

Lab1

Shreyash Singh

2024-01-21

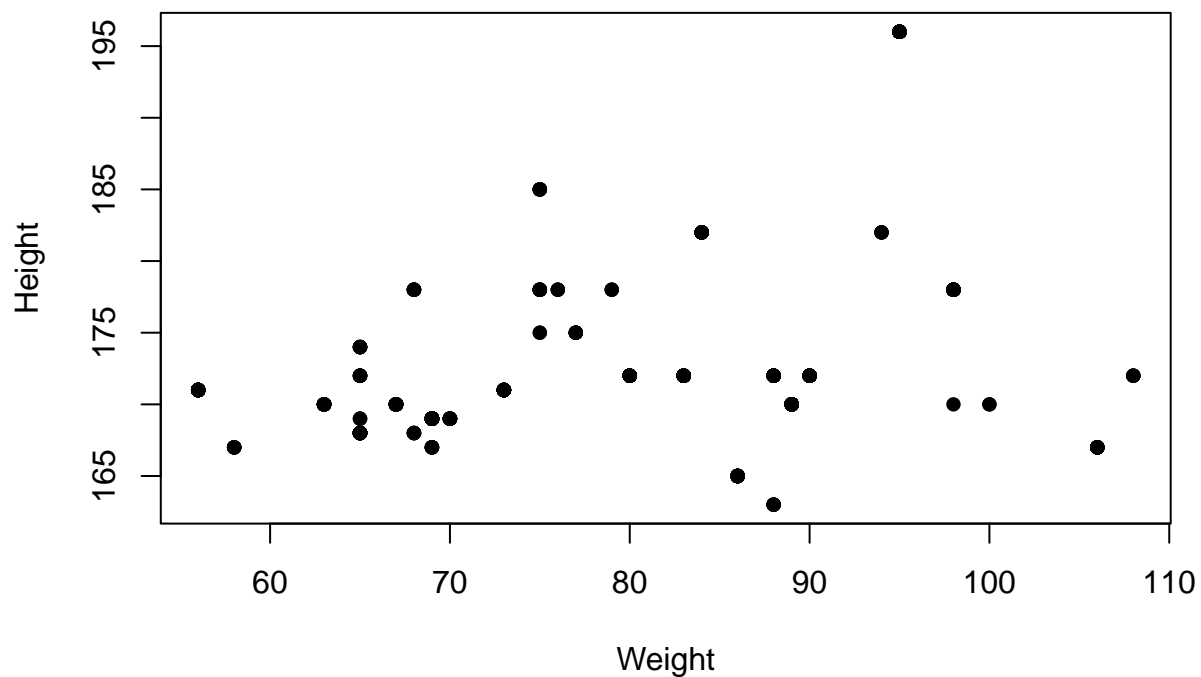
Reading the Data from the CSV File

```
my_Data <- read.csv("Absenteeism_at_work.csv", sep=";", header=TRUE)
```

1. Scatter plot of height vs. weight

```
plot(my_Data$Weight, my_Data$Height,  
     main="Scatter Plot of Height vs. Weight",  
     xlab="Weight", ylab="Height", pch=16)
```

Scatter Plot of Height vs. Weight

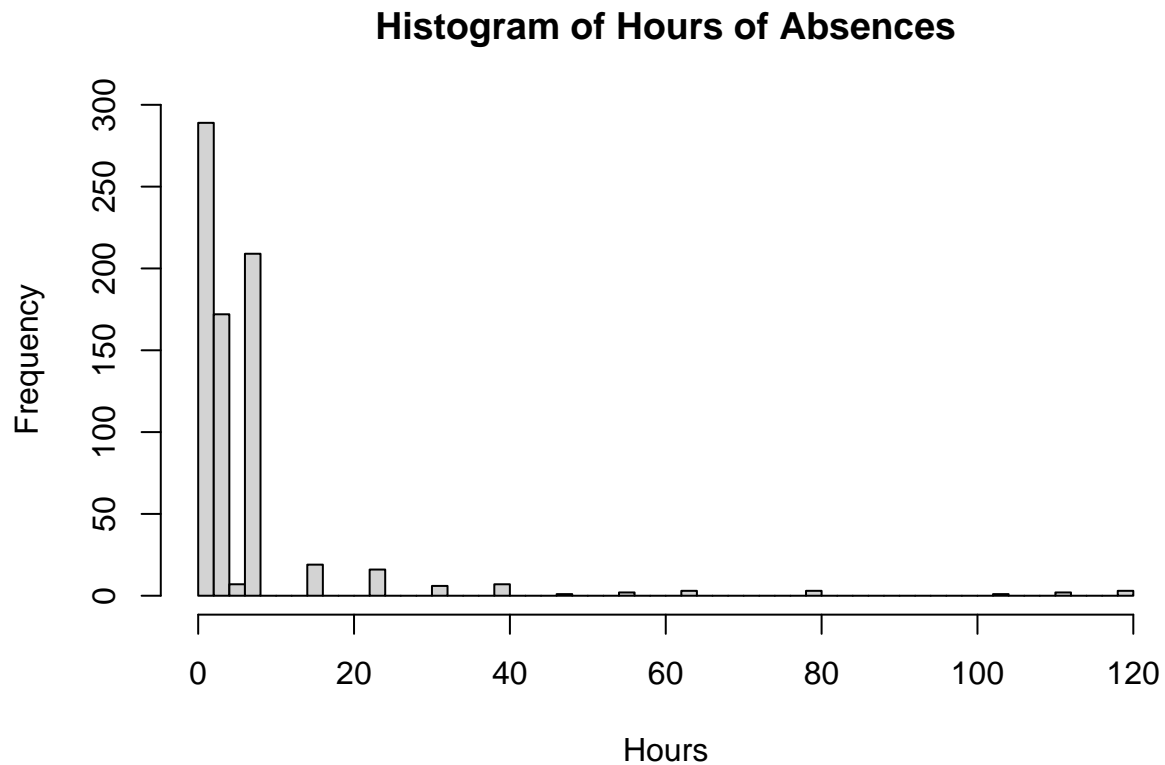


Comments:

1. There is no clear trend visible in the graph.
2. The data is has no relevance with the absenteeism.

2. Histogram of hours of absences

```
hist(my_Data$Absenteeism.time.in.hours,  
     breaks=60,  
     main="Histogram of Hours of Absences", xlab="Hours")
```

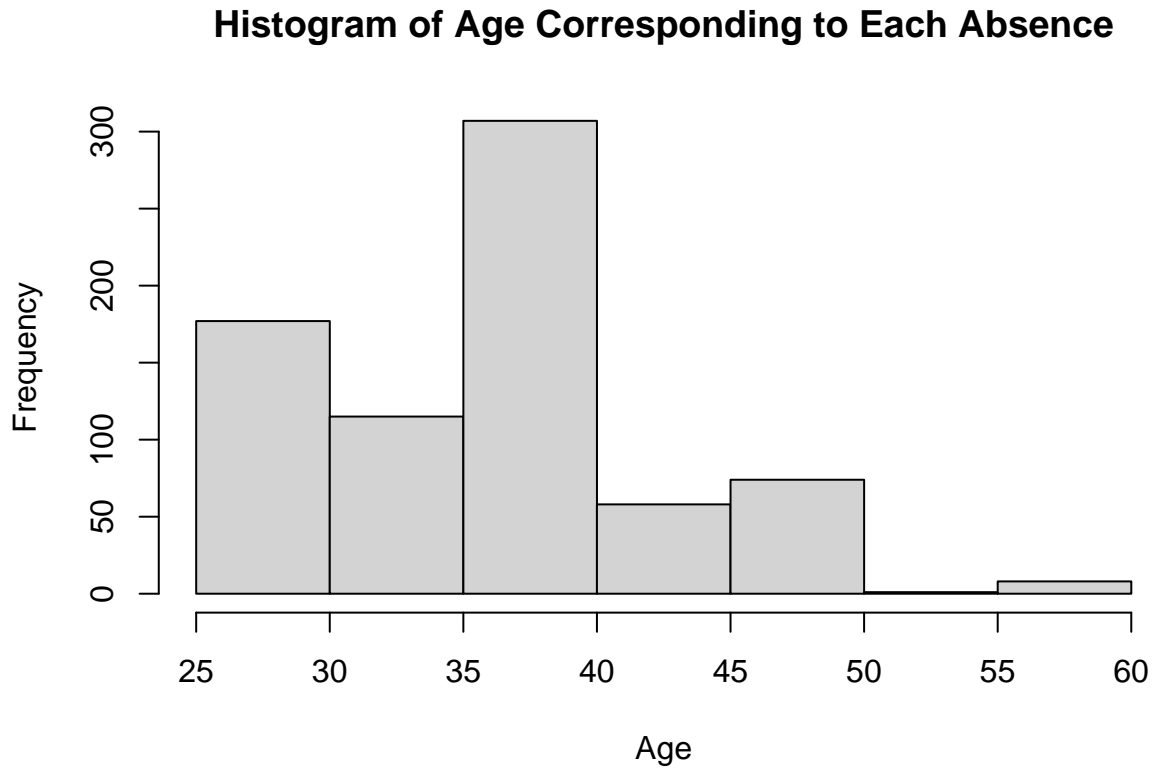


Comments:

1. Most absences fall within 0- 10 (approx) hours with frequency reaching 700.
2. It tells that most of the absences were upto 10 hours.

3. Histogram of age corresponding to each absence

```
hist(my_Data$Age,  
     main="Histogram of Age Corresponding to Each Absence",  
     xlab="Age")
```

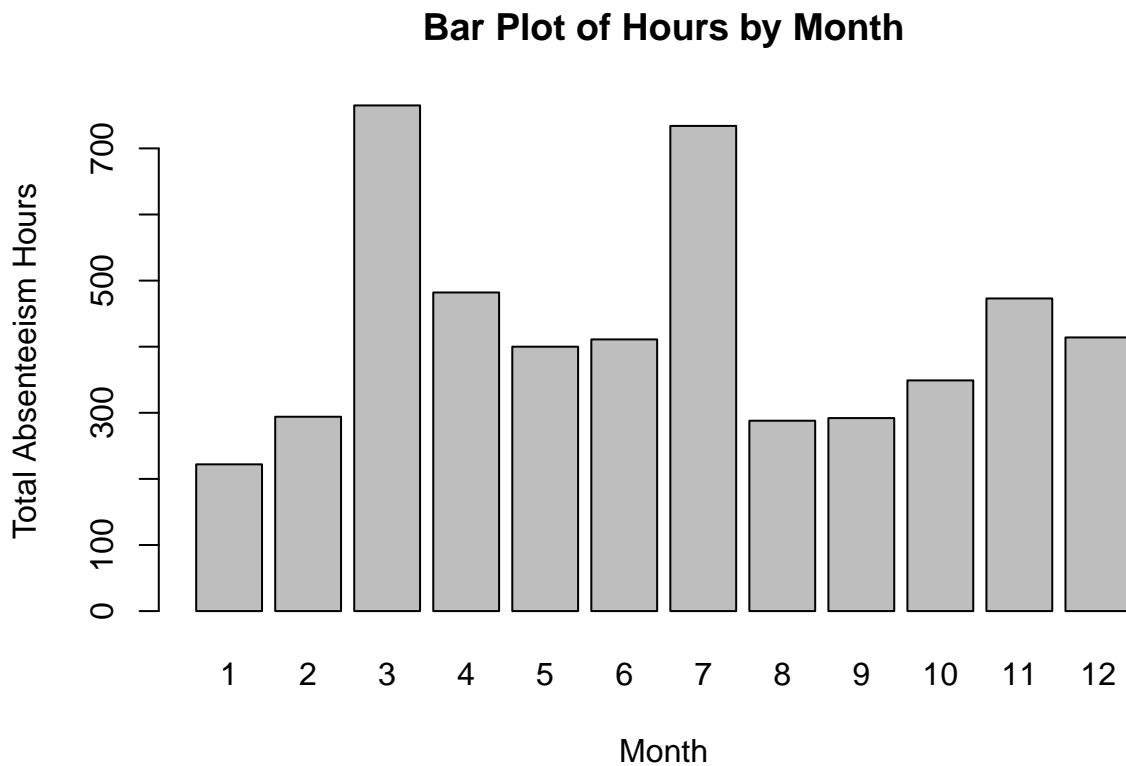


Comments:

- 1.The agegroup of 35-40 have the highest frequency reaching 300 (approx).
- 2.The agegroup of 50-55 have the lowest frequency.

4. Bar plot of hours by month

```
hours_by_month <- tapply(my_Data$Absenteeism.time.in.hours,  
                          my_Data$Month.of.absence, sum)  
hours_by_month <- hours_by_month[names(hours_by_month)!="0"]  
barplot(hours_by_month,  
        main="Bar Plot of Hours by Month",  
        xlab="Month",  
        ylab="Total Absenteeism Hours")
```



Comments:

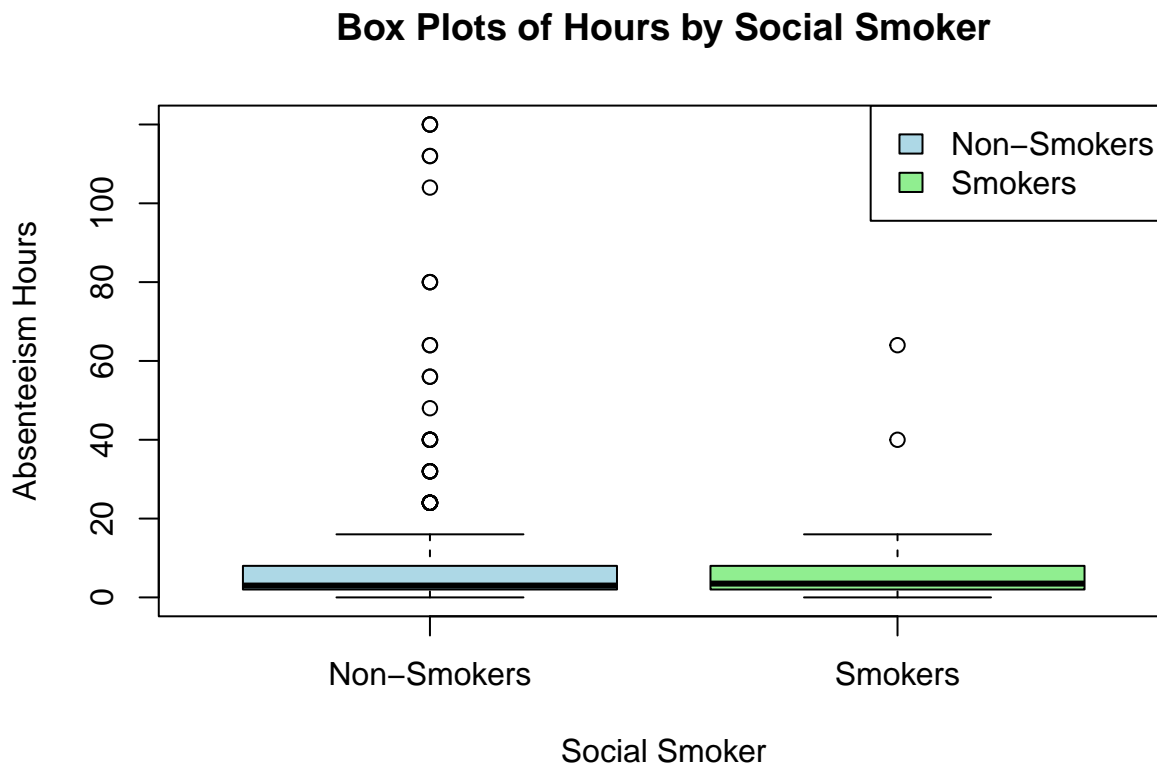
1. March and July have the most number of Absentee hours by month.
2. This graph can be used to determine which months are more or less prone to absenteeism

5. Box plots of hours by social smoker variable

```
non_smoker_data <- my_Data[my_Data$Social.smoker == 0, ]
smoker_data <- my_Data[my_Data$Social.smoker == 1, ]

# Creating a box plot for both non-smokers and smokers
boxplot(non_smoker_data$Absenteeism.time.in.hours,
        smoker_data$Absenteeism.time.in.hours,
        names = c("Non-Smokers", "Smokers"),
        main = "Box Plots of Hours by Social Smoker",
        xlab = "Social Smoker", ylab = "Absenteeism Hours",
        col = c("lightblue", "lightgreen"))

# Adding legend
legend("topright", legend = c("Non-Smokers", "Smokers"),
      fill = c("lightblue", "lightgreen"))
```



Comments:

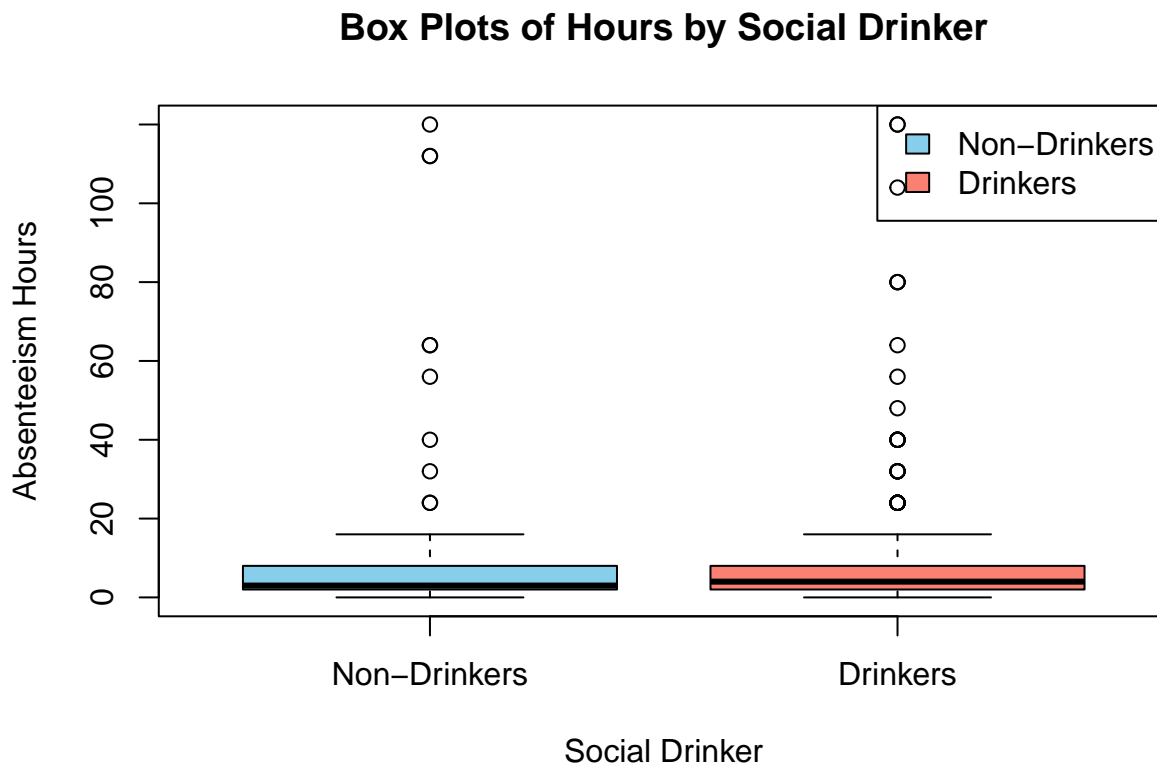
- 1.This plot represents the amount of Absentee hours for smokers and non smokers and can be used for finding trends.
2. Both the categories have almost the same Min,Max Inner quartile rance but looking at outliers non smokers have a larger variation.

6. Box plots of hours by social drinker variable

```
non_drinker_data <- my_Data[my_Data$Social.drinker == 0,]
drinker_data <- my_Data[my_Data$Social.drinker == 1,]

# Creating a box plot for both non-drinkers and drinkers
boxplot(non_drinker_data$Absenteeism.time.in.hours,
        drinker_data$Absenteeism.time.in.hours,
        names = c("Non-Drinkers", "Drinkers"),
        main = "Box Plots of Hours by Social Drinker",
        xlab = "Social Drinker", ylab = "Absenteeism Hours",
        col = c("skyblue", "salmon")) # You can change the colors here

# Adding legend
legend("topright", legend = c("Non-Drinkers", "Drinkers"),
      fill = c("skyblue", "salmon"))
```



Comments:

- 1.This plot represents the amount of Absentee hours for drinkers and non drinkers and can be used for finding trends.
- 2.Both the categories have almost the same Min,Max Inner quartile rance but looking at outliers drinkers have a larger variation.