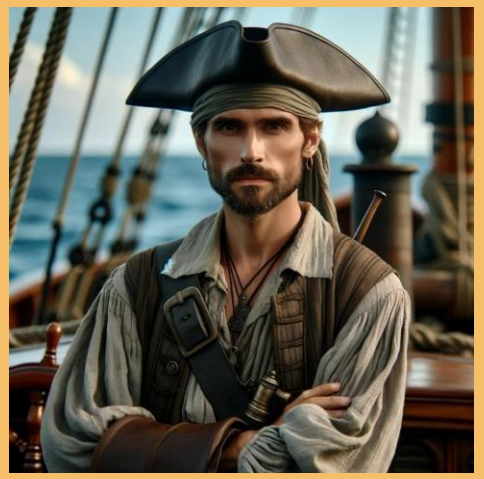
Tackling univariate and bivariate statistics

Let’s take the temperature

Your enthusiasm for 2 days of stats!

desperate



Great!

0

1

Round table

Name

What do you feel about statistics ?

Are you an R user ?

What is your favorite movie ?

Nancy Rebout

Before returning to board the data

- Identifying the nature of the variables = the most important thing

Quantitative variable = digital

Qualitative variable = text

Quantitative variable

- Discrete variable: limited number of possible outcomes (no commas)
  - *Number of drinks per pirate*
- Continuous variable: numbers that can be decimal.
  - *Salary*
  - *Parrot size*
  - *Quantity of alcohol*

Too much conflicts in pirates

Testimony capitain DataRisk

Nothing's the same on the deck of Captain DataRisk's proud ship !

It's total chaos. The sailors are tearing each other to shreds.

But our cunning captain has a whim: collect data like buried treasure and use statistics as a compass to navigate through the tumult.

Hoist the sails and let the adventure begin!

Too much conflict on the pirate ship

- Potential explanations:
  - Fatigue
  - Headache
  - Disputes between managers and employees

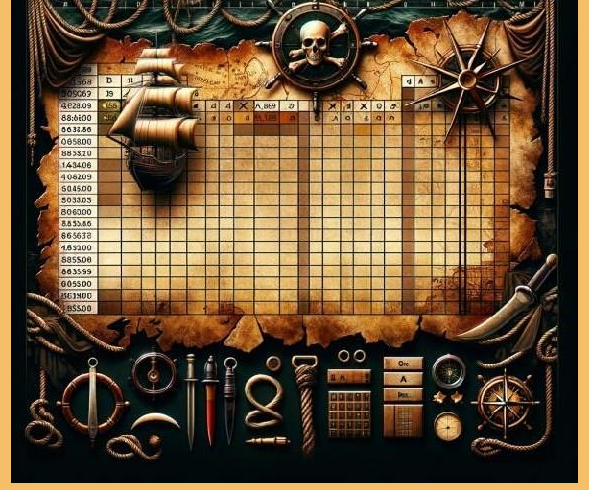
The captain, closely observing his crew, notices that fatigue has taken hold of his sailors. To better understand this, he decides to look at several numerical variables:

- **Age of the Sailors**.

- **Blood Iron Level**

- **Blood pressure**

These measurements, like nautical charts, will guide the captain towards solutions to reinvigorate his crew.



Dataset

We have respecte d the principles of the RGPD

```
> head(pirates)
      Name Age Iron_Level Systolic_Tension Diastolic_Tension
1 Pirate_1  63       63.4               128                 84
2 Pirate_2  69       71.3               112                 80
3 Pirate_3  66      138.5               114                 78
4 Pirate_4  73      109.1               120                 75
5 Pirate_5  64      145.8               109                 75
6 Pirate_6  70      136.9               126                 76
```

Univariate statistics:

Variable description

Quantitative variable

Variables taken one by one → univariate

Quantitative variable → Graph: Histogram

Quantitative variable → Metric: Mean, standard deviation, quantiles

Central tendency indicators

The mean/average :

the most widely used central tendency indicator.

We sum the values

We divide by the number of elements

$$\bar{x} = \frac{1}{n}\sum_{i} x_i$$

Pirate competition

- Number of treasures (classic average)
- Number of treasures recovered weighted by value of treasure (weighted average)

Central tendency indicators

Weighted average: weights/coefficients are associated

We sum the values multiplied by their weights

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

We divide by the sum of weights

Meaning of average

- Corresponds to the value that all observations would take if they were identical.

- Indicate a centrality

Description of variable averages

```
> print(averages)
  Average_Age Average_Iron_Level Average_Systolic_Tension Average_Diastolic_Tension
1      64.975           111.2325                    118.5                     79.75
```

Normal iron values :
60 to 170 µg/dL for men

Description of variable averages

```
> print(averages)
  Average_Age Average_Iron_Level Average_Systolic_Tension Average_Diastolic_Tension
1      64.975           111.2325                    118.5                     79.75
```

Normal blood pressure values:
120/80 mmHg for men

Description of variable averages

```
> print(averages)
  Average_Age Average_Iron_Level Average_Systolic_Tension Average_Diastolic_Tension
1      64.975           111.2325                    118.5                    79.75
```

Normal blood pressure values:
Must be less than 80 mmHg

Description of variable averages

```
> print(averages)
  Average_Age Average_Iron_Level Average_Systolic_Tension Average_Diastolic_Tension
1      64.975           111.2325                    118.5                     79.75
```

Not so young

Conclusion: pirates are old...

The captain feels bitter, saddened by his crew's retirement.

Recruiting new sailors is tough, requiring patience and adapting to Generation Z.

He realizes that the average doesn't capture reality on board. The median, more representative of his crew, might be a better guide.

At sea and in statistics, discernment is essential.

Central tendency indicators

median: separates the range of variation between 2 parts of equal size.

207 cm

40 cm

100 cm

74 cm

50% smaller characters than the median

132 cm

median

178 cm

50% larger characters than the median

400 cm

Average height = 116 cm

median = 116 cm

Central tendency indicators

Median: separates the range of variation between 2 parts of equal size.

$$\tilde{x} = x_{(n+1)/2}$$

if n odd

$$\tilde{x} = \frac{x_{n/2} + x_{(n/2)+1}}{2}$$

if n even

Description of variable medians

```
> print(medians)
  Median_Age Median_Iron_Level Median_Systolic_Tension Median_Diastolic_Tension
1       63.5               110                   117.5                     78.5
```

Not so young

[60 to 170 µg/dL].

[80 to 120 µg/dL]

< 80 µg/dL

The captain, grumpy and skeptical, mutters under his breath, "The median, shipmates, it doesn't shed much light on us!" He accuses us of a touch of bad faith.

He sees the crew as a motley bunch, with differences as vast as the ocean. To grasp these variations, he decides to explore dispersion indicators.

"Let's dive into the standard deviation, the variance, maybe even the interquartile range," he says.

When navigating the fog of data, it's best to have all the tools at hand.

Dispersion parameters

- The position parameters do not describe the dispersion of the distribution.
- Scattering of values around central values: to see whether the population is homogeneous or heterogeneous.
- Min, max
- Standard deviations
- Quartiles

Understanding dispersion indicators

On average, we scored in the center of the target, yes!

variation +++

variation ---

Dispersion indicators

Extent (= amplitude):

This is the difference between the maximum and minimum values.

Used when you're not interested in dispersion details.

scope: $x_{max} - x_{min}$

range: 400–40 = 360 cm

Dispersion indicators

Quantiles :

Quartiles subdivide the range of variation into 4 intervals of equal size.

Size average = 87 cm

Size average = 192.5 cm

40 cm

74 cm

100 cm

132 cm

178 cm

207 cm

400 cm

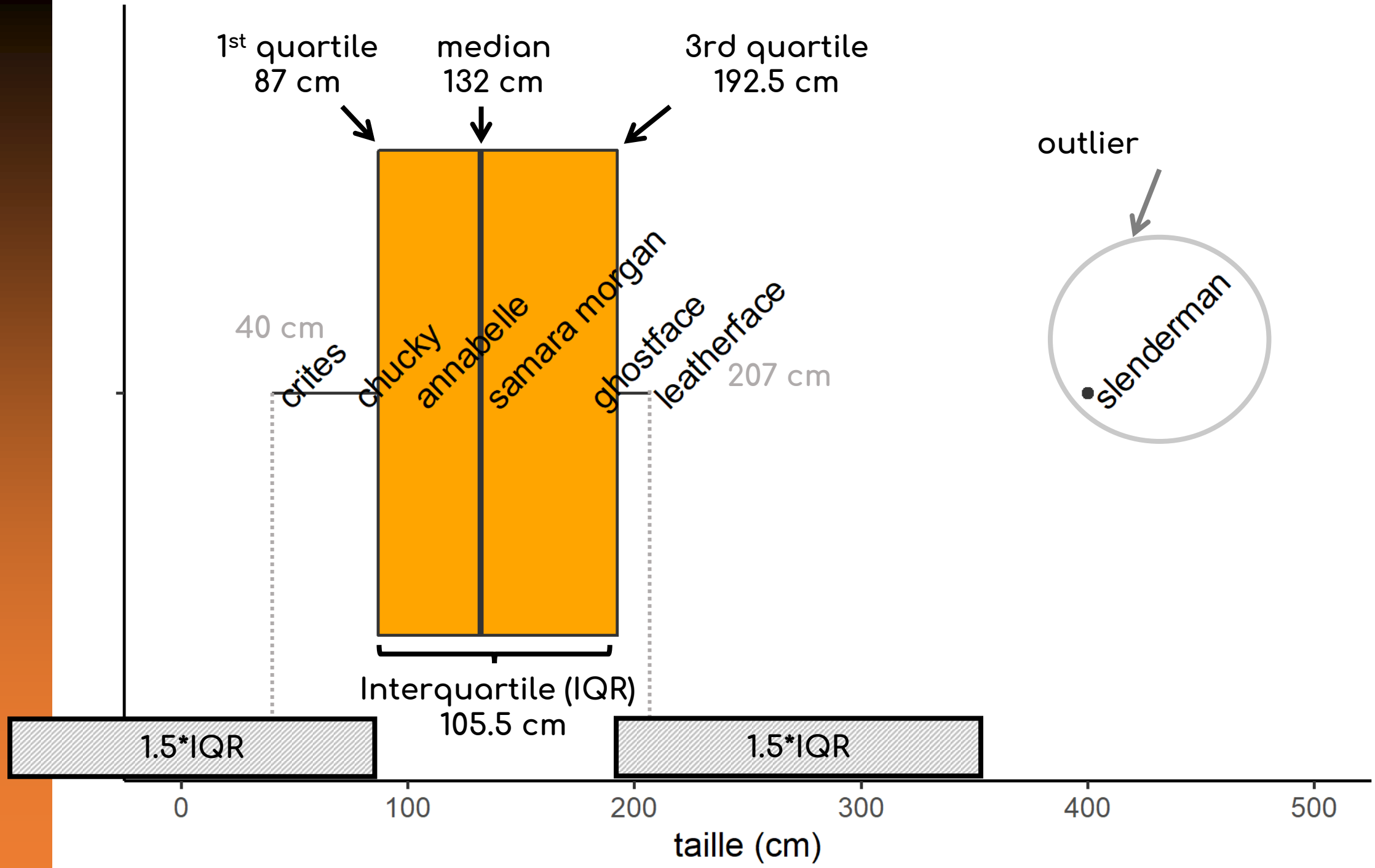
25% of characters smaller than the 1st quartile

25% of characters taller than the 3d quartile

1er quartile

median

3e quartile


1st quartile
87 cm
median
132 cm
3rd quartile
192.5 cm
outlier
slenderman
40 cm
crites
chucky
annabelle
samara morgan
ghostface
leatherface
207 cm
Interquartile (IQR)
105.5 cm
1.5*IQR
1.5*IQR
0
100
200
300
400
500
taille (cm)

Dispersion indicators

Variance: The average of the squares of the deviations from the mean.

$$s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$s^2 = \frac{1}{7}[(100 - 162)^2 + (132-162)^2 + (74-162)^2 + (178-162)^2 + (207-162)^2 + (400-162)^2 + (40-162)^2]$$

$$s^2 = 14383$$

Dispersion indicators

Standard deviation: The square root of the variance

$$s = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$$

$S = \sqrt{s^2} = \sqrt{14383} = 120$ cm

Meaning of standard deviation

- Standard deviation cannot be negative
- The higher it is, the greater the dispersion around the average.
- standard deviation = 0: total homogeneity.
- standard deviation ~ 0: values with very little dispersion around the mean.
- The further the values are from the mean, the greater the standard deviation/dispersion.

What is the impact of an outlier for mean and standard deviation ?

outliers

- **Without Slenderman**
- Mean : 121.83 cm
- Standard deviation : 57.70 cm

- **With Slenderman**
- Mean : 161.57 cm
- Standard deviation : 111.03 cm

Description of standard deviations

```
> print(averages)
  Average_Age Average_Iron_Level Average_Systolic_Tension Average_Diastolic_Tension
1      64.975           111.2325                    118.5                     79.75
```

```
> print(standard_deviations)
  Standard_Deviation_Age Standard_Deviation_Iron_Level Standard_Deviation_Systolic_Tension Standard_Deviation_Diastolic_Tension
1               9.531291                       32.25231                          9.562212                           5.19245
```

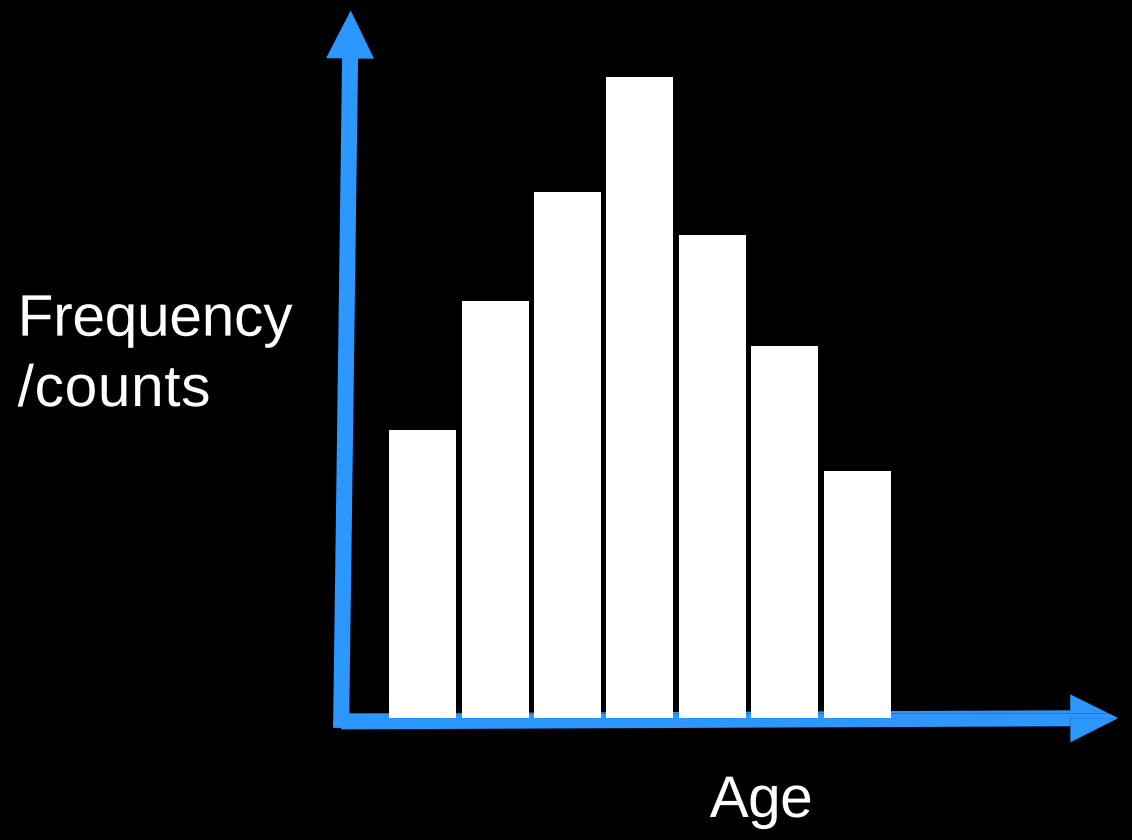
Not so young

[60 to 170 µg/dL].

[80 to 120 µg/dL]

< 80 µg/dL

Visualizing a quantitative variable: the histogram

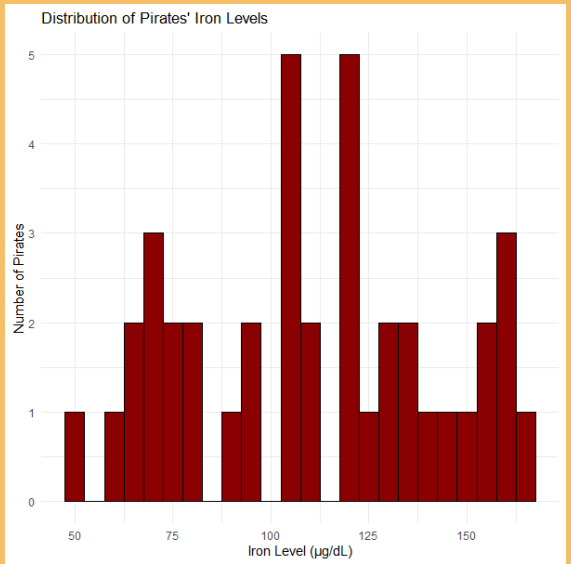


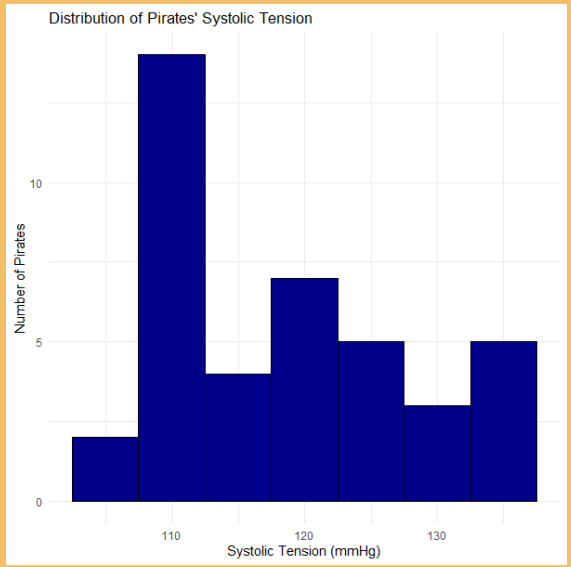
We visualize the distribution of a quantitative variable

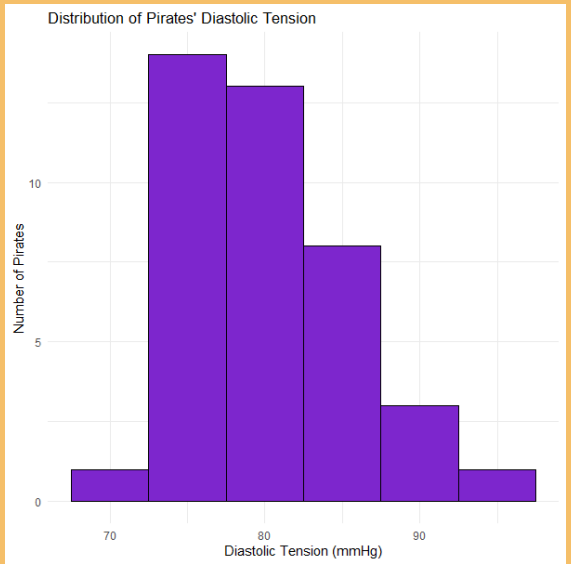
Visual of central tendency + dispersion (with symmetry)

Strengths:
Only tool for observing distribution of a quantitative variable

Histogram to describe the evolution of a quantitative variable







The captain, after analyzing the data, concludes melancholically, "Alright, I get it, my brave crew has aged and gained experience."

This realization, though difficult, leads to deeper reflections on managing an aging crew with care while keeping the ship ready for future adventures.

Qualitative variable

- Textual
  - Binary (2 modalities: M/F)
  - Nominal (several non-hierarchical modalities)
  - Ordinal if there is an order (Likert scale, growth stage), age classes)

What is
it's chirping!

The captain, exasperated by the morning commotion, exclaims, "Oh dear, there are nothing but parrots on this ship this morning! That must be what's tiring my sailors! What an incessant hubbub!

I'm convinced it's because we have too many females among them!" He suspects the large number of parrotlets is contributing to the particularly noisy atmosphere, impacting the well-being and rest of his crew.

Qualitative variable

1variable → univariate

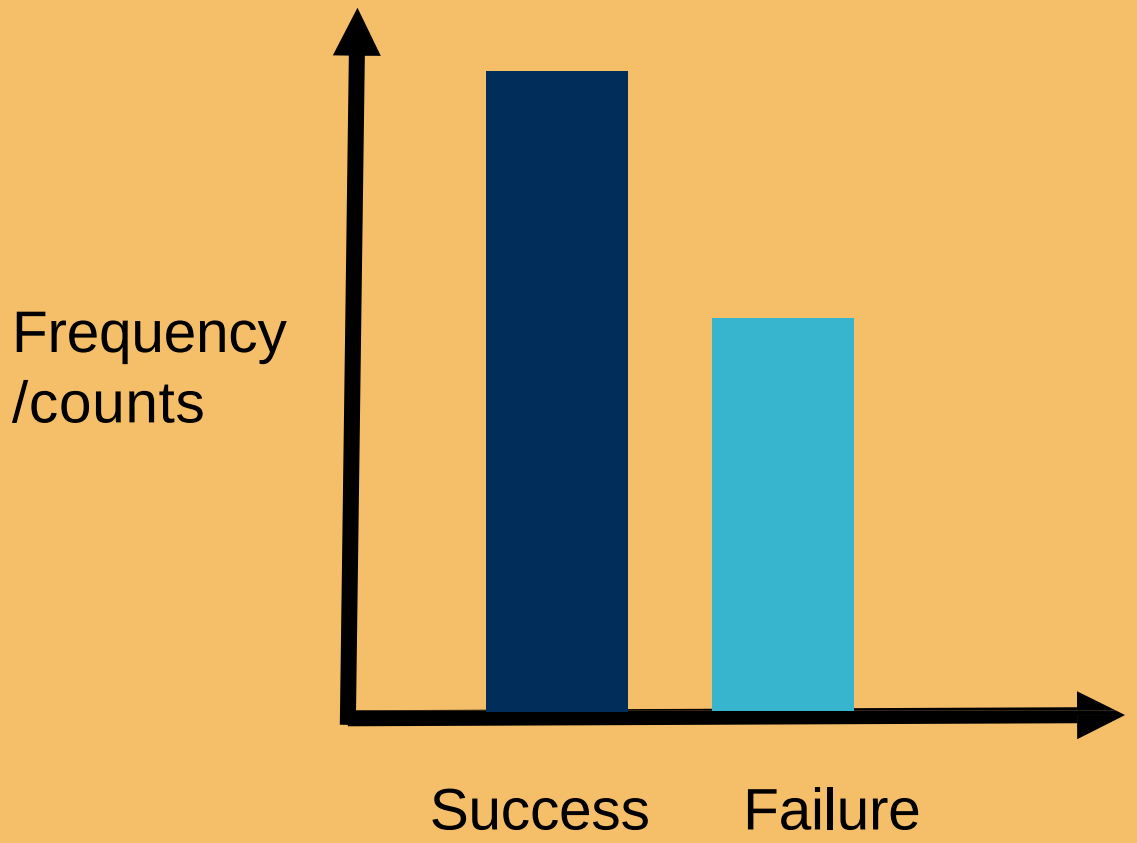
Qualitative variable → Chart: pie chart or bar chart

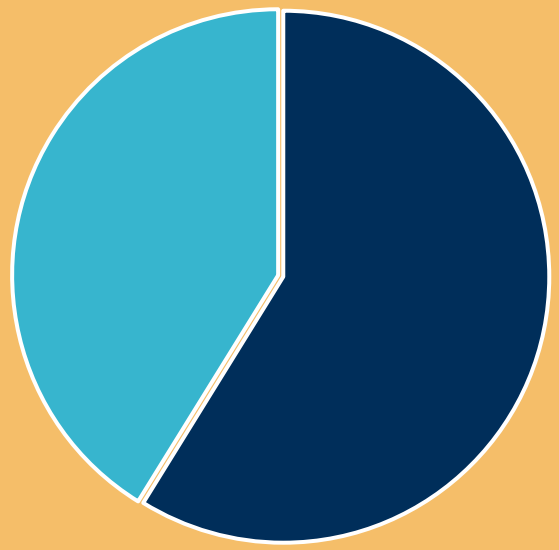
Qualitative variable → Table of frequencies, %ages

A little camembert?

Pie chart and bar graph

	Workforce	Frequency (relative)	percentage
Success	10	0.588	58.8
Failure	7	0.412	41.2



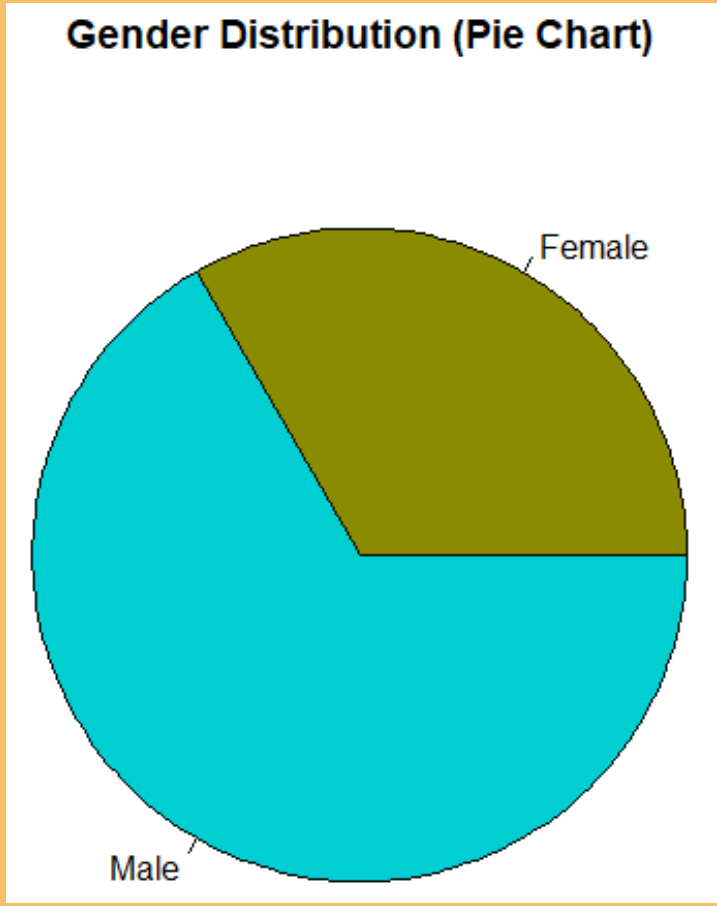


Charts / Camemberts / Bar graphs

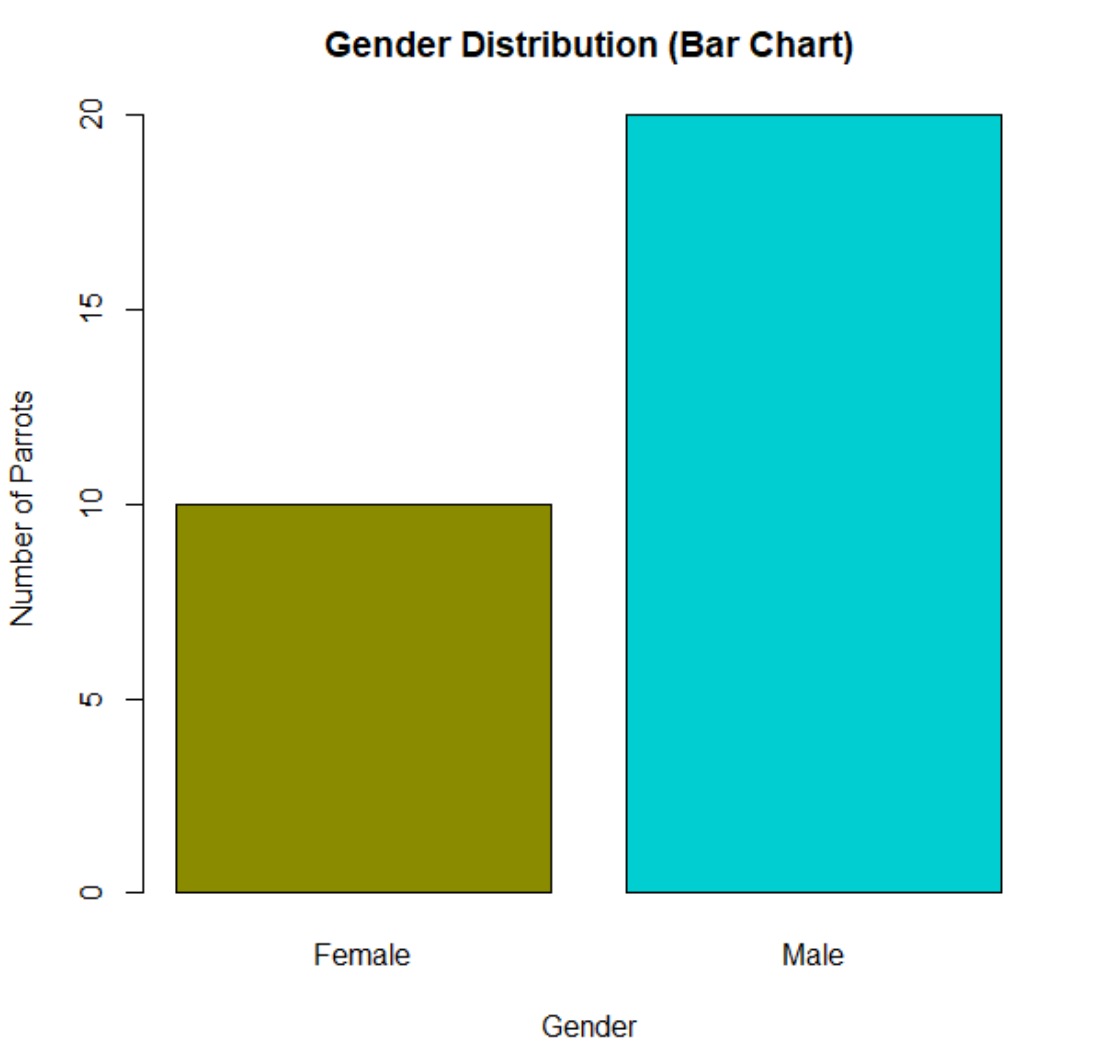
```
> gender_counts

Female   Male
    10     20
```

table



Pie chart



Bar graph

The captain, after a moment's reflection, admits with a hint of realism, "Okay, well, maybe female parrots aren't the problem after all..."

He realizes the source of his crew's fatigue may be more complex, prompting him to explore other hypotheses and better understand the true needs of his crew.

This marks a turning point in his approach. He will consider bivariate statistics. But This is another story.

Summary description of variables (univariate statistics)

	Metrics	graphs
Quanti	Means, standard deviations, quartiles	histogram
Quali	Table of frequencies/effectives/percentages	Pie chart/ bar charts

Statistics
descriptive, it's useless!

My own research example

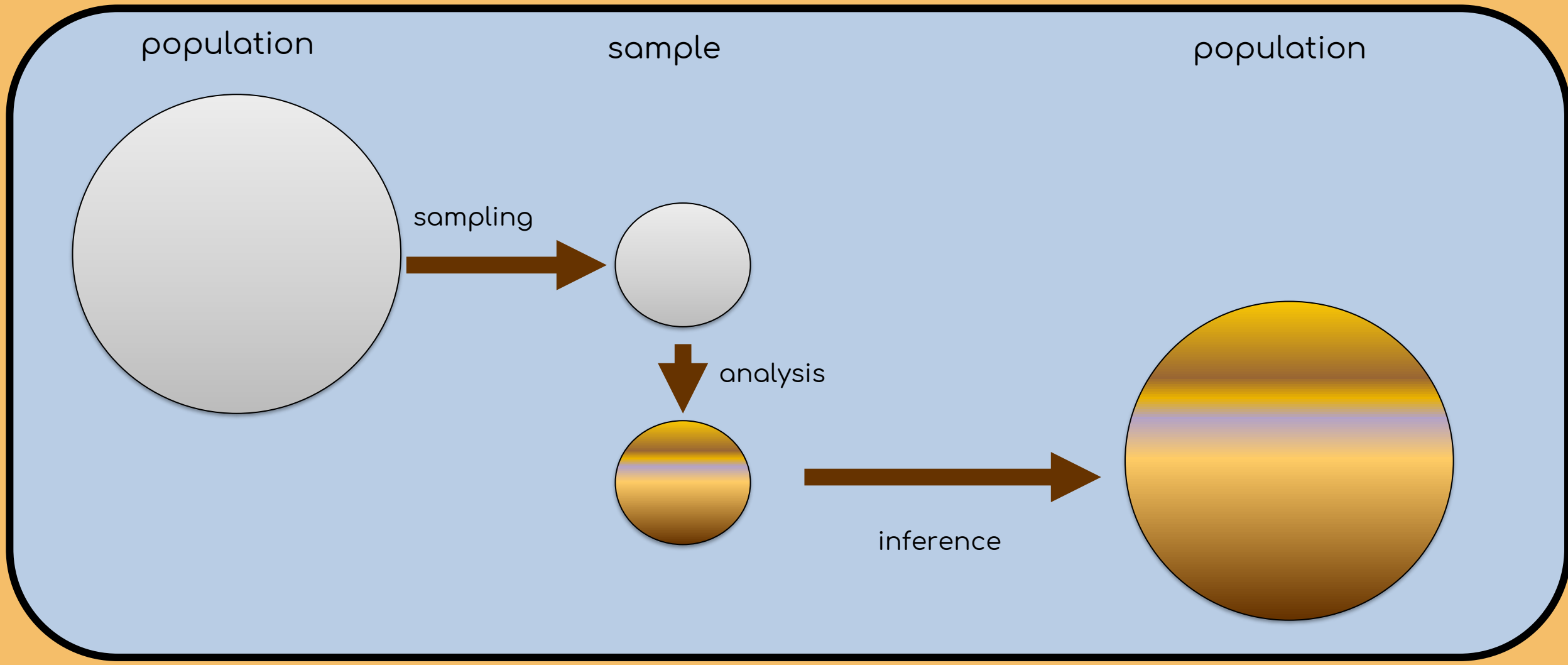
Bad data, bad conclusion

Rotten wood

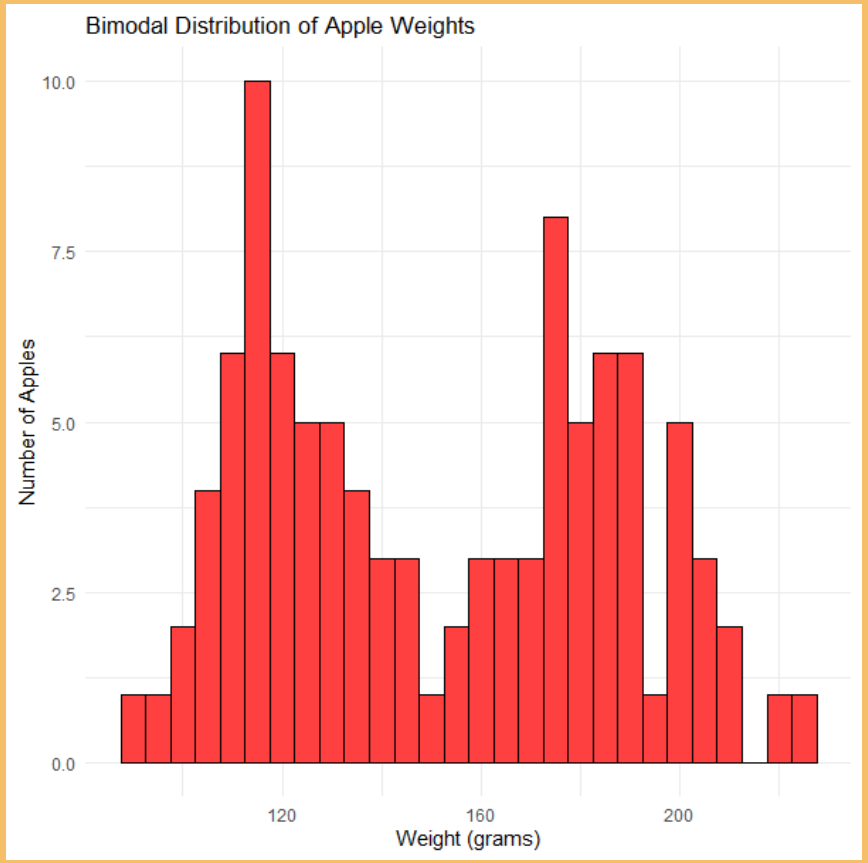
Rotten boat

In your field

Next course : caracterize a population based on a sample.



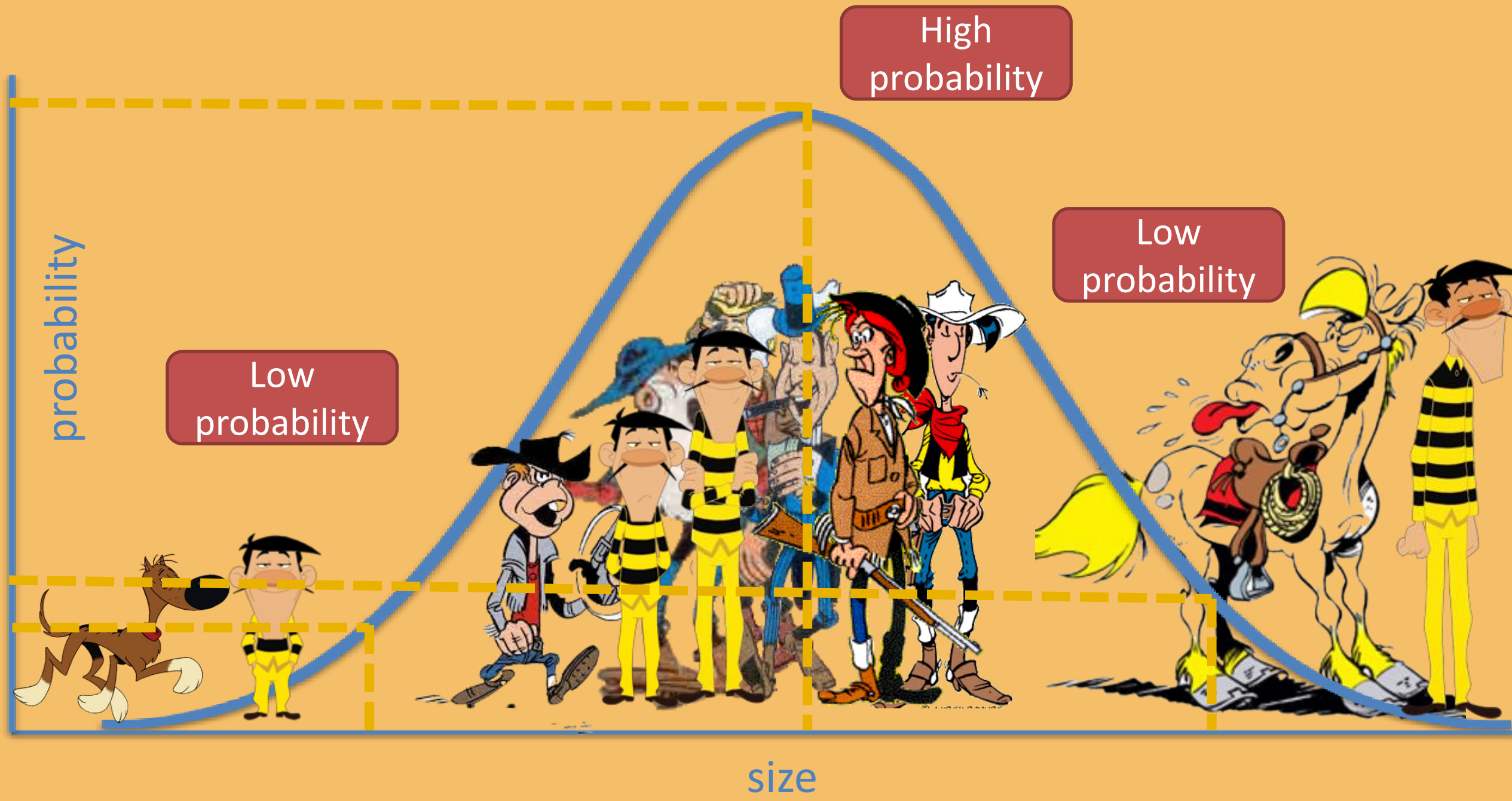
Calculation based on means.



Does calculating the mean make sense for this distribution?

Being normal,

what is it ?


High probability
Low probability
Low probability
probability
size

MEN
WOMEN

ROOMS

ALOON

Check normal distribution for a variable

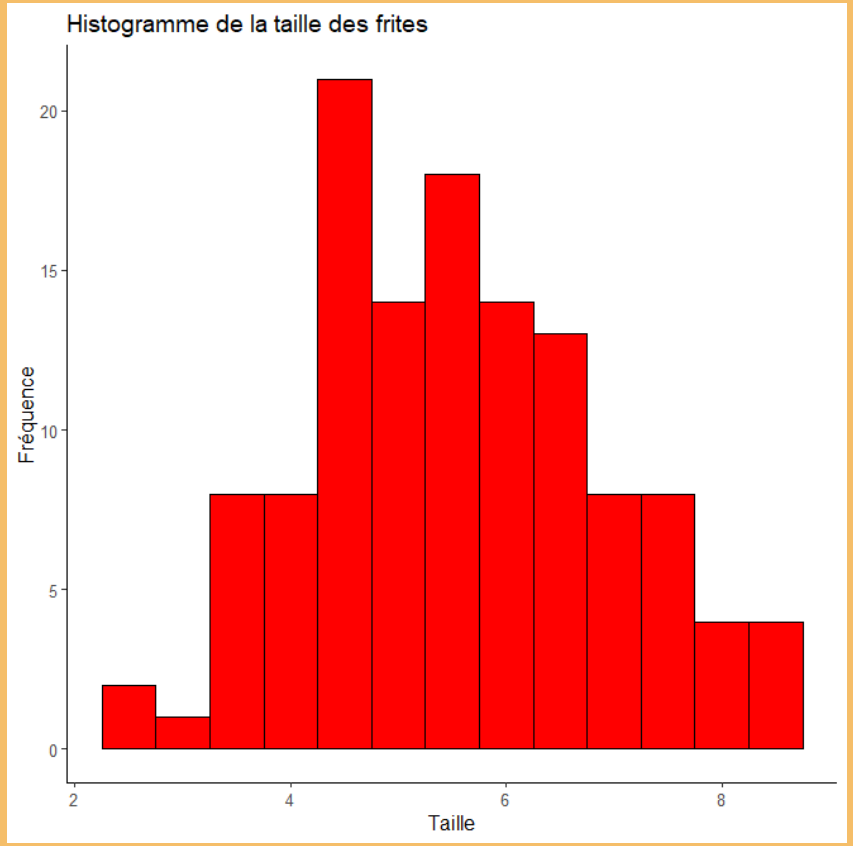
Will be usefull afterwards for several tests

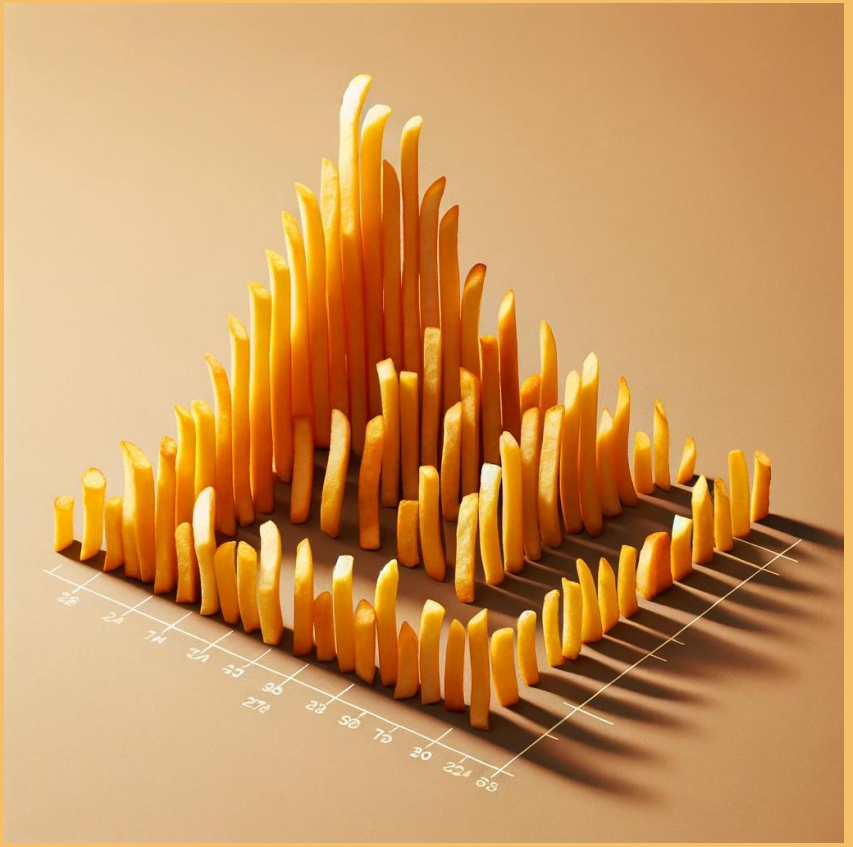
Do the sizes of french fries from north of France follow a normal distribution?

Do the sizes of french fries from north of at France follow a normal distribution?

size
4.2
7.5
4.6
7.8
3.9
6.9
5.6
4.0
7.3
5.9
8.2
2.6
4.6
4.7
7.2
5.9

1st possibility : plot an histogram

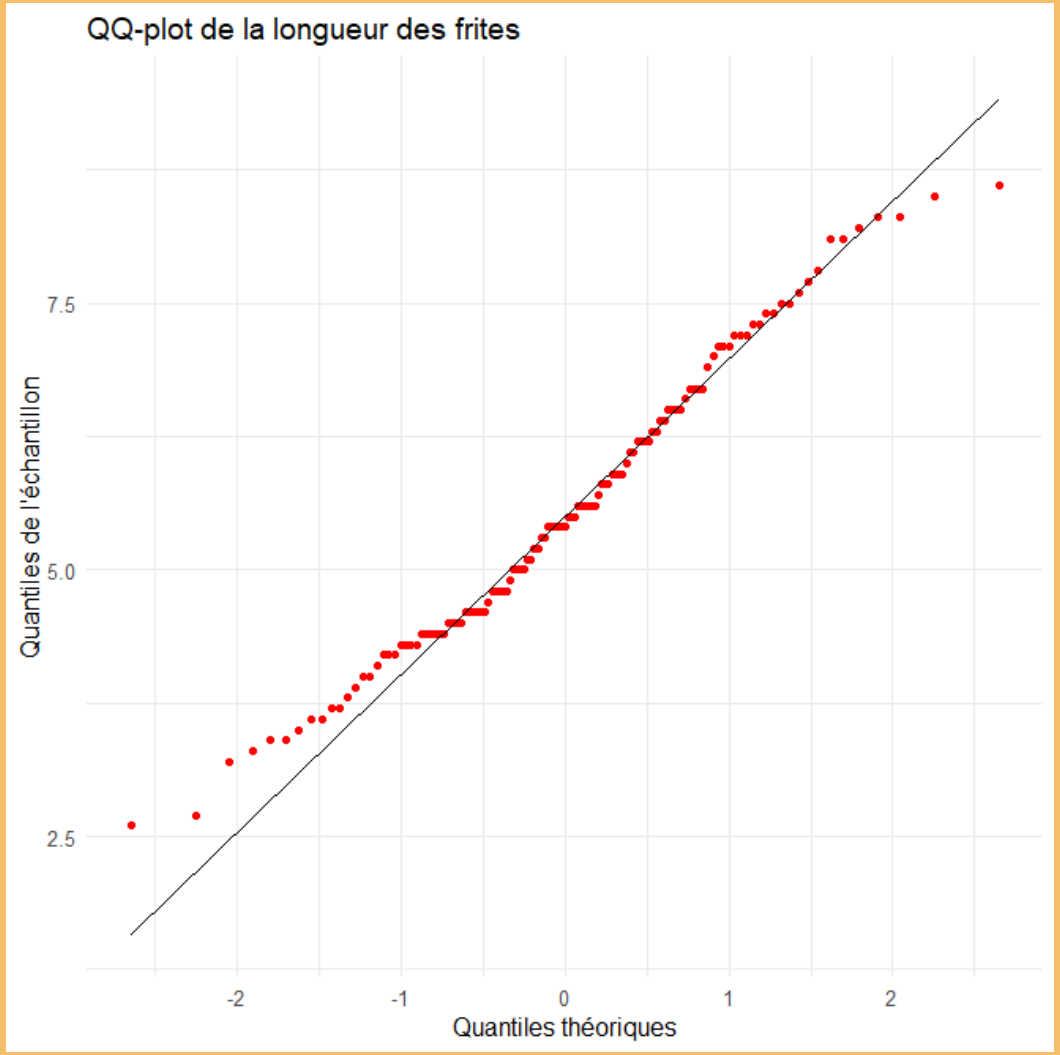




Do the sizes of french fries from north of at France follow a normal distribution?

2nd possibility :
Plot the qqplot

the line must be globally on the 1st bisector and ~80% of the points must be aligned



Do the sizes of french fries from north of at France follow a normal distribution?

3rd possibility : shapiro's test

```
> shapiro.test(data$size)

        Shapiro-Wilk normality test

data:  data$size
W = 0.98256, p-value = 0.1134
```

The p-value

~ is whether or not we believe in a difference

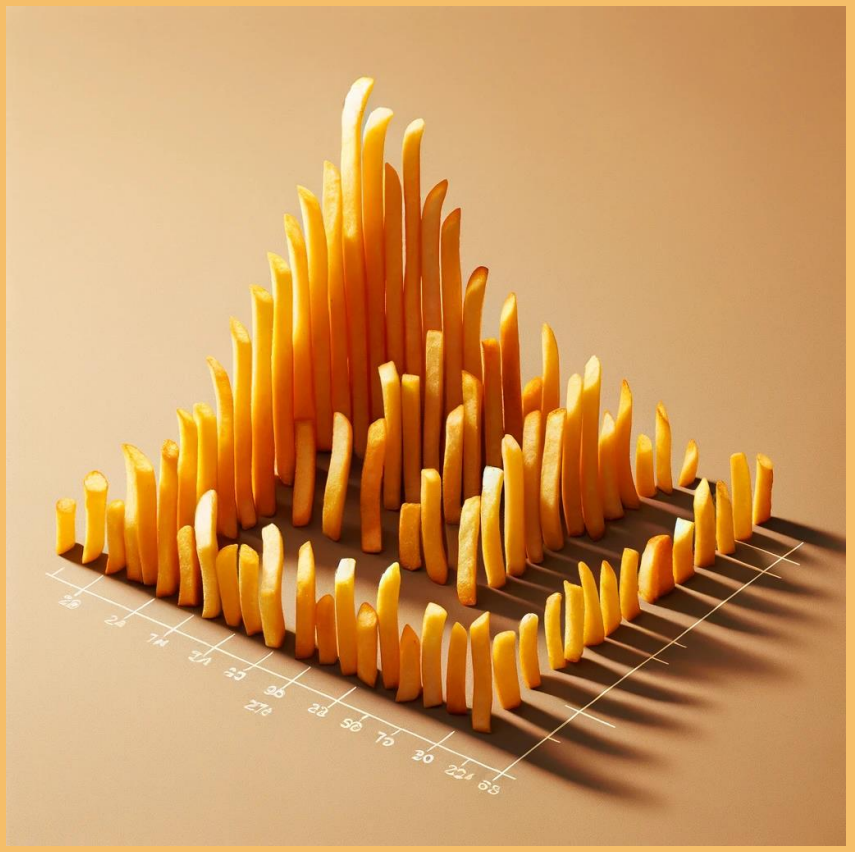
~ is the distribution of the variable different from a normal distribution?

Criteria:

- p-value < 0.05: we believe in the difference
- p-value > 0.05: we do not believe in the difference

French fries from north of France follow a normal distribution !

Could we have expected that result ?

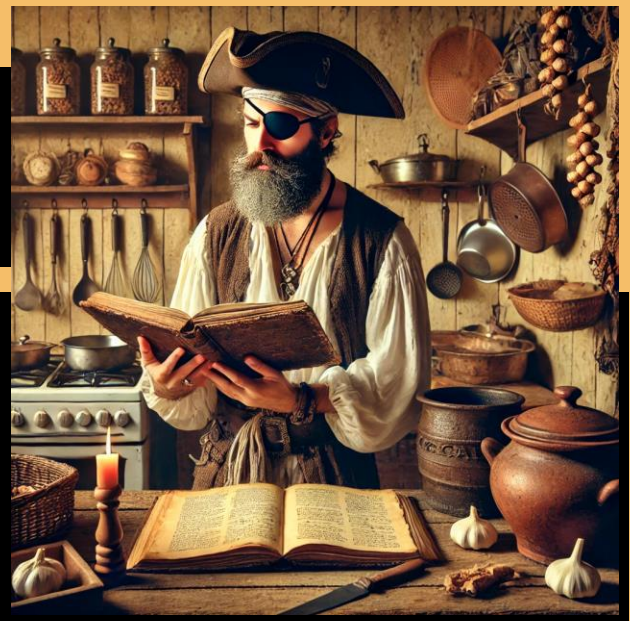


Introduction to R

What is R ?

- R is a free, open-source programming language and software environment for statistical computing and graphics.
- Powerful statistical analysis toolkit.
- Extensive graphical capabilities.
- Highly extensible through packages.

What is a package?



- Package ~ recipe book.

- Download a package :
  ‘order and wait for delivery’

```
Install.packages('package_name')
```

Once

- Load a package :
  ‘open the book to cook’

```
library('package_name')
```

Each time you use it

Install R and R studio

R : software

R Studio : an integrated development environment (IDE) that enhances the use of the R (fancy)

Rcmdr ('Rcommander') : a package allowing click on a button with a drop-down menu

Quartz

Wanna learn about coding in R while not comfortable with coding ?

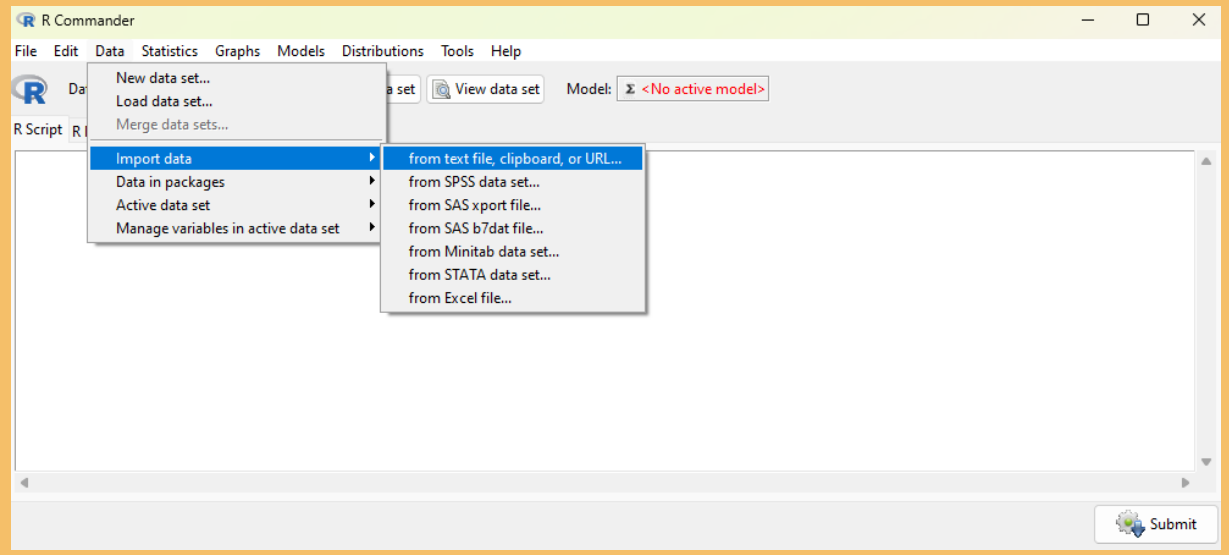
Note that in the Rcmdr window, the R code is displayed.

This will enable you to adapt the code when you'll feel comfortable.

2 teams for this tutorial

CLICK/PRESS BUTTON TEAM

CODING TEAM




```
# Fitting linear model ------------------------------------------------
mod <- lm(speed ~ height_m + attack + type1 + percentage_male + base_happiness, data = data)
data2 <- data[-85, ]  # Excluding the 85th row for robustness check
mod_updated <- lm(speed ~ height_m + attack + type1 + percentage_male + base_happiness, data

# Model diagnostics ---------------------------------------------------
Anova(mod_updated)
summary(mod_updated)

# Checking for collinearity -------------------------------------------
vif(mod_updated)  # Variance inflation factors

# Post hoc analysis ---------------------------------------------------
tcm <- emmeans(mod_updated, pairwise ~ type1)
print(tcm)

# Base R functions for scatter plots with points and lines -------------
boxplot(data$speed ~ data$type1, xlab = "Type", ylab = "Speed", main = "Speed by Type")

plot(data$speed ~ data$attack, xlab = "Attack", ylab = "Speed", main = "Speed vs. Attack")
```

Learning R is copy pasting !

YES !

R very basics

dataset$column

function_name(colum1 ~ colum2, data= dataset_name)

Avoid spaces in your columns' names and special features in your dataset