EDA

Group 10

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

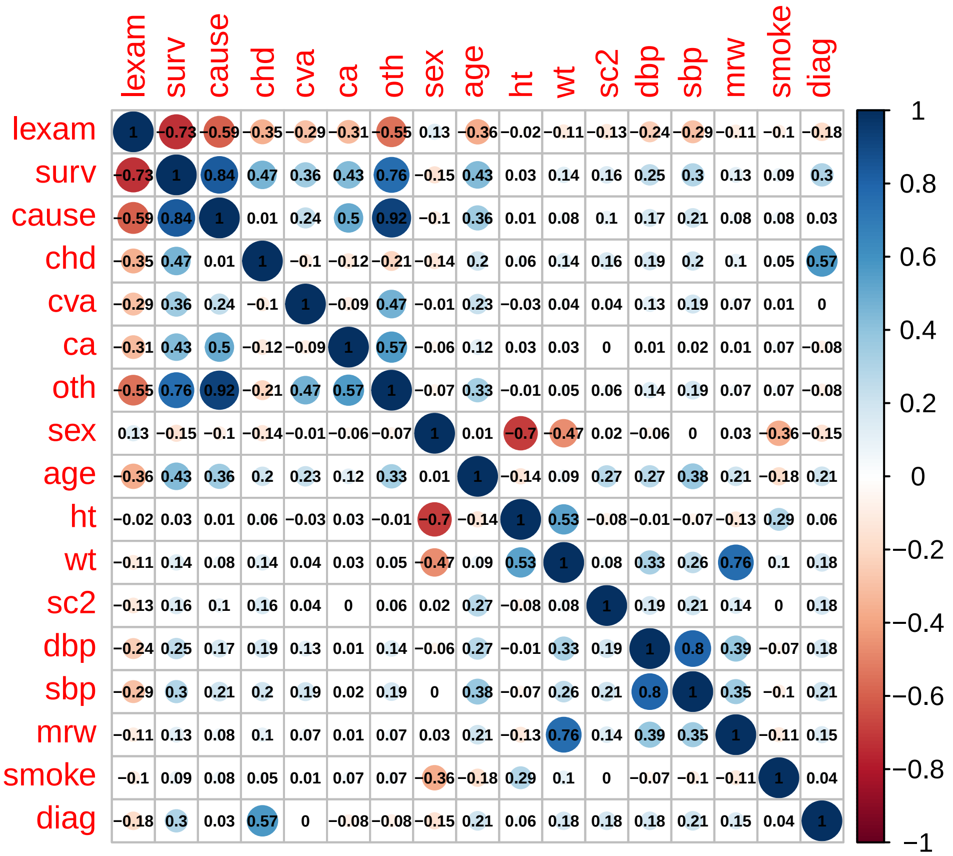
```
data <- data %>% mutate(diag = if_else(cexam == 0, 0, 1))
```

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



	lexam	surv	cause	chd	cva	ca	oth	sex	age	ht	wt	sc2	dbp	sbp	mrw	smoke	diag
lexam	1	−0.73	−0.59	−0.35	−0.29	−0.31	−0.55	0.13	−0.36	−0.02	−0.11	−0.13	−0.24	−0.29	−0.11	−0.1	−0.18
surv	−0.73	1	0.84	0.47	0.36	0.43	0.76	−0.15	0.43	0.03	0.14	0.16	0.25	0.3	0.13	0.09	0.3
cause	−0.59	0.84	1	0.01	0.24	0.5	0.92	−0.1	0.36	0.01	0.08	0.1	0.17	0.21	0.08	0.08	0.03
chd	−0.35	0.47	0.01	1	−0.1	−0.12	−0.21	−0.14	0.2	0.06	0.14	0.16	0.19	0.2	0.1	0.05	0.57
cva	−0.29	0.36	0.24	−0.1	1	−0.09	0.47	−0.01	0.23	−0.03	0.04	0.04	0.13	0.19	0.07	0.01	0
ca	−0.31	0.43	0.5	−0.12	−0.09	1	0.57	−0.06	0.12	0.03	0.03	0	0.01	0.02	0.01	0.07	−0.08
oth	−0.55	0.76	0.92	−0.21	0.47	0.57	1	−0.07	0.33	−0.01	0.05	0.06	0.14	0.19	0.07	0.07	−0.08
sex	0.13	−0.15	−0.1	−0.14	−0.01	−0.06	−0.07	1	0.01	−0.7	−0.47	0.02	−0.06	0	0.03	−0.36	−0.15
age	−0.36	0.43	0.36	0.2	0.23	0.12	0.33	0.01	1	−0.14	0.09	0.27	0.27	0.38	0.21	−0.18	0.21
ht	−0.02	0.03	0.01	0.06	−0.03	0.03	−0.01	−0.7	−0.14	1	0.53	−0.08	−0.01	−0.07	−0.13	0.29	0.06
wt	−0.11	0.14	0.08	0.14	0.04	0.03	0.05	−0.47	0.09	0.53	1	0.08	0.33	0.26	0.76	0.1	0.18
sc2	−0.13	0.16	0.1	0.16	0.04	0	0.06	0.02	0.27	−0.08	0.08	1	0.19	0.21	0.14	0	0.18
dbp	−0.24	0.25	0.17	0.19	0.13	0.01	0.14	−0.06	0.27	−0.01	0.33	0.19	1	0.8	0.39	−0.07	0.18
sbp	−0.29	0.3	0.21	0.2	0.19	0.02	0.19	0	0.38	−0.07	0.26	0.21	0.8	1	0.35	−0.1	0.21
mrw	−0.11	0.13	0.08	0.1	0.07	0.01	0.07	0.03	0.21	−0.13	0.76	0.14	0.39	0.35	1	−0.11	0.15
smoke	−0.1	0.09	0.08	0.05	0.01	0.07	0.07	−0.36	−0.18	0.29	0.1	0	−0.07	−0.1	−0.11	1	0.04
diag	−0.18	0.3	0.03	0.57	0	−0.08	−0.08	−0.15	0.21	0.06	0.18	0.18	0.18	0.21	0.15	0.04	1

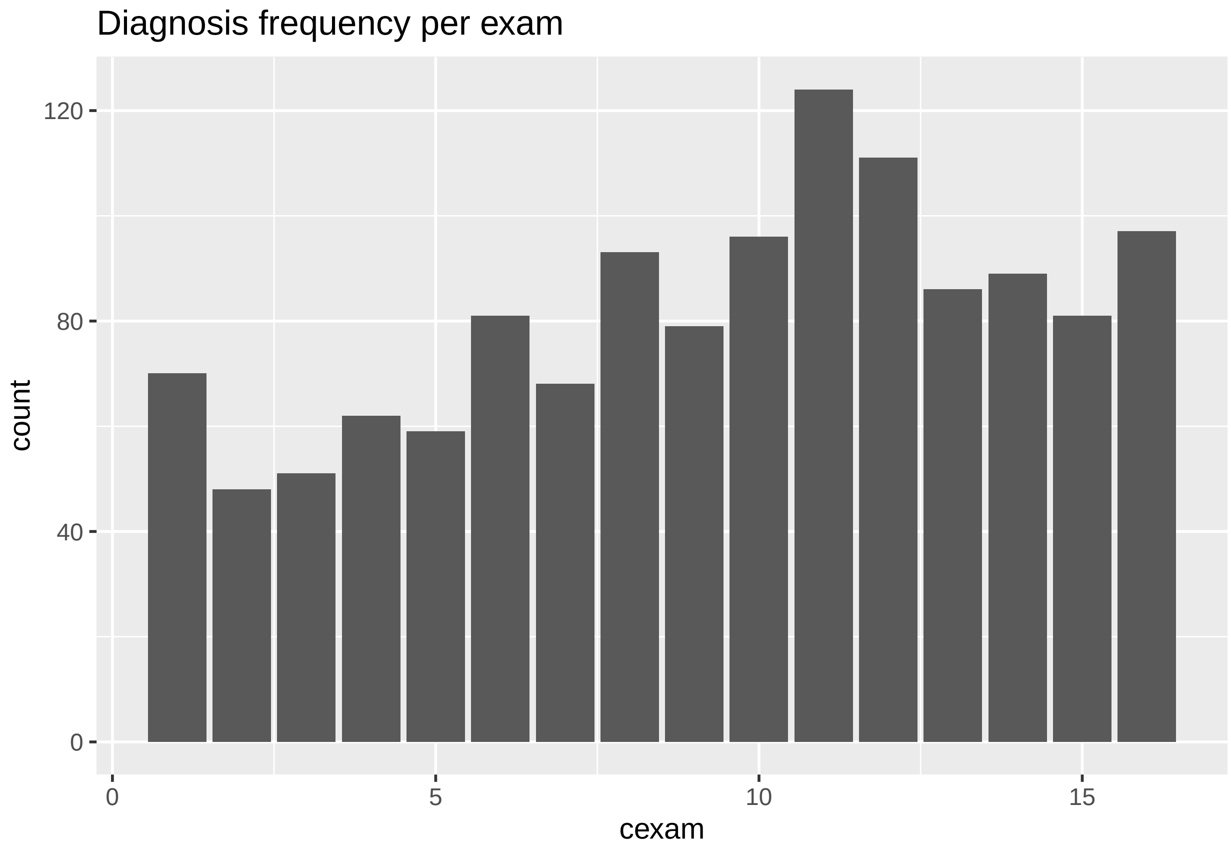
The plot makes sense as we see very high correlation between variable pairs [`weight`, `mrw`, cor=0.76] and [`dbp`, `sbp`, cor=0.8]. `mrw` - Metropolitan Relative Weight- can be calculated by taking ratio of that person's body weight to the reference weight for that person's height, and systolic and 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

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

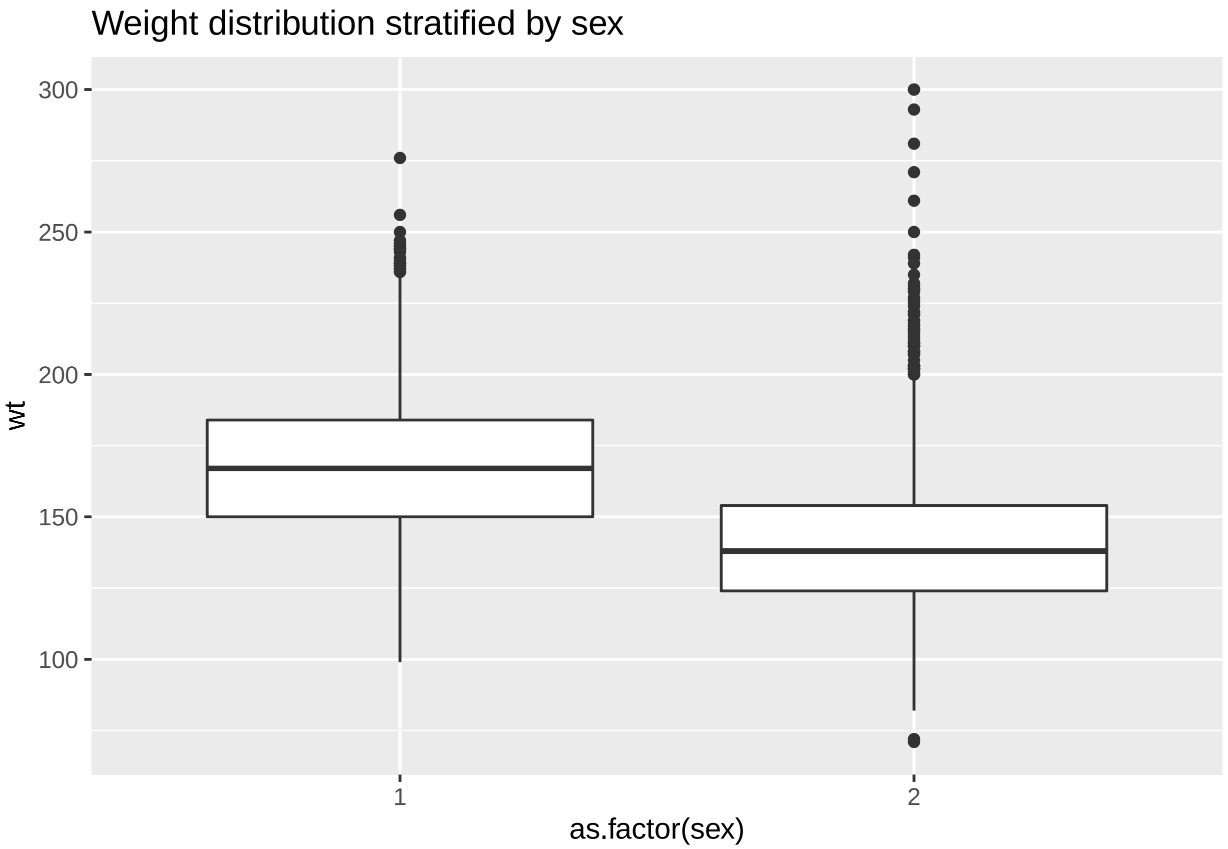
Diagnosis frequency per exam



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
```

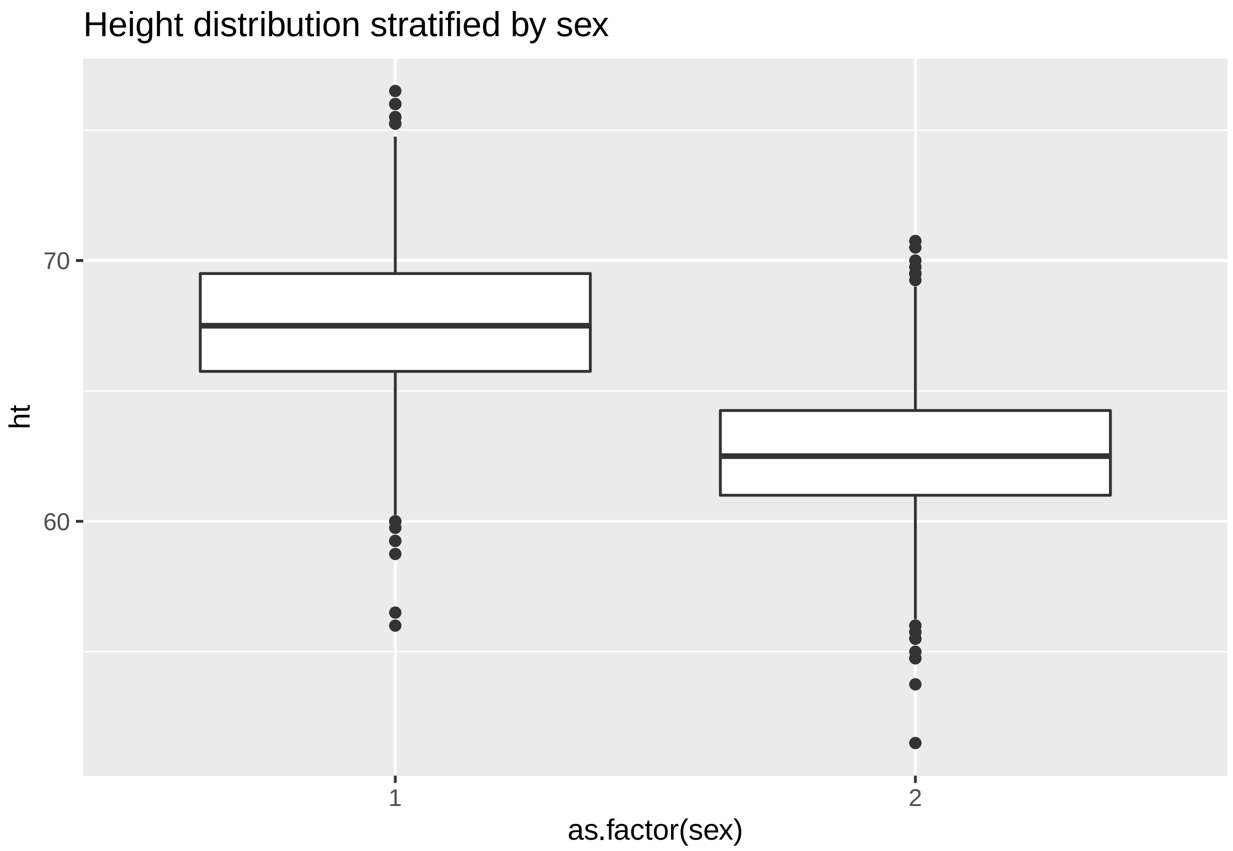
```
summary(data[data$sex==2,]$wt)
```

```
##    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
```

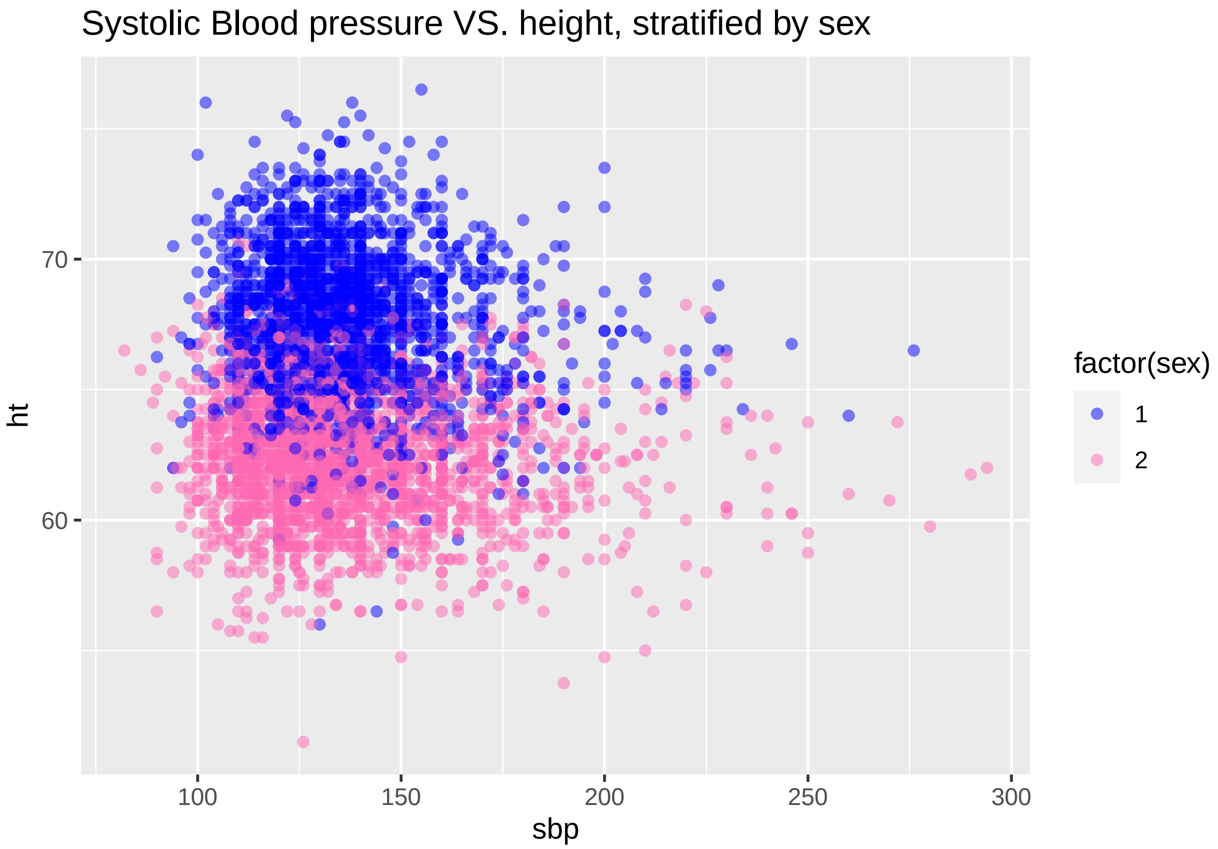
```
summary(data[data$sex==2,]$ht)
```

```
##    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.

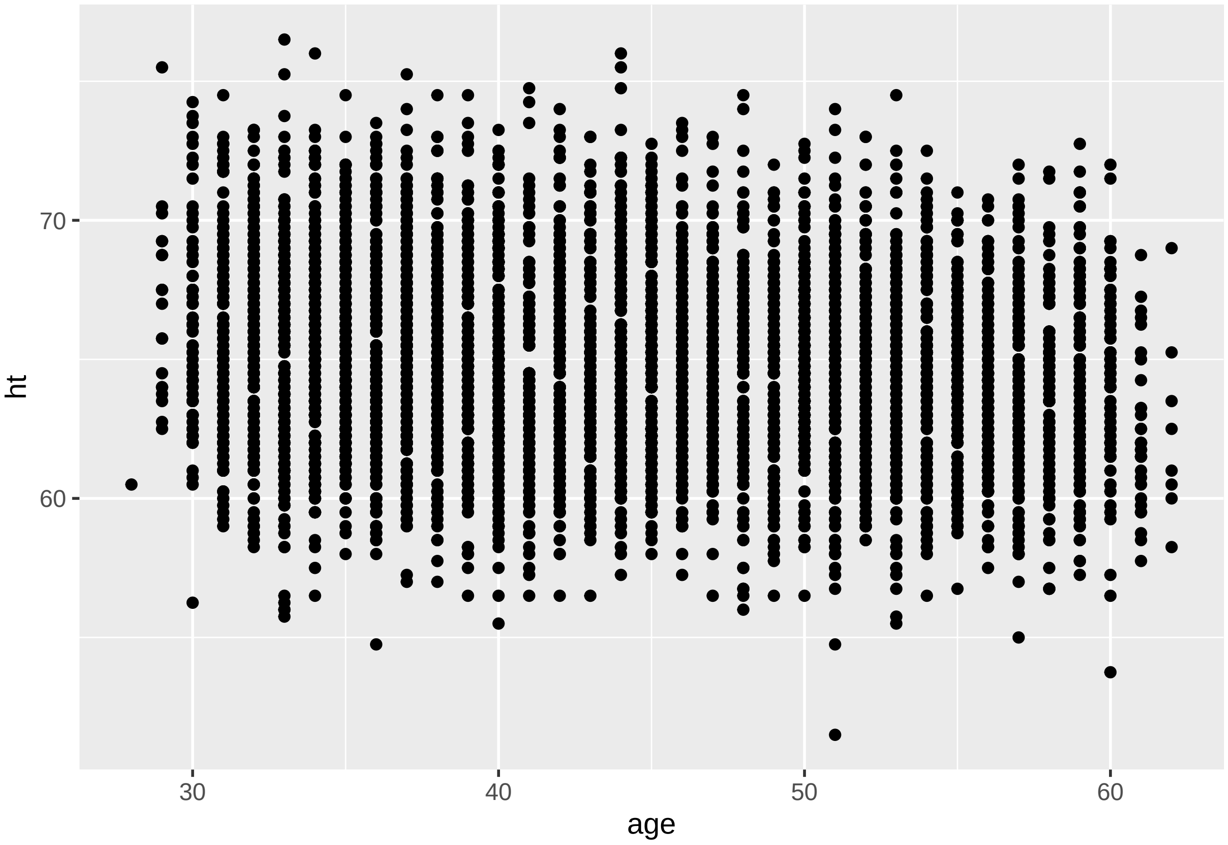
```
ggplot(data = data, aes(x = sbp, y = ht, color = factor(sex))) + geom_point(alpha=0.5) + scale_color_mar
ggtitle('Systolic Blood pressure VS. height, stratified by sex')
```

Systolic Blood pressure VS. height, stratified by sex

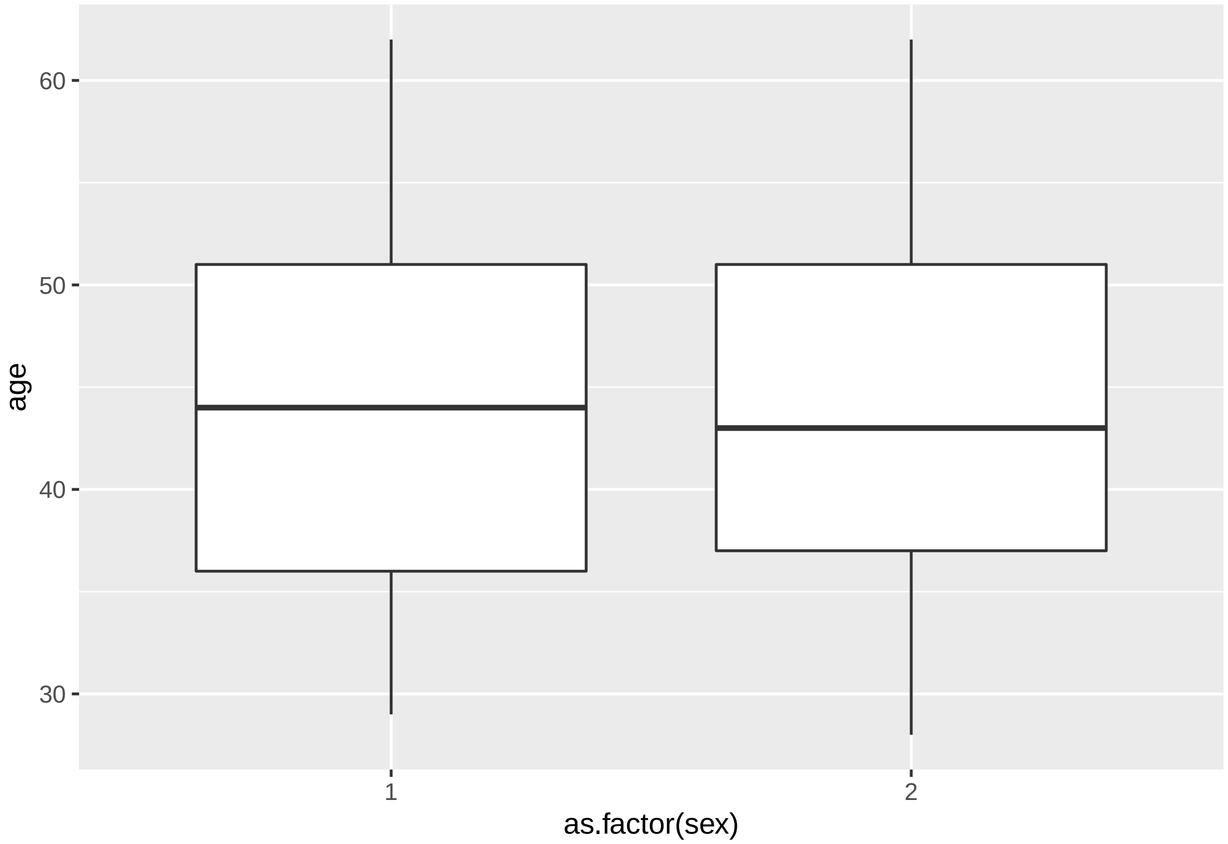


From the plot, we can see male and females have generally similar distribution in terms of systolic blood pressure. However, for subjects with `sbp` larger than 200, we see `sbq` increase with height in females while `sbq` decrease with height in males. Difference in height distribution among males and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

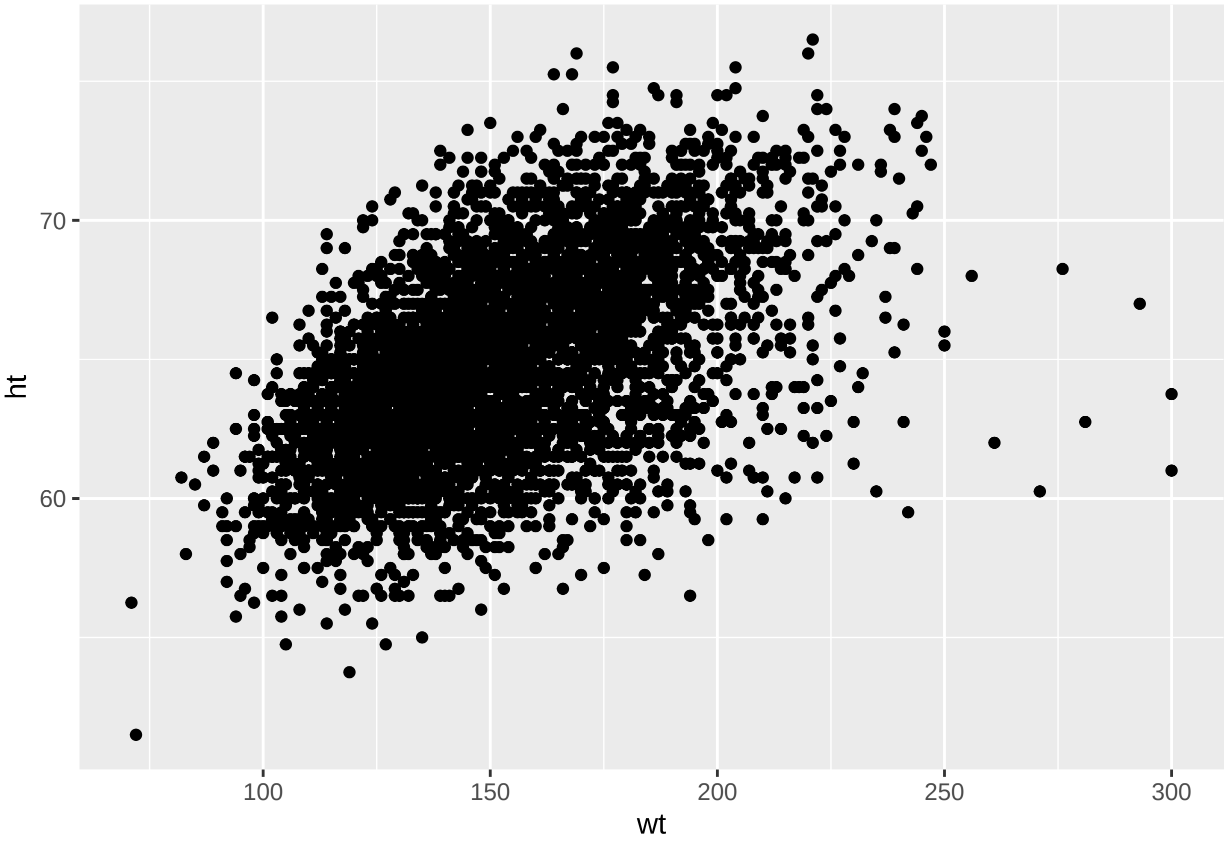
```
library(ggplot2)

ggplot(data, aes(x=age, y=ht)) + geom_point()
```



```
ggplot(data, aes(x=as.factor(sex), y=age)) + geom_boxplot()
```



```
ggplot(data, aes(x=wt, y=ht)) + geom_point()
```


70
60
ht
100
150
200
250
300
wt