Math 208: Analysis of Heart Disease

Lily Samuel

2021-11-07

```
library(languageserver)
library(readr)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v stringr   1.5.0
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(knitr)

heart <- read_csv("/Users/lilysamuel/Desktop/heart 2.csv")
```

```
## Rows: 918 Columns: 12
## -- 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.
```

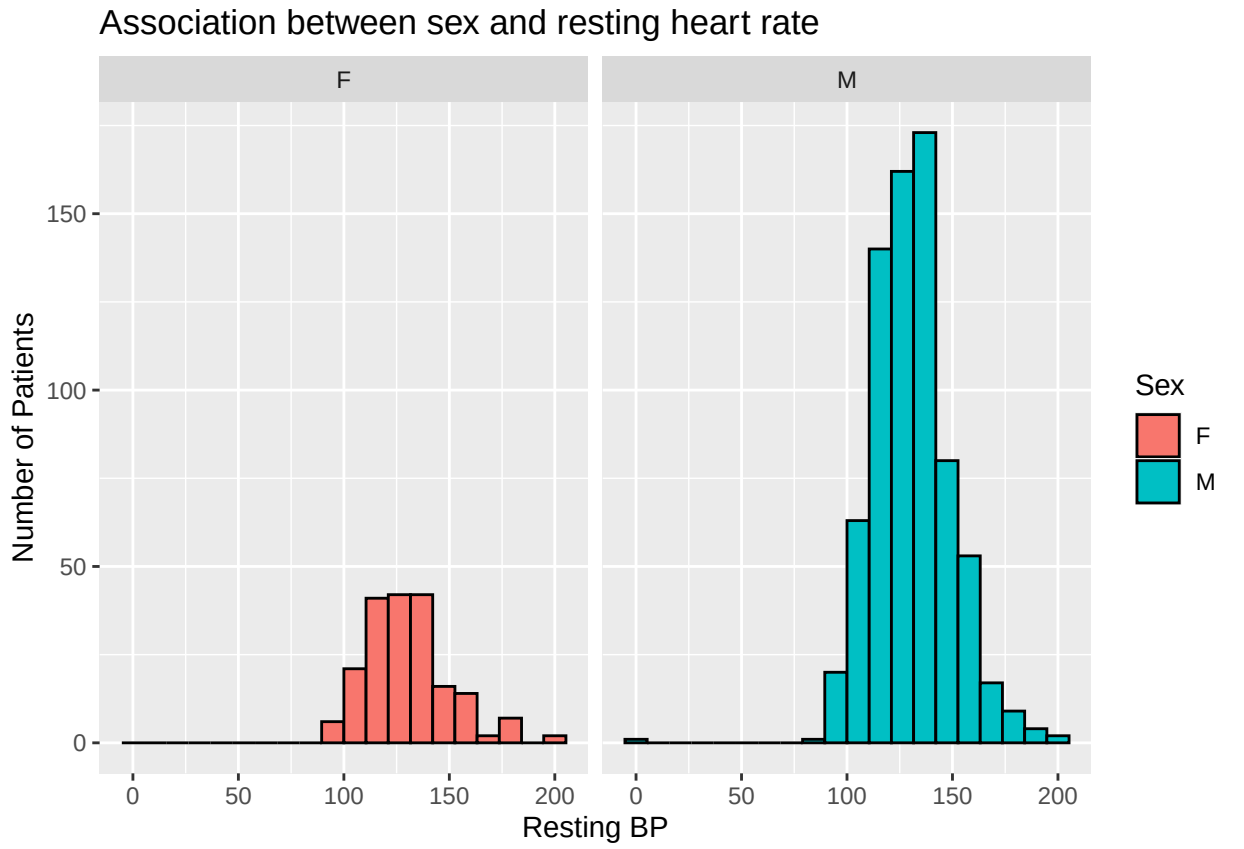
```
head(heart)
```

```
## # A tibble: 6 x 12
##     Age Sex   ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
##   <dbl> <chr> <chr>              <dbl>       <dbl>     <dbl> <chr>      <dbl>
## 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
## 6    39 M     NAP                  120         339         0 Normal       170
## # i 4 more variables: ExerciseAngina <chr>, Oldpeak <dbl>, ST_Slope <chr>,
## #   HeartDisease <dbl>
```

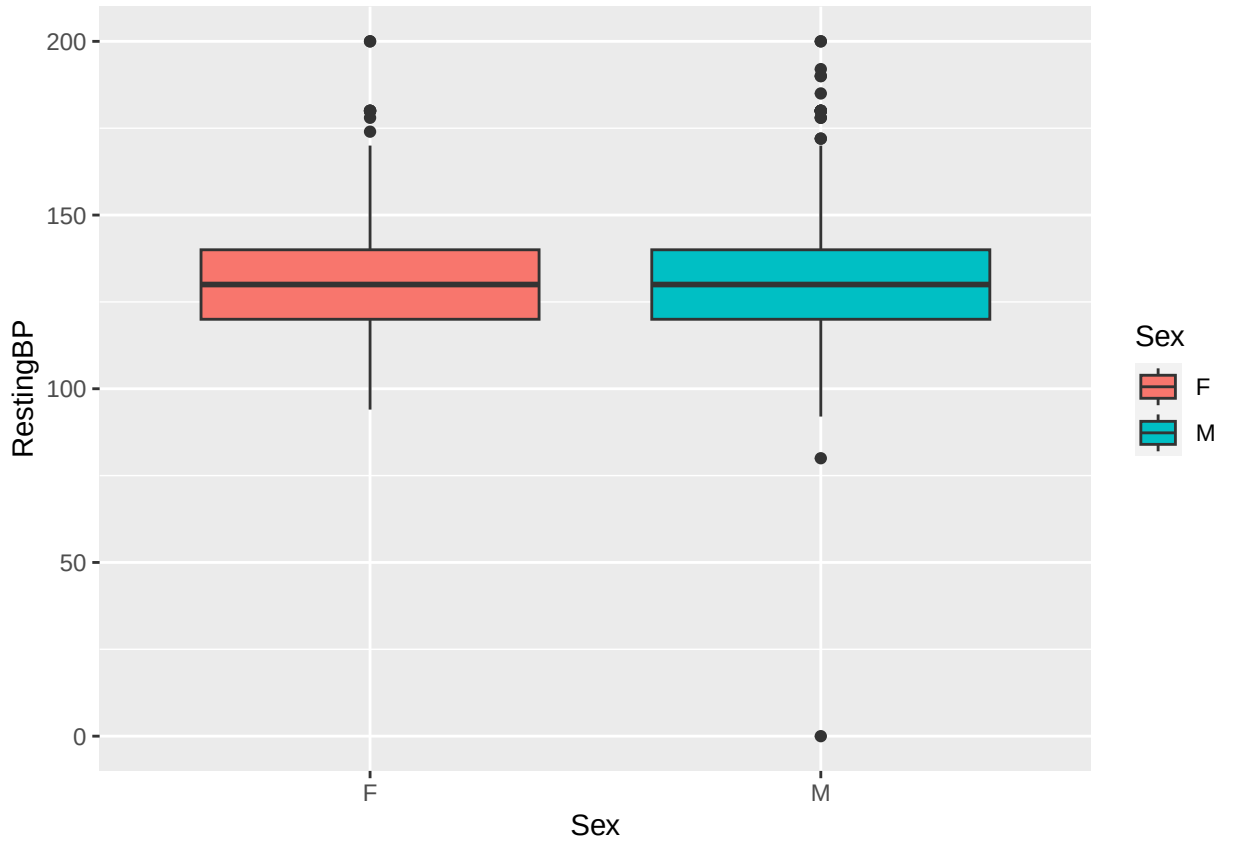
Task 1: Assess whether there is an association between the sex of the patient and their resting heart rates, i.e. is there a difference in distribution of the resting heart rates across the sexes?

Solution:

```
ggplot(heart, aes(x=RestingBP, group=Sex,fill= Sex)) +
  geom_histogram(bins=20,col="black") + facet_wrap(~Sex)+
  labs(x="Resting BP",y="Number of Patients",
       title="Association between sex and resting heart rate")
```




```
ggplot(heart, aes(x=Sex, y=RestingBP, fill=Sex)) + geom_boxplot()
```

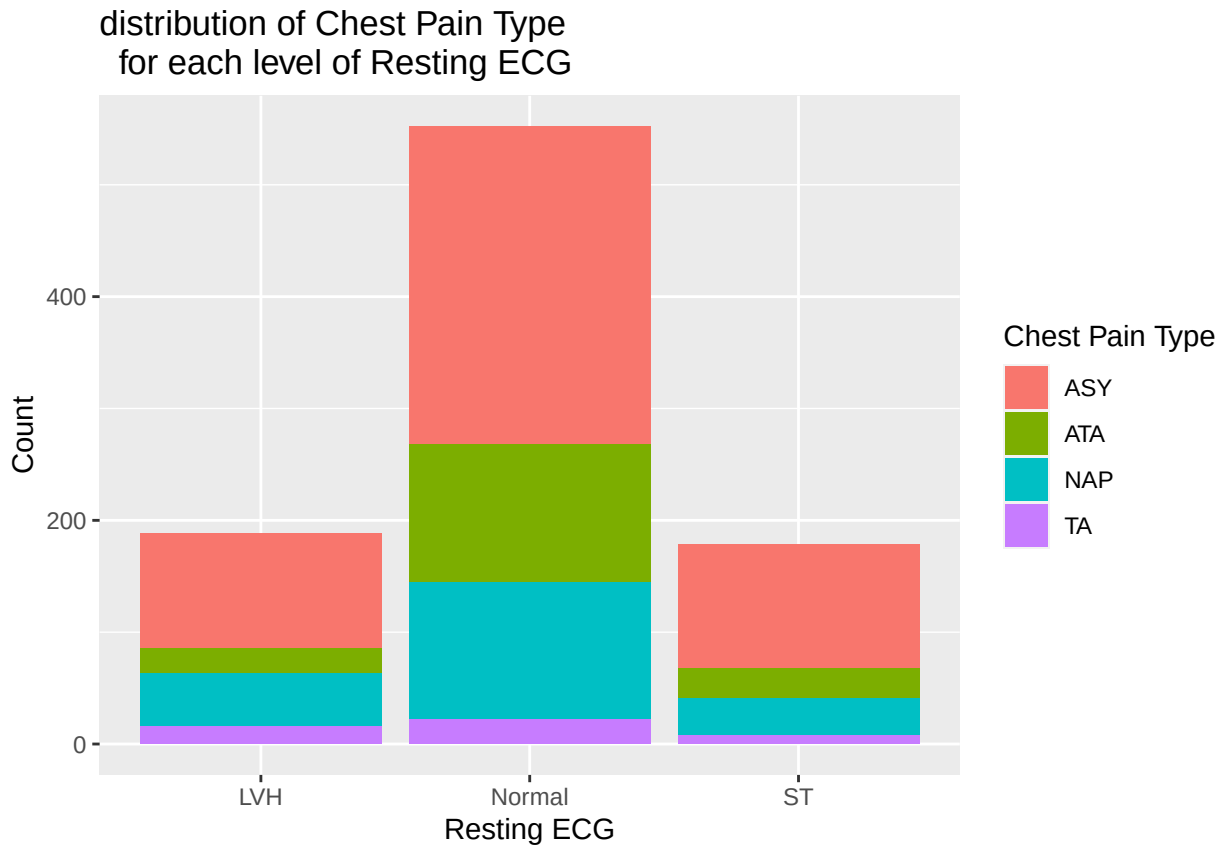


The distribution of Resting BP between male and female is very similar, which is an indication that there is no association between the sexes and Resting BP

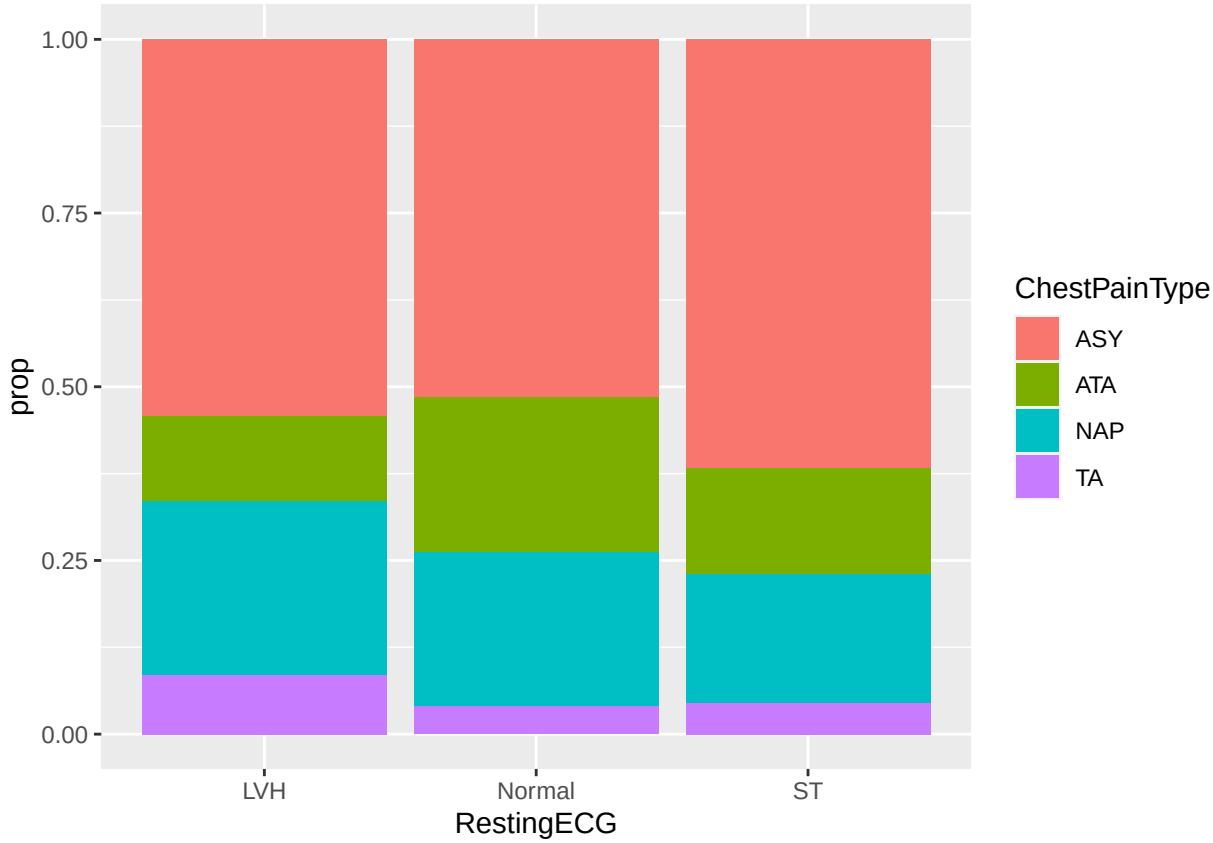
Task 2: Produce a stacked barplot showing the distribution of Chest Pain Type for each level of RestingECG

Solution:

```
ggplot(heart,aes(x=RestingECG,fill=ChestPainType)) +
  geom_bar() + labs(x="Resting ECG",y="Count",
  title="distribution of Chest Pain Type
  for each level of Resting ECG",
  fill="Chest Pain Type")
```




```
ggplot(heart %>% count(ChestPainType,RestingECG) %>%
    group_by(RestingECG) %>%
    reframe(ChestPainType=ChestPainType,prop=n/sum(n)),
  aes(y=prop,x=RestingECG,fill=ChestPainType)) +
geom_bar(stat="identity")
```



Task 3: Produce a summary table containing counts and proportions of RestingECG category for each sex/ChestPainType factor combination.

Solution:

```
summary_table<-heart%>%mutate(ChestPainTypeImp=fct_explicit_na(ChestPainType))%>%
  group_by(Sex, ChestPainTypeImp, RestingECG)%>%
  count() %>%
  group_by(RestingECG) %>%
  mutate(prop = n/sum(n))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `ChestPainTypeImp = fct_explicit_na(ChestPainType)`.
## Caused by warning:
## ! `fct_explicit_na()` was deprecated in forcats 1.0.0.
## i Please use `fct_na_value_to_level()` instead.
```

```
summary_table%>% slice(sample(1:nrow(.), 15))
```

```
## # A tibble: 17 x 5
## # Groups:   RestingECG [3]
##    Sex   ChestPainTypeImp RestingECG     n   prop
##    <chr> <fct>            <chr>      <int>  <dbl>
##  1 M     ASY              LVH           80 0.426
##  2 F     ASY              LVH           22 0.117
##  3 F     ATA              LVH            9 0.0479
```

```
##  4 F     NAP             LVH           15 0.0798
##  5 M     TA              LVH           15 0.0798
##  6 F     ATA             Normal        42 0.0761
##  7 F     ASY             Normal        38 0.0688
##  8 M     NAP             Normal        92 0.167
##  9 M     ASY             Normal       246 0.446
## 10 M     ATA             Normal        81 0.147
## 11 F     NAP             Normal        31 0.0562
## 12 M     TA              Normal        15 0.0272
## 13 F     ASY             ST            10 0.0562
## 14 M     TA              ST             6 0.0337
## 15 F     NAP             ST             7 0.0393
## 16 F     ATA             ST             9 0.0506
## 17 M     ASY             ST           100 0.562
```

Task 4:
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.

Solution:

```
heart %>% group_by(ChestPainType) %>%
  summarise_at(vars(RestingBP,Cholesterol,FastingBS,MaxHR),
  list(mean=mean, median=median, IQR=IQR)) %>%
  pivot_longer(cols=c(ends_with("mean"),ends_with("median"), ends_with("IQR")),
  names_to="Var_Statistic")%>%
  pivot_wider(id_cols="Var_Statistic",names_from="ChestPainType")
```

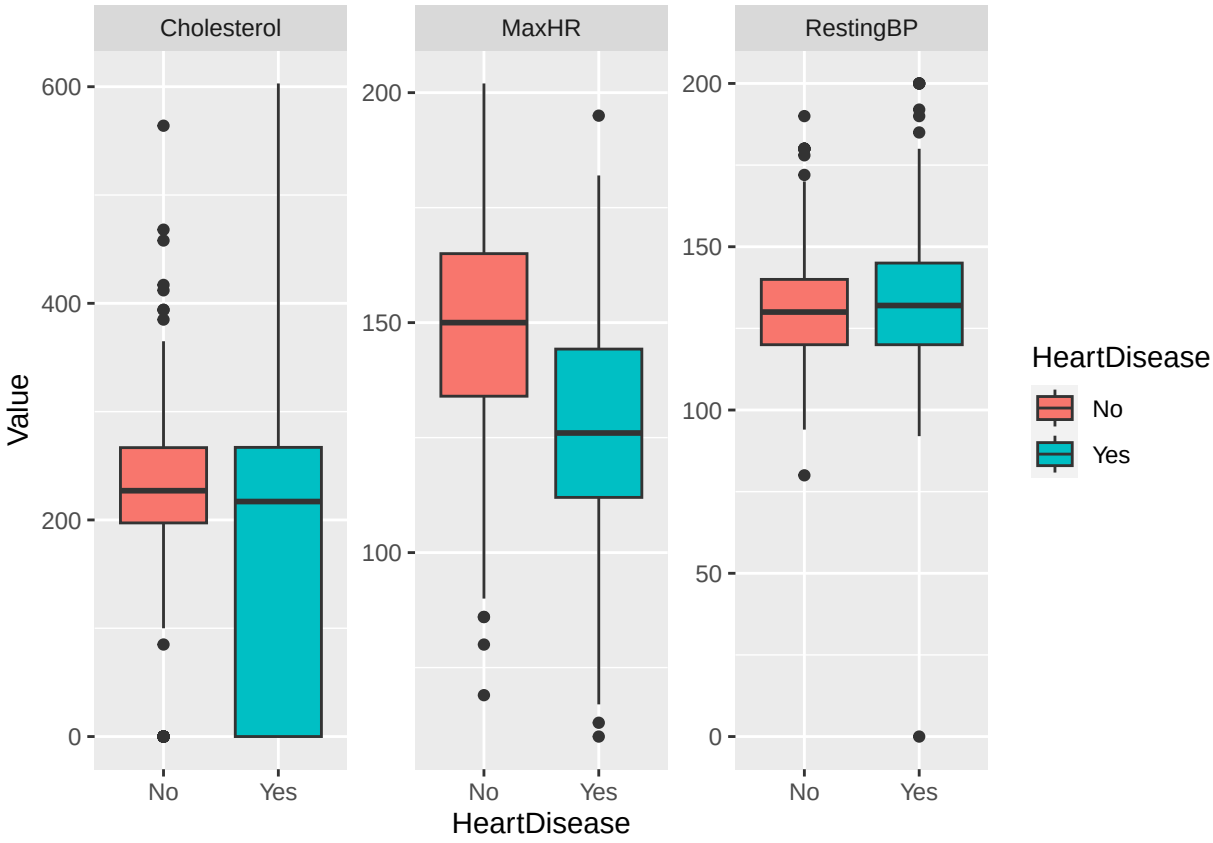
```
## # A tibble: 12 x 5
##    Var_Statistic          ASY     ATA     NAP      TA
##    <chr>                <dbl>   <dbl>   <dbl>   <dbl>
##  1 RestingBP_mean     133.    131.    131.    136.
##  2 Cholesterol_mean   187.    233.    197.    207.
##  3 FastingBS_mean       0.284   0.110   0.202   0.283
##  4 MaxHR_mean         128.    150.    143.    148.
##  5 RestingBP_median   130     130     130     140
##  6 Cholesterol_median 220.    237     218     229
##  7 FastingBS_median     0       0       0       0
##  8 MaxHR_median       128     152     147     145
##  9 RestingBP_IQR       23      20      20      27.2
## 10 Cholesterol_IQR    268.     70      76.5    74.8
## 11 FastingBS_IQR        1       0       0       1
## 12 MaxHR_IQR           32      28      39.5    35.5
```

Task 5: Using plots, explain which of the following measurements seem most strongly associated with Heart Disease (heart disease vs. normal) : RestingBP, Cholesterol, FastingBS, and MaxHR.

```
quant_data<-heart %>%
 select(HeartDisease,RestingBP,Cholesterol,MaxHR) %>%
 pivot_longer(cols=RestingBP:MaxHR,values_to="Value") %>%
 mutate(HeartDisease=ifelse(HeartDisease==1,"Yes","No"))
head(quant_data)
```

```
## # A tibble: 6 x 3
##   HeartDisease name        Value
##   <chr>        <chr>       <dbl>
## 1 No           RestingBP     140
## 2 No           Cholesterol   289
## 3 No           MaxHR         172
## 4 Yes          RestingBP     160
## 5 Yes          Cholesterol   180
## 6 Yes          MaxHR         156
```

```
ggplot(quant_data, aes(x=HeartDisease,
                       y=Value,
                       fill=HeartDisease,
                       group=HeartDisease)) +
  geom_boxplot() + facet_wrap(~name,scales="free")
```

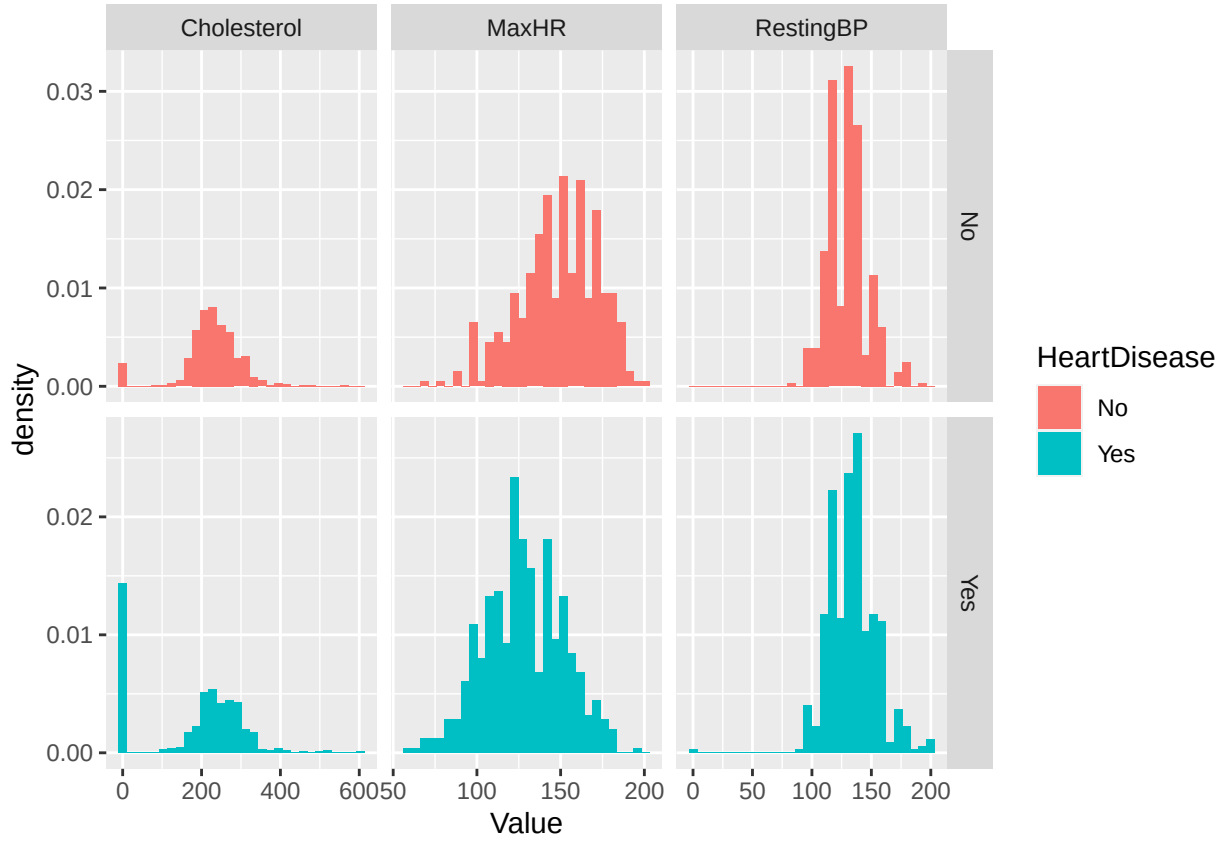



```
ggplot(quant_data, aes(x=Value,
                       fill=HeartDisease,
                       group=HeartDisease)) +
  geom_histogram(aes(y=..density..)) +
  facet_grid(HeartDisease~name,scales="free")
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
```

```
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

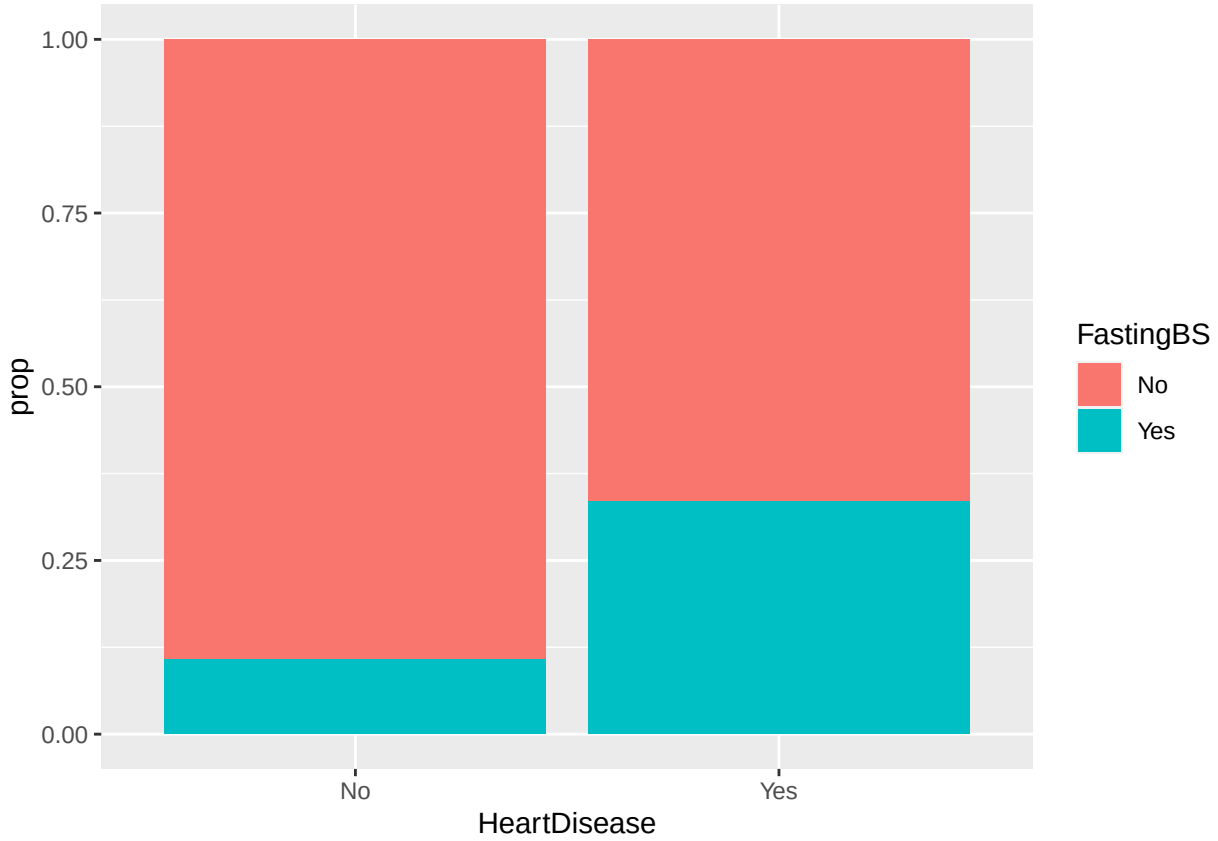
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



Based on the boxplot and bar graph above, the maximum heart rate is more strongly associated with heart disease that the other quantitative variables. This is evident as we can see the heart disease group has lower heart rates than the other group of those without heart disease. The distributions of the other two variables seem quite similar

```
catergorical_data<-ggplot(heart %>%
        mutate(FastingBS=ifelse(FastingBS==1,"Yes","No"),
               HeartDisease=ifelse(HeartDisease==1,"Yes","No")) %>%
        count(HeartDisease,FastingBS) %>%
        group_by(HeartDisease) %>%
        reframe(FastingBS=FastingBS,prop=n/sum(n)),
    aes(y=prop,x=HeartDisease,fill=FastingBS)) +
geom_bar(stat="identity")

catergorical_data
```

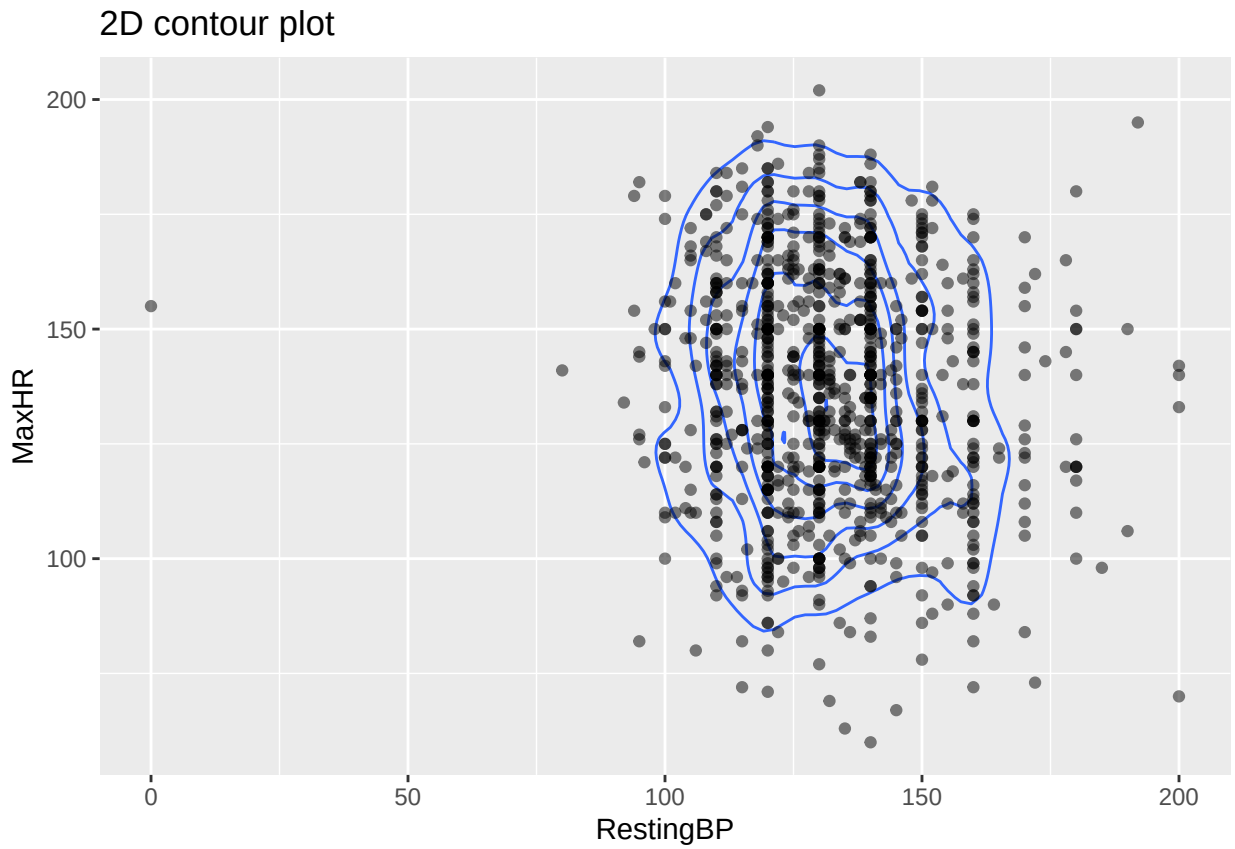


Heart Disease has a larger percentage of patients whose blood sugars were measured after fasting.

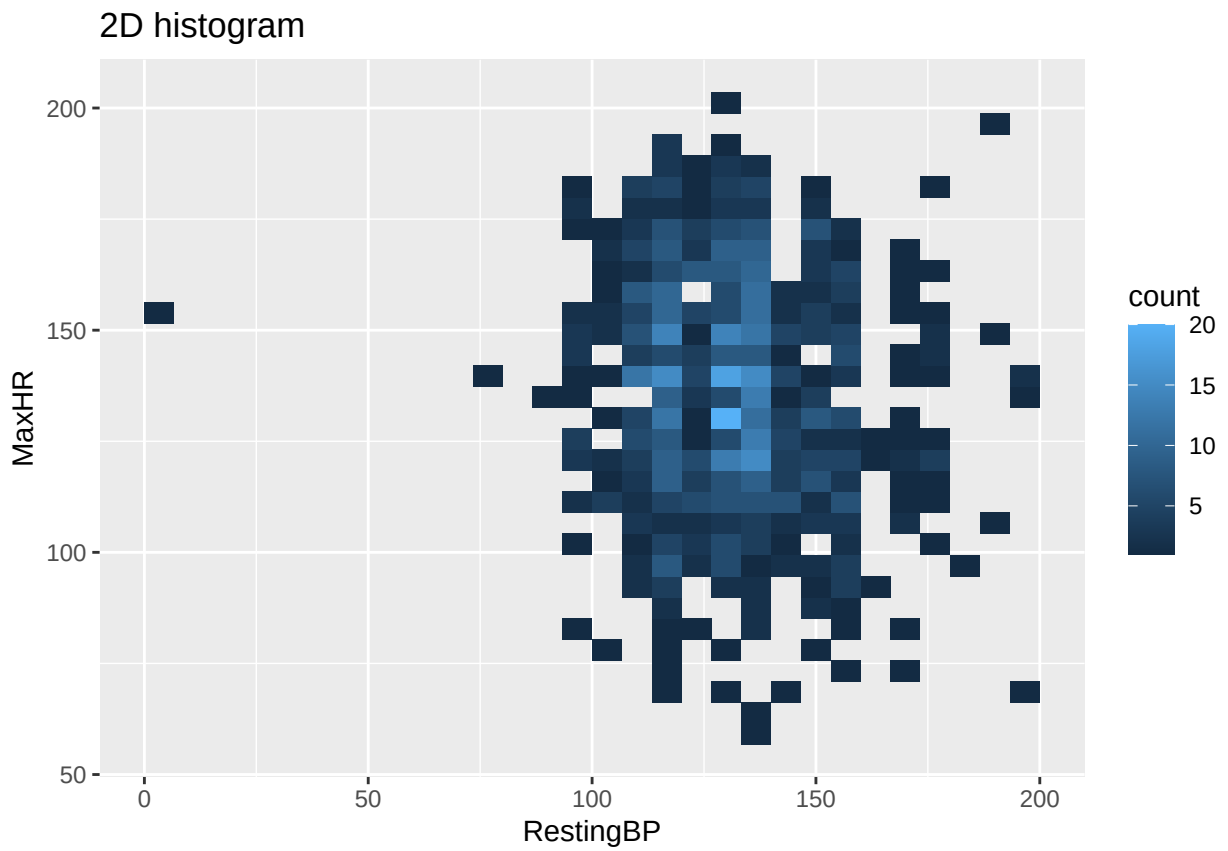
Task 6: Create both a 2-d histogram and a 2-d contour plot to assess the association between RestingBP and MaxHR. Describe this association and also explain which plot you think shows the association most clearly (or explain why they are about the same).

Solution:

```
ggplot(heart, aes(x=RestingBP,y=MaxHR)) + geom_density_2d()+ geom_point(alpha=0.5)+ ggtitle("2D contour
  theme(legend.position = "none")
```



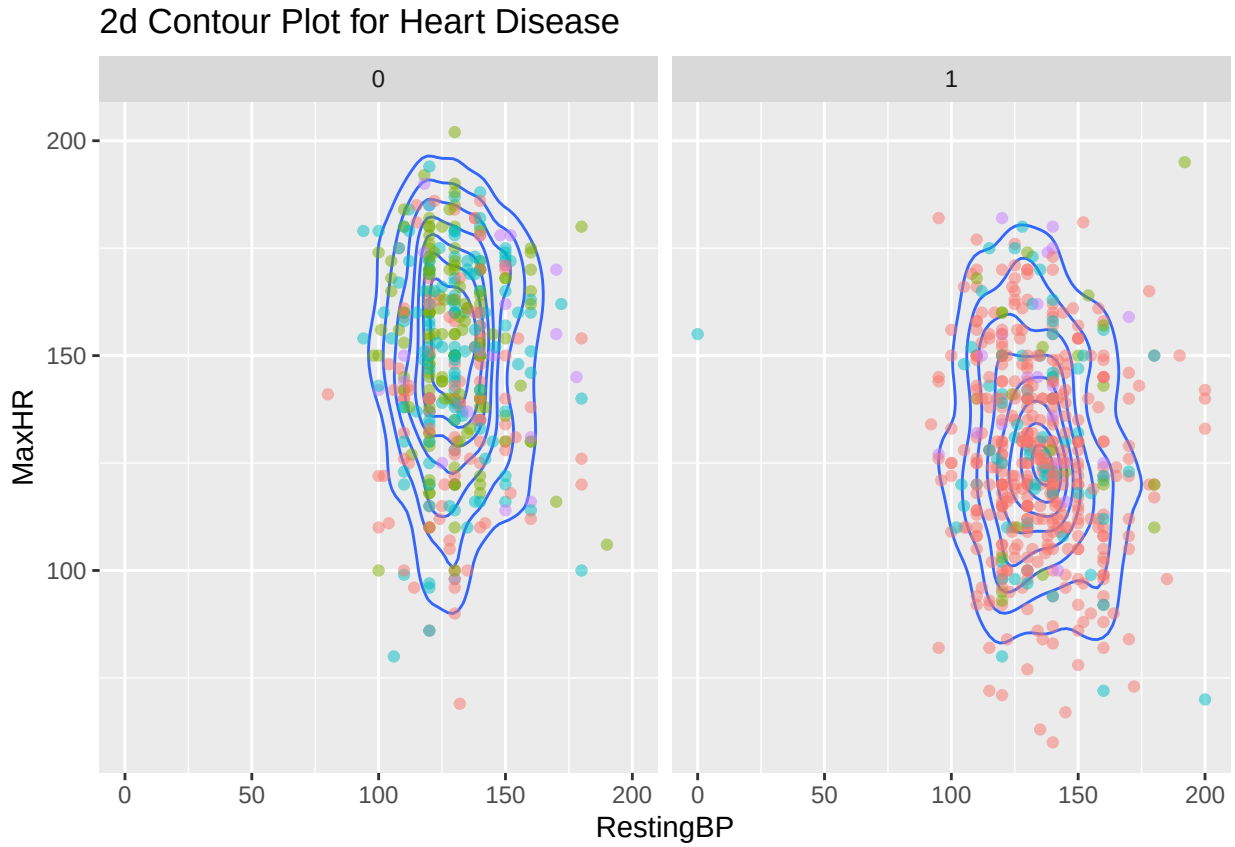

```
ggplot(heart, aes(x=RestingBP,y=MaxHR)) + geom_bin2d() + ggtitle("2D histogram")
```



Task 7: Determine whether the association between RestingBP and MaxHR depends on either the Chest Pain Type or the Heart Disease status (or both)

Solution

```
ggplot(heart, aes(x=RestingBP,y=MaxHR)) +
  geom_density2d()+
  geom_point(alpha=0.5,aes(col=ChestPainType))+
  facet_grid(~HeartDisease)+
  ggtitle("2d Contour Plot for Heart Disease") +
  theme(legend.position="none")
```




```
ggplot(heart, aes(x=RestingBP,y=MaxHR))+
  geom_density2d()+
  geom_point(alpha=0.5,aes(col=ChestPainType))+
  facet_grid(~ChestPainType)+
  ggtitle("2d Contour Plot for Chest Pain Type")+
  theme(legend.position="none")
```

Based on these figures, it seems like the association does depend on these two variables and their interaction.