assignment2

shuo wang

06/11/2021

```
library(kableExtra)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v readr   2.0.2
```

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

```
library(forcats)
heart_tbl<-read_csv("heart.csv")
```

```
## Rows: 918 Columns: 12
```

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldp
28	M	ATA	130	132	0	LVH	185	N	
29	M	ATA	120	243	0	Normal	160	N	
29	M	ATA	140	263	0	Normal	170	N	
29	M	ATA	130	204	0	LVH	202	N	
30	F	TA	170	237	0	ST	170	N	
31	M	ASY	120	270	0	Normal	153	Y	
31	F	ATA	100	219	0	ST	150	N	
32	M	ATA	125	254	0	Normal	155	N	
32	M	ATA	110	225	0	Normal	184	N	
32	M	ASY	118	529	0	Normal	130	N	

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Sex, ChestPainType, RestingECG, ExerciseAngina, ST_Slope
## dbl (7): Age, RestingBP, Cholesterol, FastingBS, MaxHR, Oldpeak, HeartDisease

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(heart_tbl)
```

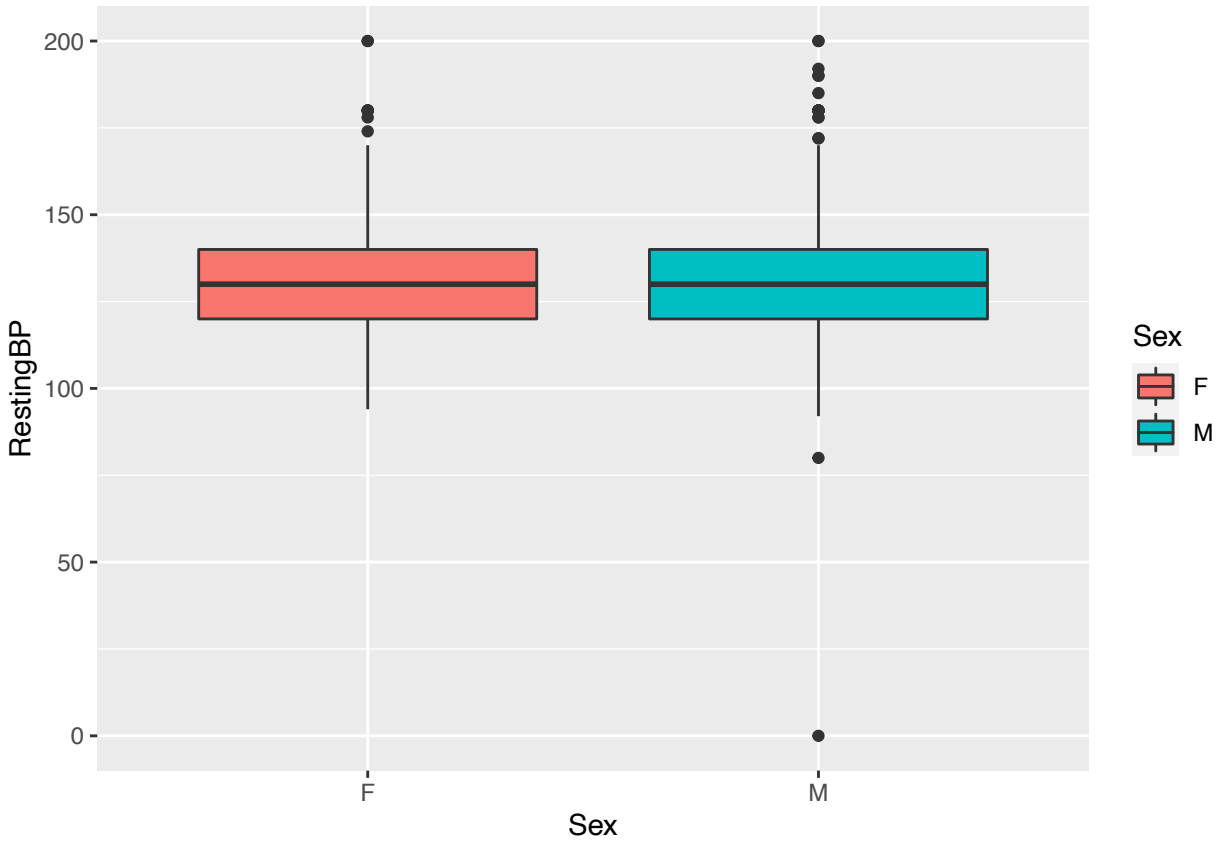
```
## Rows: 918
## Columns: 12
## $ Age            <dbl> 40, 49, 37, 48, 54, 39, 45, 54, 37, 48, 37, 58, 39, 49,~
## $ Sex            <chr> "M", "F", "M", "F", "M", "M", "F", "M", "M", "F", "F", ~
## $ ChestPainType  <chr> "ATA", "NAP", "ATA", "ASY", "NAP", "NAP", "ATA", "ATA",~
## $ RestingBP      <dbl> 140, 160, 130, 138, 150, 120, 130, 110, 140, 120, 130, ~
## $ Cholesterol    <dbl> 289, 180, 283, 214, 195, 339, 237, 208, 207, 284, 211, ~
## $ FastingBS      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ RestingECG     <chr> "Normal", "Normal", "ST", "Normal", "Normal", "Normal",~
## $ MaxHR          <dbl> 172, 156, 98, 108, 122, 170, 170, 142, 130, 120, 142, 9~
## $ ExerciseAngina <chr> "N", "N", "N", "Y", "N", "N", "N", "N", "Y", "N", "N", ~
## $ Oldpeak        <dbl> 0.0, 1.0, 0.0, 1.5, 0.0, 0.0, 0.0, 0.0, 1.5, 0.0, 0.0, ~
## $ ST_Slope       <chr> "Up", "Flat", "Up", "Flat", "Up", "Up", "Up", "Up", "Fl~
## $ HeartDisease   <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1~
```

```
heart_tbl %>% arrange(Age) %>% head(10) %>% kable %>% kable_styling()
```

a. Using the plot(s) of your choice, assess whether there is an association between the sex of the patient and their resting blood pressure, i.e. is there a difference in distribution of the resting heart rates across the sexes? Explain your answer.

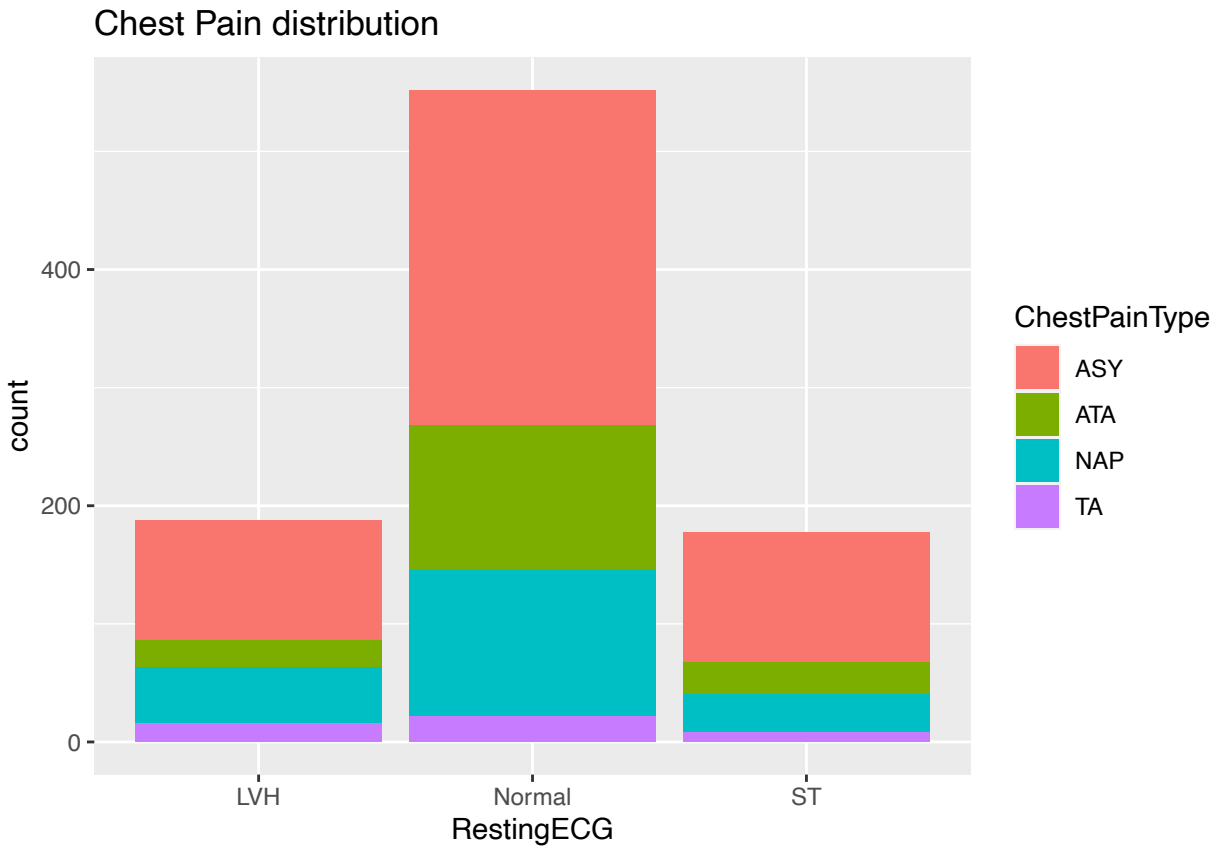
ans: from the box-plot below, there is no significant association between the sex of the patient and their resting blood pressure

```
 A<-heart_tbl %>% select(Sex, RestingBP) %>% drop_na()
ggplot(data = A, aes(x=Sex,y=RestingBP)) + geom_boxplot(aes(fill=Sex))
```



b. Produce a stacked barplot showing the distribution of Chest Pain Type for each level of RestingECG.

```
B<- heart_tbl %>% group_by(ChestPainType) %>% select(ChestPainType, RestingECG)
ggplot(data=B, mapping=aes(x=RestingECG, fill=ChestPainType))+geom_bar()+labs(title="Chest Pain distri
```



c. Produce a summary table containing counts and proportions of RestingECG category for each sex/ChestPainType factor combination.

```
heart_tbl %>% group_by(RestingECG) %>% count(Sex, ChestPainType, sort = TRUE) %>% mutate(proportion =
```

d.Create a summary table that finds the **mean, median and IQR** of RestingBP, Cholesterol, FastingBS, and MaxHR for each of the Chest Pain Types and report those results in a tibble where the columns are the levels of Chest Pain Types and the summary statistics are in the rows.

```
D<- heart_tbl %>% group_by(ChestPainType) %>%select(RestingBP, Cholesterol,FastingBS,MaxHR) %>% summar
  list(mean = ~mean(.), median = ~median(.), IQRs = ~IQR(.)),
  na.rm=TRUE)
```

```
## Adding missing grouping variables: `ChestPainType`
```

```
 t(D) %>% kable %>% kable_styling()
```

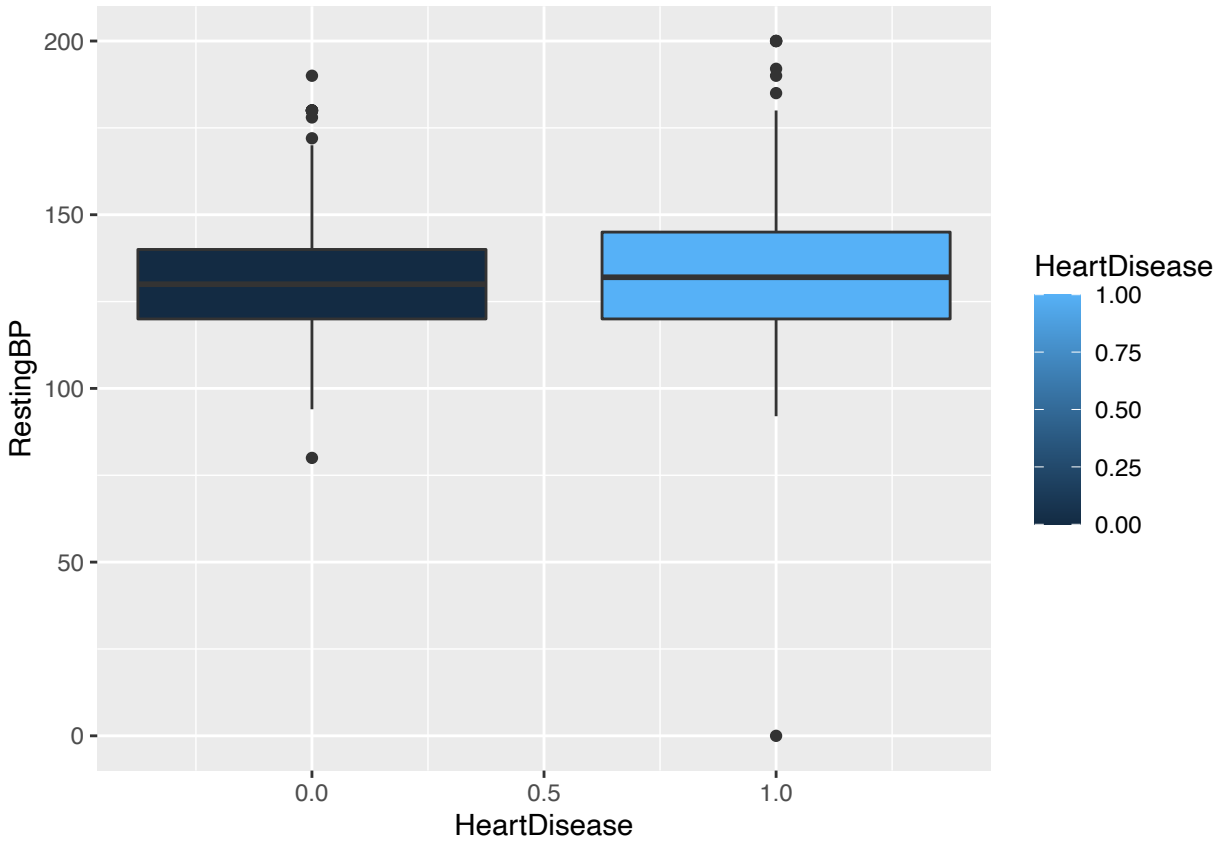
e.Using the plot(s) of your choice, explain which of the following measurements seem most strongly associated with Heart Disease (heart disease vs. normal) : RestingBP, Cholesterol, FastingBS, and MaxHR.

ANS:Fasting blood sugar seems have the most strong association with Heart Disease

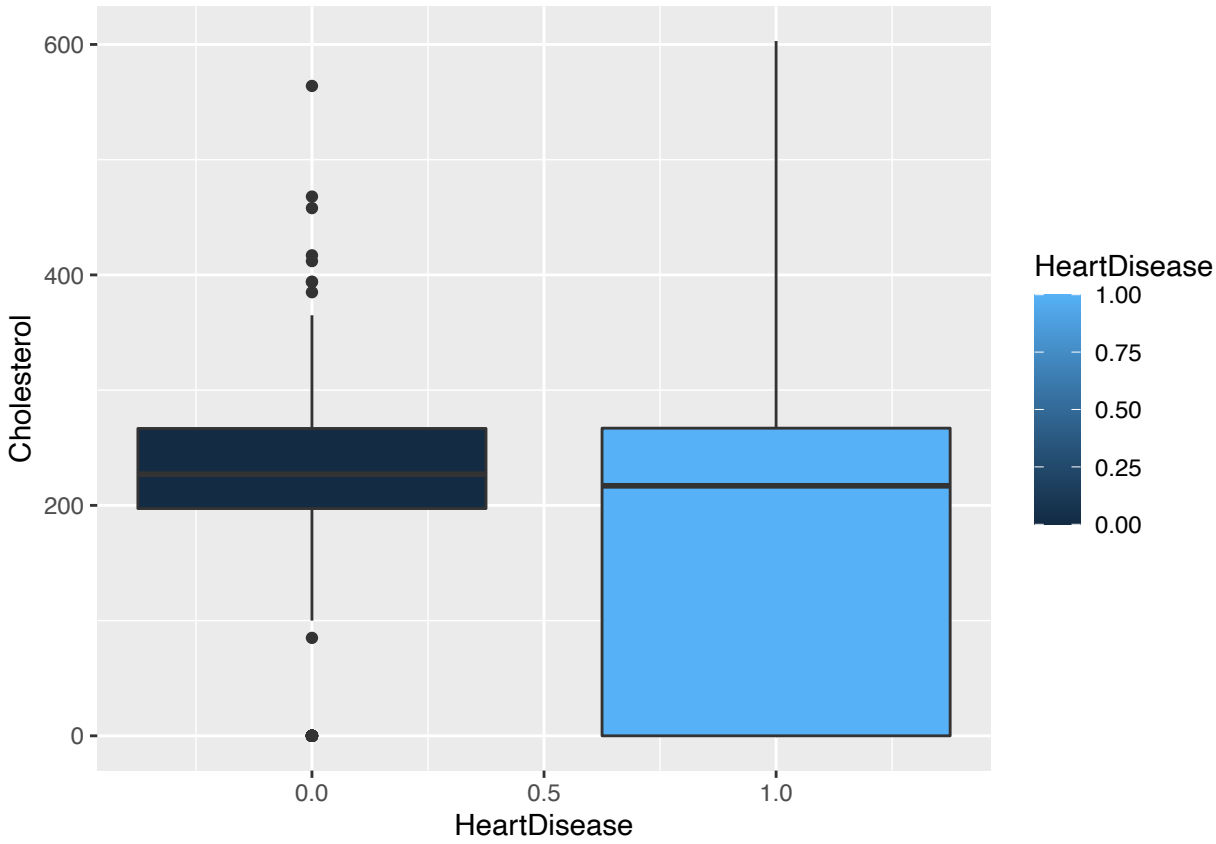
```
ggplot(data = heart_tbl, aes(x=HeartDisease,y=RestingBP,group=HeartDisease)) + geom_boxplot(aes(fill=H
```

RestingECG	Sex	ChestPainType	n	proportion
LVH	M	ASY	80	0.4255319
LVH	M	NAP	32	0.1702128
LVH	F	ASY	22	0.1170213
LVH	F	NAP	15	0.0797872
LVH	M	TA	15	0.0797872
LVH	M	ATA	14	0.0744681
LVH	F	ATA	9	0.0478723
LVH	F	TA	1	0.0053191
Normal	M	ASY	246	0.4456522
Normal	M	NAP	92	0.1666667
Normal	M	ATA	81	0.1467391
Normal	F	ATA	42	0.0760870
Normal	F	ASY	38	0.0688406
Normal	F	NAP	31	0.0561594
Normal	M	TA	15	0.0271739
Normal	F	TA	7	0.0126812
ST	M	ASY	100	0.5617978
ST	M	NAP	26	0.1460674
ST	M	ATA	18	0.1011236
ST	F	ASY	10	0.0561798
ST	F	ATA	9	0.0505618
ST	F	NAP	7	0.0393258
ST	M	TA	6	0.0337079
ST	F	TA	2	0.0112360

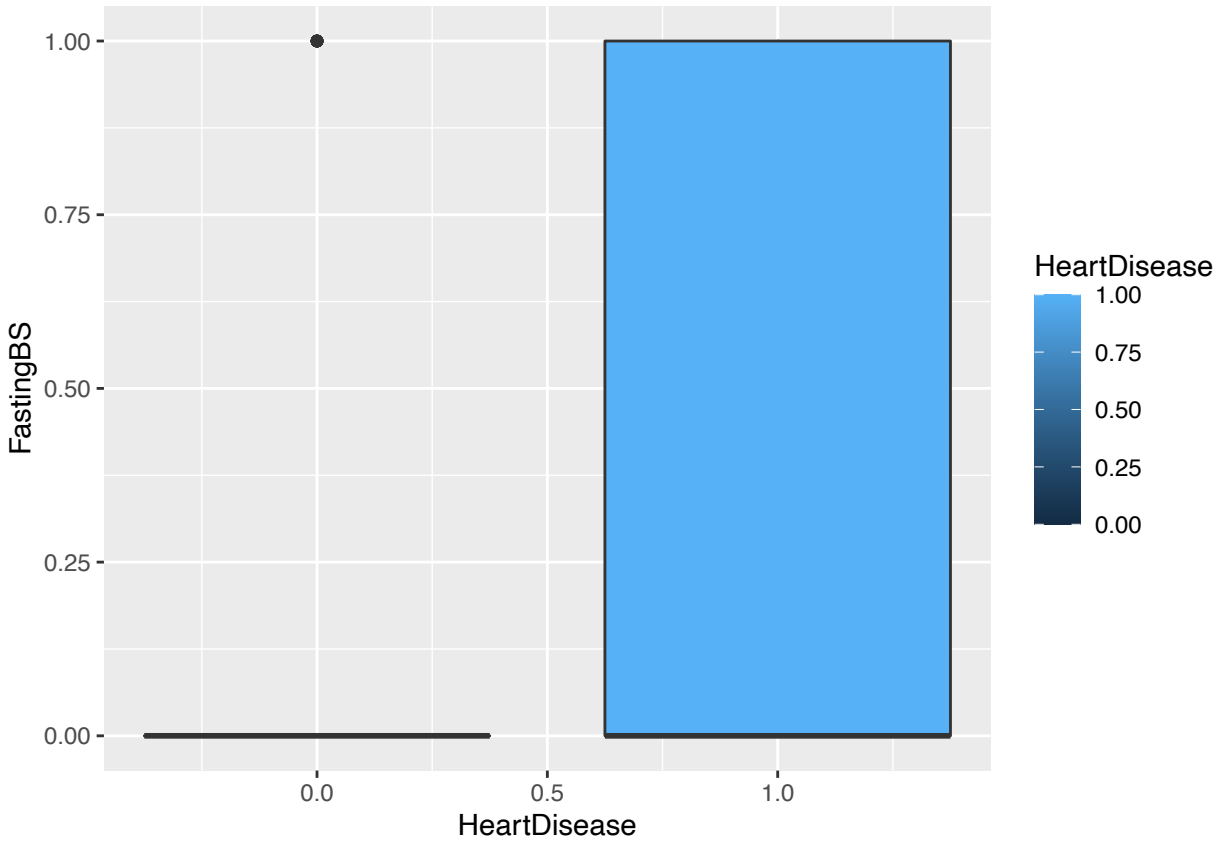
ChestPainType	ASY	ATA	NAP	TA
RestingBP_mean	133.2298	130.6243	130.9606	136.4130
Cholesterol_mean	186.6452	233.0462	197.4384	207.0652
FastingBS_mean	0.2842742	0.1098266	0.2019704	0.2826087
MaxHR_mean	128.4778	150.2081	143.2365	147.8913
RestingBP_median	130	130	130	140
Cholesterol_median	220.5	237.0	218.0	229.0
FastingBS_median	0	0	0	0
MaxHR_median	128	152	147	145
RestingBP_IQRs	23.00	20.00	20.00	27.25
Cholesterol_IQRs	268.25	70.00	76.50	74.75
FastingBS_IQRs	1	0	0	1
MaxHR_IQRs	32.0	28.0	39.5	35.5



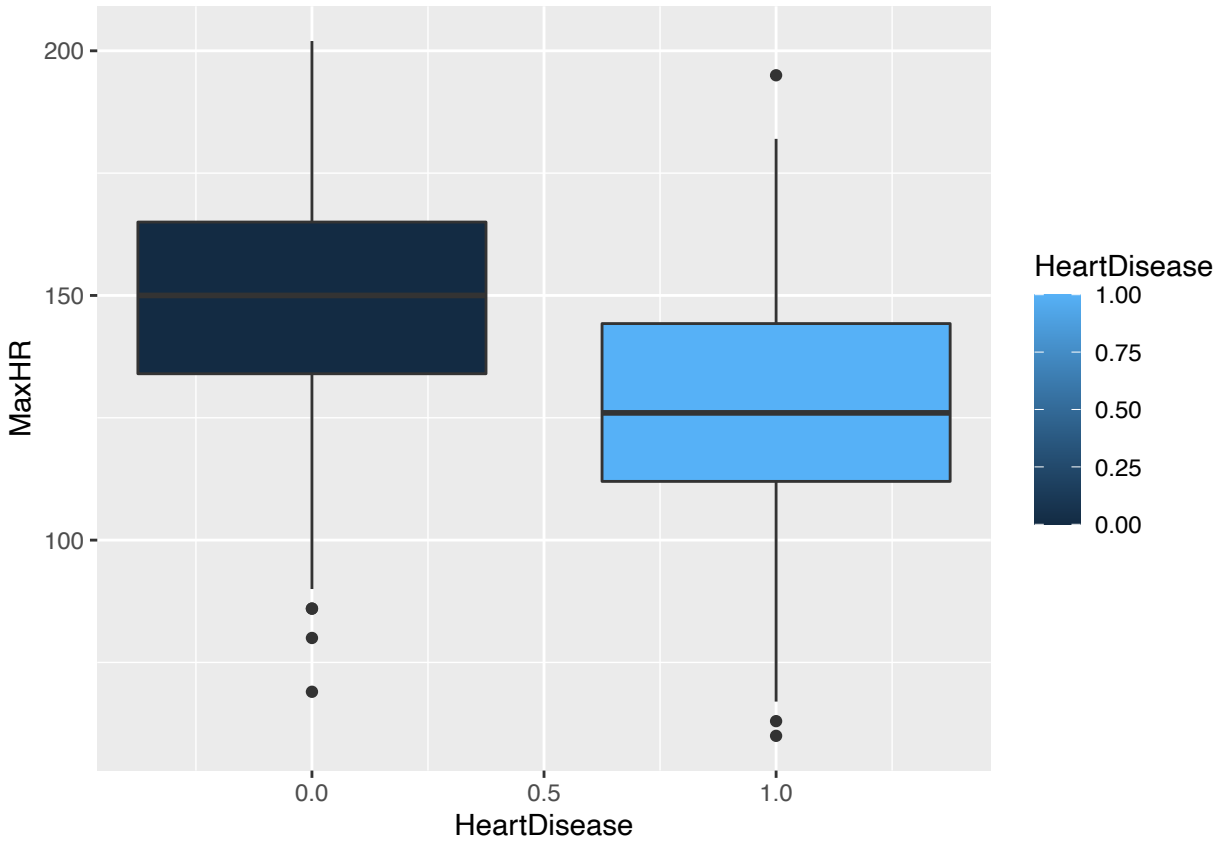

```
ggplot(data = heart_tbl, aes(x=HeartDisease,y=Cholesterol,group=HeartDisease)) + geom_boxplot(aes(fill
```




```
ggplot(data = heart_tbl, aes(x=HeartDisease,y=FastingBS,group=HeartDisease)) + geom_boxplot(aes(fill=H
```



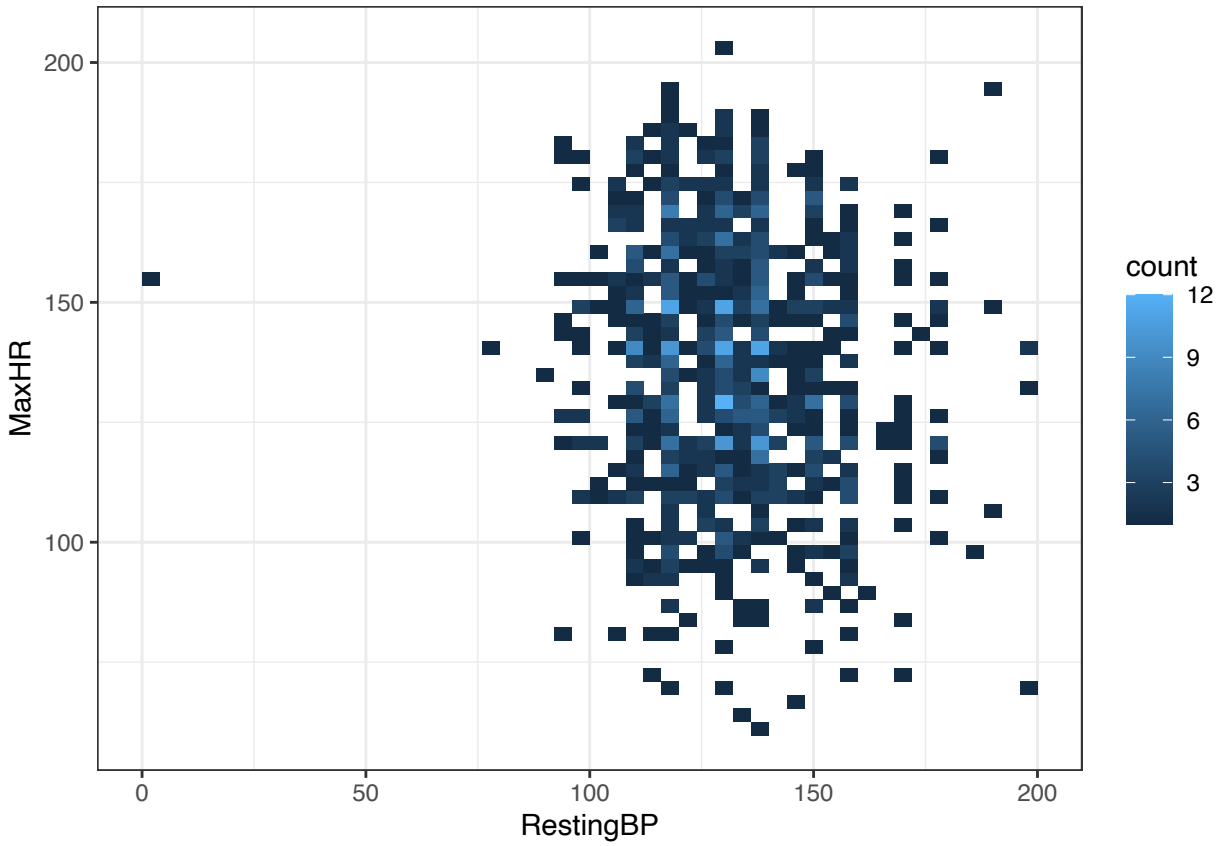

```
ggplot(data = heart_tbl, aes(x=HeartDisease,y=MaxHR,group=HeartDisease)) + geom_boxplot(aes(fill=Heart
```



f.Create both a 2-d histogram and a 2-d contour plot to assess the association betwen RestingBP and MaxHR. Describe this association and also explain which plot you think shows the association most clearly (or explain why they are about the same).

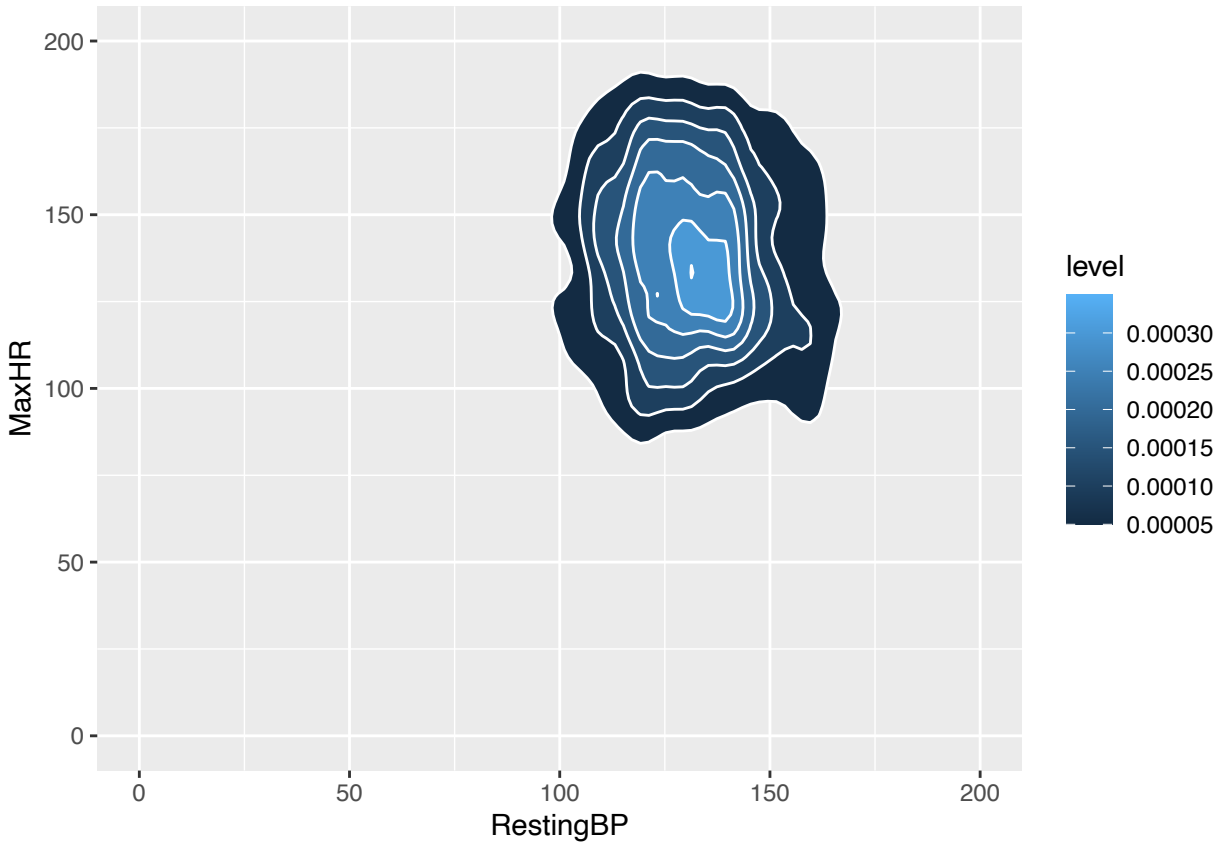
ANS: the contour plot shows more clearly that most people with MAX heart rate around 125 to150 ,RestingBP 125 to 140

```
ggplot(heart_tbl,aes(x=RestingBP, y=MaxHR))+geom_bin2d(bins = 50 )+theme_bw()
```




```
ggplot(heart_tbl,aes(x=RestingBP, y=MaxHR))+stat_density_2d(aes(fill = ..level..), geom = "polygon", c
```

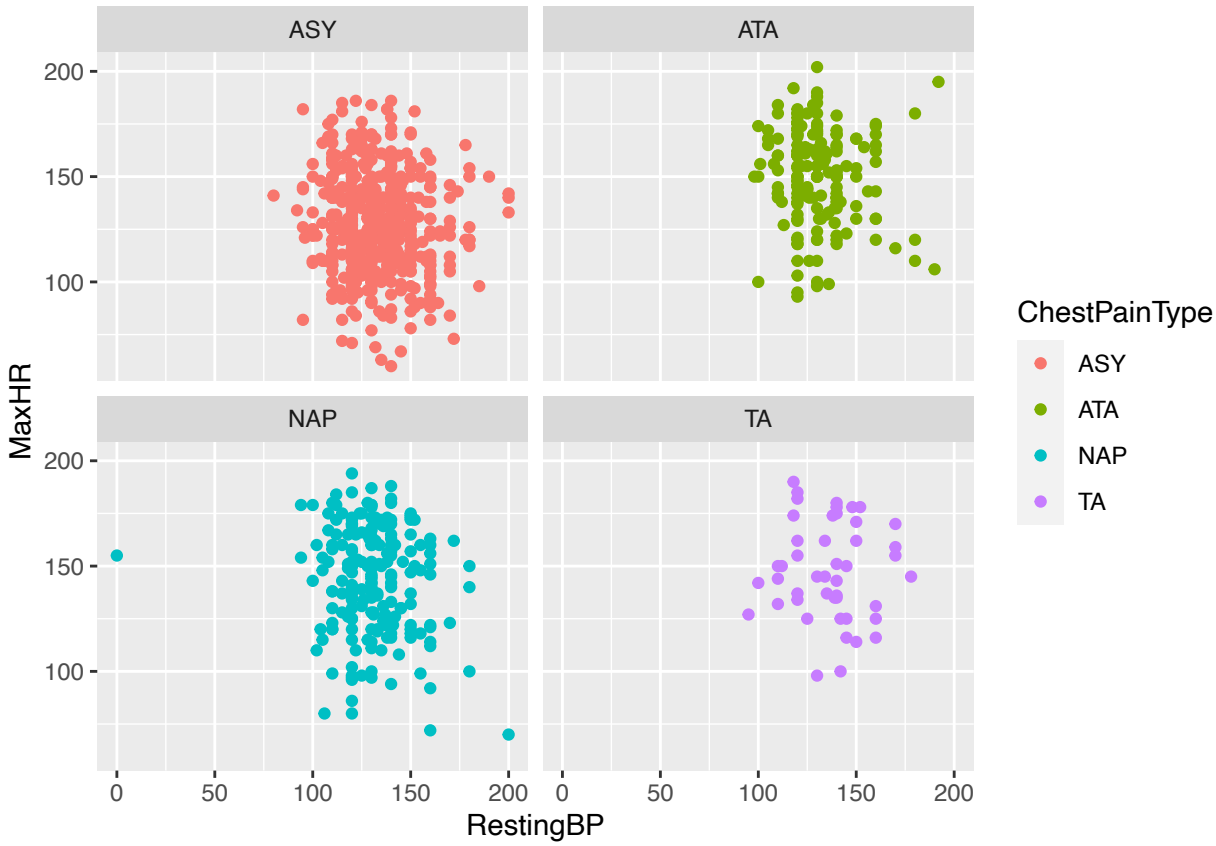
```
## Warning: Removed 1 rows containing non-finite values (stat_density2d).
```



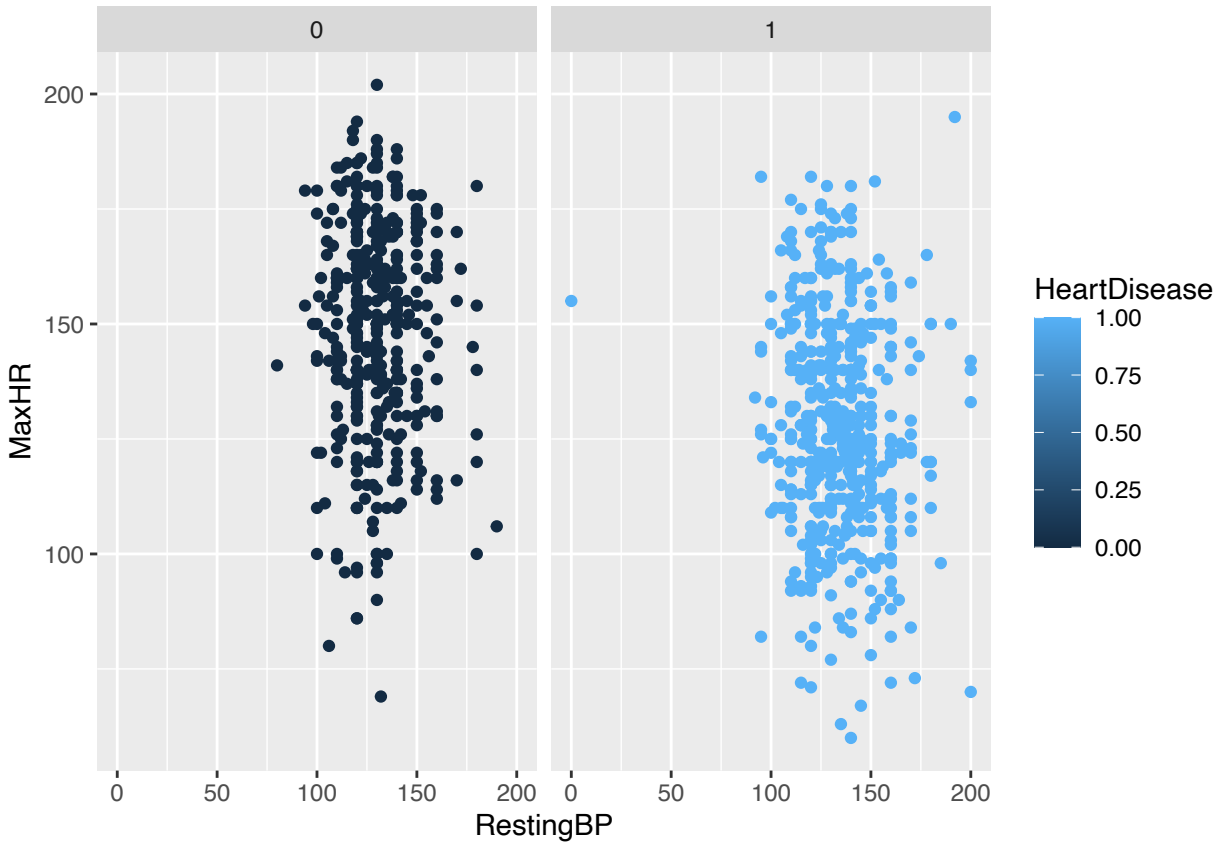
g.Using the plot(s) of your choice, determine whether the association in (f) depends on either the Chest Pain Type or the Heart Disease status (or both).

ANS: from plot2, people who have heart disease have higher max heart rate. however, in plot1, cant find significant influence of the association with different type of chest pain.

```
ggplot(data = heart_tbl, aes(x=RestingBP,y=MaxHR,group=ChestPainType,col=ChestPainType)) + geom_point(
```




```
ggplot(data = heart_tbl, aes(x=RestingBP,y=MaxHR,group=HeartDisease,col=HeartDisease)) + geom_point()+
```


0
1
200
150
100
MaxHR
0
50
100
150
200
RestingBP
HeartDisease
1.00
0.75
0.50
0.25
0.00