Data Visualization

Business Analytics

Stefan Feuerriegel

Today's Lecture

Objectives

1. Visualizing data in R graphically
2. Choosing appropriate plots in a given context

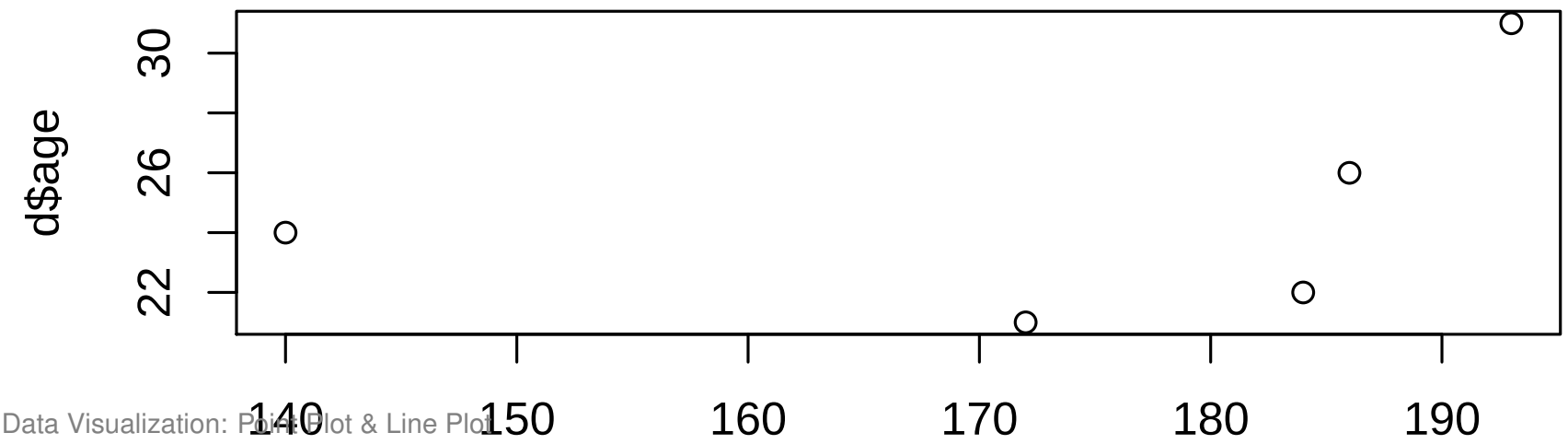
Outline

Outline

Point Plot

- Creating simple point plots (also named scatter plots) via `plot(...)`
- Relies upon vectors denoting the x-axis and y-axis locations
- Various options can be added to change appearance

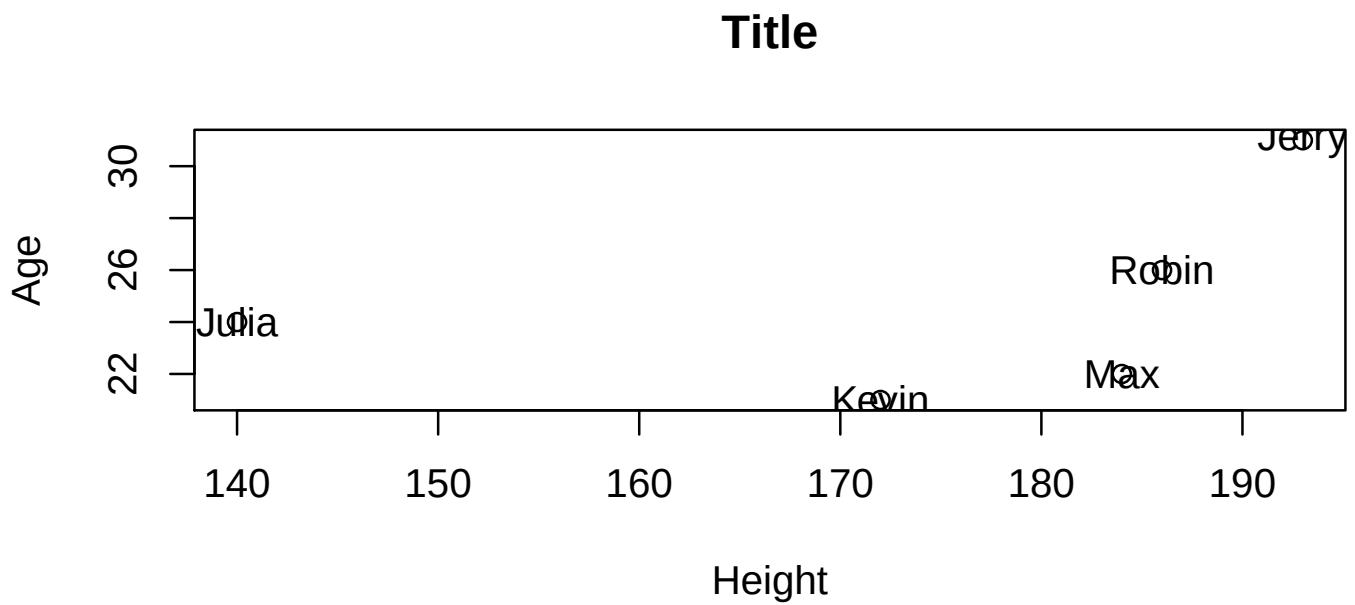
```r
d <- data.frame(name = c("Julia", "Robin", "Kevin", "Max", "Jerry"),
                height = c(140, 186, 172, 184, 193),
                shoesize = c(39, 44, 41, 43, 45),
                age = c(24, 26, 21, 22, 31))
plot(d$height, d$age)
```



Adding Title, Labels and Annotations

- Title is added through additional parameter `main`
- Axis labels are set via `xlab` and `ylab`
- Annotations next to points with `text(...)`

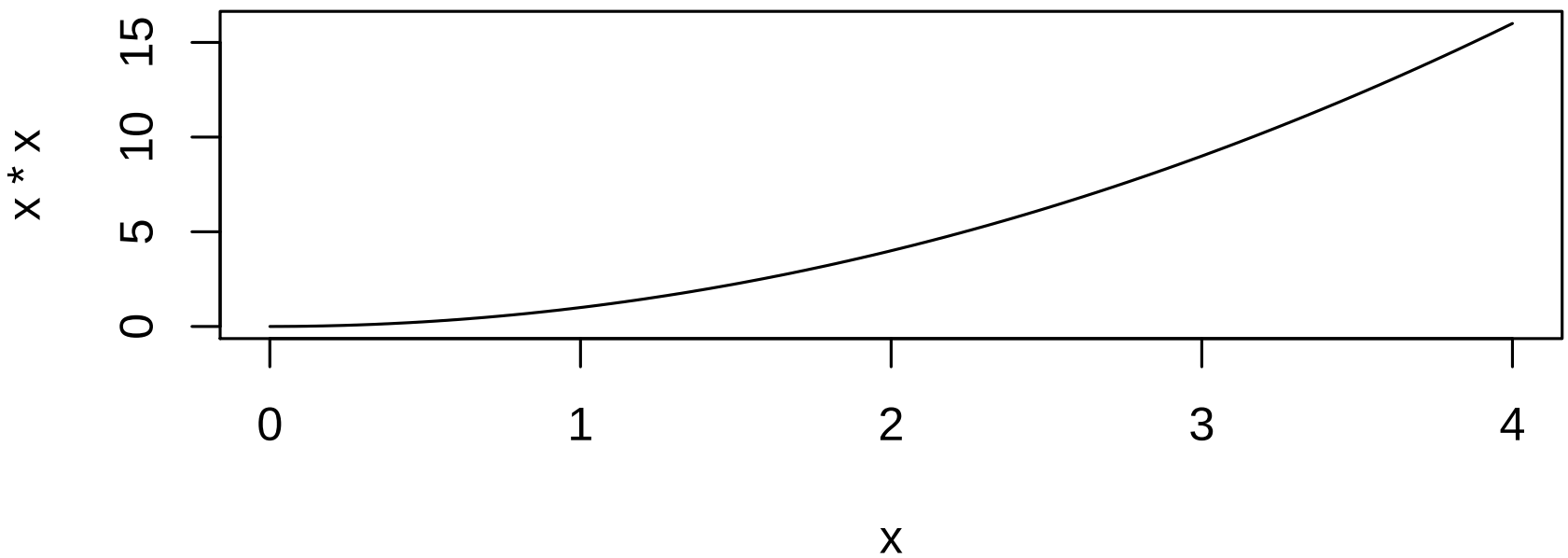
```r
plot(d$height, d$age,
  main="Title",                 # an overall title for the plot
  xlab="Height", ylab="Age") # titles for x and y axis
text(d$height, d$age, d$name) # d$name are labels
```



Line Plot

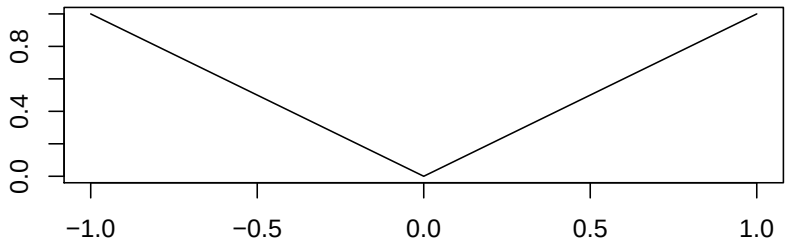
Generate line plot using the additional option `type="l"`

```
x <- seq(0, 4, 0.01)
plot(x, x*x, type="l")
```



Exercise: Plotting

```
x <- seq(-1, +1, 0.01)
```



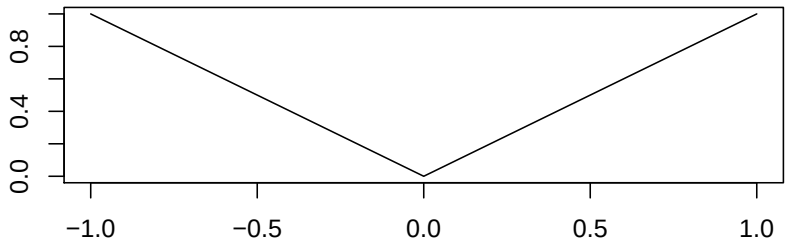
Question

- How would you reproduce the above plot?
  - `plot(x, kink(x), type="l", main="")`
  - `plot(x, kink(x), type="l", lab="")`
  - `plot(x, abs(x), type="l", ylab="", xlab="")`

Solution: 3

Exercise: Plotting

```
x <- seq(-1, +1, 0.01)
```



Question

- How would you reproduce the above plot?
  - `plot(x, kink(x), type="l", main="")`
  - `plot(x, kink(x), type="l", lab="")`
  - `plot(x, abs(x), type="l", ylab="", xlab="")`

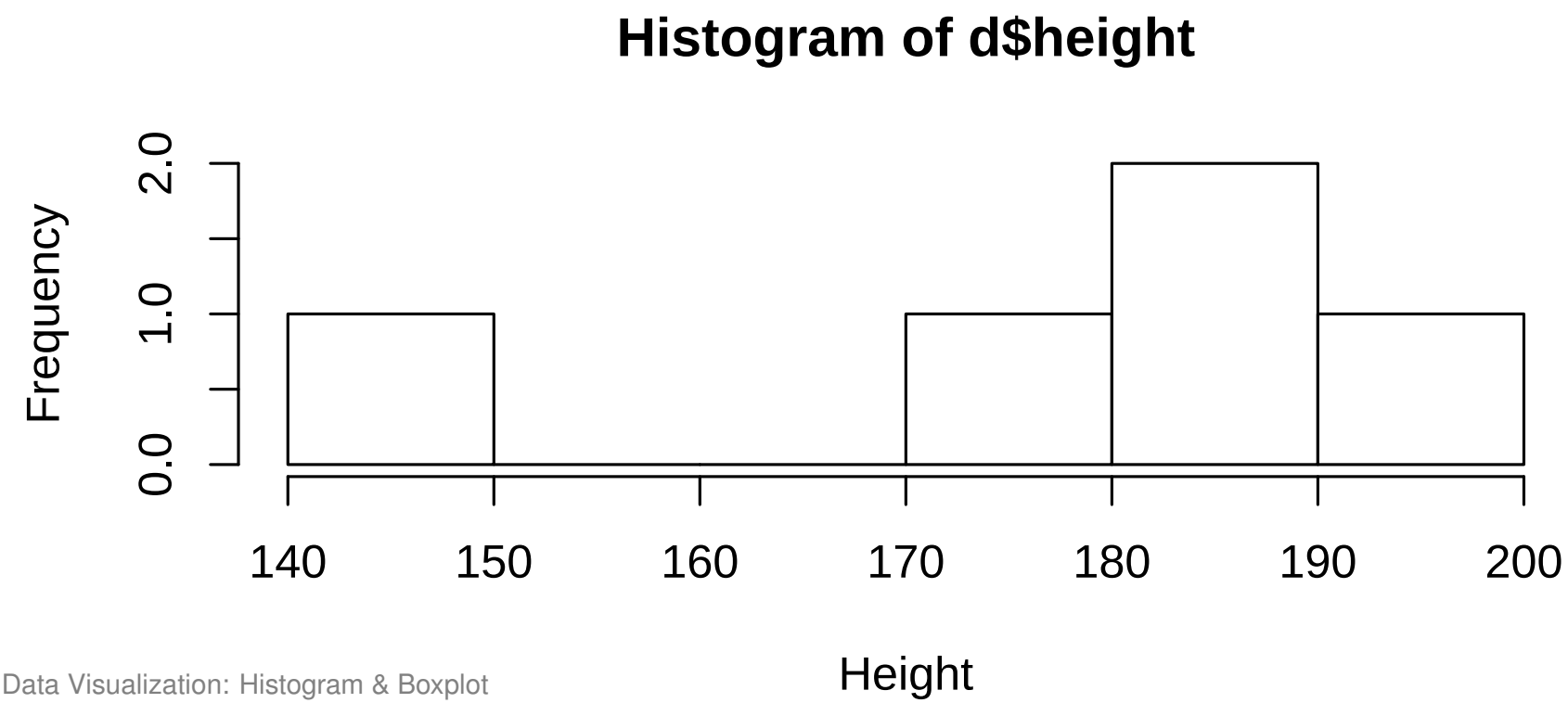
Solution: 3

Outline

Histograms with Frequencies

- Histograms are a graphical representation of the distribution of data
- Created via `hist(data)` to get fixed width of classes
- *y*-axis gives frequency → estimating probability distribution

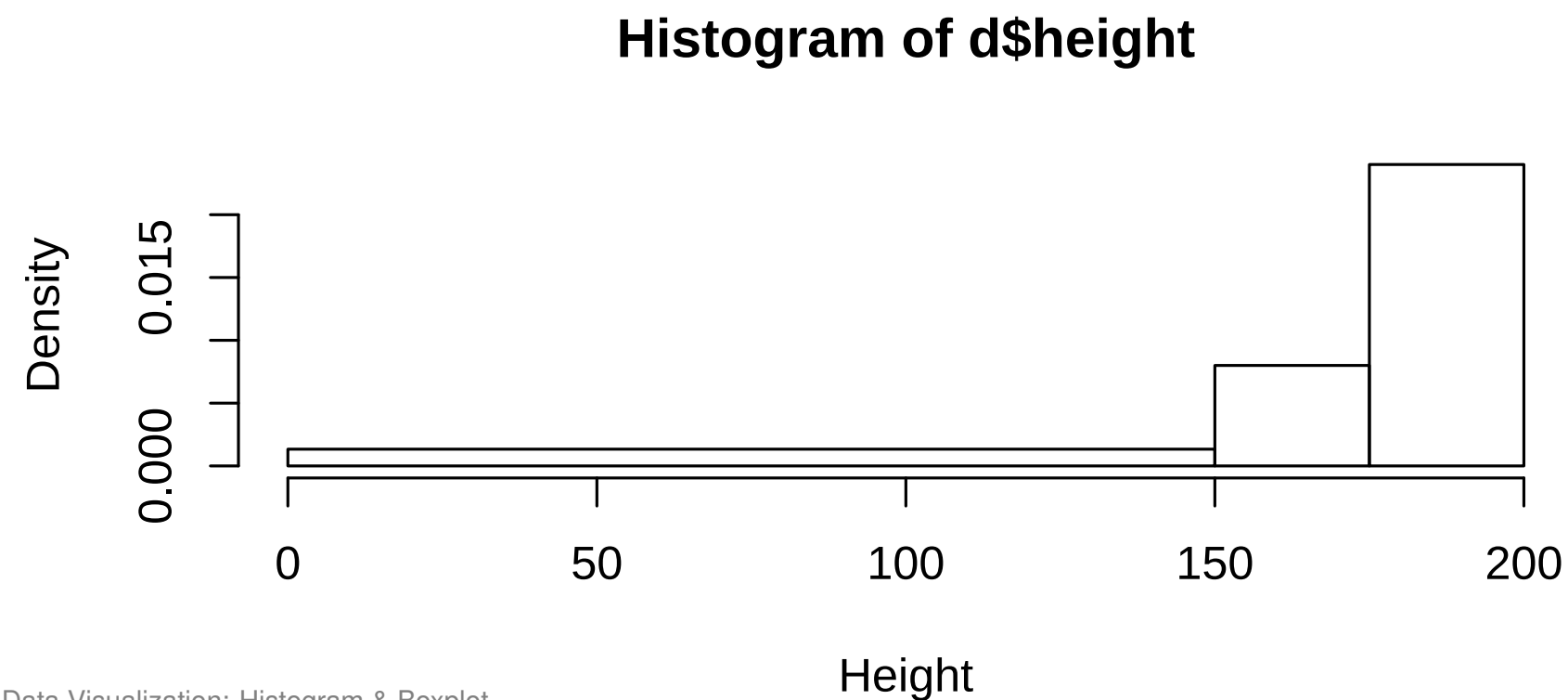
```
hist(d$height,
     xlab = "Height")
```



Histograms with Densities

- Density (1.00 $\widehat{=}$ 100 %) on *y*-axis via `hist(data, freq=FALSE)`
- Parameter `breaks=b` gets a variable width of classes

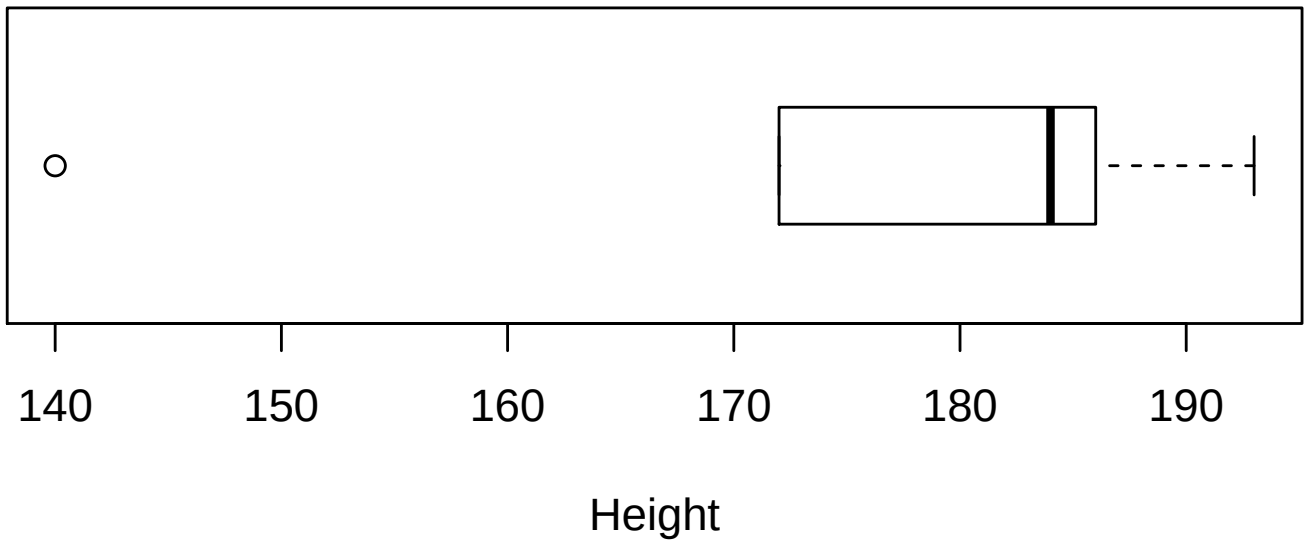
```
b <- c(0, 150, 175, 200)
hist(d$height, breaks = b,
     xlab = "Height")
```



Boxplot

- Use `boxplot(...)` to draw boxplot visualizing outliers (as circles), range and quartiles
- Default is vertical mode (`horizontal=FALSE`)

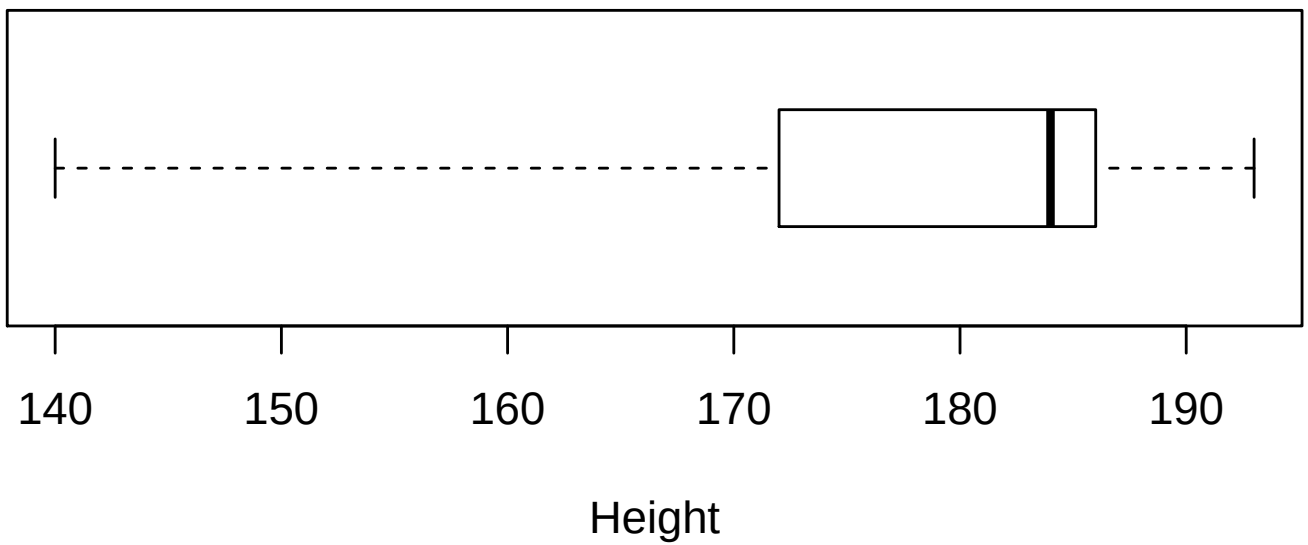
```r
boxplot(d$height, horizontal=TRUE,
        xlab="Height")
```



Boxplot

- To prevent highlighting of outliers, use `range=0`

```
boxplot(d$height, horizontal=TRUE,
        xlab="Height", range=0)
```



Outline

Comparing Distributions

BI Case Study

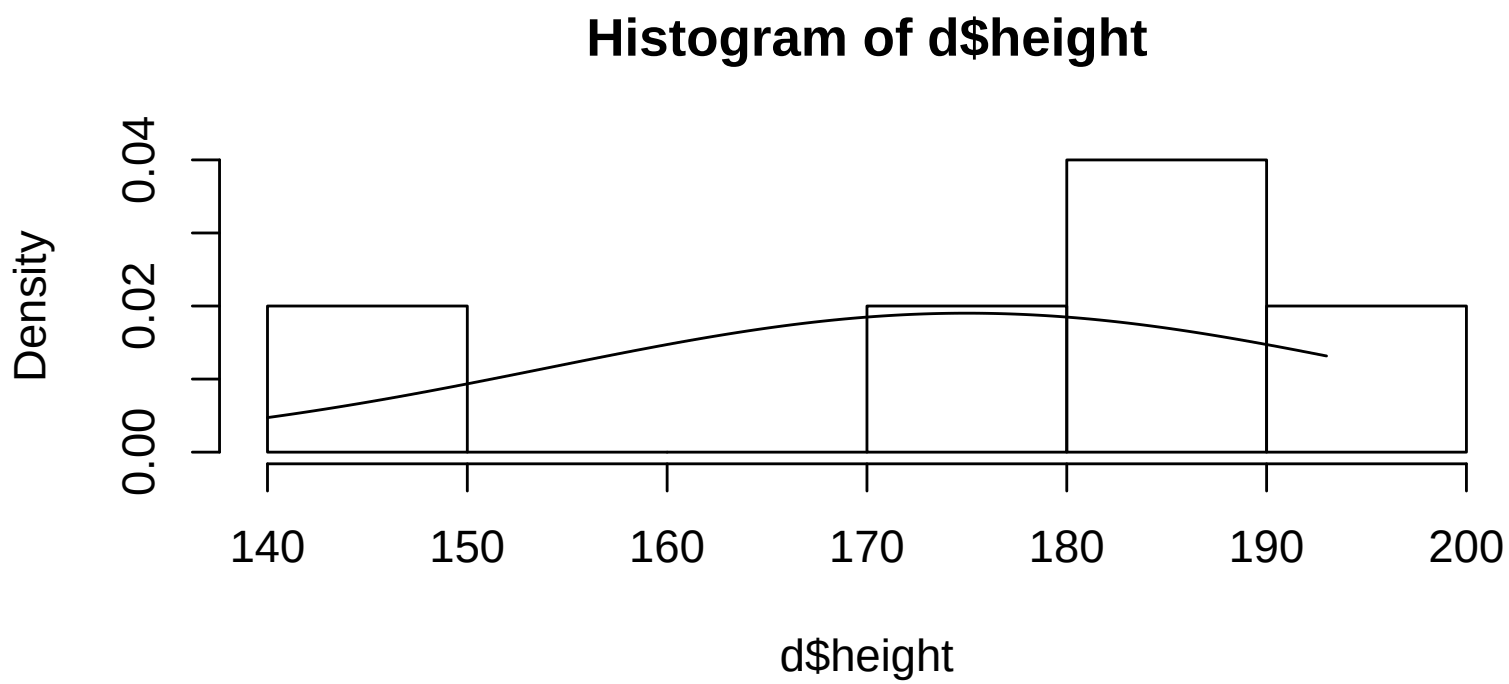
Is the duration of lawsuits normally distributed?

Solutions:

1. Histogram (also showing baseline distribution)
2. Q-Q plot

Comparing Distributions: Histogram

- Not recommended: Compare histogram and corresponding normal distribution by overlapping plot
-
```
hist(d$height, freq=FALSE)
xx <- seq(min(d$height), max(d$height), 0.01)
lines(xx, dnorm(xx, mean=mean(d$height),
                    sd=sd(d$height)))
```

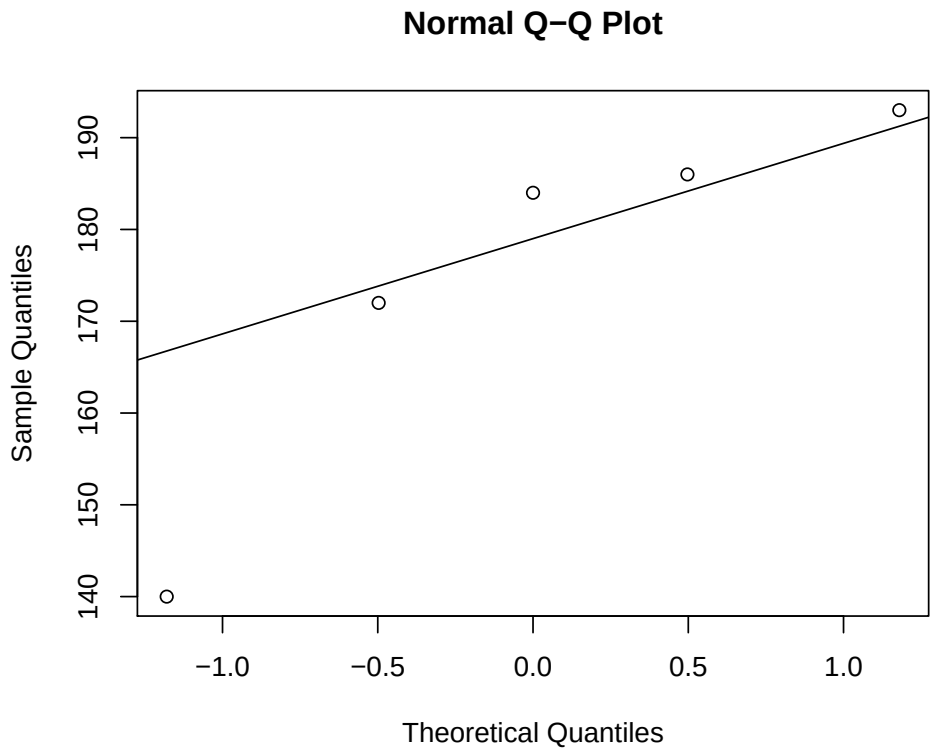


Q-Q Plot

- Q-Q plot ("Q" stands for quantile) compares two probability distributions by plotting their quantiles against each other
- `qqnorm(d)`, `qqline(d)` use standard normal distribution

```r
# plot sample against
# theoretical standard
# normal distribution
qqnorm(d$height)

# line that represents
# true normal distribution
qqline(d$height)
```

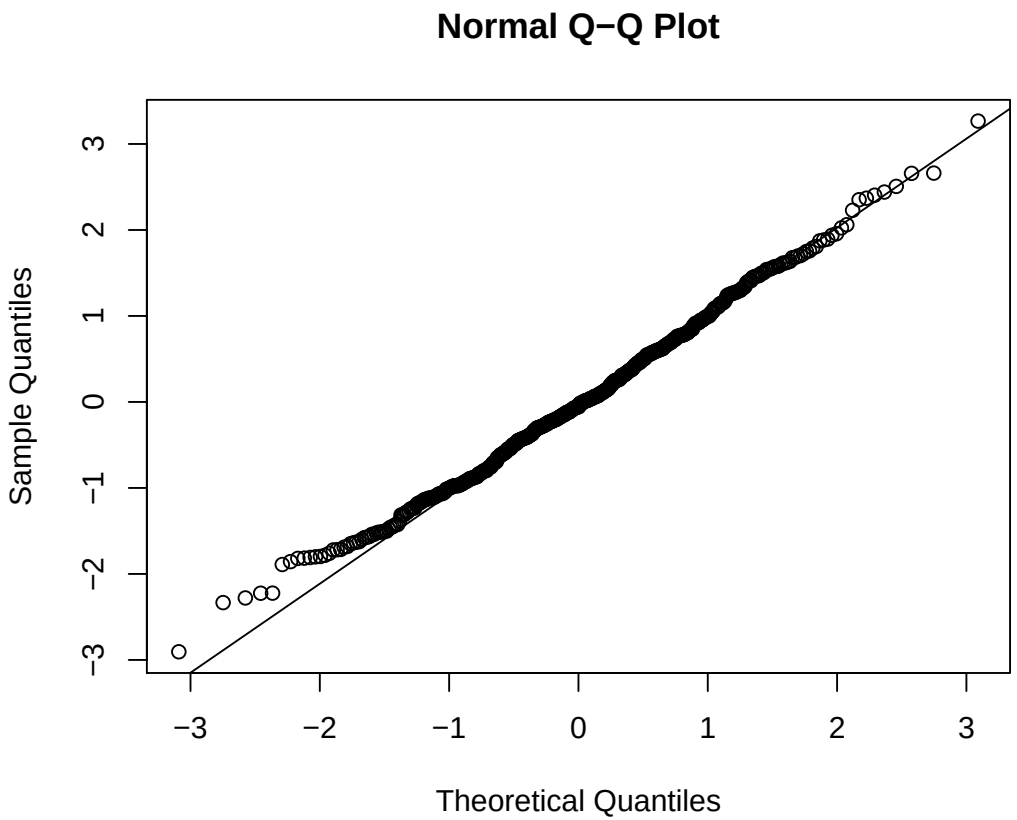


→ No standard normal distribution because of strong offset at tails

Exercise: Q-Q Plot

Question

- True or false: are the values normally distributed?

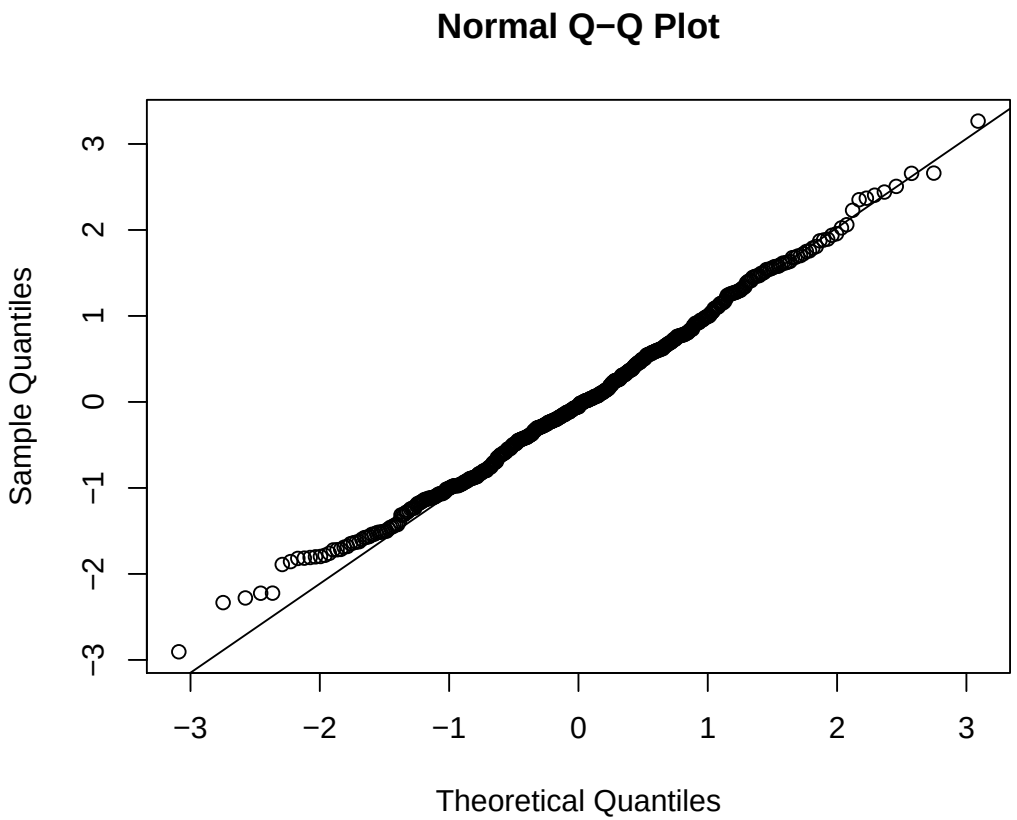


→ Strong linear pattern suggests standard normal distribution

Exercise: Q-Q Plot

Question

- True or false: are the values normally distributed?



→ Strong linear pattern suggests standard normal distribution

Outline

Guideline to Choosing Plots

Data Structure	**Plot**	**R Command**
Relationship (2-dim.)	Point Plot	`plot(x, y)`
Evolving Time Series	Line Plot	`plot(x, y, type="l")`
Frequencies (Fixed Ranges)	Histogram	`hist(d)`
Densities (Variable Ranges)	Histogram	`hist(d, freq=FALSE, breaks=b)`
Distribution Variation	Boxplot	`boxplot(d)`
Distribution Comparison	Q-Q Plot	`qqnorm(d),qqline(d)`