Heart Disease

Choudetsanakis Marios

6/2/2020

Introduction

Heart disease is the leading cause of death in the United States.The term "heart dissease" refers to several types of heart conditions.The most common type is coronary artery disease,which can cause heart attack.The symptoms vary depending on the type of heart disease.For many people,chest discomfortor a heart attack is the first sign.

The data set that will be used is the Heart Disease Data Set from the Machine Learning Repository(https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/) and especially the "processed cleveland data".This data set contains data from 303 individuals with 14 attributes.The creator of this data set is Robert Detrano, M.D., Ph.D. from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

- Goal of the project

The original goal of the data set was to predict the presence of heart disease which is the variable num of the set.In this project,the goal is the prediction of the sex of the individual given the other 13 variables including the variable for the heart disease presence.There is going to be used three methods for the analysis: the linear regression,the k-nearest neighbors and a decision tree.The results of each model will be saved in a data frame and in the end, it is going to be a evaluation of all models.

Description of the variables

- Age: the age of the individual
- Sex: the gender of the individual
- Cp : chest pain type (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- Trestbps : resting blood pressure (in mm Hg on admission to the hospital)
- Chol : serum cholestoral in mg/dl
- fbs : fasting blood sugar $>$ 120 mg/dl(1 = yes; 0 = no)
- restecg : resting electrocardiographic results (Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
- 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 (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- ca : number of major vessels (0-3) colored by flourosopy
- thal : 3 = normal, 6 = fixed, defect , 7 = reversable defect
- num : diagnosis of heart disease(from 0 (no presence) to 4)

Data importing and cleaning

```
#download file from the 'https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/'
#we use the file.choose() function to read the file from your pc
#the file is the processed.cleveland.data
my_data=read.delim(file.choose(),header = FALSE)
#a vector with the header of the table
attr=c("Age","Sex","Cp","Trestbps","Chol","fbs","restecg","thalach","exang",
       "oldpeak","slope","ca","thal","num")
#import the header to the table
my_data=separate(my_data,col = 1,sep = ",", remove = TRUE,
                 convert = FALSE, extra = "warn", fill = "warn",into = attr)

#convert the variables to numeric from character
my_data=my_data%>%mutate(Age=as.numeric(Age),
                 Sex=as.numeric(Sex),
                 Cp=as.numeric(Cp),
                 Trestbps=as.numeric(Trestbps),
                 Chol=as.numeric(Chol),
                 fbs=as.numeric(fbs),
                 restecg=as.numeric(restecg),
                 thalach=as.numeric(thalach),
                 exang=as.numeric(exang),
                 oldpeak=as.numeric(oldpeak),
                 slope=as.numeric(slope),
                 ca=as.numeric(ca),
                 thal=as.numeric(thal),
                 num=as.numeric(num))
#removing the na's and replaciong them with zeros
my_data=my_data%>%mutate(thal=ifelse(is.na(thal),0,thal),
                         num=ifelse(is.na(num),0,num),
                         ca=ifelse(is.na(ca),0,ca))
```

Data exploration and visualization

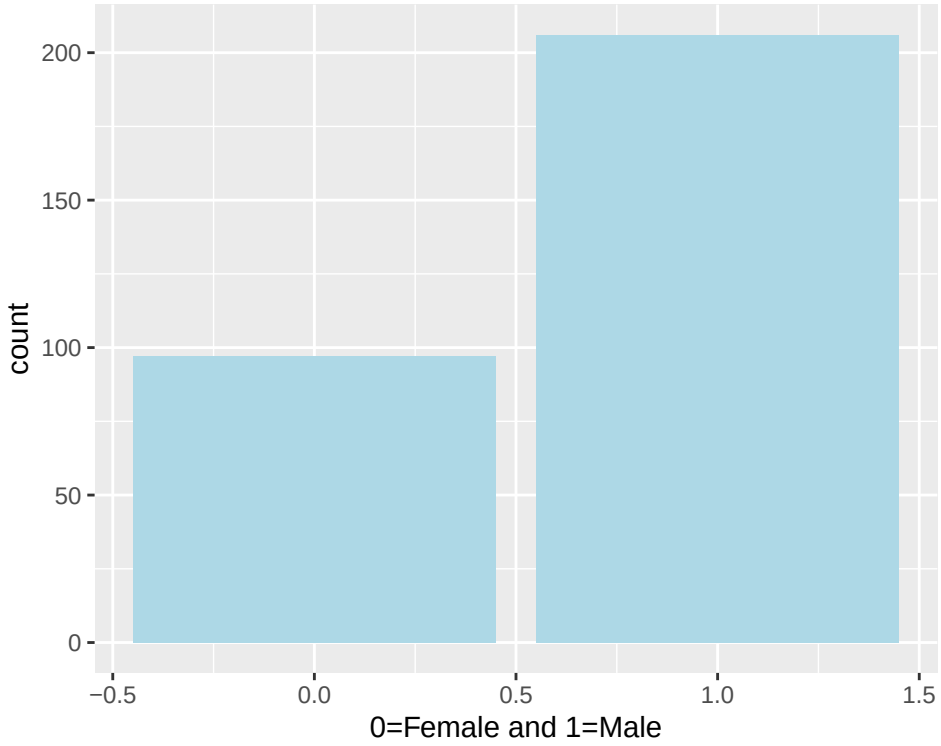
- data dimensions

```
dim(my_data)
```

```
## [1] 303  14
```

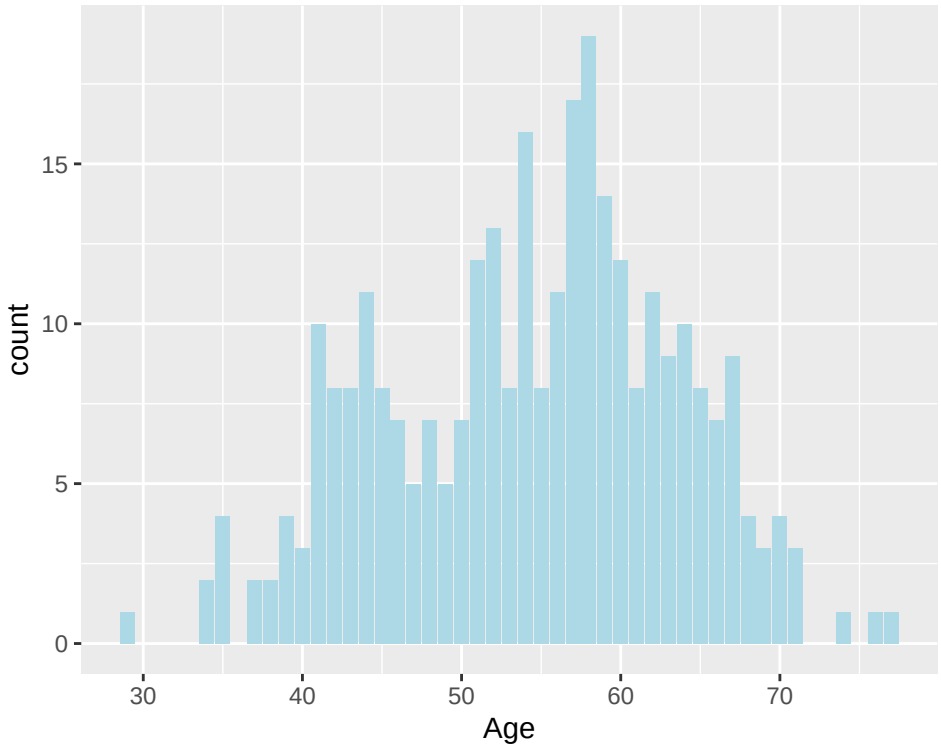
- the number of males and females

```
my_data%>%ggplot(aes(Sex))+geom_bar(fill="lightblue")+xlab("0=Female and 1=Male")
```



- age

```
my_data%>%ggplot(aes(Age))+geom_bar(fill="lightblue")
```



```r
max(my_data$Age)
```

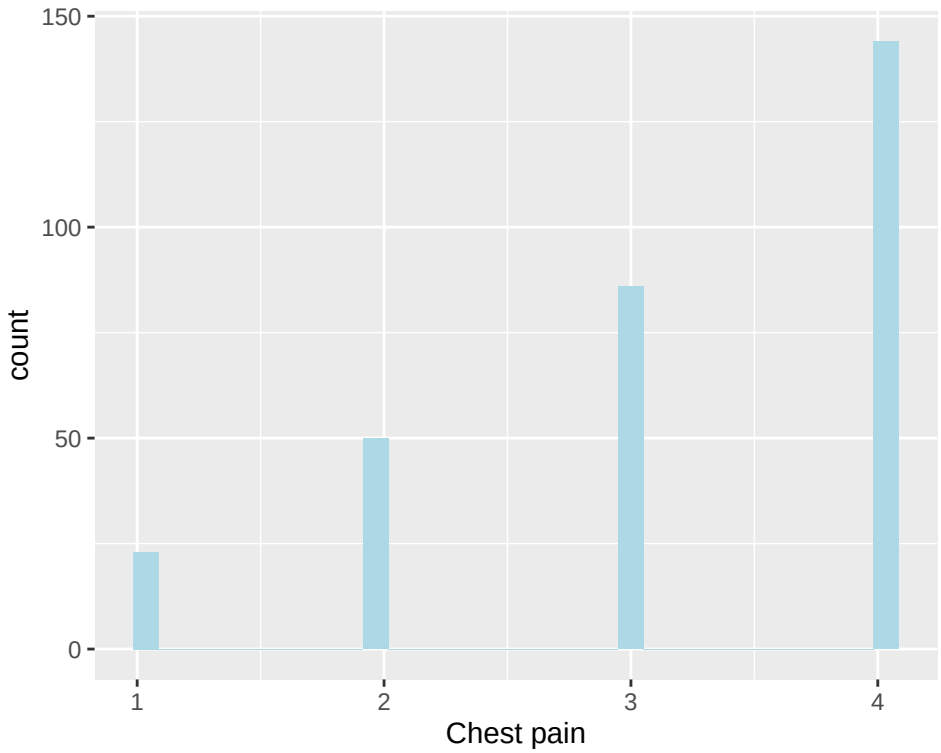
```
## [1] 77
```

```r
min(my_data$Age)
```

```
## [1] 29
```

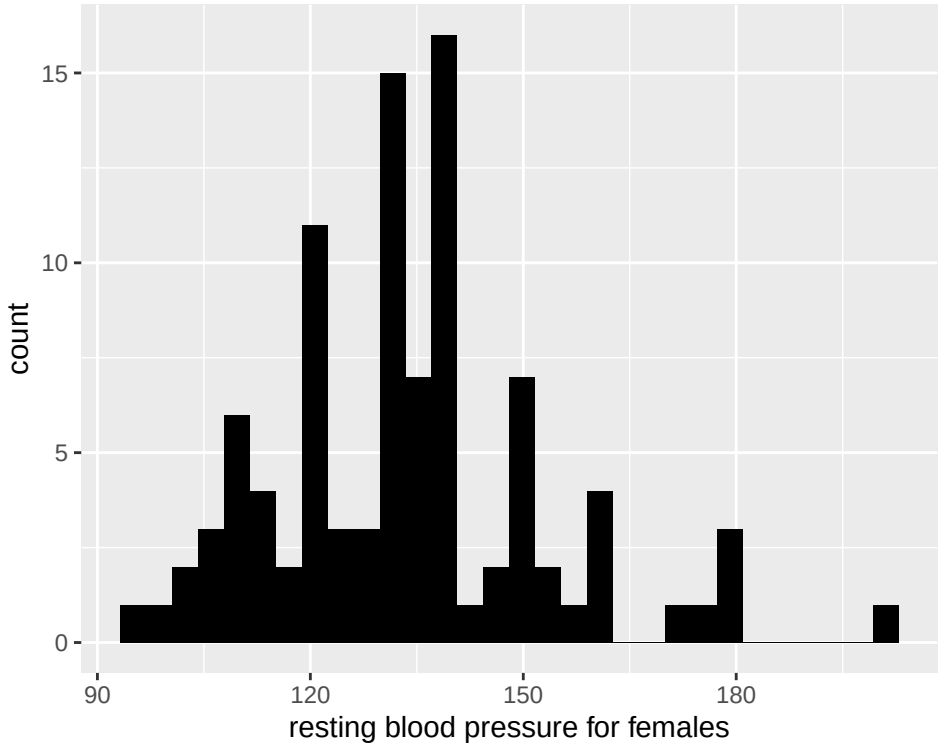
- chest pain

```r
my_data%>%ggplot(aes(Cp))+geom_histogram(fill="lightblue")+xlab("Chest pain")
```

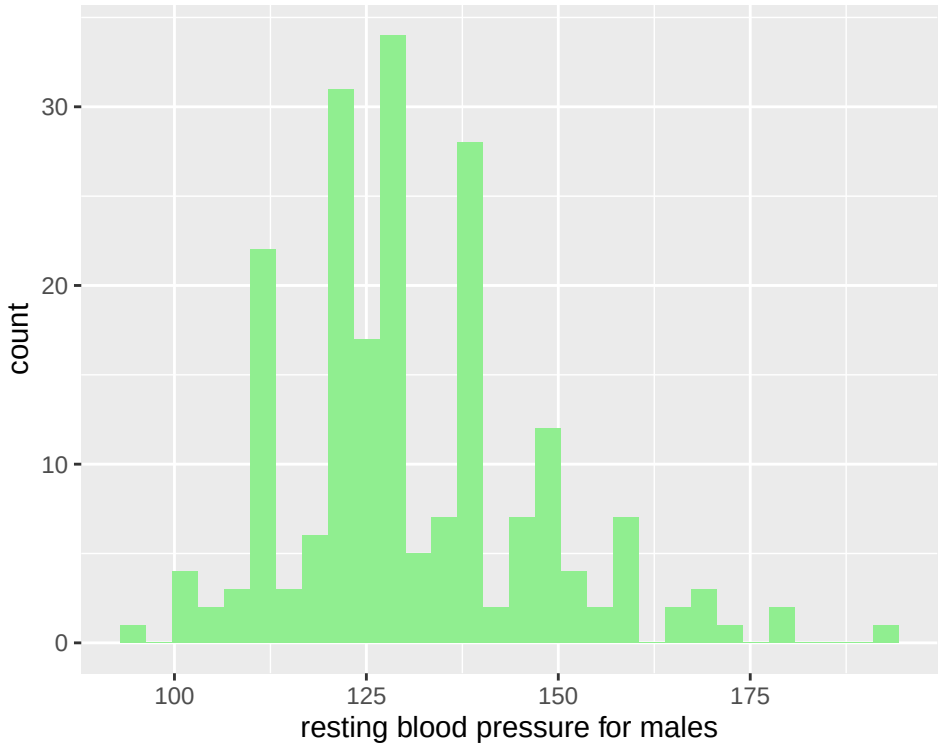


- resting blood pressure

```r
my_data%>%filter(Sex==0)%>%ggplot(aes(Trestbps))+geom_histogram(fill="black")+xlab("resting blood pressu
```

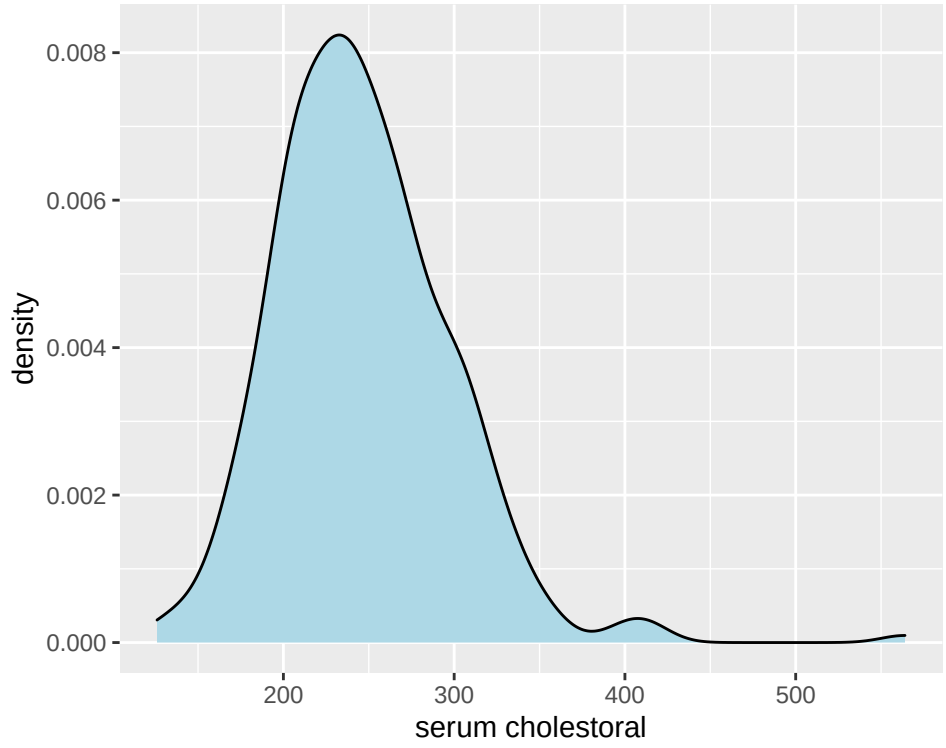



```
my_data%>%filter(Sex==1)%>%ggplot(aes(Trestbps))+geom_histogram(fill="lightgreen")+xlab("resting blood p
```

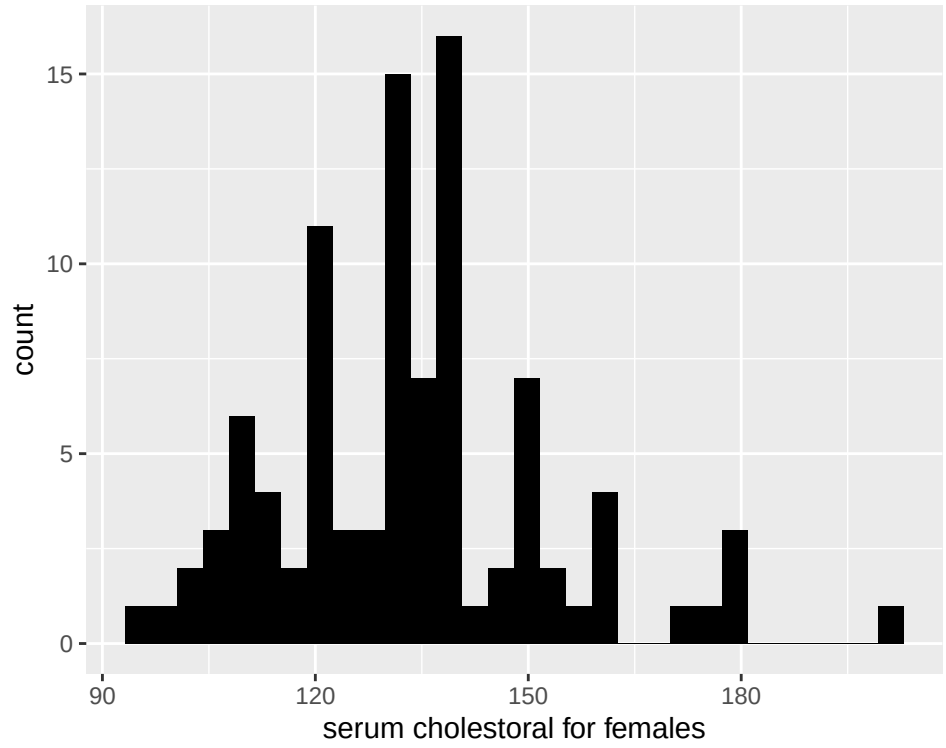


- serum cholestoral in mg/dl

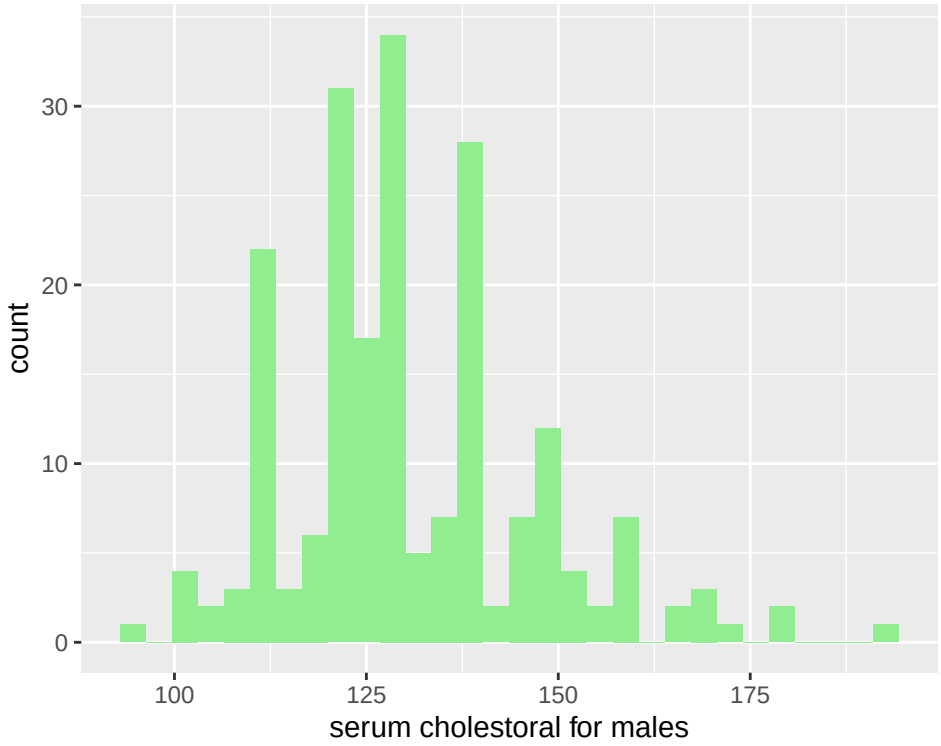
```
my_data%>%ggplot(aes(Chol))+geom_density(fill="lightblue")+xlab("serum cholestoral")
```




```
my_data%>%filter(Sex==0)%>%ggplot(aes(Trestbps))+geom_histogram(fill="black")+xlab("serum cholestoral fo
```

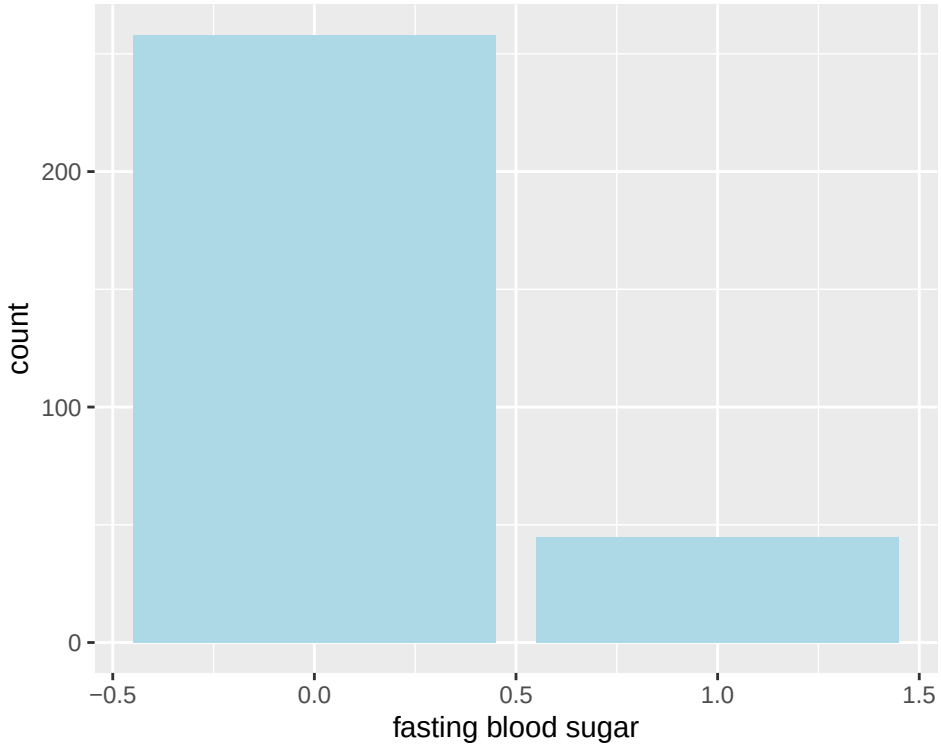


```
my_data%>%filter(Sex==1)%>%ggplot(aes(Trestbps))+geom_histogram(fill="lightgreen")+xlab("serum cholesto
```



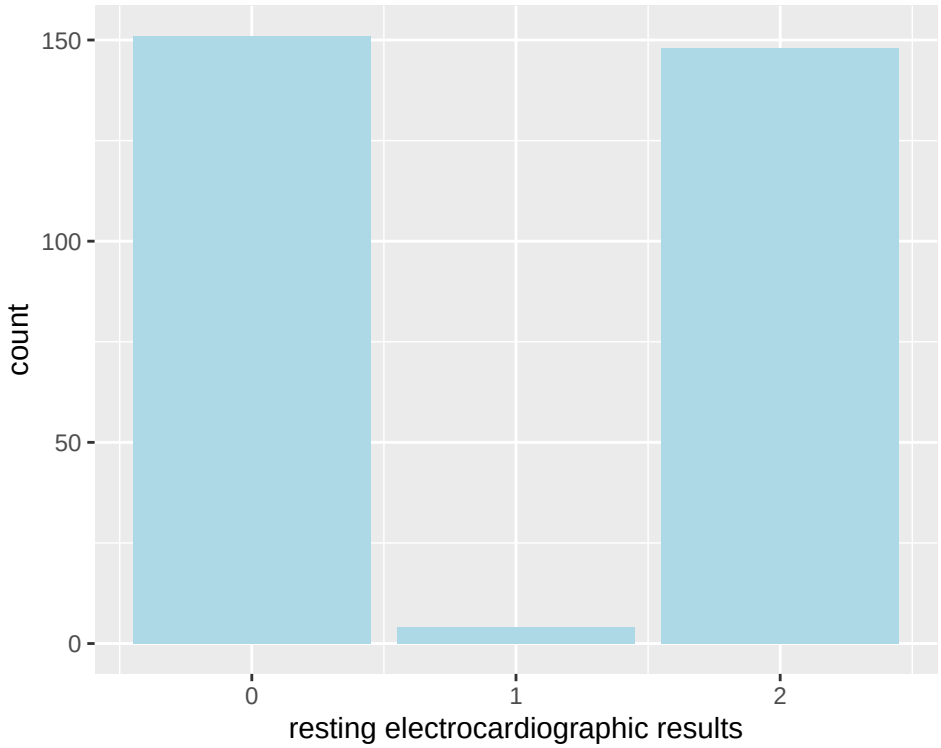
- fasting blood sugar > 120 mg/dl($1 =$ yes, $0 =$ no)

```
my_data%>%ggplot(aes(fbs))+geom_bar(fill="lightblue")+xlab("fasting blood sugar")
```



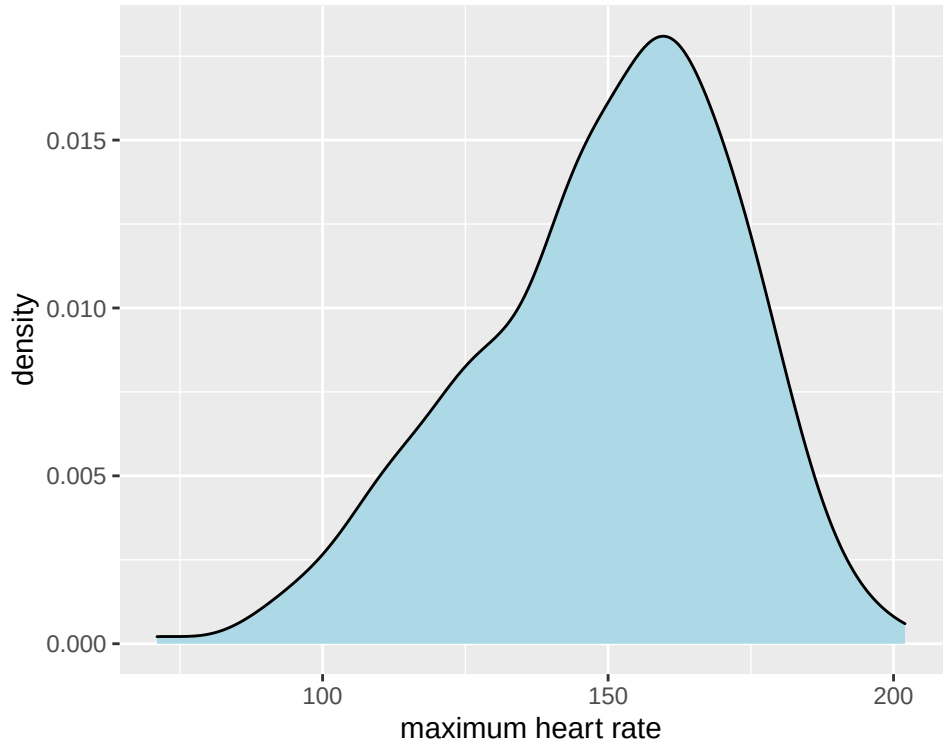
- resting electrocardiographic results

```
my_data%>%ggplot(aes(restecg))+geom_bar(fill="lightblue")+xlab("resting electrocardiographic results")
```



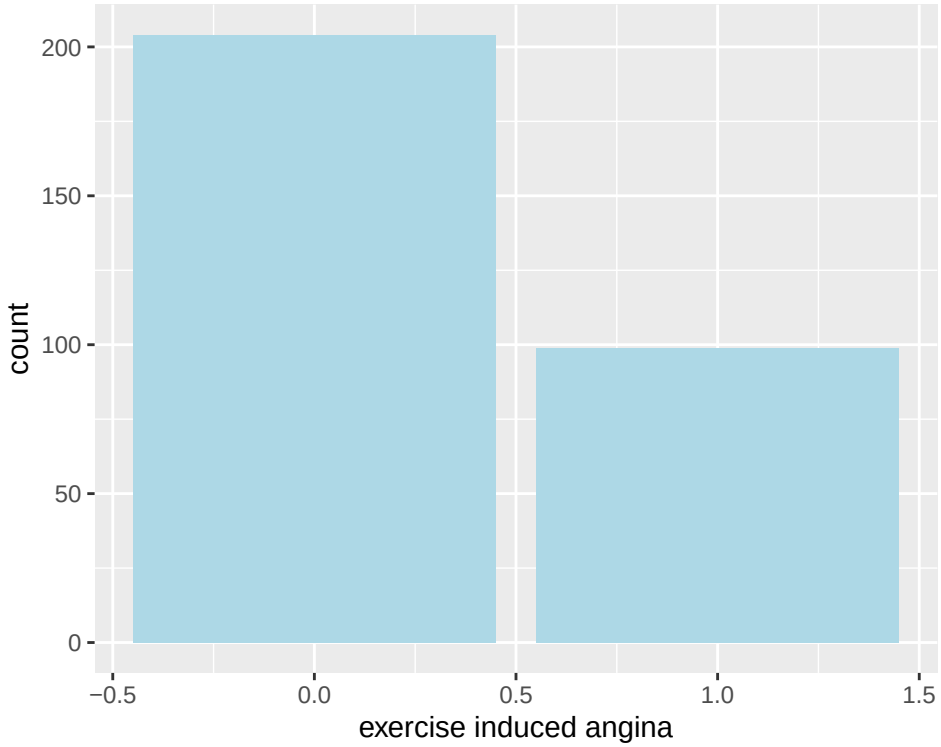
- maximum heart rate achieved

```
my_data%>%ggplot(aes(thalach))+geom_density(fill="lightblue")+xlab("maximum heart rate")
```



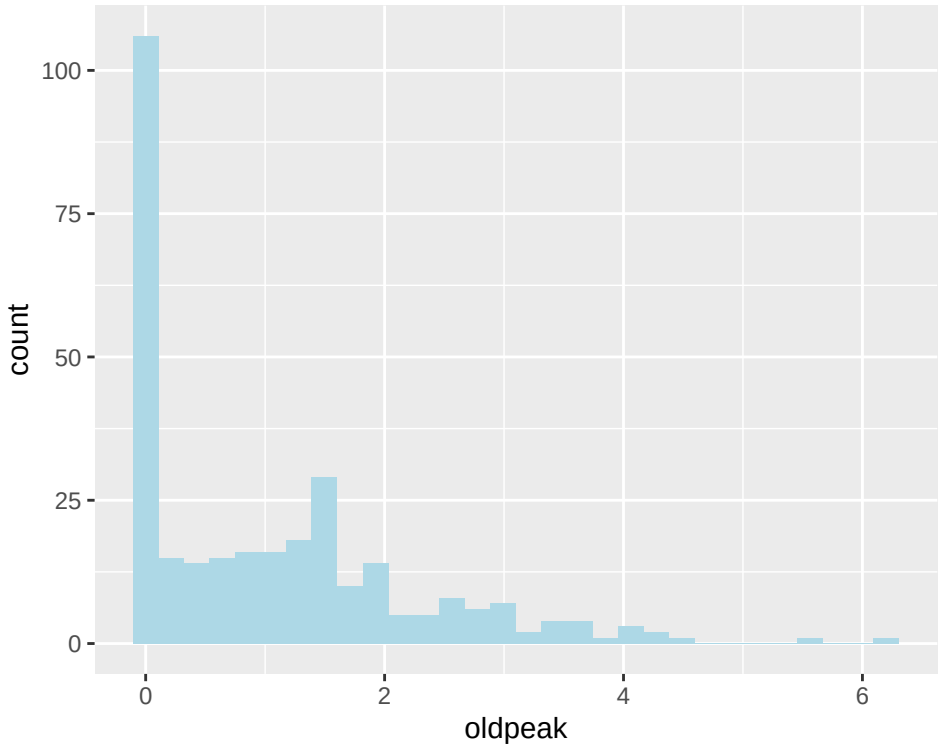
- exercise induced angina

```
my_data%>%ggplot(aes(exang))+geom_bar(fill="lightblue")+xlab("exercise induced angina")
```



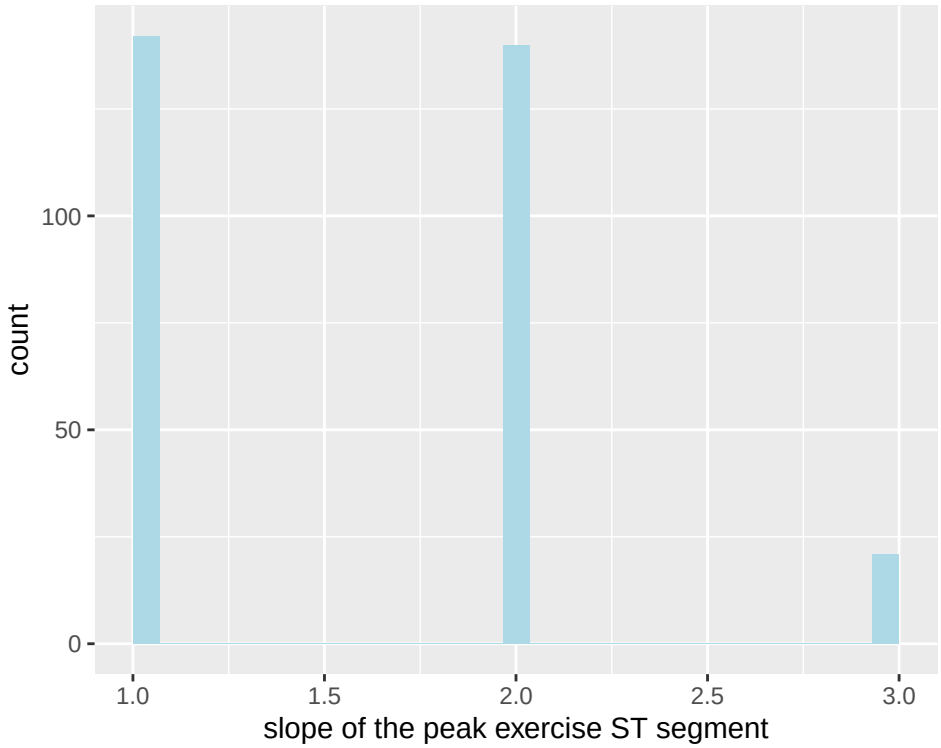
- ST depression induced by exercise relative to rest

```
my_data%>%ggplot(aes(oldpeak))+geom_histogram(fill="lightblue")
```



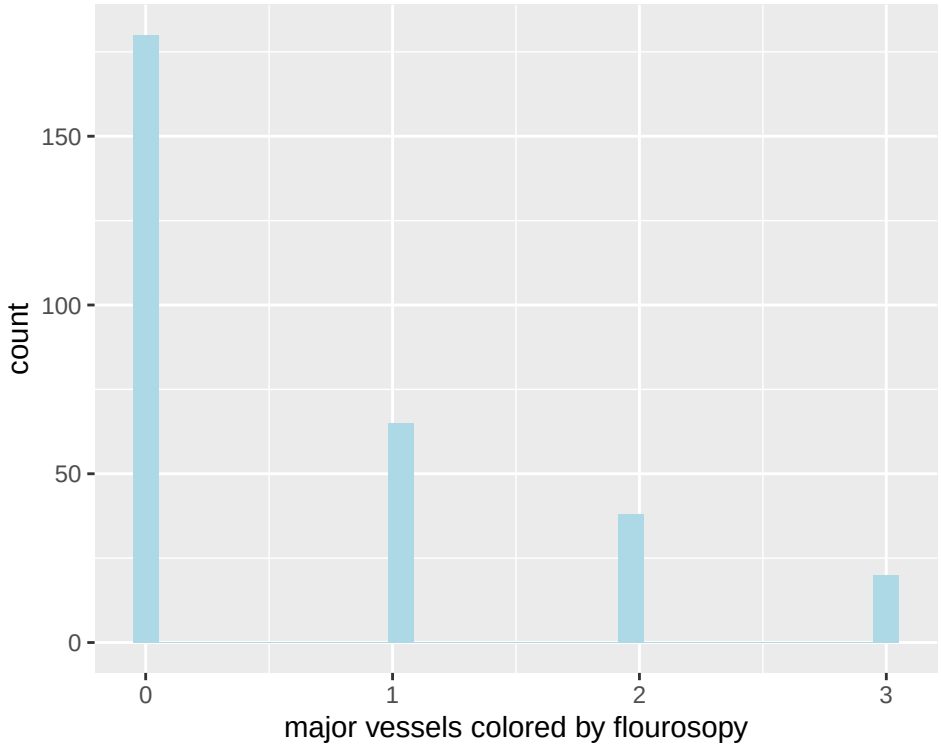
- slope of the peak exercise ST segment

```
my_data%>%ggplot(aes(slope))+geom_histogram(fill="lightblue")+xlab("slope of the peak exercise ST segme
```



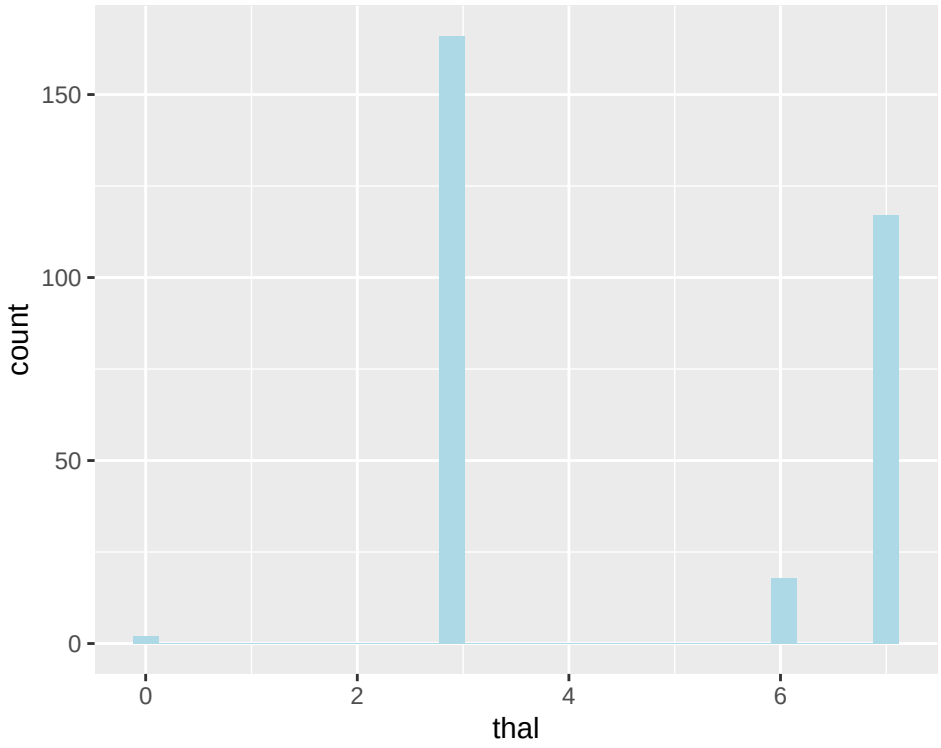
- number of major vessels (0-3) colored by flourosopy

```
my_data%>%ggplot(aes(ca))+geom_histogram(fill="lightblue")+xlab("major vessels colored by flourosopy")
```



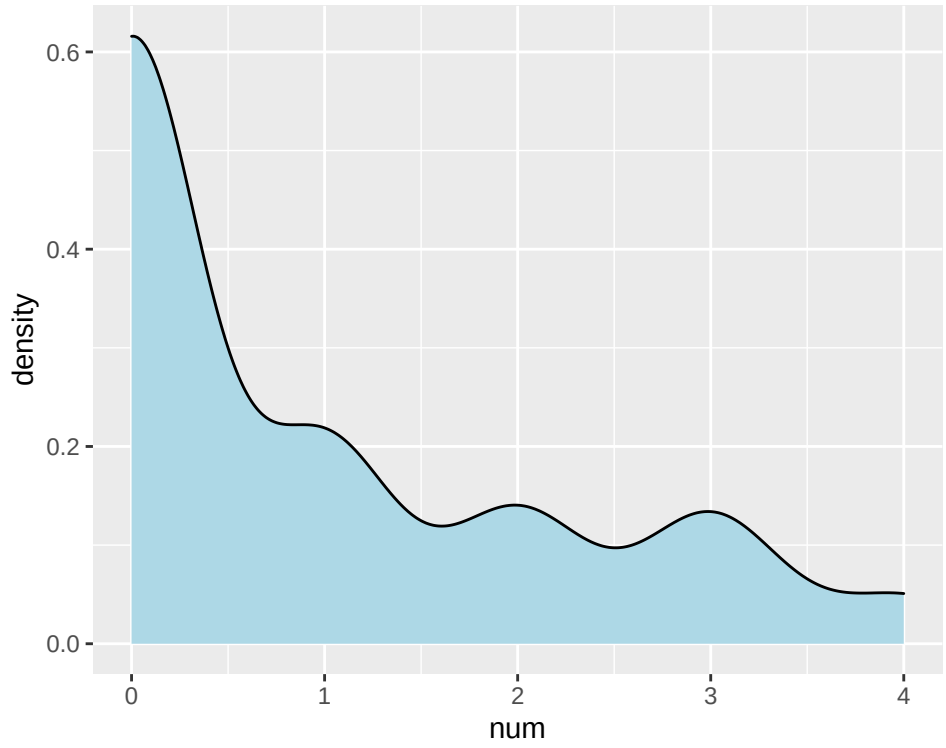
- thal: 3 = normal , 6 = fixed defect , 7 = reversable defect

```
my_data%>%ggplot(aes(thal))+geom_histogram(fill="lightblue")
```



- diagnosis of heart disease (angiographic disease status)

```
my_data%>%ggplot(aes(num))+geom_density(fill="lightblue")
```



Relationship between the variables

Here we are going to see the relationship between the Sex and the other variables.We will use the chisq.test function with null hypothesis that there is no relationship between the other variables in a significance level of 0.05%.

- Sex vs Age

```
chisq.test(my_data$Sex,my_data$Age)$p.value
```

```
## [1] 0.3716642
```

- Sex vs chest pain

```
chisq.test(my_data$Sex,my_data$Cp)$p.value
```

```
## [1] 0.07885024
```

- Sex vs resting blood sugar

```r
chisq.test(my_data$Sex,my_data$Trestbps)$p.value
```

```
## [1] 0.7491636
```

- Sex vs serum cholestoral

```r
chisq.test(my_data$Sex,my_data$Chol)$p.value
```

```
## [1] 0.2621938
```

- Sex vs fasting blood sugar

```r
chisq.test(my_data$Sex,my_data$fbs)$p.value
```

```
## [1] 0.5092599
```

- sex vs resting electrocardiographic results

```r
chisq.test(my_data$Sex,my_data$restecg)$p.value
```

```
## [1] 0.1665345
```

- Sex vs maximum heart rate achieved

```r
chisq.test(my_data$Sex,my_data$thalach)$p.value
```

```
## [1] 0.0362739
```

- Sex vs ST depression induced by exercise relative to rest

```r
chisq.test(my_data$Sex,my_data$oldpeak)$p.value
```

```
## [1] 0.7254774
```

- Sex vs the slope of the peak exercise ST segment

```r
chisq.test(my_data$Sex,my_data$slope)$p.value
```

```
## [1] 0.6951351
```

- Sex vs number of major vessels (0-3) colored by flourosopy

```r
chisq.test(my_data$Sex,my_data$ca)$p.value
```

```
## [1] 0.1587968
```

- Sex vs thal

```r
chisq.test(my_data$Sex,my_data$thal)$p.value
```

```
## [1] 6.091214e-10
```

- Sex vs presence of heart disease

```r
chisq.test(my_data$Sex,my_data$num)$p.value
```

```
## [1] 0.0001041059
```

- Results

As we see the variables: Age,Chest pain(Cp),resting blood sugar(Trestbps),serum cholestoral(Chol),fasting blood sugar(fbs),resting electrocardiographic results(restecg),ST depression induced by exercise relative to rest(oldpeak),the slope of the peak exercise ST segment(slope),number of major vessels (0-3) colored by flourosopy(ca) have a p-value bigger than 0.05 so we have not to reject the null hypothesis ,which is that there is no relationship between the variables in a significance level of 0.05%.On the other hand the other variables maximum heart rate achieved (thalach),thal and presence of heart disease(num) have a p-value smaller than 0.05 so we have to reject the null hypothesis in a significance level of 0.05%.

Analysis

- Data spliting

The data will be splited in two sets,the train set(80% of my_data) and the test set(20% of my_data).The function createDataPartition from caret package will make the data spliting.

```r
set.seed(3,sample.kind="Rounding")
test_index=createDataPartition(my_data$Age,times=1,p=0.2,list=FALSE)
test_set=my_data[test_index,]
train_set=my_data[-test_index,]
```

Now we see the dimenions of these two datasets

```r
dim(train_set)
```

```
## [1] 241  14
```

```r
dim(test_set)
```

```
## [1] 62 14
```

- Linear Regression Model

The first model will be a multivariate linear regression.The outcome of this linear regression is a categorical variable which fits in our problem because we want to predict the sex of the individual which is either male or female.It will be used the function round to round the y_hat _lm to the nearest interger , either 0 or 1.

```r
#fitting the model
model_lm=lm(Sex ~ .,data=train_set)
#prediction
y_hat_lm=predict(model_lm,test_set)
#confusion matrix and round the y_hat lm
cm_lm=confusionMatrix(data=factor(round(y_hat_lm)),reference=factor(test_set$Sex))
cm_lm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0  9  3
##          1 13 37
##
##                Accuracy : 0.7419
##                  95% CI : (0.615, 0.8447)
##     No Information Rate : 0.6452
##     P-Value [Acc > NIR] : 0.06970
##
##                   Kappa : 0.3722
##
##  Mcnemar's Test P-Value : 0.02445
##
##             Sensitivity : 0.4091
##             Specificity : 0.9250
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.7400
##              Prevalence : 0.3548
##          Detection Rate : 0.1452
##    Detection Prevalence : 0.1935
##       Balanced Accuracy : 0.6670
##
##        'Positive' Class : 0
##
```

```r
results[1,]%>%knitr::kable()
```

method	Accuracy	Sensitivity	Specificity
Linear Regression	0.7419355	0.4090909	0.925

The overall accuracy is almost 74.2%.Sensitivity is 40% specificity is 92.5% which are both good.

- K-nearest neighbors

Now we build a kNN model with k=5.K-nn model can be used on classes which fits in our problem because we want to determine the sex of an individual which is either male or female.

```r
#fitting the model
model_knn=knn3(Sex~.,data=train_set,k=5)
#prediction
y_hat_knn=predict(model_knn,test_set)
#store predictions as 0 or 1 based on the distance that have with #them
p_hat_knn=ifelse(y_hat_knn[,1]>0.5,0,1)
#confusion matrix
cm_knn=confusionMatrix(data=factor(p_hat_knn),reference=factor(test_set$Sex))
cm_knn
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0  5  4
##          1 17 36
##
##                Accuracy : 0.6613
##                  95% CI : (0.5299, 0.7767)
##     No Information Rate : 0.6452
##     P-Value [Acc > NIR] : 0.452323
##
##                   Kappa : 0.1468
##
##  Mcnemar's Test P-Value : 0.008829
##
##             Sensitivity : 0.22727
##             Specificity : 0.90000
##          Pos Pred Value : 0.55556
##          Neg Pred Value : 0.67925
##              Prevalence : 0.35484
##          Detection Rate : 0.08065
##    Detection Prevalence : 0.14516
##       Balanced Accuracy : 0.56364
##
##        'Positive' Class : 0
##
```

```r
results[2,]%>%knitr::kable()
```

method	Accuracy	Sensitivity	Specificity
Knn	0.6612903	0.2272727	0.9

Here the accuracy, sensitivity and specificity are lower than the previous model (66%, 22% and 90%).

- Tree Model

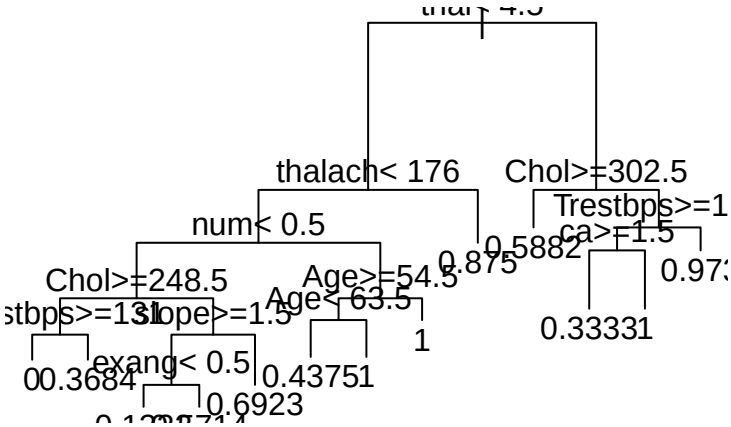
For the decision tree ,we will use rpart function from the rpart library.

```r
#fitting the model
model_rpart=rpart(Sex~ .,data=train_set)
#predictions
y_hat_rpart=predict(model_rpart,test_set)
#confusion matrix
cm_rpart=confusionMatrix(data=factor(round(y_hat_rpart)),reference = factor(test_set$Sex))
cm_rpart
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 14  7
##          1  8 33
##
##                Accuracy : 0.7581
##                  95% CI : (0.6326, 0.8578)
##     No Information Rate : 0.6452
##     P-Value [Acc > NIR] : 0.03937
##
##                   Kappa : 0.4661
##
##  Mcnemar's Test P-Value : 1.00000
##
##             Sensitivity : 0.6364
##             Specificity : 0.8250
##          Pos Pred Value : 0.6667
##          Neg Pred Value : 0.8049
##              Prevalence : 0.3548
##          Detection Rate : 0.2258
##    Detection Prevalence : 0.3387
##       Balanced Accuracy : 0.7307
##
##        'Positive' Class : 0
##
```

The plot of the desicion tree

```r
plot(model_rpart)
text(model_rpart)
```




```
results[3,]%>%knitr::kable()
```

method	Accuracy	Sensitivity	Specificity
Rpart	0.7580645	0.6363636	0.825

Here we observe a big increase in sensitivity(63.6%),the other two are almost same with the other two models(accuracy 75.8% and specificity 82.5%).

Results

Here is the complete table of all the models.

```
results%>%knitr::kable()
```

method	Accuracy	Sensitivity	Specificity
Linear Regression	0.7419355	0.4090909	0.925
Knn	0.6612903	0.2272727	0.900
Rpart	0.7580645	0.6363636	0.825

Conclusion

In this project are used real-life data from machine learning repository and tried to make a prediction of the sex using three different models.Heart disease models are useful for medical reasearch and of course for non

professionals to understand this disease which is the cause of death of thousands people around the world.

Reference

Introduction of Data Science by Rafael A. Irizarry.