Untitled

Joseph Harty

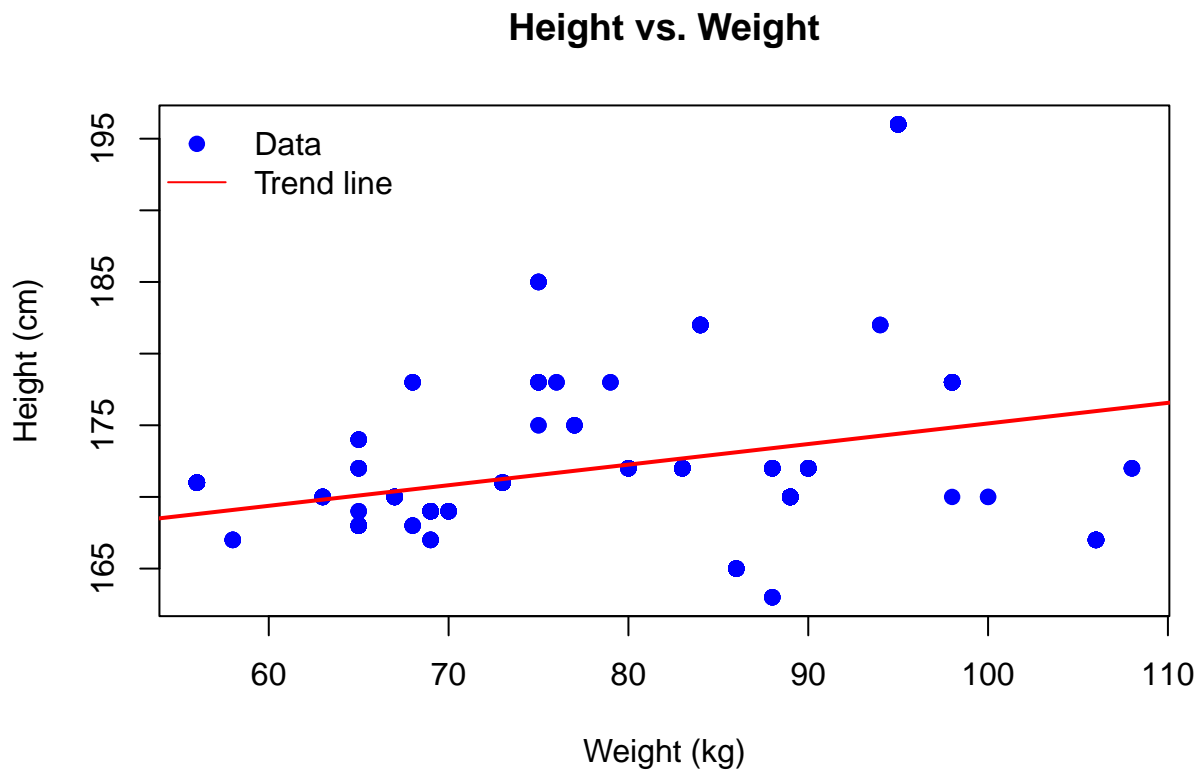## 2025-11-02

Load Data

```r
# (no change) load & pick columns
df <- read.csv("Absenteeism_at_work.csv", sep = ";", header = TRUE)
hours   <- df$Absenteeism.time.in.hours
age     <- df$Age
weight  <- df$Weight
height  <- df$Height
month   <- df$Month.of.absence
smoker  <- df$Social.smoker
drinker <- df$Social.drinker
```

#1 Height vs. Weight

```r
ok <- complete.cases(weight, height)
plot(weight[ok], height[ok],
     main = "Height vs. Weight",
     xlab = "Weight (kg)", ylab = "Height (cm)",
     pch = 19, col = "blue")

if (sum(ok) >= 2) {
  abline(lm(height ~ weight, data = data.frame(height = height[ok], weight = weight[ok])),
         col = "red", lwd = 2)
}
legend("topleft", legend = c("Data", "Trend line"),
       pch = c(19, NA), lty = c(NA, 1),
       col = c("blue", "red"), bty = "n")
```
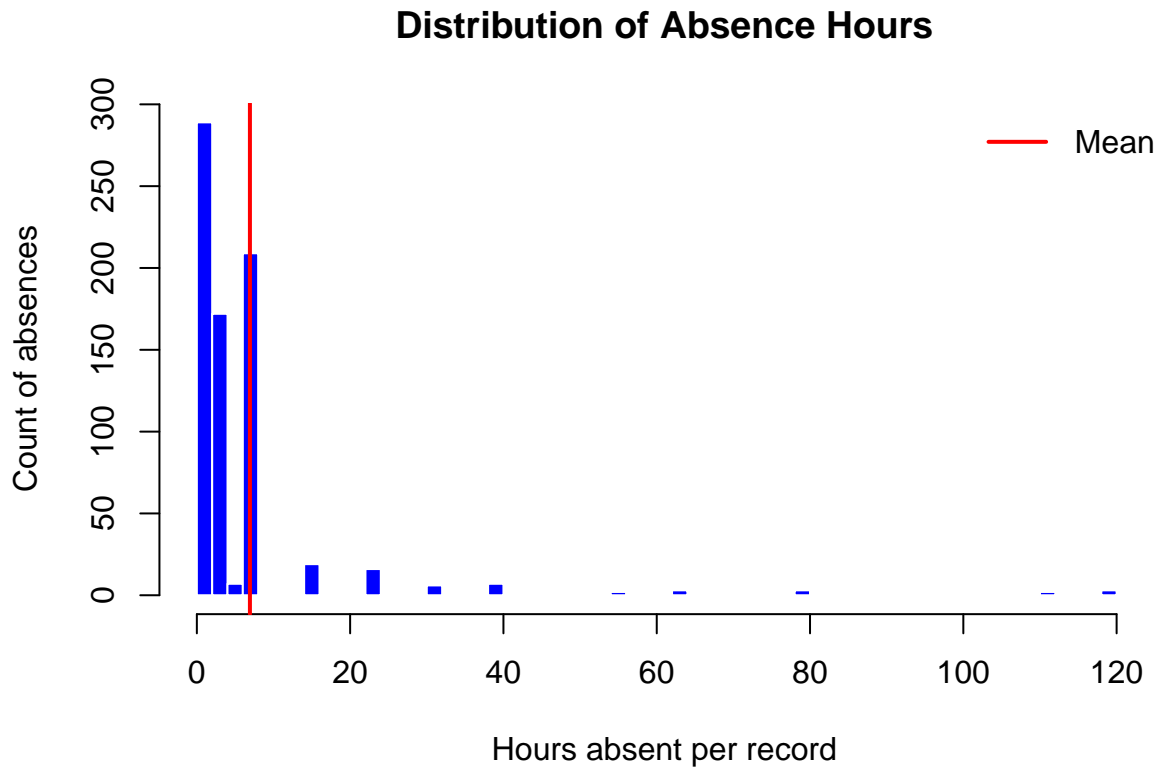
Height vs. Weight

In this graph there is a clear correlation of height to weight. The more someone is tall, the more their weight would be. There are a few outliars but the ammount is not signifant.

#2 Histogram: Absence Hours

```r
hist(hours,
     main = "Distribution of Absence Hours",
     xlab = "Hours absent per record",
     ylab = "Count of absences",
     col = "blue", border = "white",
     breaks = seq(0, max(hours, na.rm = TRUE) + 2, by = 2))
abline(v = mean(hours, na.rm = TRUE), col = "red", lwd = 2)
legend("topright", legend = "Mean", col = "red", lwd = 2, bty = "n")
```

Distribution of Absence Hours

#Most people will take half days off and full days off assuming a full work day is 8 hours long. If people take more than a half day, they will just take the full day off. The most they will miss is the first few hours, could be due to being late, and appoint or some other reason like that.

3# Histogram: Employee Age

```
hist(age,
     main = "Distribution of Employee Age",
     xlab = "Age (years)",
     ylab = "Number of records",
     col = "green", border = "white")
abline(v = mean(age, na.rm = TRUE), col = "red", lty = 2, lwd = 2)
legend("topright", legend = "Mean age", col = "red", lty = 2, bty = "n")
```
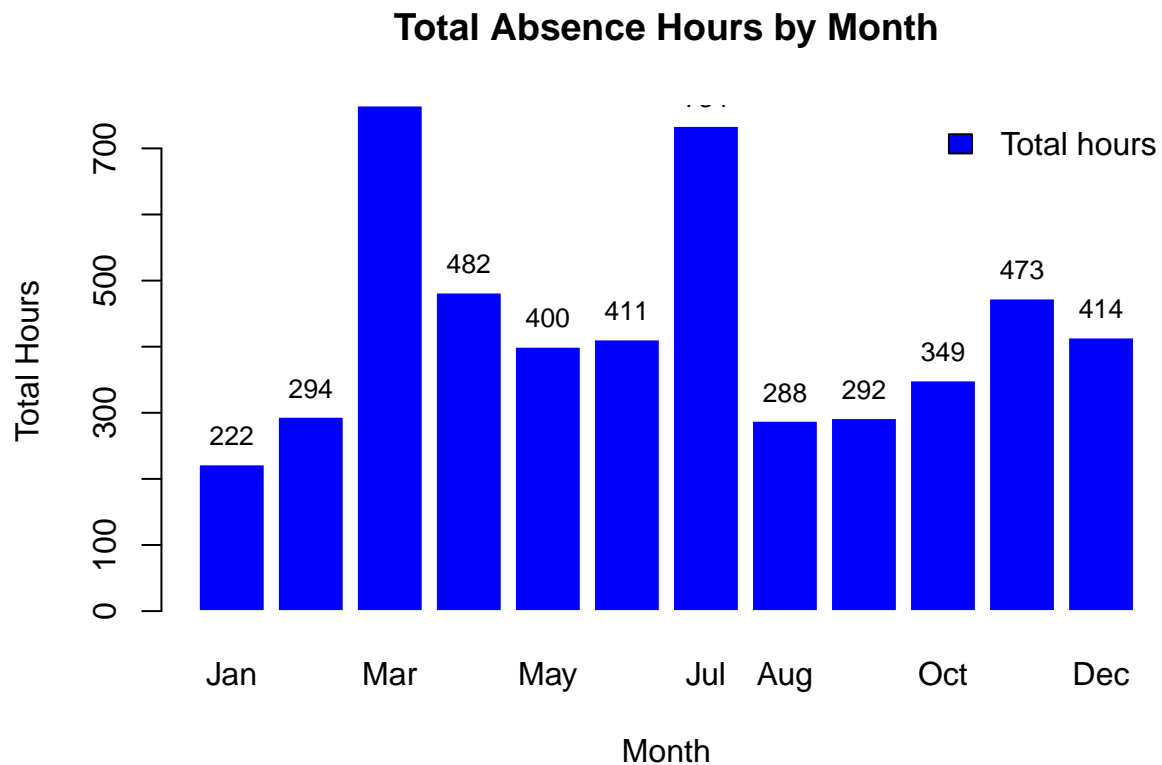
Distribution of Employee Age

This graph would suggest that the age who misses the most is ages 36-38 by a signigant portion. This graph doesnt show us the spread of the number of people at this age so the graph coule be misleading if a signifagnt portion of employee are in the age of 36-38.

4# Total Absence Hours by Month

```r
valid_month <- month %in% 1:12
total_by_month <- tapply(hours[valid_month],
                         factor(month[valid_month], levels = 1:12),
                         sum, na.rm = TRUE)

bp <- barplot(total_by_month,
              main = "Total Absence Hours by Month",
              xlab = "Month", ylab = "Total Hours",
              names.arg = month.abb,
              col = "blue", border = "white")
text(bp, total_by_month, labels = round(total_by_month), pos = 3, cex = 0.8)
legend("topright", legend = "Total hours", fill = "blue", bty = "n")
```
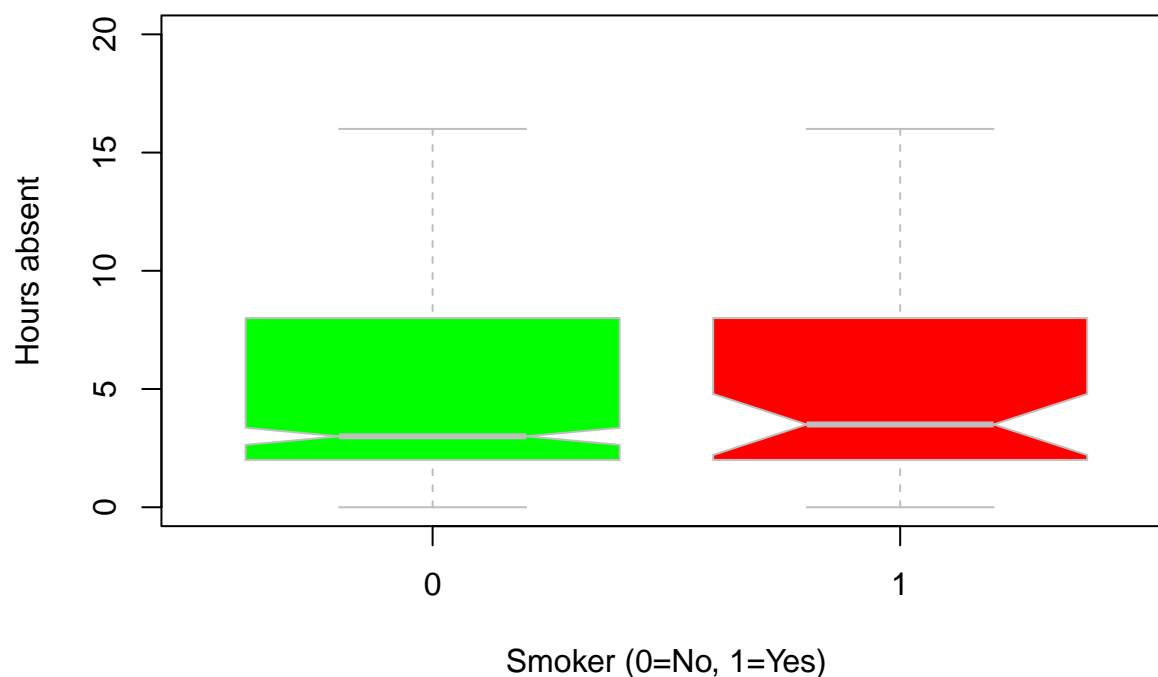
Total Absence Hours by Month

There are two months with higher than normal hours absent which are March and July. This could be for the summer for July and then people being sick for March.

5# Absence Hours by Social-Smoker

```r
boxplot(hours ~ smoker,
        main = "Absence Hours by Social-Smoker Status",
        xlab = "Smoker (0=No, 1=Yes)",
        ylab = "Hours absent",
        col = c("green","red"),
        notch = TRUE, border = "gray",
        ylim = c(0, 20))
```

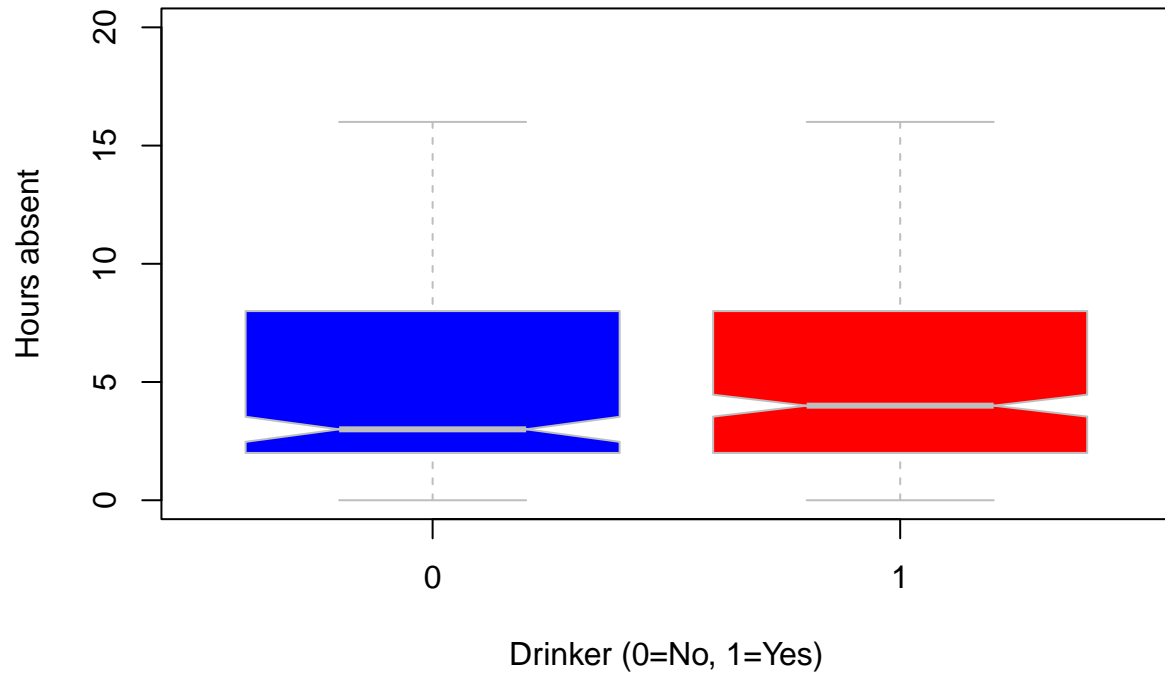## Absence Hours by Social–Smoker Status

Both groups have similar medians. With the notches overlapping shows that there is no signifiacant difference between the two.

6# Absence Hours by Social-Drinker

```r
boxplot(hours ~ drinker,
        main = "Absence Hours by Social-Drinker Status",
        xlab = "Drinker (0=No, 1=Yes)",
        ylab = "Hours absent",
        col = c("blue","red"),
        notch = TRUE, border = "gray",
        ylim = c(0, 20))
```

Absence Hours by Social–Drinker Status

Both groups have similar medians. With the notches overlapping shows that there is no signifiacant difference between the two.