EDA

Group 10: Michelle Dai, Lucy Tian, Eli Wang

2022-10-06

```
library(foreign)
data <- read.dta("fram.dta")
```

Date Cleaning

According to the documentation of the Framingham Dataset, the original data has a total of 5209 observations and 18 variables recorded. For column `scl1`, we have a total of 2, 037 missing values. Thus for the sake of retaining as much data as possible for future analysis, we will drop the `scl1` column.

To deal with other missing values, we approach with the simple method of dropping observations that contain on or more missing values. Other methods such as imputation will be used later if see fit.

```
library(dplyr)
library(ggplot2)
data <- data[,-c(13)]
data <- data[complete.cases(data),]
```

A total of 4,568 observations and 17 columns are selected for exploratory data analysis.

Exploratory Data Analysis (EDA)

We perform EDA based on our primary and secondary questions.

Primary Question What factors contribute to the diagnosis of coronary heart disease (CHD)?

Secondary Question Is there an association between physiological features (example being height, weight, a

Since we first look at the diagnosis of CHD, create a new column `diag` where observations diagnosed with CHD = 1 and 0 otherwise

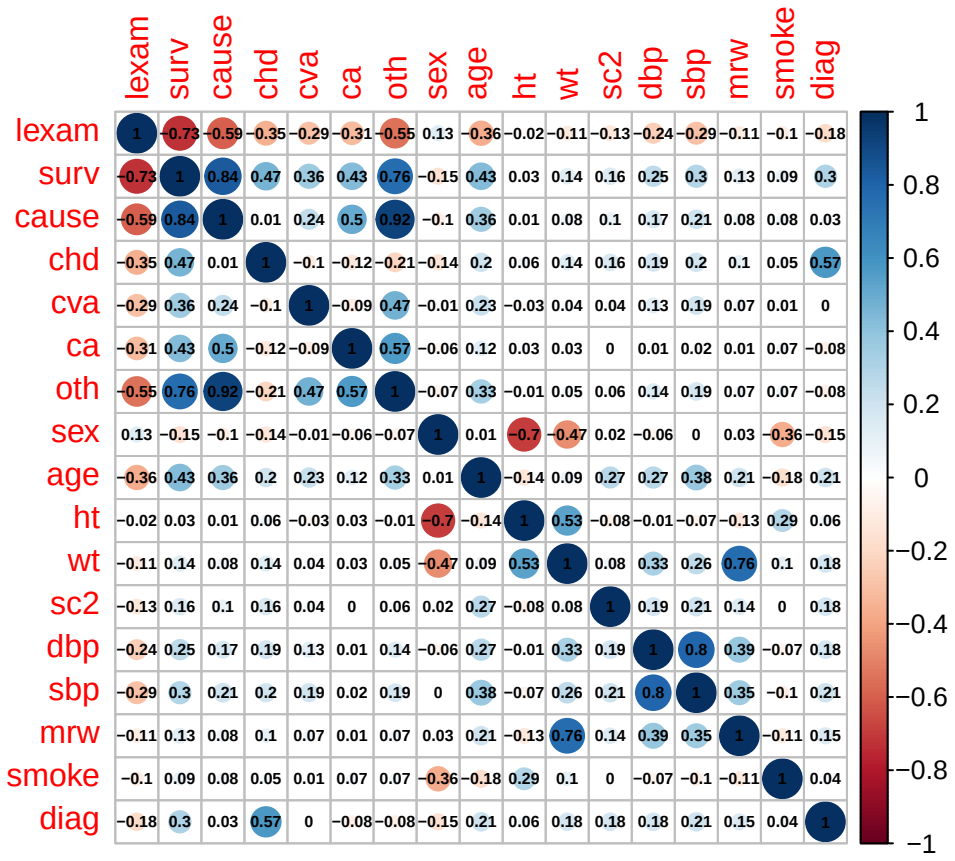
```
data <- data %>% mutate(diag = if_else(cexam == 0, 0, 1))
```

Simple calculation gives us CHD Diagnosis count = 1449 and CHD Death count = 605

We then plot a grid that calculates correlation between each variable pairs.

```
library(corrplot)
M = cor(data[, -c(4)])
corrplot(M, addCoef.col = 'black',  number.cex= 7/(ncol(data) - 4))
```

Correlation grid



The plot makes sense as we see very high correlation between variable pairs [`weight`, `mrw`, cor=0.76] and [`dbp`, `sbp`, cor=0.8]. `mrw` - Metropolitan Relative Weight- can be calculated by taking ratio of that person's body weight to the reference weight for that person's height, and systolic and diabolic pressures are highly correlated as they each represent the maximum pressure the heart exerts while beating and the amount of pressure in the arteries between beats.

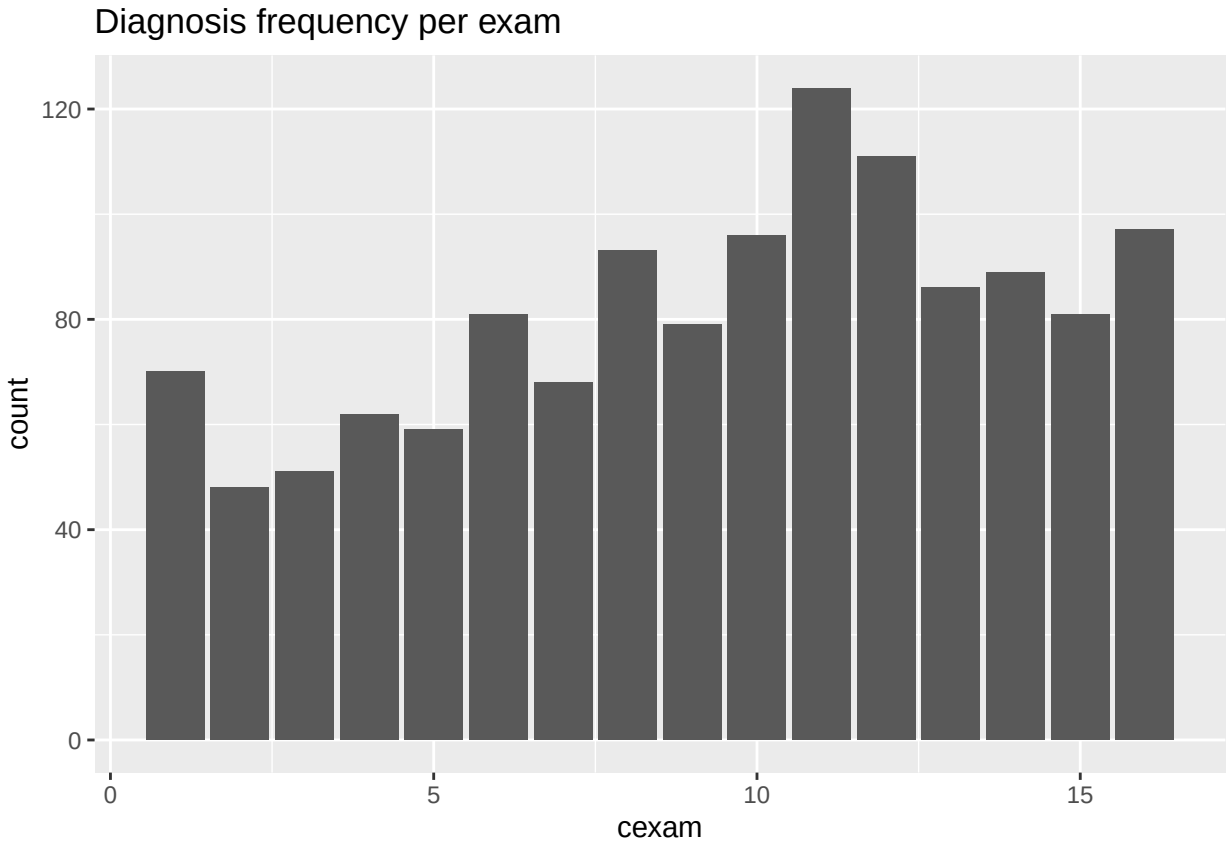
From the grid above, we see that diagnosis of CHD (`diag`) has a relatively higher correlation with `age`, `sbp` (Systolic blood pressure), and `mrw` (Metropolitan Relative Weight) with values 0.21, 0.21, and 0.15 respectively. Death from CHD (`chd`) has a relatively higher correlation with `age`, `sc2`, `sbp`, and `weight`, with values 0.2, 0.16, 0.2, and 0.14 respectively.

We proceed to examine distribution of important variables respectively

Investigating relationship within and between variables

```
ggplot(data[data$diag!=0,], aes(x=cexam)) + geom_bar() + ggtitle('Diagnosis frequency per exam')
```

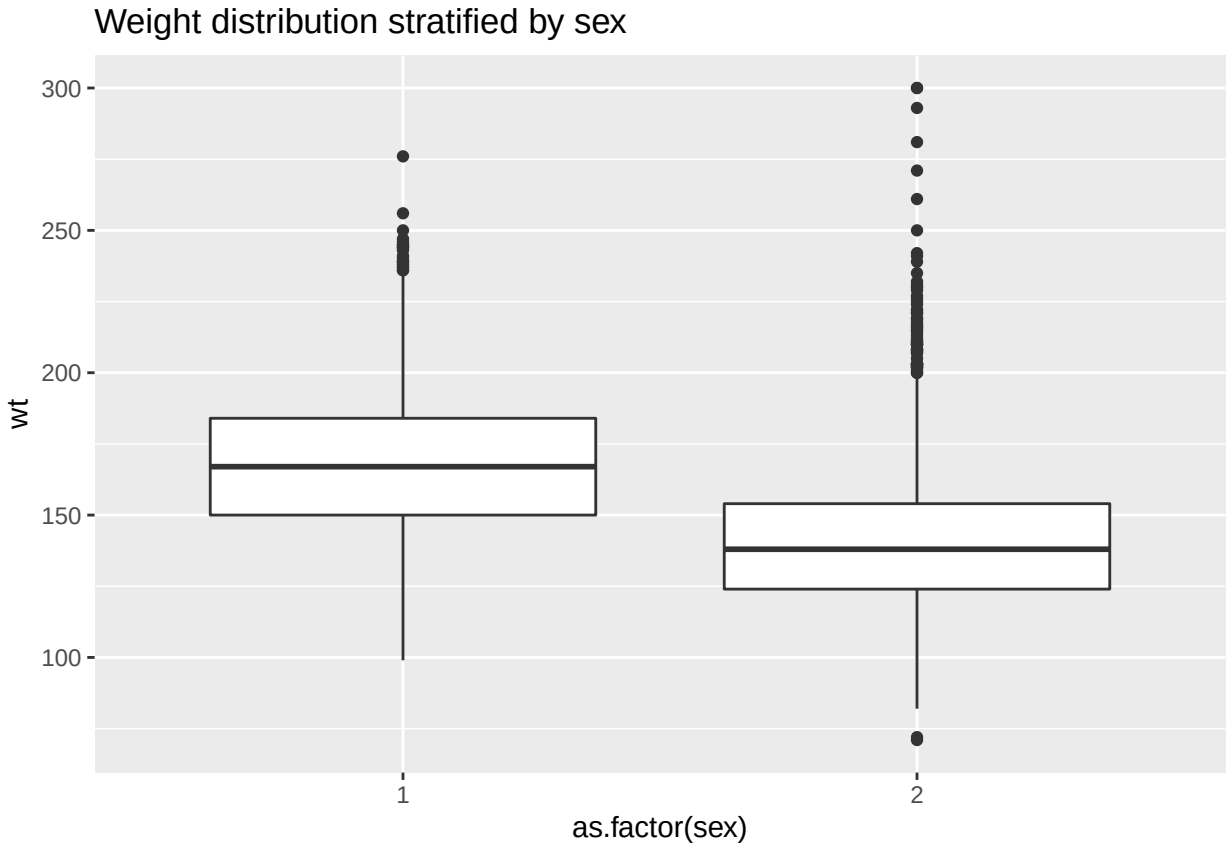
Diagnosis frequency per exam



From the bar plot above, we see for those diagnosed with CHD, the frequency of diagnosis for each exam generally increases until the 11th exam, which reached the maximum frequency o greater than 120. After the 11th exam, frequency of diagnosis show a downward trend.

```
ggplot(data, aes(x=as.factor(sex), y=wt)) + geom_boxplot()+ggtitle('Weight distribution stratified by s
```

Weight distribution stratidied by sex




```
summary(data[data$sex==1,]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    99.0   150.0   167.0   167.8   184.0   276.0
```

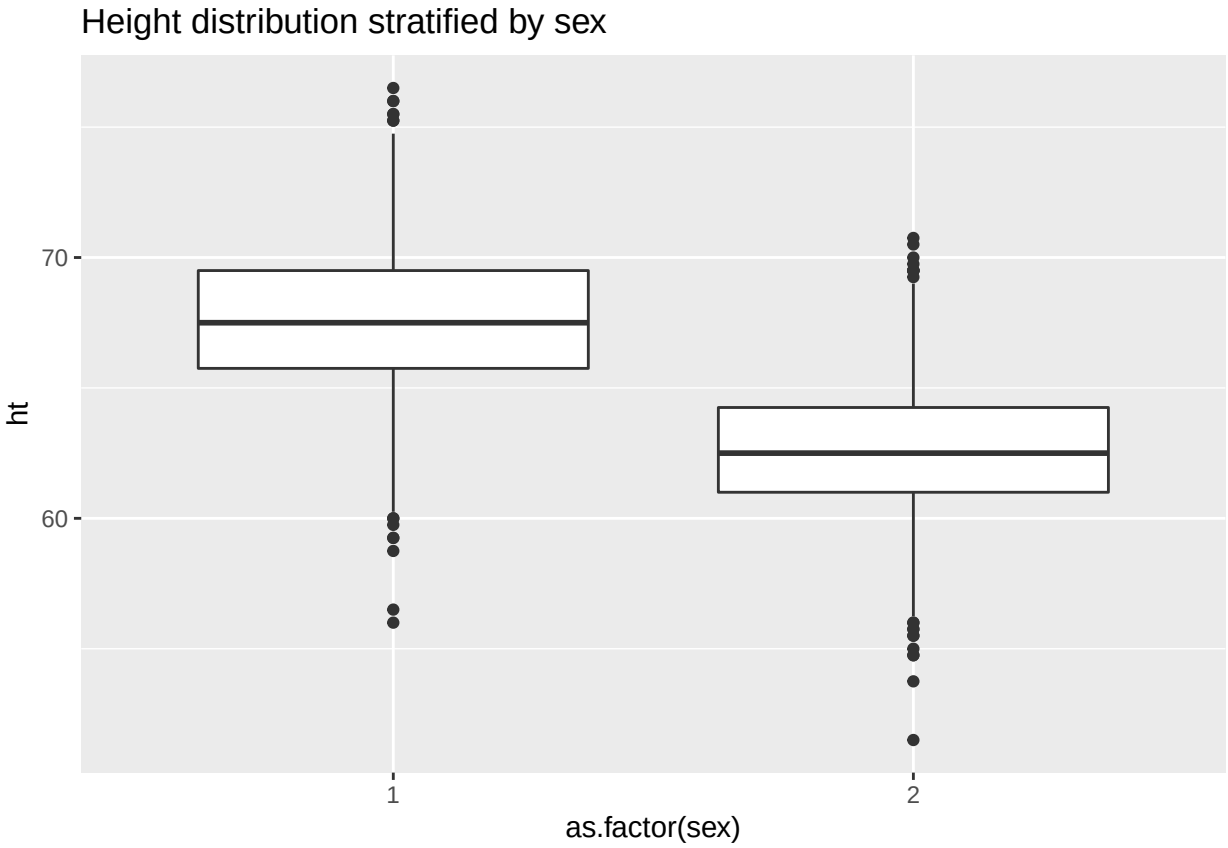
```
summary(data[data$sex==2,]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      71     124     138     141     154     300
```

From the side by side box plot above, we see females generally have a lower weight than males. Males have a median values of 167lbs and females 138lbs. In addition, females have more outliers compared to males and includes both the minimum (71lbs) and maximum (300lbs) values of the total sample of subjects studies.

```
ggplot(data, aes(x=as.factor(sex), y=ht)) + geom_boxplot()+ggtitle('Height distribution stratified by s
```

Height distribution stratified by sex




```
summary(data[data$sex==1,]$ht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   65.75   67.50   67.62   69.50   76.50
```

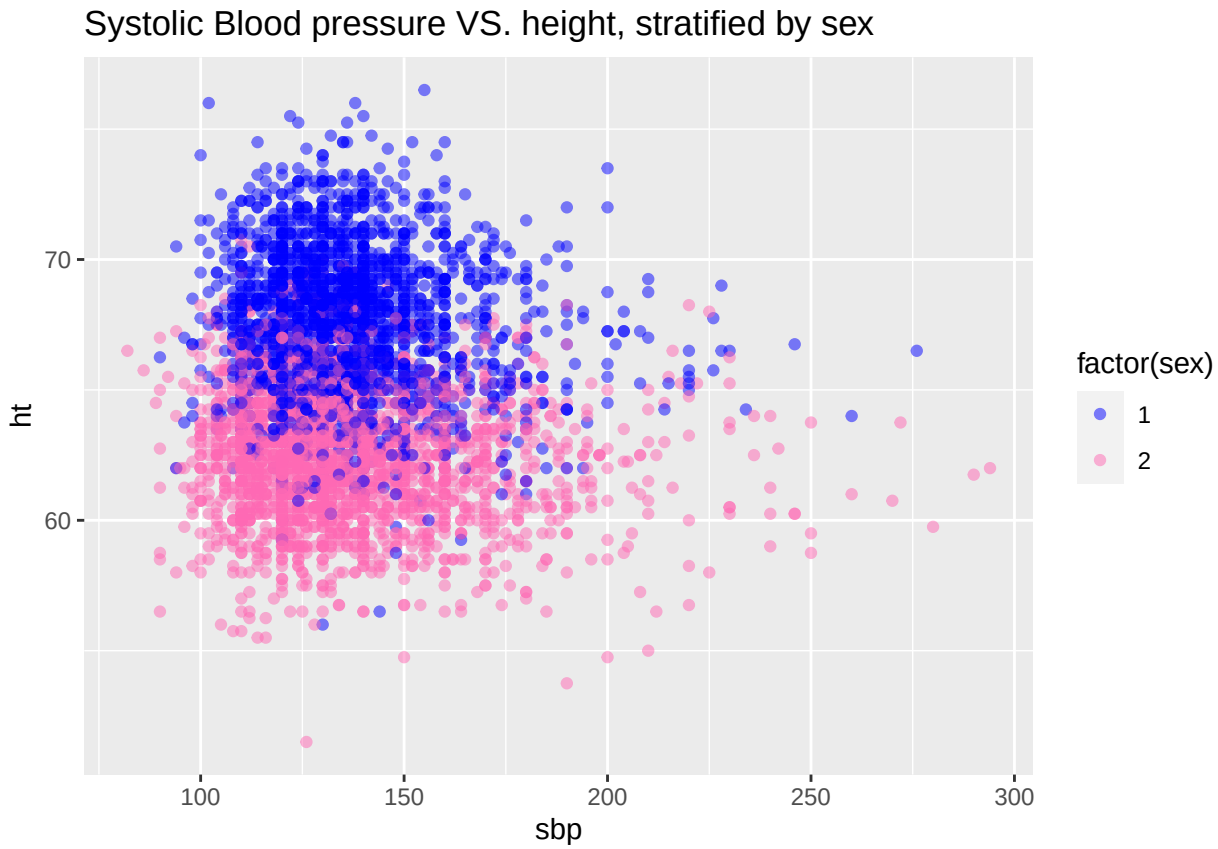
```
summary(data[data$sex==2,]$ht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.50   61.00   62.50   62.60   64.25   70.75
```

From the side by side box plot above, we see females generally have a lower height than males. Males have a median values of 67.5 inches and females 62.5 inches. The magnitude of their respective range is relatively the same.

```
ggplot(data = data, aes(x = sbp, y = ht, color = factor(sex))) + geom_point(alpha=0.5) + scale_color_ma
ggtitle('Systolic Blood pressure VS. height, stratified by sex')
```

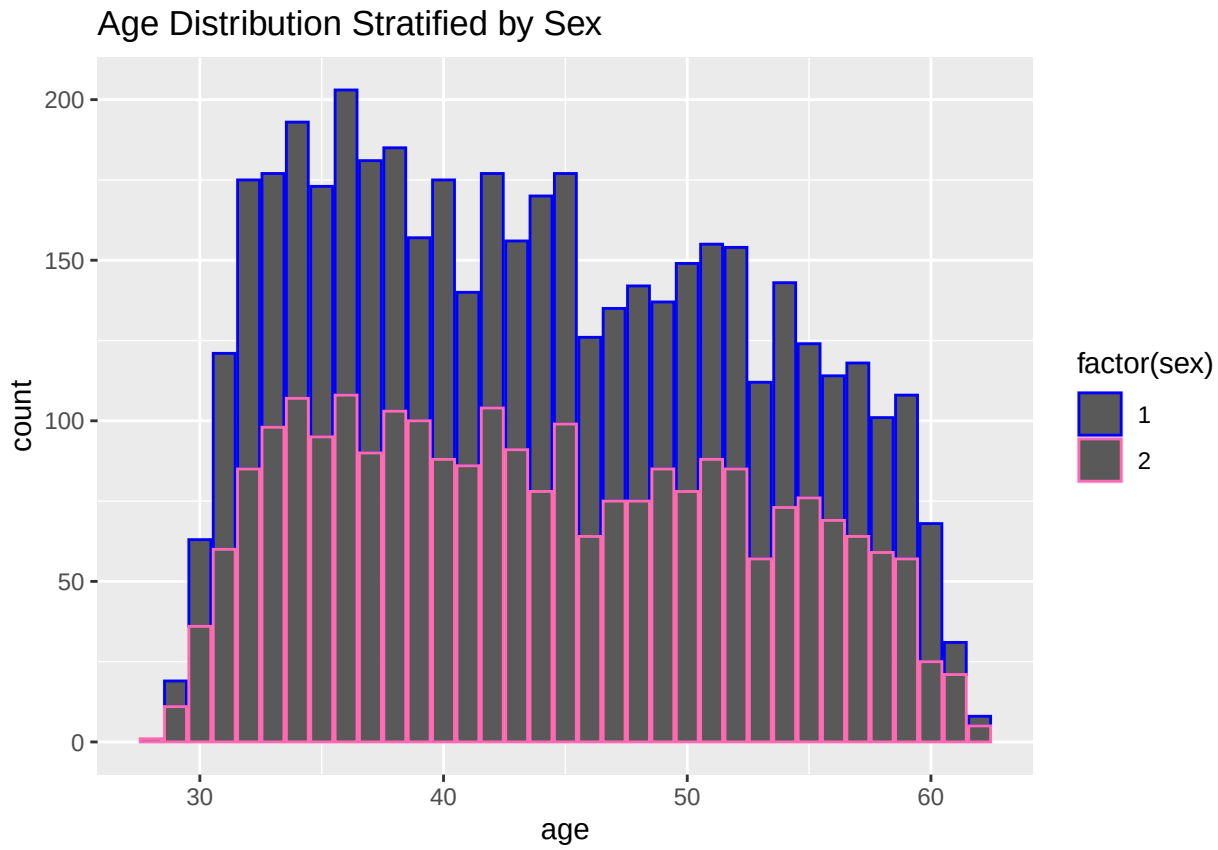
Systolic Blood pressure VS. height, stratified by sex



From the plot, we can see male and females have generally similar distribution in terms of systolic blood pressure. However, for subjects with `sbp` larger than 200, we see `sbq` increase with height in females while `sbq` decrease with height in males. Difference in height distribution among males and females is again validated.

```
ggplot(data=data, aes(x=age, color=factor(sex))) + geom_bar() + scale_color_manual(values = c("blue", "
```

Age distribution




```
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   37.00   43.00   44.02   51.00   62.00
```

```
summary(data[data$sex==1,]$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   36.00   44.00   43.95   51.00   62.00
```

```
summary(data[data$sex==2,]$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   37.00   43.00   44.08   51.00   62.00
```

From the bar plot above stratified by sex, we see male and female having generally similar, slightly right-skewed age distribution. The population age mean is 44.02, and mean age for male is slightly higher than that of female, with respective values 43.95 and 44.08. No obvious outliers observed.

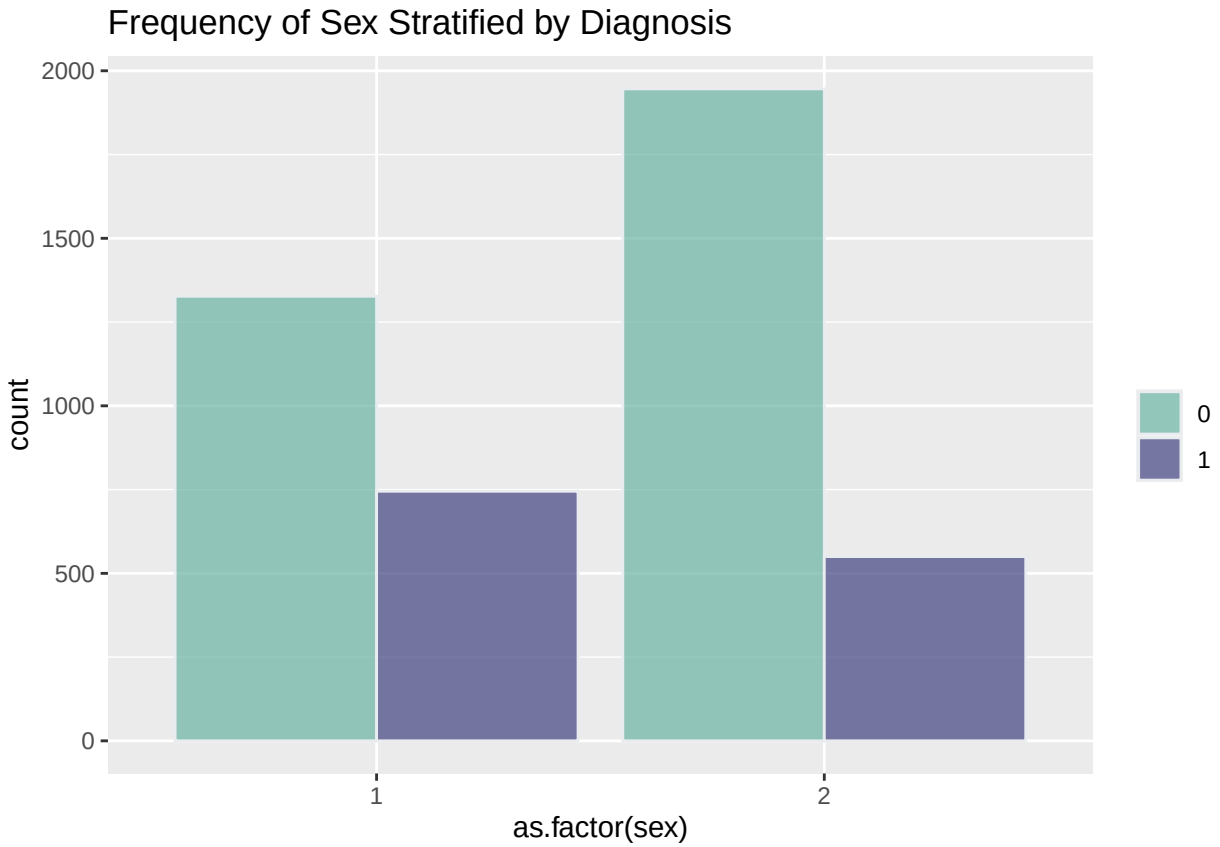
After investigating the relationship between each feature, we then delve into each one of them to compare the distribution among CHD diagnosed patients and CHD non-diagnosed patients.

Investigating relationship between variables and outcome (Diagnosis of CHD)

```
ggplot(data, aes(x=as.factor(sex), fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge', stat="count") +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Frequency of Sex Stratified by Diagnosis")
```

Distribution of Sex Stratified by Diagnosis

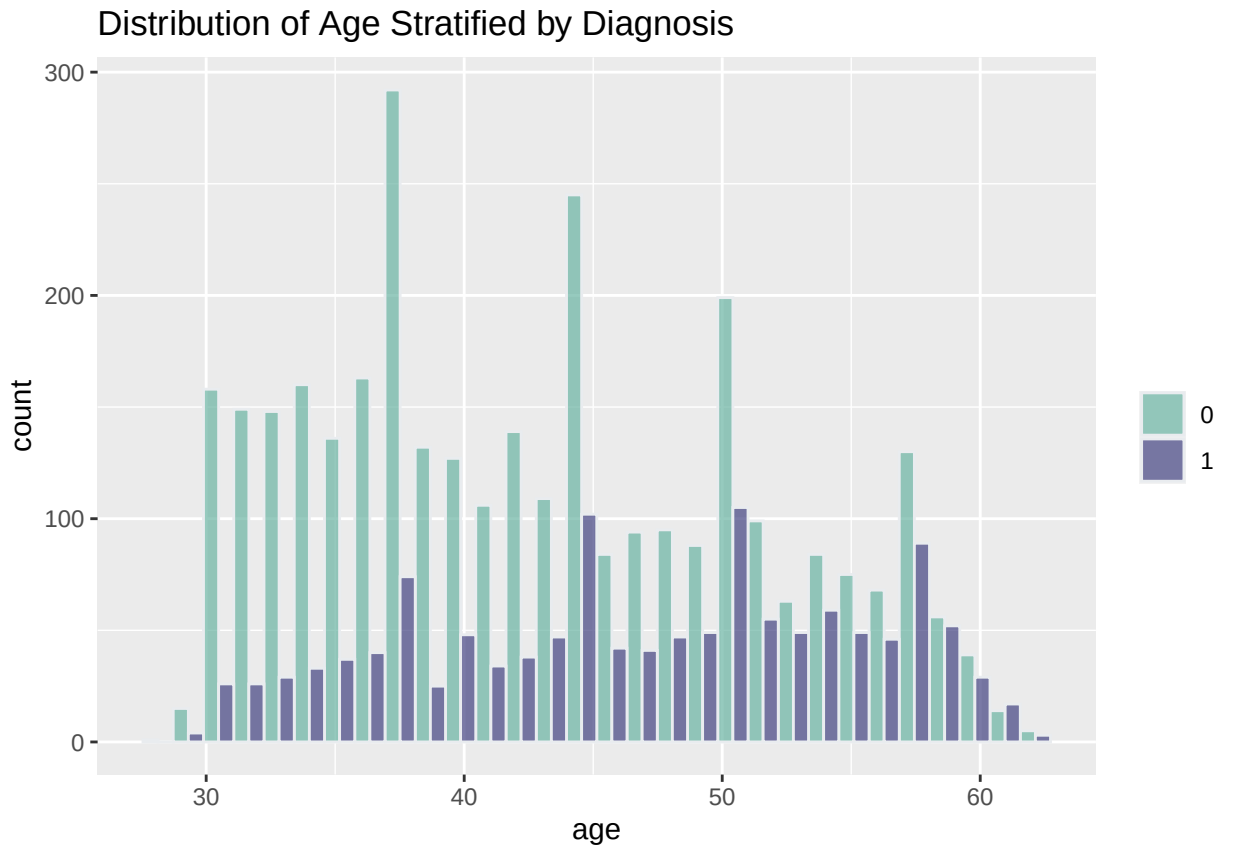
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



From the histogram above, we see that higher proportion of males (sex=1) are diagnosed compared to females. This suggest sex may be a potential contributor to diagnosis and require further investigation.

```
ggplot(data, aes(x=age, fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Age Stratified by Diagnosis")
```

Distribution of Age Stratified by Diagnosis




```
summary(data[data$diag == 0,]$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    28.0    36.0    42.0    42.9    50.0    62.0
```

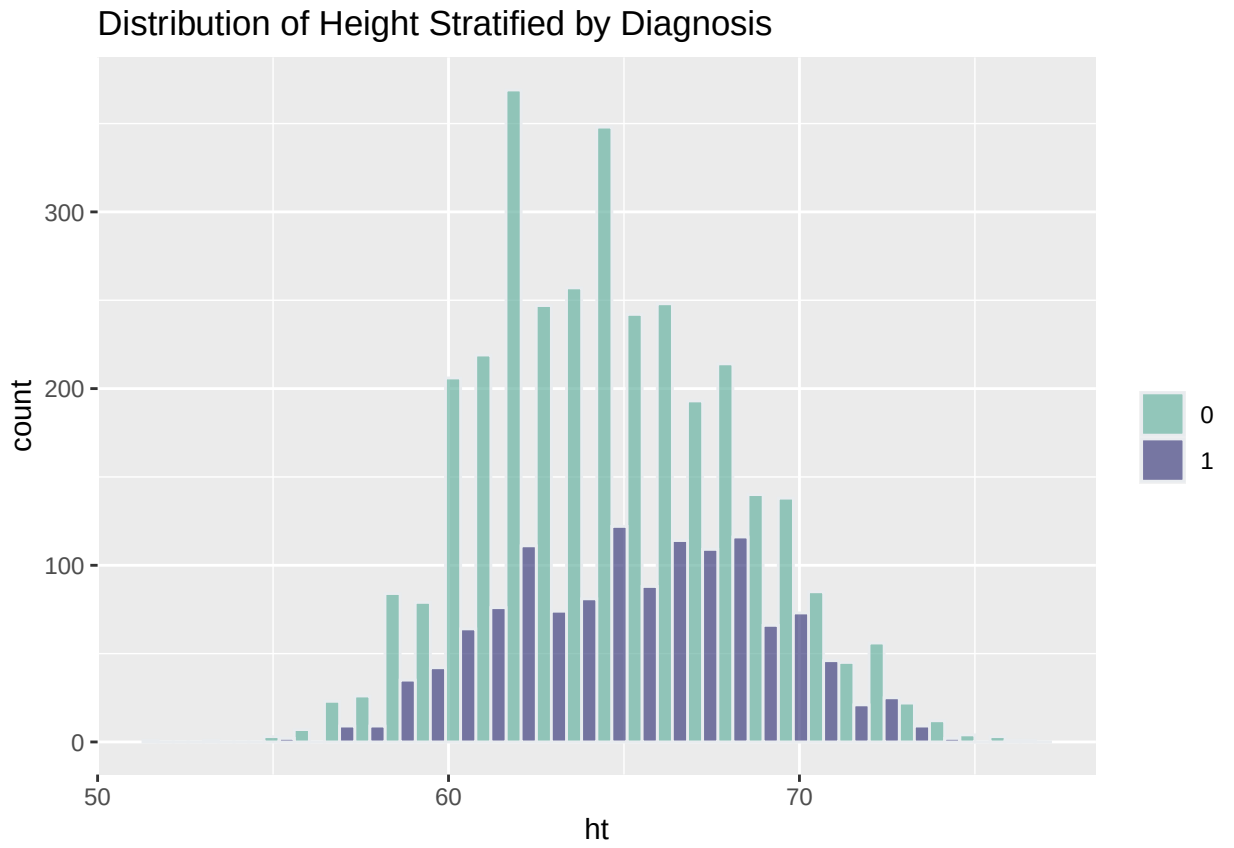
```
summary(data[data$diag == 1,]$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   40.00   48.00   46.84   54.00   62.00
```

From the histogram, we can observe that the distribution of age among CHD diagnosed patients is more skewed to the left as compared with the distribution of age among non-diagnosed patients. The mean of age among diagnosed patients is 46.84402, and the mean of age among non-diagnosed patients is 42.90192.

```
ggplot(data, aes(x=ht, fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Height Stratified by Diagnosis")
```

Distribution of Height Stratified by Diagnosis




```
summary(data[data$diag == 0,]$ht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.50   62.00   64.50   64.73   67.25   76.50
```

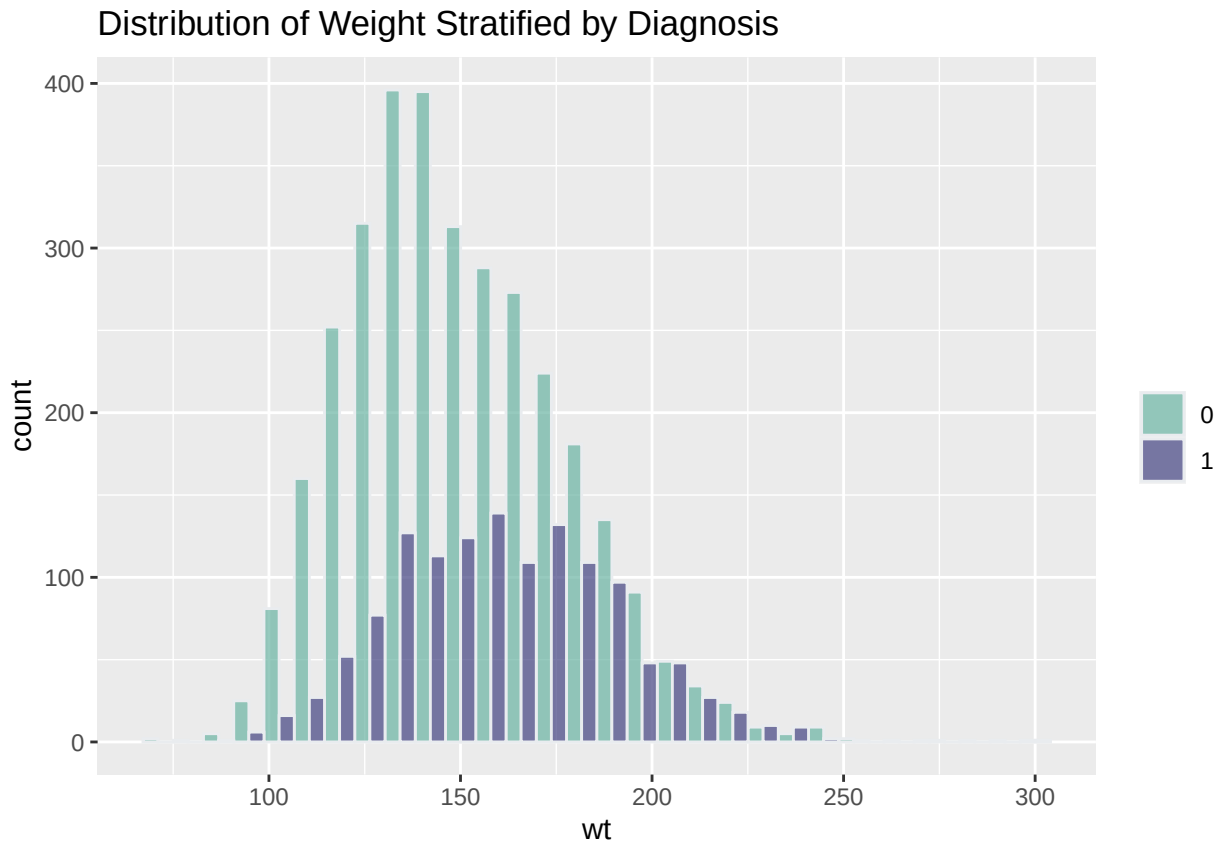
```
summary(data[data$diag == 1,]$ht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.75   62.50   65.25   65.25   68.00   75.50
```

The distribution of Height stratified by diagnosis looks pretty normal for both groups. The mean of height among diagnosed patients is 65.24691, and the mean of age among non-diagnosed patients is 64.73006.

```
ggplot(data, aes(x=wt, fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Weight Stratified by Diagnosis")
```

Distribution of Weight Stratified by Diagnosis




```
summary(data[data$diag == 0,]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   130.0   147.0   149.9   168.0   300.0
```

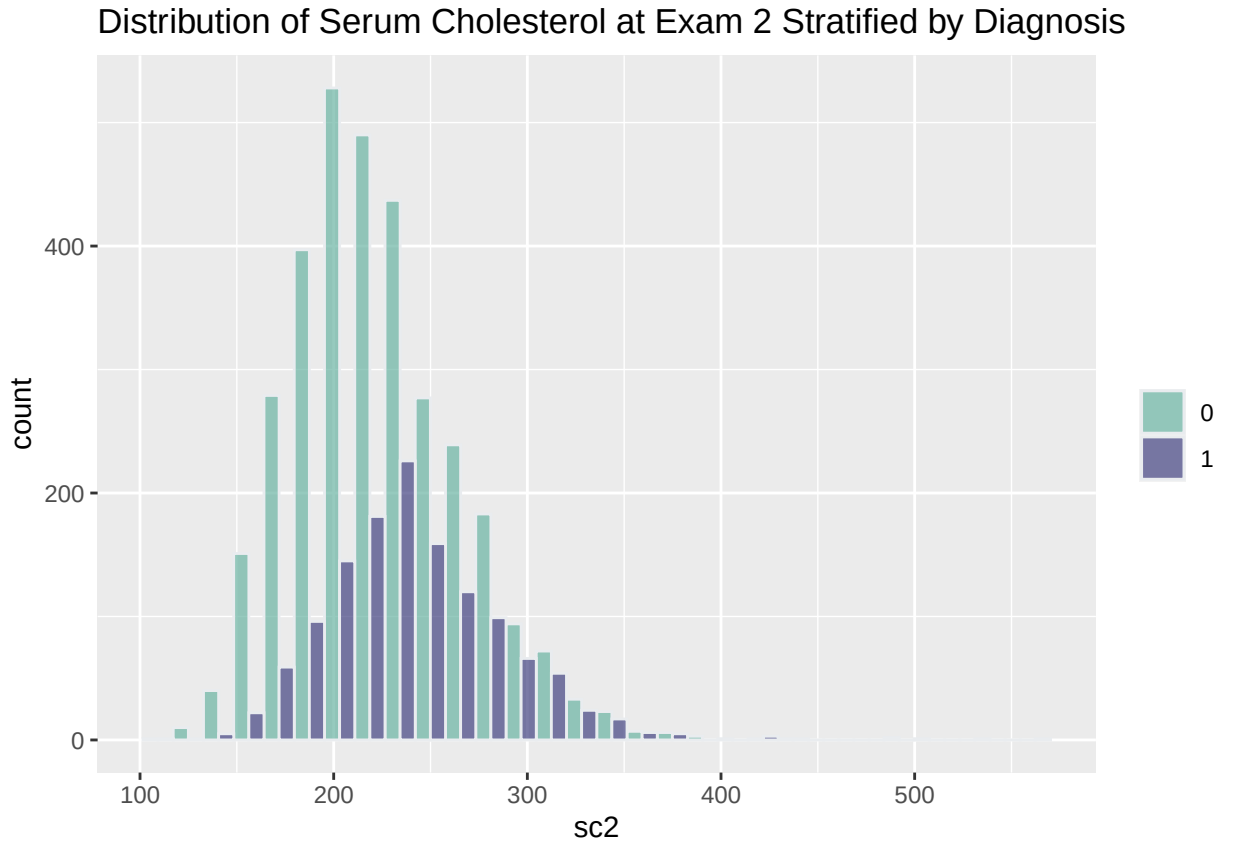
```
summary(data[data$diag == 1,]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    83.0   139.0   159.0   161.4   180.0   300.0
```

The distribution of weight for CHD diagnosed patients is more normal as compared with the distribution of weight for CHD non-diagnosed patients. We observe a right-skewed distribution on the weight for non-diagnosed patients with mean of 149.9206 (161.3761 for diagnosed).

```
ggplot(data, aes(x=sc2, fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Serum Cholesterol at Exam 2 Stratified by Diagnosis")
```

Distribution of Serum Cholesterol at Exam 2 Stratified by Diagnosis




```
summary(data[data$diag == 0,]$sc2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     115     194     219     223     248     435
```

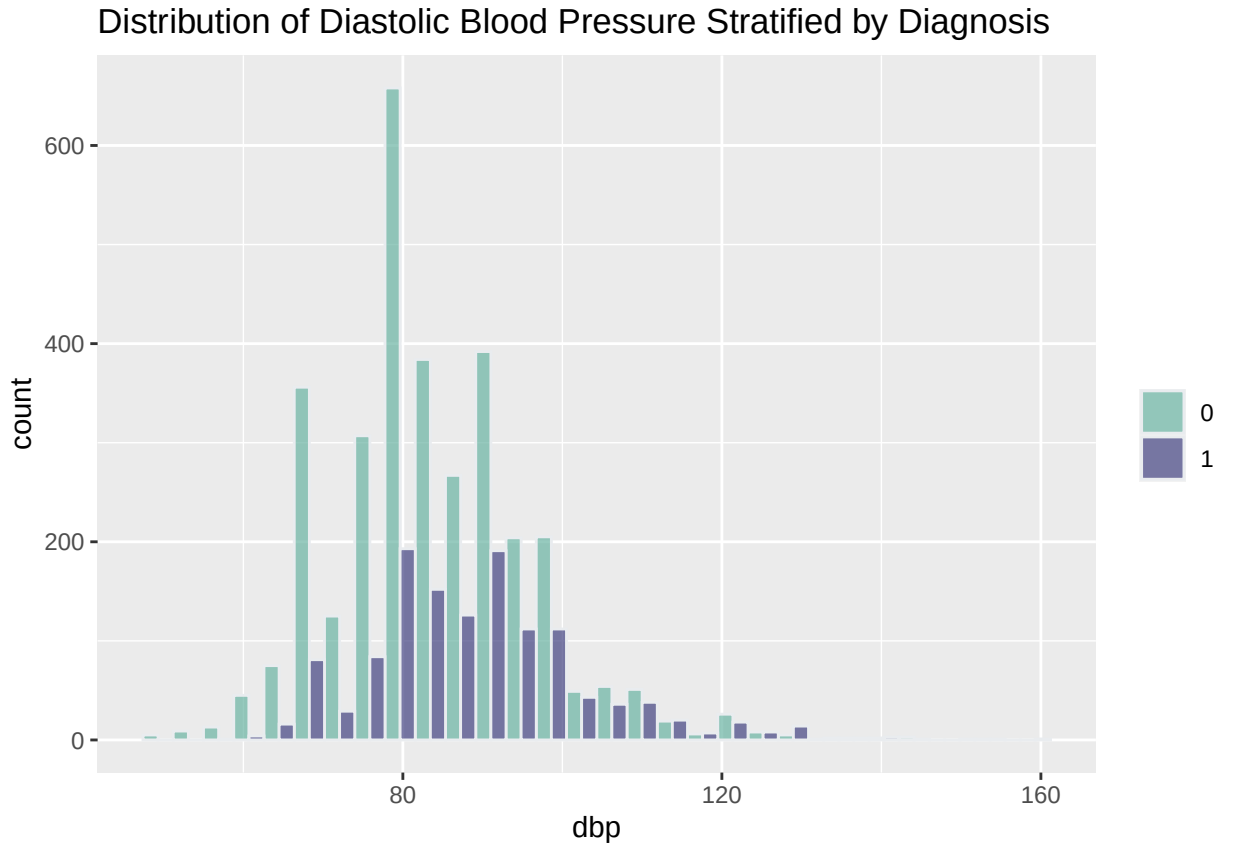
```
summary(data[data$diag == 1,]$sc2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   117.0   210.0   236.0   241.3   267.0   568.0
```

From the histogram, we can observe that the distribution for Serum Cholesterol at Exam 2 among CHD non-diagnosed patients is more skewed to the right as compared with the distribution of Serum Cholesterol at Exam 2 among diagnosed patients. The mean of Serum Cholesterol at Exam 2 among diagnosed patients is 241.3, and the mean of Serum Cholesterol at Exam 2 among non-diagnosed patients is 223.

```
ggplot(data, aes(x=dbp, fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Diastolic Blood Pressure Stratified by Diagnosis")
```

Distribution of Diastolic Blood Pressure Stratified by Diagnosis




```
summary(data[data$diag == 0,]$dbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.00   76.00   82.00   83.78   90.00  155.00
```

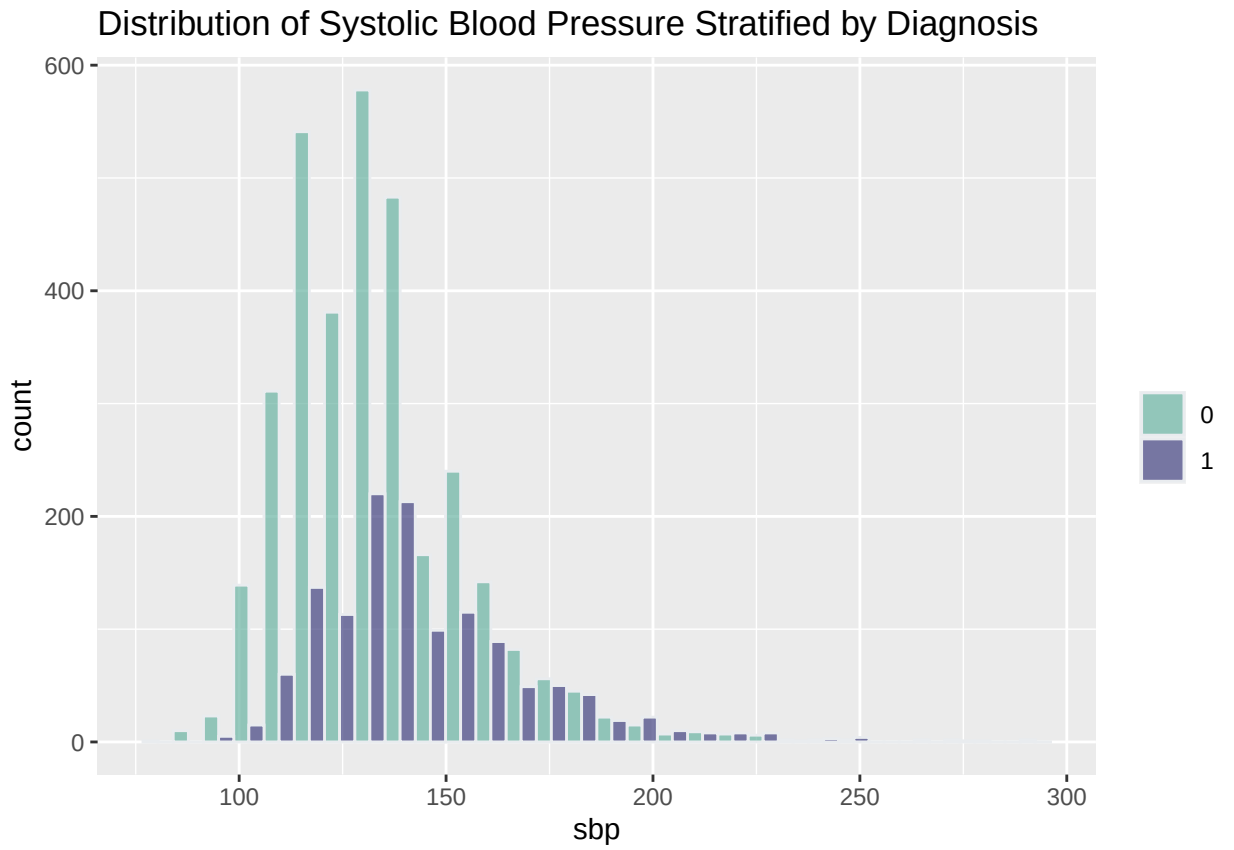
```
summary(data[data$diag == 1,]$dbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   80.00   88.00   89.06   96.00  160.00
```

From the histogram, the distributions for Diastolic Blood Pressure among both groups are pretty normal. The distribution for Diastolic Blood Pressure among CHD non-diagnosed patients is comparatively more skewed to the right as compared with the distribution of Diastolic Blood Pressure among diagnosed patients. The mean of Diastolic Blood Pressure among diagnosed patients is 89.06, and the mean of Diastolic Blood Pressure among non-diagnosed patients is 83.78.

```
ggplot(data, aes(x=sbp, fill=as.factor(diag))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Systolic Blood Pressure Stratified by Diagnosis")
```

Distribution of Systolic Blood Pressure Stratified by Diagnosis




```
summary(data[data$diag == 0,]$sbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    82.0   120.0   130.0   133.5   142.0   294.0
```

```
summary(data[data$diag == 1,]$sbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   126.0   140.0   144.2   156.0   276.0
```

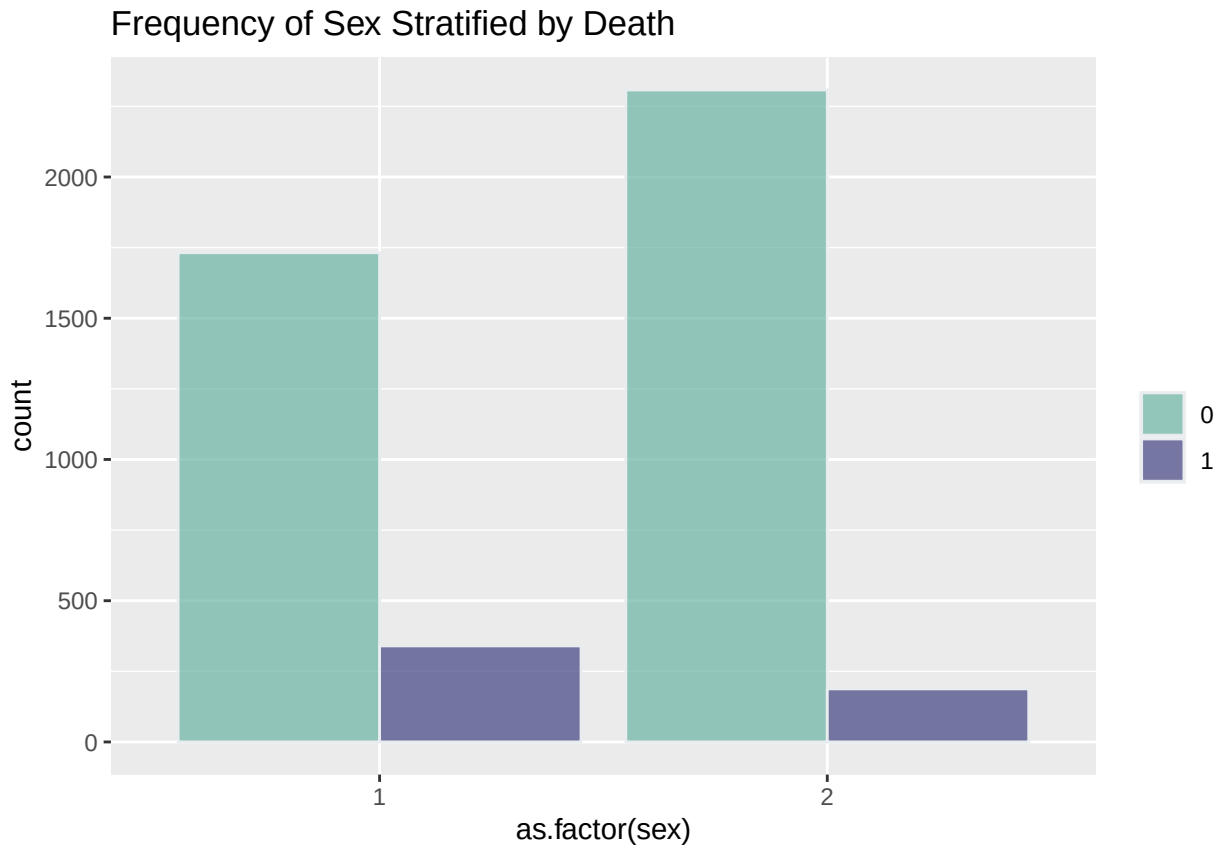
From the histogram, we can observe that the distribution for Systolic Blood Pressure among CHD non-diagnosed patients is more skewed to the right as compared with the distribution of Systolic Blood Pressure among diagnosed patients. The mean of Systolic Blood Pressure among diagnosed patients is 144.2, and the mean of Systolic Blood Pressure among non-diagnosed patients is 133.5.

Investigating relationship between variables and outcome (Death by CHD)

```
ggplot(data, aes(x=as.factor(sex), fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge', stat="count") +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Frequency of Sex Stratified by Death")
```

Frequency of Sex Stratified by Death

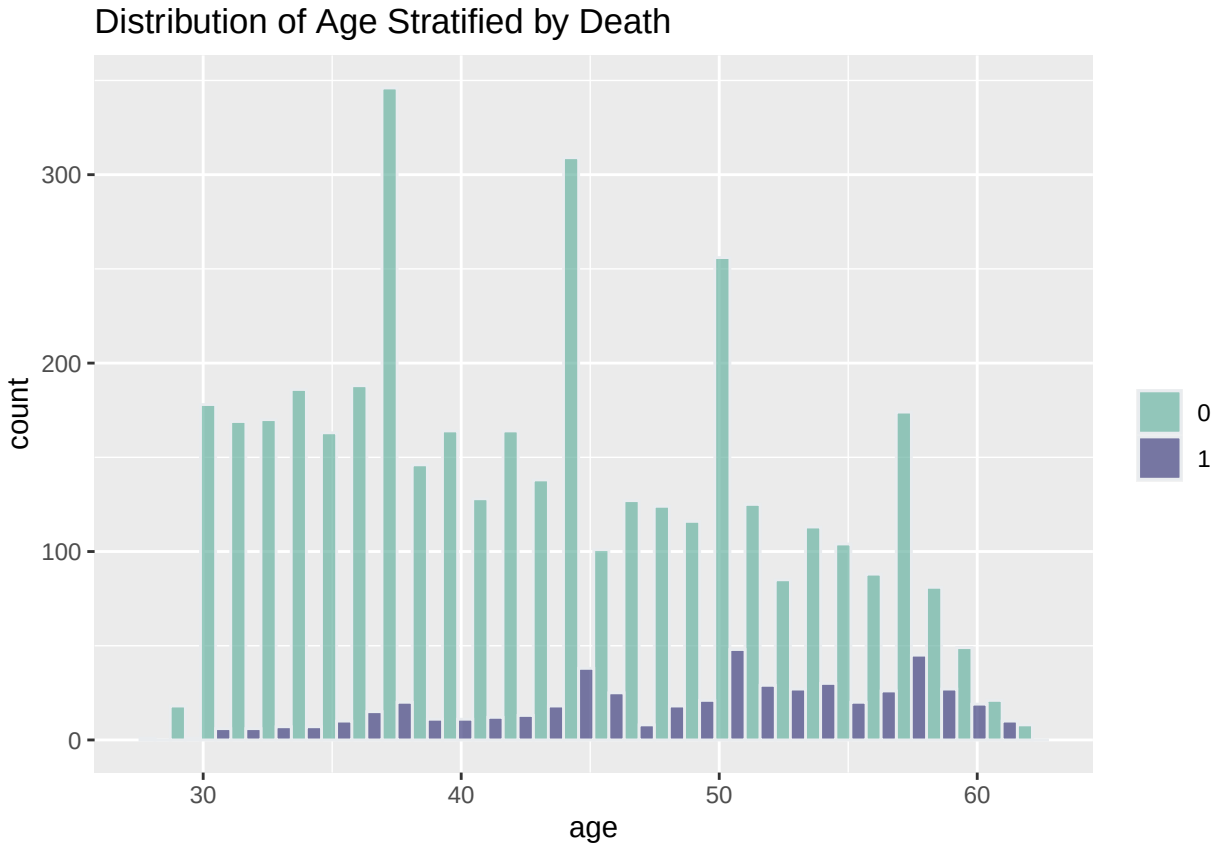
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



From the histogram above, we see that higher proportion of males (sex=1) result in death from CHD compared to females. This suggest sex may be a potential contributor to death from CHD require further investigation.

```
ggplot(data, aes(x=age, fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Age Stratified by Death")
```

Distribution of Age Stratified by Death




```
summary(data[data$chd==0,]$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.00   36.00   42.00   43.39   50.00   62.00
```

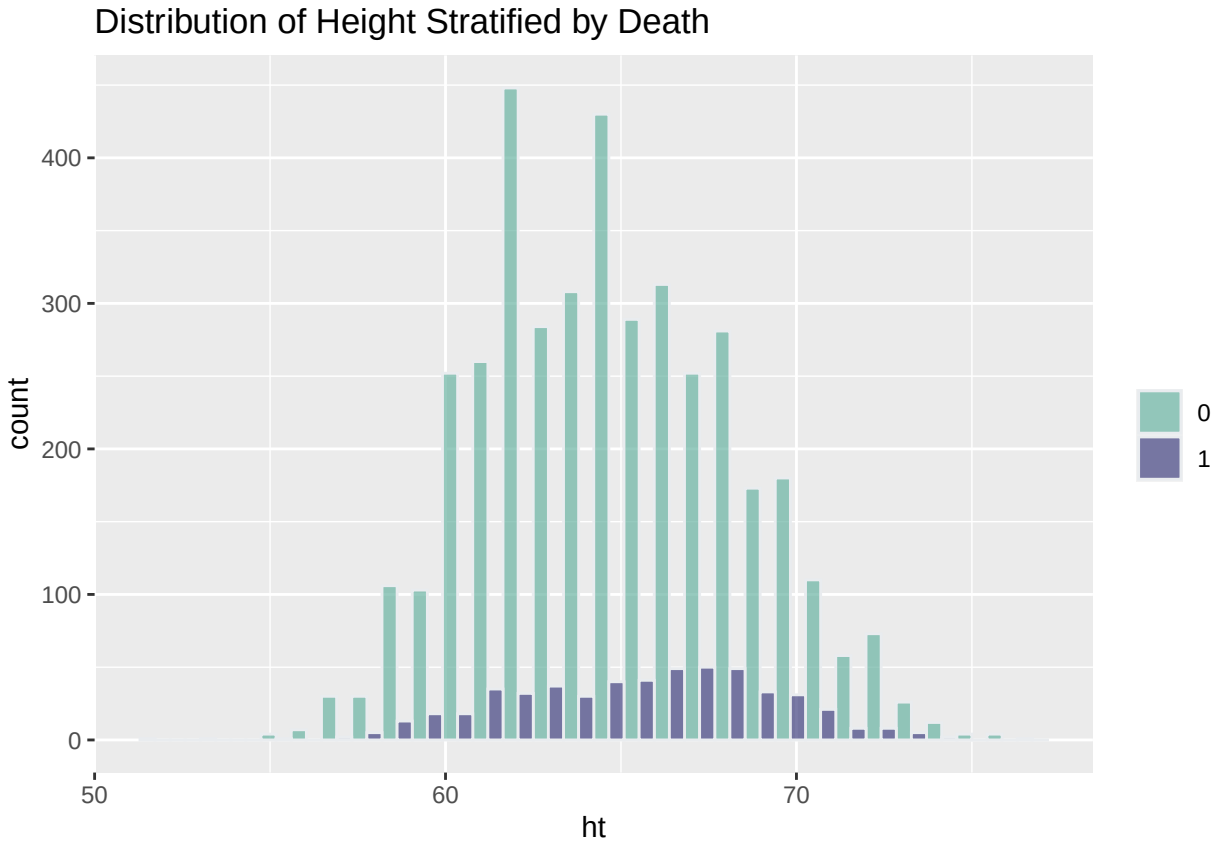
```
summary(data[data$chd==1,]$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   43.00   50.00   48.81   55.00   61.00
```

From the histogram above, we see clearly different distribution of age among the alive and dead. For those who are alive with CHD (`chd`=0), the distribution of age is right skewed. For those who resulted in death from CHD (`chd`=1), the distribution of age is right skewed. This suggest that age may have a positive correlation with death from CHD. This can be further validated by the fact that alive patients have an age mean smaller that dead patients (43.49 vs 48.81)

```
ggplot(data, aes(x=ht, fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Height Stratified by Death")
```

Distribution of Height Stratified by Death




```
summary(data[data$chd==0,]$ht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.50   62.25   64.50   64.80   67.25   76.50
```

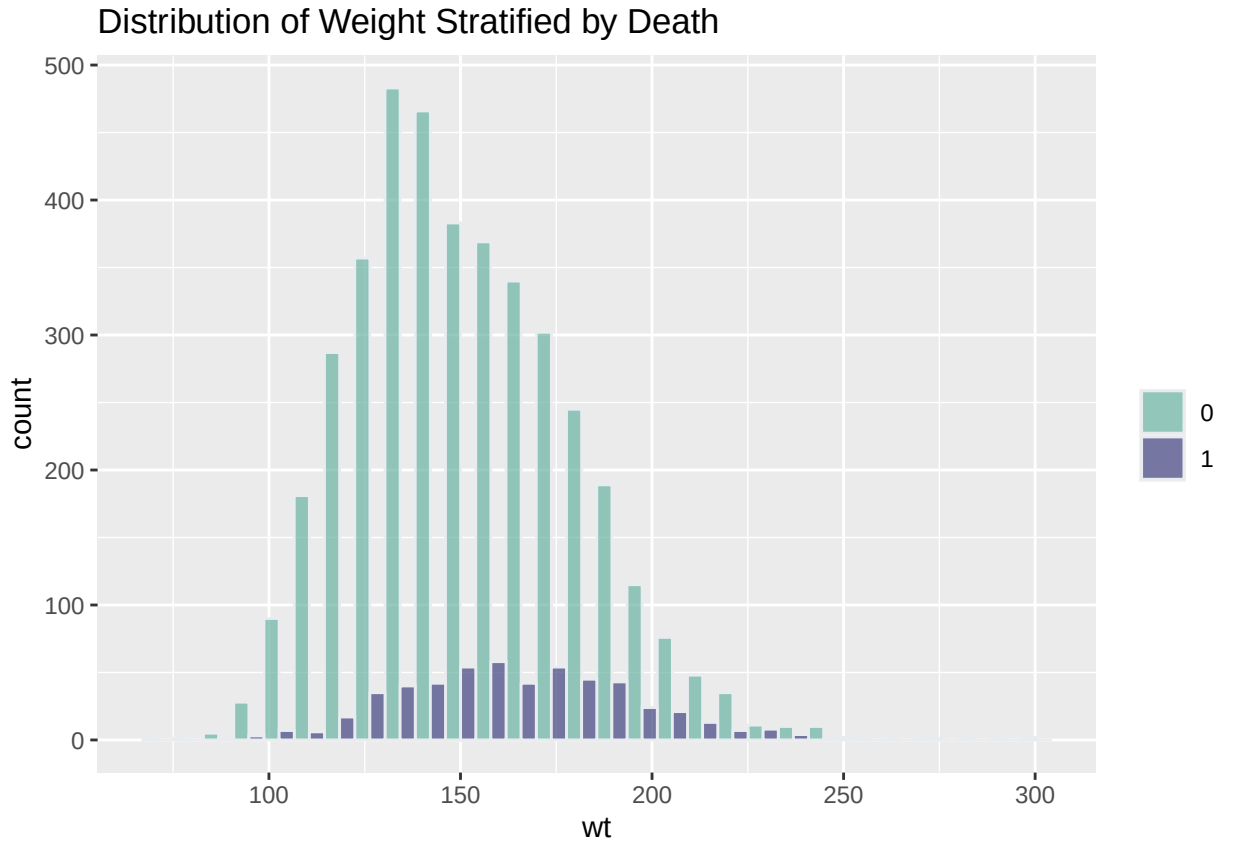
```
summary(data[data$chd==1,]$ht)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   62.75   65.75   65.47   68.25   74.50
```

From the histogram above, we see that alive CHD patients' height follow a roughly normal distribution, while those with death from CHD have a height distribution slightly left skewed. This suggest that height may have a positive correlation with death from CHD.

```
ggplot(data, aes(x=wt, fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Weight Stratified by Death")
```

Distribution of Weight Stratified by Death




```
summary(data[data$chd==0,]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   132.0   149.0   151.8   170.0   300.0
```

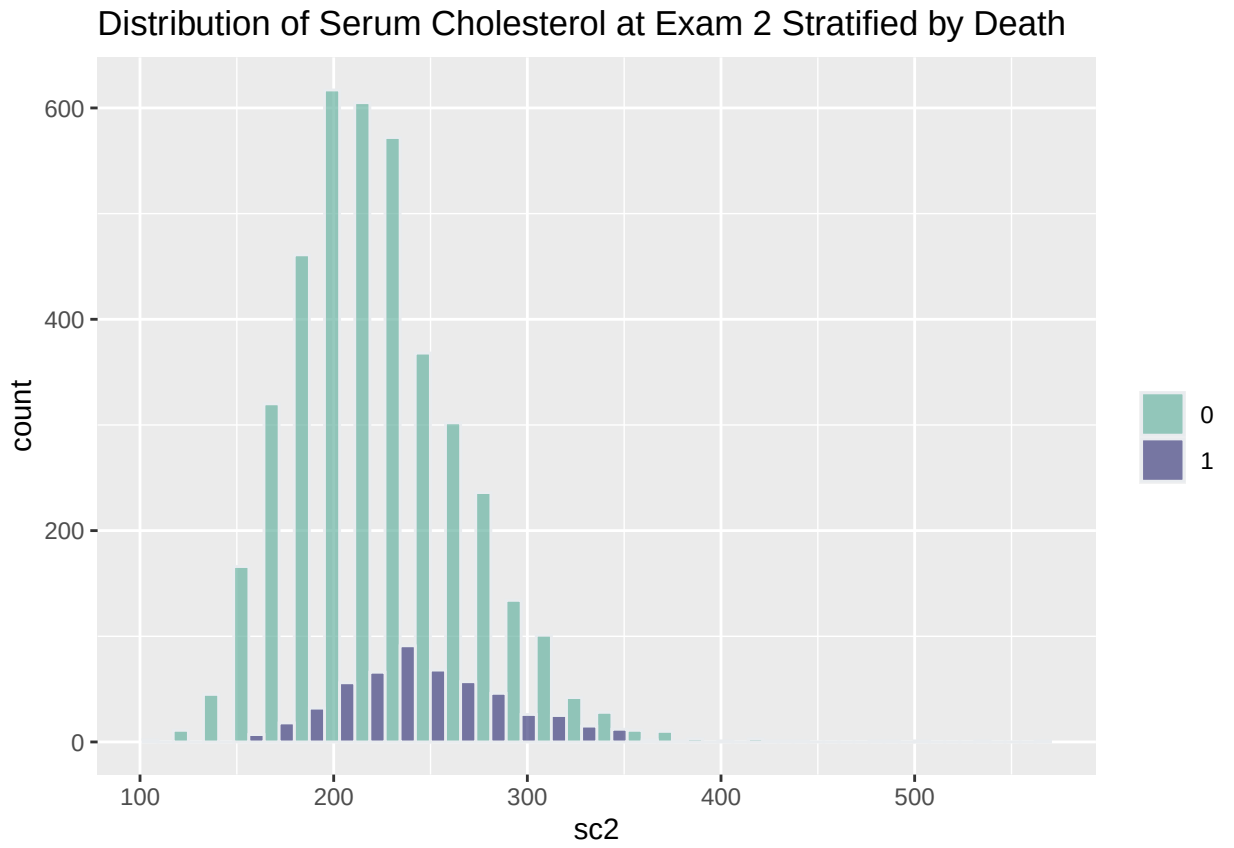
```
summary(data[data$chd==1,]$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    83.0   142.0   162.0   163.9   183.0   271.0
```

From the histogram above, we see that alive CHD patients' weight follow a slightly right-skewed distribution, while those with death from CHD have a height distribution slightly left skewed. This suggest that weight may have a positive correlation with death from CHD. We validate this by further comparing the respective means where alive patients have an average weight of 151.8lbs compared to that of dead patients with 163.9lbs.

```
ggplot(data, aes(x=sc2, fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Serum Cholesterol at Exam 2 Stratified by Death")
```

Distribution of Serum Cholesterol at Exam 2 Stratified by Death




```
summary(data[data$chd==0,]$sc2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   115.0   196.0   221.0   225.6   250.0   492.0
```

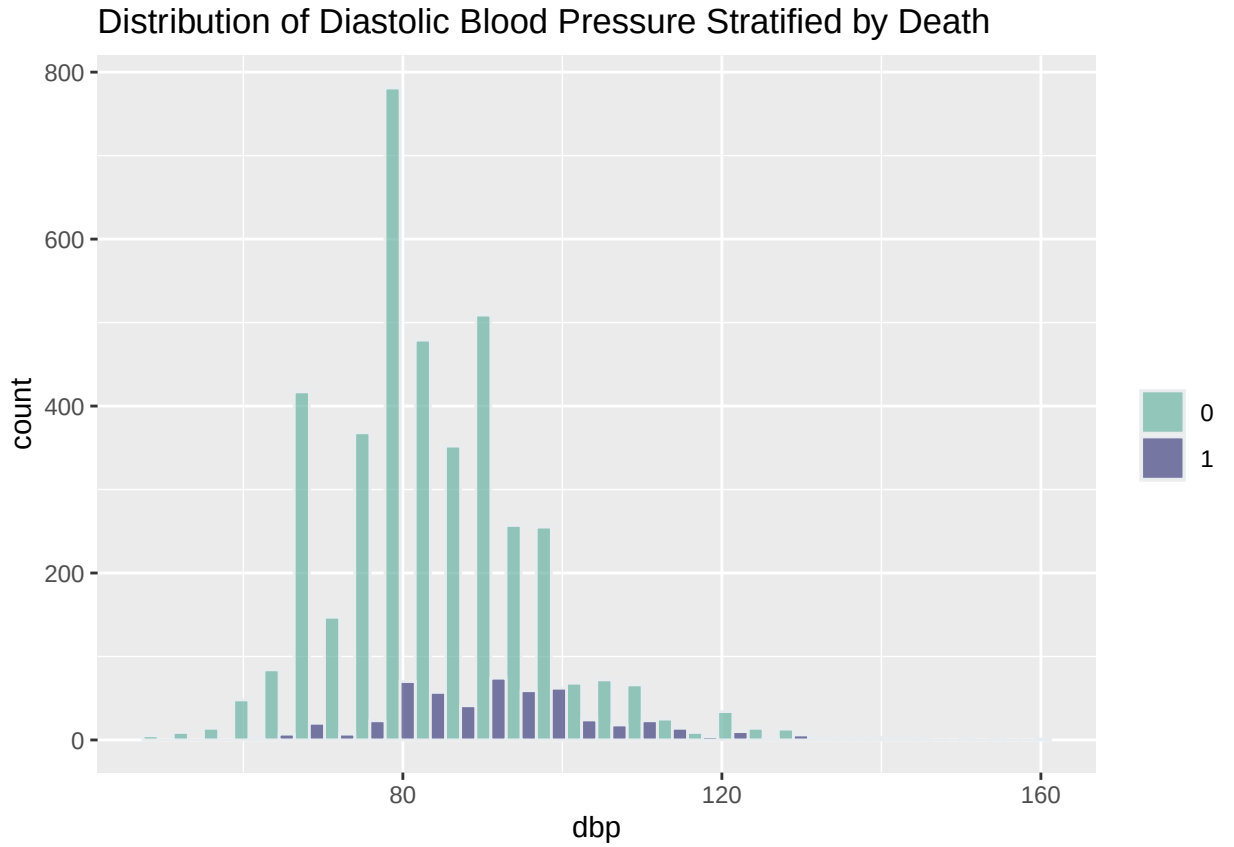
```
summary(data[data$chd==1,]$sc2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   150.0   217.8   242.0   248.2   274.2   568.0
```

From the histogram above, we see that alive CHD patients' Serum Cholesterol at Exam 2 follow a slightly right-skewed distribution, while those with death from CHD have a Serum Cholesterol at Exam 2 distribution roughly normal. This suggest that Serum Cholesterol at Exam 2 may have a positive correlation with death from CHD. We validate this by further comparing the respective means where alive patients have an average `sc2` of 225.6 compared to that of dead patients with 248.2.

```
ggplot(data, aes(x=dbp, fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Diastolic Blood Pressure Stratified by Death")
```

Distribution of Diastolic Blood Pressure Stratified by Death




```
summary(data[data$chd==0,]$dbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.00   76.00   82.00   84.41   90.00  155.00
```

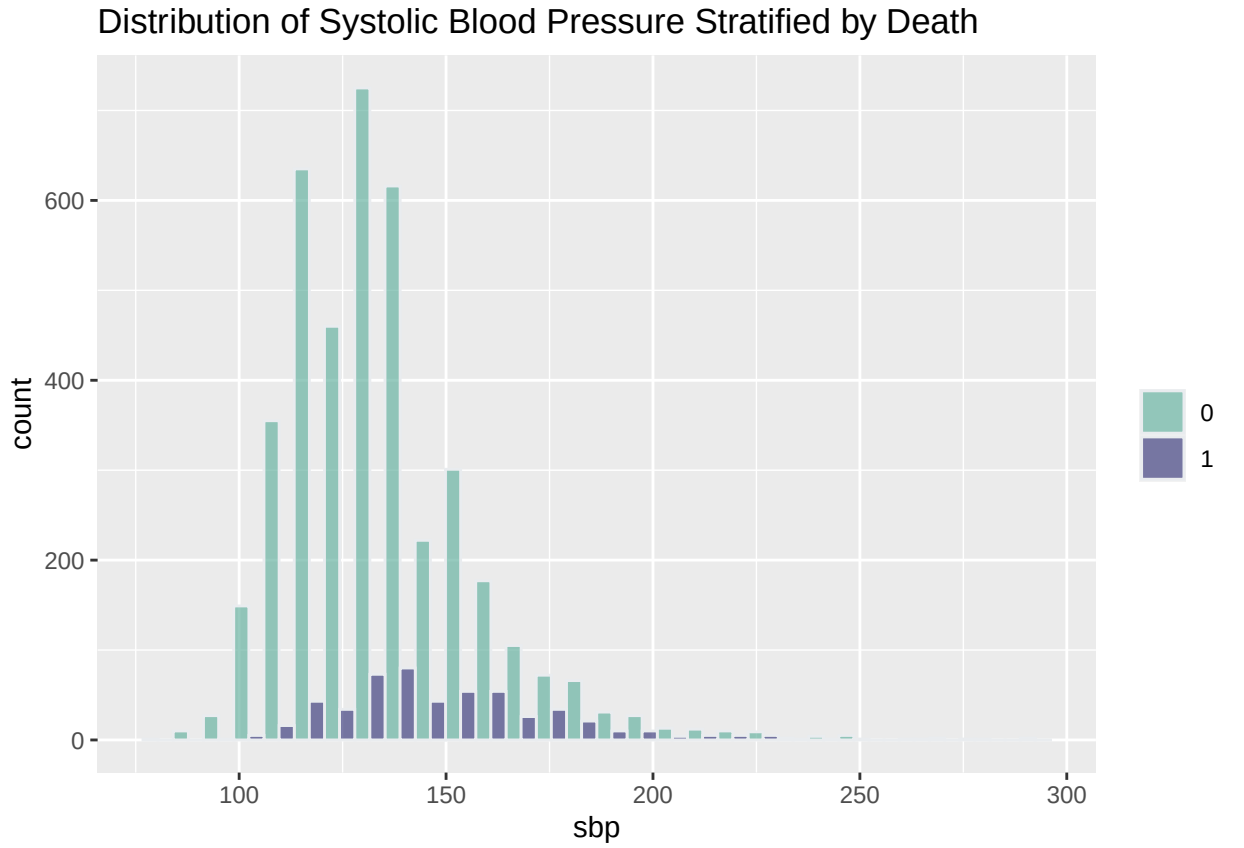
```
summary(data[data$chd==1,]$dbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.00   82.00   90.00   91.95  100.00  160.00
```

From the histogram above, we see that both alive and dead CHD patients' Diastolic Blood Pressure follow a generally normal distribution. However, the range for dead CHD patients' Diastolic Blood Pressure is more to the right. The mean Diastolic Blood Pressure for alive patients is lower than that compared to dead patients (84.41 vs. 91.95). This suggest that `dbp` may have a positive correlation with death from CHD.

```
ggplot(data, aes(x=sbp, fill=as.factor(chd))) +
    geom_histogram( color="#e9ecef", alpha=0.7, position = 'dodge') +
    scale_fill_manual(values=c("#69b3a2", "#404080")) +
    labs(fill="", title= "Distribution of Systolic Blood Pressure Stratified by Death")
```

Distribution of Systolic Blood Pressure Stratified by Death




```
summary(data[data$chd==0,]$sbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    82.0   120.0   130.0   134.8   144.0   294.0
```

```
summary(data[data$chd==1,]$sbp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   130.0   144.0   149.5   164.0   276.0
```

From the histogram above, we see that both alive CHD patients' Systolic Blood Pressure follow a slightly right-skewed distribution, while dead CHD patients' Systolic Blood Pressure follow a normal distribution. The range value for dead CHD patients' Systolic Blood Pressure is more to right. The mean Systolic Blood Pressure for alive patients is lower than that compared to dead patients (134.8 vs. 149.5). This suggest that `sbp` may have a positive correlation with death from CHD.