Lab 1: Absenteeism at Work

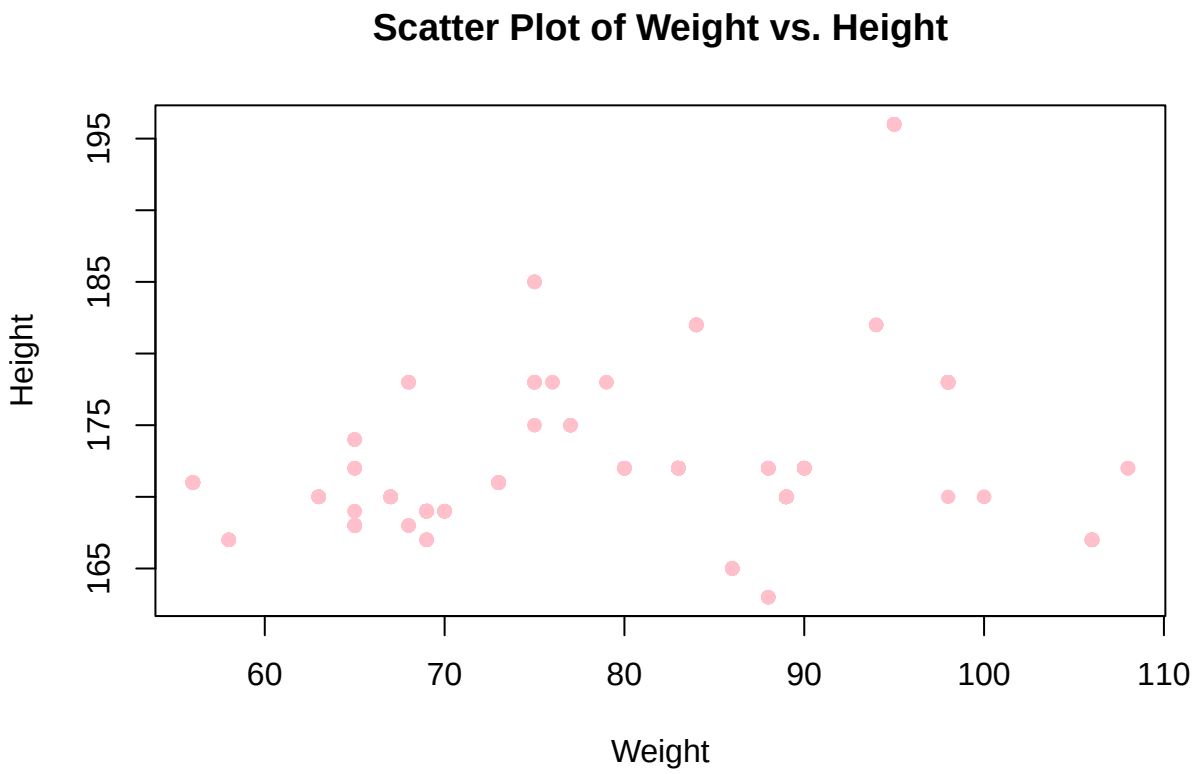
Kaytlyn Daffern

2024-09-27

Question 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,
     main="Scatter Plot of Weight vs. Height",
     xlab="Weight",
     ylab="Height",
     pch=16, col="pink")
```

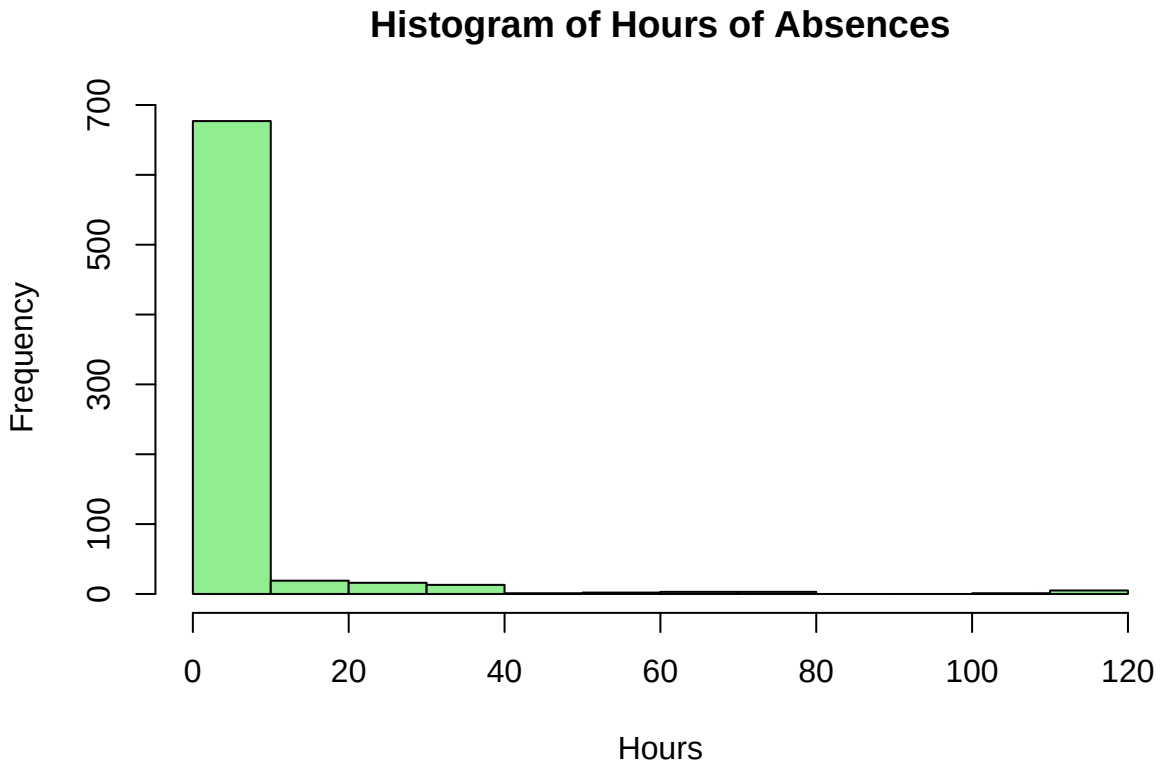


This plot shows a lot of variation and not a set correlation. The plot could suggest similar heights around 160.

Question 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,
     main="Histogram of Hours of Absences",
     xlab="Hours",
     col="lightgreen",
     border="black")
```

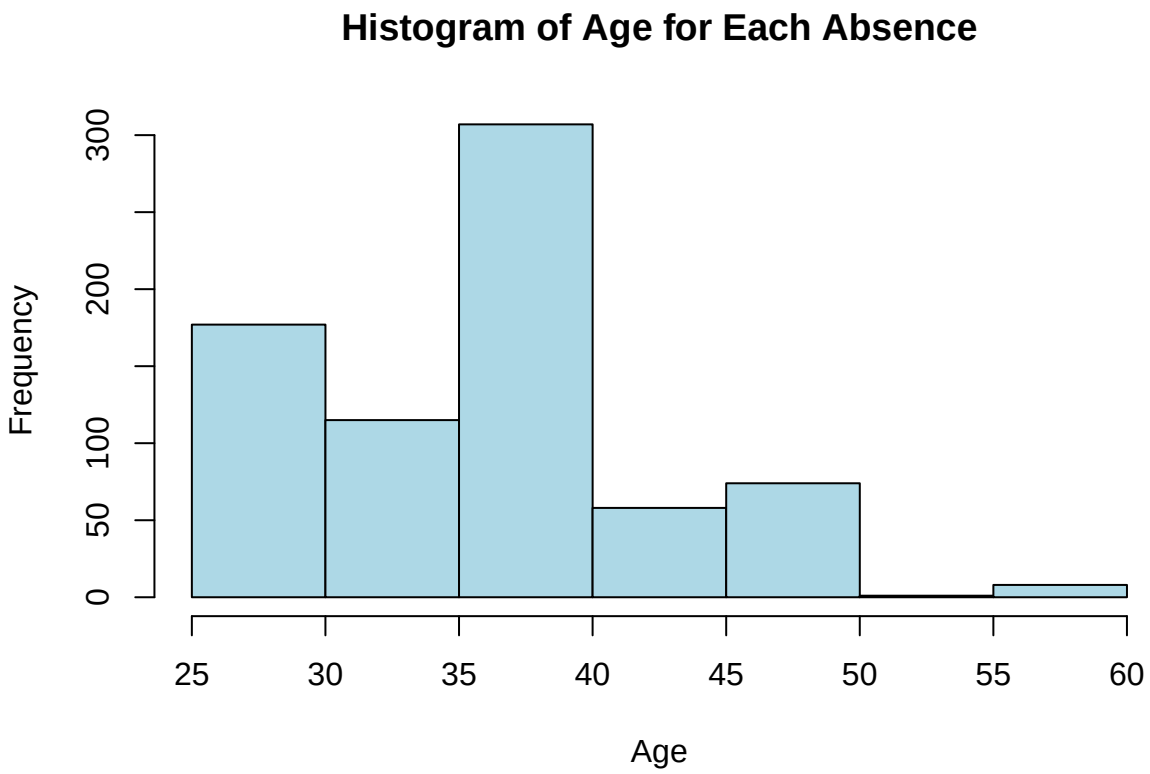


This histogram shows that the most absences are around the lower hours between 0-40 with a few random cases that are higher.

Question 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,
     main="Histogram of Age for Each Absence",
     xlab="Age",
     col="lightblue",
     border="black")
```

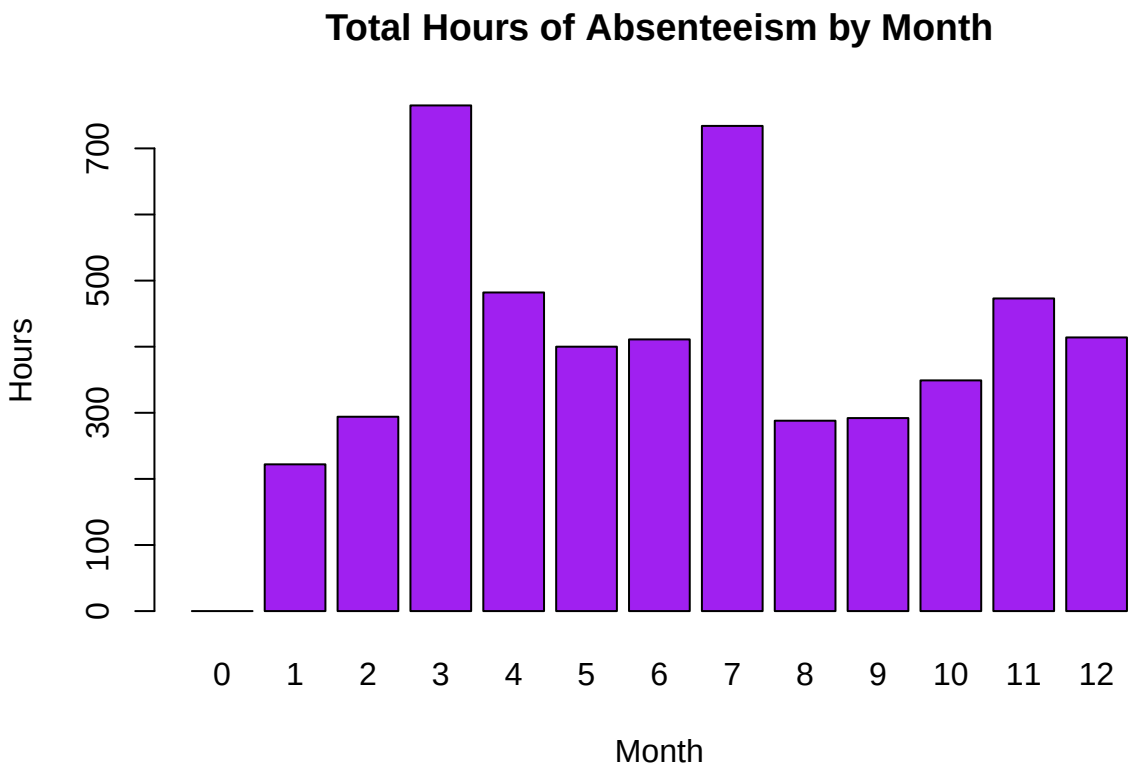


This histogram shows that most absences are 25-40 but there is still a good spread even up to the age 60. Between 35-40 the plot shows the highest.

Question 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.

```
barplot(tapply(df$Absenteeism.time.in.hours, df$Month.of.absence, sum),
        main="Total Hours of Absenteeism by Month",
        xlab="Month",
        ylab="Hours",
        col="purple",
        border="black")
```

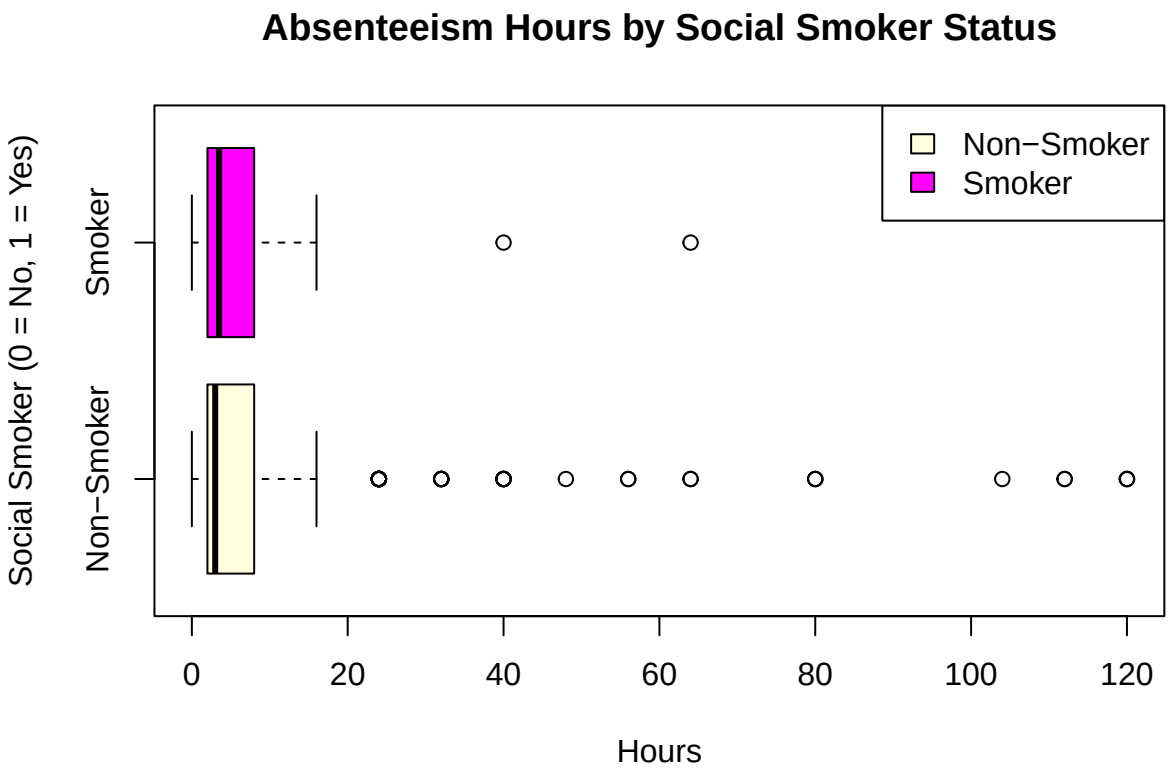


This bar plot shows that March(3) and July(7) have the most absences, but there is still an even spread between the other months.

Question 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,
        main="Absenteeism Hours by Social Smoker Status",
        horizontal = TRUE,
        xlab="Hours",
        ylab="Social Smoker (0 = No, 1 = Yes)",
        col=c("lightyellow", "magenta"),
        names=c("Non-Smoker", "Smoker")
        )
legend("topright", legend=c("Non-Smoker", "Smoker"), fill=c("lightyellow", "magenta"))
```

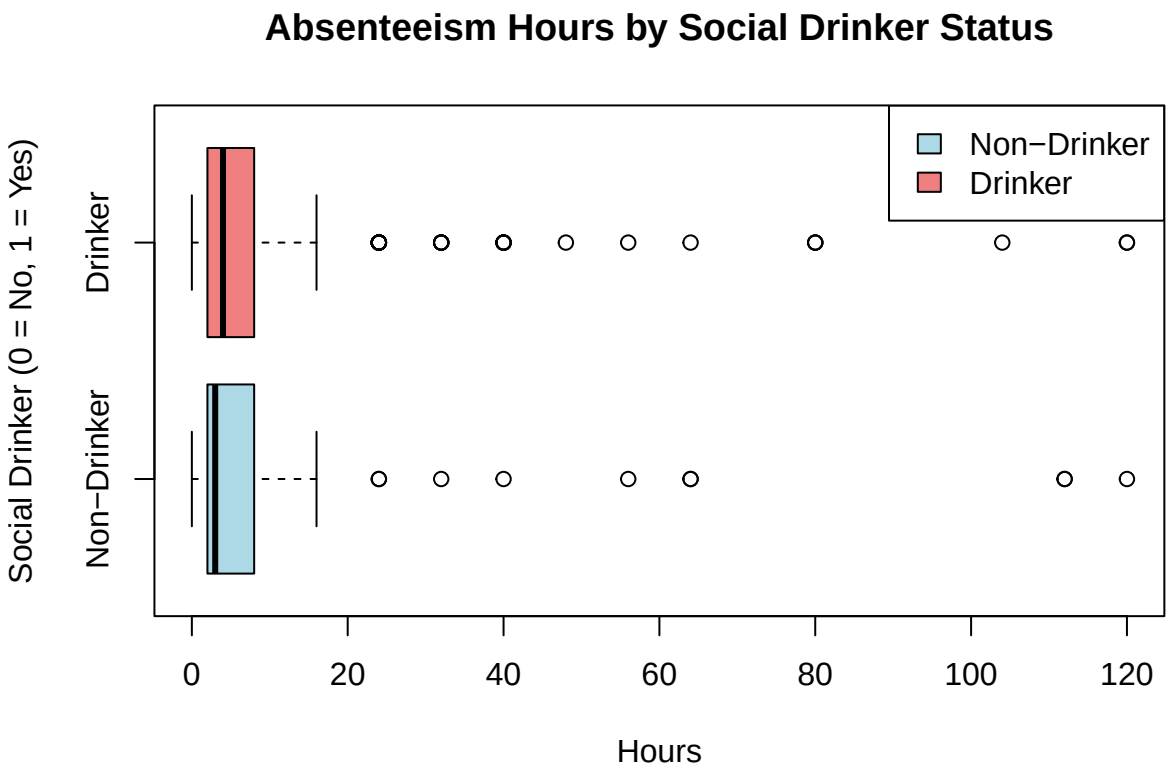


This box plot shows that social smokers have a higher median of absenteeism hours but non-smokers have more outliers.

Question 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,
        main="Absenteeism Hours by Social Drinker Status",
        horizontal = TRUE,
        xlab="Hours",
        ylab="Social Drinker (0 = No, 1 = Yes)",
        col=c("lightblue", "lightcoral"),
        names=c("Non-Drinker", "Drinker")
        )
legend("topright", legend=c("Non-Drinker", "Drinker"), fill=c("lightblue", "lightcoral"))
```



This box plot shows that social drinkers have a higher median of absenteeism hours but they share a similar amount of outliers.