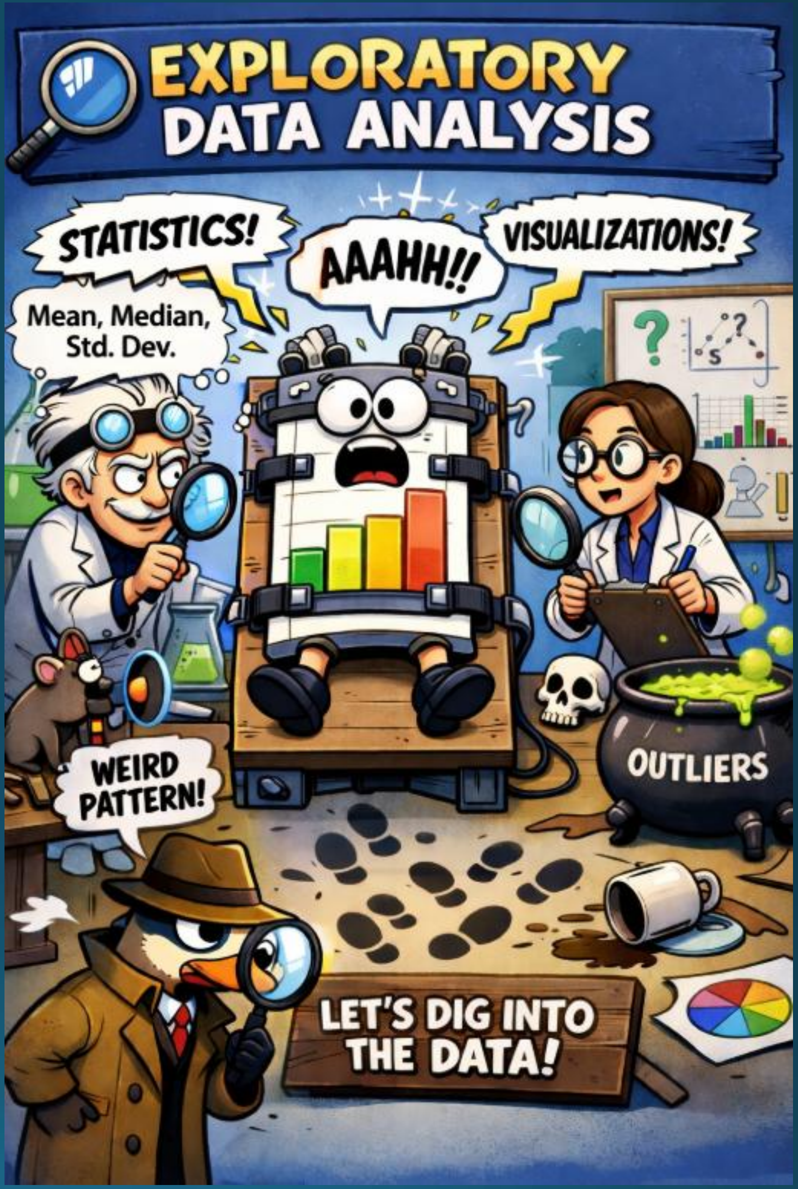
Data Analysis and Knowledge Discovery

Lecture 3

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

Sergiu Limboi, PhD Teaching Assistant

Motto: “Before building models, understand the data.”



Exploratory Data Analysis & Data Understanding

AGENDA

- Warm-Up
- What is Exploratory Data Analysis (EDA)
- Types of EDA
- Visualization in Data Analysis
- Detecting patterns
- Detecting anomalies
- Industry case study
- Teamwork time
- EDA best practices & mistakes
- Key Takeaways

FACULTATEA DE MATEMATICA SI INFORMATICA • BABEŞ-BOLYAI • ROMÂNIA

Warm-Up

Faculty of Mathematics and Computer Science

Warm-Up

Go to www.menti.com and enter the code **1636 7869**

or use the QR code

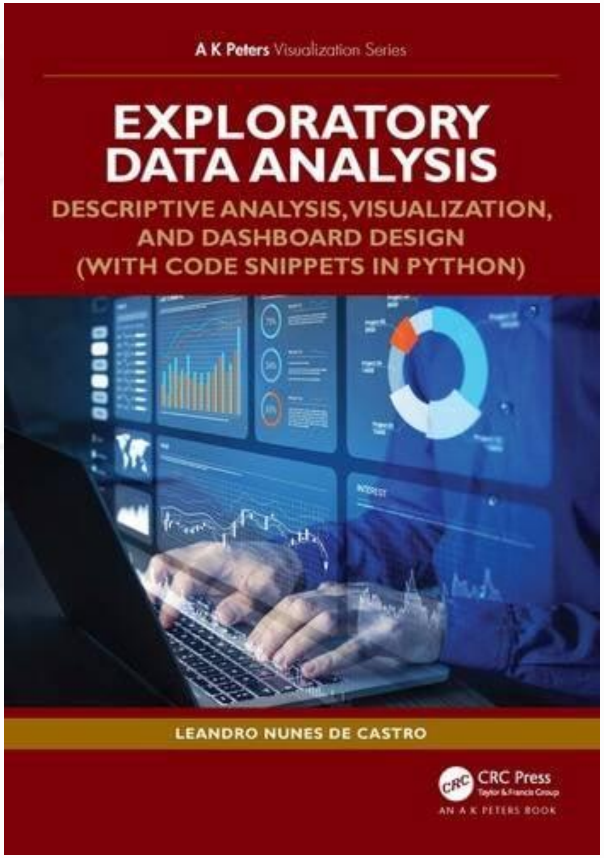
What is Exploratory Data Analysis (EDA)

Faculty of Mathematics and Computer Science

What is Exploratory Data Analysis (EDA)

- It is a preliminary step in Data Analysis to:
  - Gain better understanding of the data set
  - Summarize main characteristics of the data
  - Uncover relationships between variables
  - Extract important variables

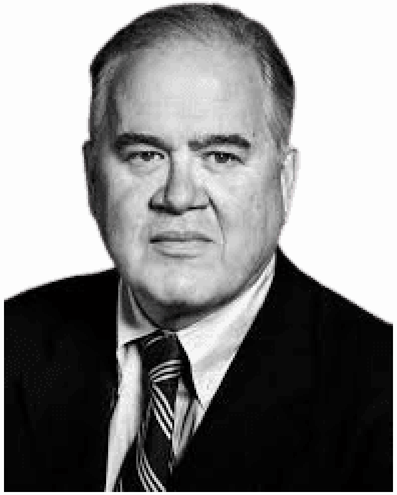
- Goal:
  - Understand the structure of the data
  - Detect anomalies
  - Discover patterns
  - Generate hypotheses
  - Validate assumptions


A K Peters Visualization Series
EXPLORATORY DATA ANALYSIS
DESCRIPTIVE ANALYSIS, VISUALIZATION, AND DASHBOARD DESIGN (WITH CODE SNIPPETS IN PYTHON)
LEANDRO NUNES DE CASTRO
CRC Press
Taylor & Francis Group
AN A K PETERS BOOK

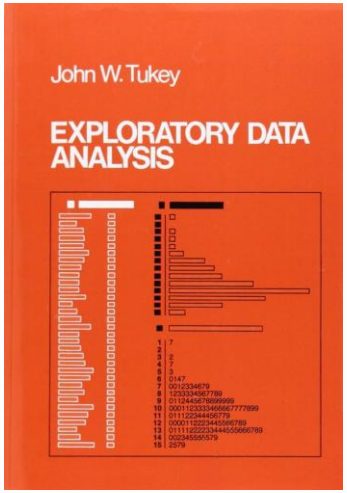
Exploratory Data Analysis (EDA)

- What does the dataset look like?
- What distribution do variables have?
- Are there hidden patterns?
- What are the characteristics which have the most impact on the business?
- Are there anomalies?

Exploratory Data Analysis (EDA)

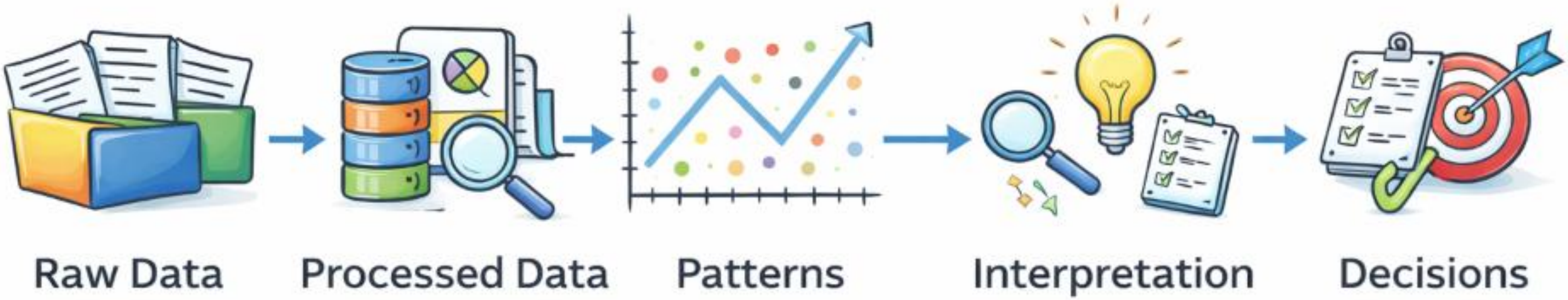


- Historical context
  - EDA was popularized by **John Tukey (1977)**
  - Tukey was an American mathematician and statistician who revolutionized Data Analysis.



  - His legacy:
    - EDA
    - Data visualization tools (e.g., he created the box plot)
    - Statistical methods
    - Fast Fourier Transform (FFT)→ algorithm for digital signal processing

The role of EDA in Knowledge Discovery



The role of EDA in Knowledge Discovery

- EDA is where **patterns start to appear**.

Stage	Goal
Data preprocessing	Make data usable
EDA	Understand patterns
Modelling	Predict
Interpretation	Knowledge

Exploratory Data Analysis (EDA)

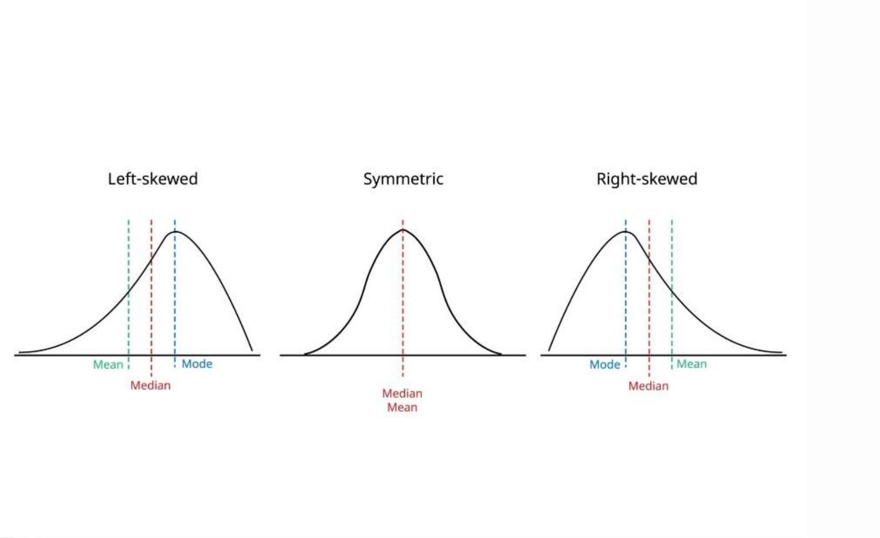
- “Exploratory data analysis is detective work.”

- What a data detective looks for:
  - Clues in the data
  - Suspicious values
  - Missing evidence
  - Hidden relationships
  - Understanding the story

Exploratory Data Analysis (EDA)

Customer	Age	Income	Country	Purchased
1	25	3000 RON	RO	Yes
2	40	800 $	USA	No

- Distribution → how values are spread
    - Example:
        - Low salaries → many
        - High salaries → few



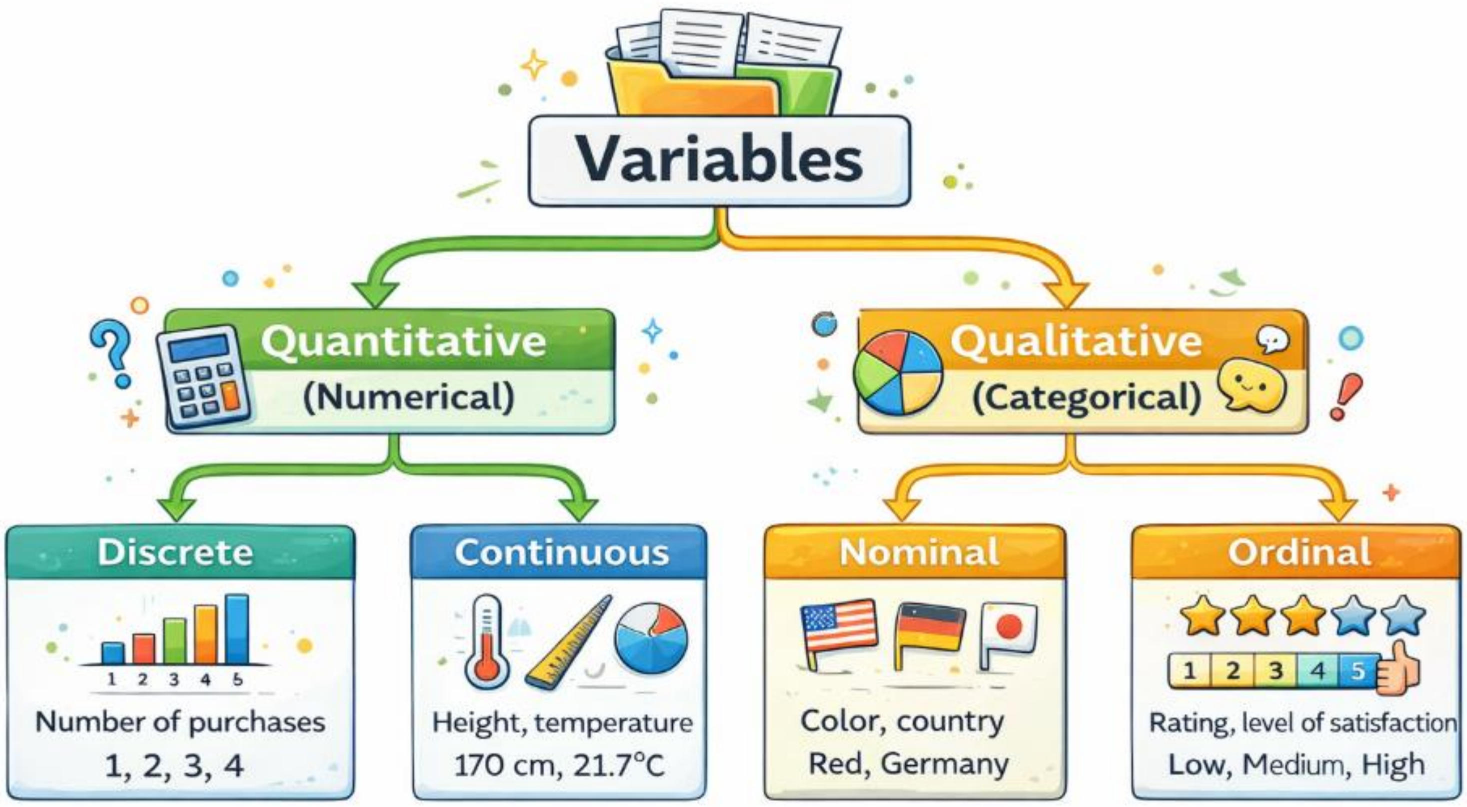
        - This often produces skewed distributions

Exploratory Data Analysis (EDA)

- Variable Types
  - Age → numerical
  - Income → numerical
  - Purchased → binary
  - Country → categorical

- Qualitative vs. quantitative variables
  - Gender?
  - Country?
  - Temperature?
  - Income?
  - Payment method?
  - Number of purchases?

Exploratory Data Analysis (EDA)



Variables

Quantitative (Numerical)

Qualitative (Categorical)

Discrete
Number of purchases
1, 2, 3, 4

Continuous
Height, temperature
170 cm, 21.7°C

Nominal
Color, country
Red, Germany

Ordinal
Rating, level of satisfaction
Low, Medium, High

Exploratory Data Analysis (EDA)

- Types of quantitative variables:
  - Discrete variables → A **discrete variable** can take **only specific separate values**, usually **countable integers**.
    - Groups of students (e.g., 224, 221)
    - Number of clicks
    - Number of purchases
    - Number of cars (e.g., cannot be 2.6)
  - Continuous variables → measured values that can take and real value withing a range
    - Weight
    - Temperature
    - time

Exploratory Data Analysis (EDA)

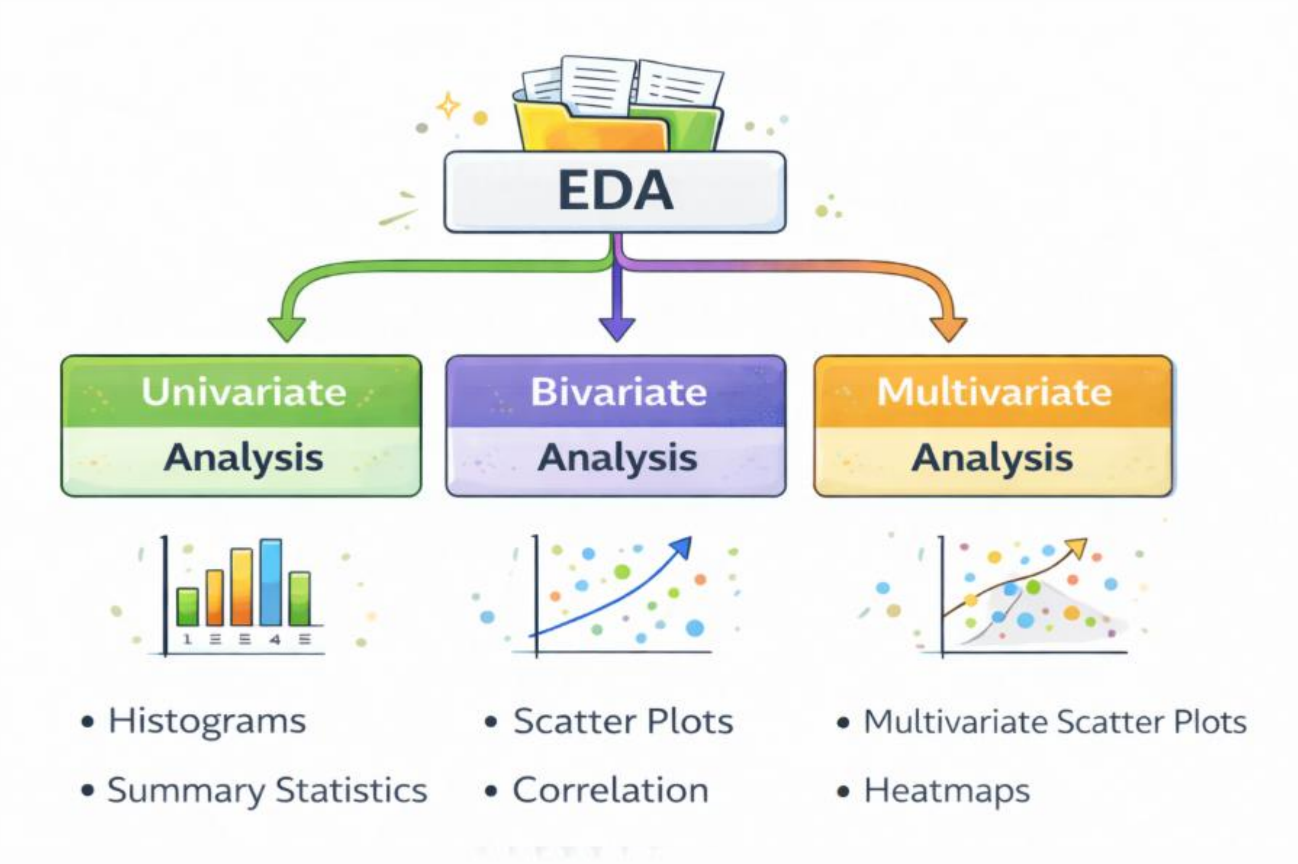
- Relationships→ Do variables influence each other?
  - Examples:
    - Education → Income
    - Temperature → Energy usage
    - Skills → Level in the company

- Anomalies → Detect unusual values
  - Example:
    - Passenger fare: 7,8,10,9,11,8,7,512
    - 512= anomaly

Types of EDA

Faculty of Mathematics and Computer Science

Types of EDA



EDA

Univariate Analysis

- Histograms
- Summary Statistics

Bivariate Analysis

- Scatter Plots
- Correlation

Multivariate Analysis

- Multivariate Scatter Plots
- Heatmaps

Univariate analysis

- It studies one variable at a time.
- Goal: understand its distribution

- Summary statistics:
  - Mean
  - Median
  - Variance
  - Standard deviation
  - Minimum
  - Maximum
  - Quartiles

Univariate analysis- mean & median

Exam scores
50
60
65
70
70
75
80
90

4-th

5-th

- Mean→ the average of the exam scores.
- Mean = $\frac{(50+60+65+70+70+75+80+90)}{8} = 70$
- Median → it is the middle value
  - Order the data
  - Odd number of data → select the middle number
    - Odd dataset (1,3,3,6,7,8,9)→ middle value is 6
  - Even number of data -> take the two middle numbers, add them, and divide them by 2.
- Median =$\frac{70+70}{2} = 70$

Univariate analysis- minimum & maximum

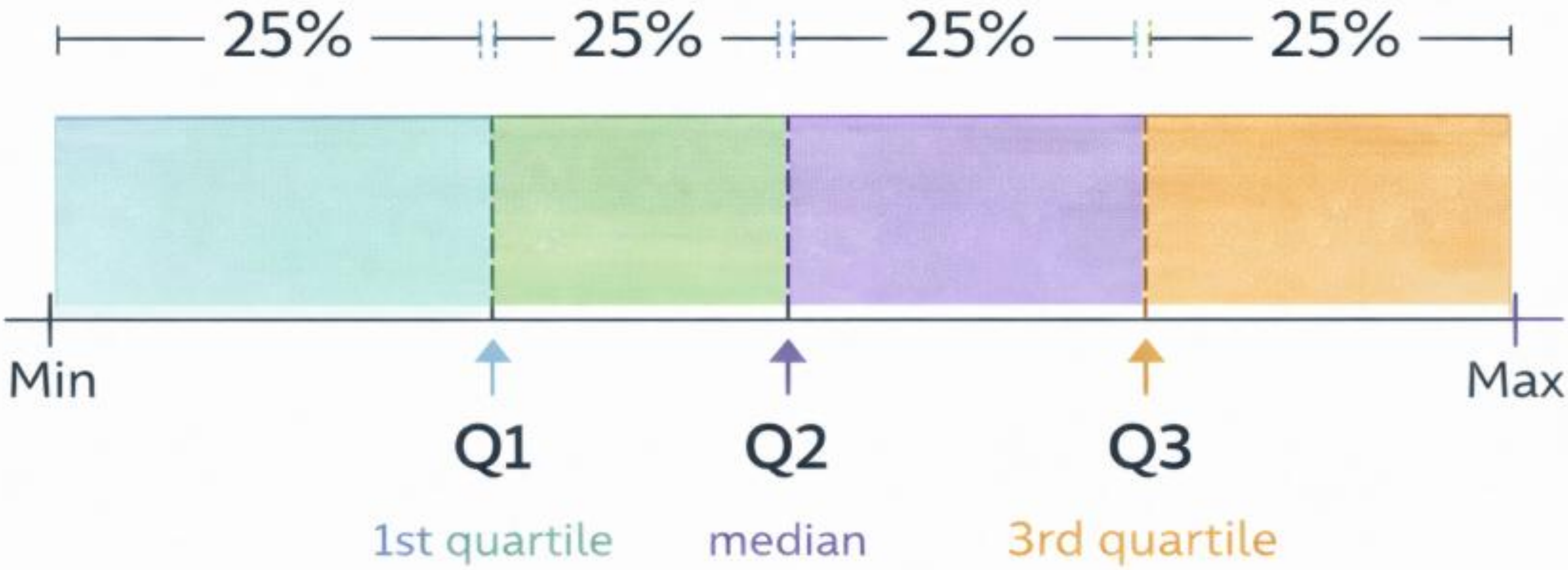
Exam scores
50
60
65
70
70
75
80
90

- Minimum → smallest value
  - Minimum =50

- Maximum → largest value
  - Maximum = 90

Univariate analysis- Quartiles

- Quartiles divide the data into 4 parts



Univariate analysis- Quartiles

Exam scores
50
60
65
70
70
75
80
90

- Q2 is the median (Q2=70)

- Q1 take the lower half and computes the median
  - Lower half (50,60,65,70)
  - Median = $\frac{60+65}{2} = 62.5$
  - Q1=62.5

- Q3 takes the upper half and computes the median
  - Upper half (70,75,80,90)
  - Median = $\frac{75+80}{2} = 77.5$
  - Q3=77.5

- 25% of the values are below 62.5
- 50% of the values are below 70
- 75% of the values are below 77.5

Univariate analysis- Variance

- Variance measures how spread out the values are around the mean.
- Formula:
  - $\sigma^2 = \frac{\sum(xi\_\mu)^2}{n}$
  - the symbol **μ** represents the **mean (average) of the dataset**. (here the mean is 70)

Value	x_i-70	$(x_i - 70)^2$
50	-20	400
60	-10	100
65	-5	25
70	0	0
70	0	0
75	5	25
80	10	100
90	20	400

Univariate analysis- Variance

Exam scores
50
60
65
70
70
75
80
90

- Variance = $\frac{400+100+25+0+0+25+100+400}{8} = 131.25$
- If variance is 0 → no variability (all values are identical)
- Variance = 131.25 indicates that the scores are moderately spread around the mean of 70.

Univariate analysis- Standard deviation

Exam scores
50
60
65
70
70
75
80
90

- Standard deviation is the square root of the variance.
- $\sigma = \sqrt{131.25} \approx 11.46$
- Interpretation: On average, scores are about **11.46 points away from the mean.**
- Variance and standard deviation both measure **how spread out the data is around the mean.**
- The key difference is **how they express that spread.**

- variance = 131.25 points^2
- standard deviation = $\sqrt{131.25} \approx 11.46$ points

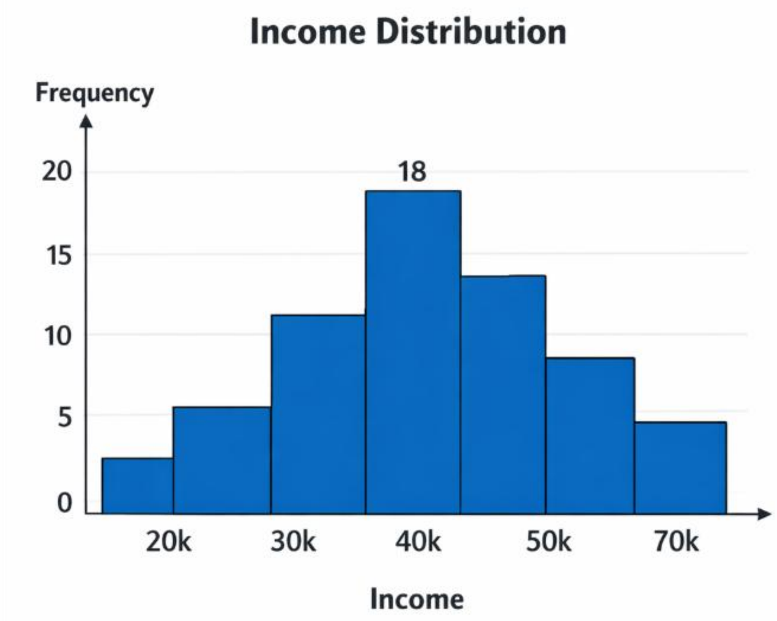
Univariate analysis- summary statistics in Python

```
import pandas as pd

data = {
        "exam_score": [50, 60, 65, 70, 70, 75, 80, 90]
}

df = pd.DataFrame(data)
df["exam_score"].describe()
```

```
count     8.000
mean     70.000
std      11.456
min      50.000
25%      62.500
50%      70.000
75%      77.500
max      90.000
```

Univariate analysis

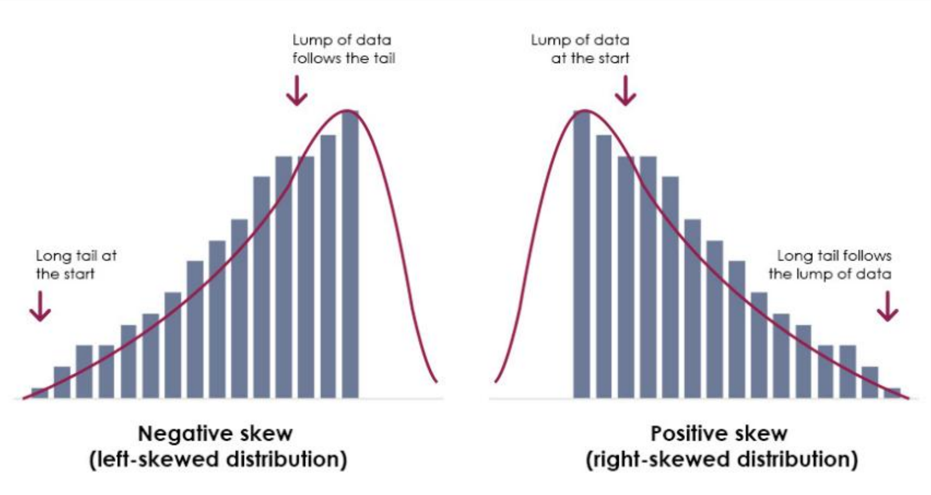
- Histogram→ visualizes distribution

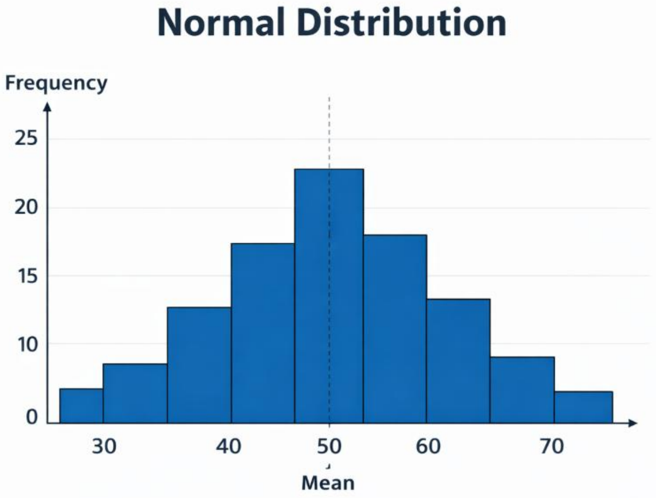


- Most people have incomes around 40k-50k
- Fewer people earn between 20k-30k

- A histogram helps us quickly understand:
  - where **most values are concentrated**
  - how **spread out the values are**
  - whether **extreme values are common or rare**

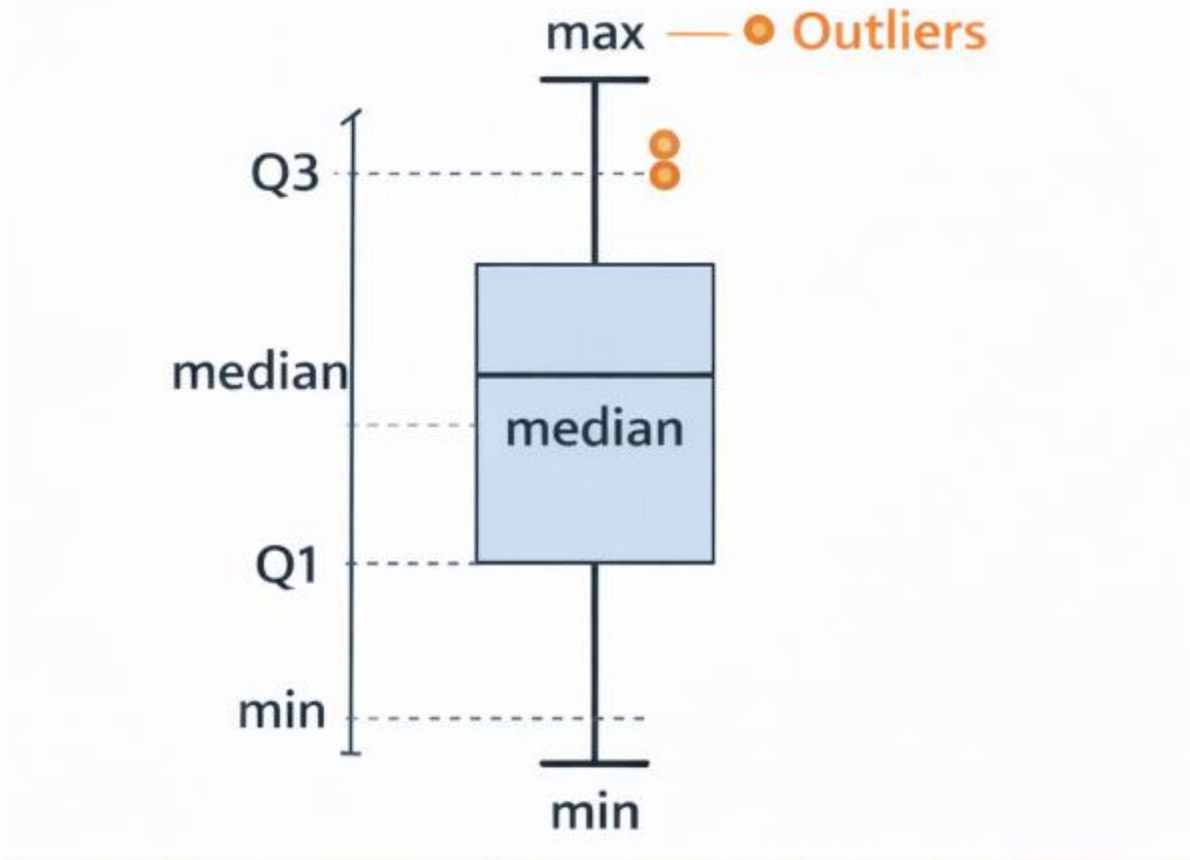
Univariate analysis





Univariate analysis

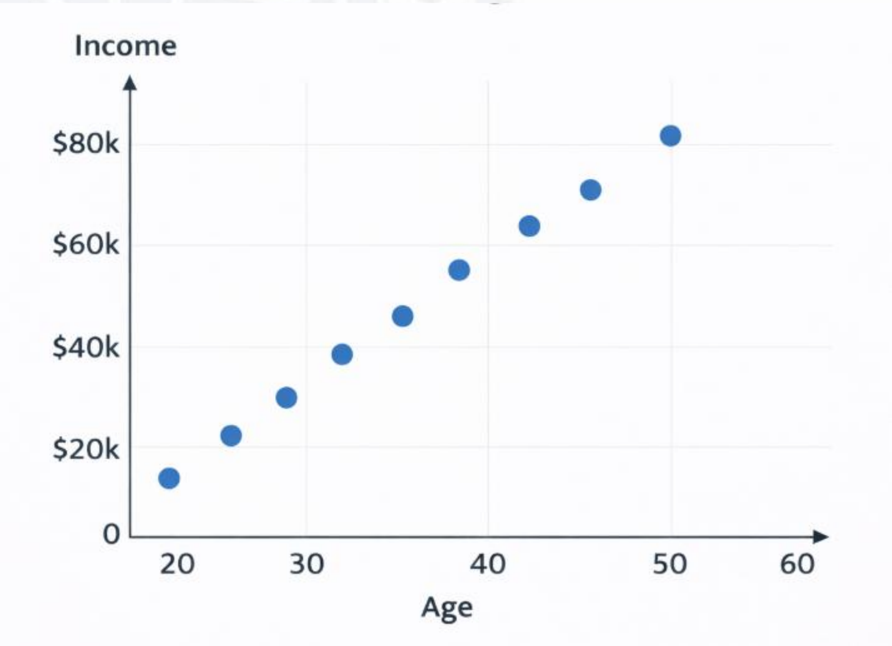
- Boxplot→ it is used to detect outliers.



Bivariate analysis

- It studies the relationship between two variables.
- Example: age vs. income
- Scatter plots show the relationship between two variables.

- Positive relationship
- Older people earn more money.

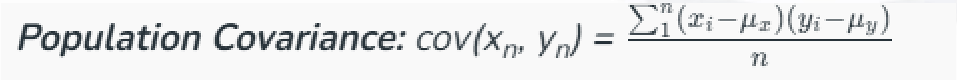


Bivariate analysis

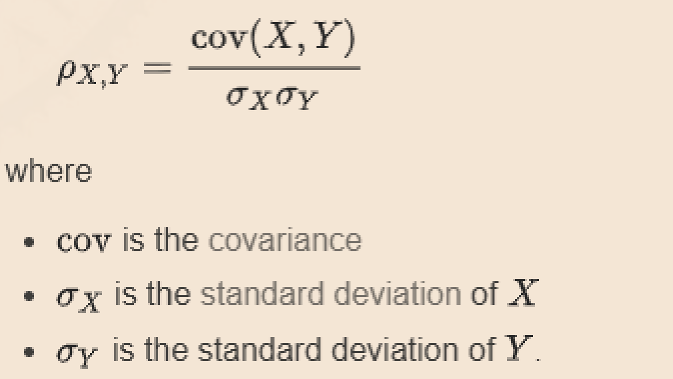
- Correlation measures the relationship strength.
- The strength can be a value from [-1,1]
  - 1→ strong/perfect positive correlation
  - 0→ no relationship
  - -1→weak/perfect negative correlation

- For example:
  - Lung cancer→ smoking
  - Rain→ Umbrella

- Pearson' s correlation coefficient



Population Covariance: $cov(x_n, y_n) = \frac{\sum_1^n (x_i-\mu_x)(y_i-\mu_y)}{n}$



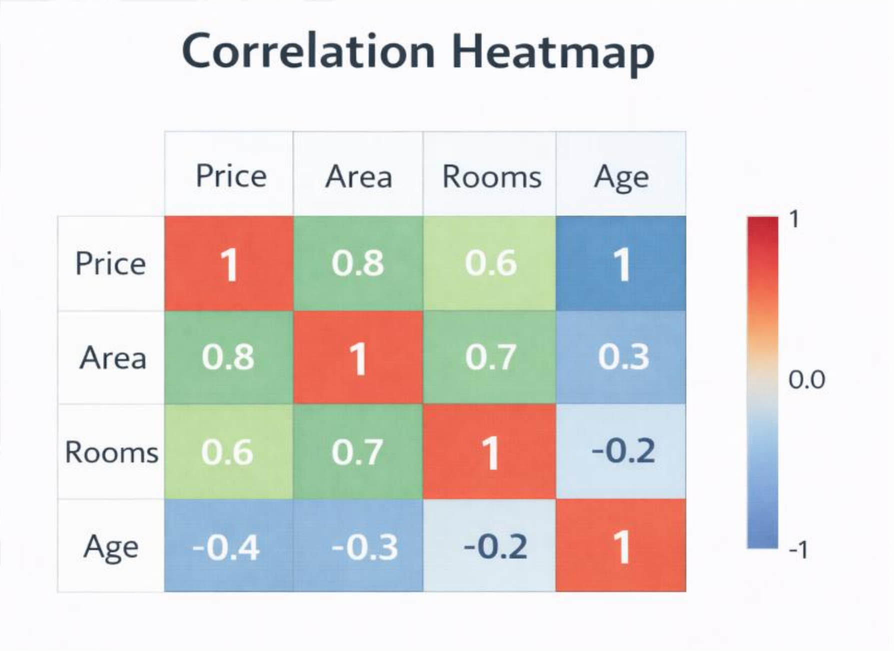
$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

Multivariate analysis

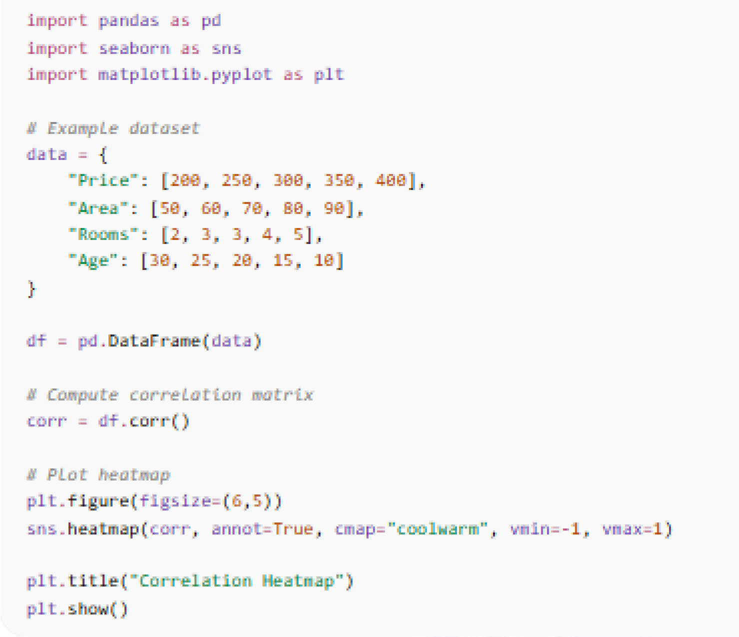
- It studies **multiple variables simultaneously**.
- Example: house price dataset
  - Price strongly related to **Area**.



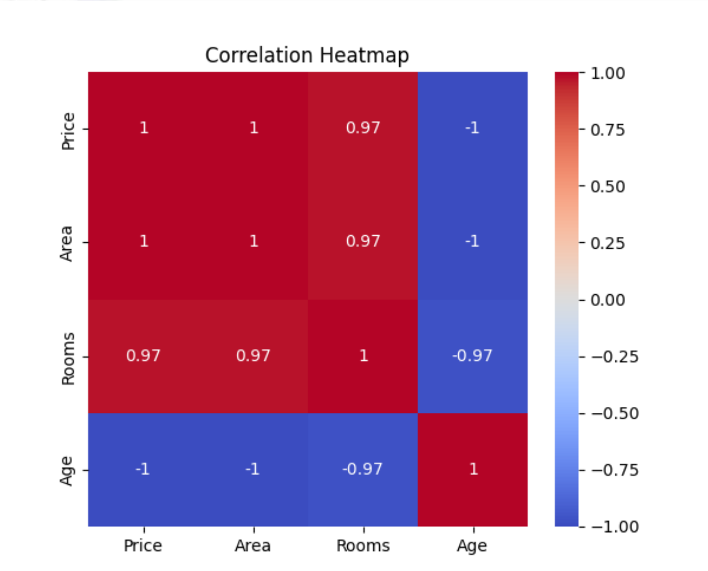
Correlation Heatmap

	Price	Area	Rooms	Age
Price	1	0.8	0.6	1
Area	0.8	1	0.7	0.3
Rooms	0.6	0.7	1	-0.2
Age	-0.4	-0.3	-0.2	1

Multivariate analysis- Correlation heatmap in Python



```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Example dataset
data = {
    "Price": [200, 250, 300, 350, 400],
    "Area": [50, 60, 70, 80, 90],
    "Rooms": [2, 3, 3, 4, 5],
    "Age": [30, 25, 20, 15, 10]
}

df = pd.DataFrame(data)

# Compute correlation matrix
corr = df.corr()

# Plot heatmap
plt.figure(figsize=(6,5))
sns.heatmap(corr, annot=True, cmap="coolwarm", vmin=-1, vmax=1)

plt.title("Correlation Heatmap")
plt.show()
```



Multivariate analysis

- Correlation ≠ causation

- Causation means **one variable directly influences another**.

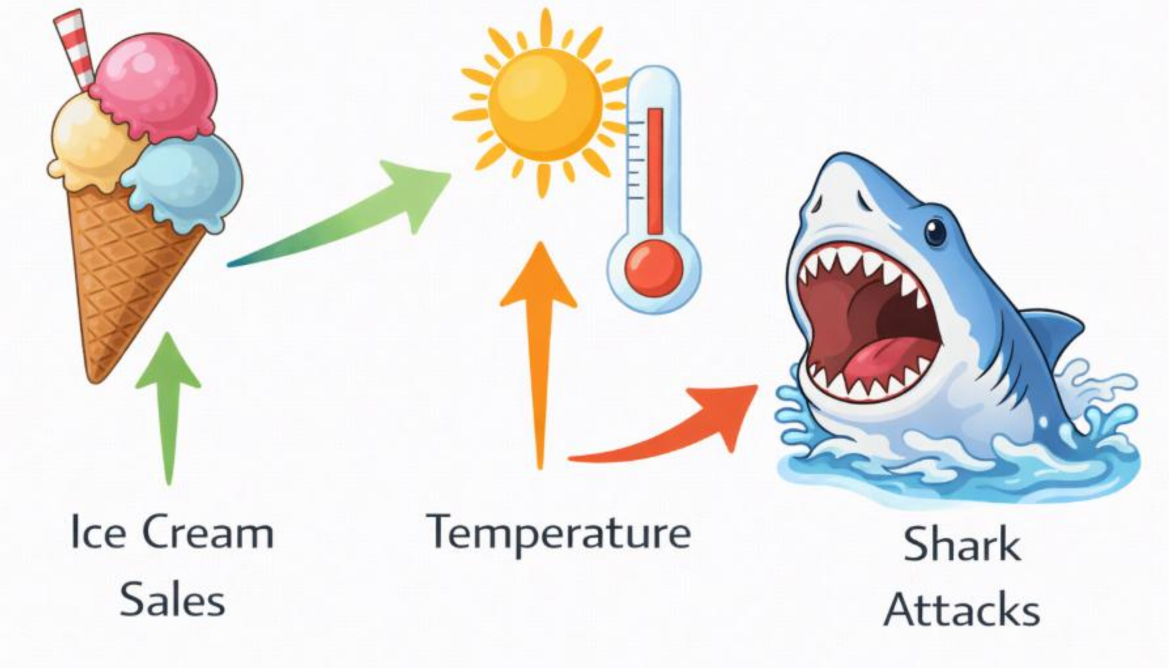
- Research dataset found correlation between:
  - number of people who drowned in swimming pools
  - number of films Nicolas Cage appeared in

spurious correlation

- Sometimes variables correlate **purely by chance**.

Multivariate analysis

- Two variables may appear correlated because **a third variable affects both.**



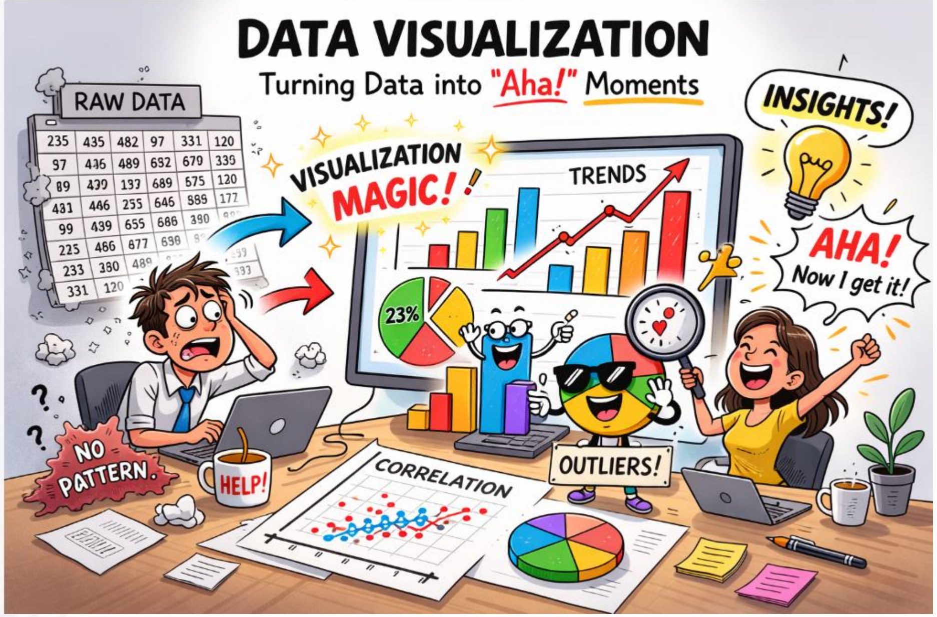
Visualization in Data Analysis

Faculty of Mathematics and Computer Science

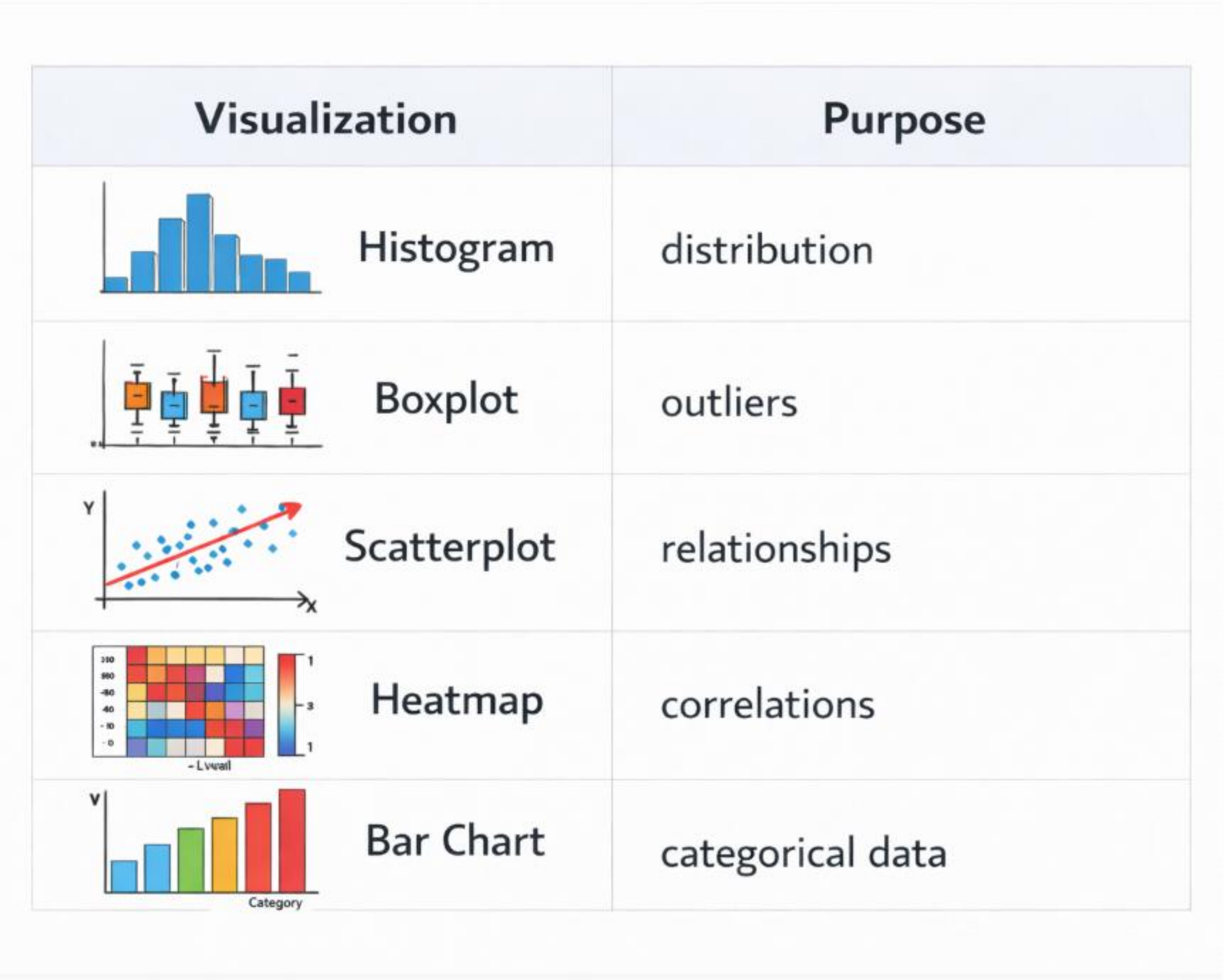
Visualization in Data Analysis

- Visualization allows the brain to detect patterns quickly.

- Visualization helps with:
  - Understanding data quickly
  - Identifying patterns and trends
  - Detecting outliers or anomalies
  - Explaining results to others
  - Supporting decision making



Visualization in Data Analysis



Visualization	Purpose
Histogram	distribution
Boxplot	outliers
Scatterplot	relationships
Heatmap	correlations
Bar Chart	categorical data

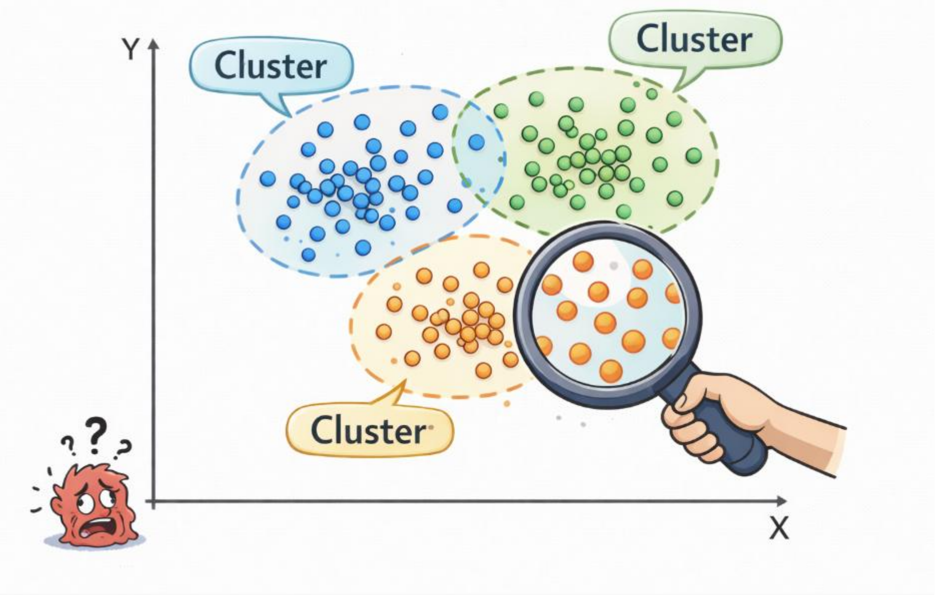
Y
X
- Lvwail
V
Category

Detecting patterns

Faculty of Mathematics and Computer Science

Detecting patterns

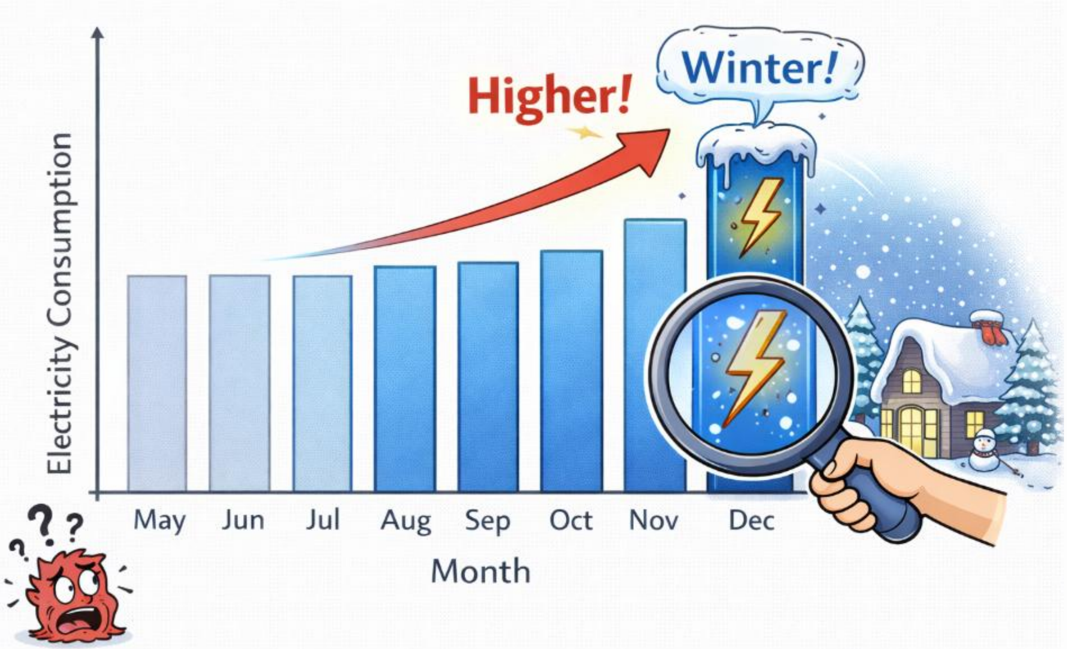
- EDA helps discover:
  - Clusters
  - Trends
  - Seasonality



- Clusters
  - Example: customer segments
  - Cluster 1 → young customers
  - Cluster 2→ families
  - Cluster 3→ retires

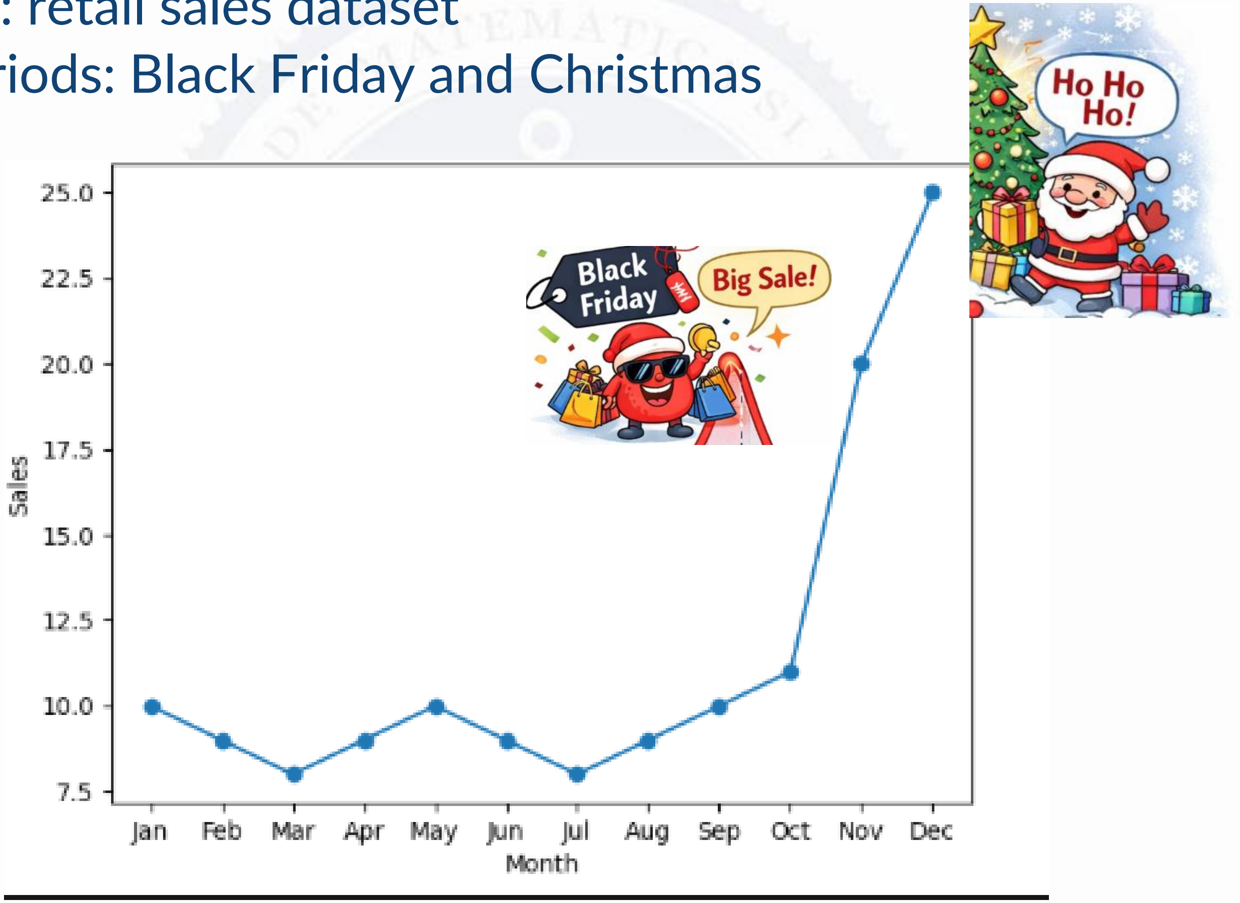
Detecting patterns

- Trend
  - Example: Electricity consumption
  - Pattern: Higher consumption in winter.



Detecting patterns

- Seasonality
  - Example: retail sales dataset
  - Peak periods: Black Friday and Christmas



Detecting anomalies

Faculty of Mathematics and Computer Science

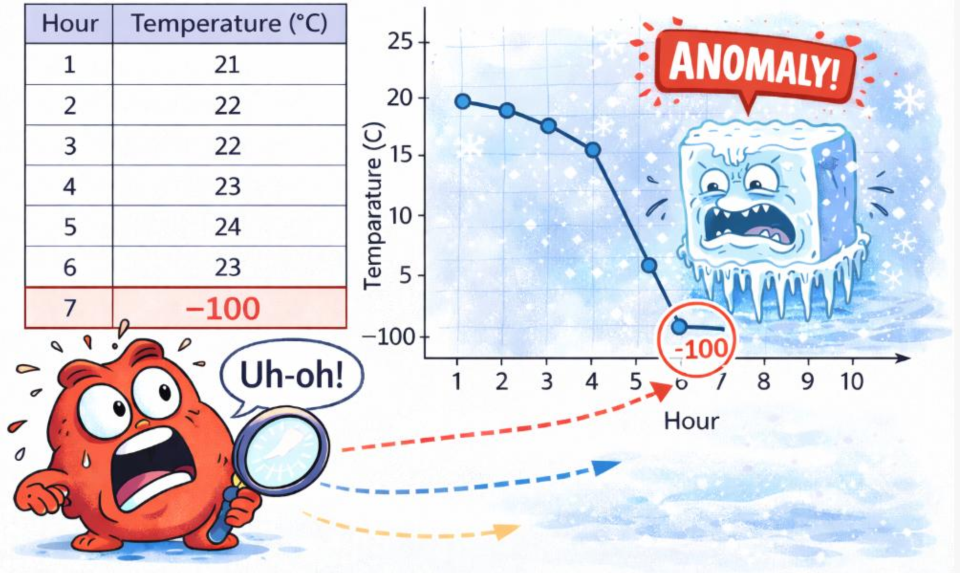
Detecting anomalies

- Anomalies may represent:
  - Errors
  - Rare events
  - Fraud
  - Interesting discoveries

- Example: Sensor Measurement Anomaly
  - A temperature sensor records hourly data.
  - Temperature suddenly becomes **−100°C**.

Hour	Temperature
1	21
2	22
3	22
4	23
5	24
6	23
7	-100

Detecting anomalies



Hour	Temperature (°C)
1	21
2	22
3	22
4	23
5	24
6	23
7	−100

- Possible explanations:
  - Sensor malfunction
  - Transmission error
  - Incorrect measurement unit

- In real-world systems (IoT, weather stations), these anomalies occur frequently.

- EDA helps detect such values before machine learning models are trained.

Industry case study

Faculty of Mathematics and Computer Science

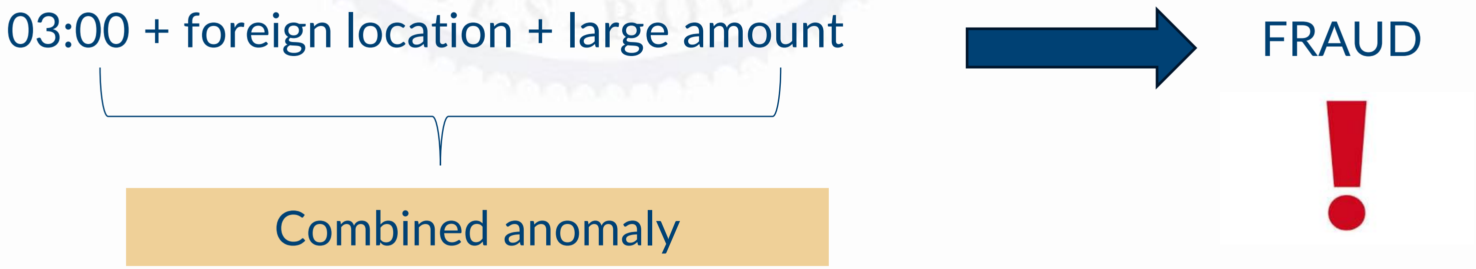
Industry case study-Credit Card Fraud Detection

- Banks process millions of credit card transactions every day.
- Example statistics (approximate industry scale):
  - Visa processes ~65,000 transactions per second
  - Fraud losses globally exceed tens of billions of dollars per year
- Because of this, banks use data analysis and machine learning systems to detect suspicious transactions.
- Before building models, analysts perform Exploratory Data Analysis (EDA) to understand patterns in the data.

Industry case study-Credit Card Fraud Detection

Transaction id	Customer id	Amount ($)	Country	Time	Merchant	Fraud
1	A123	25	RO	14:12	Supermarket	0
2	A123	40	RO	16:230	Restaurant	0
3	A123	2000	Brazil	03:30	Online store	1

- During EDA, analysts visualize distributions and identify anomalies.
- Typical Fraud Patterns Discovered in EDA:
    - Unusually large transaction
    - Unusual transaction location
    - Unusual transaction time



03:00 + foreign location + large amount

Combined anomaly

FRAUD

Teamwork time

Faculty of Mathematics and Computer Science

Teamwork time- Understanding Customer Purchasing Behaviour

- **Scenario**
  - You are working as data analysts for an online retail company.
  - The marketing team wants to understand which customers are most likely to spend more money on the platform.
  - You receive the following dataset describing customer behaviour.

Customer	Age	Income (€)	Visits_per_month	Avg_purchase (€)
1	22	1200	10	30
2	25	1500	9	45
3	31	3500	6	120
4	40	5000	4	200
5	29	2000	12	60
6	50	7000	2	350
7	36	4200	5	150
8	28	1800	11	50

Teamwork time- Understanding Customer Purchasing Behaviour

- Work in teams.

- Time: 10 minutes.

- Your goal is to perform conceptual exploratory data analysis.
  1. Discuss the following questions.
     - Which variables appear to influence average purchase value?
     - Identify Behavioural Patterns→ Look at visits per month.
     - Which Visualizations Would You Use? (histogram, box plot, scatter plot, etc.)

  2. Generate a Business Insight
     - Each team must propose one insight for the marketing team.

EDA best practices & mistakes

EDA best practices & mistakes

- Best practices:
  - always visualize first
  - check distributions
  - investigate anomalies
  - question correlations
  - understand context

- Common EDA mistake:
  - ignoring outliers
  - assuming correlation = causation
  - skipping visualization
  - trusting summary statistics only

Key Takeaways

Faculty of Mathematics and Computer Science

Key Takeaways

- EDA helps:
  - understand dataset structure
  - detect patterns
  - detect anomalies
- EDA is essential before modelling
- Visualization reveals hidden patterns
- Statistics summarize data
- Relationships explain behaviour
- Anomalies may reveal important insights

Thank you for your attention — questions, thoughts, or challenges?

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
BABEȘ-BOLYAI UNIVERSITY

1 Mihail Kogălniceanu Street,
Cluj-Napoca, Cluj, România

www.cs.ubbcluj.ro