Stats140XP_Final Project

Xiaocong Xuan – 705544936

11/22/2022

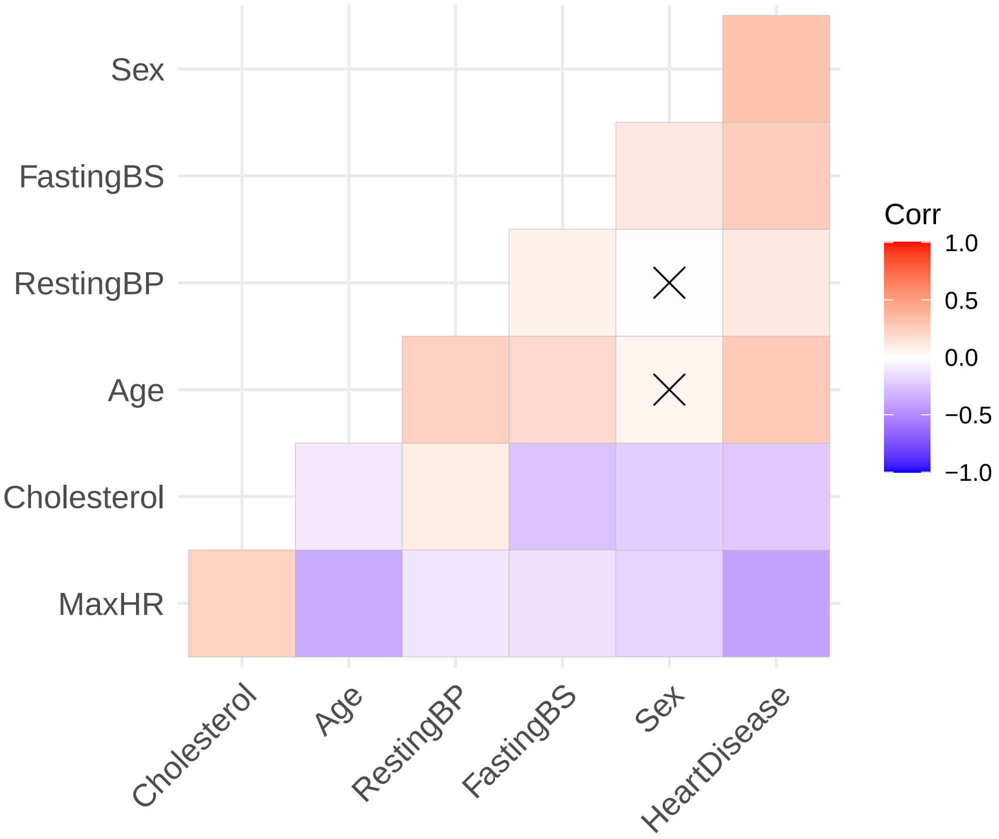
Deciding what we want to look at

We want to find out good risk factors of heart disease. Since we are not experts in medical terms, we

```
heart <- read.csv(file = 'heart.csv', stringsAsFactors = T)
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
heart$Sex <- as.numeric(heart$Sex)
heart$HeartDisease <- as.numeric(heart$HeartDisease)
index_to_investigate <- c("Age", "Sex", "RestingBP", "FastingBS", "MaxHR", "HeartDisease", "Cholesterol
corr <- cor(heart[,index_to_investigate])
p.mat <- cor_pmat(heart[,index_to_investigate])
ggcorrplot(corr,
           hc.order = TRUE,
           type = "lower",
           p.mat = p.mat)
```



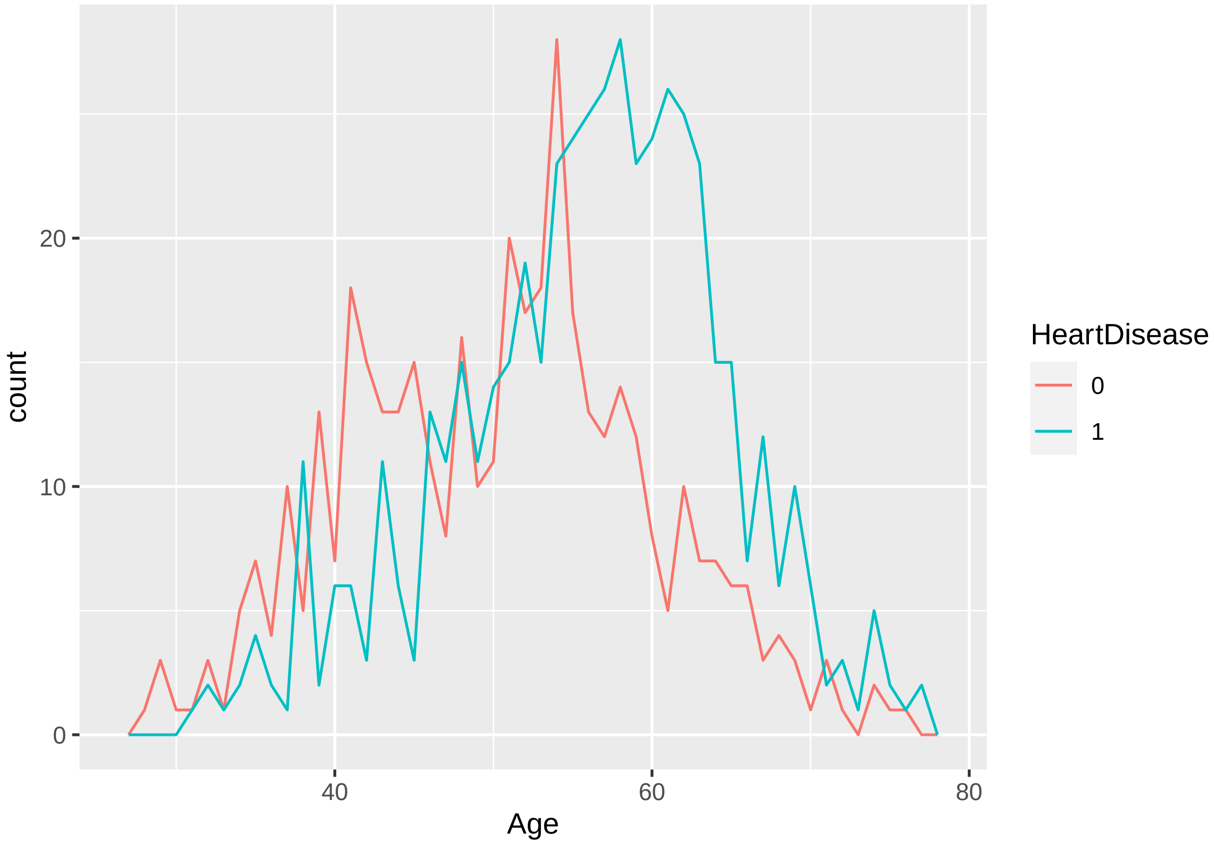
From the correlation matrix, we mainly focus on corelations between variables and Heart Disease. we can see that max Heart Rate and Cholesterol have a negative correlation with Heart disease, while sex, age, Fasting Blood Sugar level, and resting blood pressure has positive correlation with Heart Disease.

We can also look at the correlation between the interaction terms. However, they are not the main concern of this study.

We then look at the distribution for each variable to observe patterns and abnormalties.

Visualizing HeartDisease and a continuous variable.

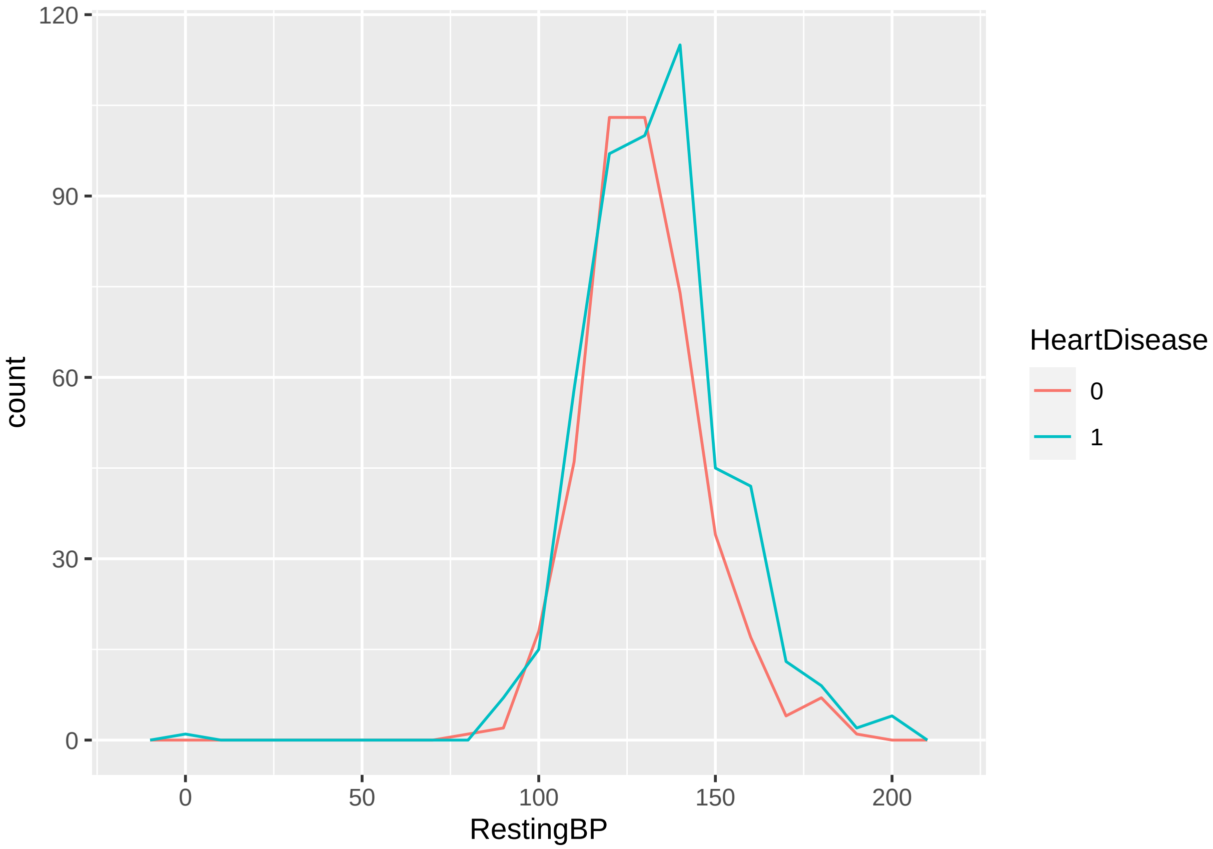
```
heart <- read.csv(file = 'heart.csv', stringsAsFactors = T)
heart$HeartDisease <- as.character(heart$HeartDisease)

# Between the ages of 55 and 65 is the high incidence period of Heart Failure.
ggplot(data = heart, mapping = aes(x = Age)) + geom_freqpoly(mapping = aes(colour = HeartDisease), binw
```



```
above_50 <- heart$Age >= 50
cat("People above 60 is ", (mean(as.integer(heart$HeartDisease[above_50]))/mean(as.integer(heart$HeartDi
```

```
## People above 60 is 65.99422% more likely to get heart disease
```

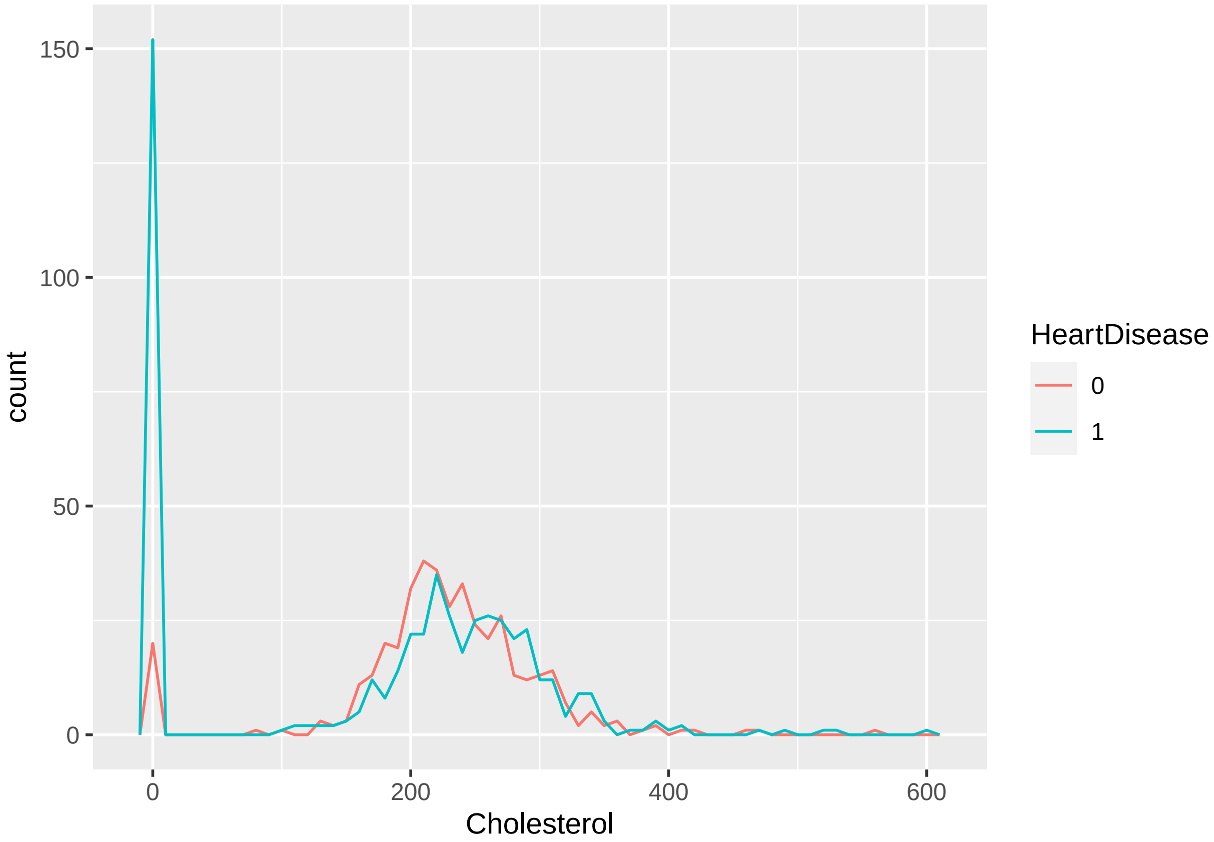
```
# RestingBP is not a significant factor to influence Heart Failure.
ggplot(data = heart, mapping = aes(x = RestingBP)) + geom_freqpoly(mapping = aes(colour = HeartDisease)
```



```
BP_above_145 <- heart$RestingBP >= 145
cat("People with Resting Blood pressure above 145 is ", (mean(as.integer(heart$HeartDisease[BP_above_145
```

```
## People with Resting Blood pressure above 145 is 26.68718% more likely to get heart disease
```

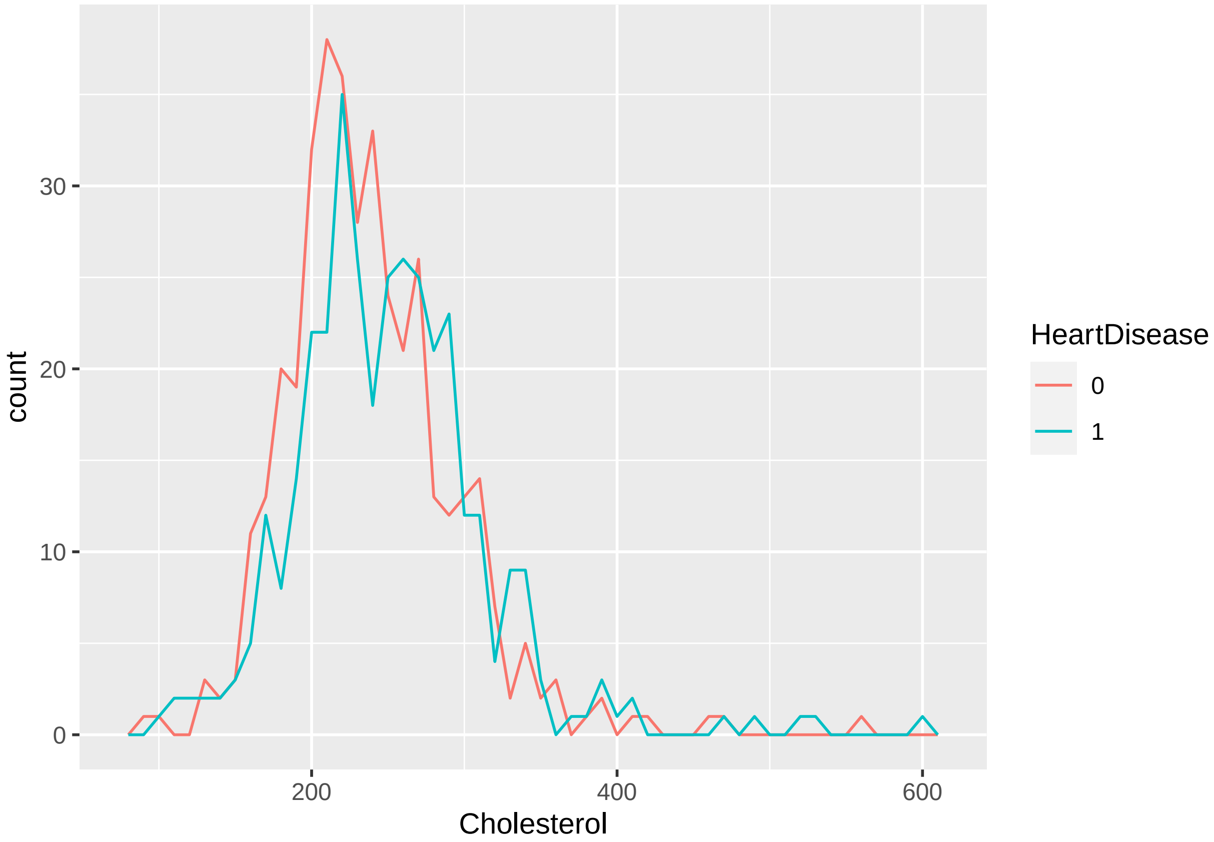
```
# Cholesterol
ggplot(data = heart, mapping = aes(x = Cholesterol)) + geom_freqpoly(mapping = aes(colour = HeartDisease
```



Note that there are many 0s recorded for Cholesterol level, which should indicate Non-recorded entries, i.e. NAs. Thus we remove the 0s, and new distribution look like this:

```
NA.index <- heart$Cholesterol == 0
heart$Cholesterol[NA.index] <- NA
ggplot(data = heart, mapping = aes(x = Cholesterol)) + geom_freqpoly(mapping = aes(colour = HeartDiseas
```

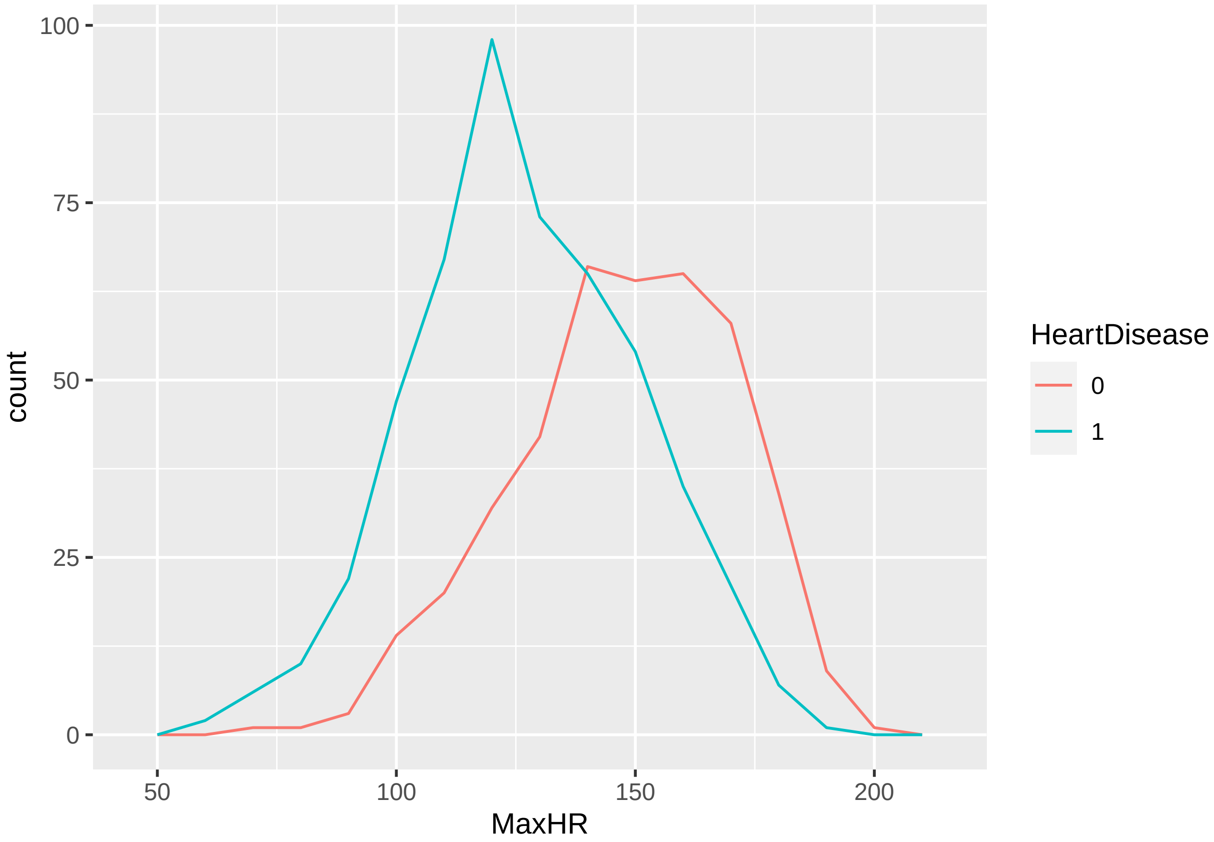
```
## Warning: Removed 172 rows containing non-finite values (‘stat_bin()‘).
```



```
cho_above_280 <- heart$Cholesterol >= 280
cat("People with Cholesterol above 280 are ", (mean(as.integer(heart$HeartDisease[cho_above_280]), na.r
```

```
## People with Cholesterol above 280 are 30.20509% more likely to get heart disease
```

```
# Higher MaxHR is associated with higher risks of Heart Failure.
ggplot(data = heart, mapping = aes(x = MaxHR)) + geom_freqpoly(mapping = aes(colour = HeartDisease), bi
```



```
MaxHR_below_150 <- heart$MaxHR <= 150
cat("People with Max Heart Rate below 150 are ", (mean(as.integer(heart$HeartDisease[MaxHR_below_150]),
```

```
## People with Max Heart Rate below 150 are 132.3906% more likely to get heart disease
```

Visualizing HeartDisease and a categorical variable.

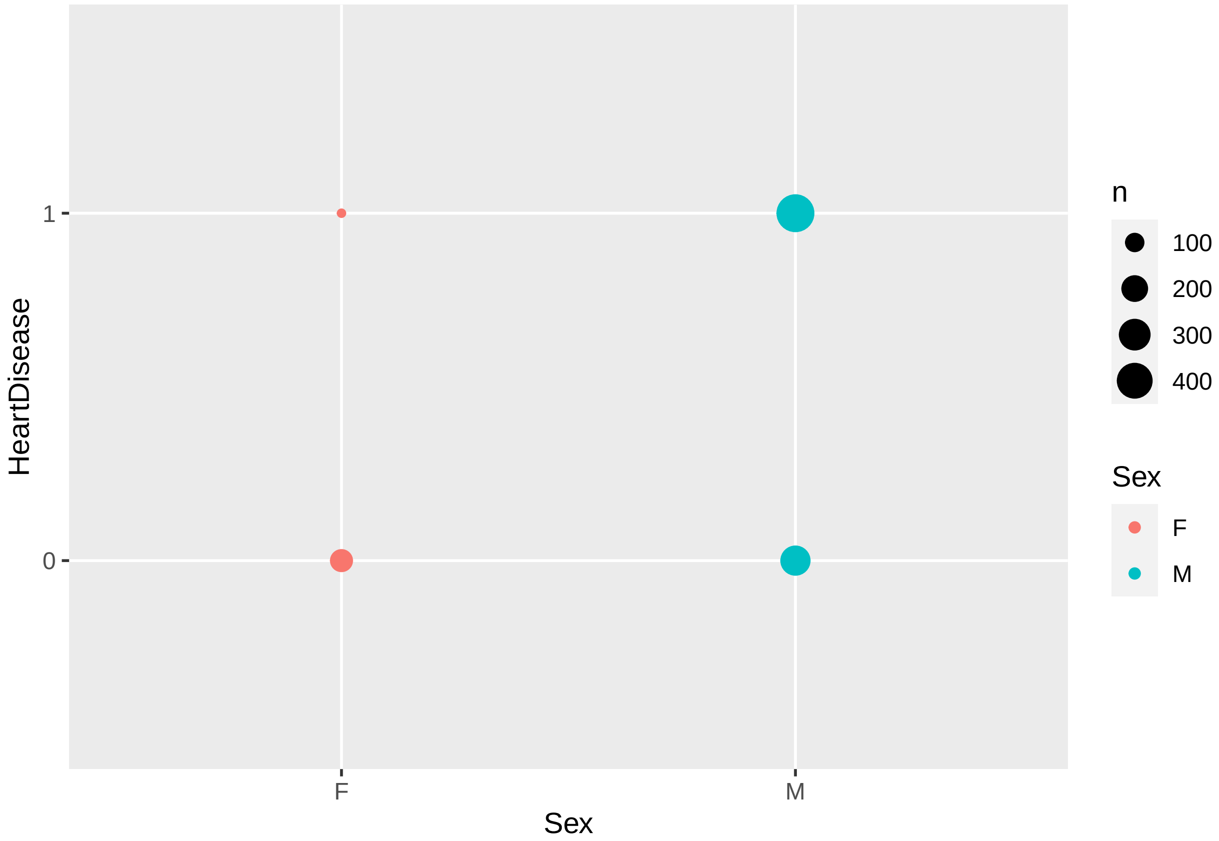
```
library(ggplot2)
library(tidyverse)
```

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

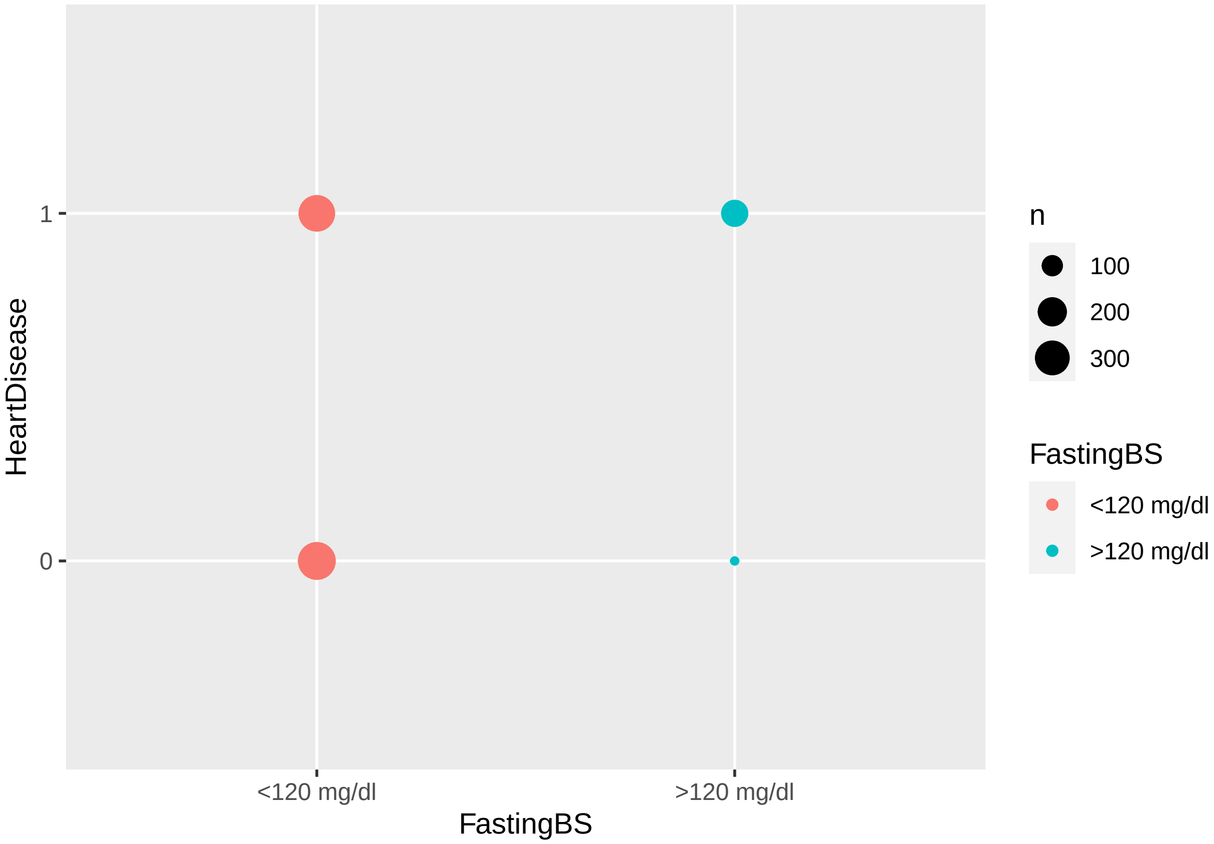
```
## v tibble  3.1.8     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4
```

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

```r
# The incidence of heart failure is higher in women.
ggplot(data = heart) +
  geom_count(mapping = aes(x = Sex, y = HeartDisease, colour = Sex))
```



```r
# The incidence of heart failure is higher in FastingBS > 120 mg/dl.
heart$FastingBS <- ifelse(heart$FastingBS == 1, ">120 mg/dl", "<120 mg/dl")
ggplot(data = heart) +
  geom_count(mapping = aes(x = FastingBS, y = HeartDisease, colour = FastingBS))
```



```
FastingBS_below_150 <- heart$FastingBS <= 150
cat("People with Max Heart Rate below 150 are ", (mean(as.integer(heart$HeartDisease[FastingBS_below_15
```

```
## People with Max Heart Rate below 150 are NaN% more likely to get heart disease
```

use ANOVA to check whether patterns we observed are are significant.

```
anova1 <- aov(HeartDisease~.,heart[,index_to_investigate])
summary(anova1)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)    
## Age            1  16.60  16.596  87.813  < 2e-16 ***
## Sex            1  14.67  14.674  77.645  < 2e-16 ***
## RestingBP      1   1.58   1.581   8.365  0.00394 ** 
## FastingBS      1   0.59   0.587   3.104  0.07852 .  
## MaxHR          1  10.95  10.950  57.938 8.27e-14 ***
## Cholesterol    1   2.06   2.060  10.897  0.00101 ** 
## Residuals    739 139.67   0.189                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 172 observations deleted due to missingness
```

```
anova2 <- aov(HeartDisease~FastingBS,heart[,index_to_investigate])
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## FastingBS      1  16.21   16.21   70.48 <2e-16 ***
## Residuals    916 210.67    0.23                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova2 <- aov(HeartDisease~Sex,heart[,index_to_investigate])
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## Sex            1  21.17  21.168   94.25 <2e-16 ***
## Residuals    916 205.72   0.225                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova2 <- aov(HeartDisease~Age,heart[,index_to_investigate])
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## Age            1  18.05  18.048   79.16 <2e-16 ***
## Residuals    916 208.84   0.228                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova2 <- aov(HeartDisease~Cholesterol,heart[,index_to_investigate])
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)   
## Cholesterol    1   2.01  2.0078   8.114 0.00451 **
## Residuals    744 184.10  0.2475                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 172 observations deleted due to missingness
```

```
anova2 <- aov(HeartDisease~MaxHR,heart[,index_to_investigate])
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## MaxHR          1  36.38   36.38   174.9 <2e-16 ***
## Residuals    916 190.51    0.21                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova2 <- aov(HeartDisease~RestingBP,heart[,index_to_investigate])
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)   
## RestingBP      1   2.63  2.6263   10.73 0.0011 **
## Residuals    916 224.26  0.2448                  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We conclude that the only insignificant indicator is Fasting Blood Sugar Level.