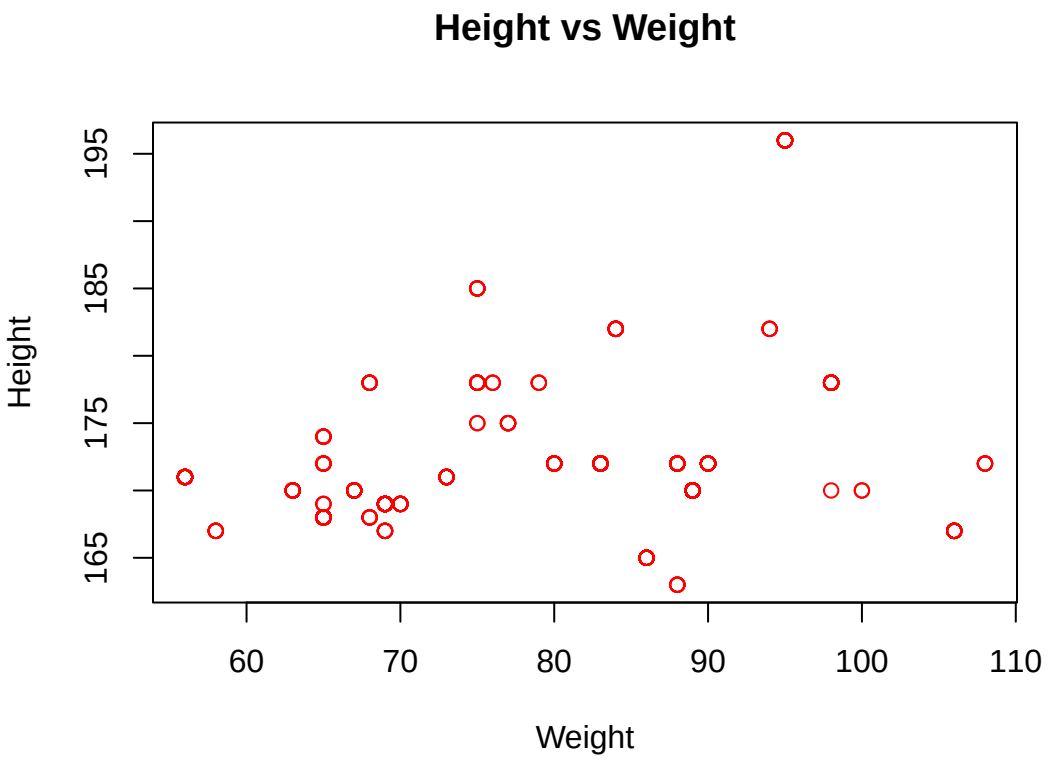
Lab - Absenteeism at Work

jungbin lee (ID : 1235143431)

2025-09-25

1. Scatter Plot: Height vs Weight

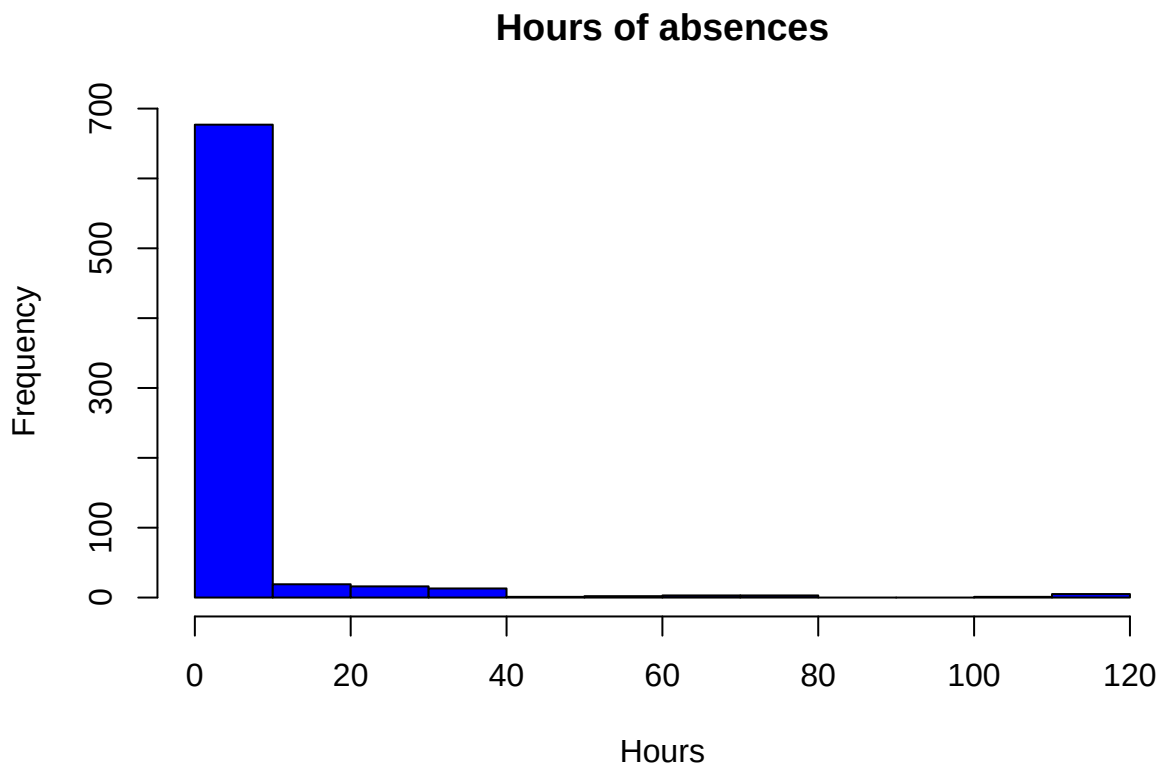
```
df = read.csv("Absenteeism_at_work.csv", sep=";", header=TRUE)
par(mfrow = c(1,1), mai = c(1,1,1,1))
#1. Scatter Plot: Height vs Weight

#remove missing data
df_1 <- df[(!is.na(df$Height) & !is.na(df$Weight)), ]

# make a plot, x is weight and y is height
plot(df_1$Weight, df_1$Height, main = "Height vs Weight", xlab = "Weight", ylab = "Height", col = 'red'
```




```
# height and weight is not relate to Absent in work
```

2. Histogram: Hours of absences

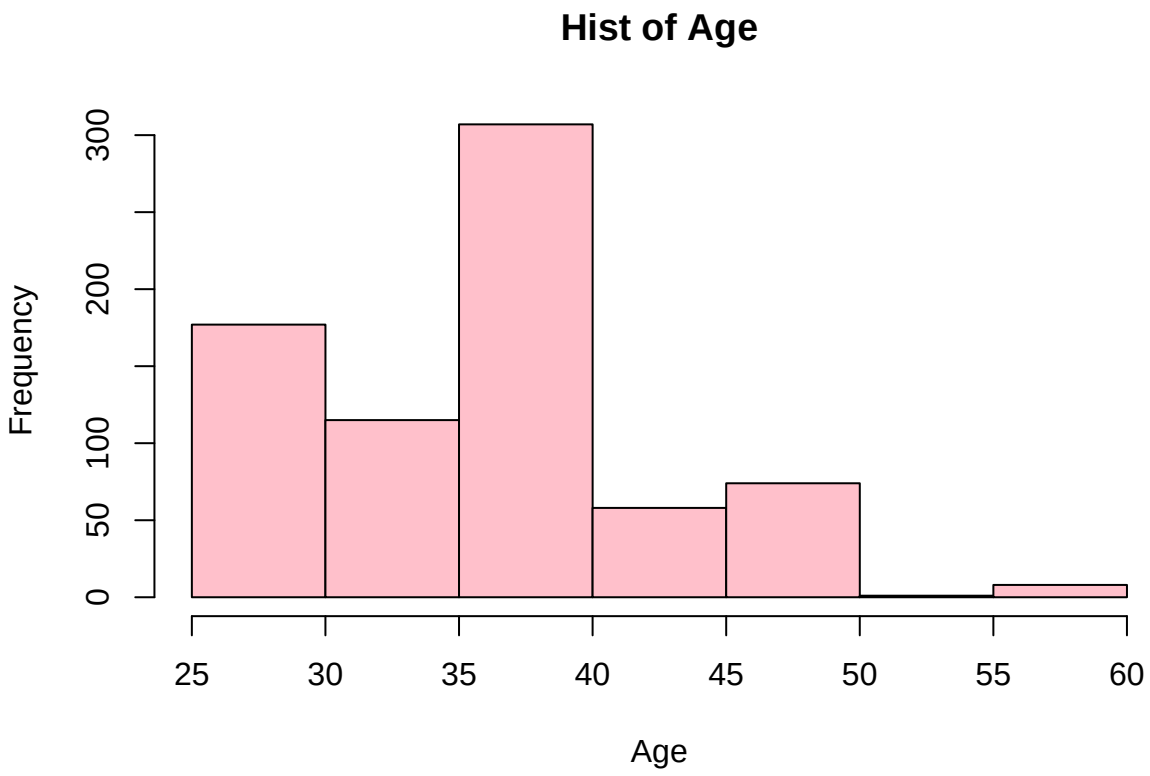
```
#2 Hist plot, hours of absences

#remove missing data, select 'Absenteeism.time.in.hours' column
df_2 <- df[!is.na(df$Absenteeism.time.in.hours), 'Absenteeism.time.in.hours']

# make a histogram, x is absent time in hours and y is frequency
hist(df_2, main = "Hours of absences", xlab = "Hours", ylab = "Frequency", col = 'blue')
```




```
# a lot of poeple absence less than 20 hours.
```

3. Histogram: Age

```
#3 Hist plot, Age

#remove missing data select 'specific 'Age' column
df_3 <- df[!is.na(df$Age), 'Age']

#make a histogram, x is age and y is frequency
hist(df_3, main = 'Hist of Age', xlab = 'Age', ylab = 'Frequency', col = 'pink')
```




```
# in this company, lot of people have age between 35 to 40 people
```

4. Bar plot: Hours by Month

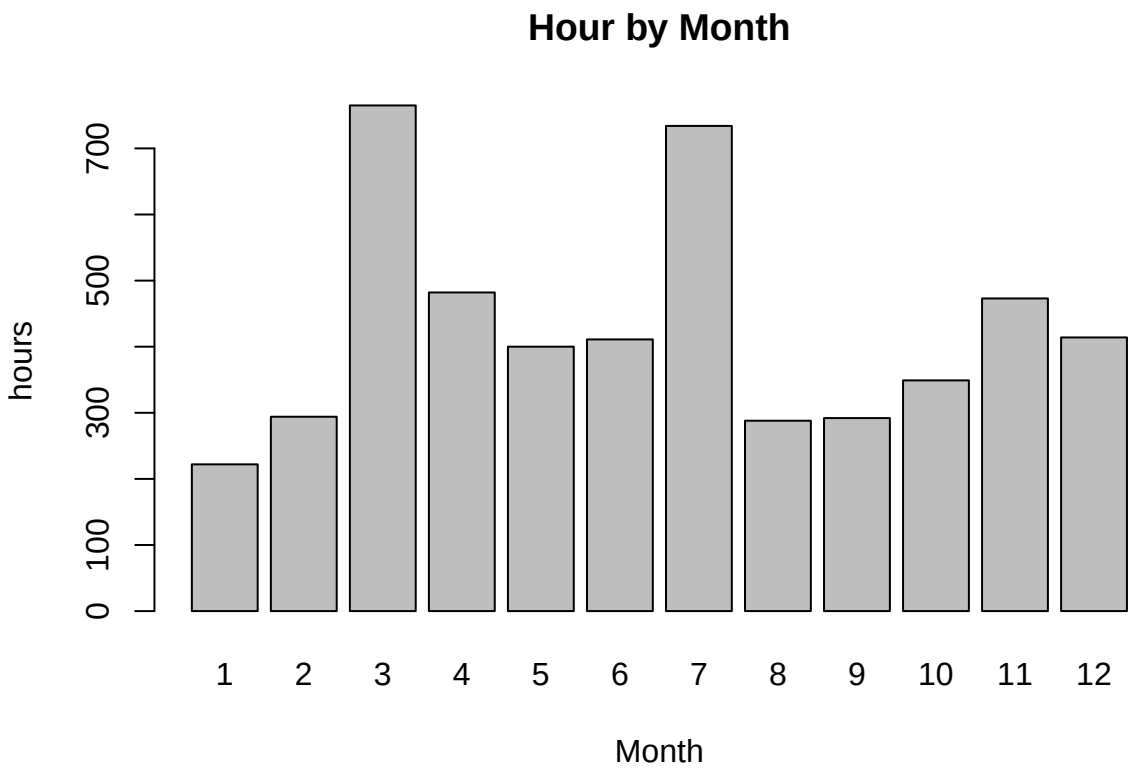
```
#4 Bar plot, Hour by Month

# find outlier
unique(df$Month.of.absence)
```

```
##  [1]  7  8  9 10 11 12  1  2  3  4  5  6  0
```

```
# month can not be 0, remove
df_4 <- df[df$Month.of.absence != 0, ]

# check if I remove month 0
unique(df_4$Month.of.absence)
```

```
##  [1]  7  8  9 10 11 12  1  2  3  4  5  6
```

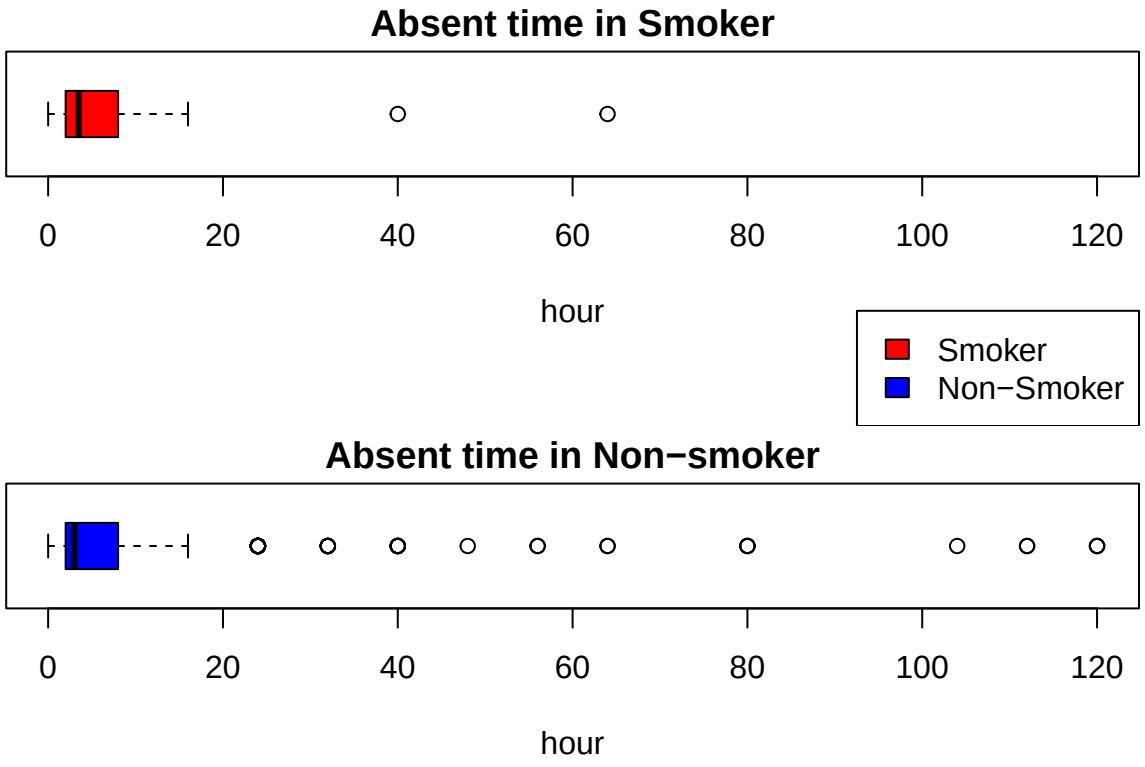
```
# using tapply, get a whole absence hour each month
df_4 <- tapply(df_4$Absenteeism.time.in.hours, df_4$Month.of.absence, sum)

# make barplot
barplot(df_4, main = "Hour by Month", xlab = "Month",ylab = "hours", col = 'gray')
```



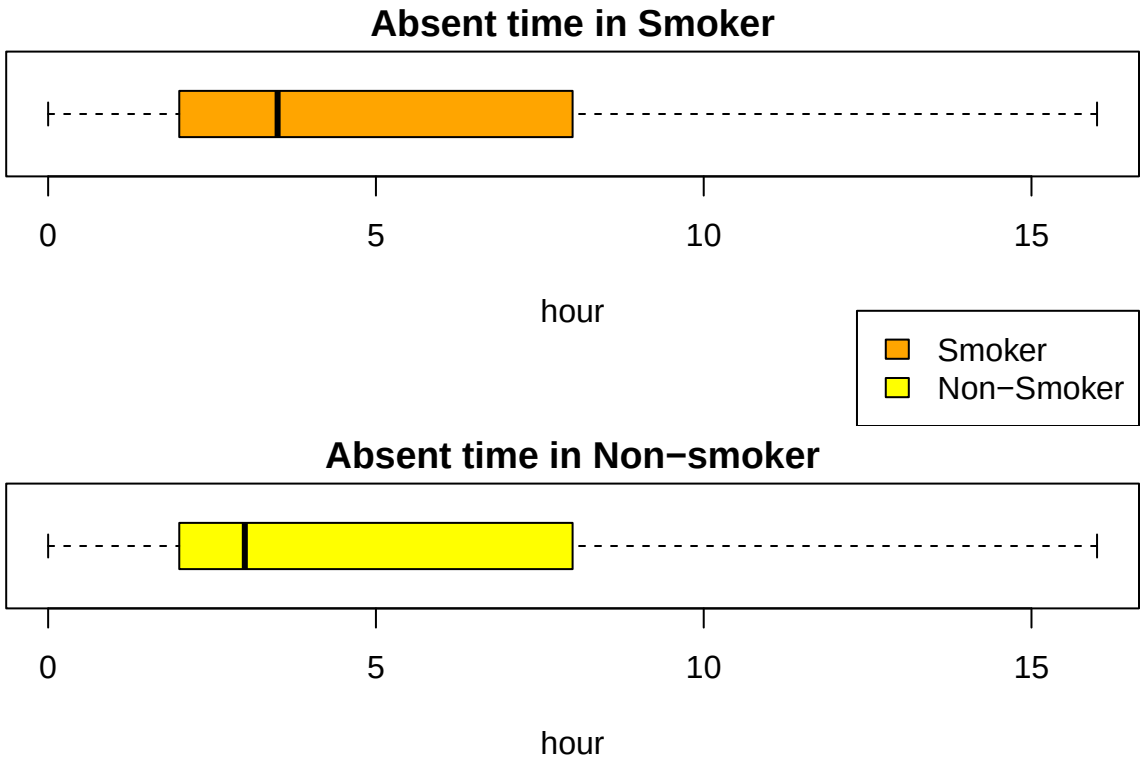

```
# on march, lot of people absence, and on January, a few people absence
```

5. Box plot: Social smoker

```
#5 Box plot, Social smoker

# remove missing data
df_5 <- df[!is.na(df$Social.smoker) & !is.na(df$Absenteeism.time.in.hours), ]

# dataframe who smoke
df_smoke <- df_5[df$Social.smoker == 1 ,]

#dataframe who are not smoke
df_no_smoke <- df_5[df$Social.smoker == 0 ,]

# make a space for two boxplot
par(mfrow = c(2,1),mai = c(1.3,.1,.3 ,.5))

# put legend outside of boxplot
par(xpd=TRUE)
# make two boxplot with legend,
boxplot(df_smoke$Absenteeism.time.in.hours, main = "Absent time in Smoker", ylim = c(0,120), horizontal
legend("bottomright", inset = c(0, -2),legend = c("Smoker", "Non-Smoker"), fill = c('red', 'blue'))
boxplot(df_no_smoke$Absenteeism.time.in.hours, main = "Absent time in Non-smoker", horizontal = T, col =
```



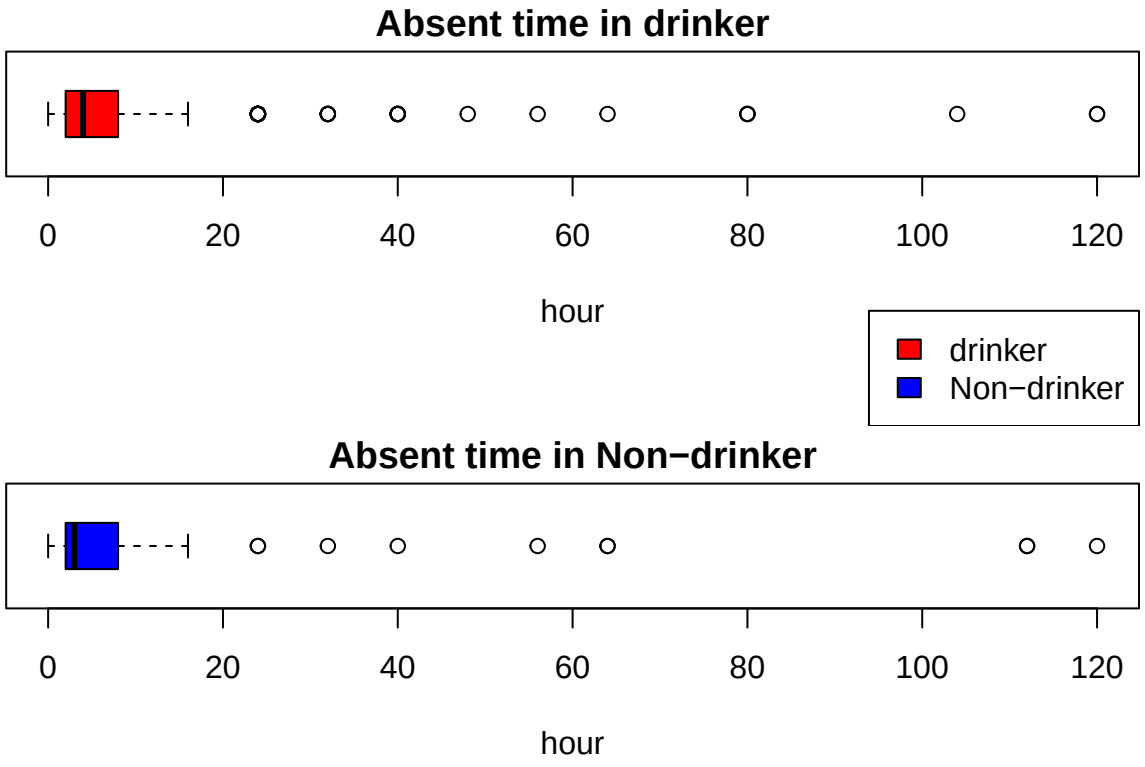

```
# since there is people who absence too much, we can not compare well both plot

# make a two boxplot with legend, remove outlier
boxplot(df_smoke$Absenteeism.time.in.hours, main = "Absent time in Smoker", horizontal = T, col = "orang
legend("bottomright", inset = c(0, -2),legend = c("Smoker", "Non-Smoker"), fill = c('orange', 'yellow'))
boxplot(df_no_smoke$Absenteeism.time.in.hours, main = "Absent time in Non-smoker", horizontal = T, col =
```



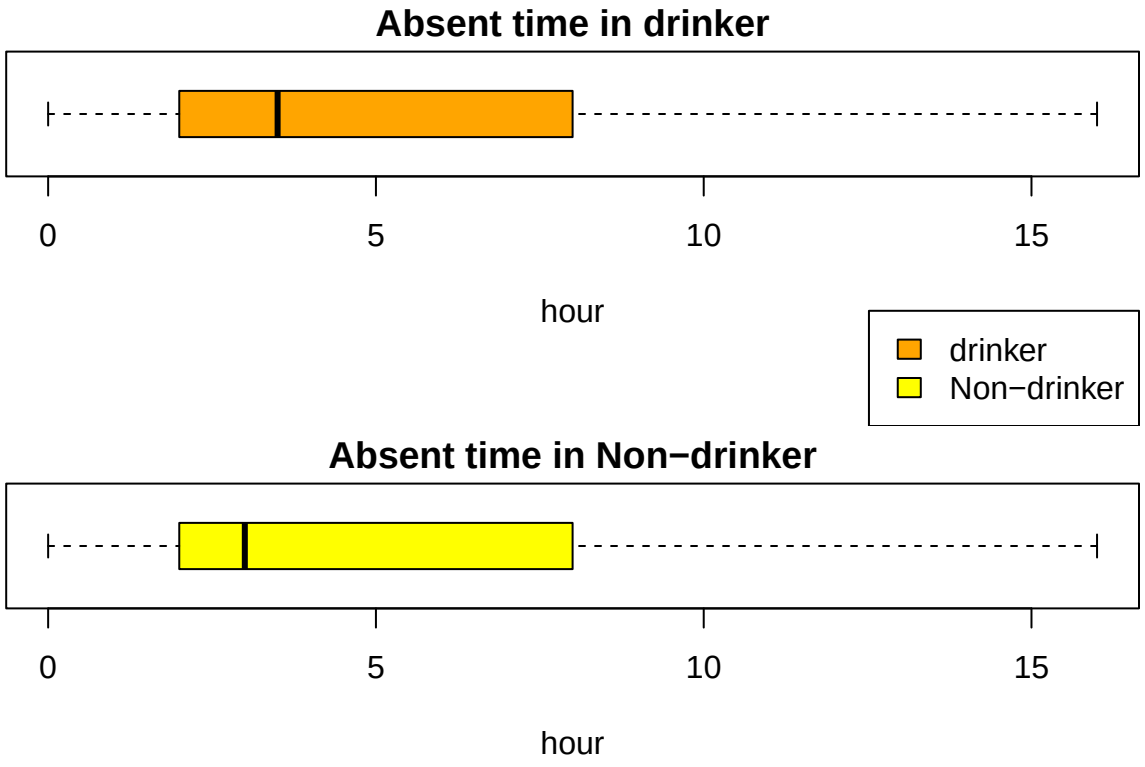

```
# As you can see in the boxplot the medium of smoke people absence hour is much higher than people who
```

6. Box plot: Social drinker

```
#6 Box plot, Social drinker

# remove missing data
df_6 <- df[!is.na(df$Social.drinker) & !is.na(df$Absenteeism.time.in.hours), ]

# dataframe who drinker
df_drinker <- df_5[df$Social.drinker == 1 ,]

#dataframe who are not drinker
df_no_drinker <- df_5[df$Social.drinker == 0 ,]

# make a space for two boxplot
par(mfrow = c(2,1),mai = c(1.3,.1,.3 ,.5))

# put legend outside of boxplot
par(xpd=TRUE)
#make a boxplot with legend
boxplot(df_drinker$Absenteeism.time.in.hours, main = "Absent time in drinker", ylim = c(0,120), horizon
legend("bottomright", inset = c(0,-2),legend = c("drinker", "Non-drinker"), fill = c('red', 'blue'))

boxplot(df_no_drinker$Absenteeism.time.in.hours, main = "Absent time in Non-drinker", horizontal = T, c
```




```
# since there is people who absence too much, we can not compare well both plot

#make a two boxplot with legend, remove outliner
boxplot(df_smoke$Absenteeism.time.in.hours, main = "Absent time in drinker", horizontal = T, col = "ora
legend("bottomright", inset = c(0, -2),legend = c("drinker", "Non-drinker"), fill = c('orange','yellow'
boxplot(df_no_smoke$Absenteeism.time.in.hours, main = "Absent time in Non-drinker", horizontal = T, col
```




```
# As you can see in the boxplot the medium of drink people absence hour is much higher than people who
```