DAT 301 Lab 1

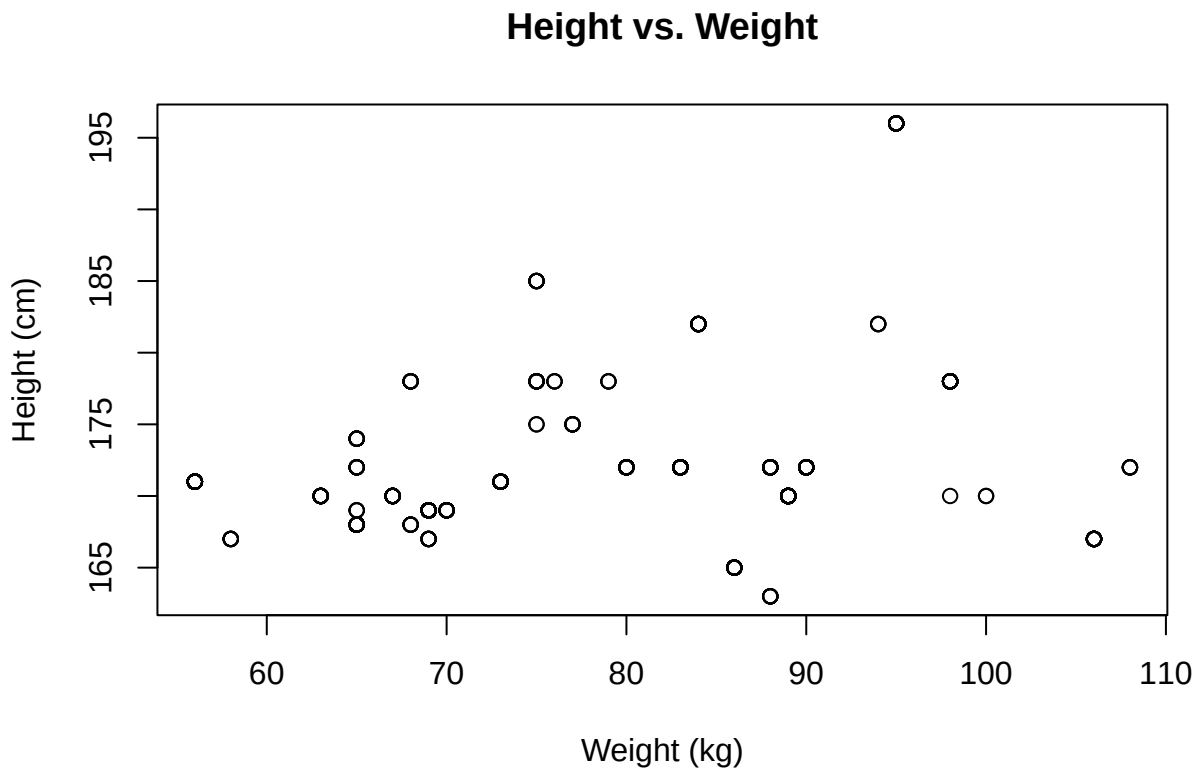
Jason Kong

2024-03-23

```
df = read.csv("Absenteeism_at_work.csv",sep=";",header=TRUE)
```

1. Plot the scatter plot of height vs. weight (so, weight on x-axis) including all the (non-missing) data.

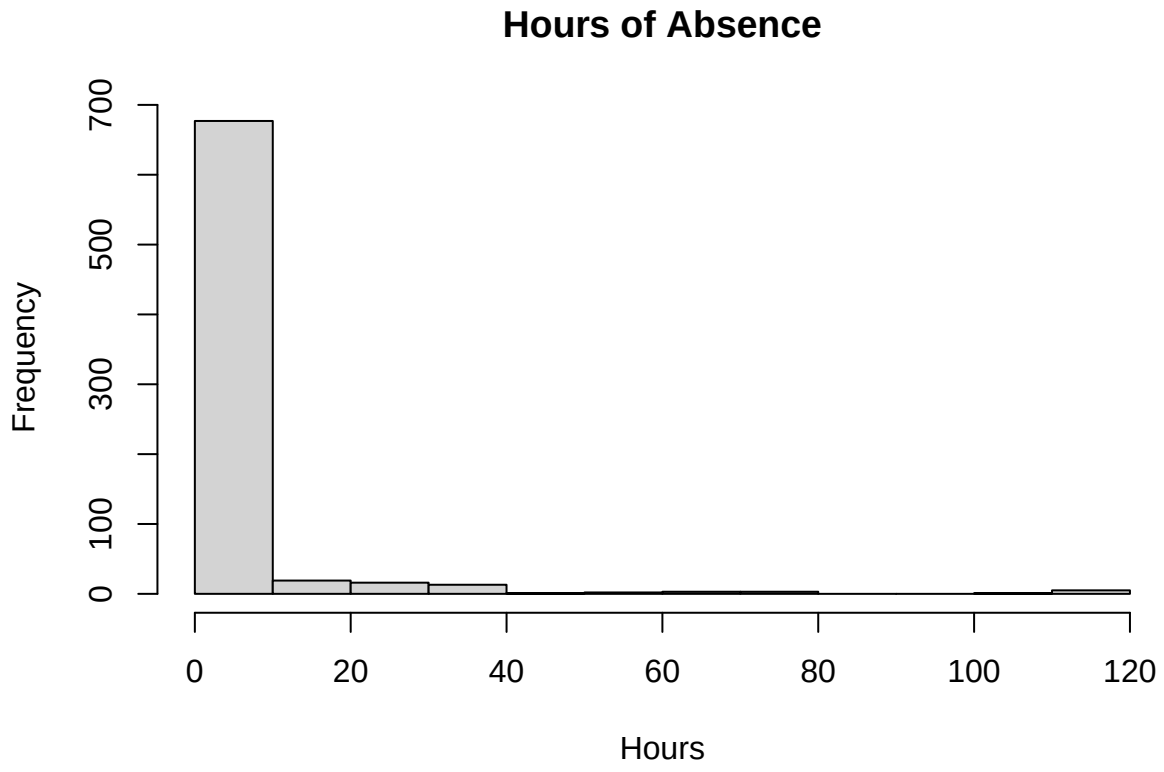
```
plot(df$Weight, df$Height,
     xlab="Weight (kg)",
     ylab="Height (cm)",
     main="Height vs. Weight")
```



The graph suggests that most absentees have a body height between 165 and 175cm.

2. Plot the histogram of hours of absences. Do not group by ID, just treat each absence as one observation.

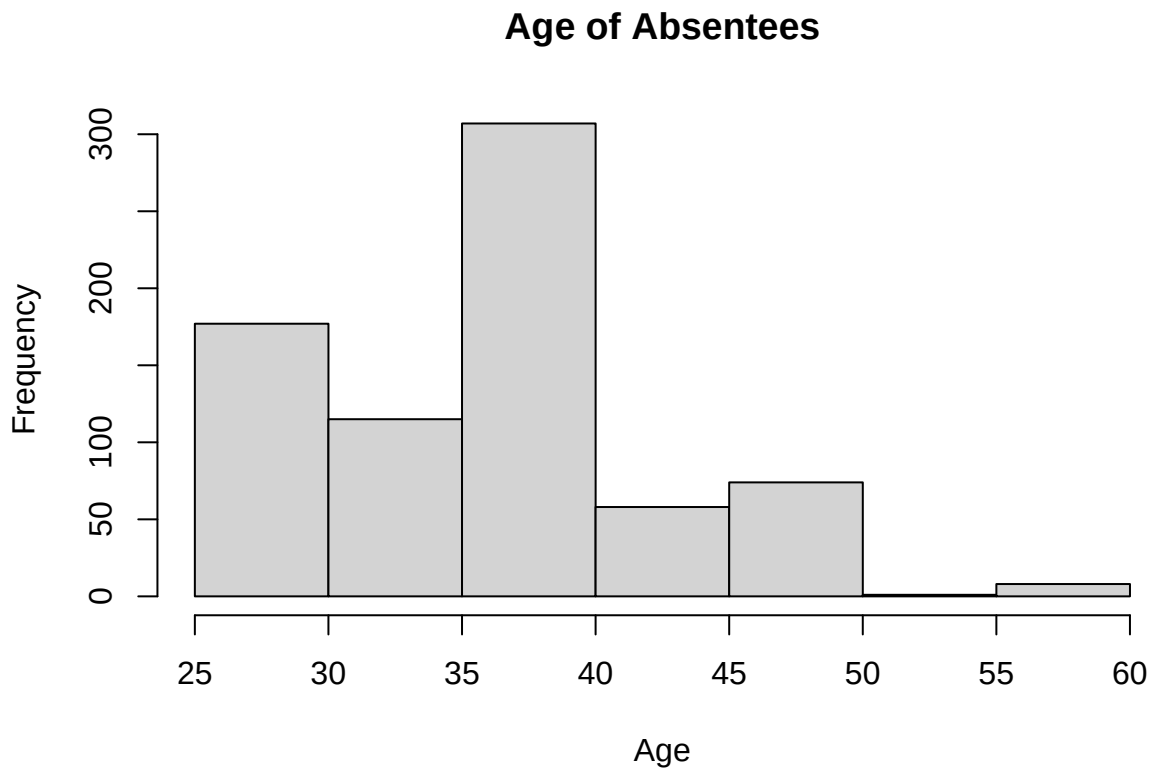
```
hist(df$Absenteeism.time.in.hours,
     xlab="Hours",
     main="Hours of Absence")
```



This graph suggests that most absentees have between 0-10 hours of absence.

3. Plot the histogram of age of a person corresponding to each absence. Do not group by ID, just treat each absence as one observation.

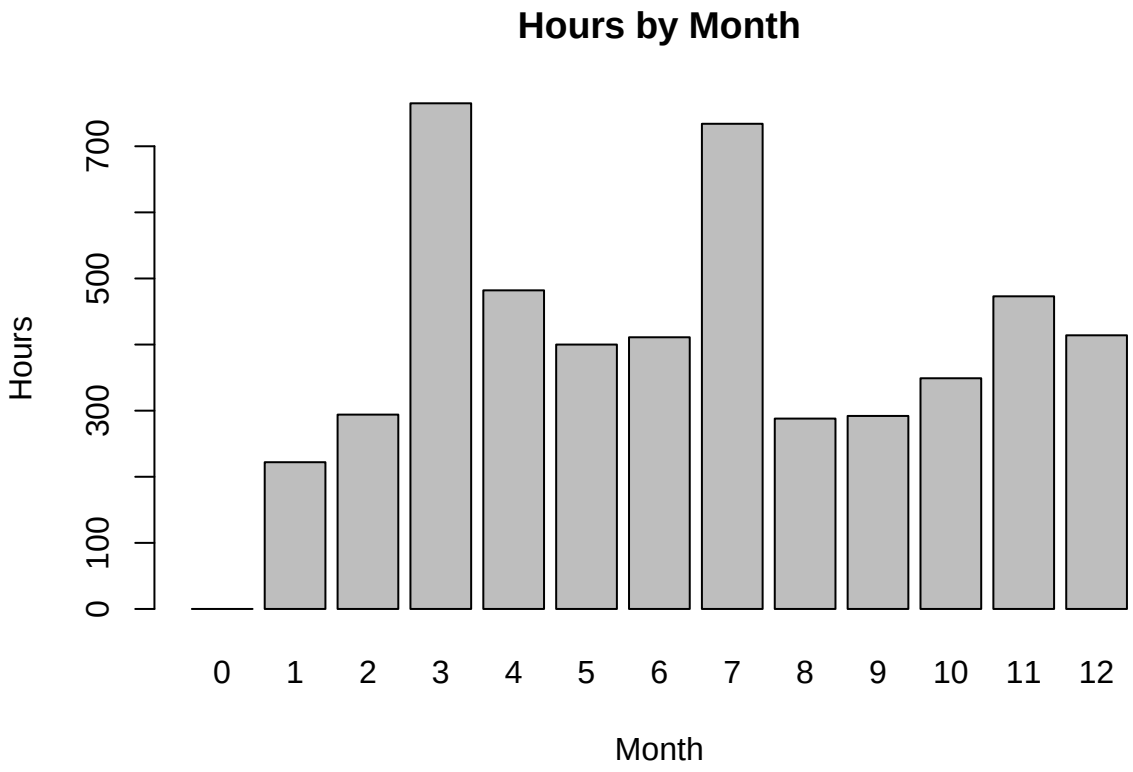
```
hist(df$Age,
     xlab="Age",
     main="Age of Absentees")
```



This graph indicates that a significant amount of absentees are between the age of 35 and 40.

4. Plot the bar plot of hours by month. So, each month is represented by one bar, whose height is the total number of absent hours of that month. (Hint: you can use tapply().)

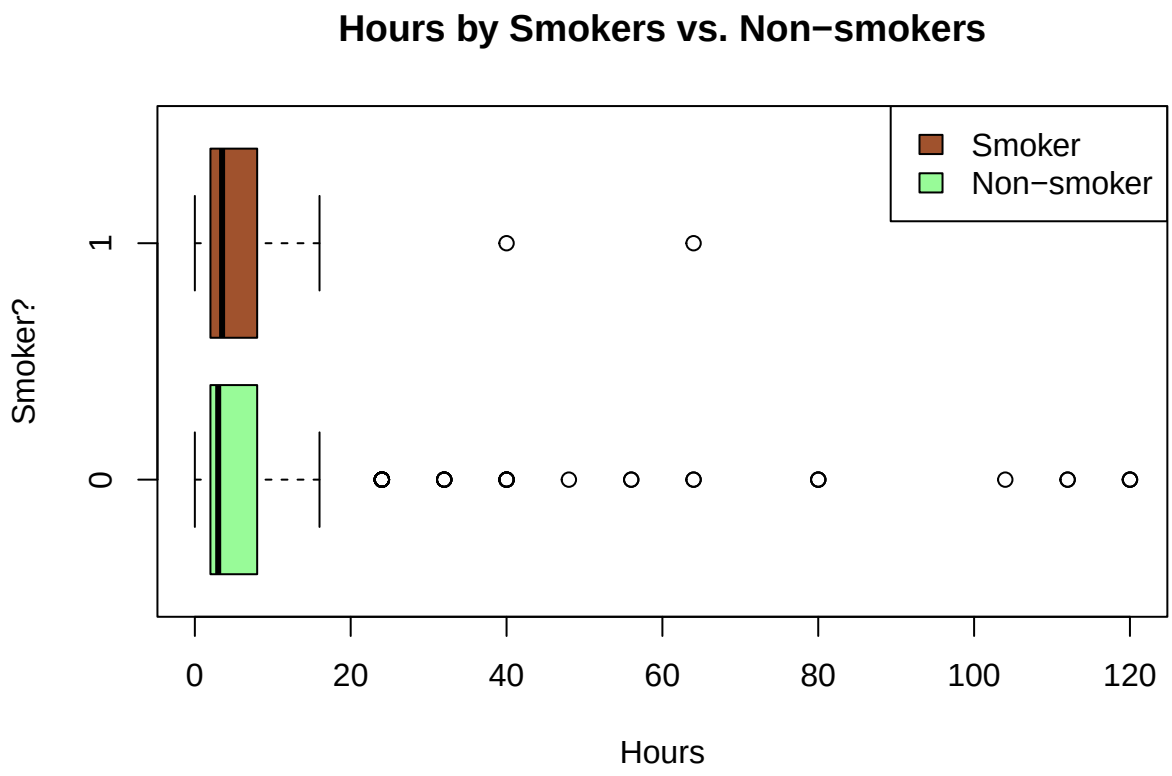
```
hoursByMonth = tapply(X=df$Absenteeism.time.in.hours, INDEX=df$Month.of.absence, FUN=sum)
barplot(hoursByMonth,
        xlab="Month",
        ylab="Hours",
        main="Hours by Month")
```



This graph suggests that the months of March and July see the most hours of absences.

5. Plot the box plots of hours by social smoker variable. So, you will have two box plots in one figure. Use the legend, labels, title. Play with colors.

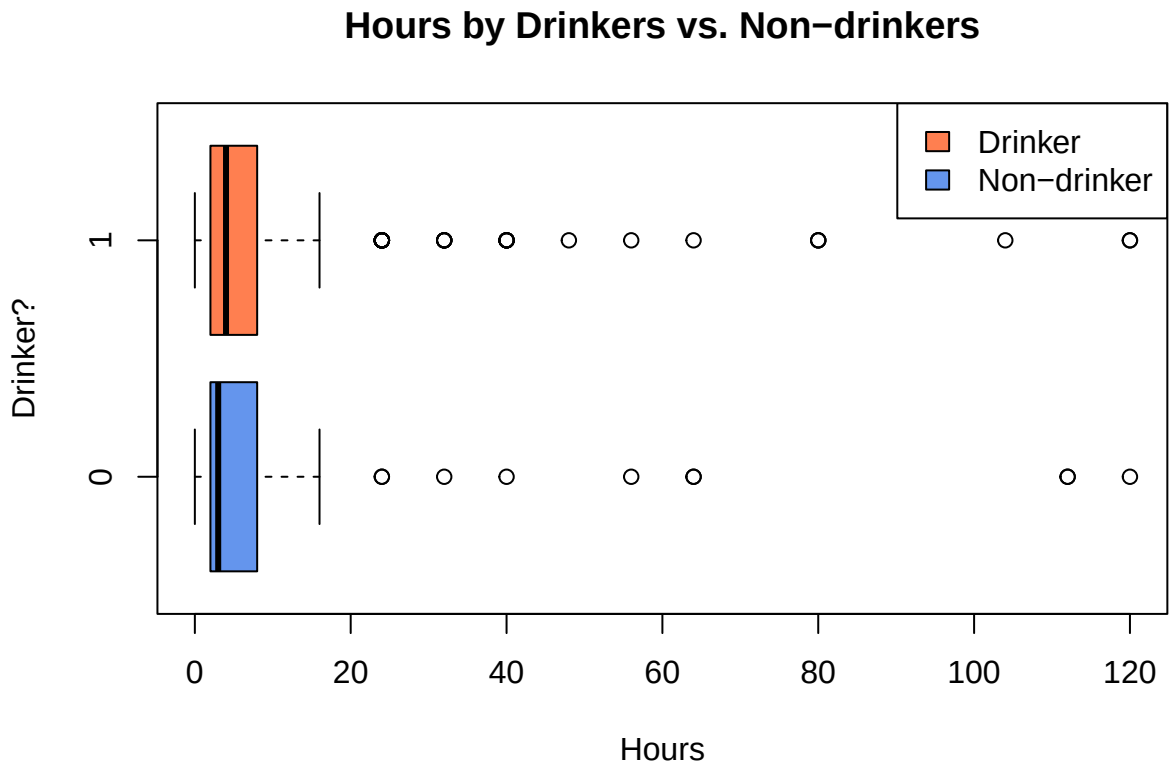
```
boxplot(df$Absenteeism.time.in.hours ~ df$Social.smoker,
        horizontal=TRUE,
        col=c("palegreen","sienna"),
        xlab="Hours",
        ylab="Smoker?",
        main="Hours by Smokers vs. Non-smokers")
legend("topright", legend=c("Smoker","Non-smoker"), fill=c("sienna","palegreen"))
```



This graph indicates that non-smokers have a lower median hours of absence.

6. Plot the box plots of hours by social drinker variable. So, you will have two box plots in one figure. Use the legend, labels, title. Play with colors.

```
boxplot(df$Absenteeism.time.in.hours ~ df$Social.drinker,
        horizontal=TRUE, col=c("cornflowerblue","coral"),
        xlab="Hours",
        ylab="Drinker?",
        main="Hours by Drinkers vs. Non-drinkers")
legend("topright", legend=c("Drinker","Non-drinker"), fill=c("coral","cornflowerblue"))
```



This graph indicates a higher median hours of absence among drinkers when compared to non-drinkers.