Fondamenti di Analisi dei Dati

from data analysis to predictive techniques

Prof. Antonino Furnari (antonino.furnari@unict.it)
Corso di Studi in Informatica
Dip. di Matematica e Informatica
Università di Catania

Data Description and Visualization

Datasets are often too large to be understood by simply printing rows or scanning tables. This lecture introduce **Descriptive statistics** and **visualisation techniques** as essential tools for data analysis.

Why Describe and Visualise?

Gather Insights

Understand the main characteristics of the dataset and each variable

Detect Anomalies

Identify missing or anomalous values that could affect the analysis

Identify Relationships

Discover connections between variables and identify trends or similar clusters

Guide Modelling

Inform the choice of subsequent pre-processing and modelling techniques

Example Dataset

```
import pandas as pd
hw=pd.read_csv('http://antoninofurnari.it/downloads/height_weight_pounds.csv')
hw.info()
hw.head()
```

	sex	height	weight
0	M	74	53.484771
1	M	70	38.056472
2	F	61	34.970812
3	M	68	35.999365
4	F	66	34.559390

Frequency Visualisations

Bar Charts

Display absolute or relative frequencies with values on the x-axis and frequencies on the y-axis

Pie Charts

Useful for a few categories, but difficult to interpret with precision

Stacked Charts

Allow comparison of distributions between different groups

⚠️ Pie charts should be used with caution - angular differences are difficult to assess

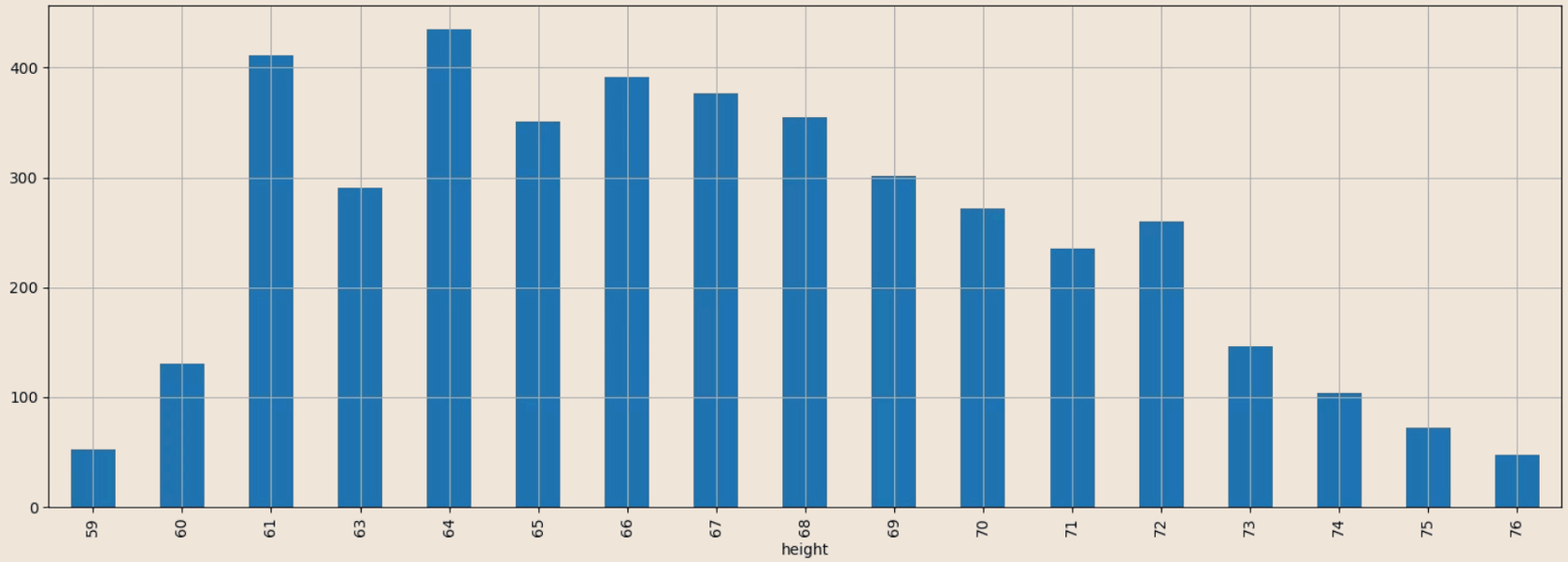
Absolute Frequencies

Absolute frequencies count how many times each value appears in the sample. For discrete variables with finite unique values:

$$a_1, a_2, ..., a_n$$

A frequency n_i us defined the number of times a_i appears in the sample. Note that:

$$\sum_i n_i = n$$



Relative Frequencies

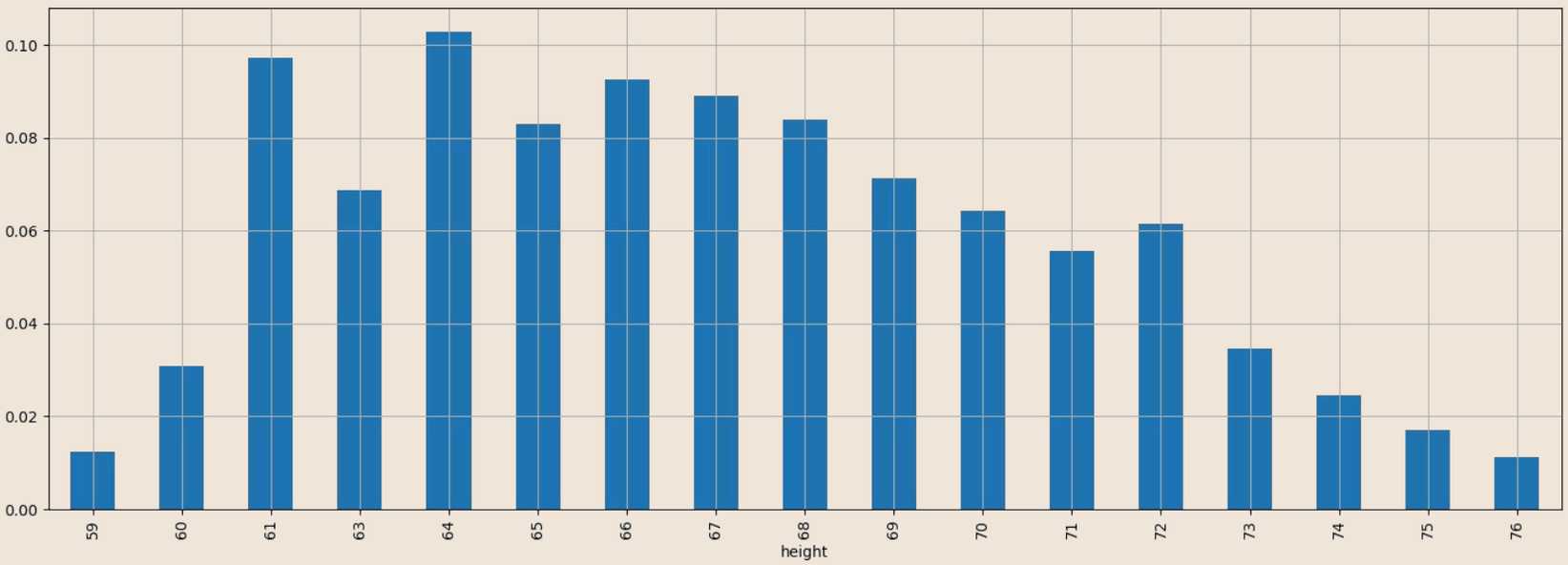
Relative frequencies are defined as follows:

$$f_j = f(a_j) = \frac{n_j}{n}, j = 1, 2, \ldots, k$$

Note that, given the definition, we will have:

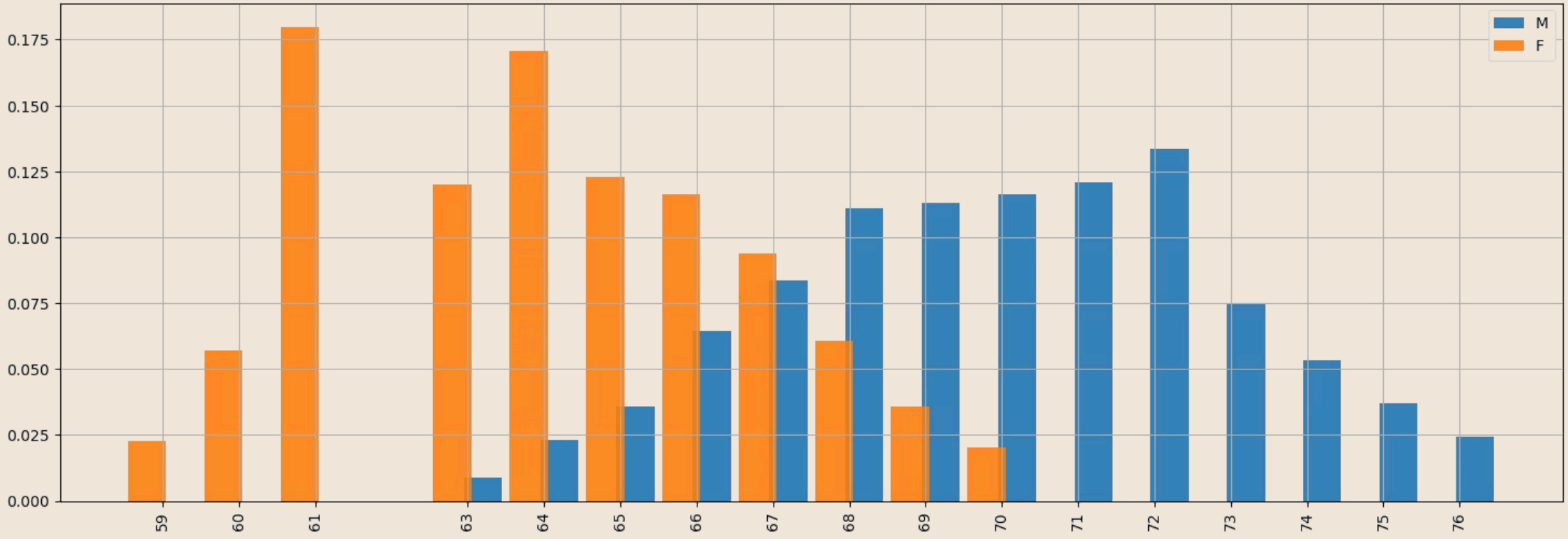
$$n_j \leq n \Rightarrow f_j \leq 1 \, \forall j$$

$$\sum_j f_j = \sum_j \frac{n_j}{n} = \frac{1}{n} \sum_j n_j = \frac{n}{n} = 1$$



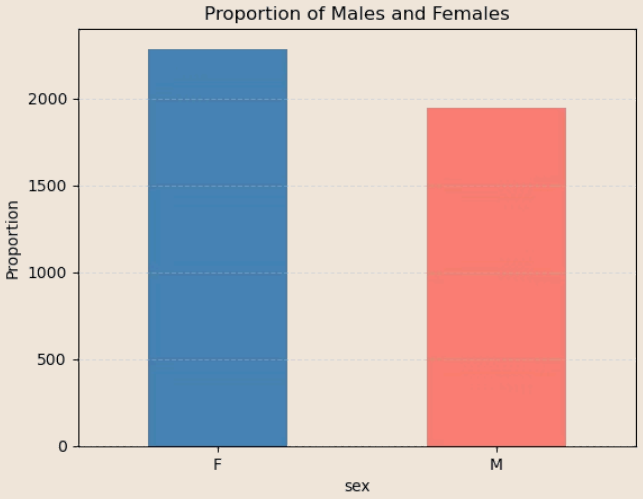
Comparing samples with relative frequencies

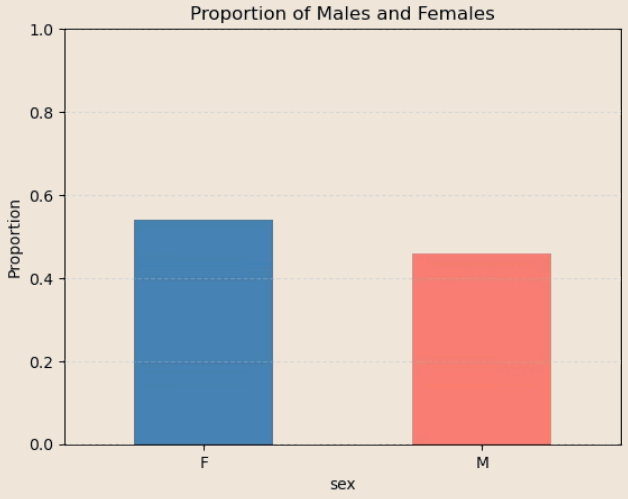
Relative frequency bar charts are useful for comparing different samples. For example, we can consider the heights of men and women in the weights-and-heights dataset.

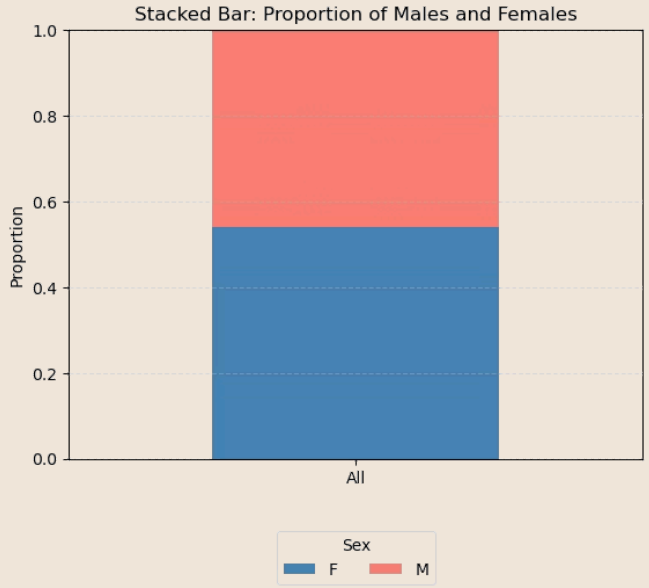


Barplots and stacked barplots

We can also used barplots and stacked barplots to count the number of elements in a given class.

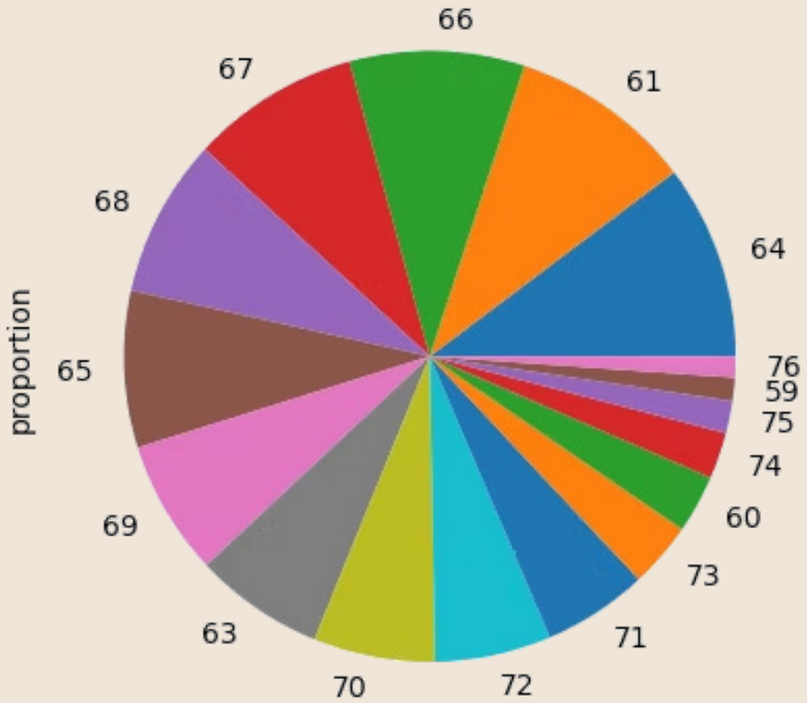


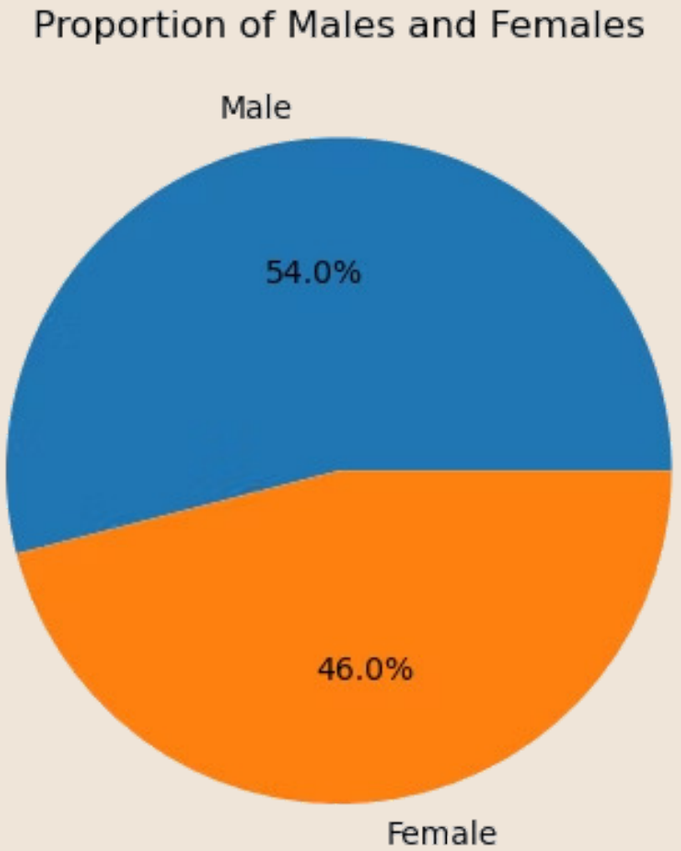




Pie Charts

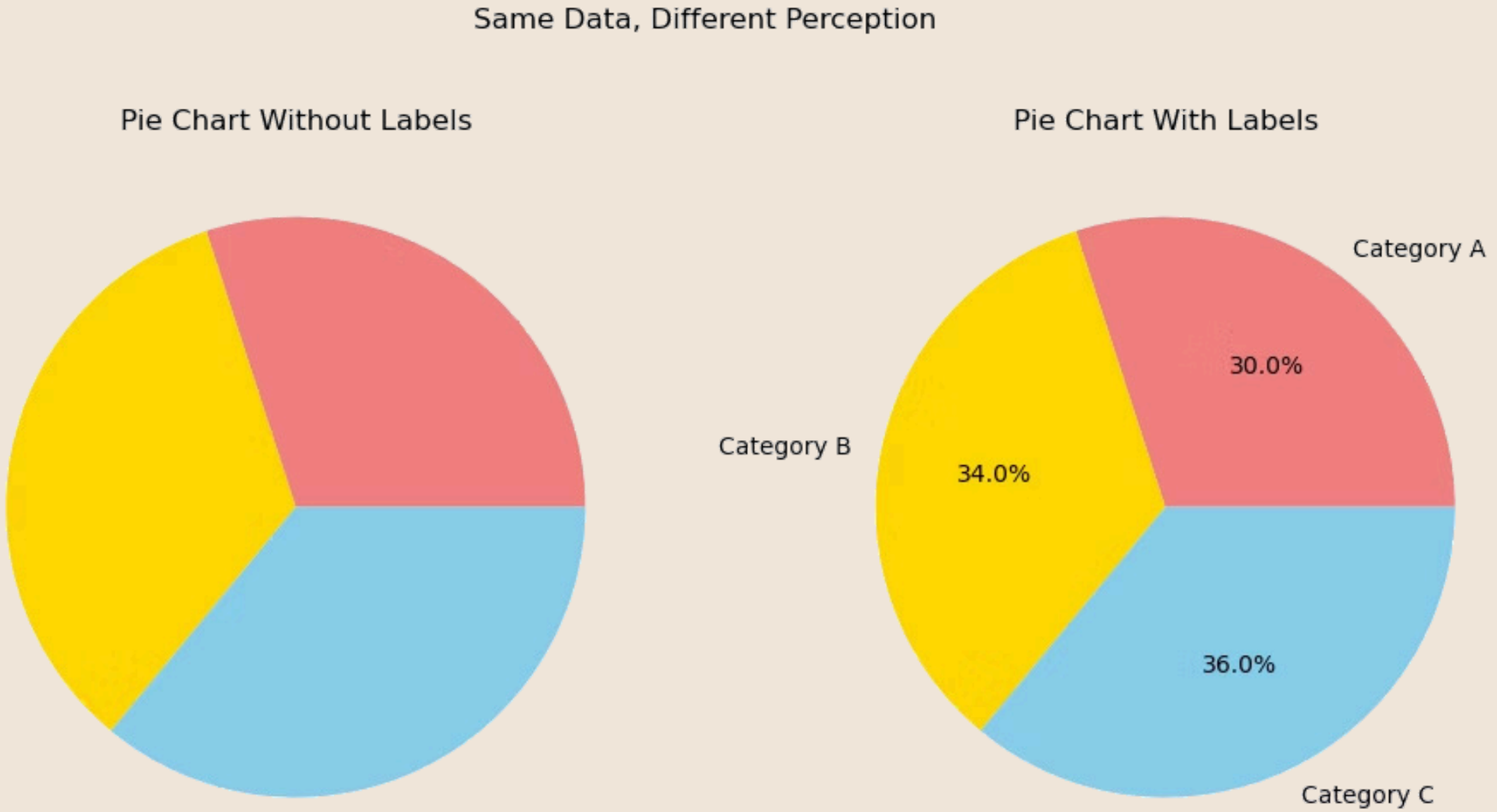
As an alternative to bar charts, relative frequencies can also be displayed using pie charts. These are most used when data cannot be meaningfully sorted (see case on the right).





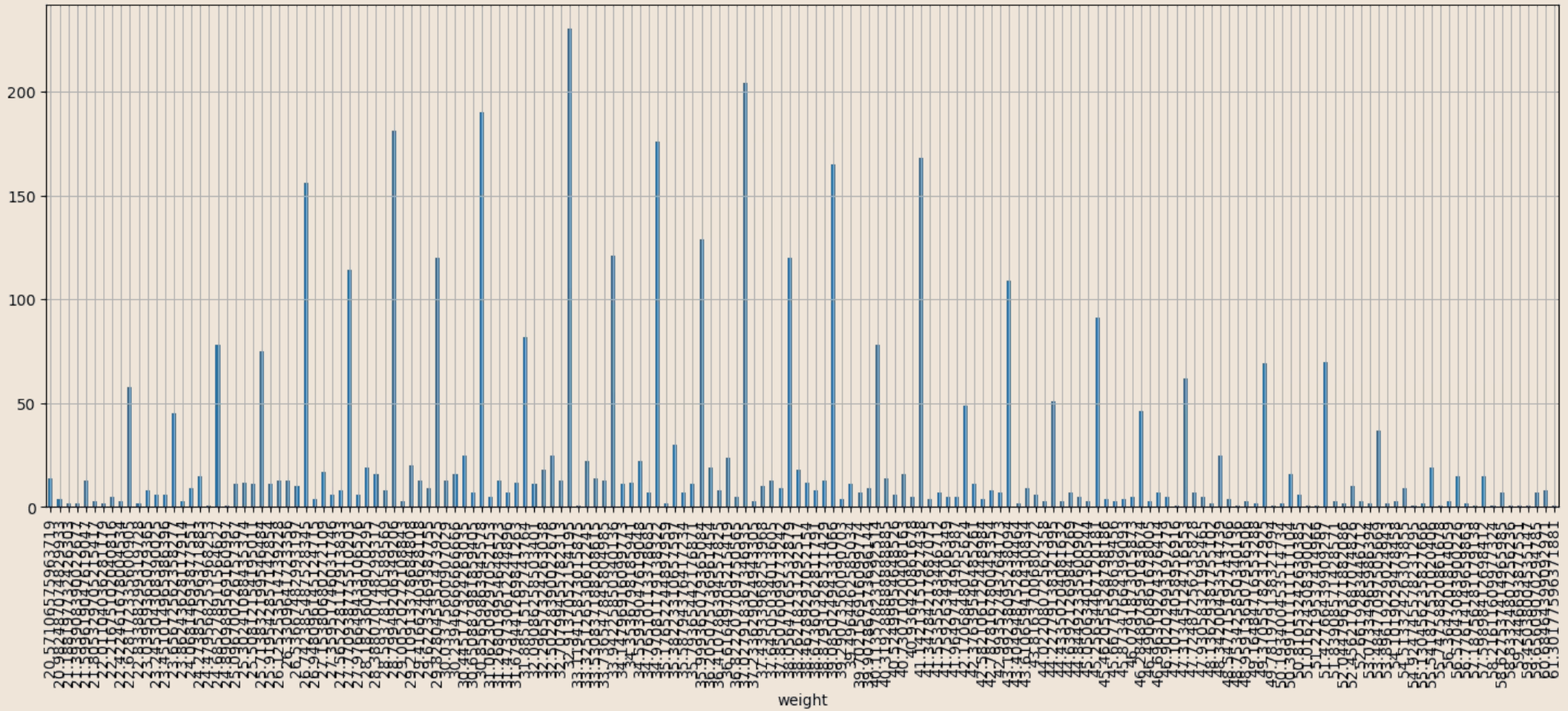
Pie Charts can be hard to read

While pie charts can be intuitive when there is no specific ordering of the elements we want to show, they often introduce more problems than other chart types. In particular, it’s difficult for the human eye to accurately compare angles or assess relative sizes—especially when slices are similar in proportion or not arranged consistently.



Displaying continuous data

When data is continuous (or not heavily quantized), bar charts of frequencies are limited. See the example below with "weights".



Empirical Cumulative Distribution Function (ECDF)

When there are many unique values, discrete frequencies become noisy. The **ECDF** solves this problem:

$$ECDF(x) = \sum_{a_j : a_j \leq x} f(a_j)$$

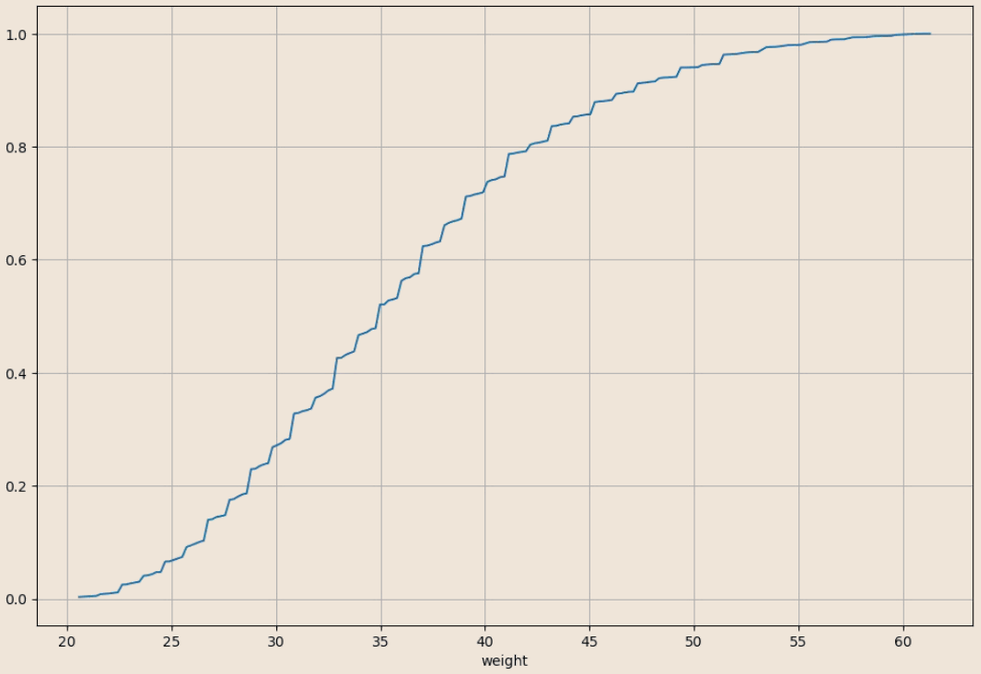
Interpretation

For a point (x,y), y% of the elements have a value ≤ x

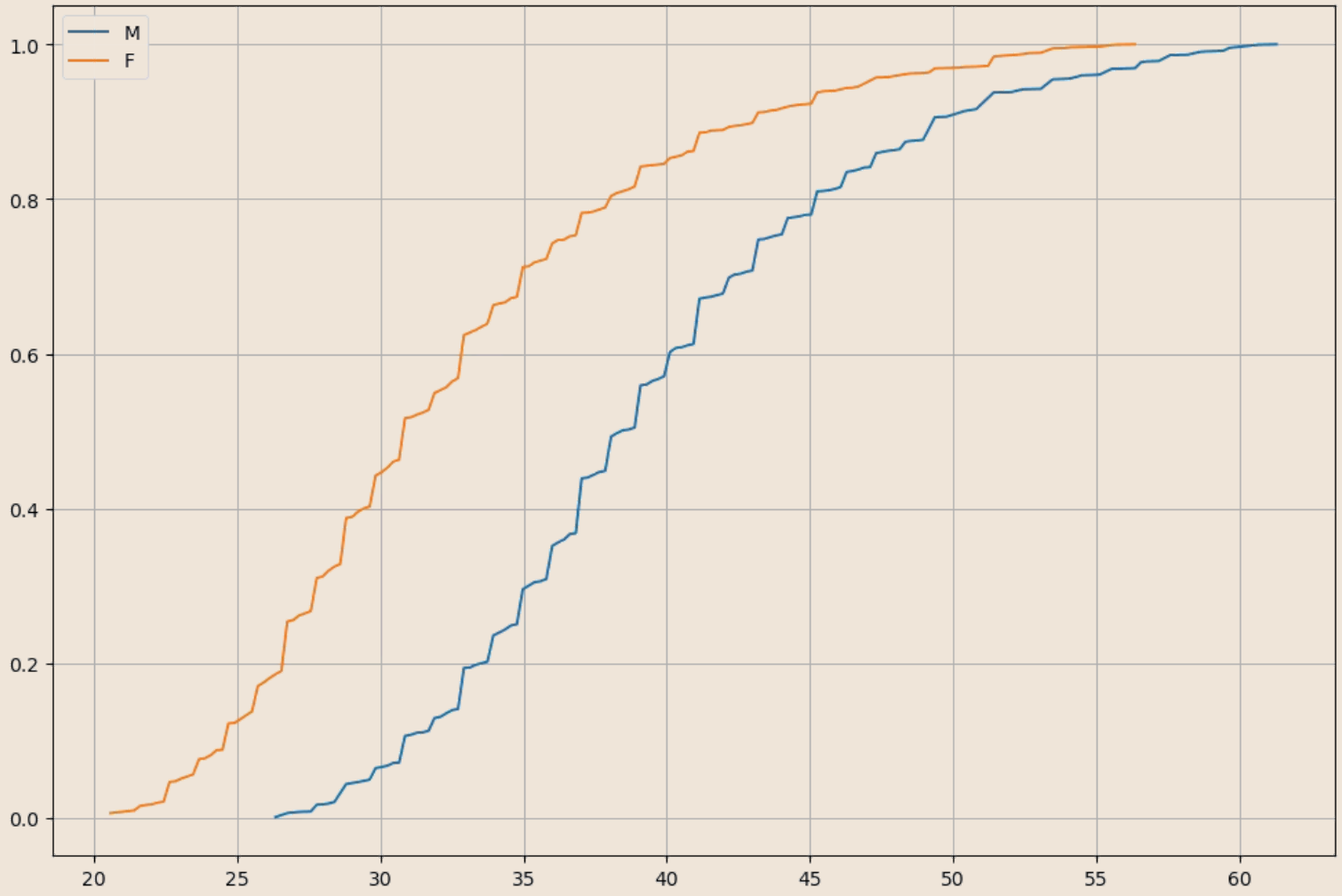
Useful for comparing distributions between groups

Characteristics

- Always increasing
- Last value = 1
- Reduces data noise



Comparing samples with ECDFs



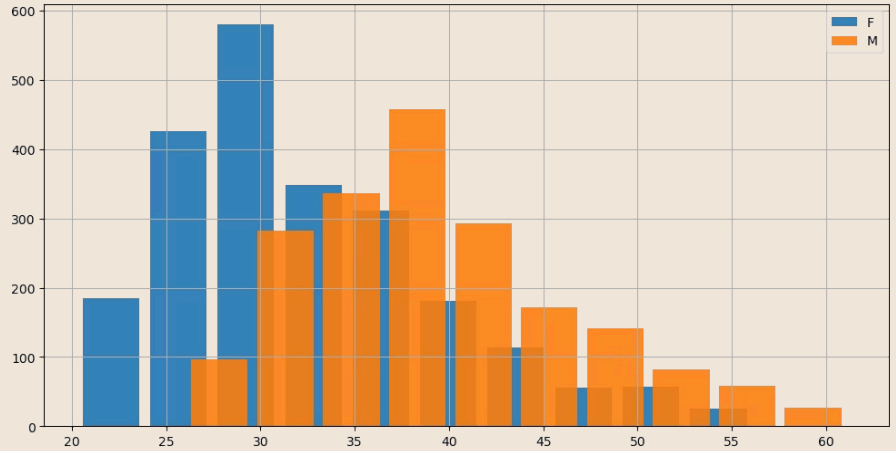
Observing the graph above, we can say that:

- Approximately of men weigh less than ;
- Approximately of women weigh less than ;
- of men weigh less than ;
- of women weigh less than .

In general, the graph tell us that men tend to be havier than women.

Histograms and Binning

Histograms group similar observations into "bins" to reduce noise. Bins are often of equal size.



01

Choosing the Number of Bins

Sturges: #bins = 3.3×log(n)

Rice: #bins = 2×n^(1/3)

02

Custom Bins

It is possible to define arbitrary or variable-sized intervals

03

Comparing Samples

Useful for comparing distributions between different groups

Density Estimation

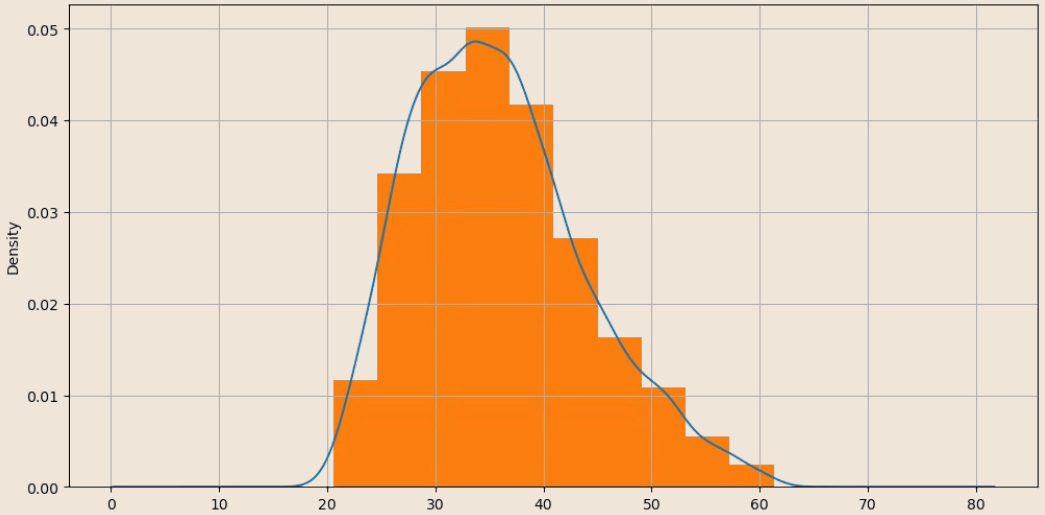
Density estimation overcomes the limitations of histograms by providing a continuous representation instead of arbitrarily categorising data.

Advantages

- Continuous representation
- Does not depend on bin choice
- Visually smoother

Bandwidth Parameter

Controls the "sensitivity to detail" of the estimation



💻 We will use plot.density() with the bw_method parameter and skip details for now

Summary Statistics

Summary statistics are quantitative measures that summarise the main characteristics of a dataset. They allow us to quickly understand the 'size' or scale of the data, as well as its distribution and central tendencies, without having to examine every single data point.

Simplification

They reduce large volumes of data to a few significant values, making them more manageable and interpretable.

Description

They provide a numerical overview of the dataset's distribution, central tendency, and variability.

Comparison

They allow for easy comparison of different datasets or subgroups, highlighting similarities and differences.

Size

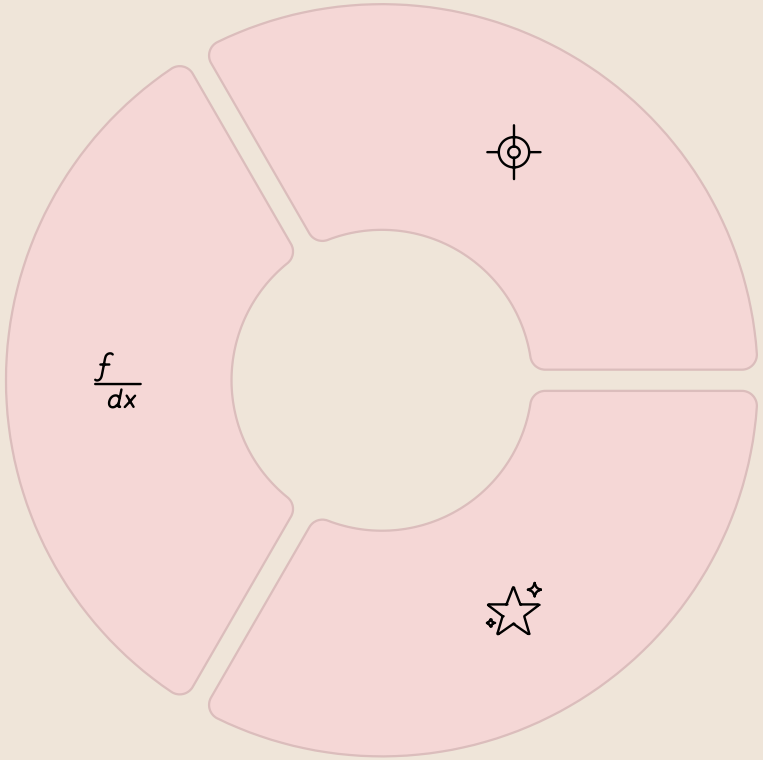
The size of a univariate sample $\{x^{(i)}\}_i^N$ is given by the number of values it contains: $|\{x^{(i)}\}_i^N| = N$. The size of each column can also be different as there may be missing values, generally indicated by `NA` or `NaN`.

Measures of Central Tendency

Mean

$$\overline{X} = \frac{1}{N}\sum_{i}^{N} x^{(i)}$$

Sum divided by sample size



Median

Element that divides the ordered sample into two equal parts

More robust to outliers

Mode

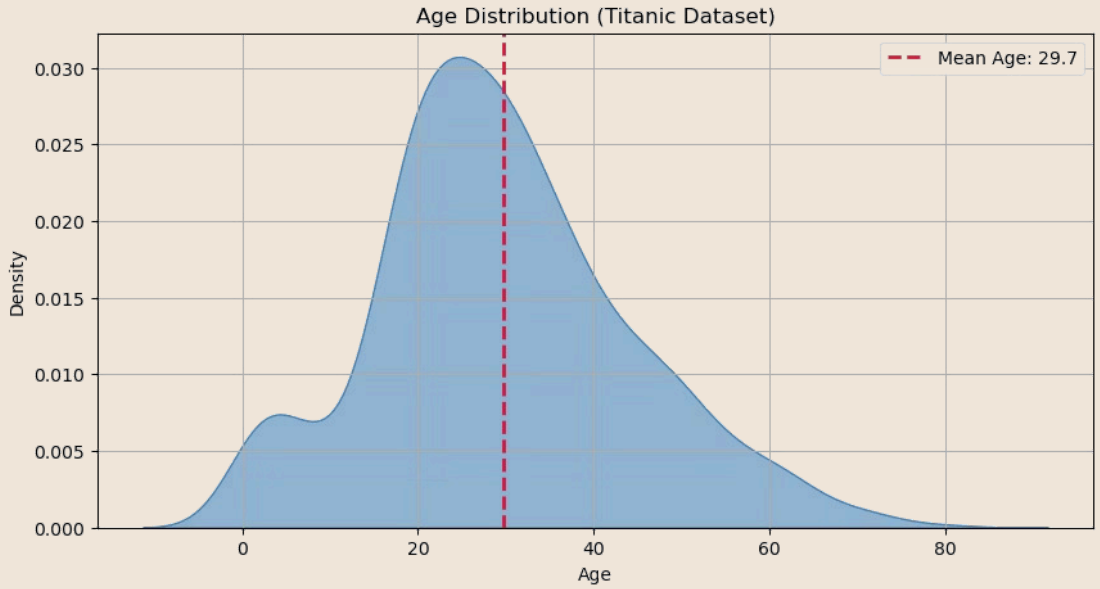
Value that appears most frequently in the sample

Peak of the distribution

Mean

The **arithmetic mean** (or average) is the most commonly used measure of central tendency. It is calculated by summing all values and dividing by the number of observations:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$



Properties:

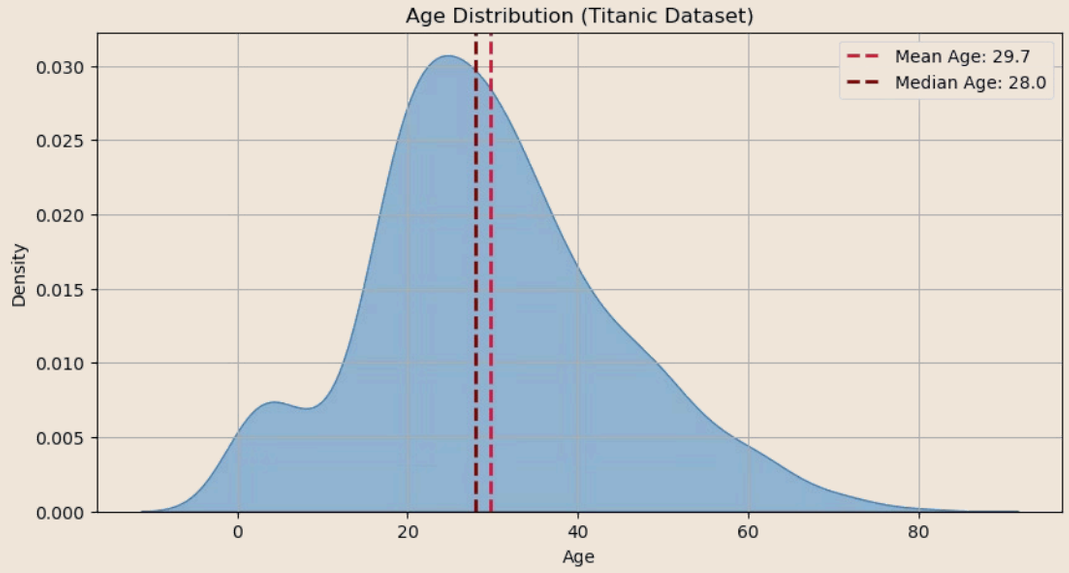
- Uses all data points in the calculation
- Sensitive to outliers and extreme values
- Best suited for symmetric distributions
- Can be influenced by skewed data

Median

The **median** is the middle value when data is arranged in ascending order. It divides the dataset into two equal halves, with 50% of values below and 50% above.

Calculation:

- For odd n: median = middle value [1,3,5] → 5
- For even n: median = average of two middle values [1,3,4,5] → (3+4)/2 =3,5



Properties:

- More robust to outliers than the mean
- Not affected by extreme values
- Better representation for skewed distributions
- Uses positional information rather than actual values

The median is particularly useful when dealing with skewed data or when outliers are present.

Quantiles, Percentiles, and Quartiles

1

Quantiles

These are values that divide an ordered dataset into equal-sized, contiguous subgroups. They are a general term for any division point in a distribution.

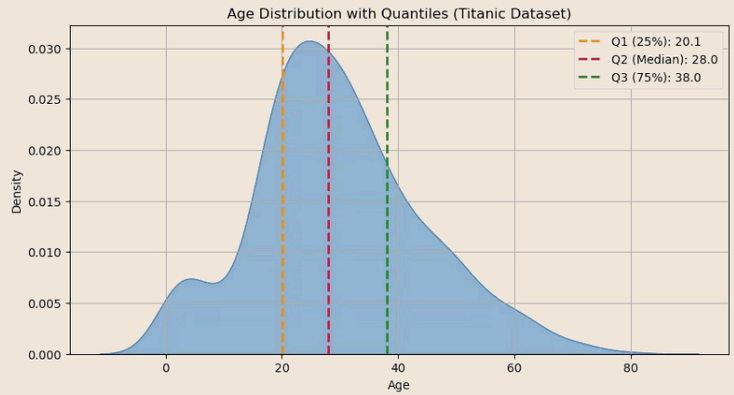
[1,2,3,3,4,5,6,6,7,8,8,9] → [1,2,3,3] [4,5,6,6,7,8,8,9]

$q_{0.25} = 3$ dividing in ¼ + ¾

2

Percentiles

Divide the data into 100 equal parts. The Pth percentile is the value below which P% of observations fall. Used widely in interpreting individual scores relative to a larger group.



3

Quartiles

Divide the data into four equal parts: Q1 (25th percentile), Q2 (the median, 50th percentile), and Q3 (75th percentile). They are crucial for understanding data spread.

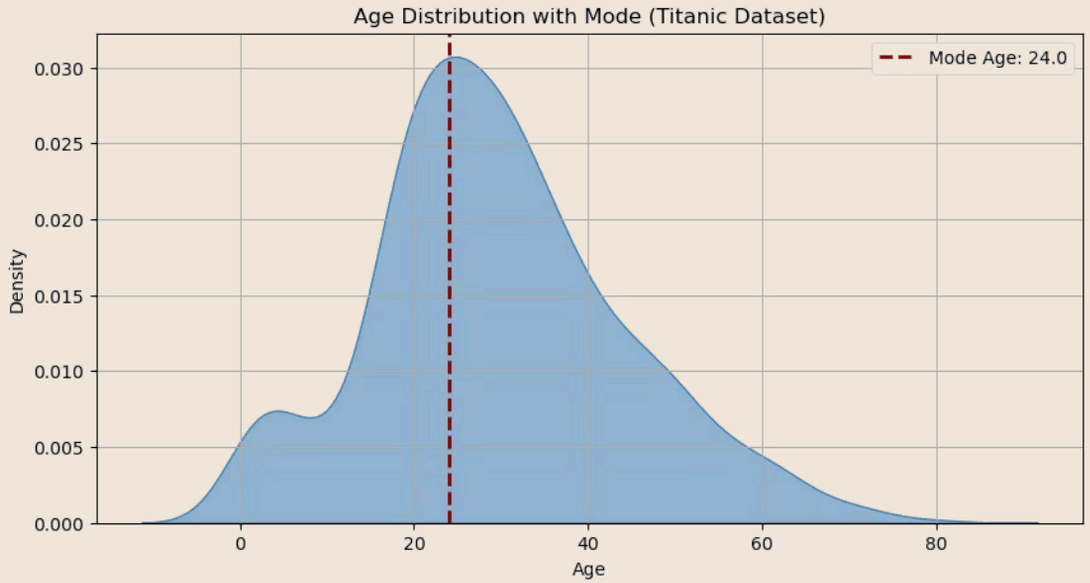
Mode

The **mode** is the value that appears most frequently in the dataset. It represents the peak of the distribution and is the only measure of central tendency that can be used with categorical data.

1

Properties:

- Can be used with any type of data (numerical or categorical)
- Not affected by extreme values
- May not exist or may not be unique
- Useful for identifying the most common category or value



The mode is particularly useful for understanding the most typical or popular value in your dataset.

Measures of Dispersion

Range and IQR

Range: max - min

IQR: Q3 - Q1 (more robust)

Variance

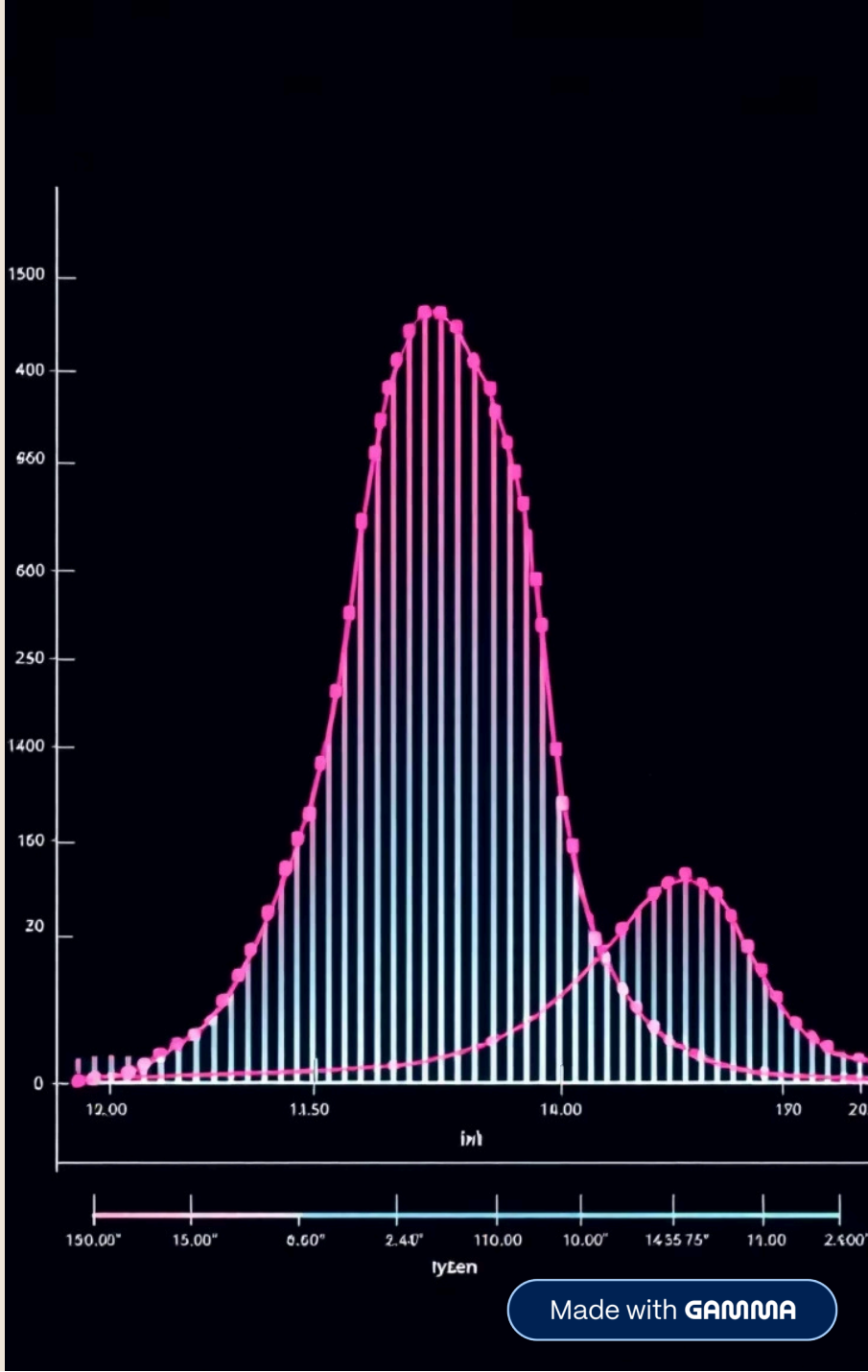
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

Mean of the squared deviations

Standard Deviation

$$s = \sqrt{s^2}$$

Same unit of measurement as the original data



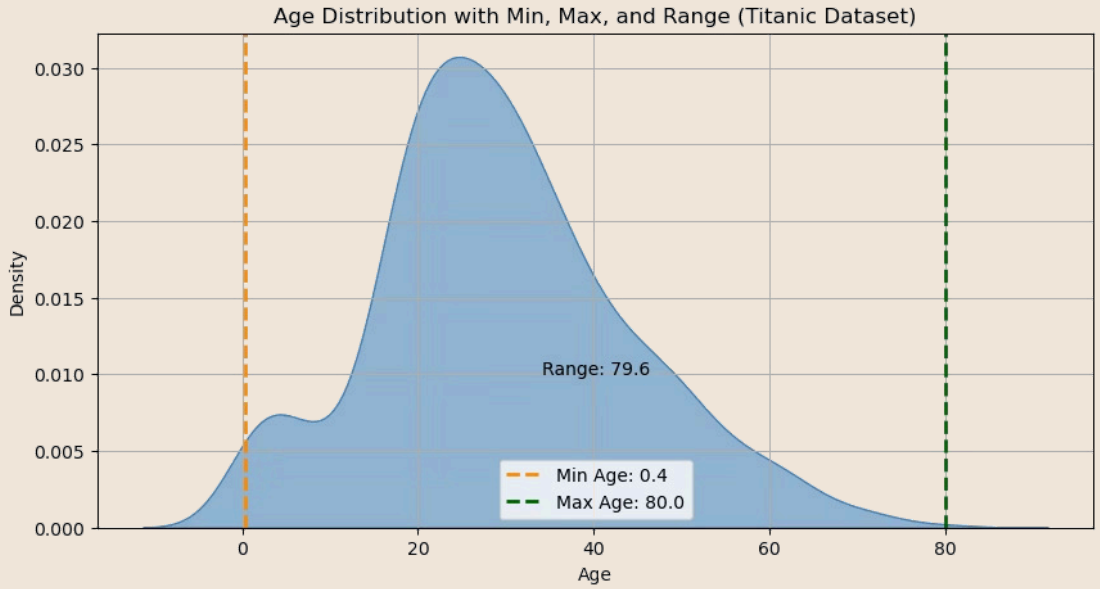
Range

Range is the simplest measure of dispersion, calculated as the difference between the maximum and minimum values in the dataset:

$$\text{Range} = \max(x) - \min(x)$$

Properties:

- Range uses only two extreme values
- Highly sensitive to outliers



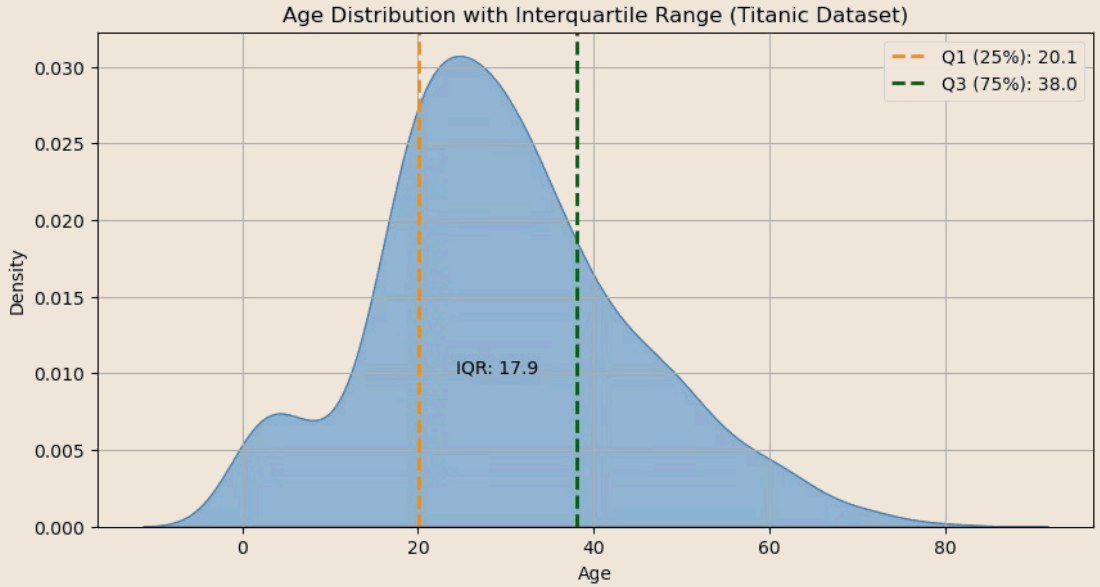
Inter-Quartile Range

Interquartile Range (IQR) is more robust to outliers, measuring the spread of the middle 50% of the data:

$$\text{IQR} = Q_3 - Q_1$$

Properties:

- IQR focuses on the central portion of data
- IQR is more stable and reliable for skewed distributions



Variance

Variance measures how much the data points deviate from the mean. It's calculated as the average of the squared differences from the mean:

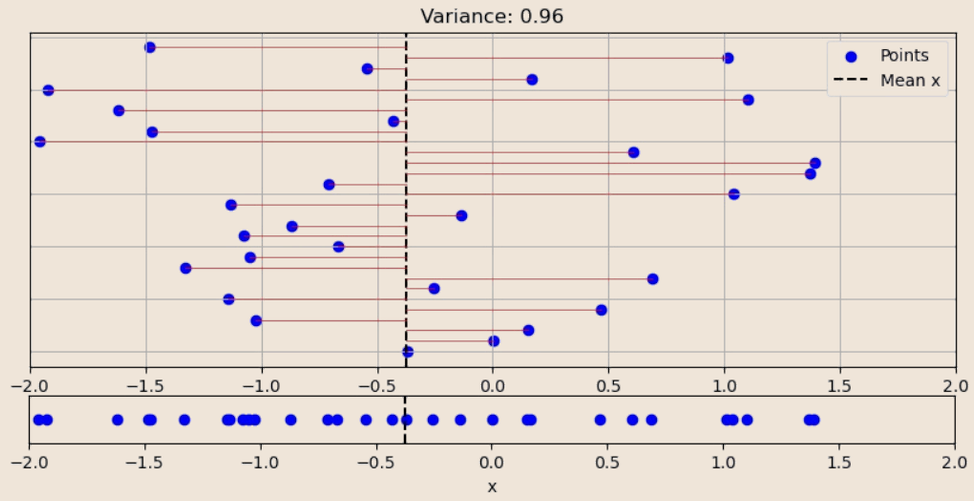
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

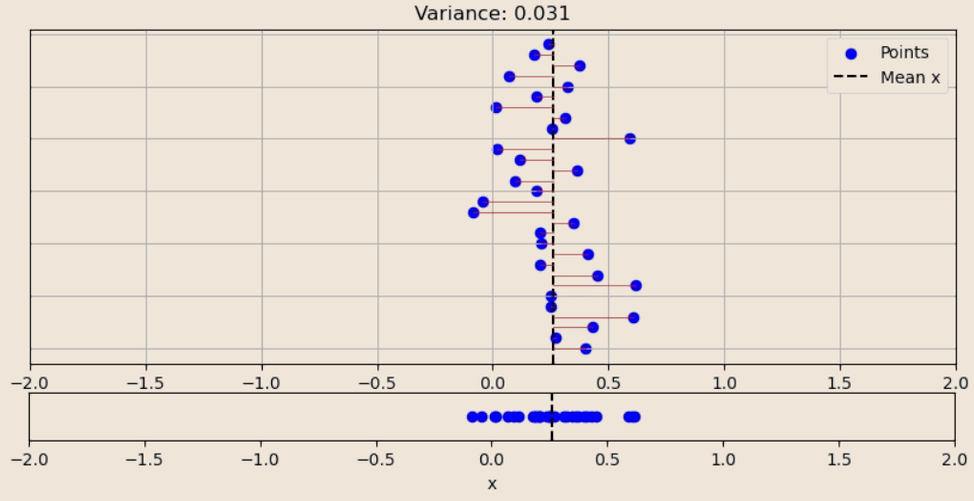
Sample Variance For sample variance (more common), we use n-1 in the denominator (will see better later):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Properties:

- Uses all data points in the calculation
- Sensitive to outliers (squared differences amplify extreme values)
- Units are squared (e.g., cm² for height data)
- Always non-negative
- Larger variance indicates more spread in the data





Standard Deviation

Standard deviation is the square root of the variance, bringing the measure back to the original units of the data:

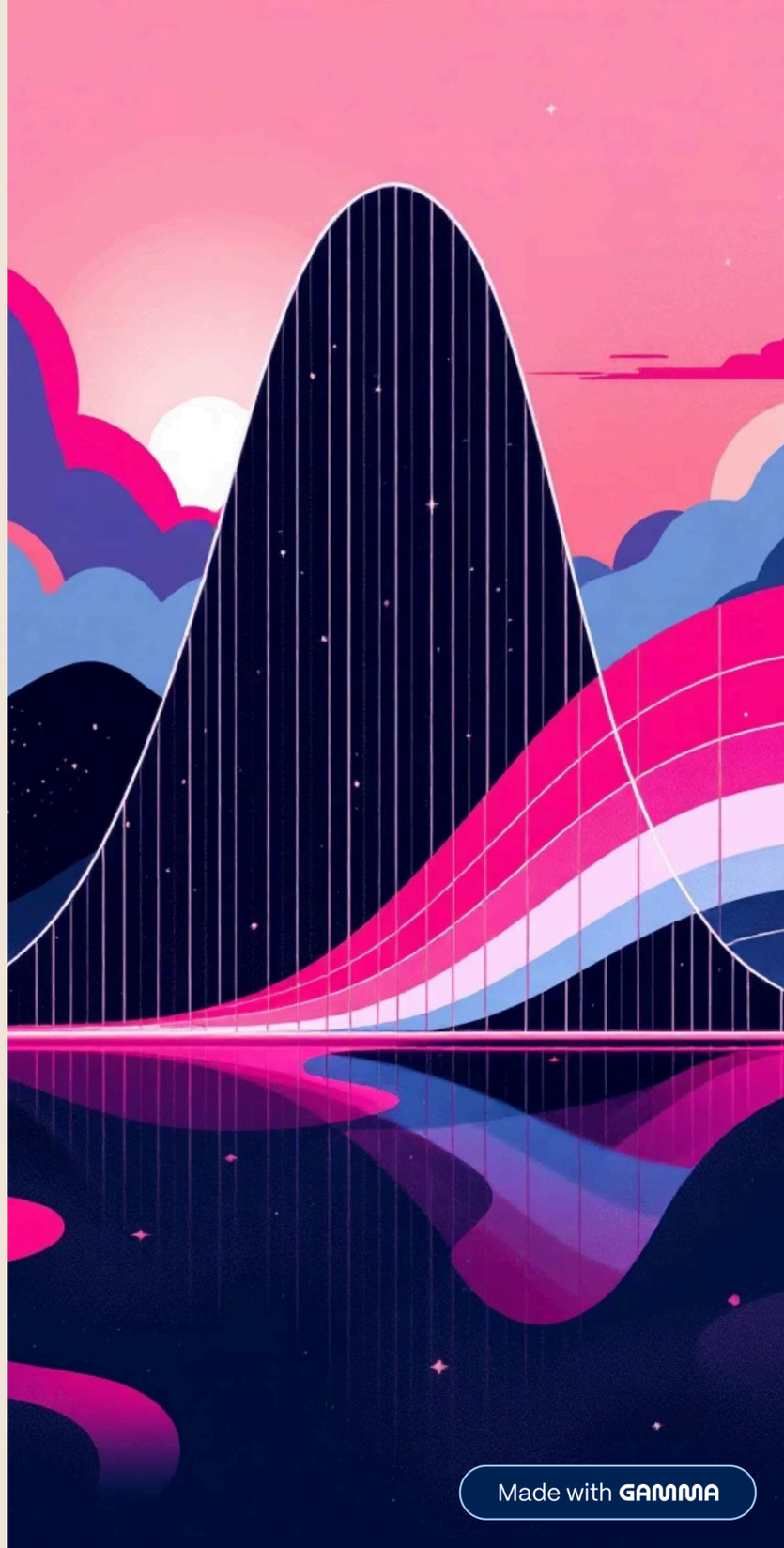
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Properties:

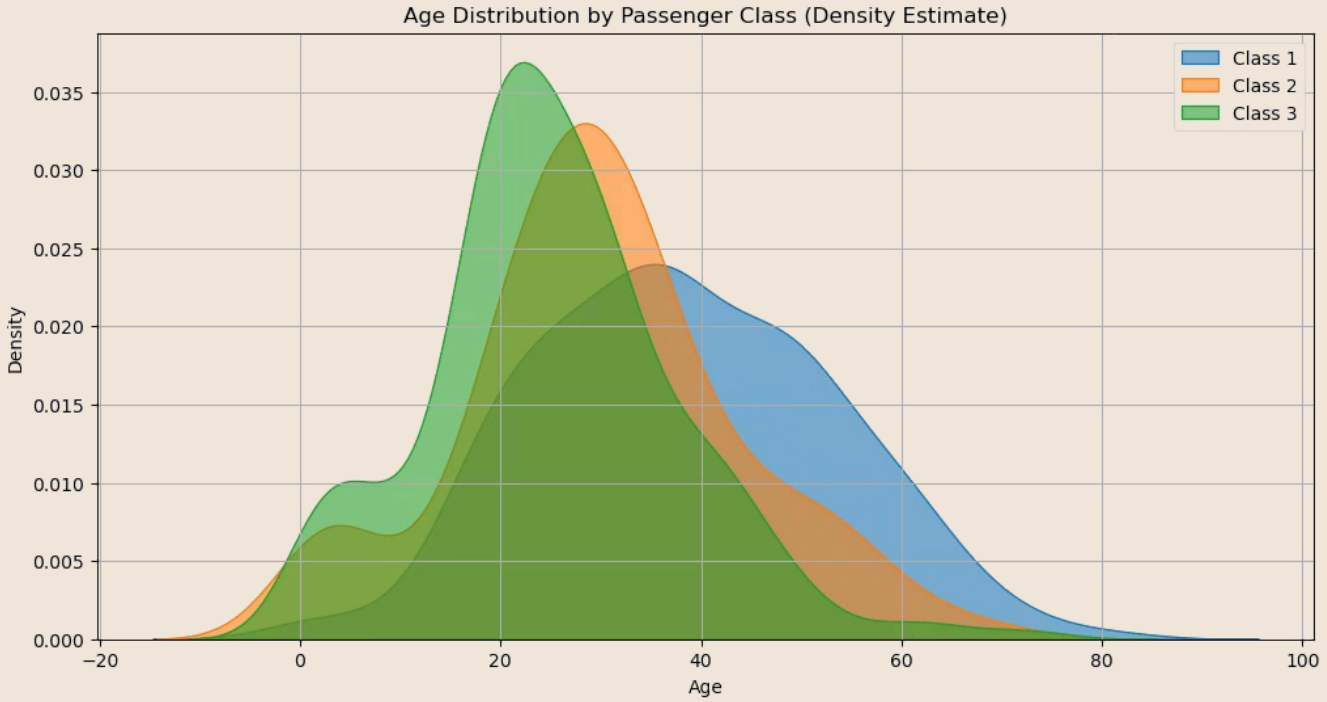
- Same units as the original data (interpretable)
- Most commonly used measure of spread
- Approximately 68% of data falls within 1 standard deviation of the mean (for normal distributions)
- Approximately 95% of data falls within 2 standard deviations
- Sensitive to outliers but less extreme than variance

Interpretation:

- Small standard deviation: data points are close to the mean
- Large standard deviation: data points are spread out from the mean
- Zero standard deviation: all values are identical



Comparing Samples by Variance



Class 1 ages:
Variance: 219.12
Std: 14.80

Class 2 ages:
Variance: 196.03
Std: 14.00

Class 3 ages:
Variance: 156.13
Std: 12.49

Data Normalization

The observed data dispersion indicators depend on the nature of the data and their unit of measurement. Normalization techniques make data based on different units of measurement comparable.

Normalization between 0 and 1:

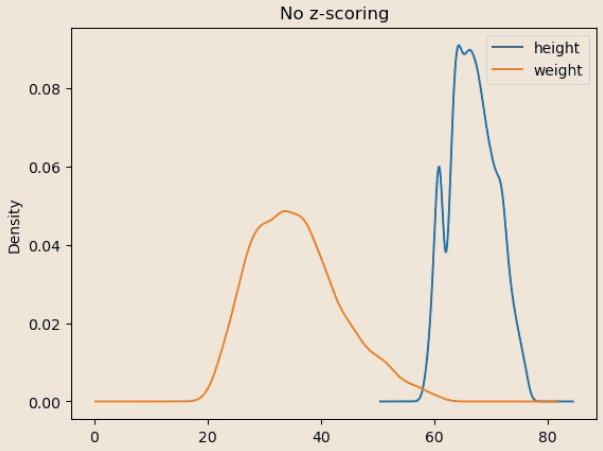
$$x_{norm} = (x - x_{min})/(x_{max} - x_{min})$$

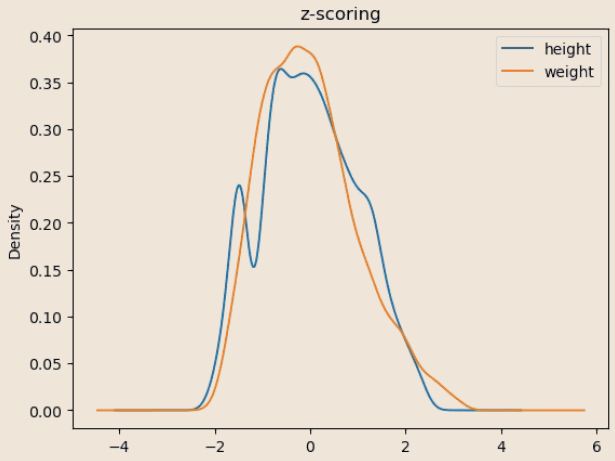
Normalization between -1 and 1:

$$x_{norm} = (x_{max} + x_{min} - 2 \cdot x)/(x_{max} - x_{min})$$

Standardization (z-scoring):

$$z_i = \frac{x_i - \overline{x}}{s_x}$$





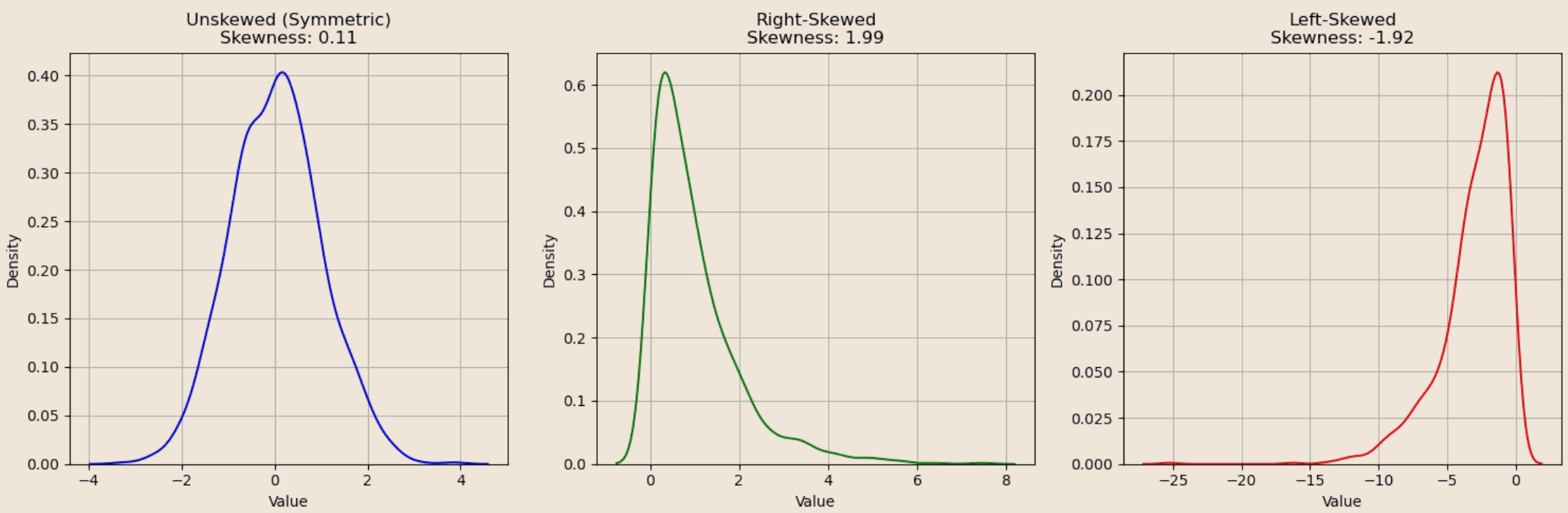
Shape Indicators: Skewness

Skewness is an indicator of the “imbalance” to the left (negative value) or to the right (positive value) of a data sample with respect to the central value. The formula for skewness is as follows:

$$\sum_{i}^{n} \frac{(x_i - \overline{x})^3}{n \cdot s_x^3}$$

The skewness values will be:

- **Negative** if the distribution is skewed to the left;
- **Positive** if the distribution is skewed to the right;
- **Close to zero** in the case of unskewed distributions.



Kurtosis

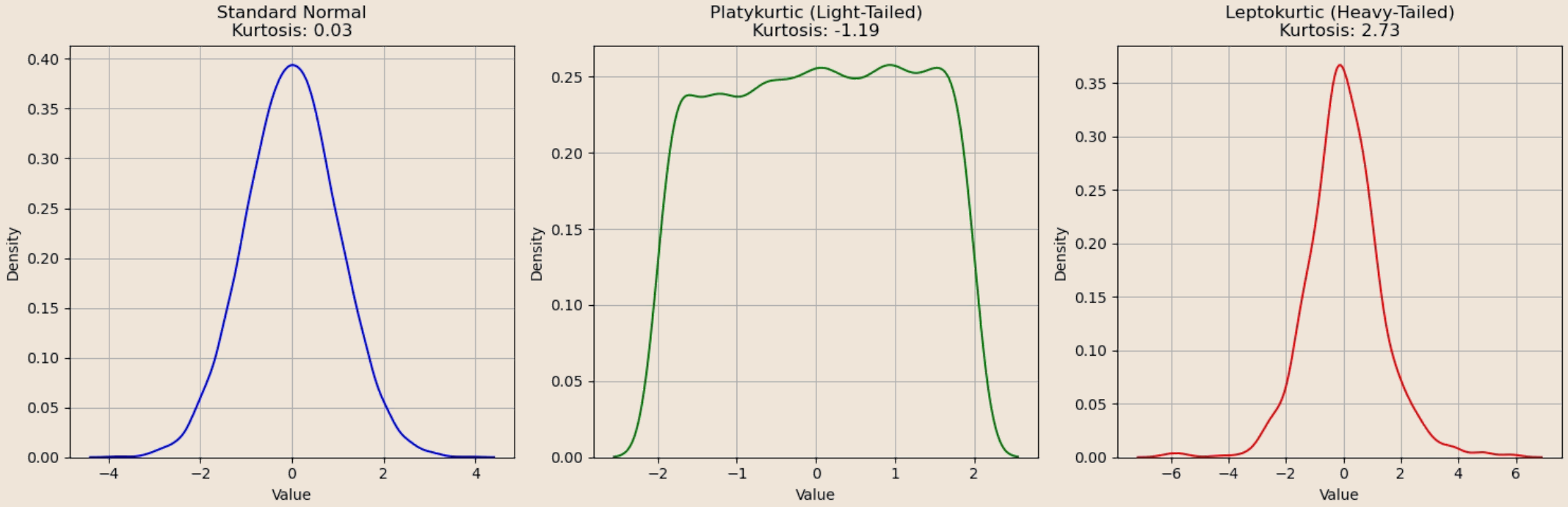
The kurtosis index measures the “thickness” of the tails of a density distribution. It is defined as follows:

$$K = \frac{1}{n}\sum_{i=1}^{n}(\frac{X_i - \overline{X}}{s})^4 - 3$$

The index is interpreted as follows:

- If it is greater than zero, the distribution is *leptokurtic*, i.e., more “peaked” than a Normal distribution;
- If it is less than zero, the distribution is *platykurtic*, i.e., more “flat” than a Normal distribution;
- If it is equal to zero, the distribution is *mesokurtic*, i.e., the tails are similar to those of a normal distribution.

We will see what a normal distribution is later in the course.



Statistical Summary

When we compute values like **mean**, **median**, **mode**, **minimum**, **maximum**, **range**, **standard deviation**, **variance**, and **interquartile range**, we're building what is known as a **statistical summary** of a dataset. These measures help us understand the **central tendency**, **spread**, and **shape** of a distribution—whether it's symmetric, skewed, concentrated, or dispersed.

In practice, we can obtain many of these summary statistics quickly using the `.describe()` method in **pandas**:

```
age.describe()
```

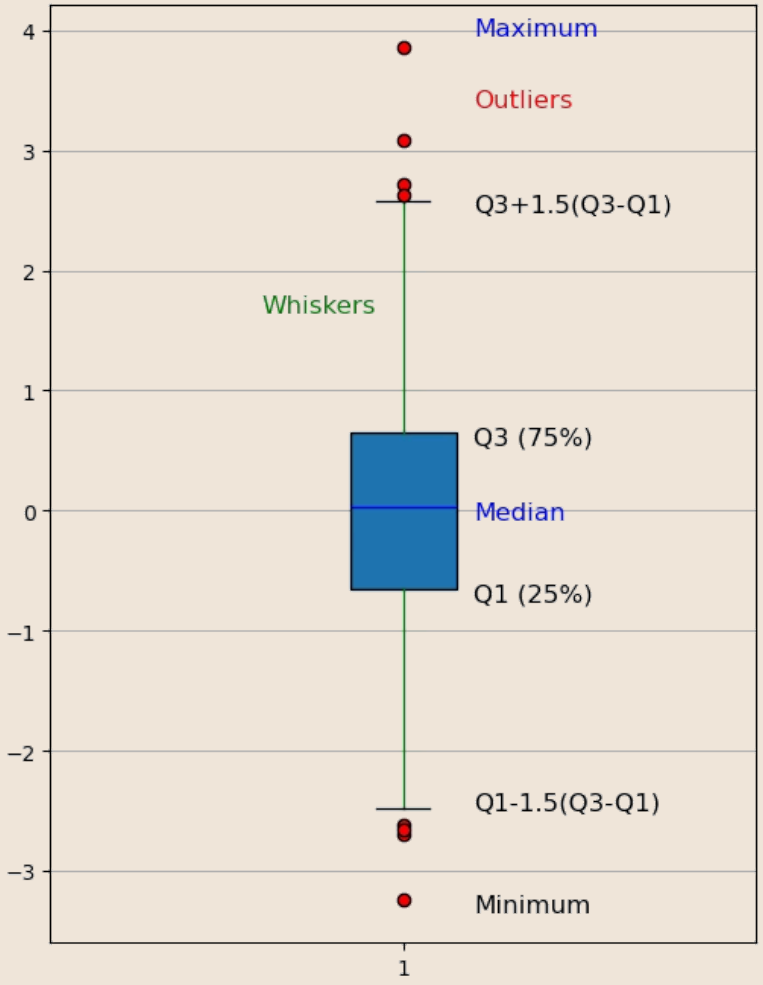
```
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
25%       20.125000
50%       28.000000
75%       38.000000
max       80.000000
Name: Age, dtype: float64
```

Boxplot

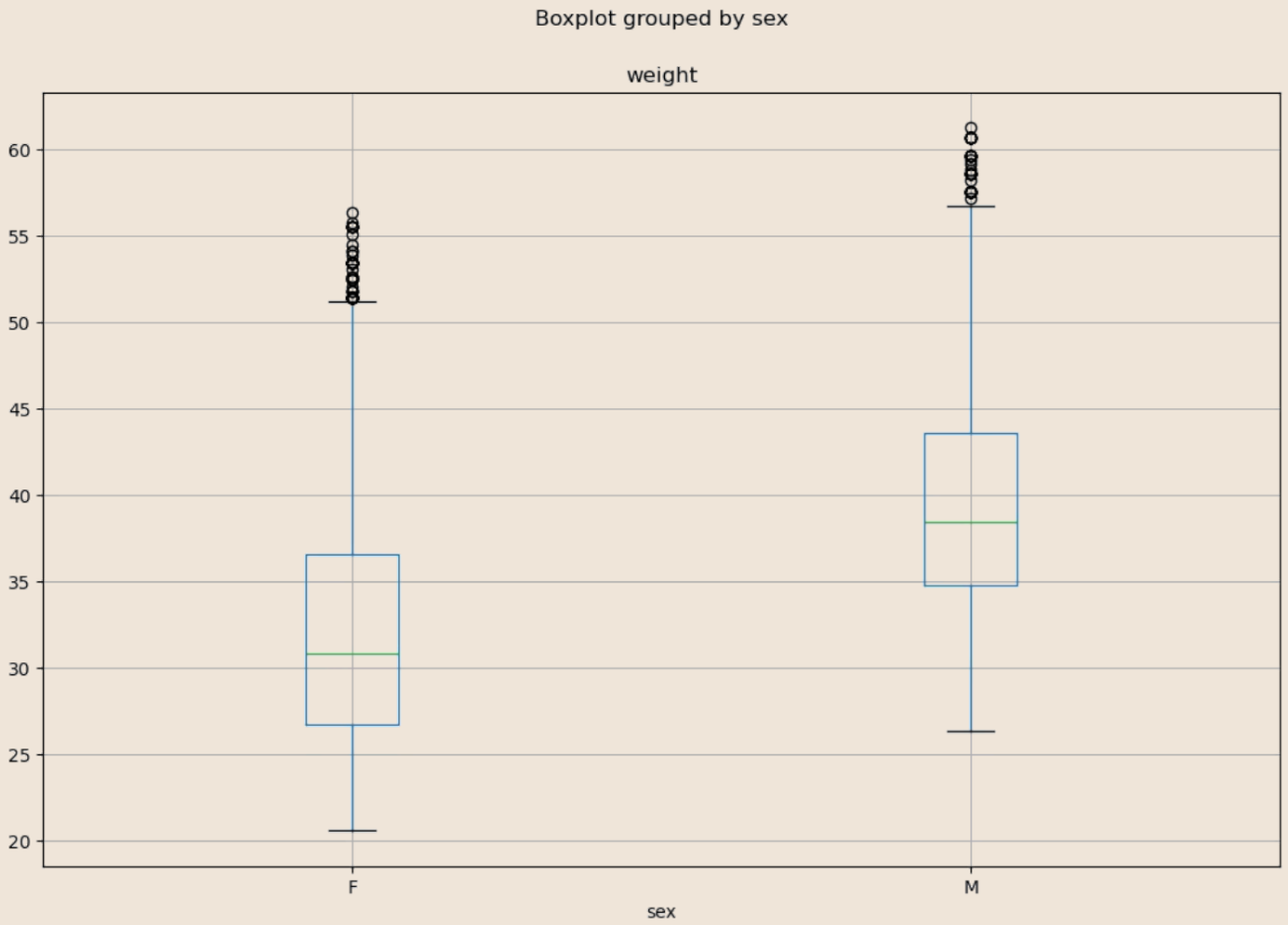
Boxplots are a compact visualization method for representing certain descriptive characteristics of the data under analysis.

In particular, given a sample, a boxplot can effectively represent the following statistics:

- Median value;
- First and third quartile;
- Depending on the version, the whiskers can denote:
  - A projection of first and third quartile;
  - Minimum and maximum;



Comparing samples with boxplots



Boxplots can be useful to compare samples. For example, the boxplots on the left compare the distributions of weights between men and women.

Conclusions and Next Steps

We Have Explored:

- Absolute and relative frequencies;
- Empirical cumulative distribution Function;
- Histograms;
- Summary Statistics;
- Statistical Summary;
- Boxplots.

In next lectures, we will look at probability for data analysis.

References

- Chapter 2 of: Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.
- Chapter 3 of: Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.