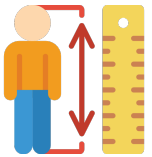
Foundations of Data Science

Lecture 2: Data and exploratory analysis

Prof. Gilles Louppe
g.louppe@uliege.be

LIÈGE université

Data



What is data?

Data are **recorded observations** about the world.

They can take many forms, including numbers, text, images, and more.

Mathematically, data can be viewed as a function f that maps real-world entities ω to measurable values x,

$$f : \Omega \to \mathcal{X},$$

where

- Ω is the sample space of possible states ω of the world,
- $\mathcal{X}$ is the measurement space of possible observations x.

Examples:

- Person's height: $\omega \in \{\textbf{all humans}\} \to x \in \mathbb{R}^+$ (cm)
- Stock price: $\omega \in \{\textbf{market states}\} \to x \in \mathbb{R}^+$ (USD)
- Image pixels: $\omega \in \{\textbf{light intensities}\} \to x \in [0, 255]^3$ (RGB values)

If the sample space Ω is equipped with a probability function p, then the data $x = f(\omega)$ can be viewed as a random variable with distribution induced by p,

$$x \sim p(x) = \int_{\omega \in \Omega} p(\omega)\delta(x - f(\omega))d\omega,$$

where δ is the Dirac delta function.

The **measurement process** is part of the data generation mechanism and can be modeled by extending the sample space to include measurement conditions θ,

$$f : \Omega \times \Theta \to \mathcal{X},$$

where Θ represents factors like instrument settings, environmental conditions, and observer effects.

Measurements can introduce quantization (continuous to discrete), noise (random perturbations), and bias (systematic deviations).

As before, if both Ω and Θ are equipped with a joint probability function p, then the data $x = f(\omega, \theta)$ can be viewed as a random variable with distribution induced by p,

$$x \sim p(x) = \iint_{\omega \in \Omega, \theta \in \Theta} p(\omega, \theta) \delta(x - f(\omega, \theta)) d\omega d\theta,$$

which now captures the variability introduced by both the underlying phenomena and the measurement process.

Data types

Atomic data are the indivisible units of information collected through measurements. It is often categorized based on its nature and the operations that can be performed on it.

- Numerical (continuous, discrete)
- Categorical (nominal, ordinal)

Numerical data

- Continuous: $\boldsymbol{x} \in \mathbb{R}$ (e.g., temperature), $\boldsymbol{x} \in \mathbb{R}^+$ (e.g., height, weight)
- Discrete: $\boldsymbol{x} \in \mathbb{Z}$ (e.g., counts)

Categorical data

- Nominal: $x \in \mathcal{C} = \{c_1, c_2, ..., c_n\}$ (e.g., colors, types, text characters) without intrinsic order
- Ordinal: $x \in \mathcal{C}$ with ordering relations $c_1 \prec c_2 \prec ... \prec c_n$ (e.g., ratings, grades)

Example: medical records

```
Patient ID: 10847          # Categorical, Nominal
Age: 34                    # Numerical, Discrete
Height: 175.2 cm           # Numerical, Continuous
Blood type: O+             # Categorical, Nominal
Pain level: 7/10           # Categorical, Ordinal
Temperature: 38.1°C        # Numerical, Continuous
```

Data structures

A measurement x can be a single atomic value or a composite structure made of multiple atomic values. Common aggregates or data structures include:

- Tabular data
- Arrays and tensors
- Sequences
- Networks and graphs

Tabular data

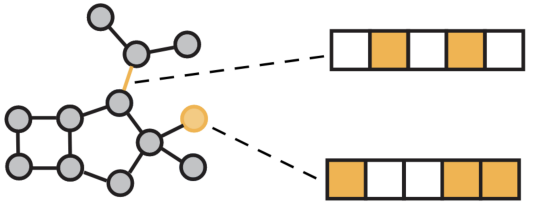
	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

Arrays and tensors

Sequences

```
['F', 'l''o', 'w', 'e', 'r']
```

Networks and graphs



Data frames $\mathbf{X}$ represents **tabular collections** of n records (rows) over d variables/atomic measurements (columns),

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}.$$

Each entry x_{ij} corresponds to the value of variable j for record i.

Variables are often heterogeneous (mixing numerical and categorical types). When all variables are numerical, the data frame can be viewed as a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$.

Collections of homogeneous measurements can be represented as **arrays** or **tensors** $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$, where the position of each atomic value in the array is usually associated to a spatial or temporal location.

- Images: 3d arrays $\mathbf{X} \in [0, 255]^{h \times w \times c}$ (height, width, channels).
- Videos: 4d arrays $\mathbf{X} \in [0, 255]^{t \times h \times w \times c}$ (time, height, width, channels).

Data can also be structured as ordered **sequences** $S = (x_1, x_2, ..., x_T)$ indexed by time or position. Each element x_t can be atomic or composite.

- Time series: $S = (x_1, x_2, ..., x_T)$ where x_t is a measurement at time t (e.g., stock prices).
- Text: $S = (w_1, w_2, ..., w_T)$ where w_t is the t-th word in a document.

Finally, data can be organized as **networks** or **graphs** $G = (V, E)$, where entities are represented as nodes V and relationships as edges E. Each may also have associated attributes, x_v for nodes and x_{uv} for edges.

- Molecular structures, where nodes represent atoms and edges represent bonds.
- Social networks, where nodes represent individuals and edges represent interactions or relationships.

Follow the tutorials in `nb02a-tables.ipynb`, `nb02b-jax.ipynb`, and `nb02c-data-wrangling.ipynb` to practice working with arrays and data frames in Python.

Data quality

Real-world data are often imperfect and may suffer from various **quality issues** that can impact analysis and modeling. Common data quality issues include:

- Missing values
- Measurement errors
- Outliers

Missing values are common in real-world datasets and can arise from various factors such as non-response in surveys, sensor malfunctions, or data corruption.

Let $\mathbf{X}_{\text{full}} \in \mathbb{R}^{n \times d}$ be a complete data matrix and $\mathbf{M} \in \{0, 1\}^{n \times d}$ be the missing data indicator matrix, where $m_{ij} = 1$ if entry ij is observed and $m_{ij} = 0$ if it is missing. The observed data can be represented as $\mathbf{X} = \mathbf{X}_{\text{full}} \odot \mathbf{M}$, where $\odot$ denotes the element-wise product.

The patterns of missingness can be modeled as part of the measurement process:

- Missing Completely at Random (MCAR): The probability of missingness is independent of both observed and masked data, $p(m_{ij} = 0|\mathbf{X}_{\text{full}}) = p(m_{ij} = 0)$.
- Missing at Random (MAR): The probability of missingness may depend on observed data but not on masked data, $p(m_{ij} = 0|\mathbf{X}_{\text{full}}) = p(m_{ij} = 0|\mathbf{X})$.
- Missing Not at Random (MNAR): The probability of missingness depends on masked data as well.

Each mechanism or assumption has implications for how to handle missing data during analysis.

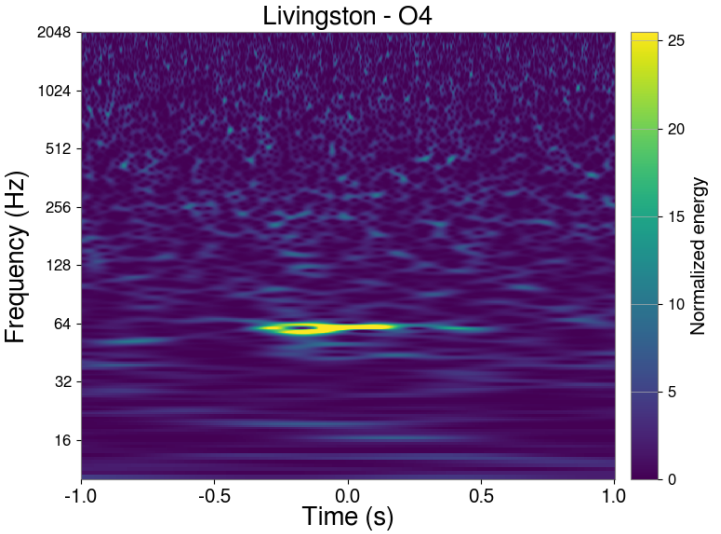
Example:

Survey of $n = 1000$ respondents, 30% do not answer the income question.

- MCAR: Randomly selected respondents skip the question.
- MAR: Younger respondents are less likely to answer.
- MNAR: High earners refuse to answer.

Imputing or discarding missing values without expliciting the assumptions about the missingness mechanism can lead to biased results.

The measurement process can also introduce acquisition errors and produce observations that deviate significantly from regular measurements. These **outliers** can arise from instrument malfunctions, data entry errors, or rare events.



Example:

Glitches in gravitational wave detectors are outliers that can mimic true signals and complicate detection efforts. They can arise from environmental disturbances, instrumental artifacts, or other non-astrophysical sources.

Treating outliers requires a model of the measurement process that either describe measurements under normal conditions or explicitly accounts for anomalies.

Outliers should not be removed blindly unless explicitly justified by the measurement model or domain knowledge.

Temperature readings in Liège:

`20.1, 19.8, 20.3, 1000.0, 20.2, 19.9`

The 1000.0 value is an outlier likely due to a sensor error.

`19.2, 19.8, 20.3, 37.8, 20.2, 19.9`

The 37.8 value is a rare but plausible measurement on a hot day.

Exploratory data analysis

Exploratory data analysis (EDA) is the systematic examination of data to understand its structure, patterns and anomalies before formal modeling. It typically involves visual and quantitative techniques to summarize key characteristics of the data.

The goal is to generate hypotheses and inform modeling decisions, not to confirm preconceived notions.

As a guiding example, we consider the **Palmer Archipelago penguins dataset**, which contains measurements for three penguin species (Adelie, Chinstrap, Gentoo) across three islands (Biscoe, Dream, Torgersen).

Switch to `nb02d-eda.ipynb` to follow along.

—

Credits: Horst et al., 2020.

Univariate analysis

Let consider a variable j from a data frame $\mathbf{X} \in \mathbb{R}^{n \times d}$, represented as the vector $\mathbf{x}_j = (x_{1j}, x_{2j}, ..., x_{nj})^T$.

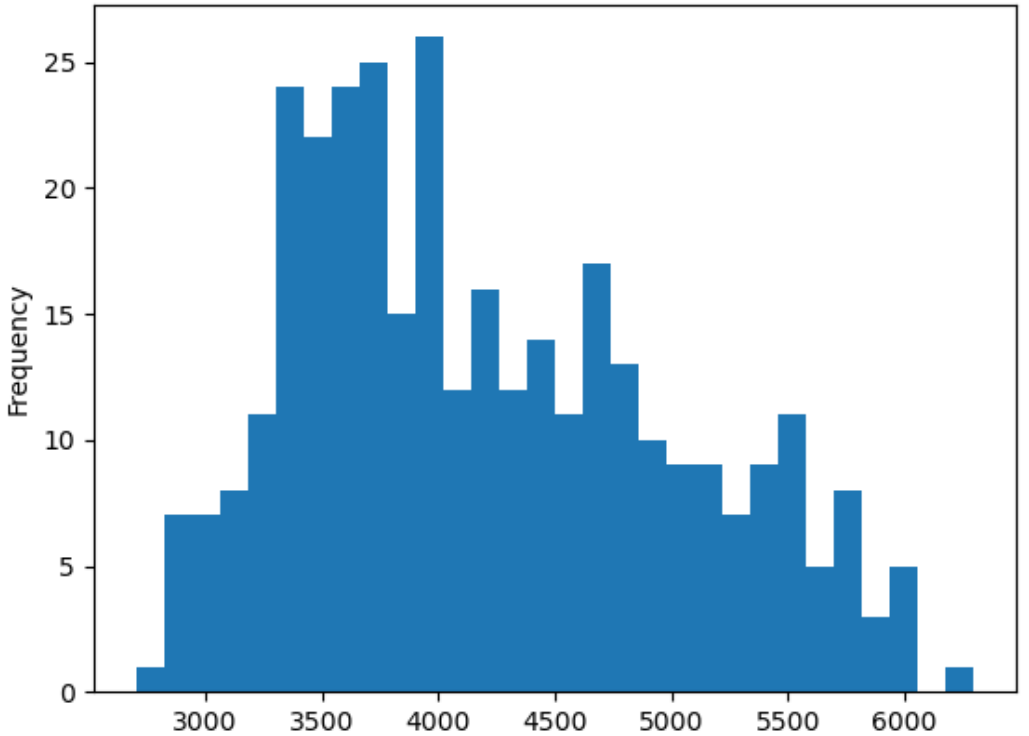
Univariate analysis focuses on understanding the distribution and characteristics of this single variable.

For numerical variables, common techniques include:

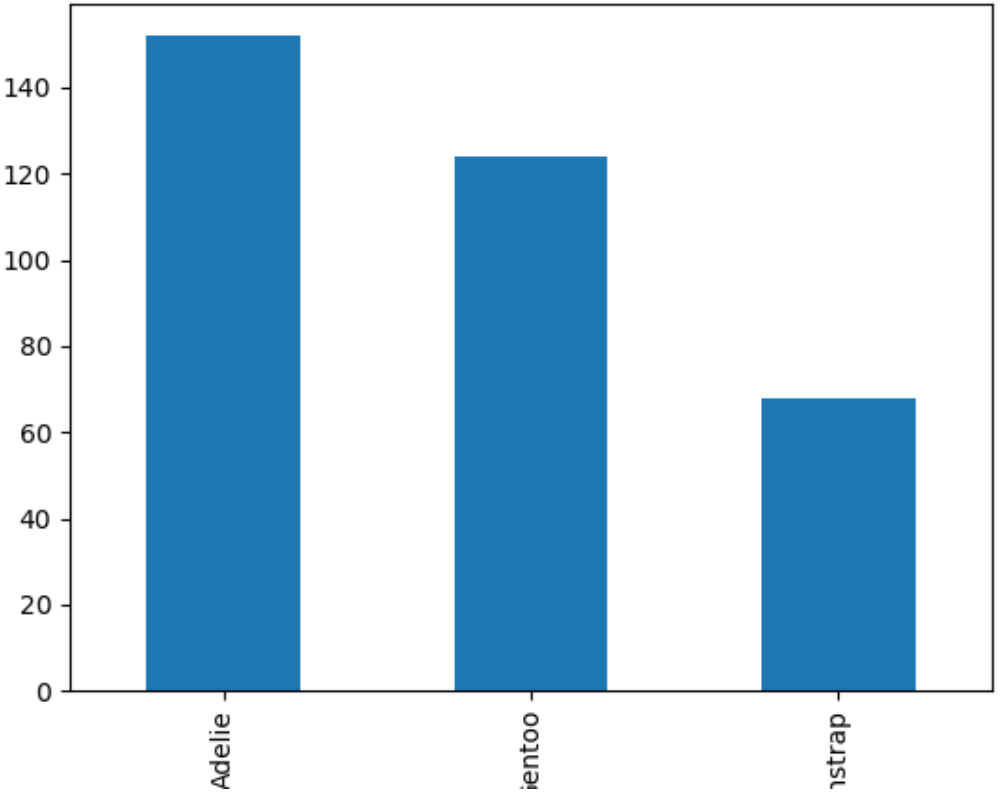
- Looking at the raw data: print values, scroll through them.
- Plotting the data: histograms reveal the distribution shape.
- Summarizing with statistics: mean, median, mode, variance, skewness, kurtosis.

For categorical variables:

- Counting occurrences of each category.



Numerical: Histograms of body mass for all penguins.



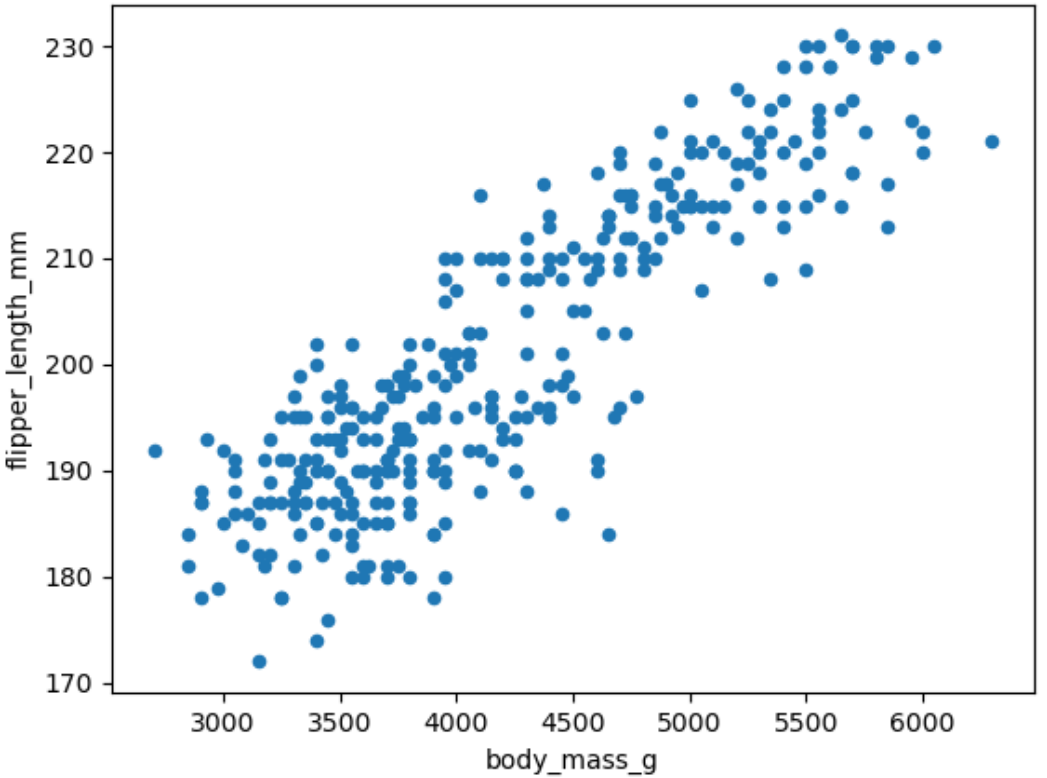
Categorical: Bar plot of species counts for all penguins.

Bivariate analysis

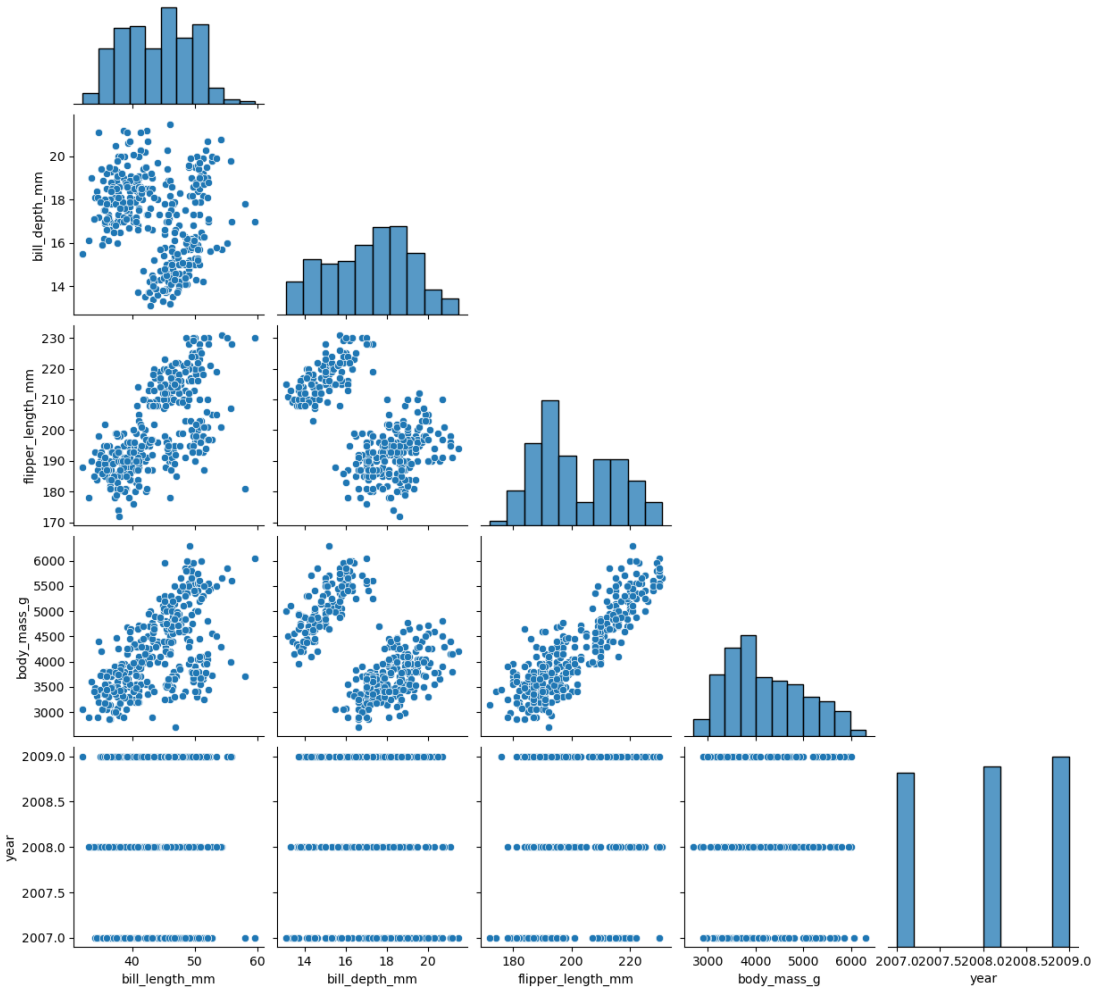
Bivariate analysis examines the relationship between two variables i and j from a data frame $\mathbf{X} \in \mathbb{R}^{n \times d}$, represented as vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

Depending on the types of variables, different techniques are used:

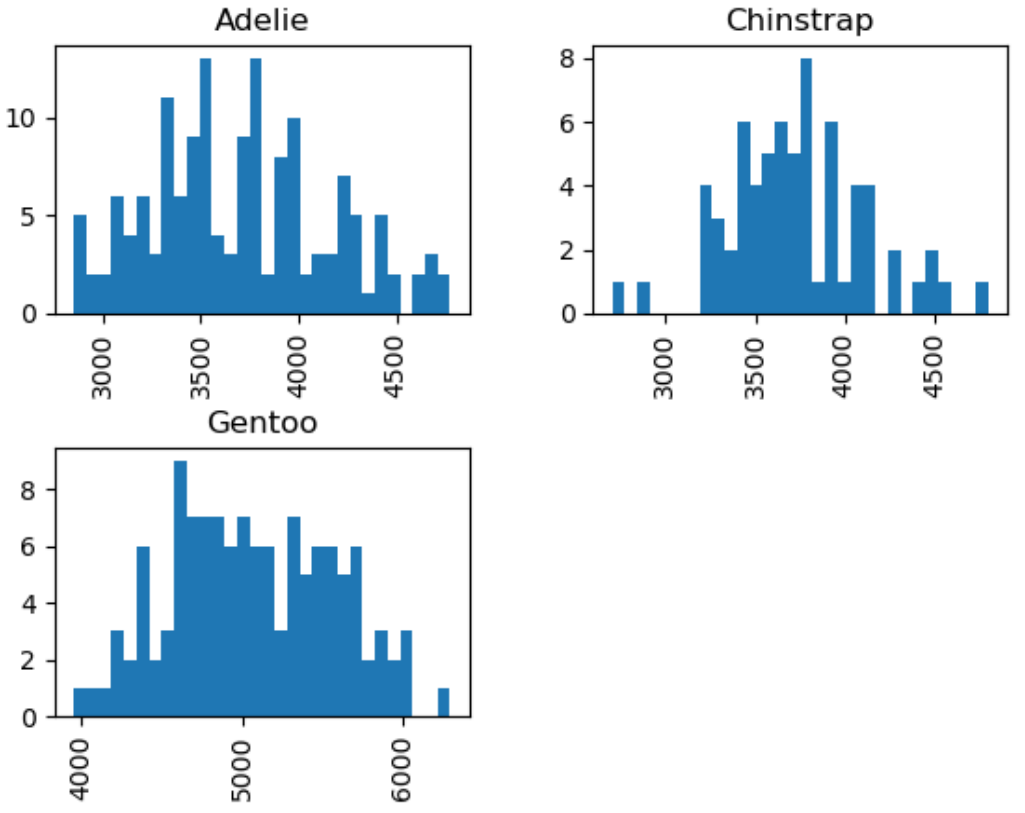
- Pair plots for two numerical variables (scatter or 2d histogram).
- Multiple histograms for categorical vs numerical variables.
- Contingency tables for two categorical variables.
- Correlation coefficients (e.g., Pearson, Spearman) for numerical variables.



Numerical vs. numerical: Scatter plot of body mass vs flipper length.



Pair plots of all numerical variables.



Categorical vs. numerical: Histograms of body mass by species.

island species	Biscoe	Dream	Torgersen
Adelie	44	56	52
Chinstrap	0	68	0
Gentoo	124	0	0

Categorical vs. categorical: Contingency table of species and island.

Correlation coefficients can quantify the dependency between two numerical variables. They are useful but come with assumptions and limitations.

- Pearson correlation

$$\rho_{ij} = \frac{\mathbf{cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j}}$$

  measures linear relationships. It ignores non-linear dependencies.

- Spearman correlation is the Pearson correlation of the rank-transformed variables. It captures monotonic relationships.
- Correlation does not imply causation and can be affected by outliers.

The **mutual information** between two variables $\mathbf{x}_i$ and $\mathbf{x}_j$ measures the reduction in uncertainty about one variable given knowledge of the other,

$$I(\mathbf{x}_i; \mathbf{x}_j) = \sum_{x_i} \sum_{x_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)},$$

where $p(x_i, x_j)$ is the joint probability distribution and $p(x_i)$, $p(x_j)$ are the marginal distributions.

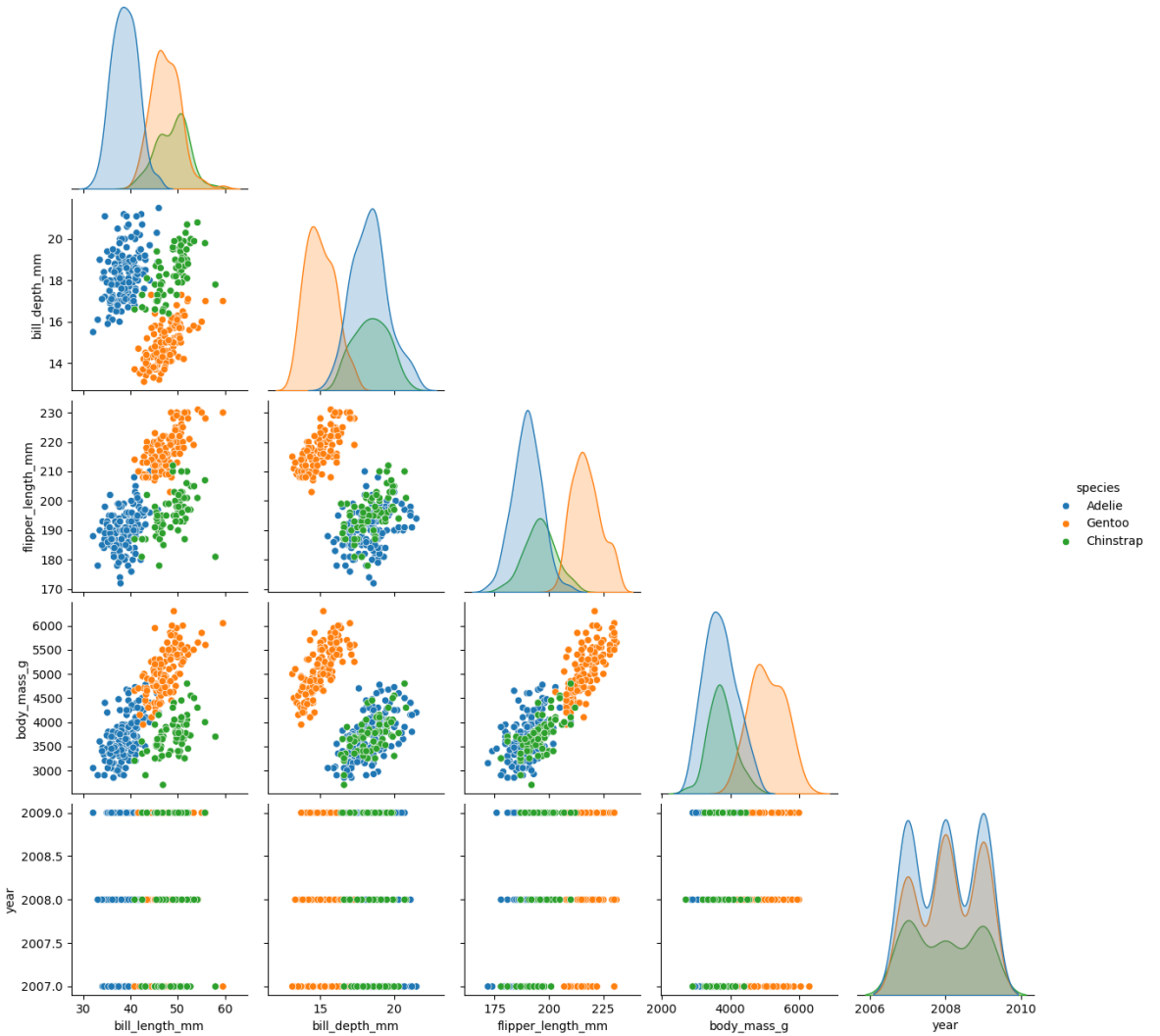
Mutual information is a more general measure of dependency that captures any statistical relationship, not just linear or monotonic ones. However, it is harder to estimate accurately from finite samples.

Multivariate analysis

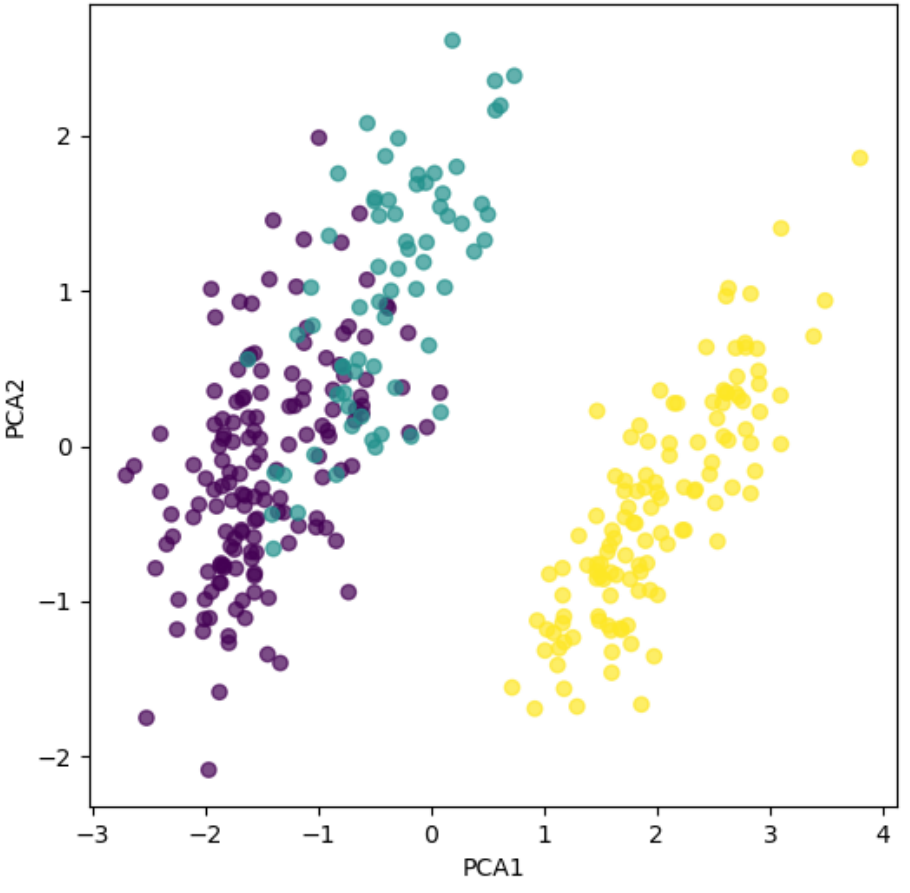
Multivariate analysis explores relationships among three or more variables in a data frame $\mathbf{X} \in \mathbb{R}^{n \times d}$.

Common techniques include:

- The same as bivariate analysis, but conditioning on a third variable (e.g., pair plots colored by species).
- Dimensionality reduction methods (e.g., PCA, t-SNE) to visualize high-dimensional data.
- Clustering algorithms (e.g., k-means, hierarchical clustering) to identify groups of similar records.



Pair plots of all numerical variables, colored by species.



PCA projection of all numerical variables, colored by species.

EDA within Box's loop

EDA is a crucial step in Box's loop for data analysis. It helps to understand the data, generate hypotheses, and **inform modeling decisions**.

- Distribution shapes can suggest appropriate model families.
- Relationships between variables can guide model structure.
- Insights about scale and variance can inform data transformations.

After building a model and computing initial results, EDA can be used to critique the model fit and identify areas for improvement.

- Residual analysis can reveal patterns not captured by the model.
- Prediction errors may point to specific data subsets that are problematic.
- Unexpected patterns in the data can suggest new features or model revisions.

The end.