2024 Spring B: DAT301

Lab 1

Juliana Perez Romero

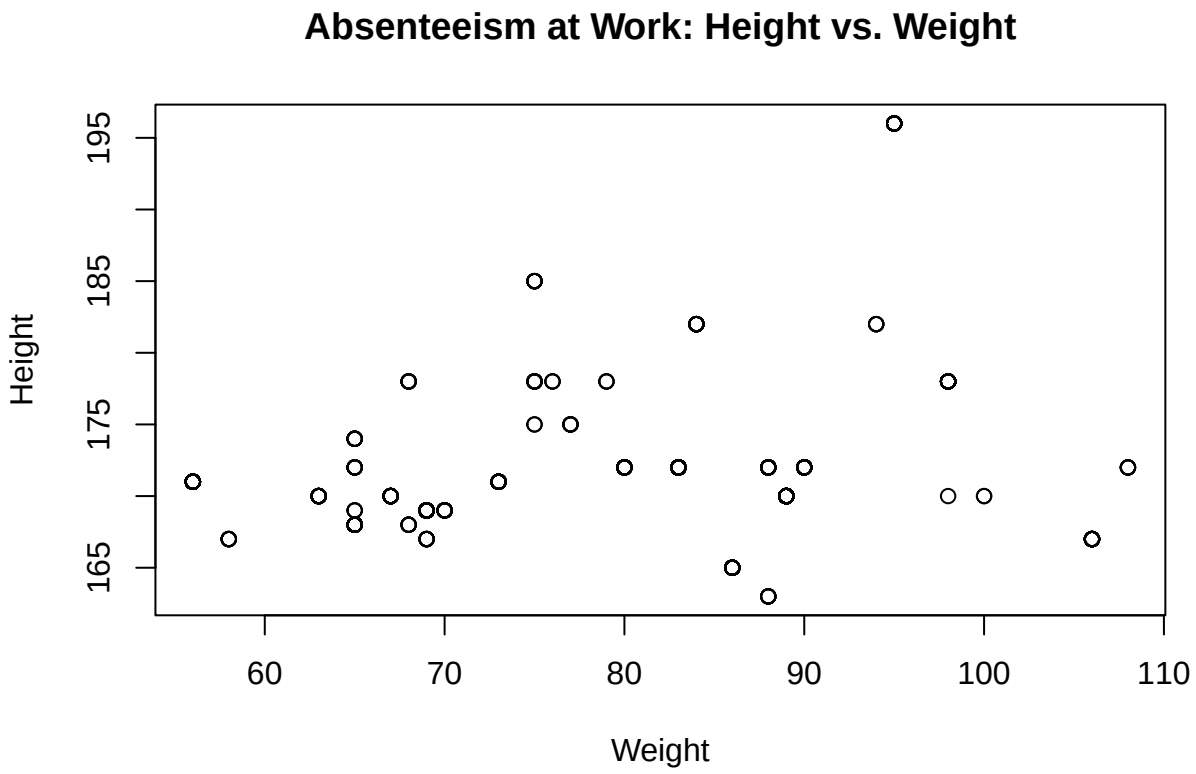
March 20th, 2024

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

```
Absenteeism_at_work <- read.csv("~/Absenteeism_at_work.csv", sep=";")

x <- Absenteeism_at_work$Weight
y <- Absenteeism_at_work$Height

plot(x,y,
     main = "Absenteeism at Work: Height vs. Weight",
     xlab = "Weight",
     ylab = "Height")
```

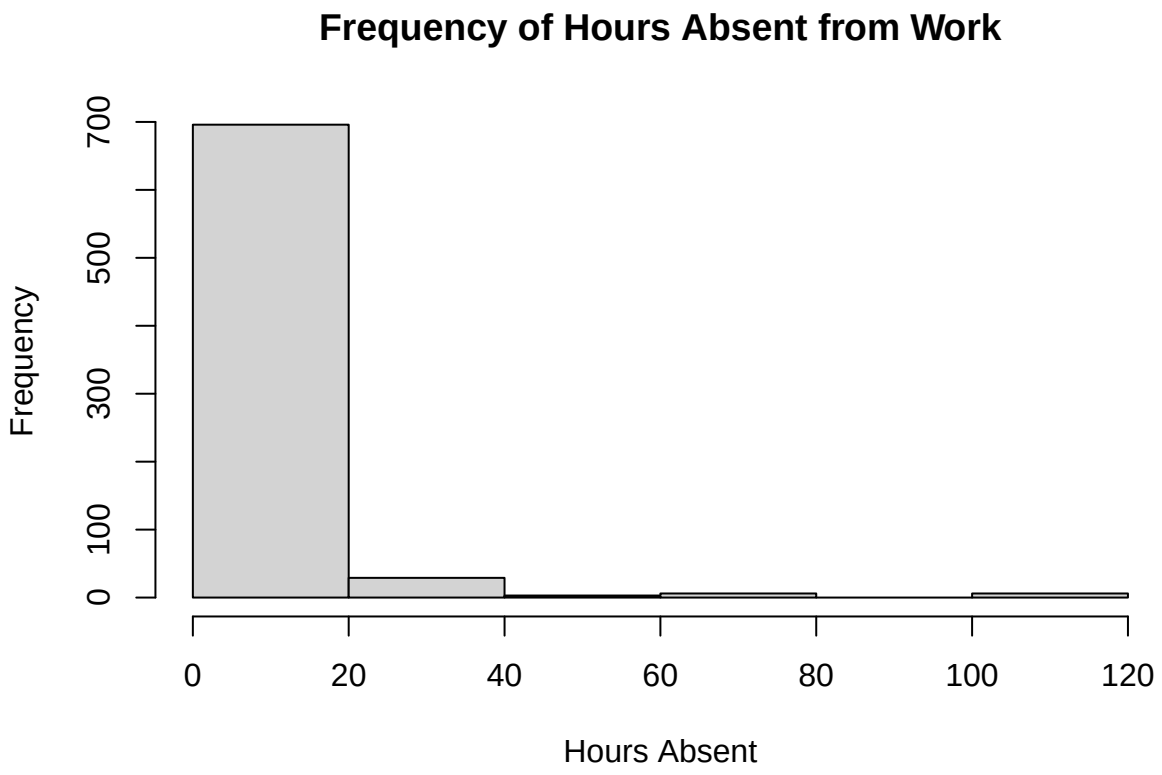


Comments: No correlation. There is no connection between the weight of the instances/observations and the corresponding heights.

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

```
hours_of_absences <- Absenteeism_at_work$Absenteeism.time.in.hours

hist(hours_of_absences,
    main="Frequency of Hours Absent from Work",
    xlab ="Hours Absent",
    ylab = "Frequency",
    breaks = 5)
```

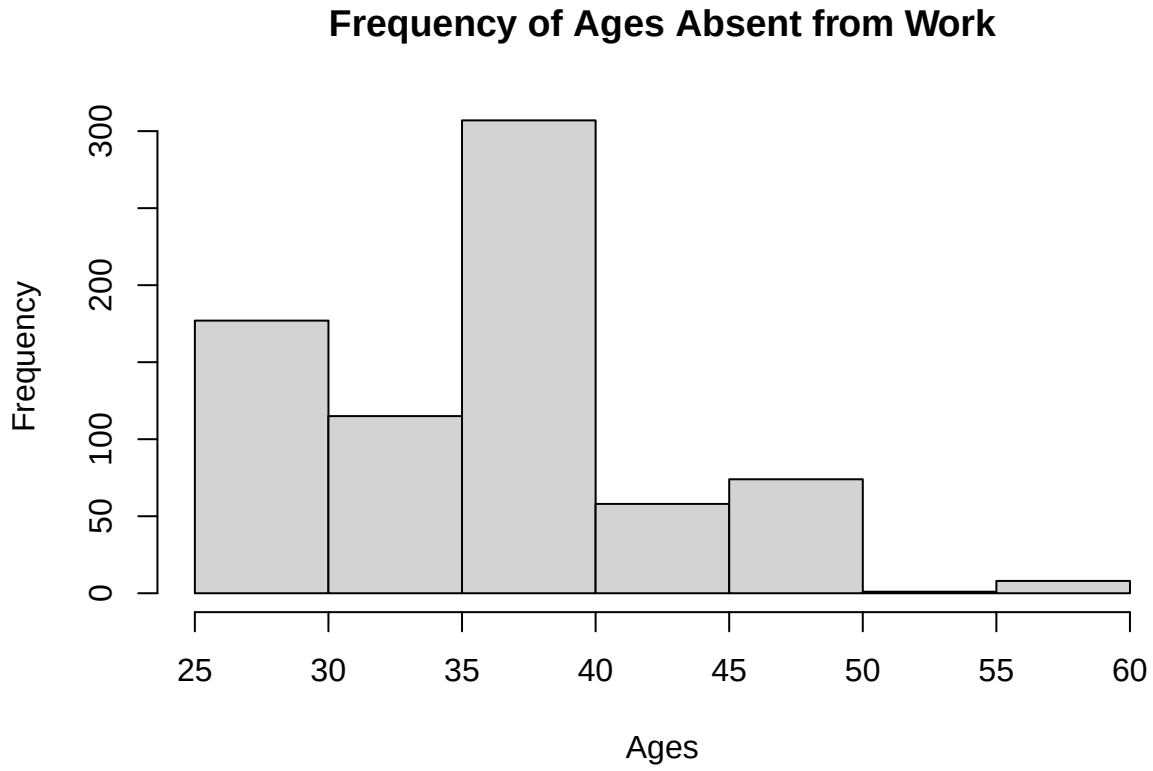


Comments: The histogram depicts positive skewness. A substantial portion of the data set, about 97% of the observations or 700 participants, were absent from work 0-20 hours. The remaining 3% or 20 participants dispersed from 21-120 hours.

(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.

```r
ages_of_absences <- Absenteeism_at_work$Age

hist(ages_of_absences,
    main="Frequency of Ages Absent from Work",
    xlab ="Ages",
    ylab = "Frequency")
```

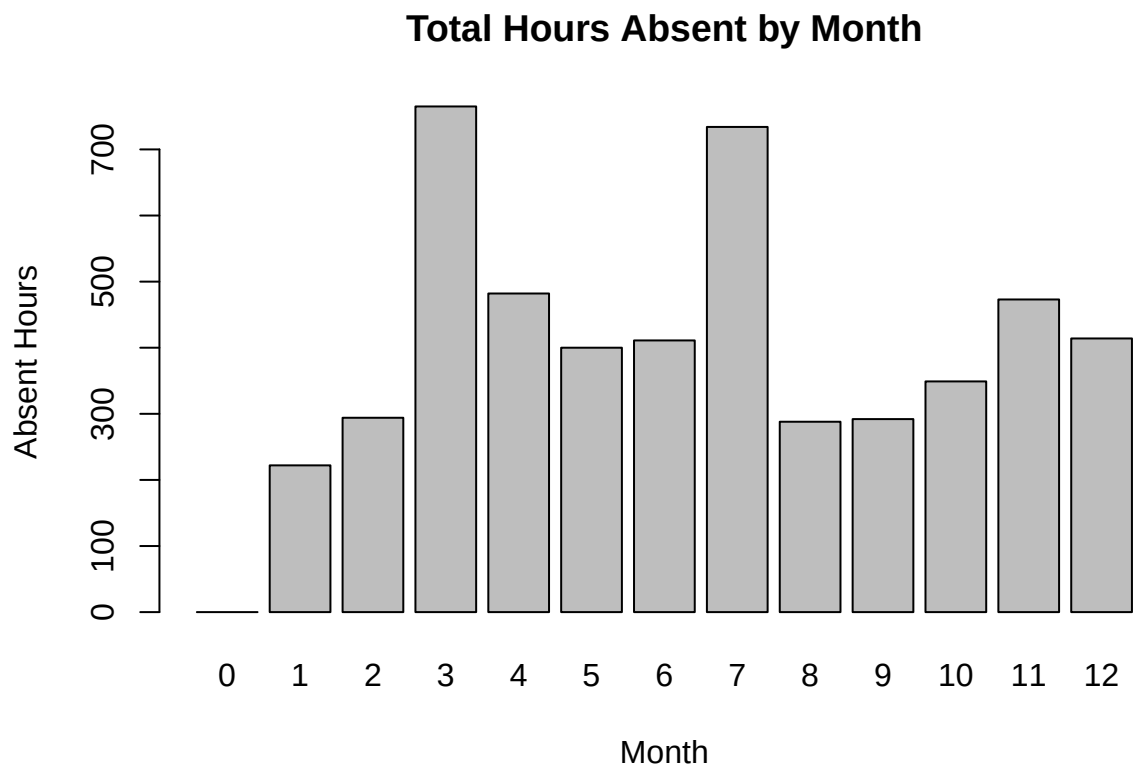


Comments: The histogram depicts a positive skew. Indicating a trend in which younger age groups tend to bt absent from work at a higer frequency than older age groups. However a notable exception is seen within the subset spanning ages 35 - 40 that accounts for about 300 participants, making it the age subset with the largest frequency.

(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().)

```
hours_by_month <- tapply(Absenteeism_at_work$Absenteeism.time.in.hours,
                         Absenteeism_at_work$Month.of.absence,
                         sum)

barplot(hours_by_month,
        main = "Total Hours Absent by Month",
        xlab = "Month",
        ylab = "Absent Hours",)
```

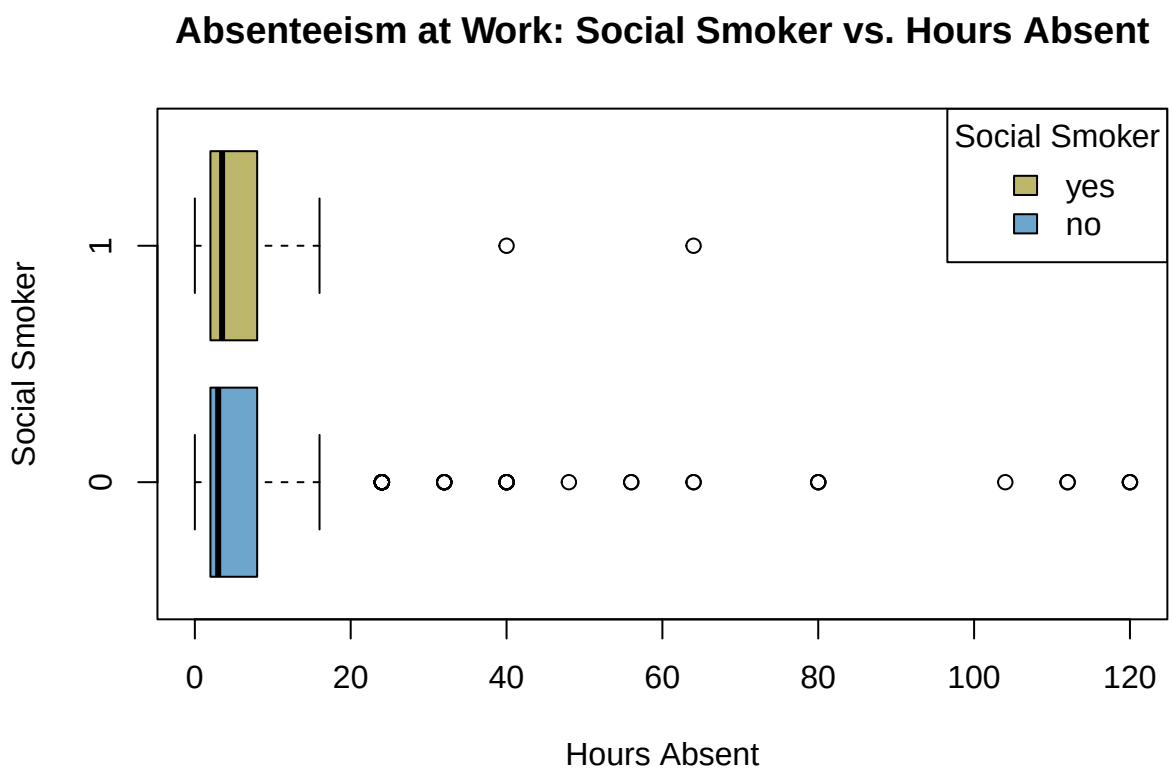


Comments: The data set shows a spike of total hours absent in the months of March and July.

(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

```
hours_absent <- Absenteeism_at_work$Absenteeism.time.in.hours
social_smoker <- Absenteeism_at_work$Social.smoker

boxplot(hours_absent ~ social_smoker,
        horizontal = TRUE,
        main = "Absenteeism at Work: Social Smoker vs. Hours Absent",
        xlab = "Hours Absent",
        ylab = "Social Smoker",
        #ylim = c(0,20),
        col = c("skyblue3","darkkhaki"))
legend("topright",
       legend = c('yes','no'),
       fill = c("darkkhaki","skyblue3"),
       title = "Social Smoker")
```



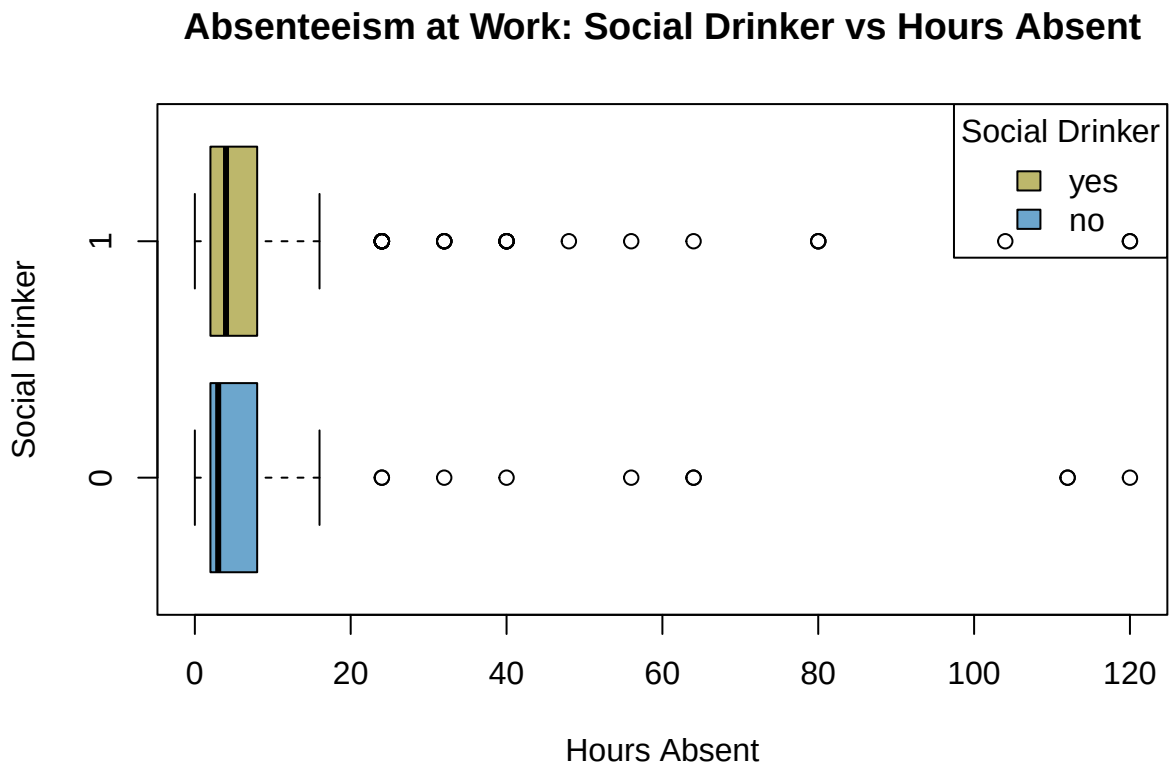
Comments: Due to the nature of the data being stored logically ,using TRUE and FALSE, the data is a bit inconclusive. The values are converted numerically into 1s and 0s (TRUE and FALSE respectively) so the data set is just a repetition of said numbers. As a result the median and quartiles don't offer any information that expand on the correlation of smoking socially and absenteeism from work.

(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

```
hours_absent <- Absenteeism_at_work$Absenteeism.time.in.hours
social_drinker <- Absenteeism_at_work$Social.drinker

boxplot(hours_absent ~ social_drinker,
        horizontal = TRUE,
        main = "Absenteeism at Work: Social Drinker vs Hours Absent",
        xlab = "Hours Absent",
        ylab = "Social Drinker",
        #ylim = c(0,20),
        col = c("skyblue3","darkkhaki"),)

legend("topright",
       legend = c('yes','no'),
       fill = c("darkkhaki","skyblue3"),
       title = "Social Drinker")
```



Comments: This graph falls victim to the same issues as the one before it, Hours Absent vs Social Smoker. Where the data set is converted into 0s and 1s and no information can be extracted from it as the original form of the data is nominal and can't be represented with meaning in a box plot.