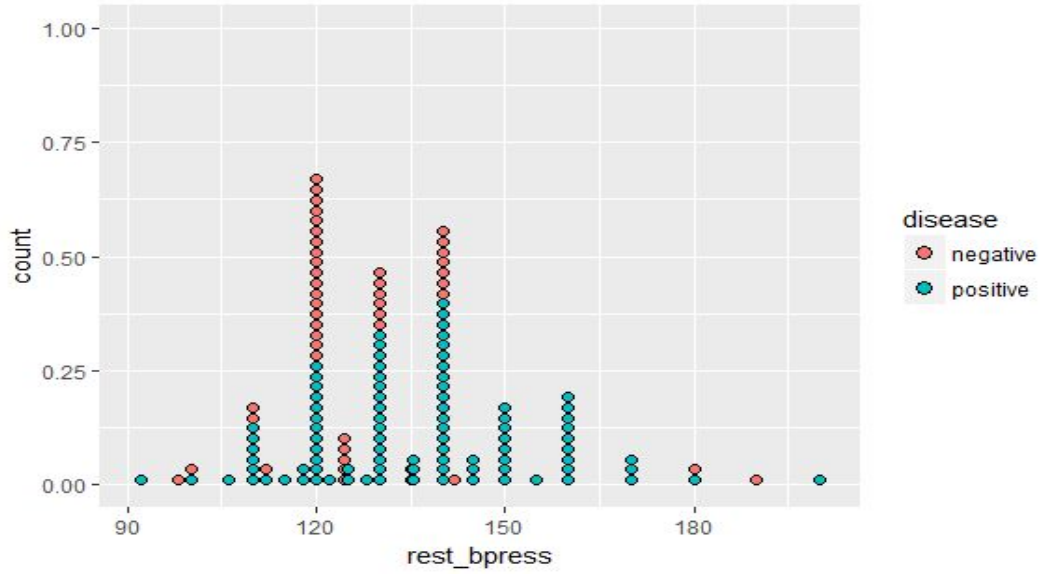
RESULT SNAPSHOTS:

1) Plotting a graph to see the occurrence of heart disease in different age groups:



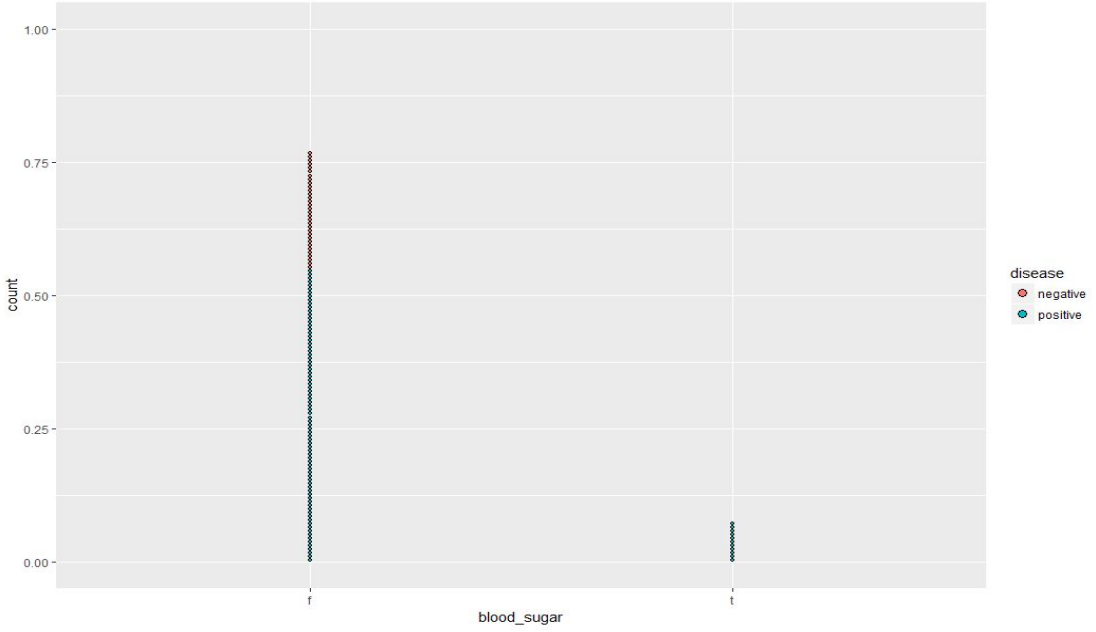
Inference:People above the age of 43 upto the age of 66 are more prone to heart disease in males than males of age group 28-42.Also,by correlation,we see that age and getting a heart disease are positively correlated(0.1870).

2)Plotting a graph to see how rest_bpress and getting a heart disease are related.



Inference: When there is resting blood pressure of 120,13 people have a heart disease and 18 have no heart disease .For blood pressure above 135 and below 120,getting a heart disease is a greater possibility.In the range of 120-135,there is a slight chance of not having a heart disease.

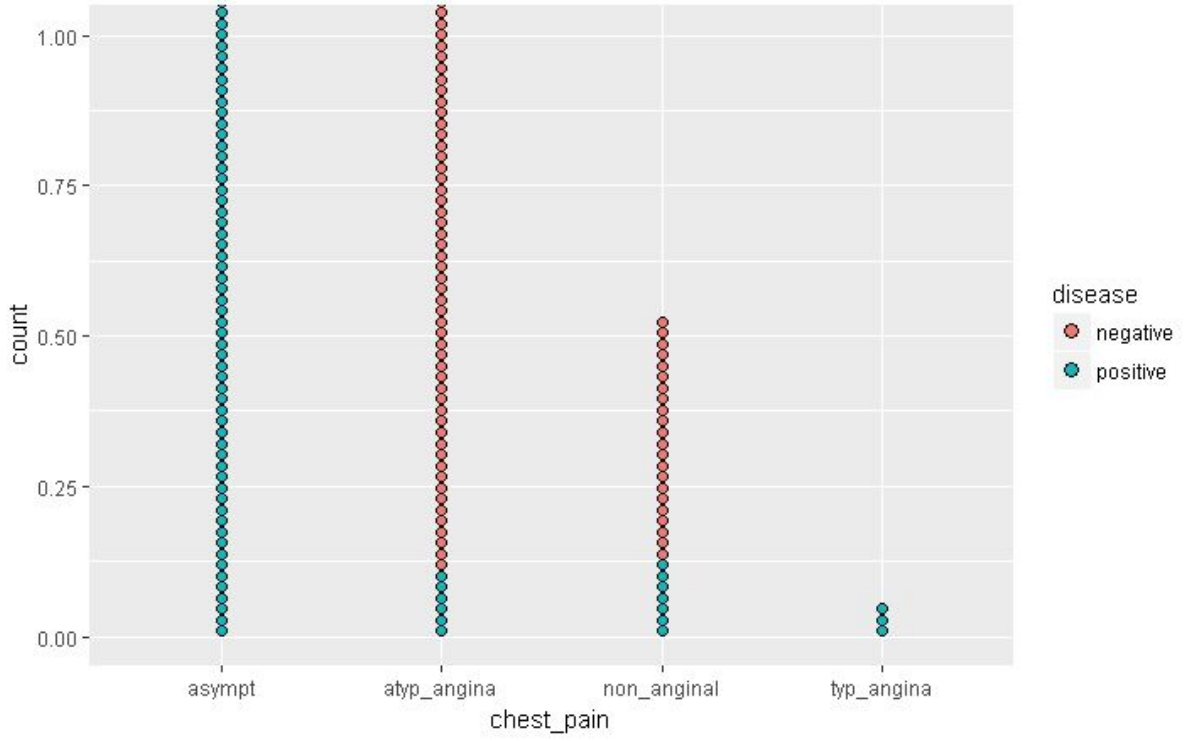
3) Plotting a graph to see if blood sugar is a factor of getting heart disease.



Inference:When a person has blood sugar,i.e according to the data set if a male has fasting blood sugar >120,then he has sugar and he will definitely get a heart disease .When there is

no blood sugar,60.9% of the people have no heart disease and 39.06% of the people have heart disease. Blood sugar and getting a heart disease are negatively correlated according to correlation obtained.Also studies tells that people with a fasting blood sugar level of 100-125 mg/dl had an adjusted nearly 300% increase higher risk of having coronary heart disease than people with a level below 79 mg/dl. This information was compiled from a cross-sectional study of nearly 2500 people.This can be one of the reasons why though the sugar level <120 and declared as false ,but still heart disease exists in those individuals.

4) Plotting a graph to see which kind of chest pain is more prominent in different age groups.



Inference: In the age below 30,everyone has atyp_angina chest pain.

Above 61,majority get asympt(almost 99%) pain.

In the age 51-60,the chances of getting asympt and non-anginal are equally probable.

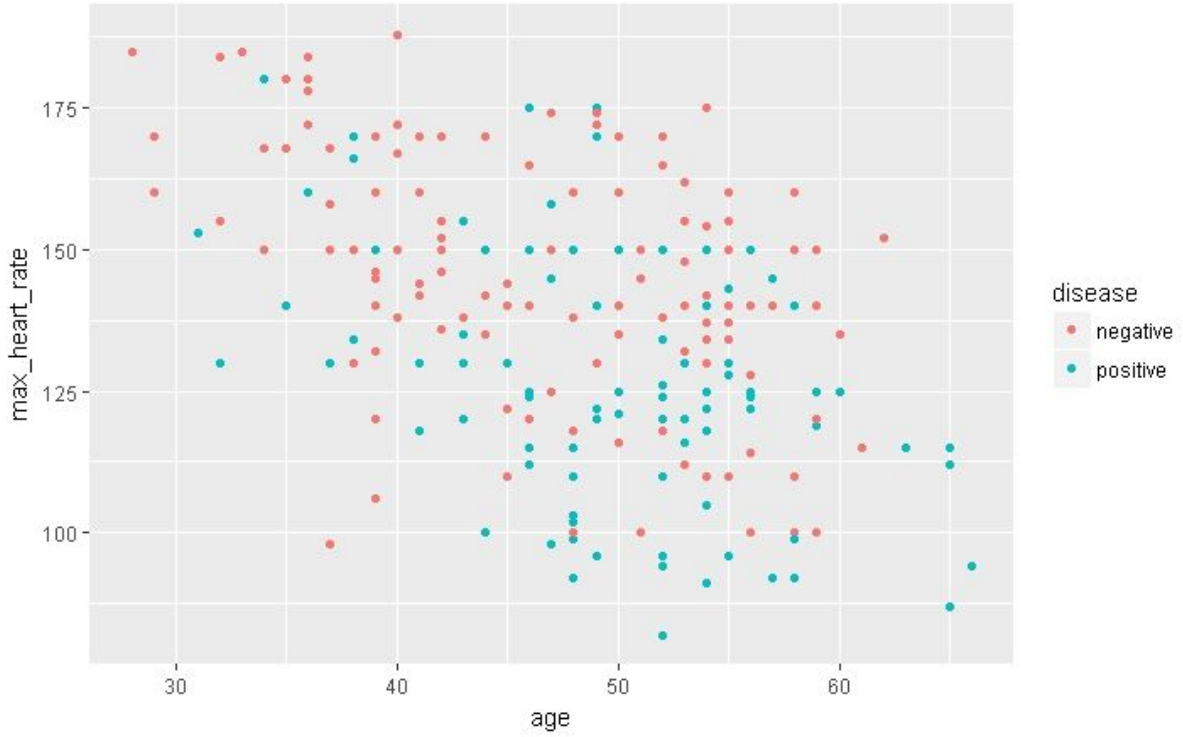
More chances of asympt pain in case of age group 41-50.

In the age group of 31-40,non_anginal chest pain is slightly prominent,next prominent one is atyp_angina followed by asympt.

Asympt pain and Typ_angina means confirms the probability of a heart disease.

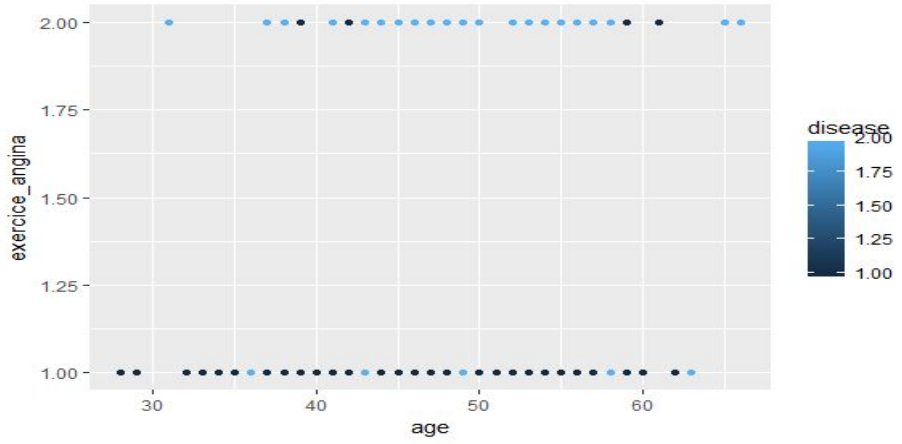
Atyp_angina and non_anginal pain means less probability of a heart disease in males.

5)Plot a graph of age v/s max_heart_rate where red dots indicate negative means no heart disease and blue dots indicate presence of heart disease.



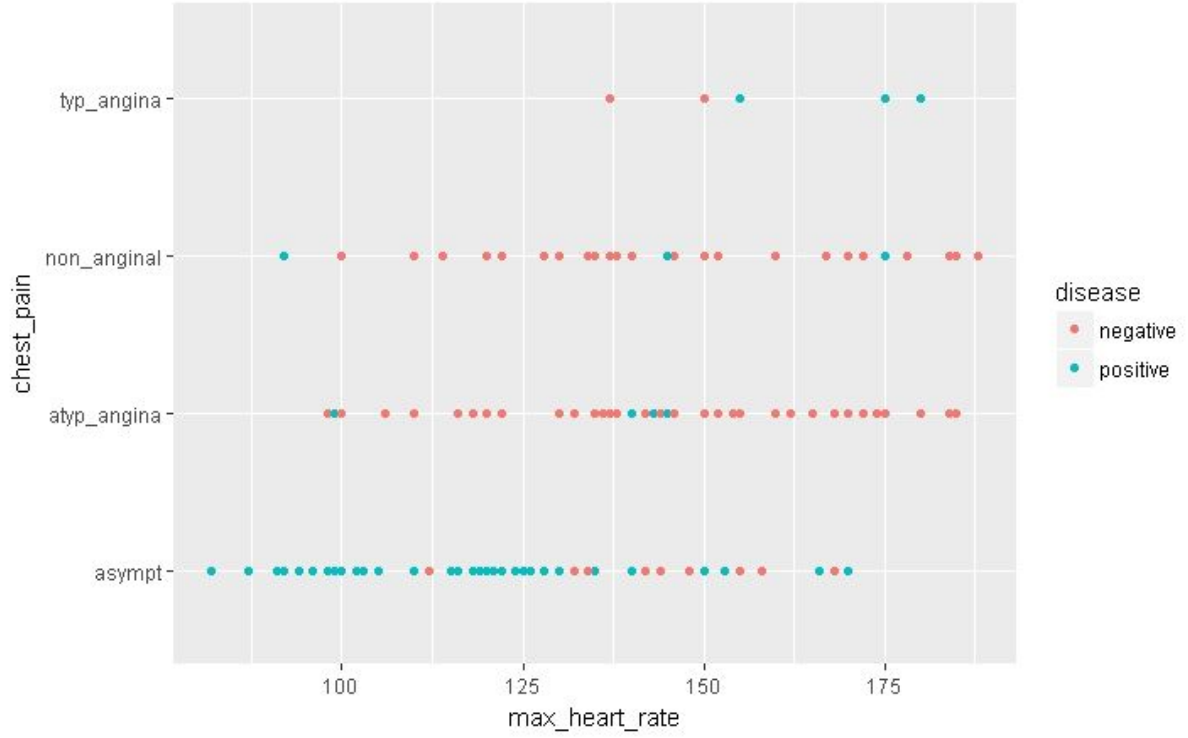
Inference:Age and max_heart_rate are negatively correlated.This can be inferred by the correlation coefficient obtained.

6)Plotting a graph to see how exercise angina and heart disease are related.



Inference:Males below the age of 40 who don't experience pain while exercising aren't prone to heart disease except a few.People belonging to the age group 40-60,if experiencing pain during exercise have disease.If these people don't experience any pain,then they don't have a heart disease.There is an equal probability of either experiencing or not experiencing pain when a male is either in range of 40-60

7)Plot of max_heart_rate v/s chest pain coloured by disease:



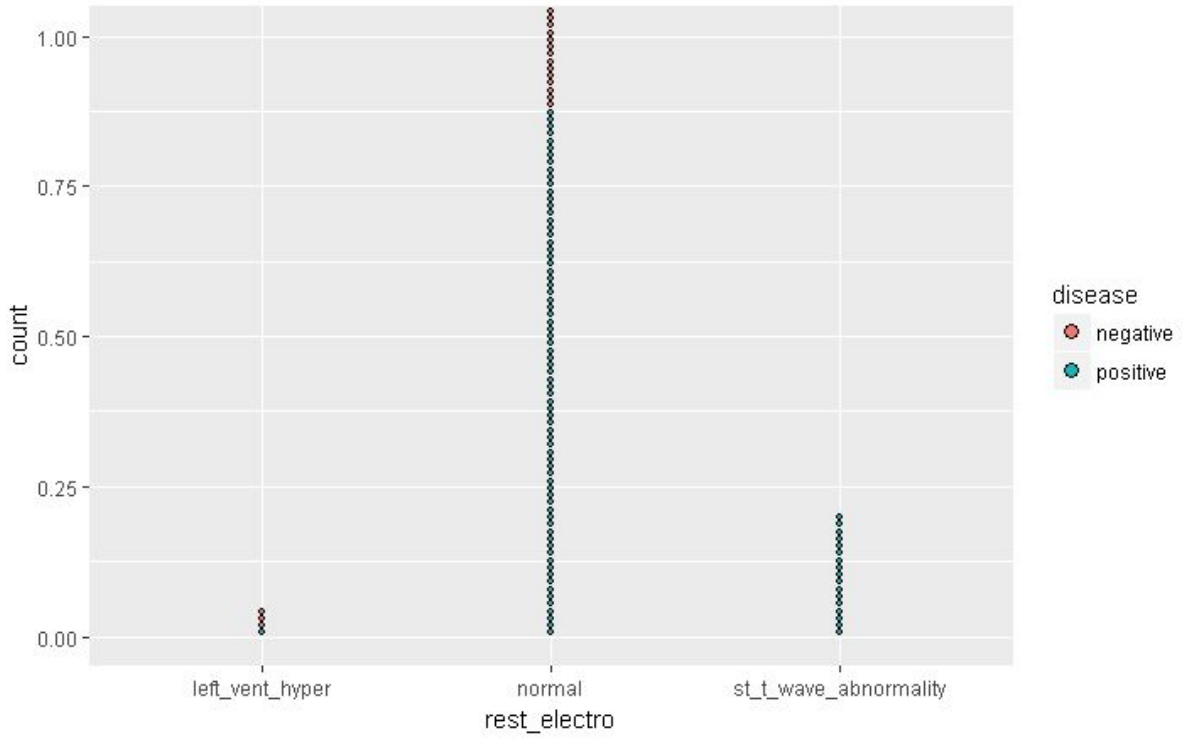
Inference:When the max-heart_rate is 100,people will experience more of asympt pain and if both these conditions satisfy ,then the person will have a heart disease.

When the max_heart rate is high i.e above 150 ,very little chances of experiencing asympt_pain and negative for heart-disease(i.e no heart disease).This also verifies the fact that heart disease and max_heart_rate are negatively correlated as shown by co-relation co-efficient(-0.360).

Asympt pain occurs mostly when the heart rate falls below 140 which is a major indicator of heart disease.

Irrespective of heart rate,whether it takes any value from100-185 atyp_anginal pain means no heart-disease (in almost 98% of the cases).

8)Plot a graph to check if ecg determines heart disease.

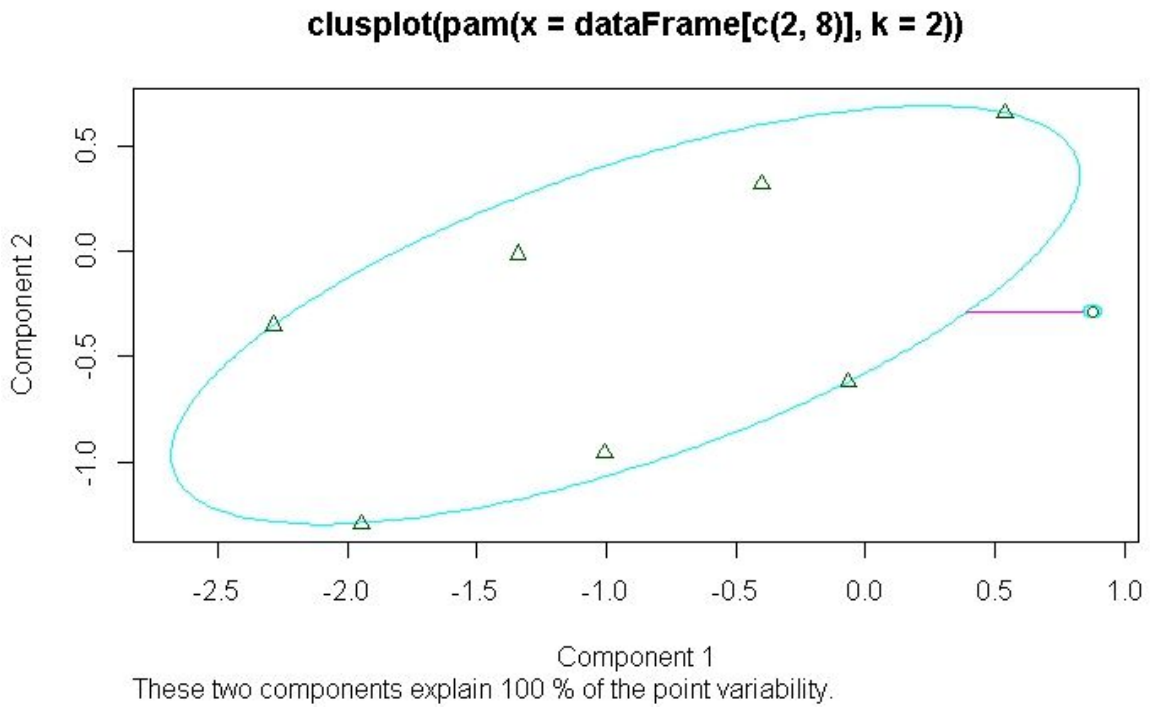


Inference: If there is any abnormality in the ecg,100% heart disease whereas 1/3rd chance of a chance of a heart disease if the ecg shows a left_vent_hyper.

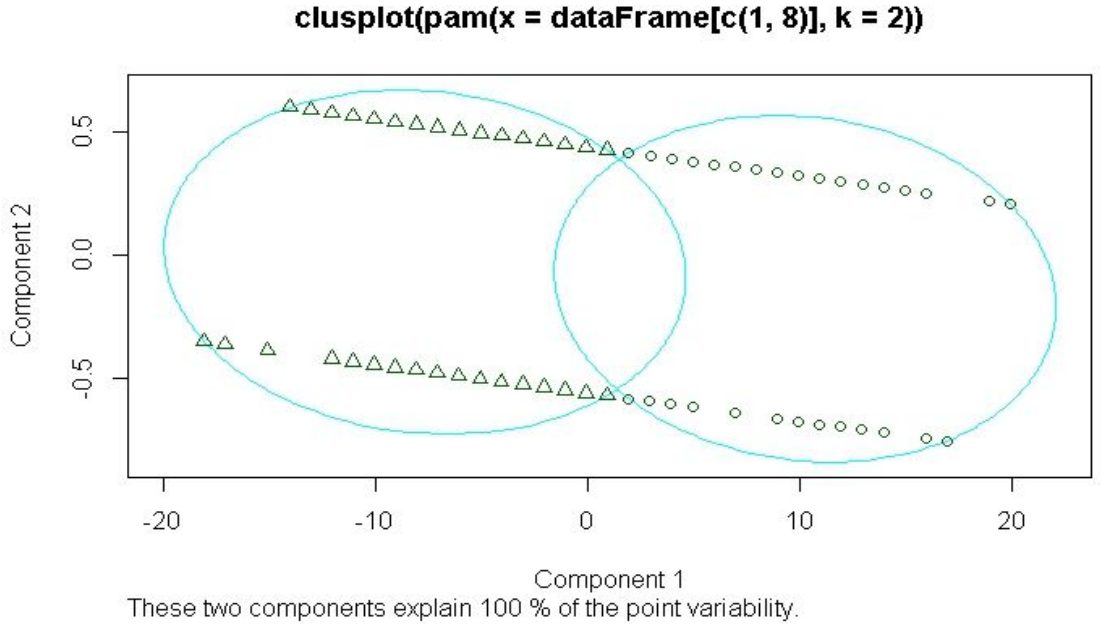
If there is a normal ecg,it doesn't mean no heart disease. A normal ecg can mean almost 60% of the times that a male has the heart disease.So we can't use the ecg attribute alone to determine the occurrence of heart disease because it can be misleading.

IMPLEMENTATION USING CLUSPLOT :

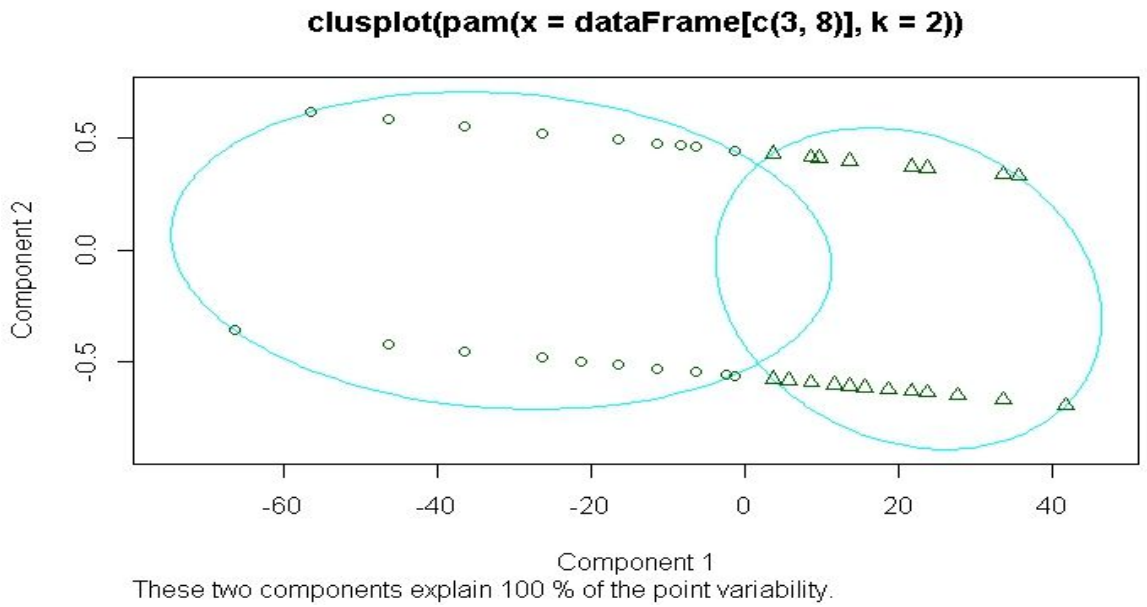
1. Clusplot between attributes chest_pain and disease.Number of clusters is chosen as 2.



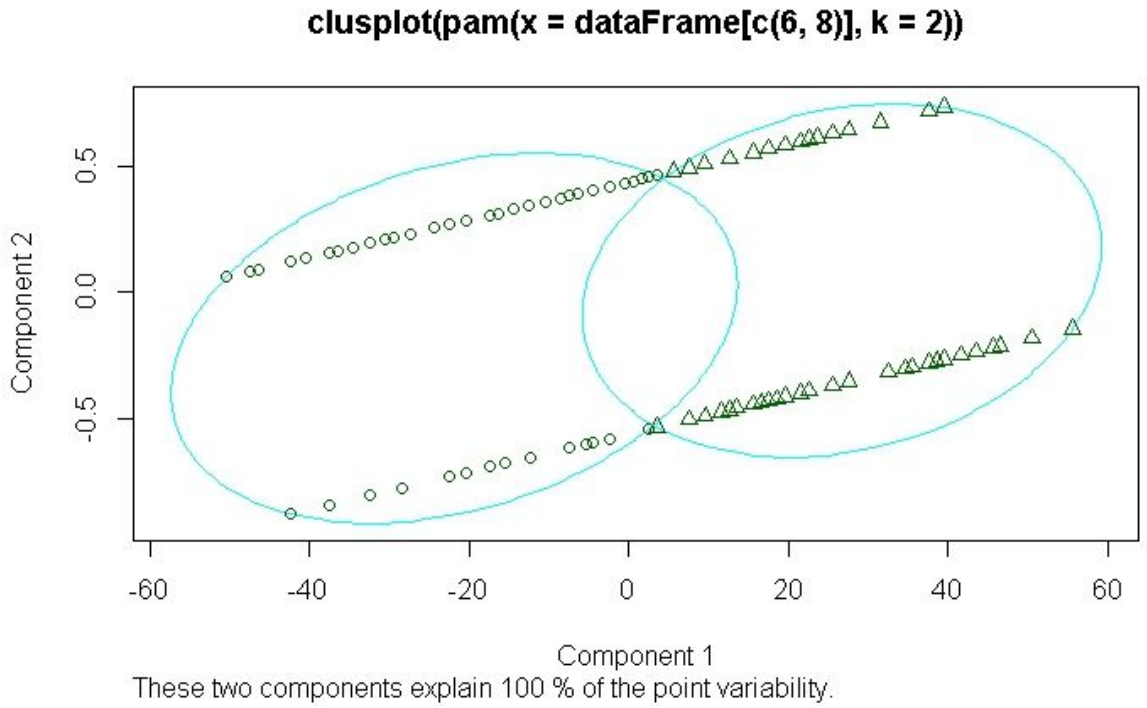
2.Clusplot between attributes age and disease.Number of clusters is chosen as 2



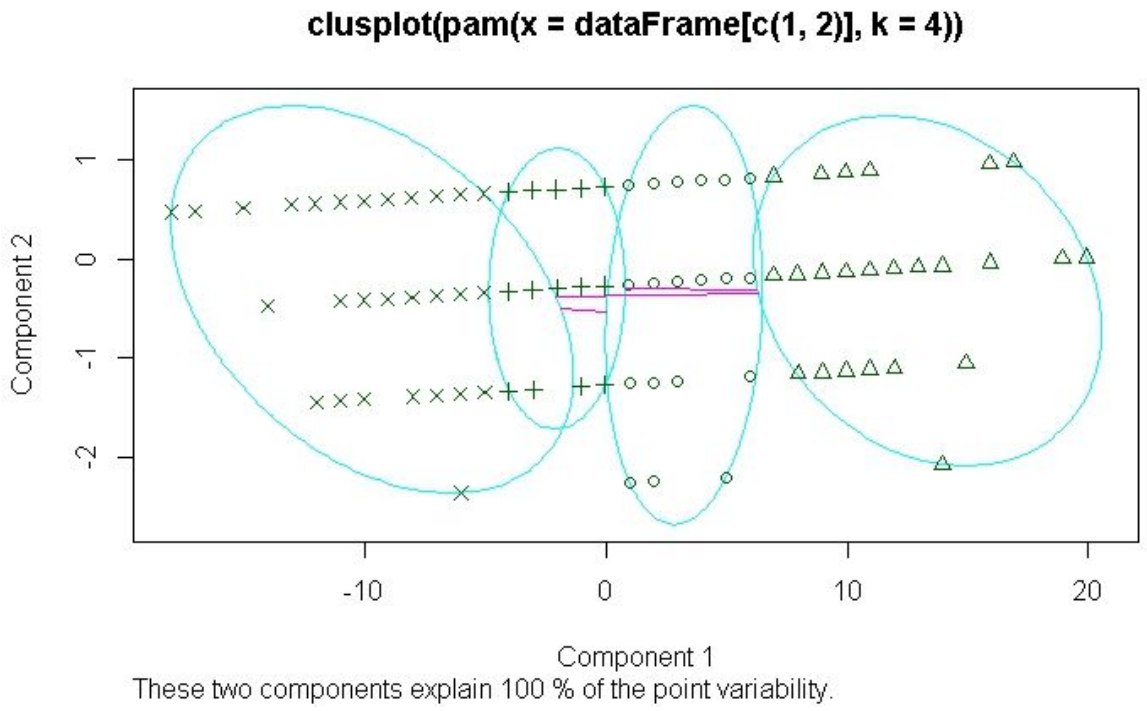
3.Clusplot between rest_bpress and disease .Clusters chosen is 2.



4.Clusplot between max_heart_rate and disease.'k' is chosen to be as 2 here.



5.Clusplot between age and chest_pain and number of clusters chosen here is 4.



IMPLEMENTATION :

The implementation of the project is on the dataset mentioned above.This is done using Rstudio.The language used to code is R.RStudio supports many inbuilt functions for plotting graphs and implementing algorithms .For k-medoids ,two algorithms have been used frequently.PAM(Partitioning Around Medoids) and CLARA(Clustering LArge Applications) are used for k-medoids .We have used PAM on our dataset.

As mentioned above,to apply PAM on our dataset,the syntax we followed is:

pam(x,k)

where x stands for the dataset with attribute column names specified in a vector form and k stands for the number of clusters required to form.

To plot graphs,the functions ggplot and clusplot have been used.The function ‘ggplot’ is present in the package ggplot2.The function ‘clusplot’ is used to draw data using PCA(Principal Component Analysis).

If the clustering algorithms like PAM are applied to a data matrix ,then a clusplot of the resulting clustering can be obtained.

The syntax for clusplot used in the implementation is:

clusplot(x,main=NULL)

Here x is an R object, i.e object of class “partition” i.e created by one of the functions PAM. main:title for the plot;when NULL (by default), a title is constructed, using x$call.

Clustering is done considering attributes as described in the snapshots and corresponding ‘clusplots’ are obtained(shown in the snapshots above).Inferences are also drawn by observing and analyzing the graphs obtained by using ‘ggplot’ function.

It is very important to find which two attributes in a dataset are positively correlated .This is implemented in our dataset by using ‘cor’ function .This gives the correlation between two attributes for our dataset which is a crucial part for our analysis.A few instances of the correlation between the attributes have been mentioned in the previous sections.
Also,to change categorical attributes to numerical attributes ,we have made use of ‘sapply’ function.
To implement the code,we have imported three libraries:

- xlsx(to read excel file)
- ggplot2(to make use of functionalities like ggplot)
- cluster (for implementing PAM)

SOCIAL IMPACT :

Heart disease is the No. 1 cause of death in the world and the leading cause of death in the United States. Half of the men who die suddenly of coronary heart disease have no previous symptoms.3 Even if you have no symptoms, you may still be at risk for heart disease.
Between 70% and 89% of sudden cardiac events occur in men.So using this analysis on our dataset ,we can predict from the medical reports a person has.This will allow a person to be more cautious and aware of his heart conditions.

In our society,there are a few myths about heart disease.

1.People usually get mislead by the fact that only old people are at a risk of getting a heart disease One in three Americans has cardiovascular disease, but not all of them are senior citizens. Since type 2 diabetes ,obesity and other risk factors are more evident in younger

generations also,we can see that being young doesn’t mean one won’t end up with a heart disease

2.People think only if they experience chest pain,they have a heart attack.This is wrong. From our analysis,we have seen that having a chest pain doesn’t always mean a heart attack and we have seen the different kinds of chest pain and people should not jump to conclusions when there is a chest pain.They should first find out what kind of chest pain are they experiencing.

3.People sometimes feel that if the heart rate increases ,he is going to experience a heart attack.This is another wrong idea or belief.Some variation in heart rate is normal like heart rate increase while exercising and decreases while sleeping.Sometimes,younger people have a higher heart rate and have no signs of heart disease.This is even inferred from our analysis.This myth has been proved wrong by our analysis performed on the dataset.

The social impact would be removing the myths people have in this field of heart disease.People will be more convinced to believe that these myths are wrong when they see the analysis results .This will also reduce the sudden deaths which happen due to cardiac arrest.To do so,medical reports should be clearly analysed .

Concluding the points made so far,to reduce the number of deaths due to heart diseases,this study will be of great importance and the inferences drawn will be helpful to determine if a person has a heart disease or not by considering the attributes mentioned in the dataset.