Assignment 2 Question 1

The data for this assignment comes from a Kaggle project which collected data on heart disease from five different sources (https://www.kaggle.com/fedesoriano/heart-failure-prediction/version/1 (https://www.kaggle.com/fedesoriano/heart-failure-prediction/version/1)). The data can be downloaded at the link below. You can load the data with the following code:

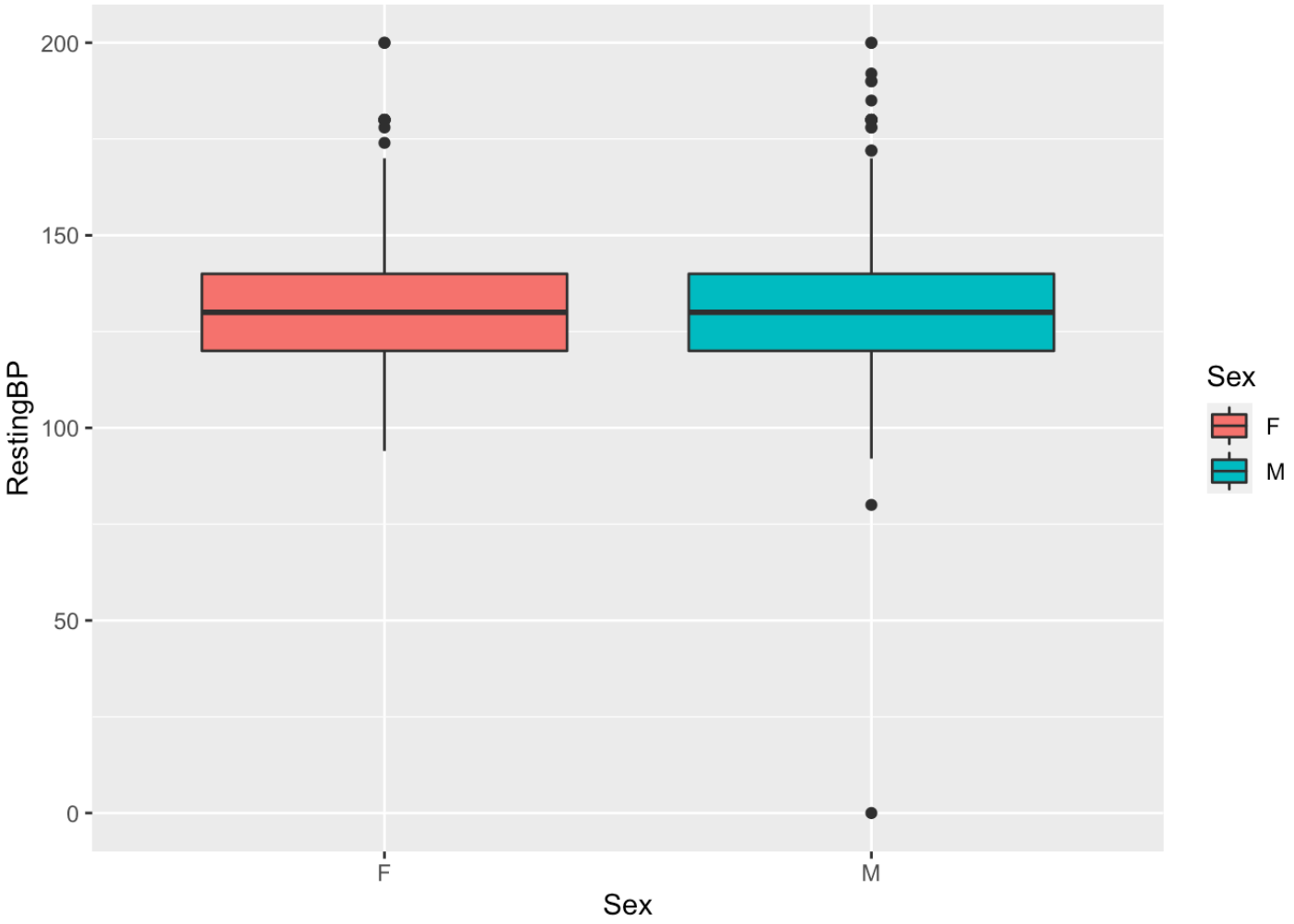
```
## Note that the heart.csv file must be in your working directory

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

a. Using the plot(s) of your choice, 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? Explain your answer.

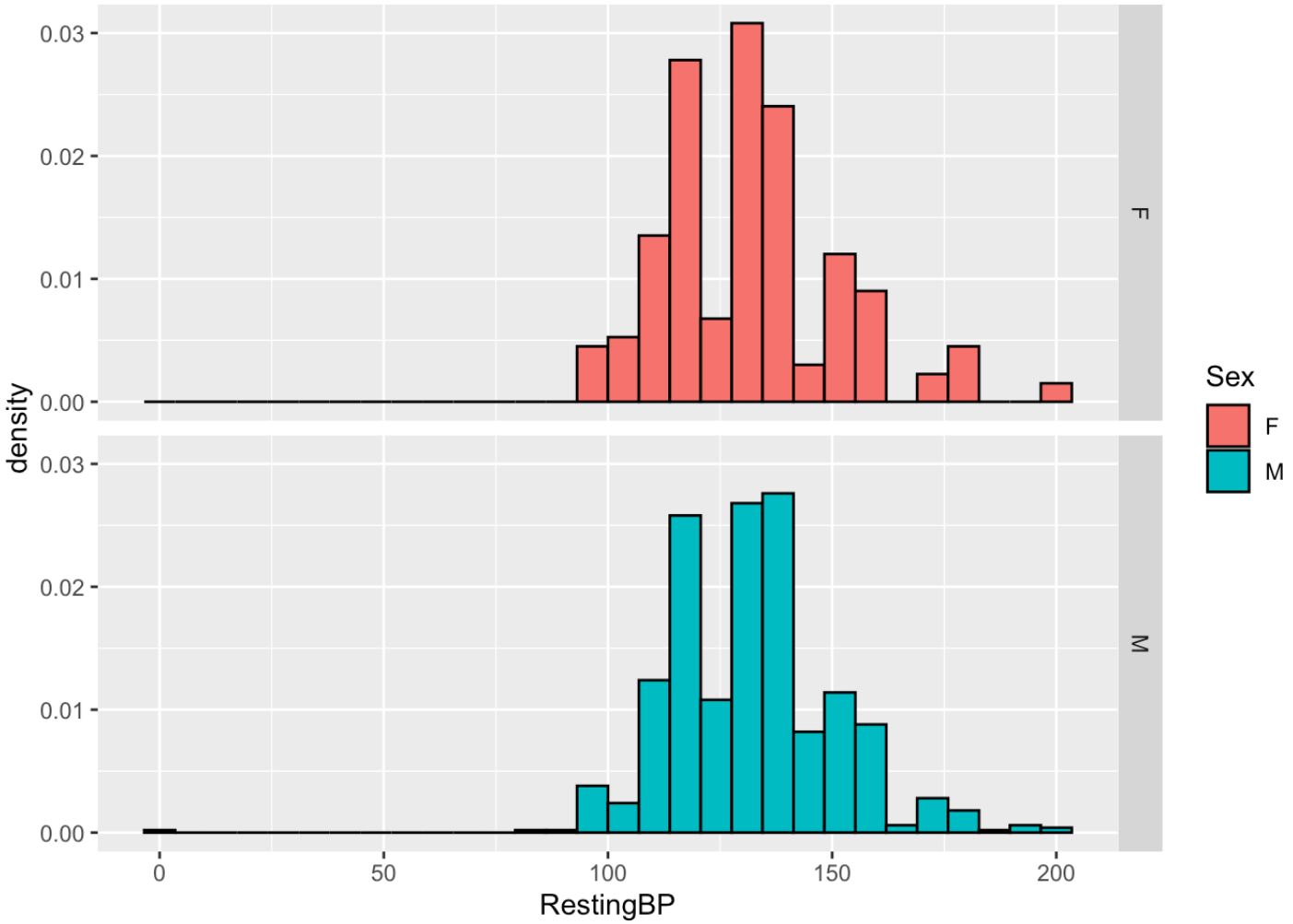
Solution:

Note that the sex of the patient is a qualitative variable and the resting BP is a quantitative variable. Therefore, the right plots to use are histograms and boxplots.

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




```
ggplot(heart_tbl,aes(x=RestingBP,fill=Sex)) + geom_histogram(aes(y=..density..),col="
black") +
  facet_grid(Sex~.)
```

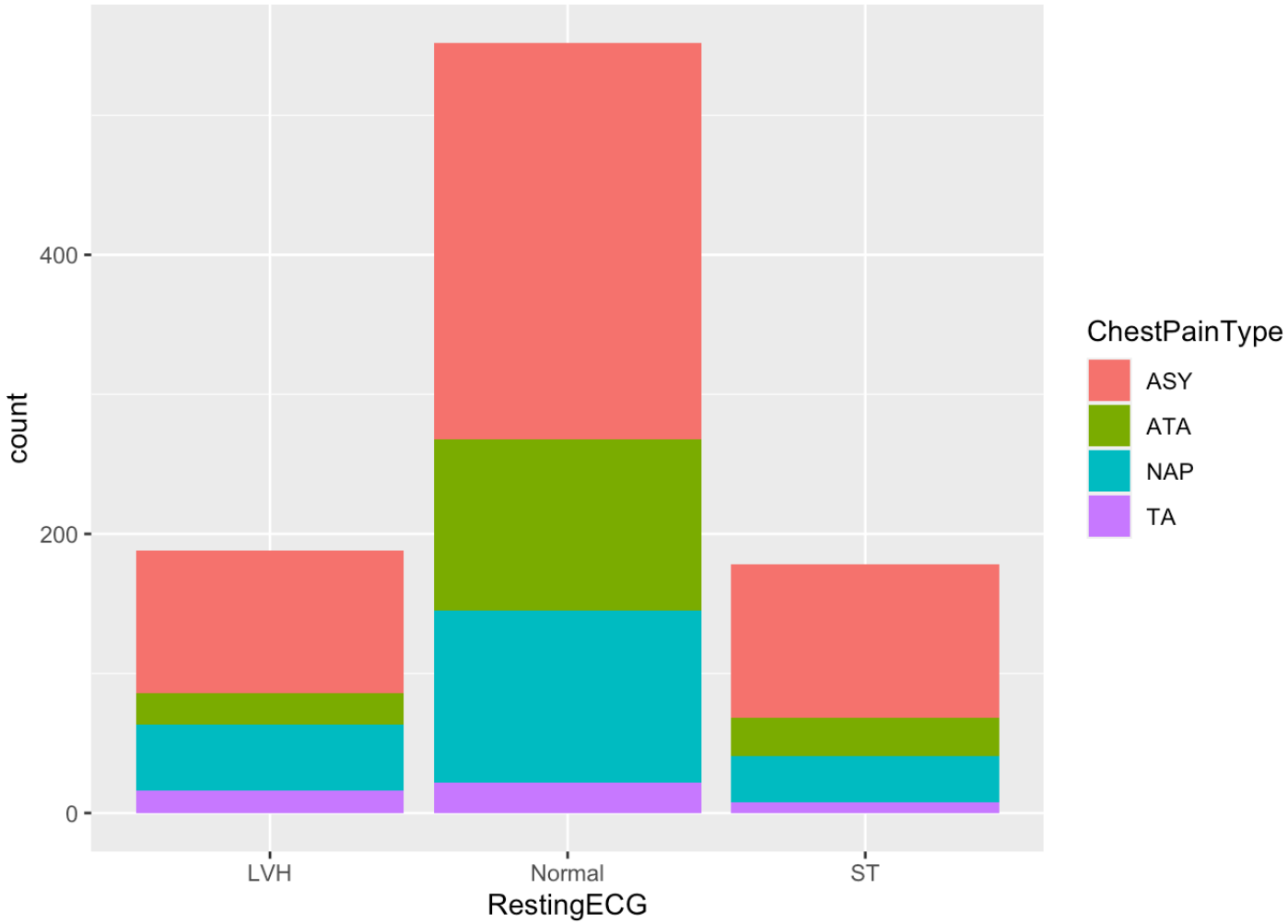


Note that it is clear that the distribution of Resting BP between the two sexes is very similar. This is a clear indication that there is no association between the sexes and Resting BP.

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

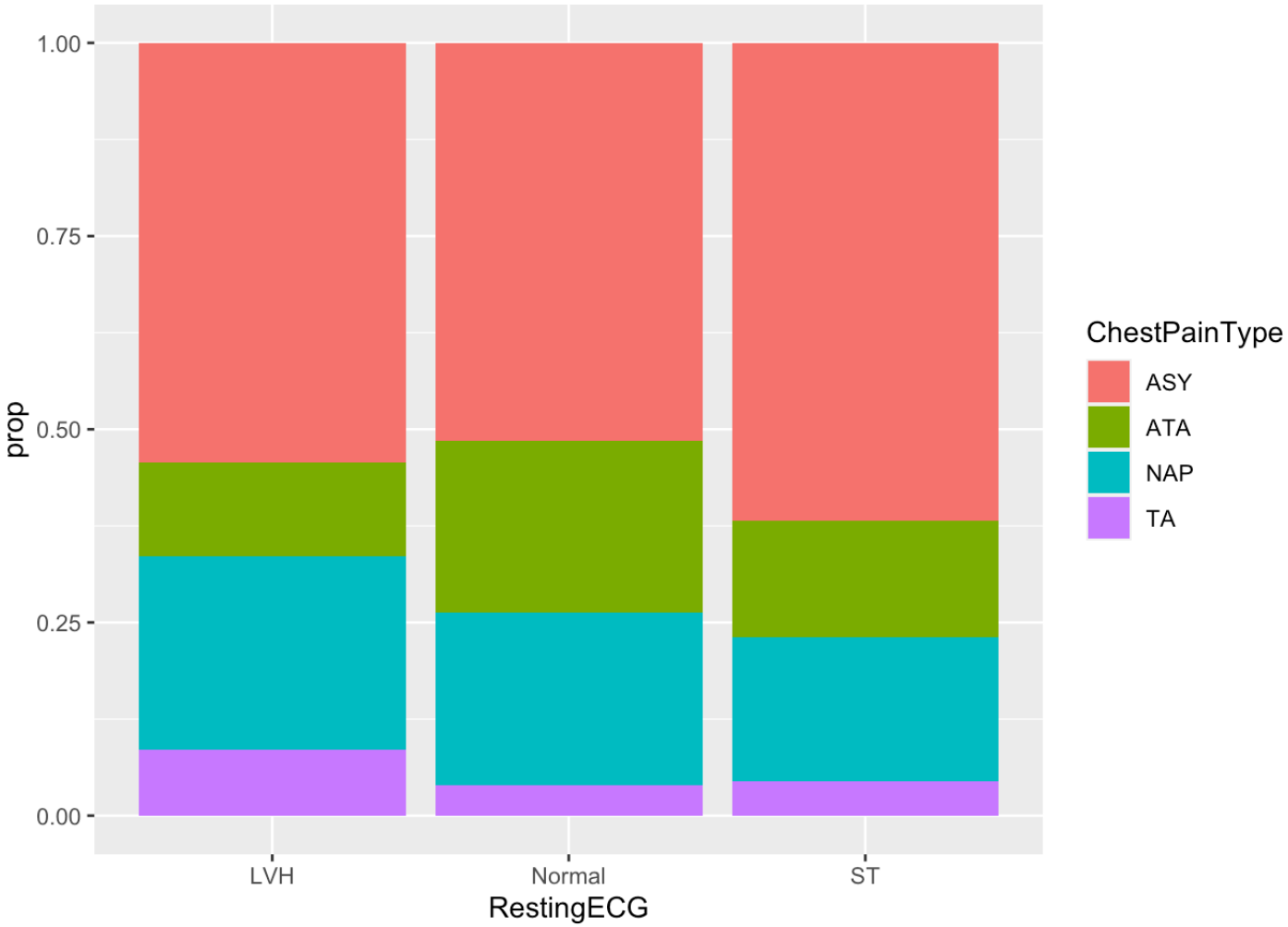
Solution: In this case we have two qualitative random variables, so a barplot makes much more sense.

```
ggplot(heart_tbl,aes(fill=ChestPainType,x=RestingECG)) + geom_bar()
```



Note that the question is not clear about whether it should be scaled within the bars or not. I had intended just to show the counts, but this allows you to see the relative distribution:

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



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

Solution:

Similar to the extra bit above, just flipped:

```
sum_table<-heart_tbl %>% count(ChestPainType,RestingECG) %>%
         group_by(ChestPainType) %>%
         summarize(RestingECG=RestingECG,ChestPainType=ChestPainType,prop=n/sum(n))
sum_table
```

```
# A tibble: 12 x 3
# Groups:   ChestPainType [4]
   ChestPainType RestingECG  prop
   <chr>         <chr>      <dbl>
 1 ASY           LVH        0.206
 2 ASY           Normal     0.573
 3 ASY           ST         0.222
 4 ATA           LVH        0.133
 5 ATA           Normal     0.711
 6 ATA           ST         0.156
 7 NAP           LVH        0.232
 8 NAP           Normal     0.606
 9 NAP           ST         0.163
10 TA            LVH        0.348
11 TA            Normal     0.478
12 TA            ST         0.174
```

Obviously looks nicer if you `pivot_wider` (but not necessary)

```
sum_table %>% pivot_wider(id_cols=ChestPainType,
                          names_from=RestingECG,values_from=prop)
```

```
# A tibble: 4 x 4
# Groups:   ChestPainType [4]
  ChestPainType   LVH Normal    ST
  <chr>         <dbl>  <dbl> <dbl>
1 ASY           0.206  0.573 0.222
2 ATA           0.133  0.711 0.156
3 NAP           0.232  0.606 0.163
4 TA            0.348  0.478 0.174
```

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.

Solution: Tricky part here is to get summaries for multiple values. The two ways to do this are to subset the data by the columns you want first, then use `summarise_all` . The other way is to use `summarise_at` on the full datset. I use the former below because it is what I used in class, but the latter is also acceptable.

```
heart_tbl %>% select(ChestPainType,RestingBP,Cholesterol,FastingBS,MaxHR) %>%
  group_by(ChestPainType) %>%
  summarise_all(.funs=list(mean=mean,median=median,IQR=IQR))
```

```
# A tibble: 4 x 13
  ChestPainType RestingBP_mean Cholesterol_mean FastingBS_mean MaxHR_mean
  <chr>                  <dbl>            <dbl>          <dbl>      <dbl>
1 ASY                     133.             187.          0.284       128.
2 ATA                     131.             233.          0.110       150.
3 NAP                     131.             197.          0.202       143.
4 TA                      136.             207.          0.283       148.
# … with 8 more variables: RestingBP_median <dbl>, Cholesterol_median <dbl>,
#   FastingBS_median <dbl>, MaxHR_median <dbl>, RestingBP_IQR <dbl>,
#   Cholesterol_IQR <dbl>, FastingBS_IQR <dbl>, MaxHR_IQR <dbl>
```

```
### Using summarize_at:
# heart_tbl  %>%
#   group_by(ChestPainType) %>%
#   summarise_at(.vars=vars(RestingBP,Cholesterol,FastingBS,MaxHR),
#   .funs=list(mean=mean,median=median,IQR=IQR))
```

This one looks better flipped around (again, not necessary):

```
heart_tbl %>% select(ChestPainType,RestingBP,Cholesterol,FastingBS,MaxHR) %>%
  group_by(ChestPainType) %>%
  summarise_all(.funs=list(mean=mean,median=median,IQR=IQR)) %>%
  pivot_longer(cols=-ChestPainType) %>%
  pivot_wider(id_cols=name,names_from=ChestPainType,values_from=value)
```

```
# A tibble: 12 x 5
   name                  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
```

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.

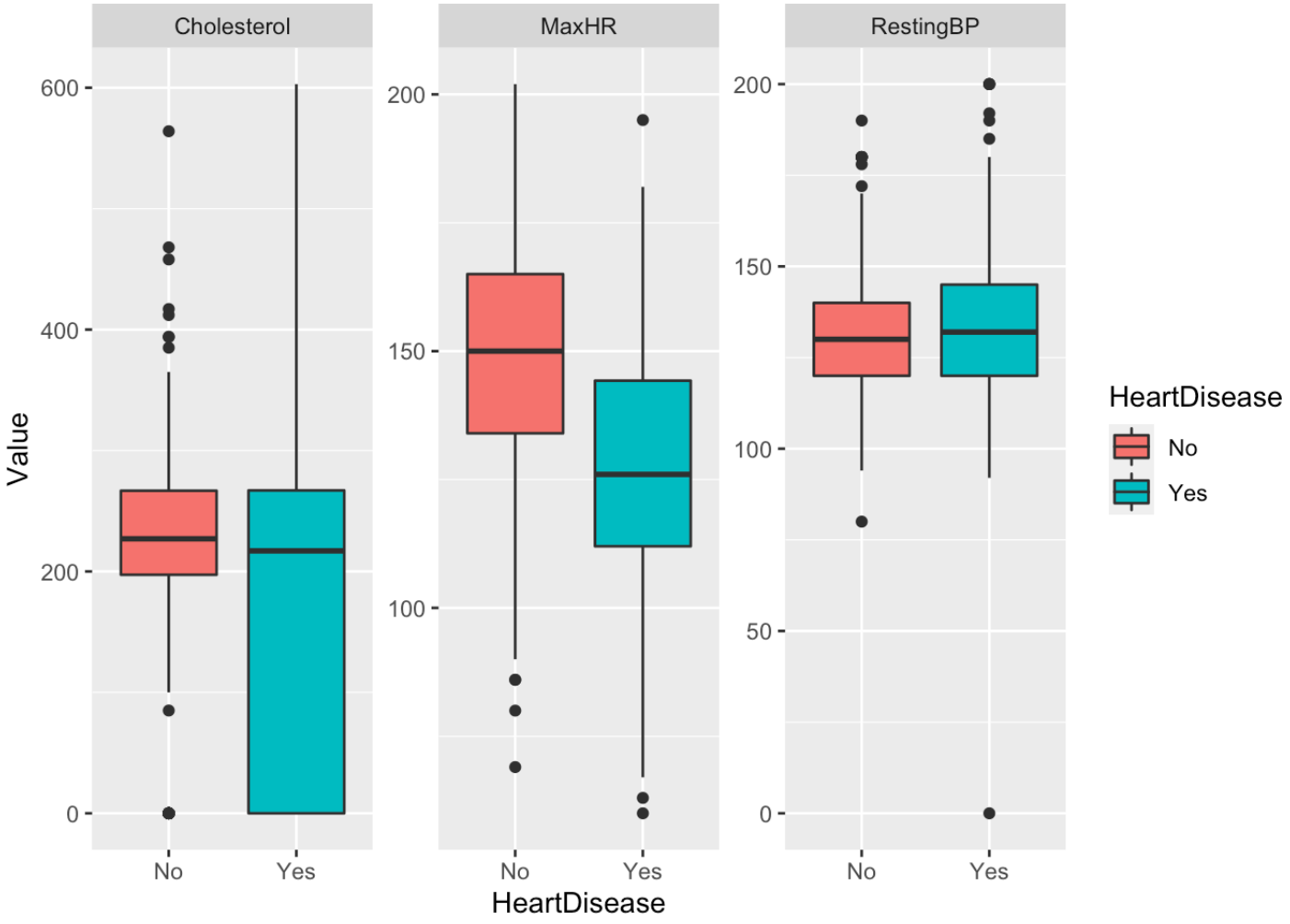
Solution: Note that FastingBS is a categorical variable, so we will treat it second. You can do the summarizing of the quantitative measures with a lot of plotting code (also OK!) but the most efficient way to do this is with a longer dataset.

```
long_version<-heart_tbl %>%
  select(HeartDisease,RestingBP,Cholesterol,MaxHR) %>%
  pivot_longer(cols=RestingBP:MaxHR,values_to="Value") %>%
  mutate(HeartDisease=ifelse(HeartDisease==1,"Yes","No")) # plotting is nicer
head(long_version)
```

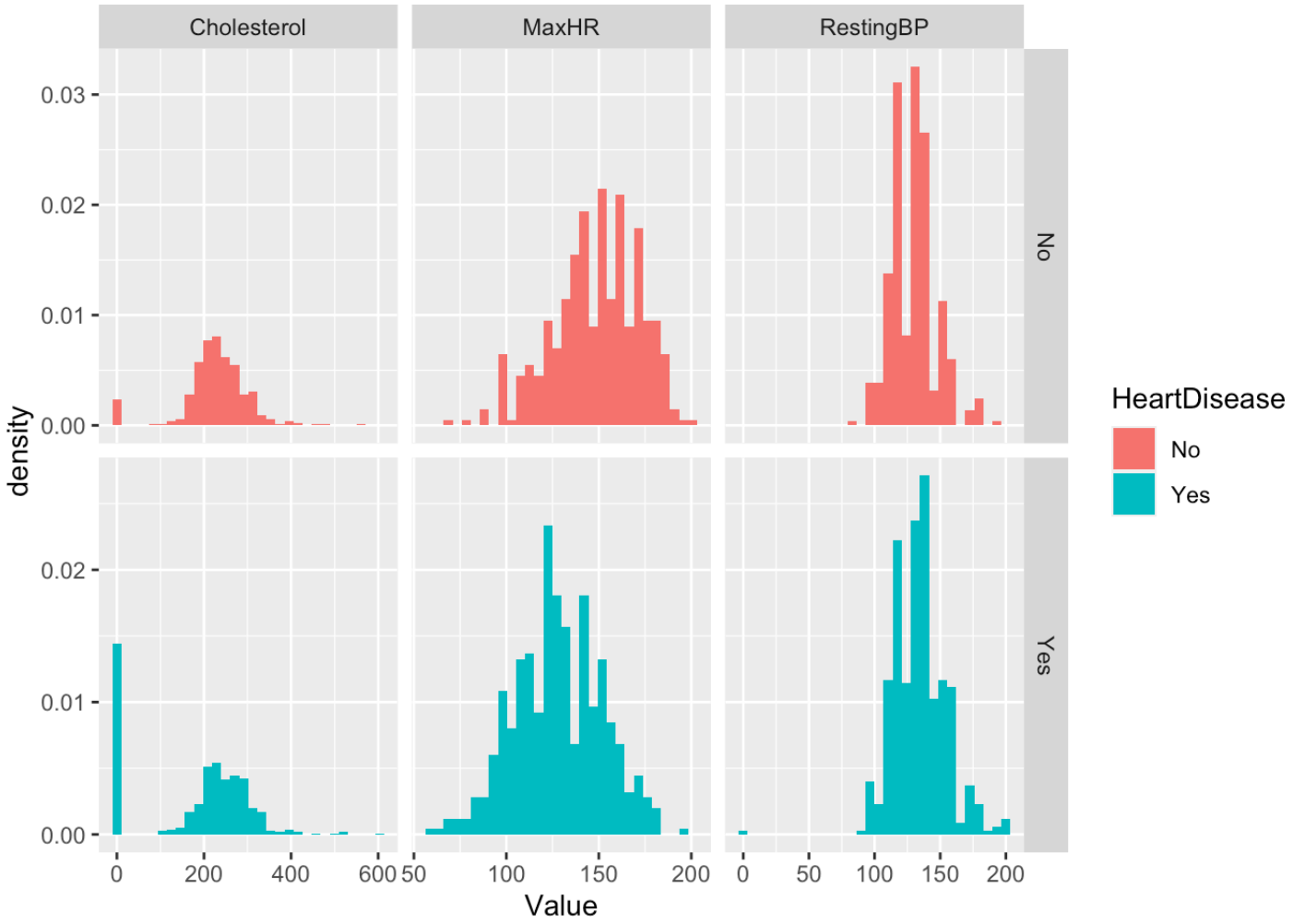
```
# A tibble: 6 x 3
  HeartDisease name        Value
  <chr>        <chr>       <dbl>
1 No           RestingBP     140
2 No           Cholesterol   289
3 No           MaxHR         172
4 Yes          RestingBP     160
5 Yes          Cholesterol   180
6 Yes          MaxHR         156
```

Now we can make the usual plots:

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



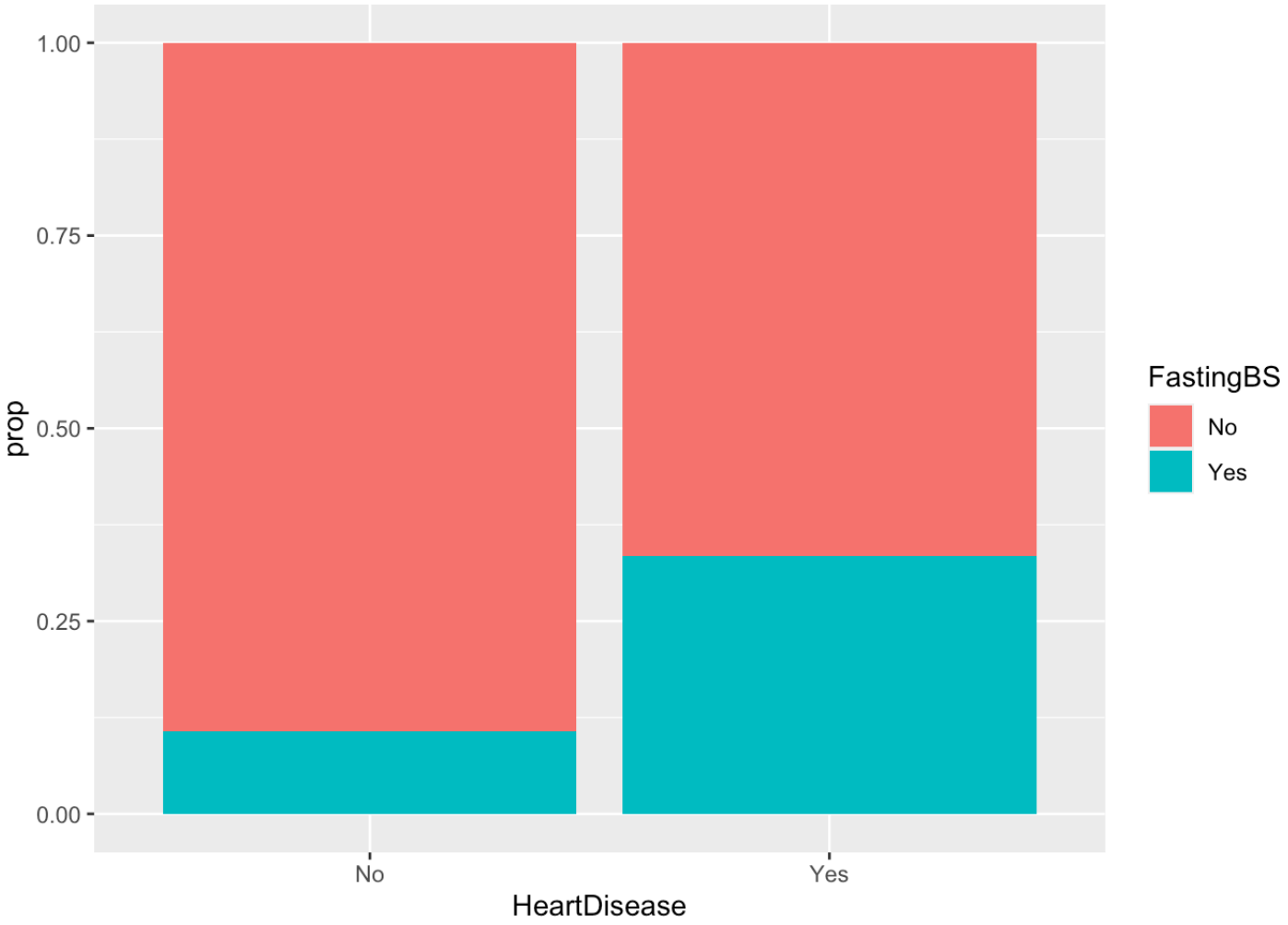
```
ggplot(long_version, aes(x=Value,fill=HeartDisease,group=HeartDisease)) +
  geom_histogram(aes(y=..density..)) + facet_grid(HeartDisease~name,scales="free")
```



Based on these plots, the maximum heart rate seems to be the most strongly associated with heart disase, as the heart disease group has lower heart rates than the normal (non-heart disease group). The distributions of the other two variables seem quite similar. Note that the boxplot for the cholesterol variable is misleading due to the zero values (very interesting that these are here – maybe should be NA?). The histogram shows that for those observations with values, the distribution in the two groups is quite similar. The resting BP is very slightly higher for the heart disease group, but not to the same extent as Max HR.

For the FastingBS variable, you really should use a stacked barplot, properly normalized:

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

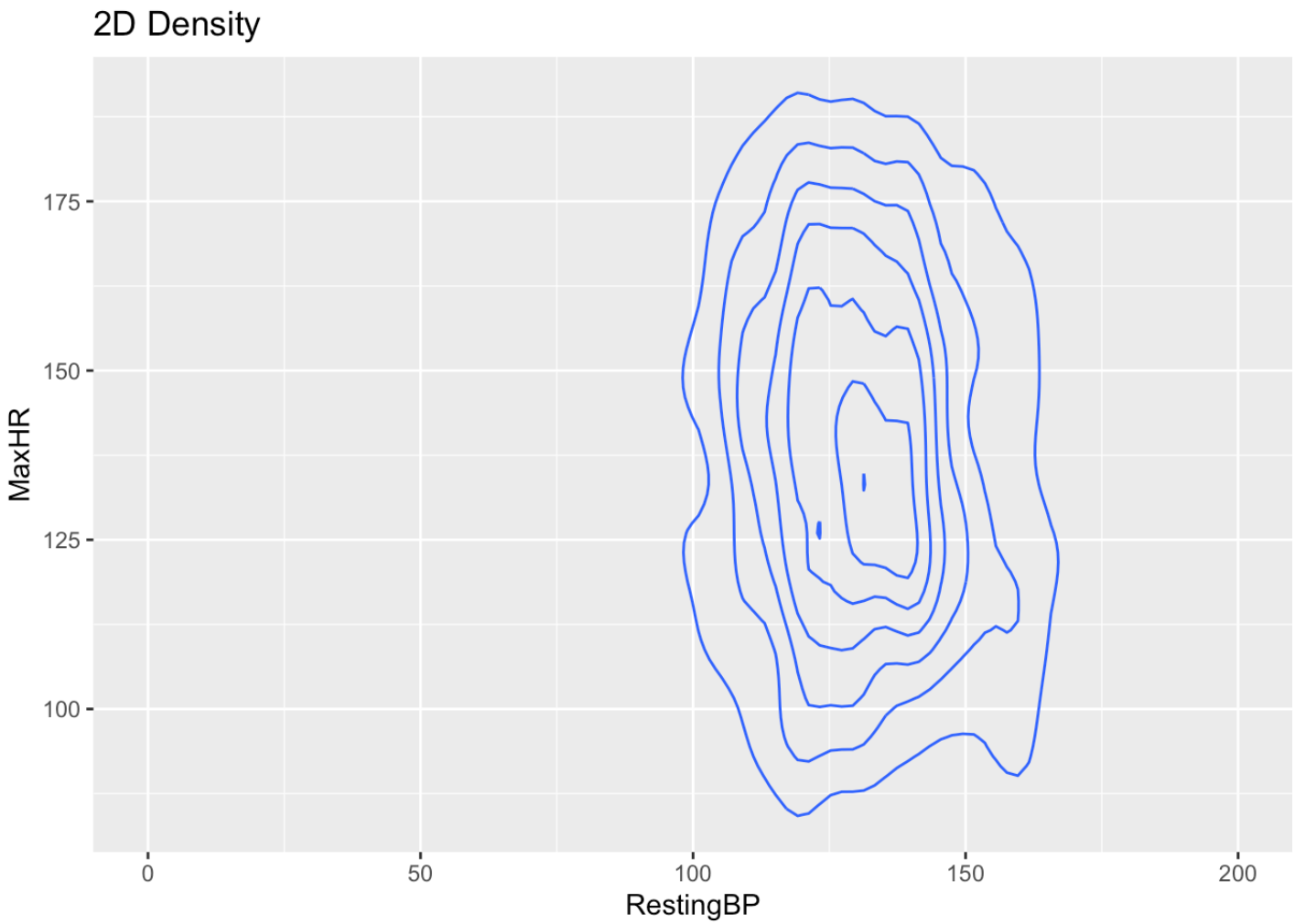


Here we see that the HeartDisease has much higher percentage of patients whose blood sugars were measured after fasting.

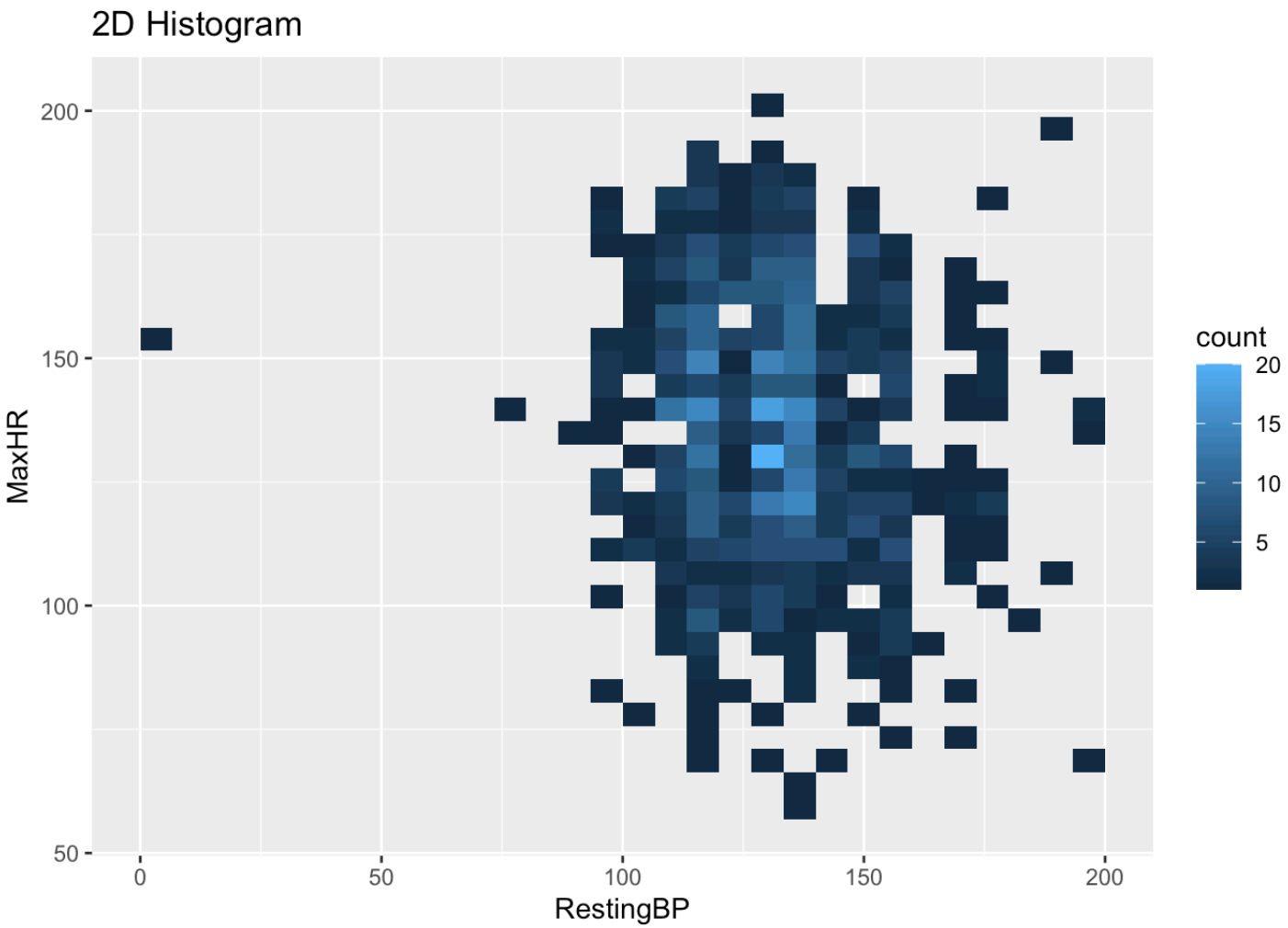
f. Create both a 2-d histogram and a 2-d contour plot to assess the association betwen RestingBP and MaxHR. Describe this association and also explain which plot you think shows the association most clearly (or explain why they are about the same).

Solution:

```
ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_density_2d() +
  ggtitle("2D Density") + xlim(c(0,200))
```




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



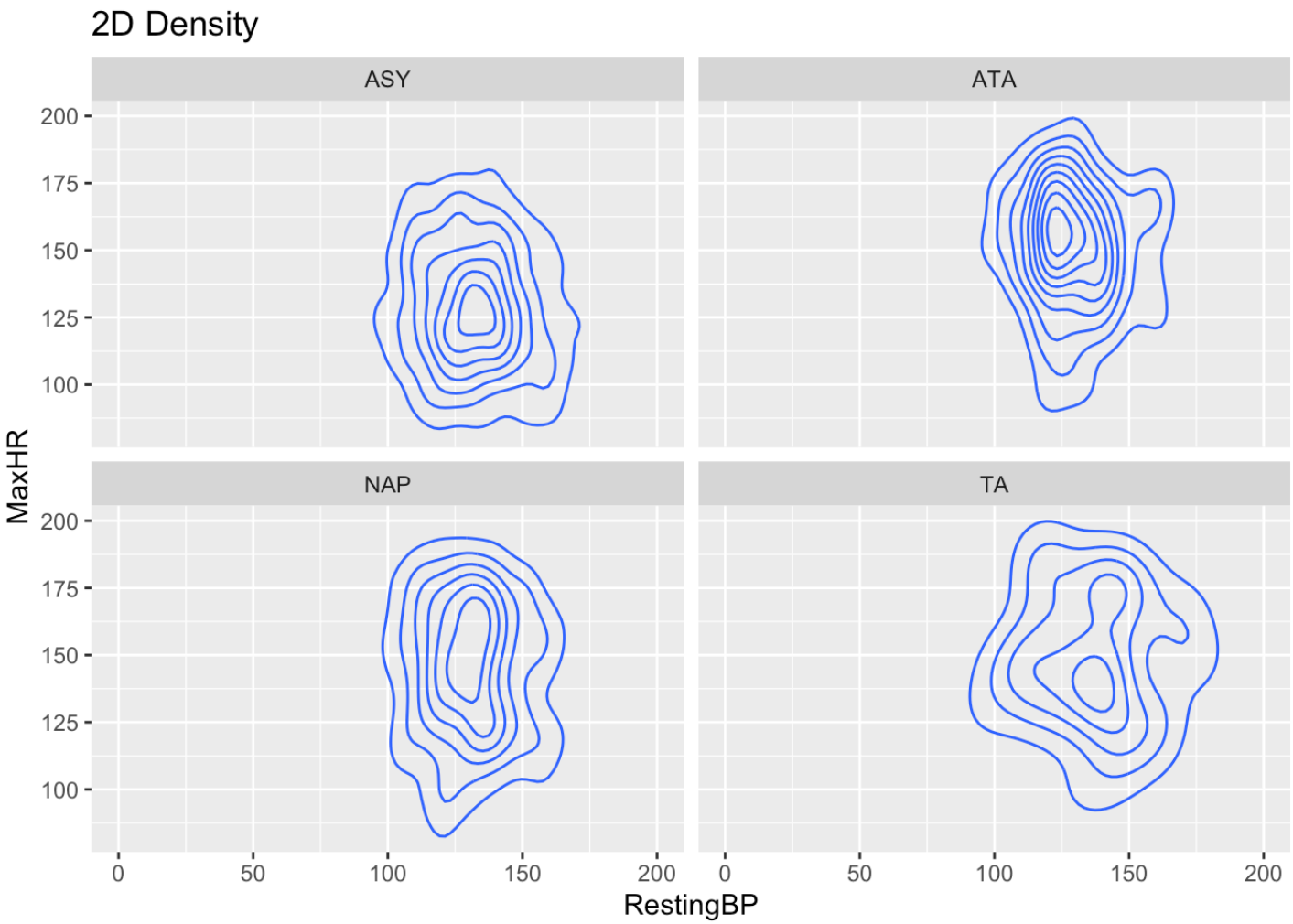
Here we see very little association between the RestingBP and MaxHR. The spread in each measure is different, but we see the distribution of one is the same as we vary the values of the other. Here the plots look very similar because I have adjusted the x-axis limits to match (note the 0 values of Resting BP cause the density plot to just cut them out). Both plots here give the same basic information once that correction is made.

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).

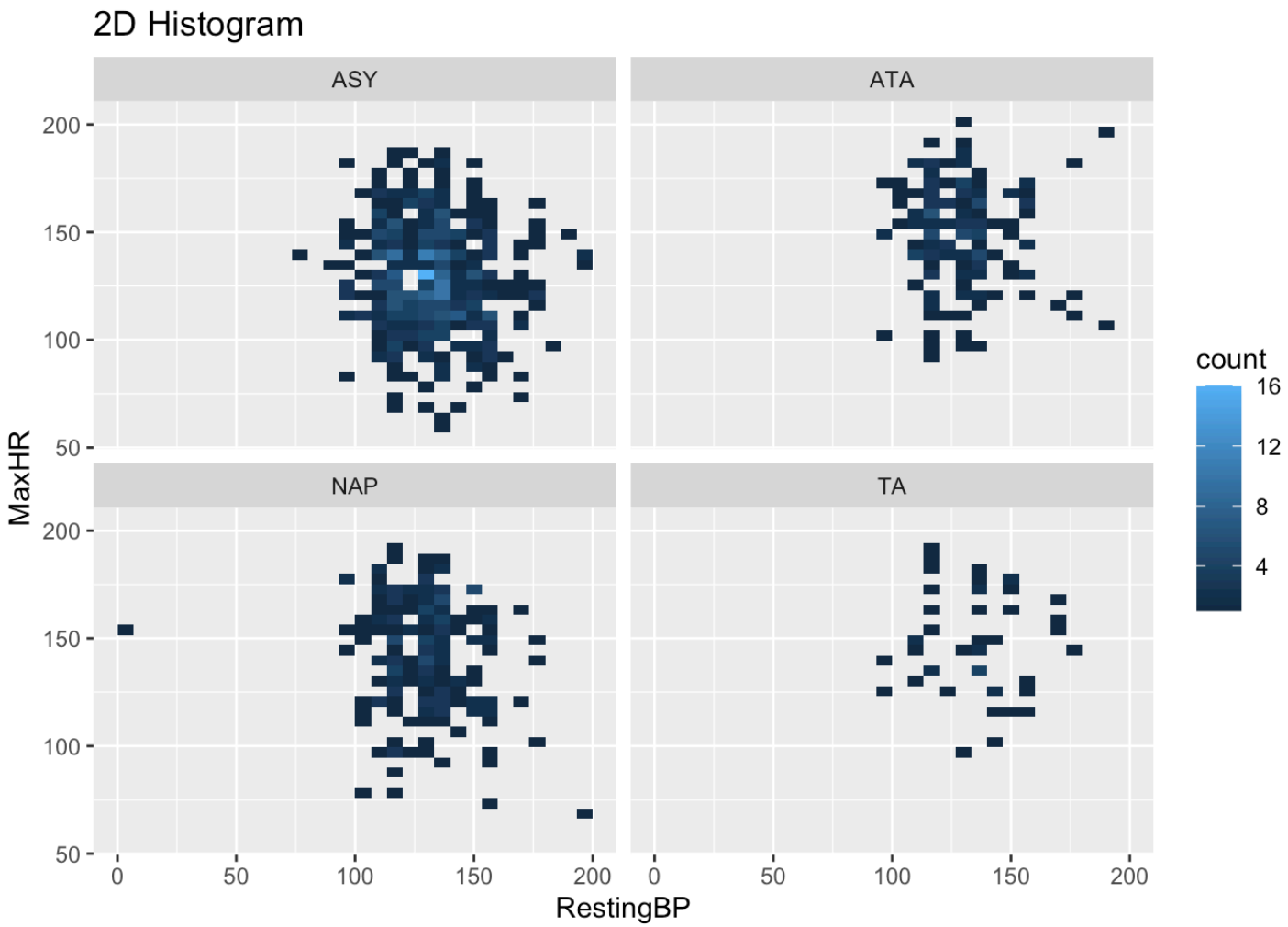
Solution:

Here we can use the same code as before, but faceting on the various factors:

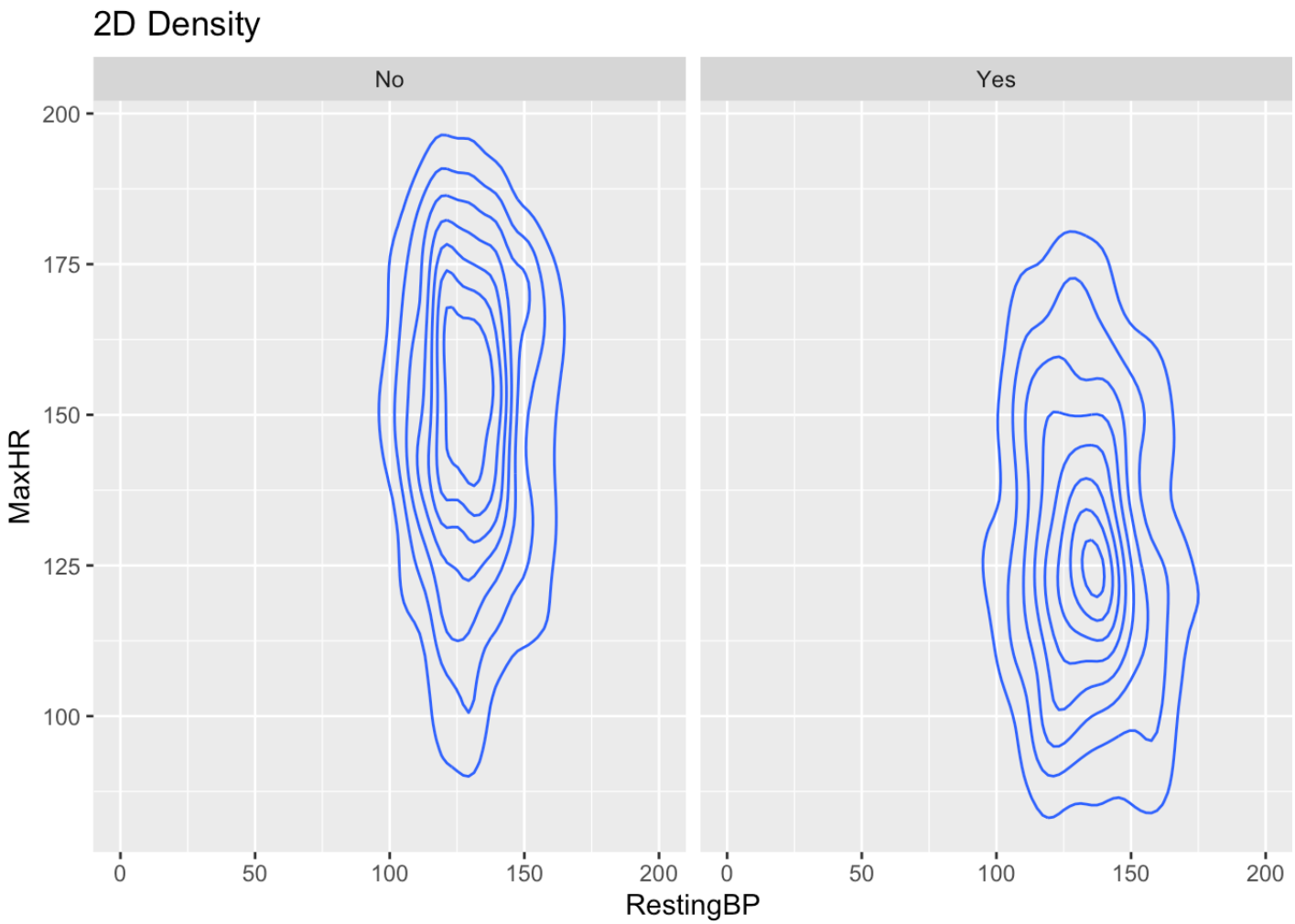
```
heart_tbl <- heart_tbl %>% mutate( HeartDisease=ifelse(HeartDisease==1,"Yes","No"))

ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_density_2d() +
  ggtitle("2D Density") + xlim(c(0,200)) + facet_wrap(~ChestPainType)
```



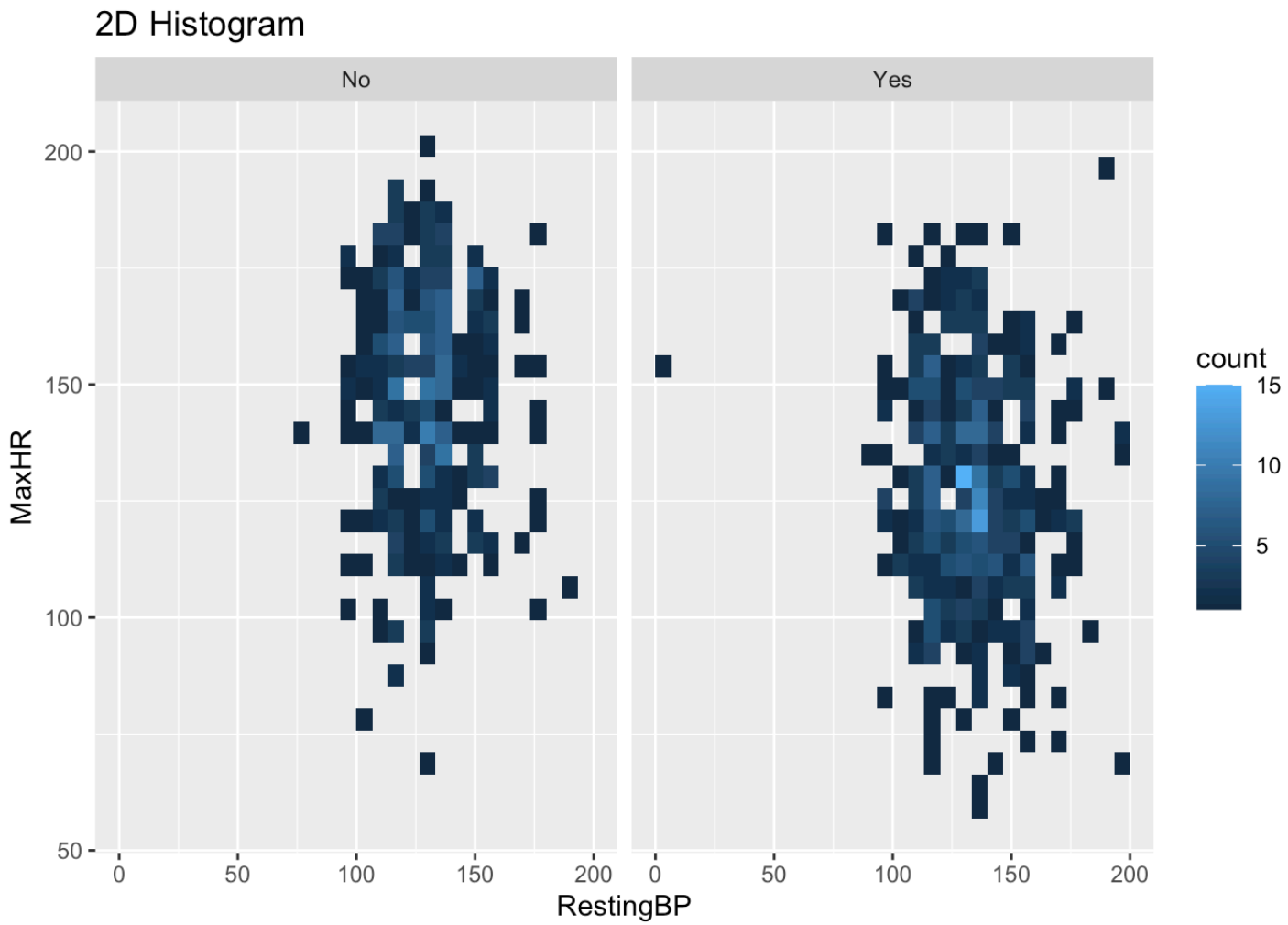

```
ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_bin2d() +
  ggtitle("2D Histogram")+ facet_wrap(~ChestPainType)
```



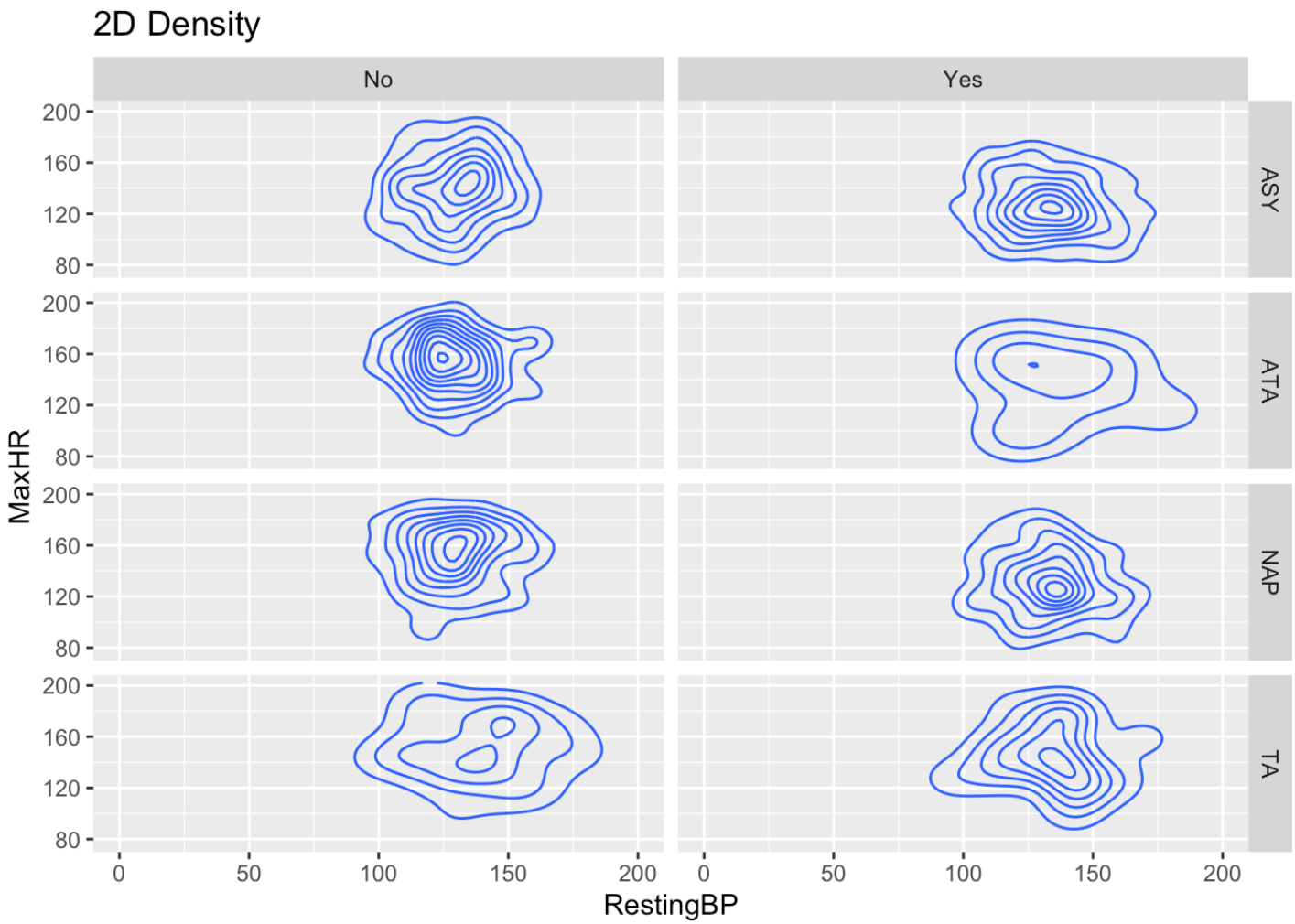

```
ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_density_2d() +
  ggtitle("2D Density") + xlim(c(0,200)) + facet_wrap(~HeartDisease)
```



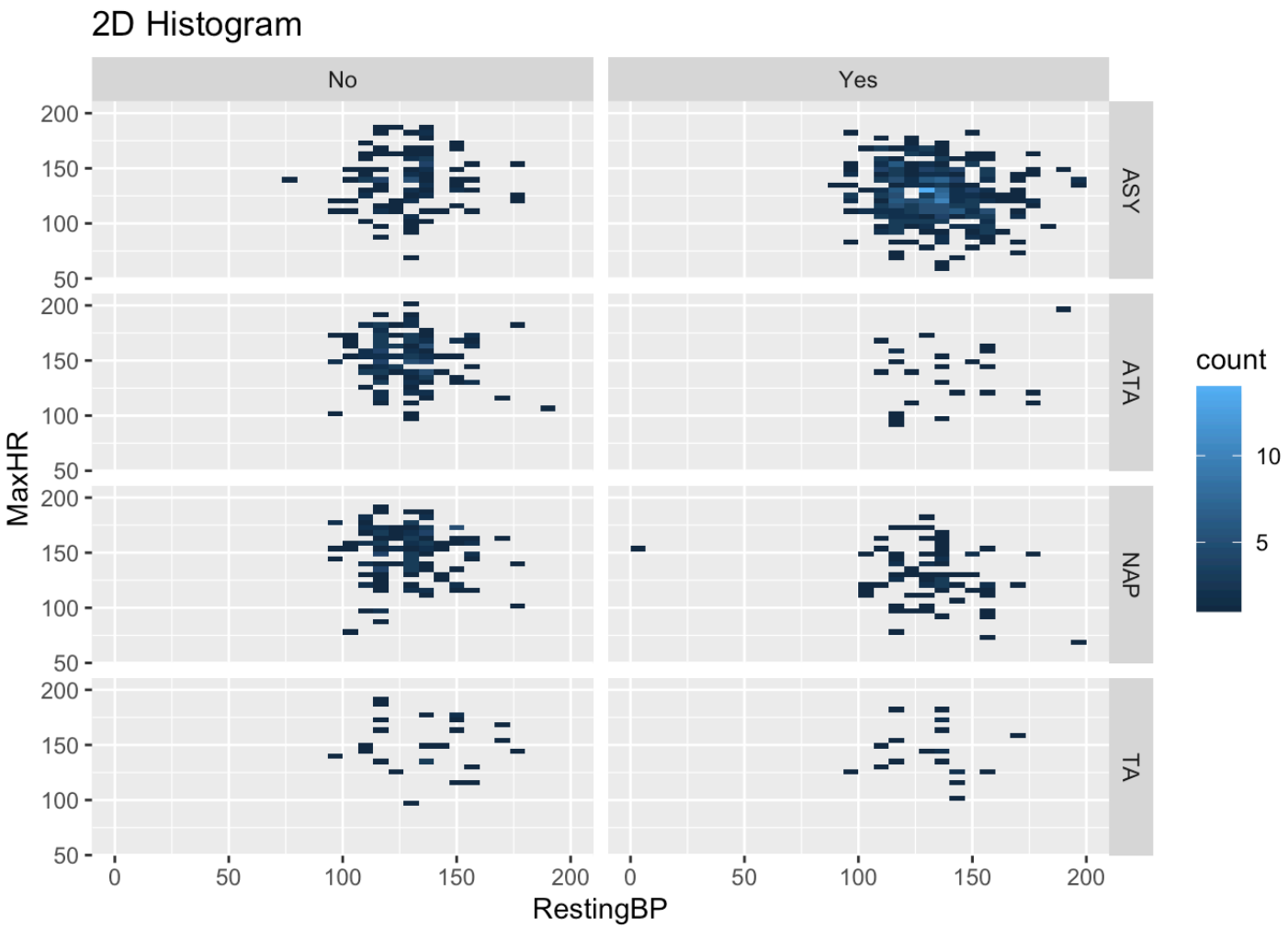

```
ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_bin2d() +
  ggtitle("2D Histogram")+ facet_wrap(~HeartDisease)
```




```
ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_density_2d() +
  ggtitle("2D Density") + xlim(c(0,200)) + facet_grid(ChestPainType~HeartDisease)
```




```
ggplot(heart_tbl,aes(x=RestingBP,y=MaxHR)) + geom_bin2d() +
  ggtitle("2D Histogram")+ facet_grid(ChestPainType~HeartDisease)
```



Based on all of these figures, it does not seem like the association doesn't depend on these two variables or their interaction.