Heart Disease - Analysis 2.0

Carlos Kelaidis

12/5/2021

Read in data set:

```
heart_dat <- read.csv("~/Documents/My Working Directory/Personal Projects/Heart Attack Prediciton & Ana
#View(heart_dat)
```

Load some libraries:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.5          v purrr   0.3.4
## v tibble  3.1.6          v dplyr   1.0.7.9000
## v tidyr   1.1.4          v stringr 1.4.0
## v readr   1.3.1          v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(ggpubr)
library(tibble)
```

#EDA

Let's look at our data set:

```
dim(heart_dat)
```

```
## [1] 918  12
```

```
heart_dat[1:5,]
```

```
##   Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
## 1  40   M           ATA       140         289         0     Normal   172
## 2  49   F           NAP       160         180         0     Normal   156
## 3  37   M           ATA       130         283         0         ST    98
## 4  48   F           ASY       138         214         0     Normal   108
## 5  54   M           NAP       150         195         0     Normal   122
##   ExerciseAngina Oldpeak ST_Slope HeartDisease
## 1              N     0.0       Up            0
## 2              N     1.0     Flat            1
## 3              N     0.0       Up            0
## 4              Y     1.5     Flat            1
## 5              N     0.0       Up            0
```

Ok so we have 918 observation and 11 variables + the repsonse (**HeartDisease**)

```r
str(heart_dat)
```

```
## 'data.frame':    918 obs. of  12 variables:
##  $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 1 2 2 1 ...
##  $ ChestPainType : Factor w/ 4 levels "ASY","ATA","NAP",..: 2 3 2 1 3 3 2 2 1 2 ...
##  $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
##  $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : Factor w/ 3 levels "LVH","Normal",..: 2 2 3 2 2 2 2 2 2 2 ...
##  $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: Factor w/ 2 levels "N","Y": 1 1 1 2 1 1 1 1 2 1 ...
##  $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : Factor w/ 3 levels "Down","Flat",..: 3 2 3 2 3 3 3 3 2 3 ...
##  $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

```r
cor(heart_dat[,-c(2,3,7,9,11,13)])
```

```
##                     Age   RestingBP Cholesterol   FastingBS      MaxHR
## Age          1.00000000  0.25439936 -0.09528177  0.19803907 -0.3820447
## RestingBP    0.25439936  1.00000000  0.10089294  0.07019334 -0.1121350
## Cholesterol -0.09528177  0.10089294  1.00000000 -0.26097433  0.2357924
## FastingBS    0.19803907  0.07019334 -0.26097433  1.00000000 -0.1314385
## MaxHR       -0.38204468 -0.11213500  0.23579240 -0.13143849  1.0000000
## Oldpeak      0.25861154  0.16480304  0.05014811  0.05269786 -0.1606906
## HeartDisease  0.28203851  0.10758898 -0.23274064  0.26729119 -0.4004208
##                 Oldpeak HeartDisease
## Age          0.25861154    0.2820385
## RestingBP    0.16480304    0.1075890
## Cholesterol  0.05014811   -0.2327406
## FastingBS    0.05269786    0.2672912
## MaxHR       -0.16069055   -0.4004208
## Oldpeak      1.00000000    0.4039507
## HeartDisease  0.40395072    1.0000000
```

Visualizing data:

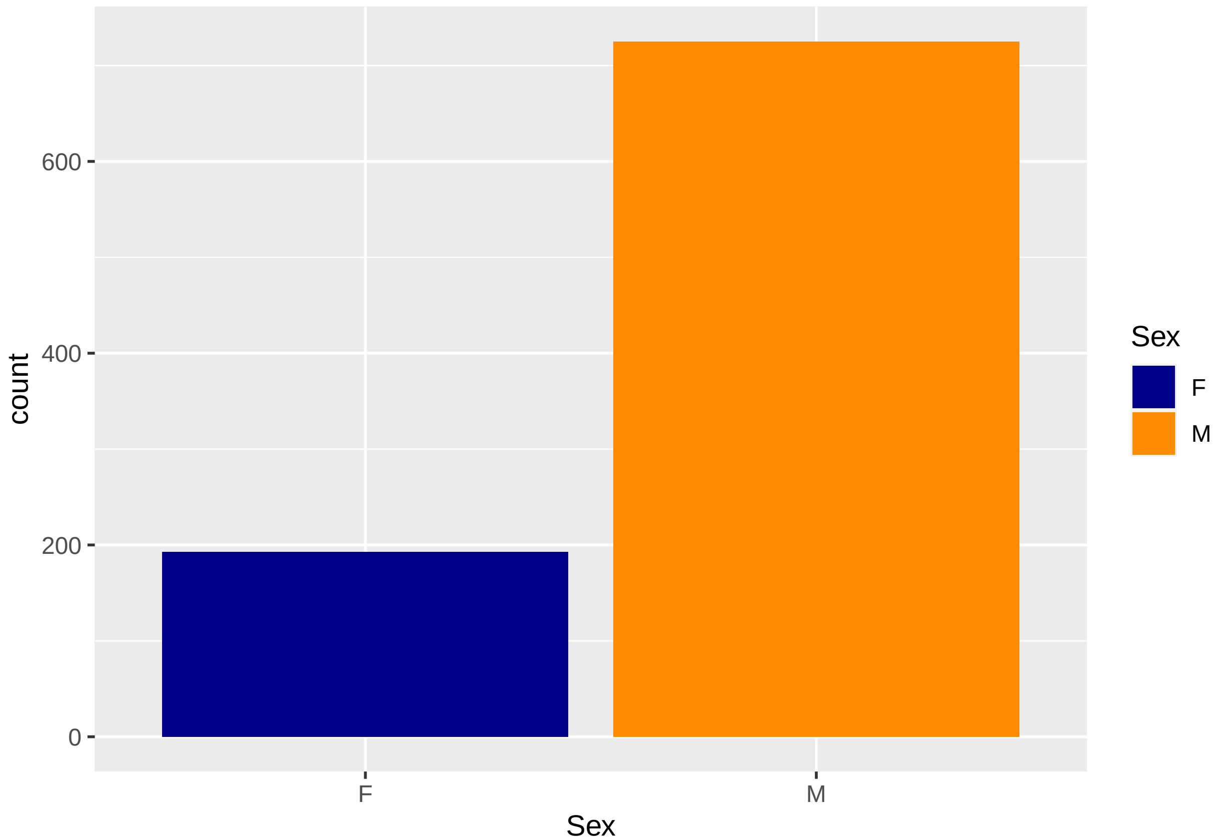
```r
#Want to observe distribution of sex
summary(heart_dat$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   47.00   54.00   53.51   60.00   77.00
```

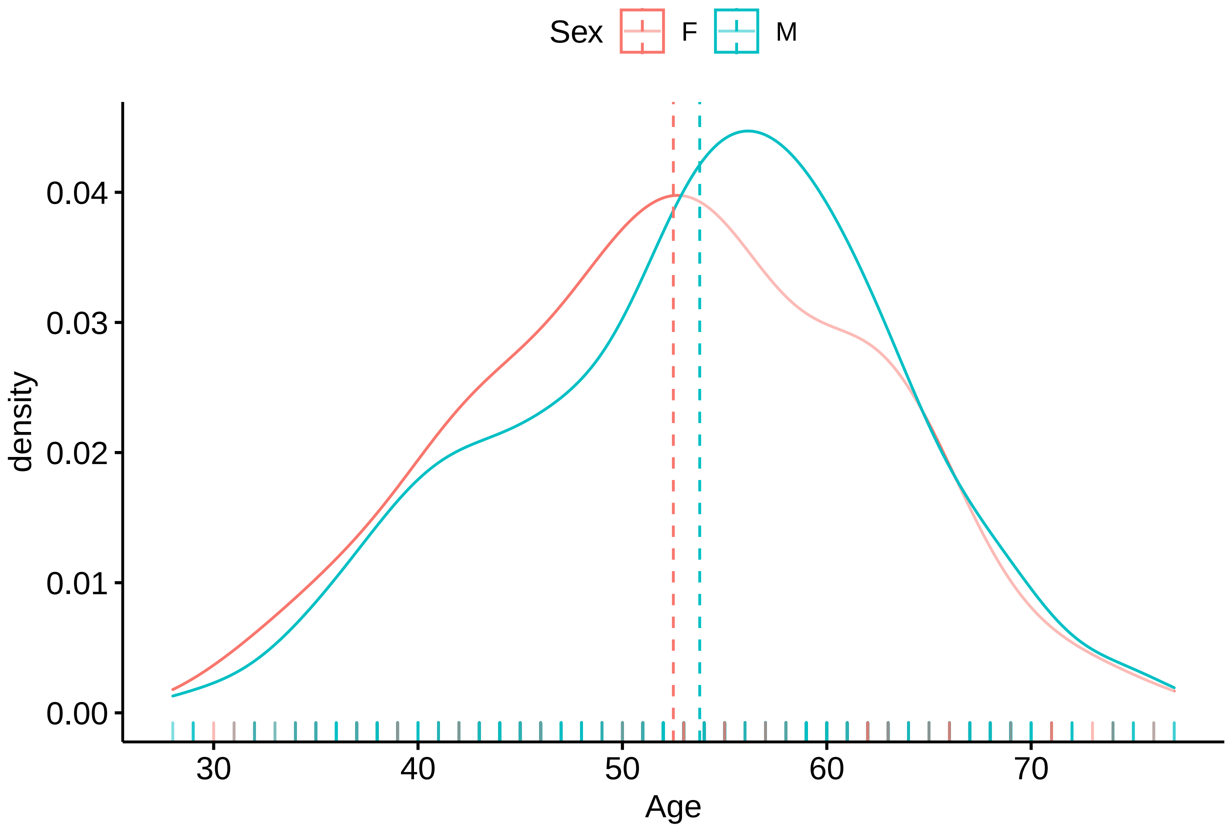
```r
summary(heart_dat$Sex)
```

```
##   F   M
## 193 725
```

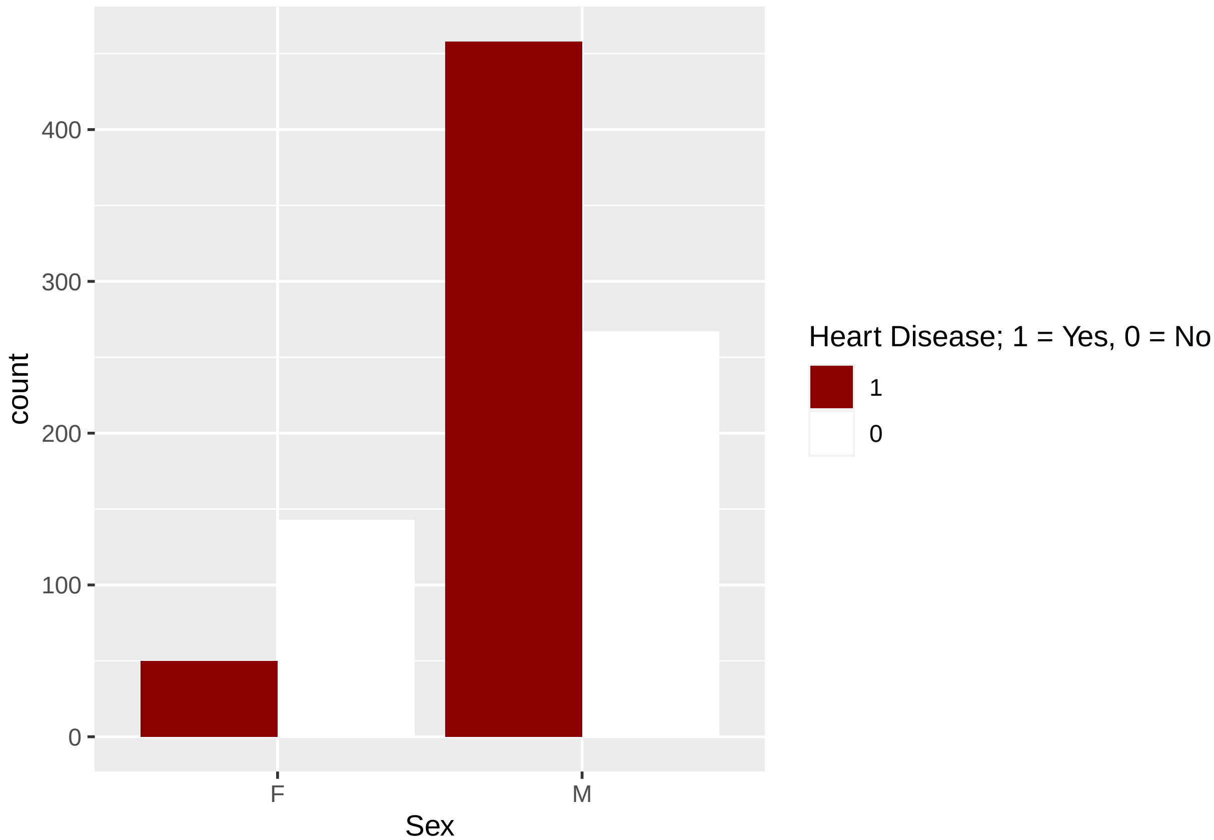
```r
#Do want to observe Age, see the distribution of young VS old, dont want no crazy weights cause by an u

#Sex distribution
heart_dat%>%
  #group_by(HeartDisease)%>%
  ggplot(aes(Sex))+
  geom_bar(aes(fill=Sex))+
  scale_fill_manual(values = c("darkblue", "darkorange"))
```




```r
#Let's look at age distribution between sexes
heart_dat%>%
  ggdensity(x="Age", add="mean", rug=T,
            color="Sex")
```




```r
#Convert HeartDisease to factor
heart_dat$fac_HD<-factor(NA, level=c("1", "0"))
heart_dat$fac_HD[heart_dat$HeartDisease==1]<-"1"
heart_dat$fac_HD[heart_dat$HeartDisease==0]<-"0"

heart_dat%>%
  #group_by(HeartDisease)%>%
  ggplot(aes(Sex))+
  geom_bar(aes(fill=fac_HD), position="dodge")+
  scale_fill_manual(values = c("darkred", "white"))+
  labs(fill="Heart Disease; 1 = Yes, 0 = No")
```



We have a much larger amount of males (725) compared to females (193).

The age distribution between sexes is similar. For women it seems pretty normal around the average age (around 52), for men it is a bit more skewed to the right (towards older ages), so men are generally older than women in this dataset => this can influence our results, as we see further down below older ages (particularly older than the mean) are more correlated to the disease than younger ages => more older ages in men => more cases in men.

We see also in males, there are more cases than not, while in females the not cases are larger in count than the cases.

We see that broken down right below (m2)

```
#m<-as.tibble(c(heart_dat%>%
  #select(Sex, HeartDisease)%>%
  #filter(Sex=="M")%>%
  #summarise(Males_with_Heart_disease=sum(HeartDisease=="1")),
  #heart_dat%>%
  #select(Sex, HeartDisease)%>%
  #filter(Sex=="M")%>%
  #summarise(Males_without_Heart_disease=sum(HeartDisease=="0")),
  #heart_dat%>%
  #select(Sex, HeartDisease)%>%
  #filter(Sex=="F")%>%
  #summarise(Females_with_Heart_disease=sum(HeartDisease=="1")),
  #heart_dat%>%
  #select(Sex, HeartDisease)%>%
  #filter(Sex=="F")%>%
  #summarise(Females_without_Heart_disease=sum(HeartDisease=="0")),
```

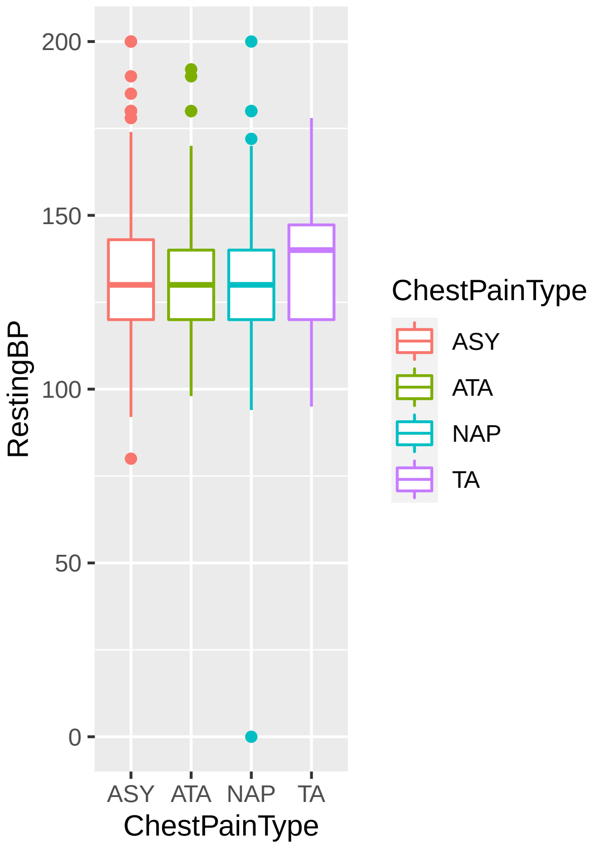
```
  #))


#m2<-pivot_longer(m,c("Males_with_Heart_disease", "Males_without_Heart_disease",
 #                     "Females_with_Heart_disease",
  #                    "Females_without_Heart_disease"),
   #               names_to="Class", values_to="Count")

#m2
heart_dat[1:5,]
```

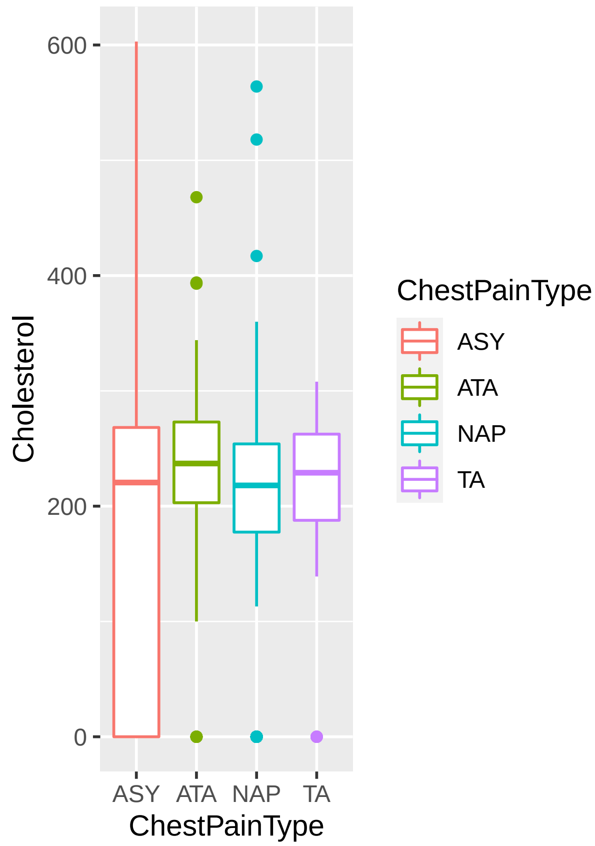
```
##   Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
## 1  40   M           ATA       140         289         0     Normal   172
## 2  49   F           NAP       160         180         0     Normal   156
## 3  37   M           ATA       130         283         0         ST    98
## 4  48   F           ASY       138         214         0     Normal   108
## 5  54   M           NAP       150         195         0     Normal   122
##   ExerciseAngina Oldpeak ST_Slope HeartDisease fac_HD
## 1              N     0.0       Up            0      0
## 2              N     1.0     Flat            1      1
## 3              N     0.0       Up            0      0
## 4              Y     1.5     Flat            1      1
## 5              N     0.0       Up            0      0
```

Above, we see the count broken down.

Let's observe **ChestPainType** and **RestingBP**, also perhaps **Age** with **RestingBP** or **Cholesterol**:

```
#CHestPainType & RestingBP
box1<-heart_dat%>%
  ggplot(aes(x=ChestPainType, y=RestingBP))+
  geom_boxplot(aes(color=ChestPainType))

box2<-heart_dat%>%
  ggplot(aes(x=ChestPainType, y=Cholesterol))+
  geom_boxplot(aes(color=ChestPainType))

ggarrange(box1, box2, ncol=2)
```

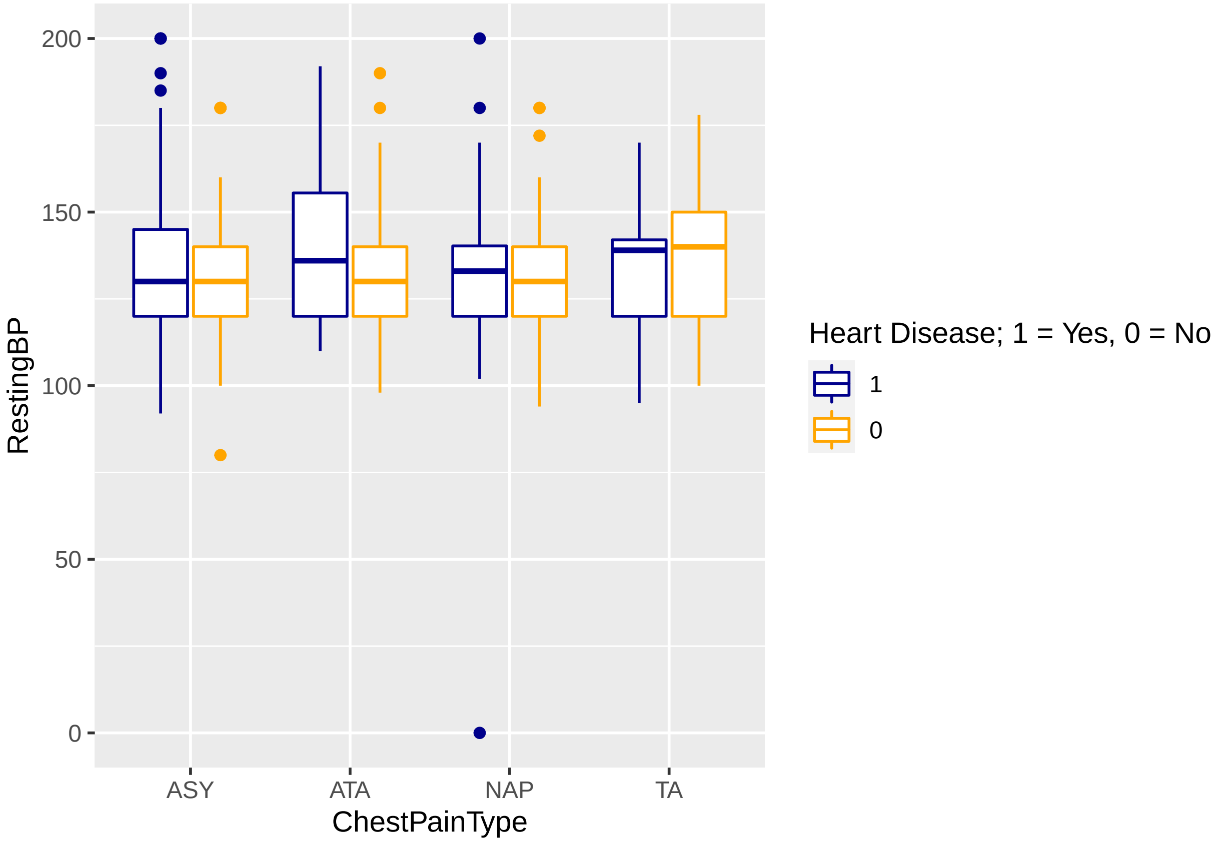




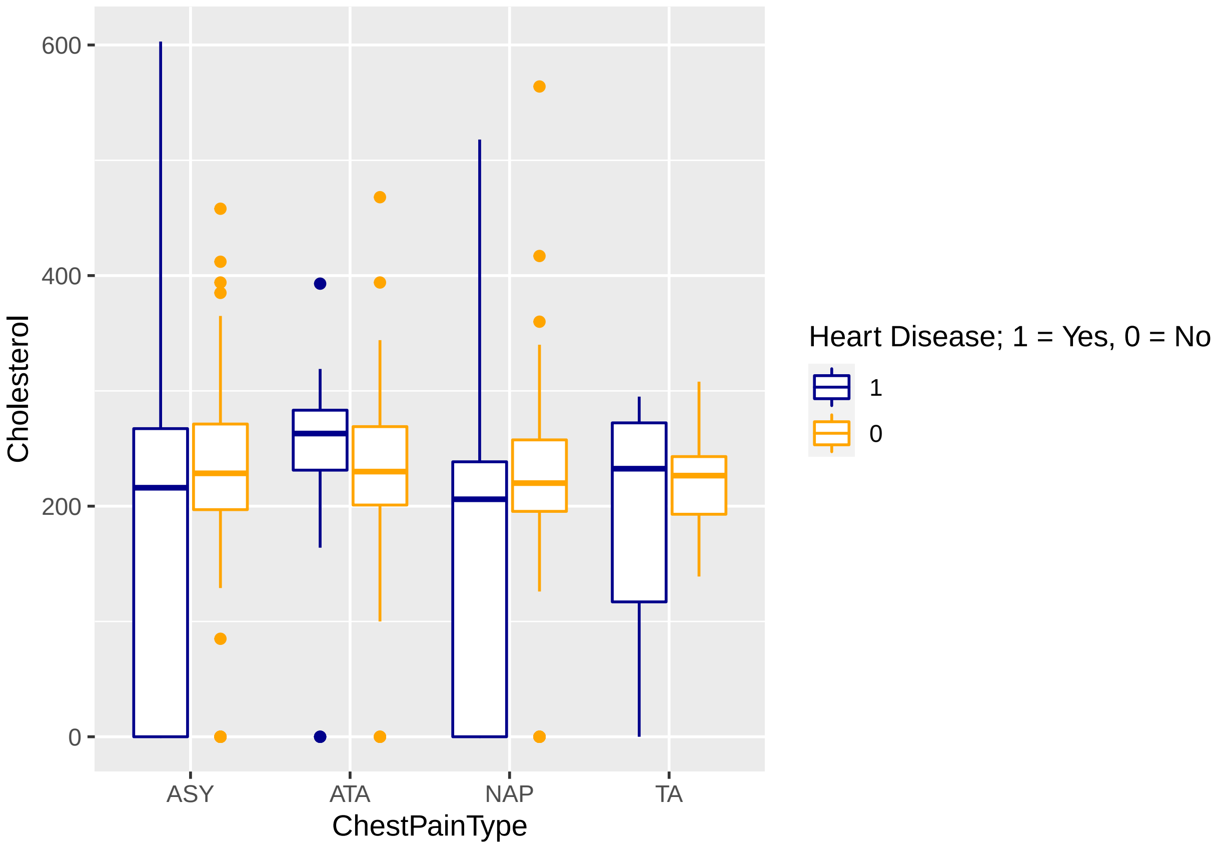
We see all boxplots at about the same level, average **RestingBP** is about the same in all four different **ChestPainType**s.

With regards to **Cholesterol** however, it appears that ChestPainType has a more uneven distribution, the means however appear to be similar however.

```
heart_dat%>%
  ggplot(aes(x=ChestPainType, y=RestingBP))+
  geom_boxplot(aes(color=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```



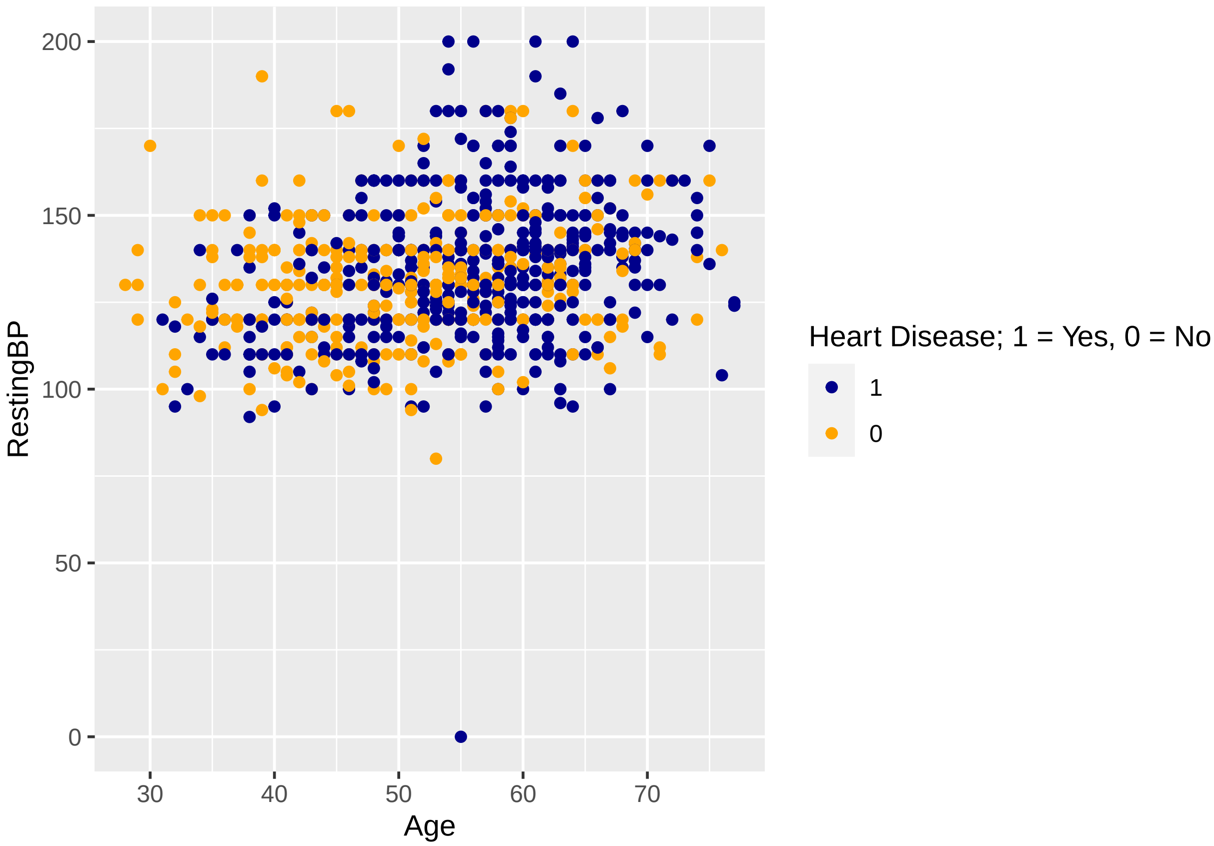

```
heart_dat%>%
  ggplot(aes(x=ChestPainType, y=Cholesterol))+
  geom_boxplot(aes(color=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```



Again, distributions of Resting Blood Pressures appear to be fairly constant among diseased and non-diseased for all Chest Pain Types.

Uneven for Cholesterol, we examine Cholesterol in further detail later on.

```
heart_dat%>%
  ggplot(aes(x=Age, y=RestingBP))+
  geom_point(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```



So we see that Resting Blood Pressure levels are fairly constant for all ages, but we do see a higher concentration of Heart Disease Cases for higher age values. => **Age**!

Want to look at **Age** a bit more - Namely its distribution:

```
summary(heart_dat$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   47.00   54.00   53.51   60.00   77.00
```

```
#Tot diseased
length(which(heart_dat$HeartDisease==1))#508
```

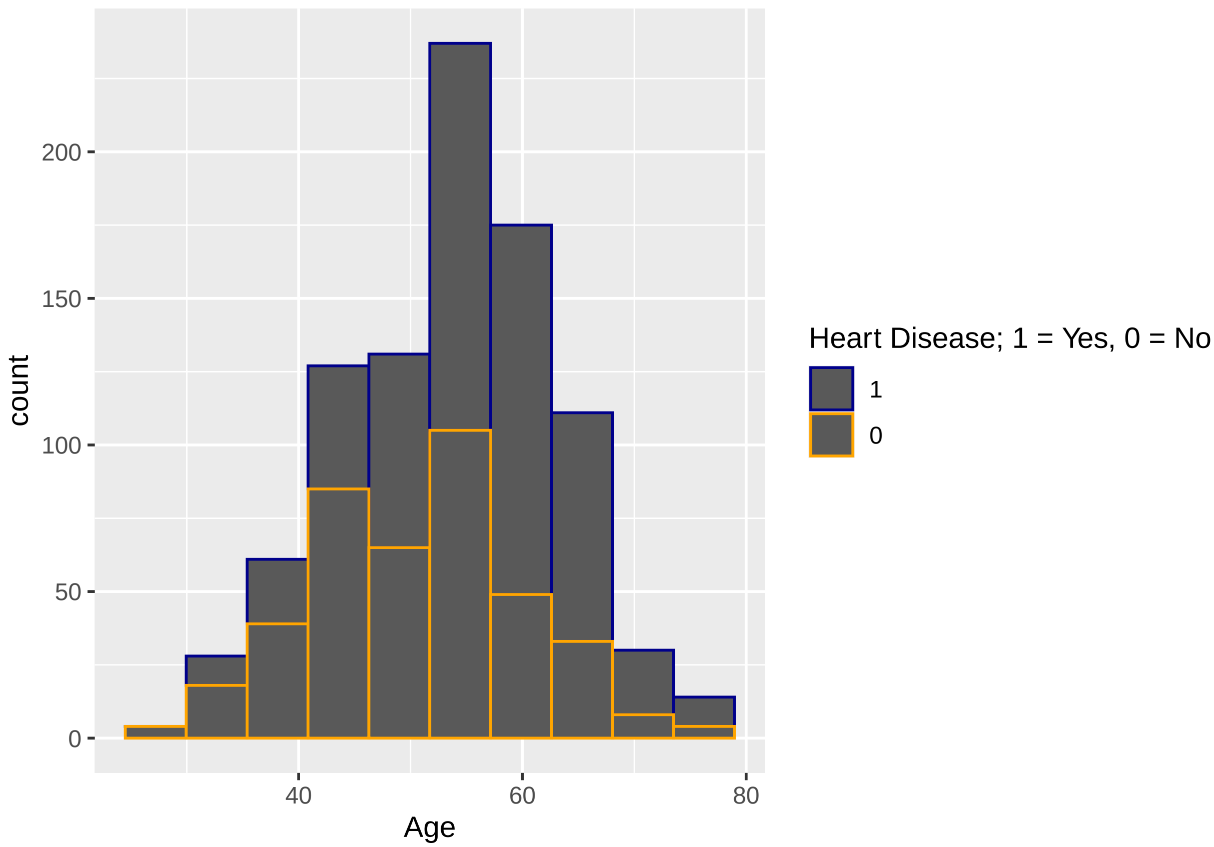
```
## [1] 508
```

```
#Tot not diseased
length(which(heart_dat$HeartDisease==0))#410
```

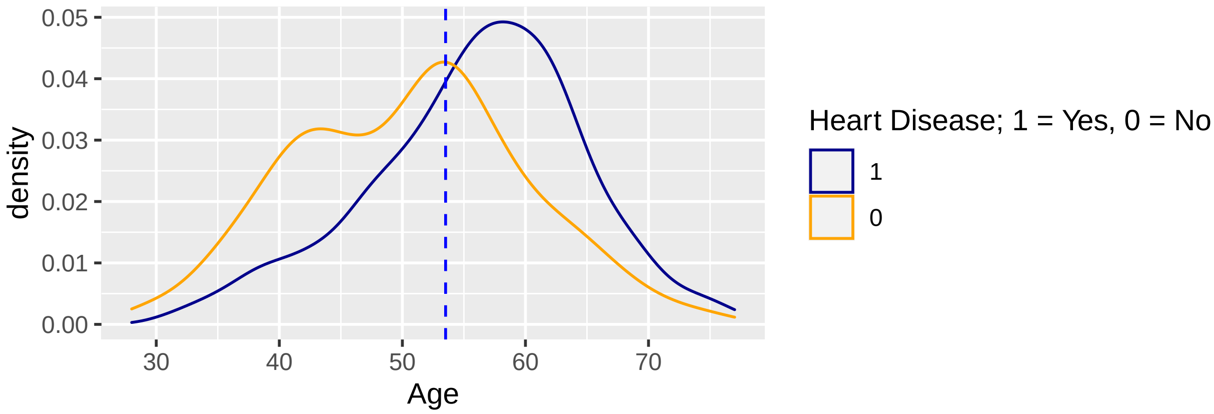
```
## [1] 410
```

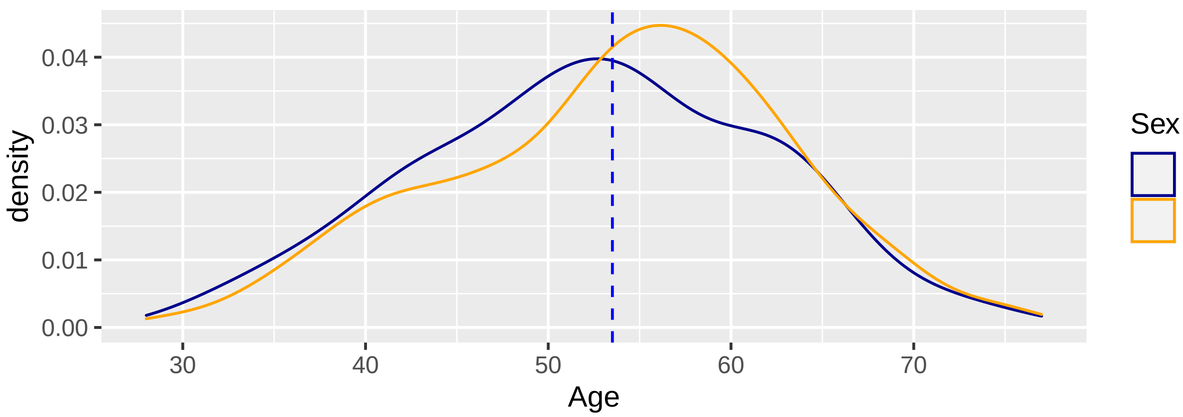
```
bar_graph<-heart_dat%>%
  ggplot(aes(Age))+
  geom_histogram(bins=10,aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No")

dens_dist<-heart_dat%>%
  ggplot(aes(Age))+
  geom_density(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange"))+
```

```r
  geom_vline(aes(xintercept=mean(Age)), col="blue", lty=2)+
  labs(color="Heart Disease; 1 = Yes, 0 = No")

dens_dist2<-heart_dat%>%
  ggplot(aes(Age))+
  geom_density(aes(col=Sex))+
  scale_color_manual(values = c("darkblue", "orange"))+
  geom_vline(aes(xintercept=mean(Age)), col="blue", lty=2)+
  labs(color="Sex")

bar_graph
```




```r
ggarrange(dens_dist, dens_dist2, nrow=2)
```





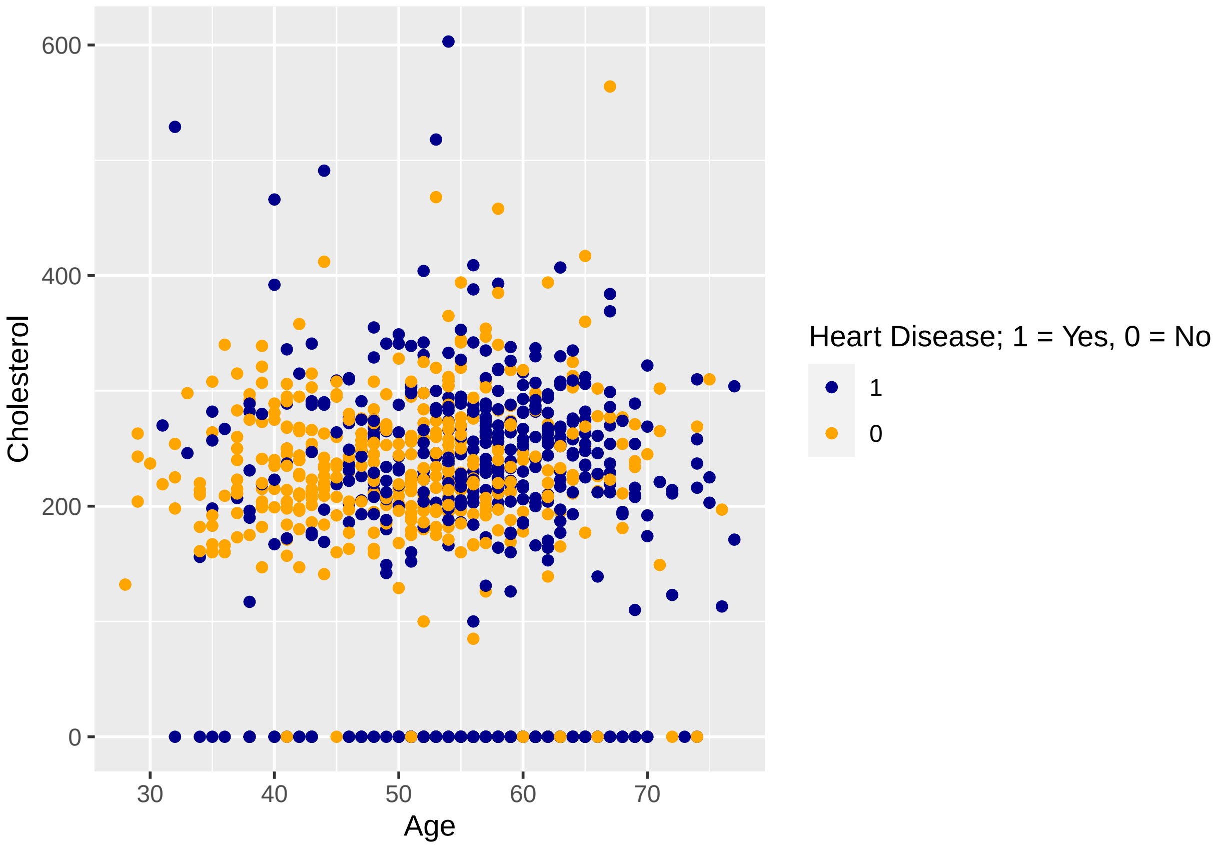
Based on the density plot wee see that, higher concentrations of diseased patients are reserved for ages higher than the mean age, while for non-diseased patients, we see a high density around the mean age, as well as for ages smaller than the mean age.

We can see, just as the dsitribution of males is skewed towards older ages, so is the distribution of diseased cases => Just as wee saw before a there is a higher concentration of cases for older ages which is why there is also more cases in men (they are older than women in this dataset).

It is not so that one sex has more risk of having heart disease, as it is that age plays a significant role.

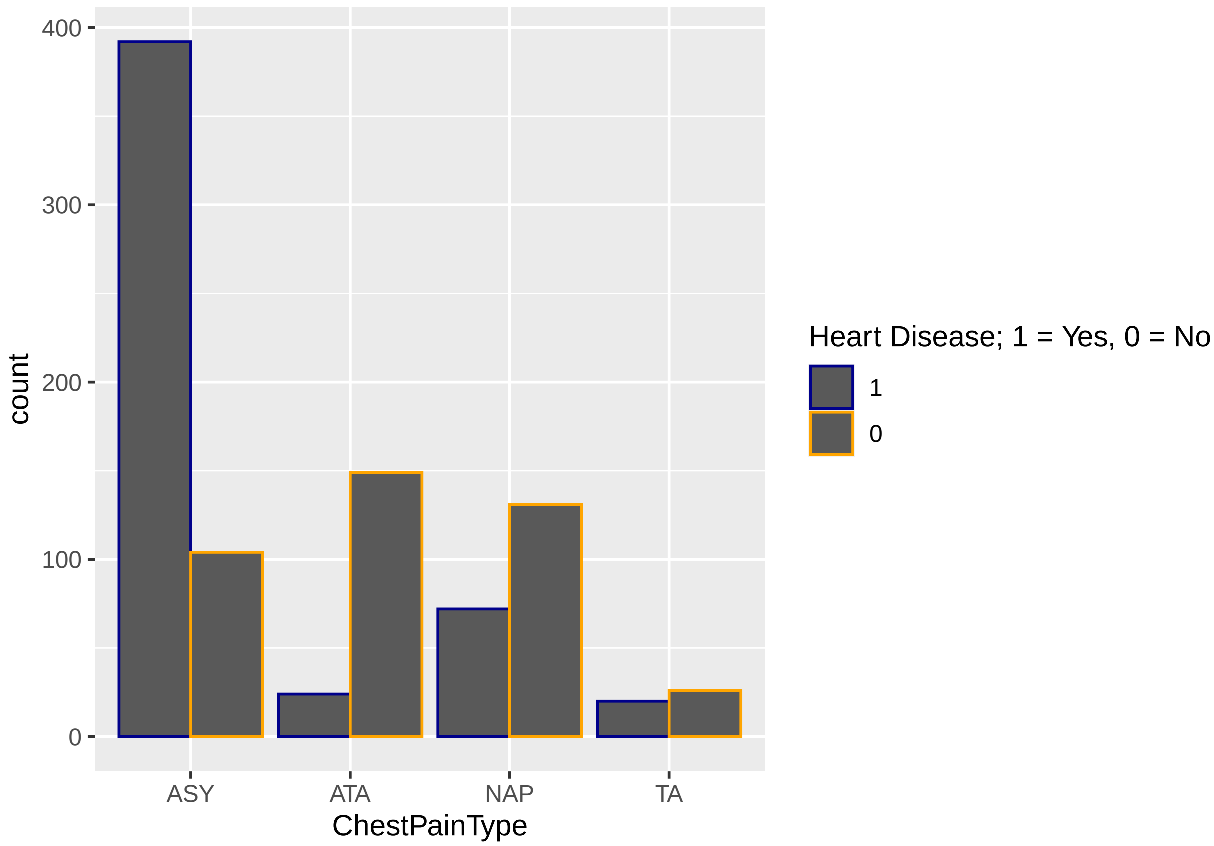
Visualizing some more variables:

```
heart_dat%>%
  ggplot(aes(x=Age, y=Cholesterol))+
  geom_point(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```



Cholesterol levels are fairly even among all age brackets, however there is a definite relationship between patients that have Cholesterol levels of 0 and that have Heart Disease. => **Cholesterol**!

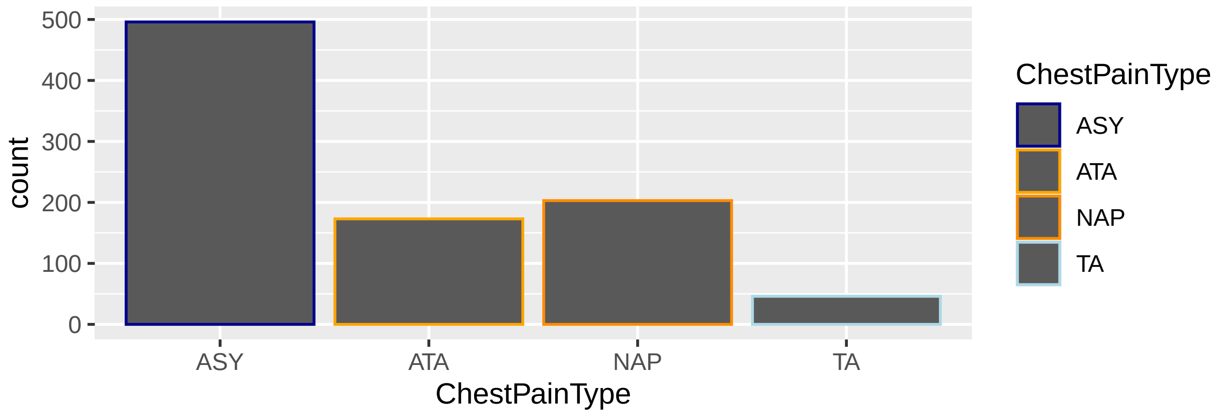
```
#Chest Pain Type and Disease status
heart_dat%>%
  ggplot(aes(x=ChestPainType))+
  geom_bar(aes(col=fac_HD), position="dodge")+
  scale_color_manual(values = c("darkblue", "orange"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```

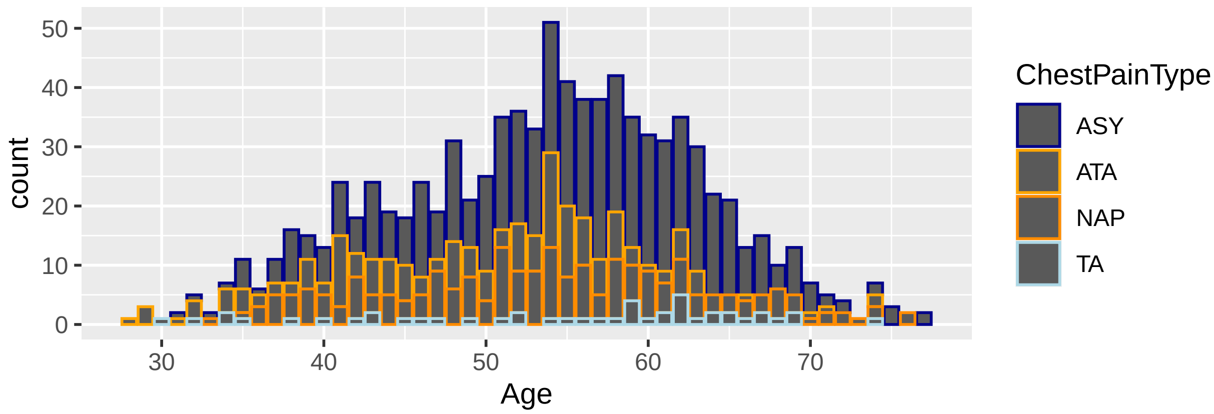


We can see that most Heart Disease cases are reserved for patients that have the **ASY** type of chest pain.

Let's see whether **Age** or any other variables are related to **ChestPainType**:

```
bar1<-heart_dat%>%
  ggplot(aes(x=Age))+
  geom_bar(aes(col=ChestPainType))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue"))

bar2<-heart_dat%>%
  ggplot(aes(ChestPainType))+
  geom_bar(aes(col=ChestPainType))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue"))

ggarrange(bar2,bar1, nrow=2)
```





The **ASY** relationship to the disease is a bit more explained. I believw the main reason for this assiociation is that we have a lot more counts of ASY chest pain type than for any other type of chest pain. Because if we look at how Chest Pain Types are distributed with regards to age, the dsitribution is fairly similar, meaning there is not one Chest Pain Type that is more prominent than others for a particualr age bracket/value, it is just that there is a much larger count of **ASY** chest pain type.

However, a good note is that Chest Pain Types are more prominent (for all chest pain types) among older age values.

ChestPainType!

Is it just that we have way more older ages than younger ones? Let's look at that:

```
#Create new factor, older >= mean(age), young < mean(age)
age<-factor(NA, levels = c("young", "mid", "aged"))

age[heart_dat$Age>=60]<-"aged"
age[heart_dat$Age<30]<-"young"
age[heart_dat$Age<60 & heart_dat$Age>=30]<-"mid"

length(which(age=='aged'))
```

```
## [1] 253
```

```
length(which(age=='mid'))
```

```
## [1] 661
```

```
length(which(age=='young'))
```

```
## [1] 4
```

With a mean age of 53/54, the data set is fairly old.

Furthermore, we only have 4 observations below 30 yrs old, 250 above 60 and the majority; 661 between 30 and 60.

Perhaps I will change the split to 50 (instead of 30), see if we get a more even split of observations, or perhaps:

```
age<-factor(NA, levels = c("young", "mid", "aged"))

age[heart_dat$Age>=60]<-"aged"
age[heart_dat$Age<50]<-"young"
age[heart_dat$Age<60 & heart_dat$Age>=50]<-"mid"

length(which(age=='aged'))
```

```
## [1] 253
```

```
length(which(age=='mid'))
```

```
## [1] 374
```

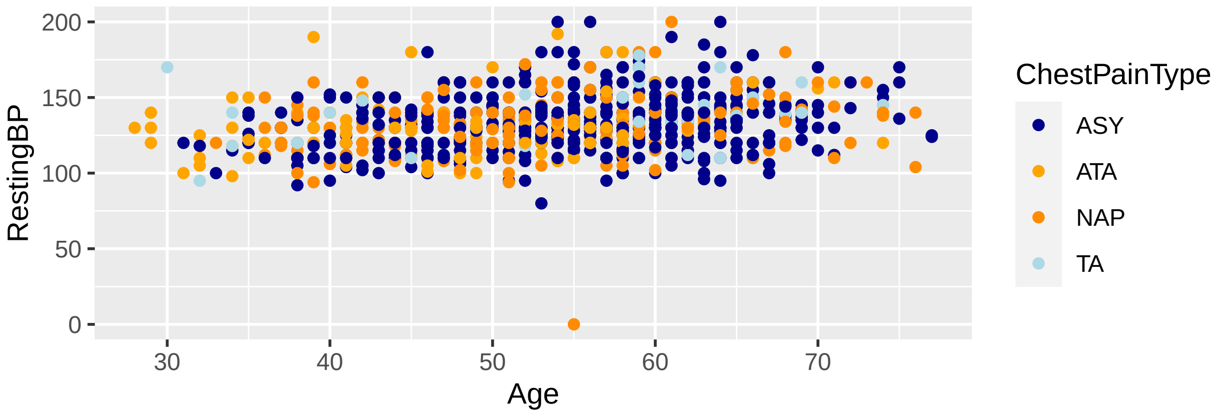
```
length(which(age=='young'))
```

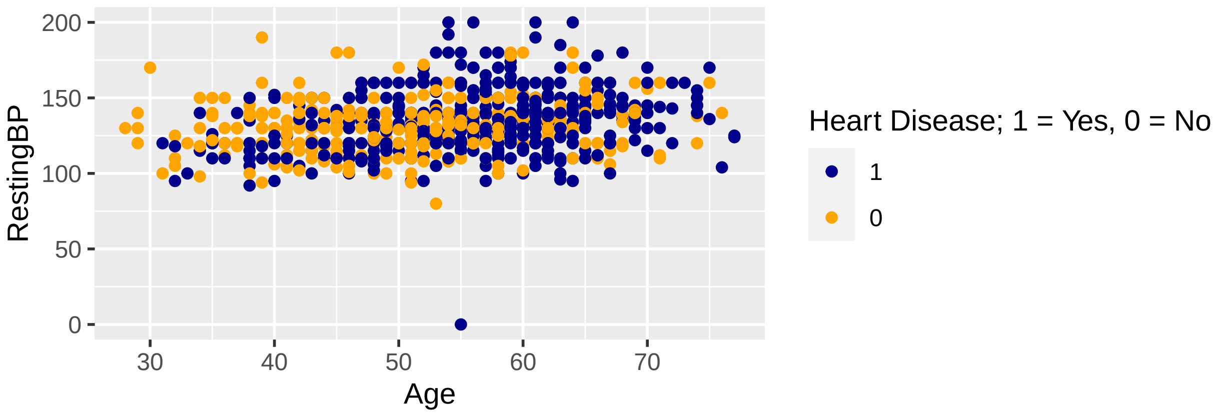
```
## [1] 291
```

The majority is between 50 and 60 years of age. We see also from the density curves above, that heart disease is more common in patients aged older than the mean age (53.5).

Are any other variables are related to **ChestPainType**:

```
#RestingBP
ggarrange(heart_dat%>%
  ggplot(aes(x=Age, y=RestingBP))+
  geom_point(aes(col=ChestPainType))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue")),
  heart_dat%>%
  ggplot(aes(x=Age, y=RestingBP))+
  geom_point(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No"), nrow=2)
```



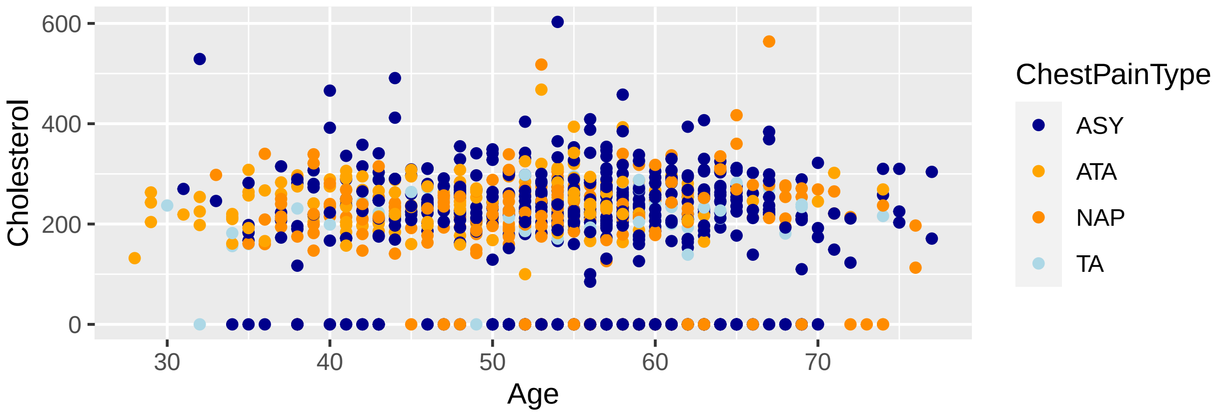


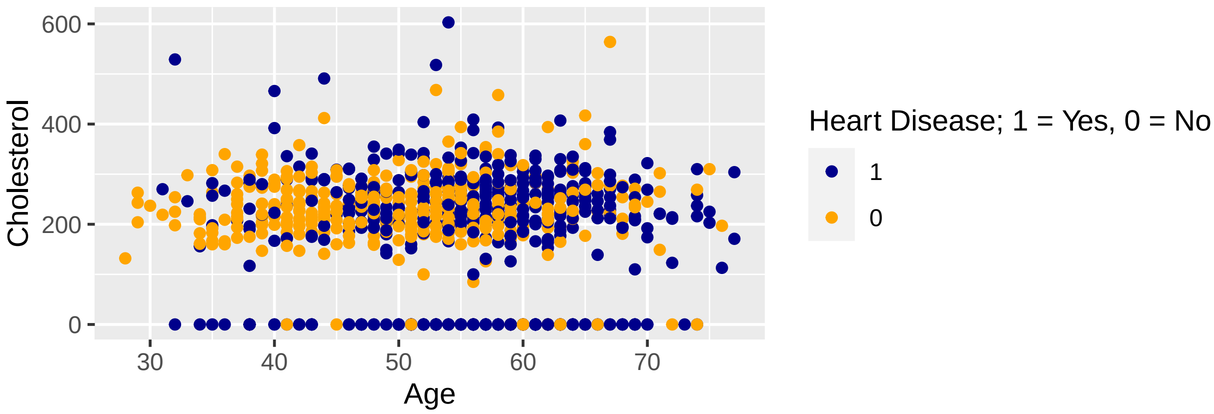
I do not see any stricking observation between RestingBP and ChestPainType or even Age. RestingBP values are evened out across the age and chestpain distributions. => From what we see here not sure how relevant Resting BP is in accounting for Heart Disease.

So far we have noted the importance, purely based on observations, of **Age**, **ChestPainType** and **Cholesterol**.

We can also note **Sex**, however there are way more men than women, furthermore the men are aged older than women => so Sex will be heavily wieghted towards having the disease as more men => more aged & more aged => more diseased.

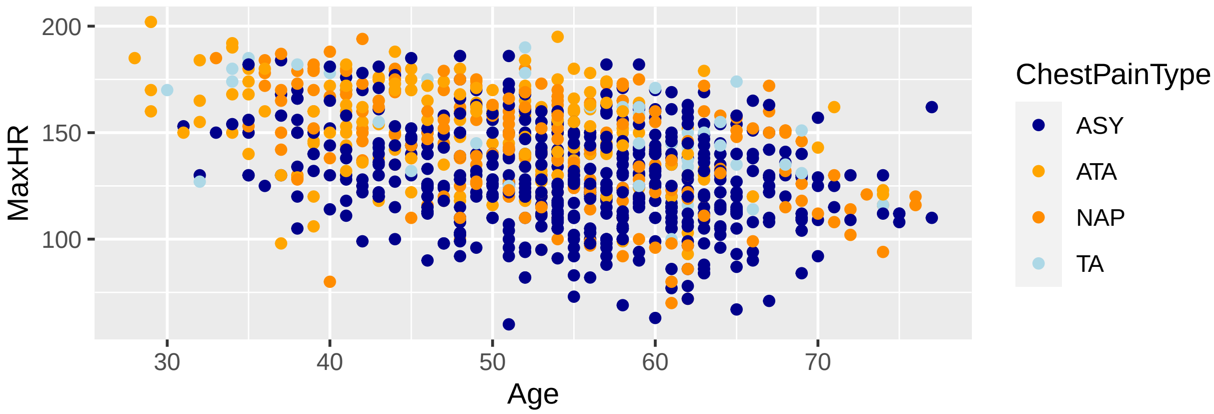
```
#Cholesterol
ggarrange(heart_dat%>%
  ggplot(aes(x=Age, y=Cholesterol))+
  geom_point(aes(col=ChestPainType))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue")),
  heart_dat%>%
  ggplot(aes(x=Age, y=Cholesterol))+
  geom_point(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No"), nrow=2)
```

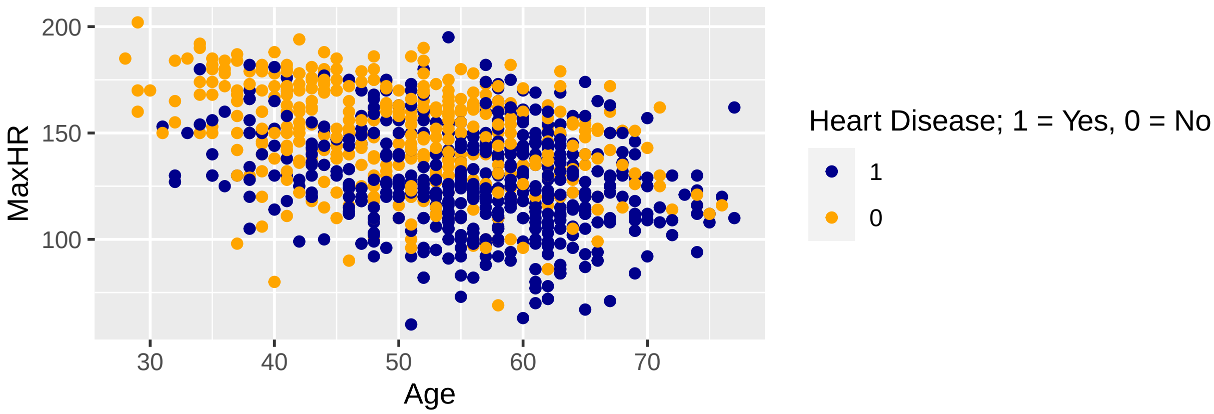




We have already noted our observations for **Cholesterol**:

```
#MaxHR
ggarrange(heart_dat%>%
  ggplot(aes(x=Age, y=MaxHR))+
  geom_point(aes(col=ChestPainType))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue")),
  heart_dat%>%
  ggplot(aes(x=Age, y=MaxHR))+
  geom_point(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No"), nrow=2)
```





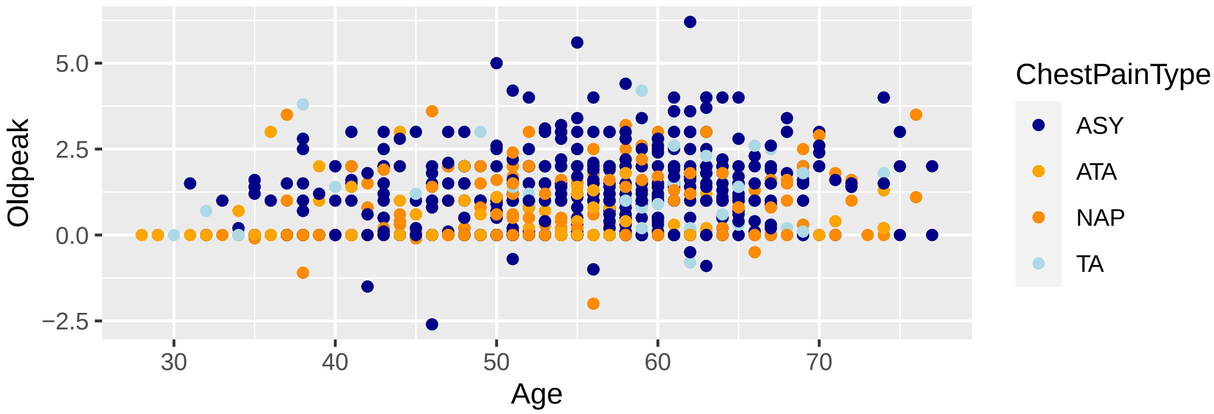
Ok so here we can see some type of relationship between MaxHR & ChestPainType (possible interaction term between them both), as well as between MaxHR & Disease status.

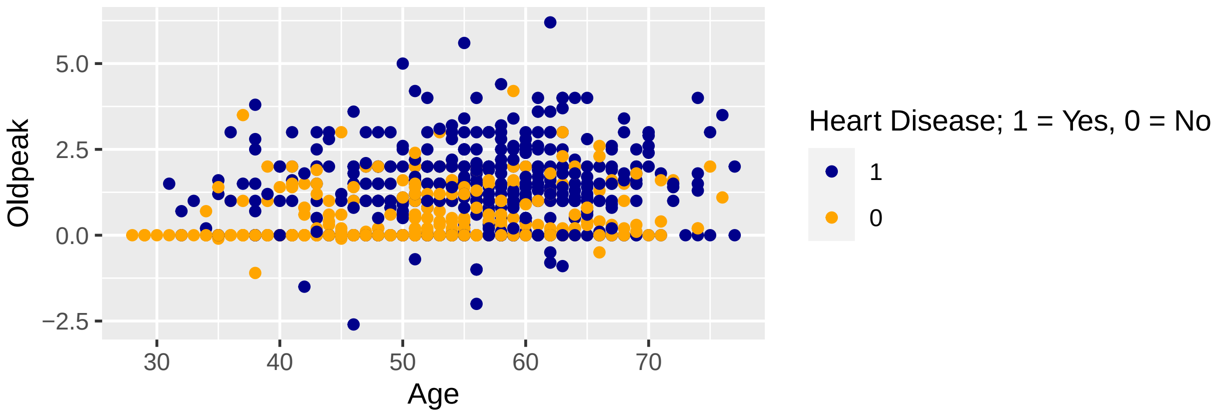
First we see that the ASY chest pain type is (additionally to being reserved for older ages) reserved for lower heart rates.

We also see that lower heart rates are associated with older ages (so possible interaction term between Age & MaxHR).

Finally, we see that lower heart rates are associated with diseased cased => **MaxHR**!

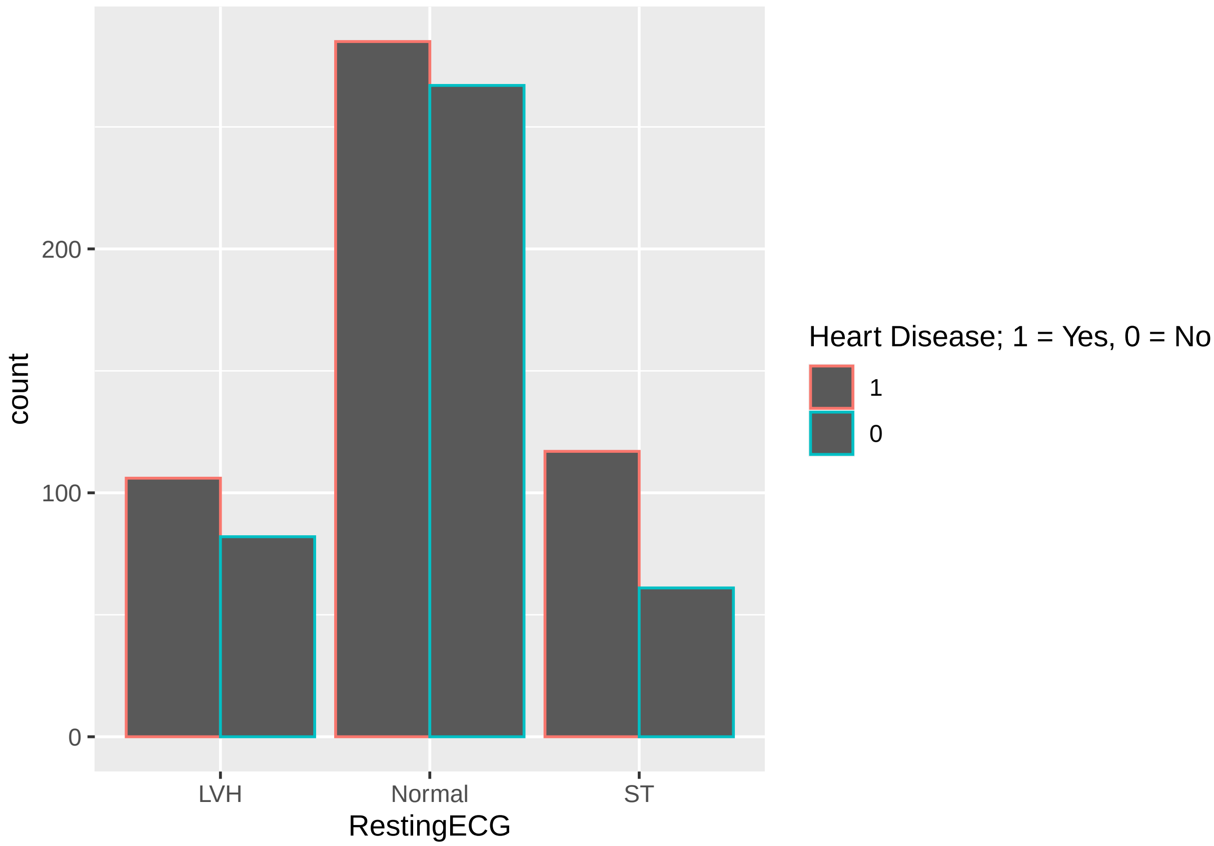
```
#Oldpeak
ggarrange(heart_dat%>%
  ggplot(aes(x=Age, y=Oldpeak))+
  geom_point(aes(col=ChestPainType))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue")),
  heart_dat%>%
  ggplot(aes(x=Age, y=Oldpeak))+
  geom_point(aes(col=fac_HD))+
  scale_color_manual(values = c("darkblue", "orange", "darkorange", "lightblue"))+
  labs(color="Heart Disease; 1 = Yes, 0 = No"), nrow=2)
```





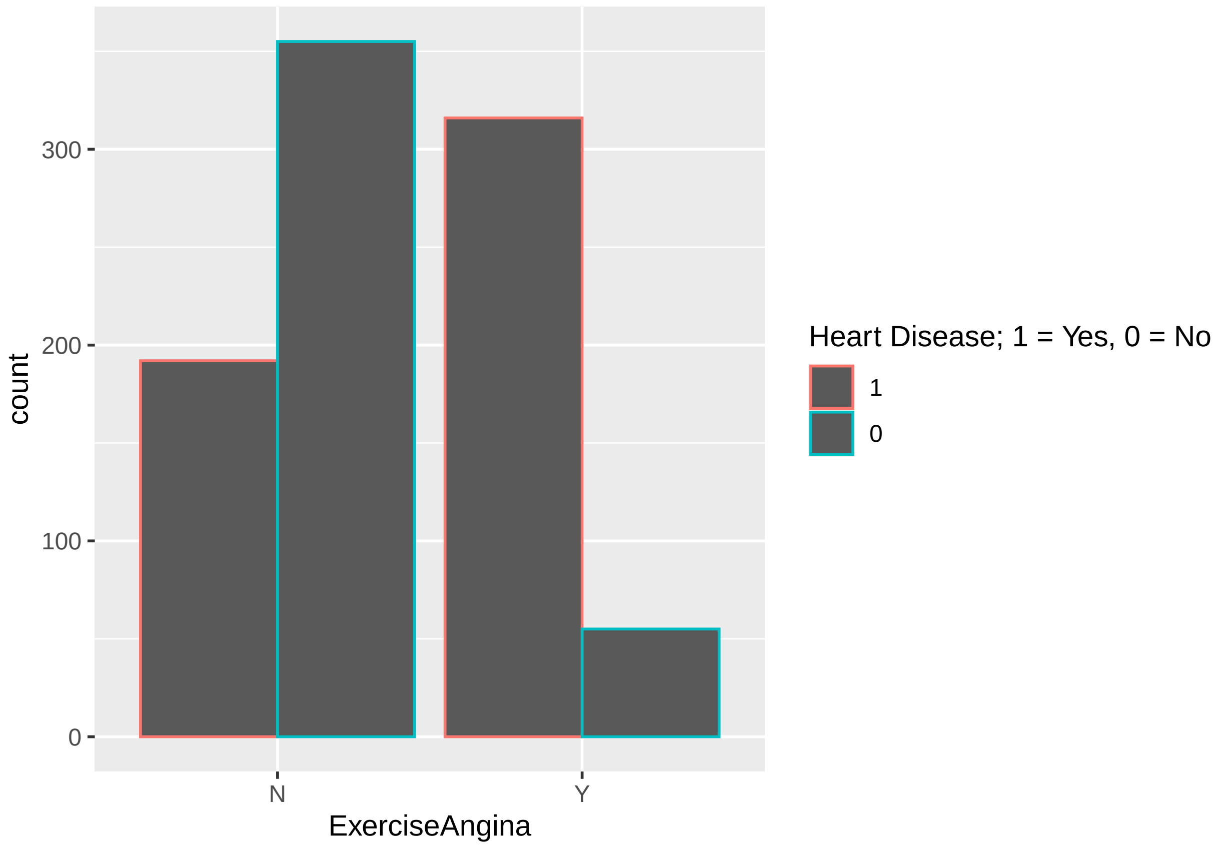
Similarly, higher values of **Oldpeak** seem to be associated with heart disease. These higher oldpeak values appear to be predominently nested in ages > mean(age), and with the ASY chest pain (although like we saw, there is much larger amount of ASY observation compared to the other chestpains, so perhaps we should not take ChestPainType so much into consideration). => **Oldpeak**!

```
#Resting ECG
heart_dat%>%
  ggplot(aes(RestingECG))+
  geom_bar(aes(col=fac_HD), position="dodge")+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```



So for **RestingECG**, we see that the diseased cases outnumber the non-diseased for all three states. I do not see any particular Resting ECG state that is more associated with the disease status compared to others.

```
#Exercise Angina
heart_dat%>%
  ggplot(aes(ExerciseAngina))+
  geom_bar(aes(col=fac_HD), position="dodge")+
  labs(color="Heart Disease; 1 = Yes, 0 = No")
```



We see however here, that **ExerciseAngina** is associated with the disease status. For patients that experienced exercise-induced angina, the cases majorely outnumber the non-cases. => **ExerciseAngina**!

Okay, so we have explored and visualized the relationships between the variables and have a pretty good understanding of the data set.

So far, purely through observations, variables associated with **HeartDisease** are:

- **ExerciseAngina**
- **Oldpeak**
- **MaxHR**
- **Age**
- **Cholesterol**
- **ChestPainType**

A few variables that seem to interact with eachother, for which there is the possibility of an interaction term, are:

- **MaxHR** & **ChestPainType** (ChestPainType is heavily weighted, so maybe not so relevant for an interaction term)
- **MaxHR** & **Age**

We will now perform some model selection approaches, to further examine which variables appear as significant.

We first need to split the data into a train set and a test set:

```
#Split data into train and test set
heart_dat<-heart_dat[,-13]
set.seed(1)
test.sample<-sample(nrow(heart_dat), nrow(heart_dat)/3)#take a third of the data for a the test sample
```

```r
heart.train<-heart_dat[-test.sample,]
heart.test<-heart_dat[test.sample,]
```

BSS:

```r
library(leaps)
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-1
```

```r
#We fit BSS on the whole data set, because we evaluate Cp, BIC and Adjusted R^2
bss.fit1<-regsubsets(HeartDisease~., data=heart_dat, nvmax=15)
sum.bss.fit<-summary(bss.fit1)
sum.bss.fit
```

```
## Subset selection object
## Call: regsubsets.formula(HeartDisease ~ ., data = heart_dat, nvmax = 15)
## 15 Variables  (and intercept)
##                    Forced in Forced out
## Age                    FALSE      FALSE
## SexM                   FALSE      FALSE
## ChestPainTypeATA       FALSE      FALSE
## ChestPainTypeNAP       FALSE      FALSE
## ChestPainTypeTA        FALSE      FALSE
## RestingBP              FALSE      FALSE
## Cholesterol            FALSE      FALSE
## FastingBS              FALSE      FALSE
## RestingECGNormal       FALSE      FALSE
## RestingECGST           FALSE      FALSE
## MaxHR                  FALSE      FALSE
## ExerciseAnginaY        FALSE      FALSE
## Oldpeak                FALSE      FALSE
## ST_SlopeFlat           FALSE      FALSE
## ST_SlopeUp             FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: exhaustive
##           Age SexM ChestPainTypeATA ChestPainTypeNAP ChestPainTypeTA RestingBP
## 1  ( 1 )  " " " "  " "              " "              " "             " "
## 2  ( 1 )  " " " "  " "              " "              " "             " "
## 3  ( 1 )  " " "*"  " "              " "              " "             " "
## 4  ( 1 )  " " "*"  "*"              "*"              " "             " "
## 5  ( 1 )  " " "*"  "*"              "*"              " "             " "
## 6  ( 1 )  " " "*"  "*"              "*"              " "             " "
## 7  ( 1 )  " " "*"  "*"              "*"              " "             " "
## 8  ( 1 )  " " "*"  "*"              "*"              "*"             " "
## 9  ( 1 )  " " "*"  "*"              "*"              "*"             " "
## 10  ( 1 ) " " "*"  "*"              "*"              "*"             " "
## 11  ( 1 ) "*" "*"  "*"              "*"              "*"             " "
```

```
## 12  ( 1 ) "*" "*"  "*"              "*"              "*"             " "
## 13  ( 1 ) "*" "*"  "*"              "*"              "*"             " "
## 14  ( 1 ) "*" "*"  "*"              "*"              "*"             " "
## 15  ( 1 ) "*" "*"  "*"              "*"              "*"             "*"
##           Cholesterol FastingBS RestingECGNormal RestingECGST MaxHR
## 1  ( 1 )  " "         " "       " "              " "          " "
## 2  ( 1 )  " "         " "       " "              " "          " "
## 3  ( 1 )  " "         " "       " "              " "          " "
## 4  ( 1 )  " "         " "       " "              " "          " "
## 5  ( 1 )  " "         " "       " "              " "          " "
## 6  ( 1 )  " "         "*"       " "              " "          " "
## 7  ( 1 )  "*"         "*"       " "              " "          " "
## 8  ( 1 )  "*"         "*"       " "              " "          " "
## 9  ( 1 )  "*"         "*"       " "              " "          " "
## 10  ( 1 ) "*"         "*"       " "              " "          " "
## 11  ( 1 ) "*"         "*"       " "              " "          " "
## 12  ( 1 ) "*"         "*"       " "              " "          "*"
## 13  ( 1 ) "*"         "*"       "*"              " "          "*"
## 14  ( 1 ) "*"         "*"       "*"              "*"          "*"
## 15  ( 1 ) "*"         "*"       "*"              "*"          "*"
##           ExerciseAnginaY Oldpeak ST_SlopeFlat ST_SlopeUp
## 1  ( 1 )  " "             " "     " "          "*"
## 2  ( 1 )  "*"             " "     " "          "*"
## 3  ( 1 )  "*"             " "     " "          "*"
## 4  ( 1 )  " "             " "     " "          "*"
## 5  ( 1 )  "*"             " "     " "          "*"
## 6  ( 1 )  "*"             " "     " "          "*"
## 7  ( 1 )  "*"             " "     " "          "*"
## 8  ( 1 )  "*"             " "     " "          "*"
## 9  ( 1 )  "*"             "*"     " "          "*"
## 10  ( 1 ) "*"             "*"     "*"          "*"
## 11  ( 1 ) "*"             "*"     "*"          "*"
## 12  ( 1 ) "*"             "*"     "*"          "*"
## 13  ( 1 ) "*"             "*"     "*"          "*"
## 14  ( 1 ) "*"             "*"     "*"          "*"
## 15  ( 1 ) "*"             "*"     "*"          "*"
```

The above summary shows the variable selection process of the Best Subset Selection method, but which model size is the most appropriate?

Let's observe Cp, BIC and AdjR2

```r
par(mfrow=c(1,3))
plot(sum.bss.fit$cp, xlab="Number of Variables", ylab="Cp", type="l")+
  abline(h=min(sum.bss.fit$cp)+.2*sd(sum.bss.fit$cp), col=2, lty=2)+
  abline(h=min(sum.bss.fit$cp)-.2*sd(sum.bss.fit$cp), col=2, lty=2)
```

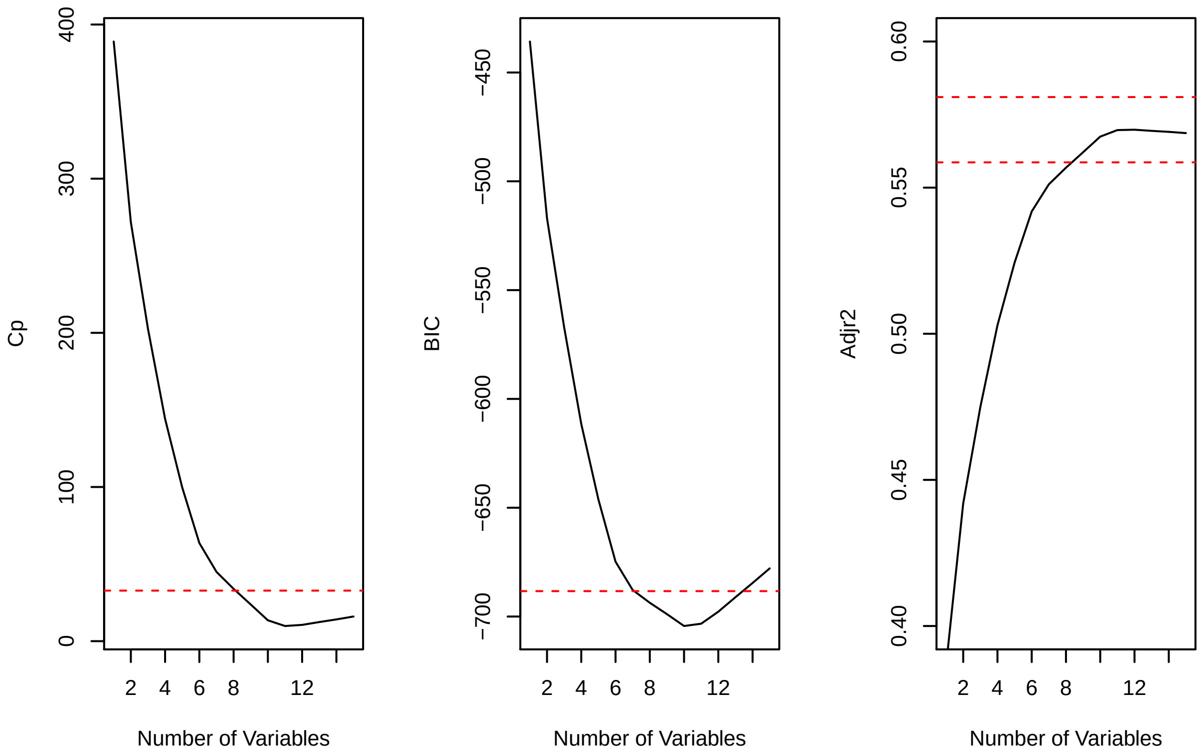
```
## integer(0)
```

```r
plot(sum.bss.fit$bic, xlab="Number of Variables", ylab="BIC", type="l")+
  abline(h=min(sum.bss.fit$bic)+.2*sd(sum.bss.fit$bic), col=2, lty=2)+
  abline(h=min(sum.bss.fit$bic)-.2*sd(sum.bss.fit$bic), col=2, lty=2)
```

```
## integer(0)
```

```r
plot(sum.bss.fit$adjr2, xlab="Number of Variables", ylab="Adjr2", type="l", ylim=c(.4, .6))+
  abline(h=max(sum.bss.fit$adjr2)+.2*sd(sum.bss.fit$adjr2), col=2, lty=2)+
```

```
abline(h=max(sum.bss.fit$adjr2)-.2*sd(sum.bss.fit$adjr2), col=2, lty=2)
```




```
## integer(0)
```

We are not necessarily looking for the extrema, we want the smaller number of parameters within the bounds (0.2 standard deviations from the optimum).

From this, all measures agree on a model of size 7, perhaps even 6. If we do not care about the number of parameters, then perhaps size 10 would be best.

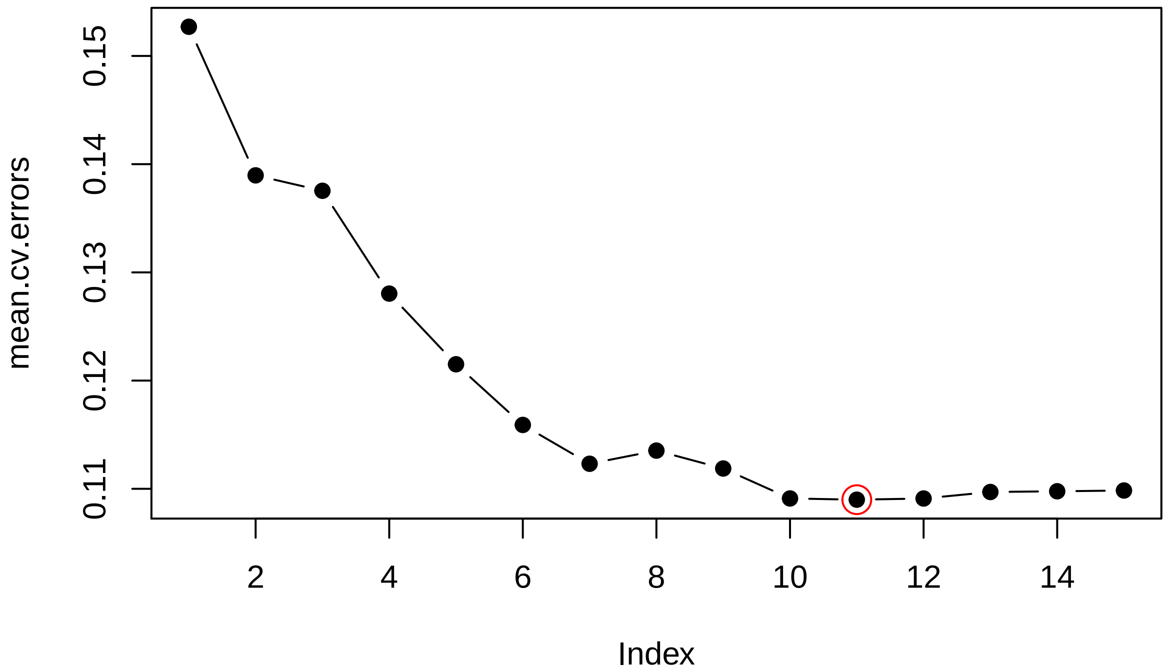
Let's see what size model 10-fold CV picks:

Need to create a predict function and then perform 10-fold CV:

```
predict.regsubsets<-function(object, newdata, id,...){
  form<-as.formula(object$call[[2]])
  mat<-model.matrix(form, newdata)
  coefs<-coef(object, id=id)
  xvars<-names(coefs)
  mat[,xvars]%*%coefs
}
```

```
#10-fold CV:
k<-10
set.seed(1)
folds<-sample(rep(1:k, length=nrow(heart_dat)))
cv.errors<-matrix(NA, k, 15, dimnames = list(NULL, paste(1:15)))

for(i in 1:k){
  bss.fit<-regsubsets(HeartDisease~., data=heart_dat[folds!=i,], nvmax=15)
  for(j in 1:15){
    preds<-predict.regsubsets(bss.fit, heart_dat[folds==i,], id=j)
```

```r
    cv.errors[i,j]<-mean((heart_dat$HeartDisease[folds==i]-preds)^2)
  }
}
mean.cv.errors<-apply(cv.errors, 2, mean)
plot(mean.cv.errors, pch=19, type="b")+
  points(which.min(mean.cv.errors), mean.cv.errors[which.min(mean.cv.errors)],
         col=2, cex=2)
```




```
## integer(0)
```

The lowest CV error is achieved by a model of size 11.

No we need to ask ourselves whether interpretability is important here. Because if it is not we should aim for the smallest error, without caring for the number of parameters.

Let's compare the CV erros of sizes 6, 7, 10 and 12.

```r
mean.cv.errors[12]
```

```
##        12
## 0.1091016
```

```r
mean.cv.errors[10]
```

```
##       10
## 0.109107
```

```r
mean.cv.errors[7]
```

```
##         7
## 0.1123142
```

```r
mean.cv.errors[6]
```

```
##         6
## 0.1159018
```

As said, the error gets worse the more we drop predictors, and 6 predictors is much simpler than 12, however for our goal which one is better? Accuracy or interpretability?

Let's check the 12 and 7 variable model picked by BSS:

```
coef(bss.fit1,12)
```

```
##      (Intercept)              Age             SexM ChestPainTypeATA 
##     0.4625159154     0.0024808827     0.1579767762    -0.2486525432 
## ChestPainTypeNAP  ChestPainTypeTA      Cholesterol        FastingBS 
##    -0.2291333133    -0.1893084841    -0.0004956440     0.1315578915 
##            MaxHR  ExerciseAnginaY          Oldpeak     ST_SlopeFlat 
##    -0.0005891636     0.1359175980     0.0502584261     0.1611650688 
##       ST_SlopeUp 
##    -0.2125559655
```

```
coef(bss.fit1,7)
```

```
##      (Intercept)             SexM ChestPainTypeATA ChestPainTypeNAP 
##     0.6831816536     0.1597063670    -0.2492698622    -0.2183257397 
##      Cholesterol        FastingBS  ExerciseAnginaY       ST_SlopeUp 
##    -0.0004747709     0.1332335080     0.1889644545    -0.4037395302
```

Gotta decide if I go with 12 or 7.