SH Dr. Soufiane Hourri

drsoufiane.com

DUT · SEMESTER 3 · COMPUTER SCIENCE

# Machine Learning

Session 01 · What machine learning is, and the statistics it needs

**Dr. Soufiane Hourri, PhD, HDR**

Associate Professor & AI Consultant

École Supérieure de Technologie de Safi · Cadi Ayyad University

**2026-2027**

# Today's session

01 What machine learning is

02 Data and variables

03 The center of a variable

04 Spread and outliers

05 The shape of a distribution

06 Two variables together

07 Lab: describing a real dataset in Python

**TODAY'S FOCUS**

A model learns from data. Before we train models, we learn to describe data with a few numbers and charts.

No statistics background is needed. Bring a calculator.

# Learning objectives

**By the end of this session, you will be able to**

1. Explain how machine learning differs from hand-written rules
2. Name the features, target and problem type of a dataset
3. Tell numerical from categorical variables
4. Compute mean, median, standard deviation and quartiles, by hand and with pandas
5. Read a histogram, a box plot and a scatter plot

**PREREQUISITES**

- Python basics: variables, lists, functions
- Secondary-school algebra: sums, squares, square roots
- No statistics course needed

# Where this session fits in the module

**01**

### Foundations

What ML is, and the statistics to describe data (today)

**02**

### Preparing data

Missing values, categories, scaling

**03**

### Regression

Predicting a number with linear models

**04**

### Classifying

Logistic regression and its metrics

**05**

### Decision trees

Splitting data into purer groups

# Why a machine learning course starts with statistics

- A model learns from data, so we first need to describe data
- Many machine learning tools are built on simple statistics
  - Scaling a feature: subtract the mean, divide by the standard deviation
  - Measuring a model's error: an average of its mistakes
  - Filling a missing value: often the median of the column

**KEY TAKEAWAY**

Today's statistics are the vocabulary of the rest of the module.

01

CHAPTER

# What machine learning is

How can a program improve from examples instead of rules written by a programmer?

# Machine learning: rules learned from examples

DEFINITION

**Machine learning lets a program find patterns in example data and use them to predict or decide on new data, without a programmer writing every rule.**

EXAMPLE

A spam filter is trained on thousands of emails already marked spam or not spam. It learns which words, links and senders signal spam.

CLASSICAL PROGRAMMING

rules + data → answers

MACHINE LEARNING

data + answers → rules (a model)

# When machine learning is a good choice

## Good fit

- The pattern is too complex for hand-written rules (faces, speech, prices)
- Many past examples with known answers exist
- Some errors are acceptable
- The pattern changes, and the model can be retrained

## Poor fit

- A simple exact rule already works (VAT, a grade average)
- There is little or no data
- A wrong answer is unacceptable and the correct rule is known

# AI, machine learning and deep learning

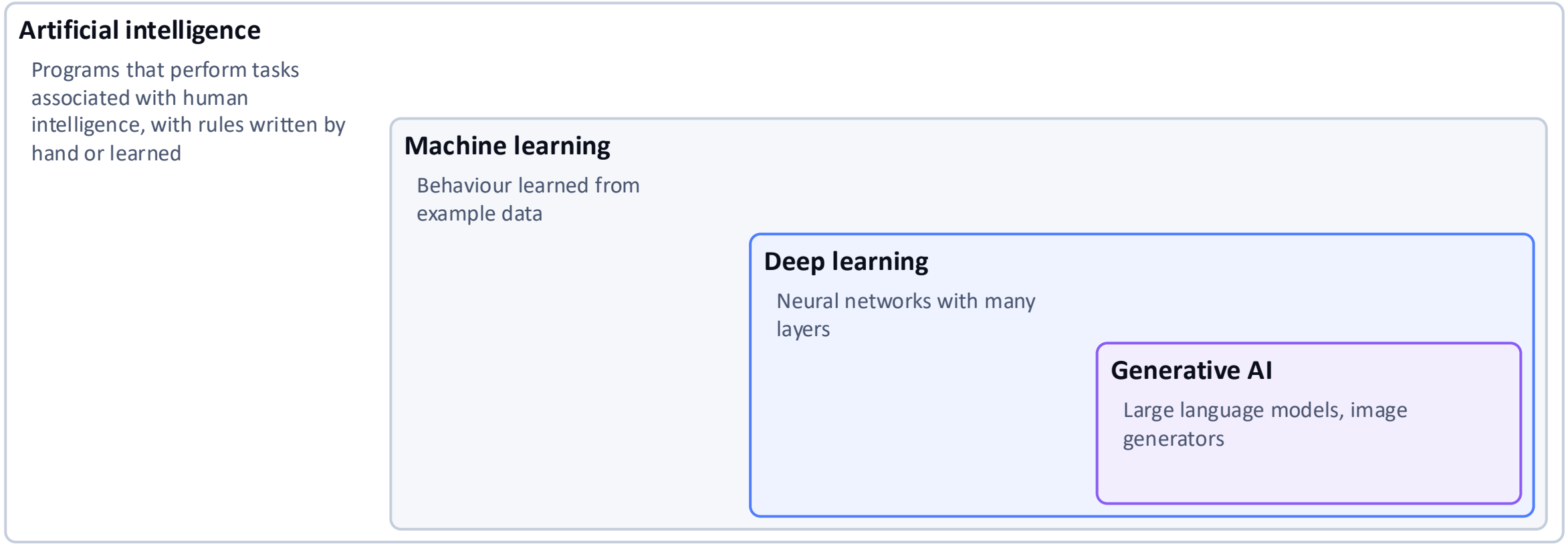



Figure 1. Each field is a part of the one around it. This module covers machine learning on tables of data.

# A real dataset: the Palmer penguins

| species | island | bill_length_mm | bill_depth_mm | flipper_length_mm | body_mass_g | sex |
|---|---|---|---|---|---|---|
| Adelie | Torgersen | 39.1 | 18.7 | 181 | 3750 | Male |
| Adelie | Torgersen | 39.5 | 17.4 | 186 | 3800 | Female |
| Adelie | Torgersen | 40.3 | 18 | 195 | 3250 | Female |
| Adelie | Torgersen | NaN | NaN | NaN | NaN | NaN |
| Adelie | Torgersen | 36.7 | 19.3 | 193 | 3450 | Female |
| Adelie | Torgersen | 39.3 | 20.6 | 190 | 3650 | Male |

Table 1. First 6 of 344 rows. NaN = missing value. Data: Gorman, Williams & Fraser (2014), via the palmerpenguins package (Horst et al., 2020).

# Observations, features and target

DEFINITION

**Each row is an observation (one penguin). Each column is a variable. The variables used to predict are features; the variable we want to predict is the target.**

EXAMPLE

Predict body mass from flipper length and species. Features: flipper_length_mm, species. Target: body_mass_g.

NOTATION

$n$ = number of observations (here 344)

$x$ = features, the inputs

$y$ = target, the value to predict

# Three families of machine learning problems

## Regression

**TARGET**

A number

**QUESTION**

How much? How many?

**PENGUINS**

Predict body mass in grams

**ELSEWHERE**

Price of an apartment in Marrakech

## Classification

**TARGET**

A category

**QUESTION**

Which class?

**PENGUINS**

Predict the species

**ELSEWHERE**

Spam or not spam

## Clustering

**TARGET**

No target

**QUESTION**

Which groups exist?

**PENGUINS**

Group penguins without their species

**ELSEWHERE**

Customer segments in a shop

Regression and classification are supervised learning (the target is known in the training data). Clustering is unsupervised.

# Knowledge check

QUESTION

## Which task is a regression problem?

A

Predicting tomorrow's maximum temperature in Safi (°C)

B

Deciding whether an email is spam or not

C

Grouping news articles by topic, with no labels

D

Recognising a handwritten digit (0 to 9) in an image

# The machine learning workflow

**01**

### Define the question

What to predict, and why it is useful

**02**

### Explore the data

Describe it with statistics and charts (today)

**03**

### Prepare the data

Clean, encode and scale the features

**04**

### Train and evaluate

Fit on training data, test on unseen data

**05**

### Use and monitor

Predict new cases, check quality over time

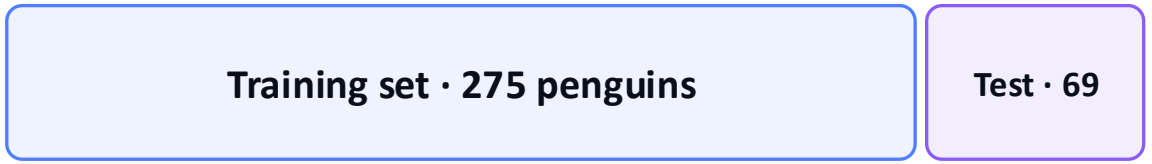
# Training data and test data

DEFINITION

**The model learns on one part of the data, the training set (often about 80%). Its quality is measured on the rest, the test set, which it never sees during training.**

EXAMPLE

An exam with exercises you memorised proves nothing. New questions show what you really learned. The test set plays the role of the new questions.




344 penguins split 80/20 at random. The test rows stay hidden until the final evaluation.

02

CHAPTER

# Data and variables

Before computing anything, decide what kind of variable each column is.

# Population and sample

## Population

- Everything we want to know about: all adult penguins of these three species in the Palmer Archipelago
- Its true values are parameters, e.g. the true mean mass $\mu$. Usually unknown.

## Sample

- The part we actually measured: 344 penguins
- Values computed from it are statistics, e.g. the sample mean $\bar{x}$. We use them to estimate the parameters.

# Types of variables

| Family | Type | What it is | Penguins | Other examples |
|---|---|---|---|---|
| Numerical | Continuous | Measured; any value in a range | `body_mass_g`, `flipper_length_mm` | Height, temperature |
| Numerical | Discrete | Counted; whole numbers | None in this dataset | Number of absences, children |
| Categorical | Nominal | Groups with no order | `species`, `island`, `sex` | City, blood type |
| Categorical | Ordinal | Groups with a natural order | None in this dataset | Bac mention, satisfaction level |

Table 2. Numerical variables can be averaged; categorical variables can only be counted (and sorted, if ordinal).

# A number is not always a numerical variable

✗ INCORRECT APPROACH

Postal codes 46000 (Safi) and 40000 (Marrakech) are numbers, so we compute their mean: 43000

✓ CORRECTION

Treat the postal code as categorical: count the students in each city, report the most frequent one

WHY IT FAILS

A code is a name written with digits. Ask: does the difference between two values mean something? If not, the variable is categorical.

# Exercise · Classify the variables

GOAL

For each variable, decide: numerical (continuous or discrete) or categorical (nominal or ordinal).

TIME

**5 min**

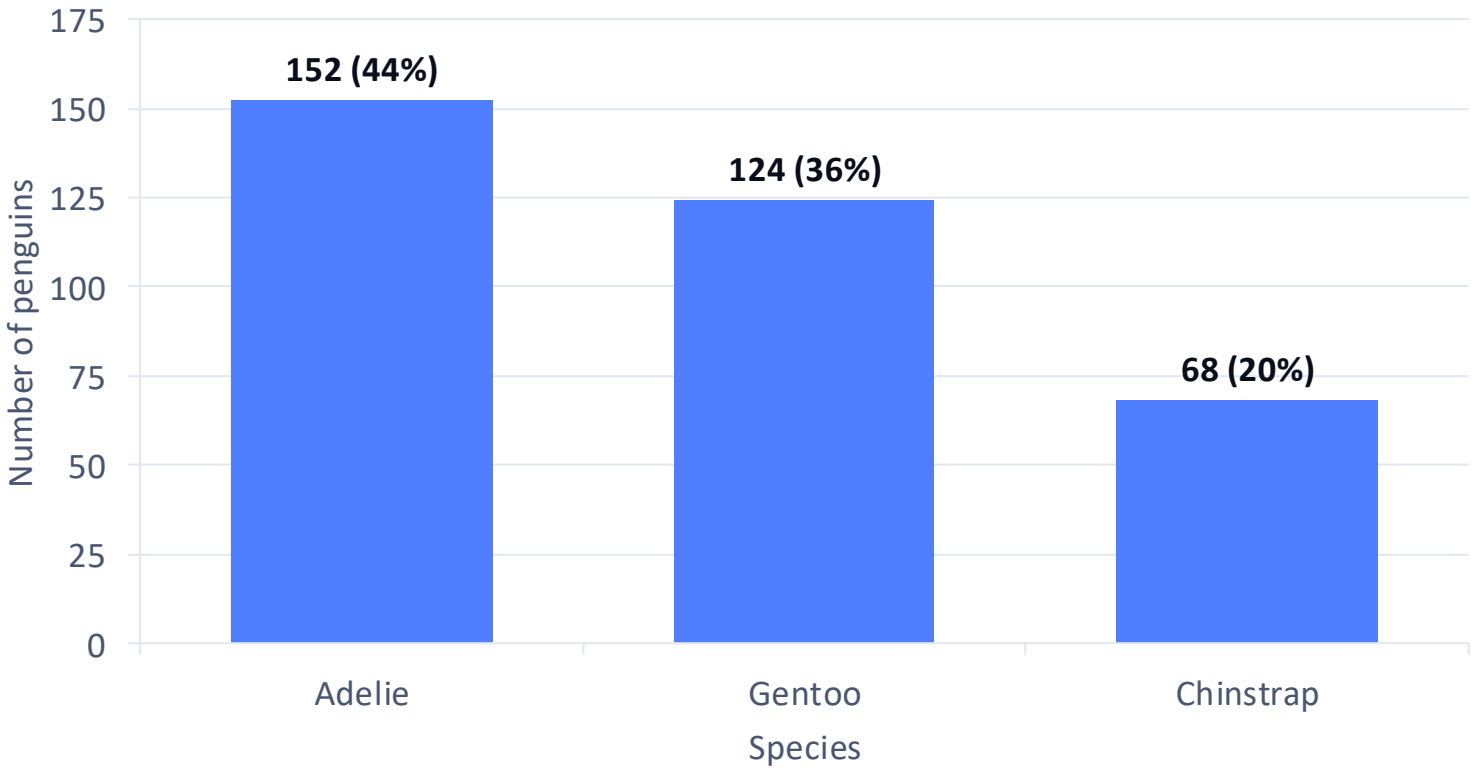
VARIABLES

1. Student ID number (CNE)
2. Number of absences in a semester
3. Bac mention (Passable, Assez bien, Bien, Très bien)
4. Mobile data used per month (GB)
5. City of origin
6. Final grade out of 20

DELIVERABLE

Your six answers, each with one word of justification. We correct together.

# Describing a categorical variable: counts




Source: Palmer penguins (n = 344), counts computed with pandas value_counts()

MAIN FINDING

Adelie is the most frequent species: it is the mode.

For a categorical variable we report counts and percentages. A mean of species makes no sense.

# 03

CHAPTER

# The center of a variable

One number that represents a typical value: mean, median and mode.

# The mean

$$\bar{x} = (x_1 + x_2 + \ldots + x_n) / n = (1/n) \Sigma_i x_i$$

## SYMBOLS

$n$ : number of observations

$x_i$ : value of observation i (i = 1 … n)

$\Sigma$ : "add up" all the terms

$\bar{x}$ : the mean, read "x bar"

## INTUITION

Add all the values, then share the total equally among the n observations.

Example: 4, 6, 11 → 21 / 3 = 7. Same unit as the data.

# The median and the mode

DEFINITION

**The median is the middle value once the data is sorted: half of the observations are below it, half above. The mode is the most frequent value.**

EXAMPLE

3, 5, 8, 10, 14 (n odd) → median 8

3, 5, 8, 10 (n even) → (5 + 8) / 2 = 6.5

RULE

1. Sort the values
2. Take the value at rank $(n + 1) / 2$
3. If that rank falls between two values, take their mean

# Mean, median and mode by hand

PROBLEM

Seven students give their commute time to school in minutes (example data): 25, 12, 20, 30, 15, 20, 18. Find the mean, the median and the mode.

REASONING

1. Sum = 140, n = 7, so the mean = 140 / 7 = 20
2. Sorted: 12, 15, 18, 20, 20, 25, 30. Rank (7 + 1) / 2 = 4: the 4th value is 20
3. 20 appears twice, every other value once

RESULT

Mean = 20 min

Median = 20 min

Mode = 20 min

# One extreme value moves the mean, not the median

PROBLEM

The student with 30 minutes is replaced by a student who travels 100 minutes by bus. Recompute the mean and the median.

REASONING

1. New sum = 140 − 30 + 100 = 210, so the mean = 210 / 7 = 30
2. Sorted: 12, 15, 18, 20, 20, 25, 100. The 4th value is still 20
3. One value changed; the mean moved by 50%, the median did not move

RESULT

Mean: 20 → 30 min

Median: 20 → 20 min

The median is robust to outliers

# Mean or median?

- Mean: uses every value; best when data is roughly symmetric, without extreme values
- Median: robust; best with outliers or a long tail (salaries, prices, waiting times)
- Report both: a large gap between them is a warning sign
  - Penguin body mass: mean 4,202 g, median 4,050 g
  - The mean is higher: a group of heavy penguins pulls it up

**KEY TAKEAWAY**

When the mean and the median disagree, look at the data before choosing.

# Knowledge check

QUESTION

## A small company pays 9 employees 6,000 MAD per month and its director 60,000 MAD. Which statement is correct?

**A**

The mean and the median are both 6,000 MAD

**B**

The mean is 11,400 MAD and the median is 6,000 MAD

**C**

The mean is 6,000 MAD and the median is 11,400 MAD

**D**

The mean is 33,000 MAD

# The center in Python

CODE · PYTHON

```python
import seaborn as sns
df = sns.load_dataset("penguins")

mass = df["body_mass_g"]
print(mass.mean())
print(mass.median())
print(df["species"].mode()[0])
```

OUTPUT

```
4201.754385964912
4050.0
Adelie
```

Missing values are skipped: the mean uses 342 penguins.

# Break · 15 minutes

After the break: how spread out are the values?

04

CHAPTER

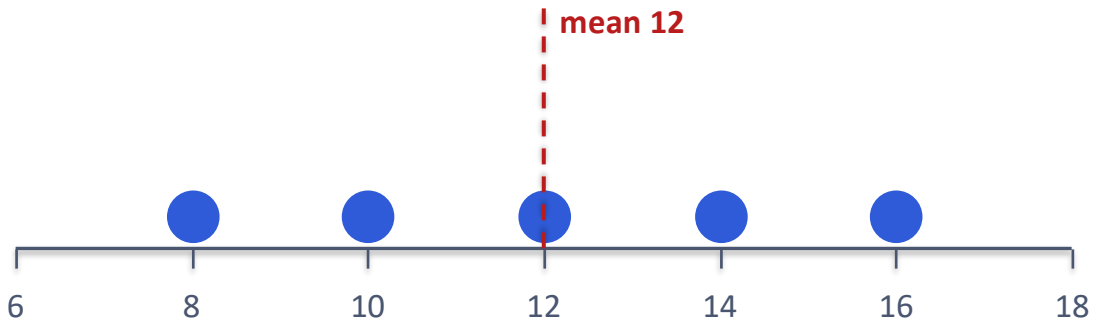
# Spread and outliers

Two groups can share the same mean and still be very different.

# Same mean, different spread

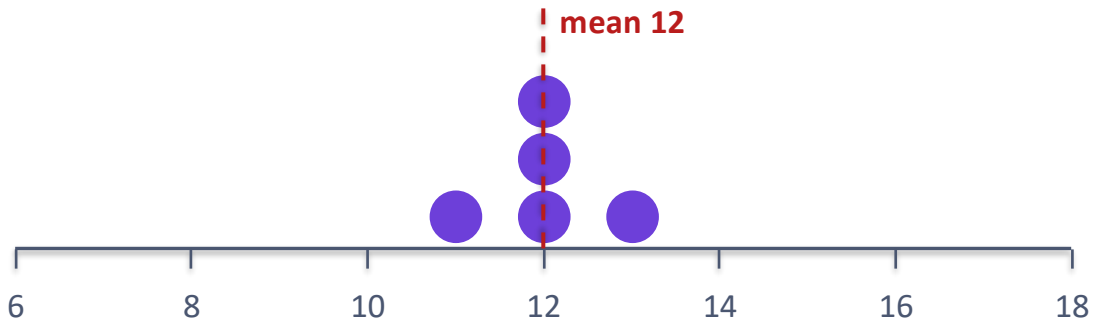
**Group A: 8, 10, 12, 14, 16**

- Mean = 12
- Grades from 8 to 16: students are very different




**Group B: 11, 12, 12, 12, 13**

- Mean = 12
- Grades from 11 to 13: students are similar



# The range

DEFINITION

**Range = maximum − minimum. It is the distance covered by the data.**

EXAMPLE

Group A: 16 − 8 = 8 points

Group B: 13 − 11 = 2 points

Penguin body mass: 6,300 − 2,700 = 3,600 g

LIMITATION

It uses only two values. One outlier changes it completely: the commute times go from a range of 18 to 88 minutes when 30 becomes 100.

# Variance and standard deviation

$$s^2 = \Sigma_i (x_i - \bar{x})^2 / (n - 1) \qquad s = \sqrt{(s^2)}$$

## SYMBOLS

$x_i - \bar{x}$: deviation from the mean

$s^2$: variance (squared units)

$s$: standard deviation (data units)

$n - 1$: divisor for a sample

## INTUITION

Measure how far each value is from the mean, square, average, then take the square root.

A typical value lies about s away from the mean.

# Standard deviation step by step: Group A

| $x_i$ (grade) | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---|---|---|
| 8 | −4 | 16 |
| 10 | −2 | 4 |
| 12 | 0 | 0 |
| 14 | 2 | 4 |
| 16 | 4 | 16 |
| **Sum: 60** | **Sum: 0** | **Sum: 40** |

$\bar{x} = 60 / 5 = 12$. $s^2 = 40 / (5 - 1) = 10$ points$^2$. $s = \sqrt{10} \approx 3.16$ points.

# Your turn: Group B

QUESTION

## Group B: 11, 12, 12, 12, 13 (mean 12). What is its standard deviation s (with n − 1)?

A

0.5

B

About 0.71

C

1

D

2

# Two libraries, two standard deviations

✗ INCORRECT APPROACH

`np.std(x)` → 2.83

`pd.Series(x).std()` → 3.16

"One of them must be wrong"

✓ CORRECTION

NumPy divides by n (ddof=0), pandas by n − 1 (ddof=1). State the divisor:

`np.std(x, ddof=1)` → 3.16

WHY IT MATTERS

Dividing by n − 1 corrects a slight underestimate when the data is a sample, which is the usual case in machine learning. For large n the difference is negligible.

# Quartiles and the interquartile range

DEFINITION

**Quartiles cut the sorted data into four parts of about 25% each: Q1, Q2 (the median), Q3. The interquartile range IQR = Q3 − Q1 contains the middle 50% of the data.**

EXAMPLE

Penguin body mass: Q1 = 3,550 g, median = 4,050 g, Q3 = 4,750 g

IQR = 4,750 − 3,550 = 1,200 g

FIVE-NUMBER SUMMARY

min · Q1 · median · Q3 · max

**2,700 · 3,550 · 4,050 · 4,750 · 6,300 g**

# Detecting an outlier with the 1.5 × IQR rule

**PROBLEM**

Weekly self-study hours of nine students (example data): 4, 7, 8, 9, 10, 11, 12, 14, 30. Is any value an outlier?

**REASONING**

1. n = 9, already sorted. Median = 5th value = 10
2. Q1 = 3rd value = 8; Q3 = 7th value = 12; IQR = 4
3. Lower fence: 8 − 1.5 × 4 = 2.  Upper fence: 12 + 1.5 × 4 = 18

**RESULT**

30 > 18: 30 is flagged as an outlier

Check it before removing it: typing error or real student?

# Reading a box plot

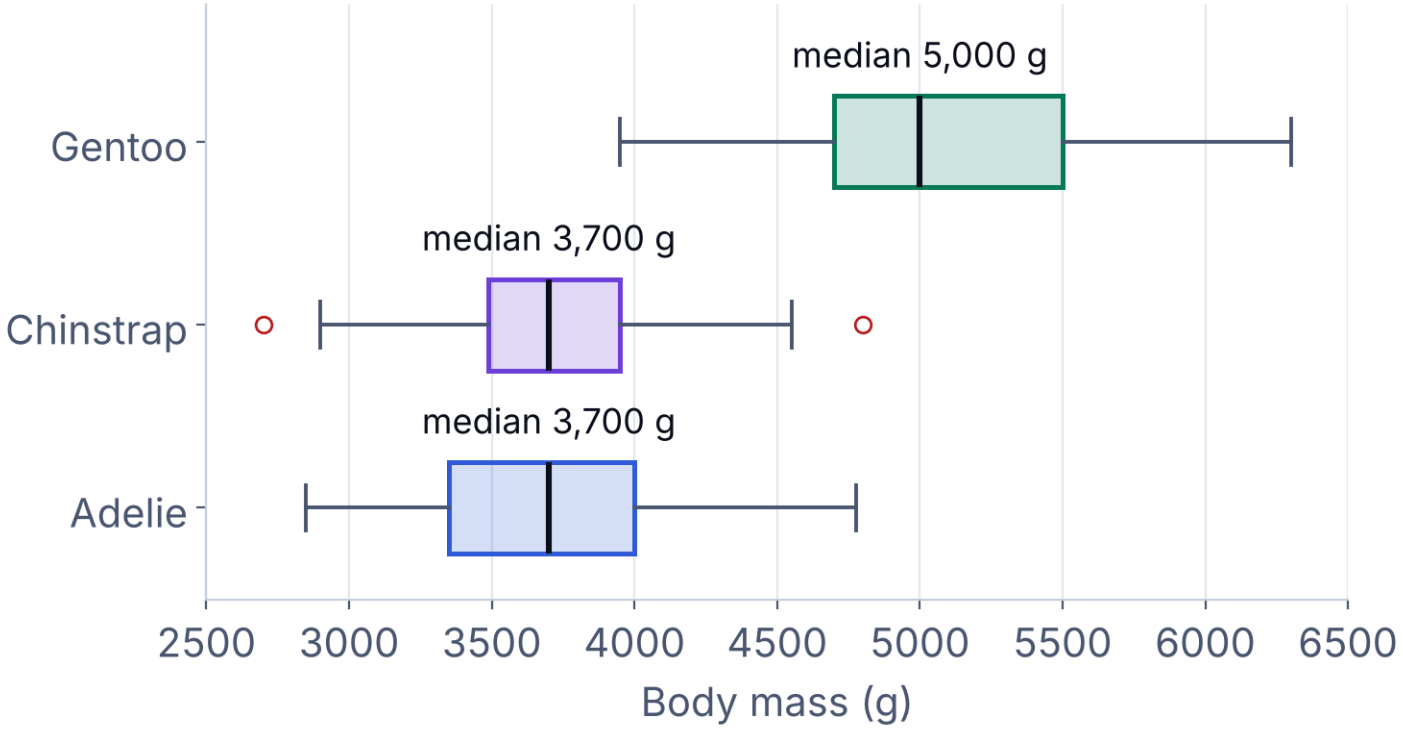



Figure 2. Body mass by species (n = 342). Data: Palmer penguins; plot made with matplotlib.

**WHAT TO NOTICE**

- Box: from Q1 to Q3, the middle 50%
- Line in the box: the median
- Whiskers: furthest points within 1.5 × IQR
- Circles: values flagged as outliers
- Gentoo are clearly heavier; Adelie and Chinstrap overlap

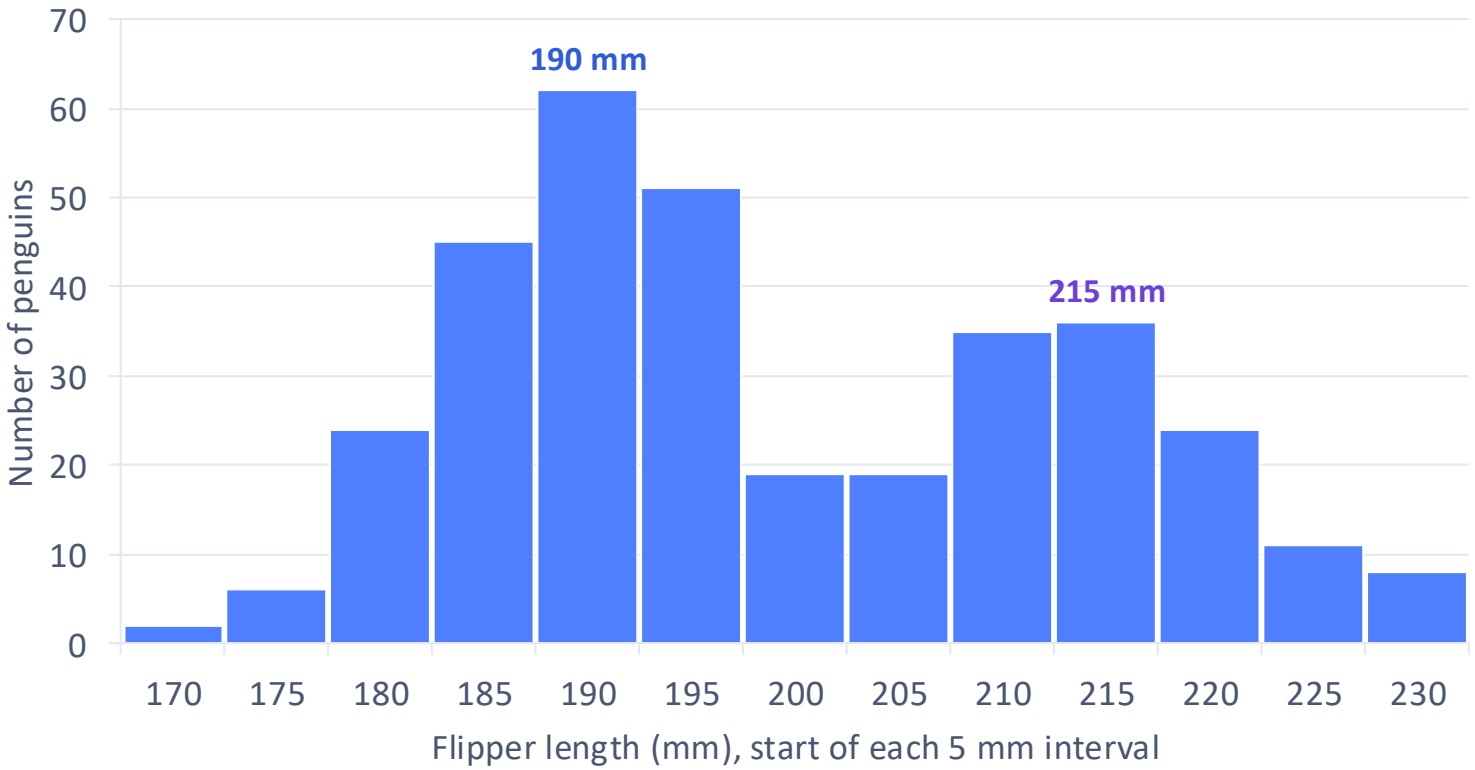
05

CHAPTER

# The shape of a distribution

Where values concentrate, and what the shape tells us about the data.

# Histogram of flipper length




Source: Palmer penguins (n = 342). Each bar counts the penguins in a 5 mm interval.

MAIN FINDING

Two peaks: small species (Adelie, Chinstrap) mixed with a large one (Gentoo).

The mean, 201 mm, falls between the peaks.

# Three common shapes

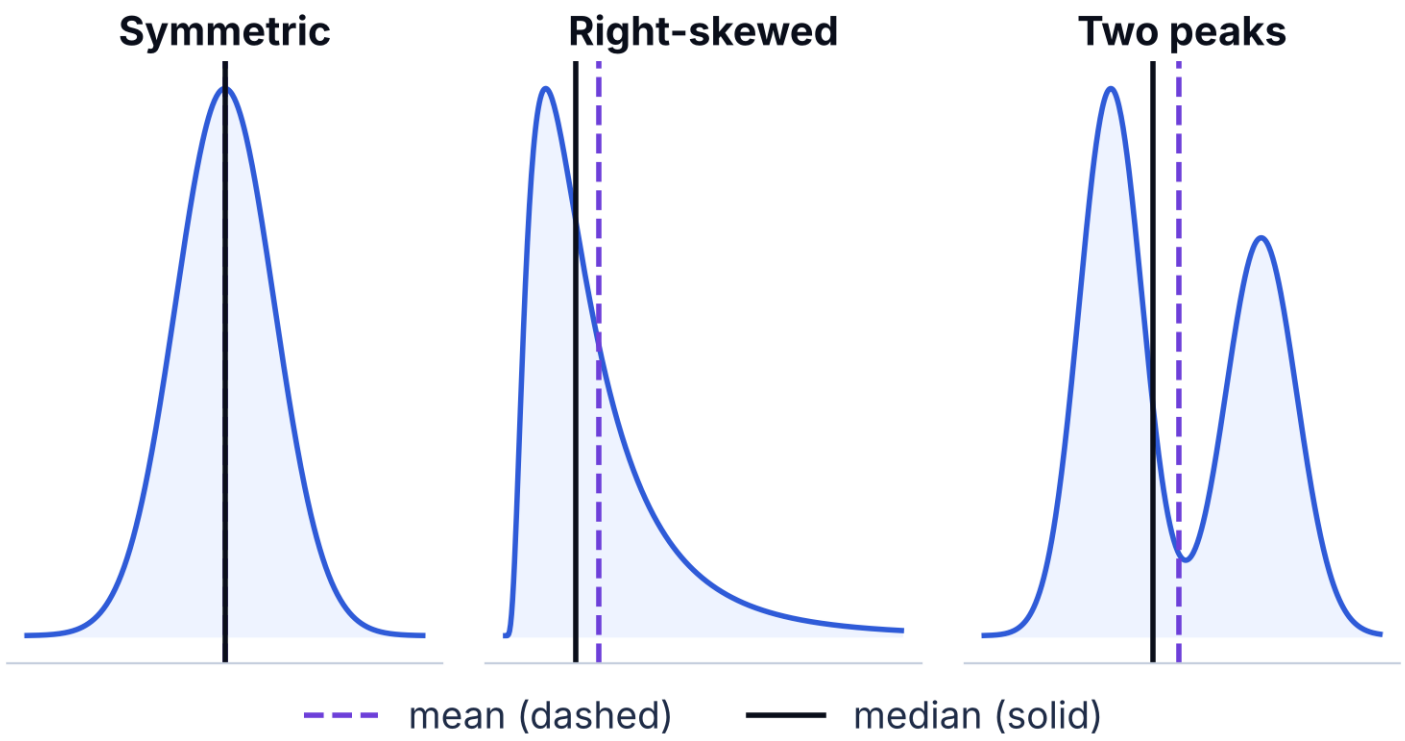



Figure 3. Illustrative distributions generated for teaching (not real data). In the symmetric case, mean = median.

WHAT TO NOTICE

- Symmetric: mean ≈ median
- Right-skewed: a long tail of large values pulls the mean above the median (salaries, prices)
- Two peaks: probably two groups mixed together

# The normal distribution and the 68-95-99.7 rule

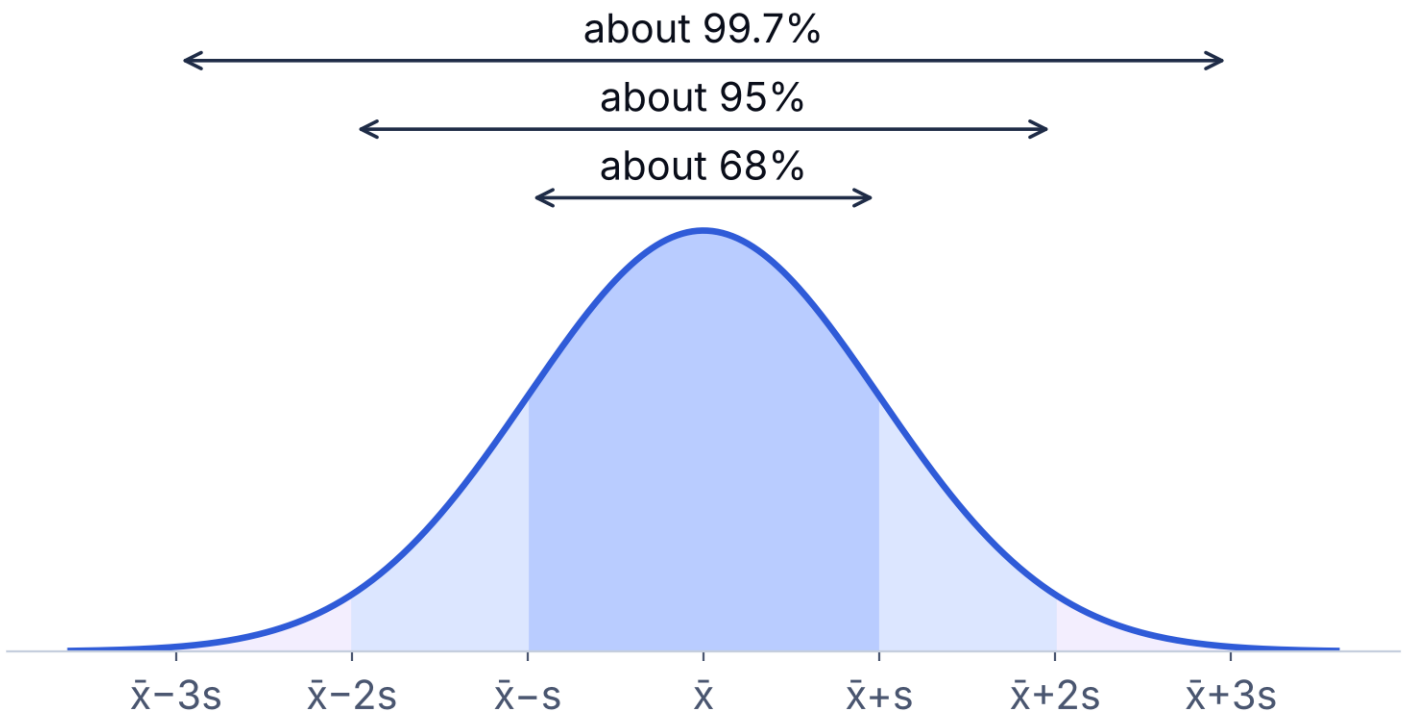



Figure 4. Theoretical normal distribution, drawn with SciPy. x̄ = mean, s = standard deviation.

**WHAT TO NOTICE**

- Bell-shaped and symmetric: mean = median
- About 68% of values within 1 s of the mean, 95% within 2 s, 99.7% within 3 s
- Penguin body mass: 67.5%, 97.4%, 100%. Close, not exact

# The z-score: how unusual is a value?

$$z = (x - \bar{x}) \,/\, s$$

## SYMBOLS

$x$ : the value to examine

$\bar{x}$, $s$ : mean and standard deviation of the variable

$z$ : number of standard deviations from the mean (no unit)

## INTUITION

z = 0 at the mean; z < 0 below it; |z| > 2 is unusual.

Penguin of 5,800 g: z = (5,800 – 4,202) / 802 ≈ 1.99

# Why z-scores matter in machine learning

- Features come in different units and scales
  - Flipper length: $s \approx 14$ mm. Body mass: $s \approx 802$ g
- Some models compare observations by distance; large-scale features would dominate
- Standardisation replaces every value by its z-score: each feature gets mean 0 and standard deviation 1
- In scikit-learn: `StandardScaler` (Session 02)

**KEY TAKEAWAY**

Standardising a feature means computing the z-score of every one of its values.

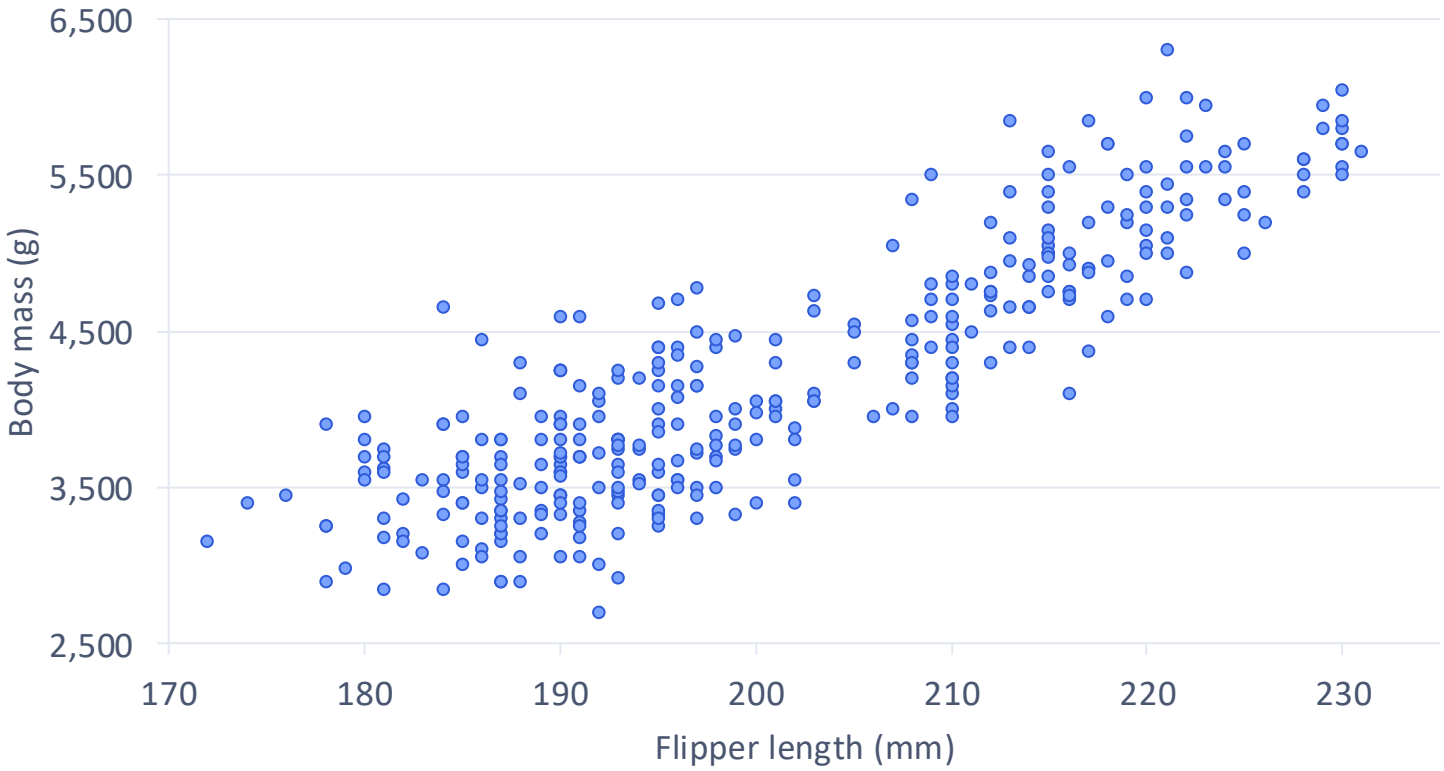
06

CHAPTER

# Two variables together

Do two measurements move together, and how strongly?

# Flipper length and body mass




Source: Palmer penguins (n = 342). One point per penguin.

MAIN FINDING

Longer flippers go with heavier penguins: a positive, roughly linear relationship.

Correlation r = 0.87

# The correlation coefficient r

$$r = \Sigma_i \, a_i \cdot b_i \, / \, (n - 1)$$

## SYMBOLS

$a_i$ : z-score of observation i on variable x

$b_i$ : z-score of observation i on variable y

$r$ : between −1 and 1, no unit

## INTUITION

Both above or both below the mean: positive product. One above, one below: negative.

r measures straight-line relationships only: always look at the plot.

# Reading a correlation

## Positive (r near +1)

PATTERN

Both increase together

EXAMPLE

Flipper length and body mass: r = 0.87

## Near zero (r ≈ 0)

PATTERN

No straight-line link (a curved one may exist)

EXAMPLE

Bill length and bill depth, all penguins: r = −0.24 (weak)

## Negative (r near −1)

PATTERN

One increases, the other decreases

EXAMPLE

Age of a used car and its price

# Correlation is not causation

✗ INCORRECT APPROACH

Ice-cream sales and drownings rise together, so ice cream causes drowning

✓ CORRECTION

Look for a third variable: hot weather increases both. Only a controlled experiment can prove a cause.

WHY IT FAILS

r measures how two variables move together, nothing more. A model can predict with correlated features, but this does not show that changing one changes the other.

# A hidden group can reverse a correlation

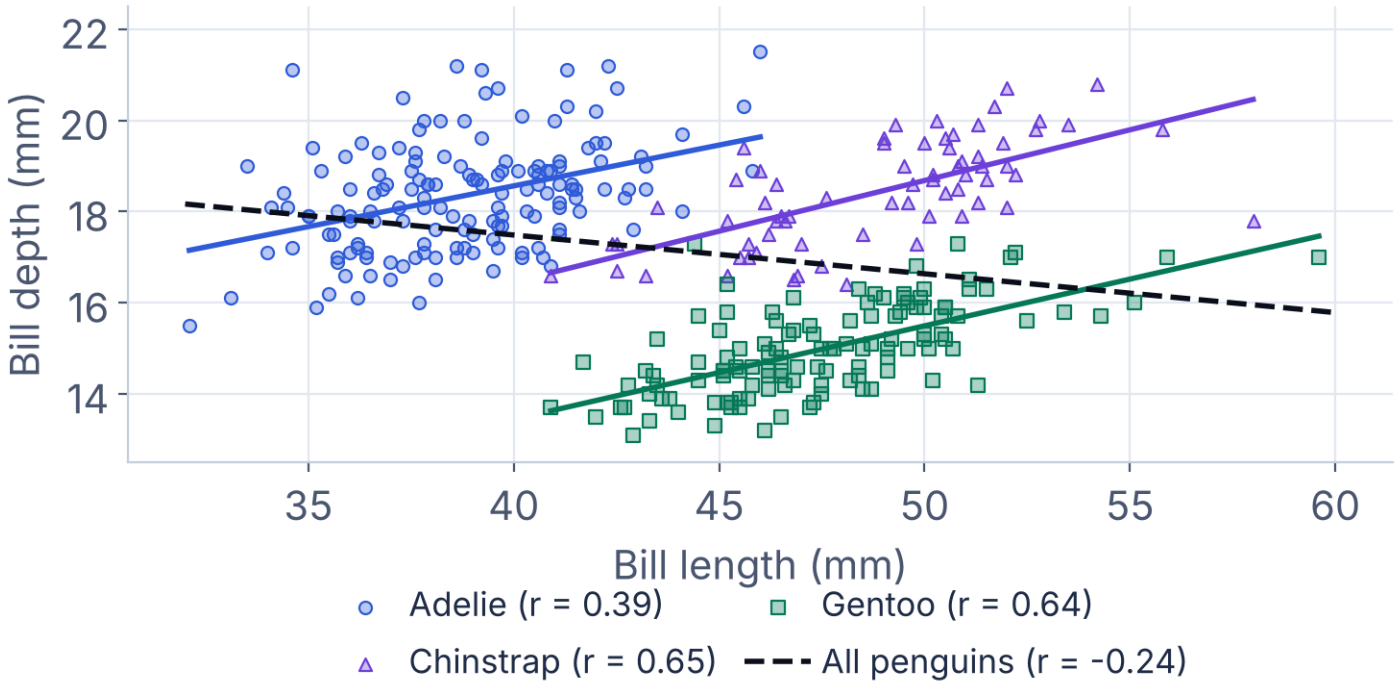



Figure 5. Bill length and depth (n = 342), with a least-squares line per group. Data: Palmer penguins.

**WHAT TO NOTICE**

- All penguins together: r = −0.24
- Inside each species: r between 0.39 and 0.65, positive
- The species is a hidden third variable (Simpson's paradox)
- Look at groups before trusting a single number

07

LAB

# Lab: describing a real dataset in Python

Reproduce today's statistics with pandas, then answer questions about the penguins.

LAB · SETUP

# Setting up Google Colab

- Open `colab.research.google.com` and sign in with a Google account
- File > New notebook. Run a cell with Shift+Enter
- pandas, seaborn and matplotlib are already installed
- The dataset is downloaded from the seaborn repository on GitHub: Internet access is required
- No network? Upload `penguins.csv` and use `pd.read_csv("penguins.csv")`

KEY TAKEAWAY

In a notebook, the last line of a cell is displayed automatically.

LAB · GUIDED CODE

# Loading the data and counting missing values

CODE · PYTHON

```python
import seaborn as sns
import pandas as pd

df = sns.load_dataset("penguins")
print(df.shape)
print(df.isna().sum())
```

OUTPUT

```
(344, 7)
species              0
island               0
bill_length_mm       2
bill_depth_mm        2
flipper_length_mm    2
body_mass_g          2
sex                 11
dtype: int64
```

LAB · GUIDED CODE

# describe(): the summary of a variable

CODE · PYTHON

```python
mass = df["body_mass_g"]
mass.describe().round(1)

# whole table, numerical columns
df.describe()
```

OUTPUT

```
count      342.0
mean      4201.8
std        802.0
min       2700.0
25%       3550.0
50%       4050.0
75%       4750.0
max       6300.0
```

Output of line 2

LAB · GUIDED CODE

# Comparing groups with groupby

CODE · PYTHON

```python
(df.groupby("species")["body_mass_g"]
   .agg(["mean", "median", "std"])
   .round(1))
```

OUTPUT

```
             mean  median    std
species
Adelie     3700.7  3700.0  458.6
Chinstrap  3733.1  3700.0  384.3
Gentoo     5076.0  5000.0  504.1
```

LAB · GUIDED CODE

# Charts and correlation

CODE · PYTHON

```python
import matplotlib.pyplot as plt
fl = df["flipper_length_mm"]

fl.plot.hist(bins=13)
plt.show()

df.boxplot("body_mass_g", by="species")
plt.show()

r = fl.corr(df["body_mass_g"])
print(round(r, 2))
```

OUTPUT

```
0.87
```

Lines 4 and 7 draw a histogram and box plots like those of Chapters 4 and 5.

LAB · EXERCISE

# Lab · Describe the bill length

GOAL

Describe bill_length_mm with numbers and charts. One sentence of interpretation per question.

TIME

**25 min**

INSTRUCTIONS

1. How many values are missing?
2. Mean, median, s, Q1, Q3 and IQR
3. Recompute s from its formula; compare with .std()
4. Histogram: one peak or several?
5. Which species has the longest bills on average?
6. r between bill and flipper length: interpret it

DELIVERABLE

Your notebook, with the code and one sentence per question.

Keep it: Session 02 starts from it.

# Exit question

QUESTION

## Apartment prices (hypothetical): mean 1.9 million MAD, median 1.2 million MAD. What does this suggest?

A

The prices are distributed symmetrically

B

A few very expensive apartments pull the mean up

C

Most apartments cost more than 1.9 million MAD

D

The standard deviation of the prices is zero

# Recap

## KEY TAKEAWAYS

- ML learns rules from examples: features x, target y
- Regression: a number. Classification: a category
- Center: mean (sensitive) and median (robust)
- Spread: s, quartiles, IQR; 1.5 × IQR flags outliers
- Look at shapes and groups; correlation ≠ causation

## NEXT

- Session 02: preparing data. Missing values, encoding categories, scaling with z-scores
- Bring your lab notebook
- Practice: OpenIntro Statistics, chapter 2

# References


(1) Gorman, K. B., Williams, T. D., & Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE, 9(3), e90081.

(2) Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. doi.org/10.5281/zenodo.3960218

(3) Diez, D., Çetinkaya-Rundel, M., & Barr, C. D. (2019). OpenIntro Statistics (4th ed.), ch. 2. openintro.org/book/os

(4) James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning, with Applications in Python. Springer. statlearning.com

(5) pandas 3.0 documentation: Series.std (ddof=1 by default); What's new in 3.0.0 (21 January 2026). pandas.pydata.org/docs

(6) NumPy 2.5 documentation: numpy.std (ddof=0 by default). numpy.org/doc/stable

(7) scikit-learn 1.9 documentation: StandardScaler. scikit-learn.org/stable

(8) Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.




Course page
learn.drsoufiane.com

SH

# Thank you · Questions?

Next session: preparing data for a model. Bring your lab notebook.

**Dr. Soufiane Hourri, PhD, HDR**

drsoufiane.com