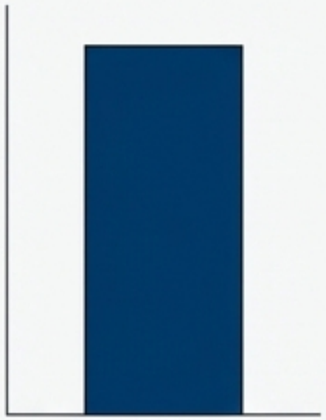
Mastering Univariate Analysis

The Foundation of Exploratory Data Analysis (EDA)

EXPLORATORY DATA ANALYSIS SERIES

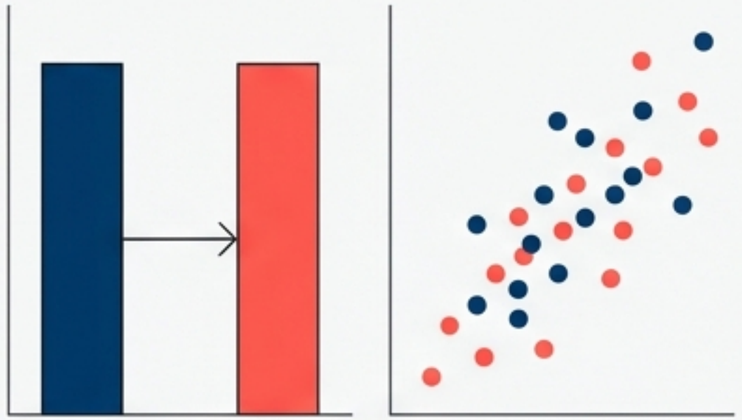
Knowing Your Data “Inside Out”

Univariate



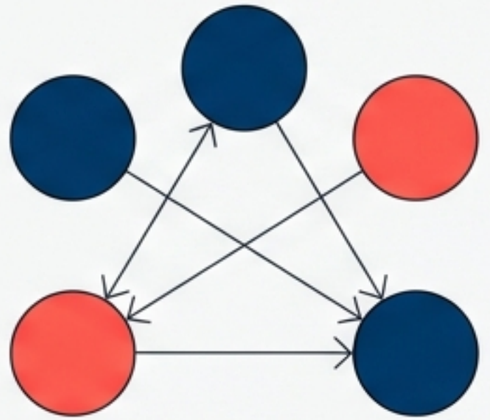
Analyzing one variable independently. Understanding raw quality and distribution.

Bivariate



Analyzing the relationship between two variables (e.g., Age vs. Price).

Multivariate



Analyzing complex interactions between three or more variables.

Core Concept: You cannot model relationships if you do not understand the raw materials first.

The First Question: Numerical or Categorical?

The analysis technique depends entirely on the data type.

Categorical Data

Discrete groups or classes.

Nationality (Country)
Gender (Male/Female)
Passenger Class (1, 2, 3)
Survived (0 or 1)

Numerical Data

Continuous measurements.

Age (Years)
Fare ($)
Height (cm)
Family Size (Count)

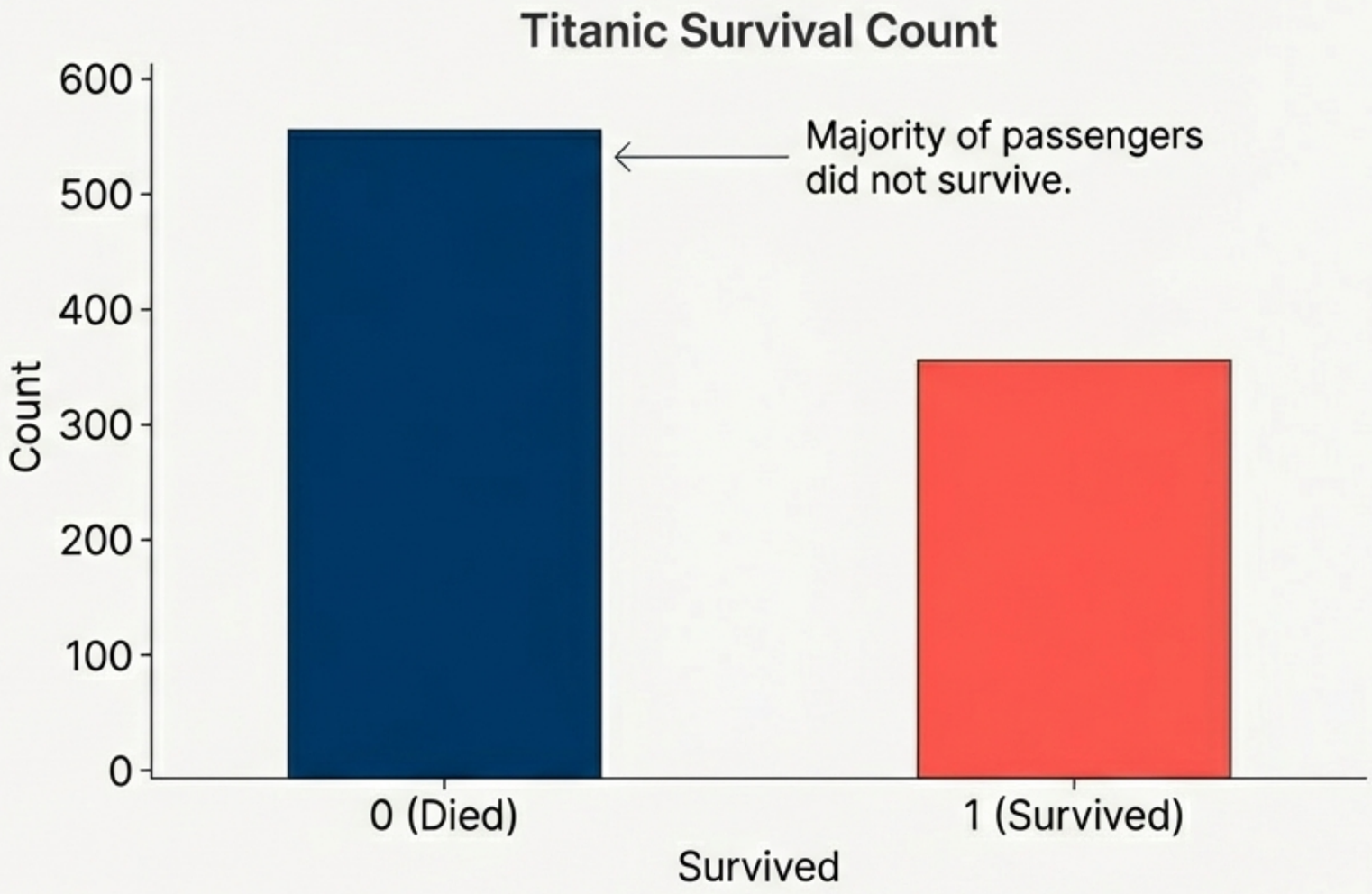
Context: We will use the Titanic dataset to demonstrate these concepts.

Categorical Analysis: Measuring Frequency

- **The Question:** How many instances of each category exist?

- **The Tool:** The Countplot (Bar Chart).

- **Insight:** Instantly reveals class imbalance.



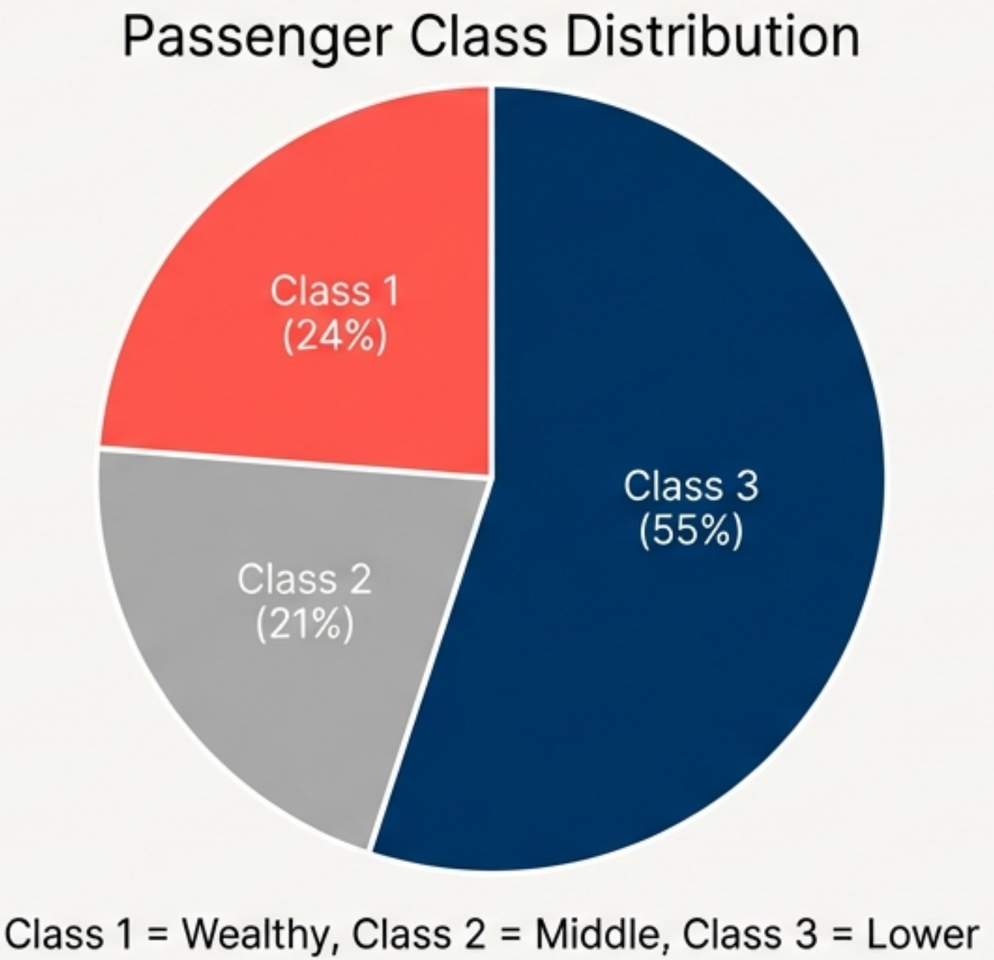
Swiss International Data Editorial

Visualising Proportions and “Market Share”

The Question: What percentage of the whole does this category represent?

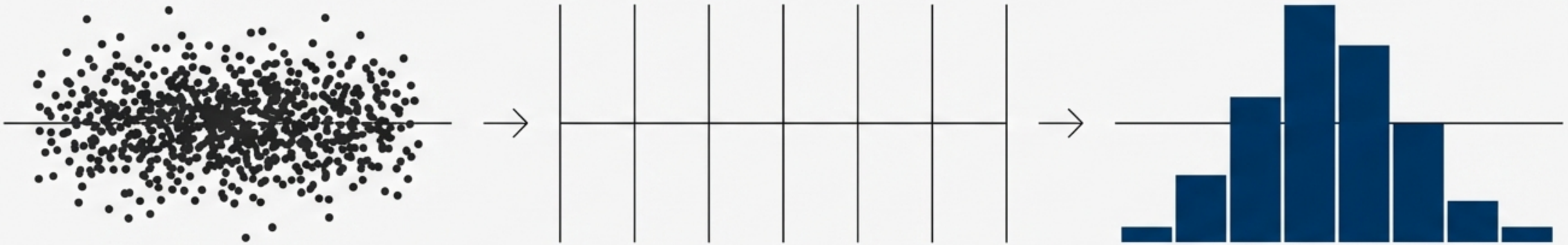
The Tool: The Pie Chart.

Technical Note: Use `autopct` to display percentage values.



Shifting to Numerical Data: The Concept of Distribution

Raw Data → Binning → Distribution



Continuous values are hard to count individually.

We create ranges called “Bins” to group values.

We count how many points fall into each bin to see the shape.

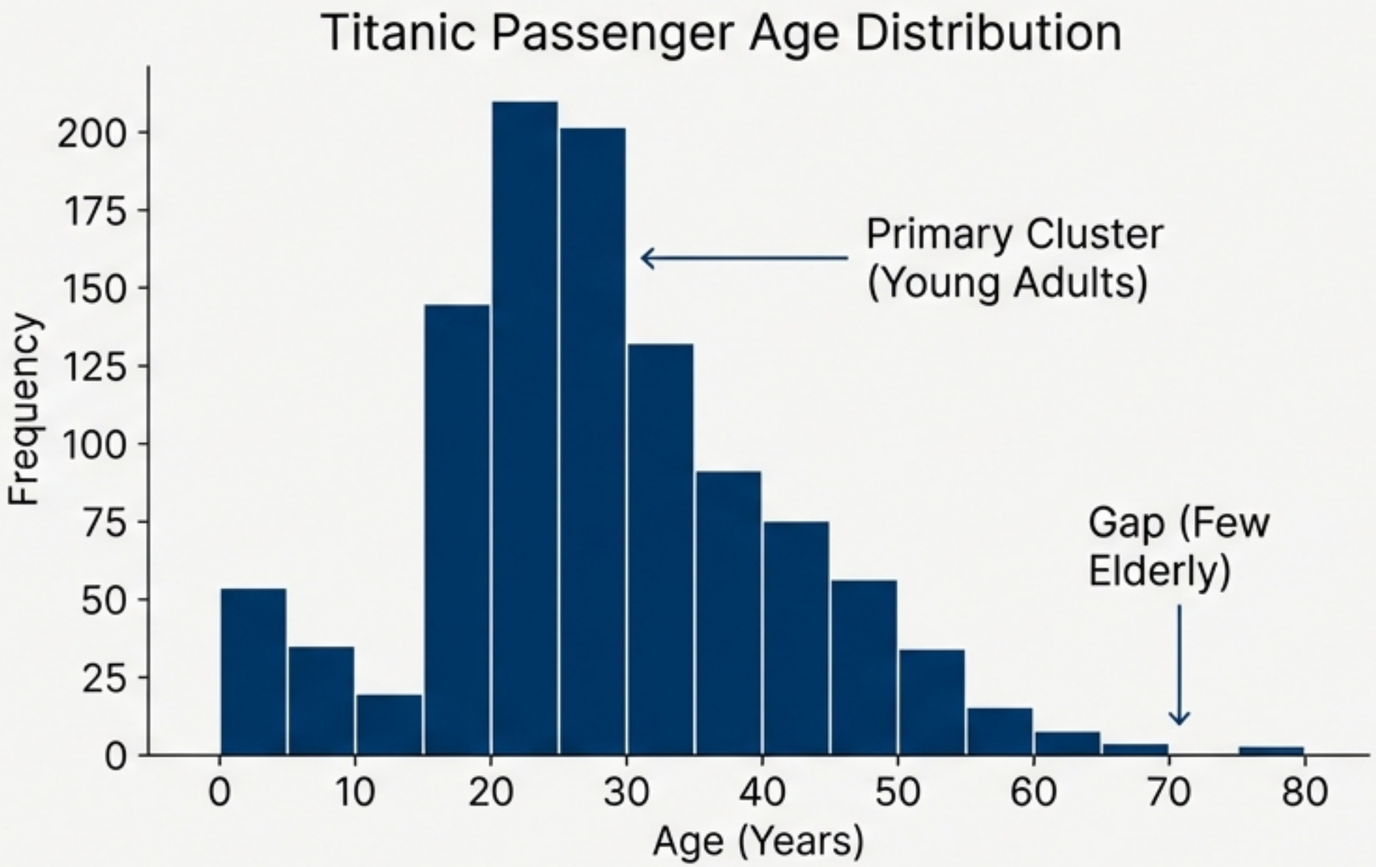
Key Metric Questions: Where is the center? What is the range? Is it clustered?

The Histogram: Visualising Visualising Frequency Ranges

The Tool: Histogram.
Short sharts count of frequency ranges. Identifying stequences and variatined bins.

Method: Plots the count of records within specific bins.

Insight: Identifying clusters and gaps.

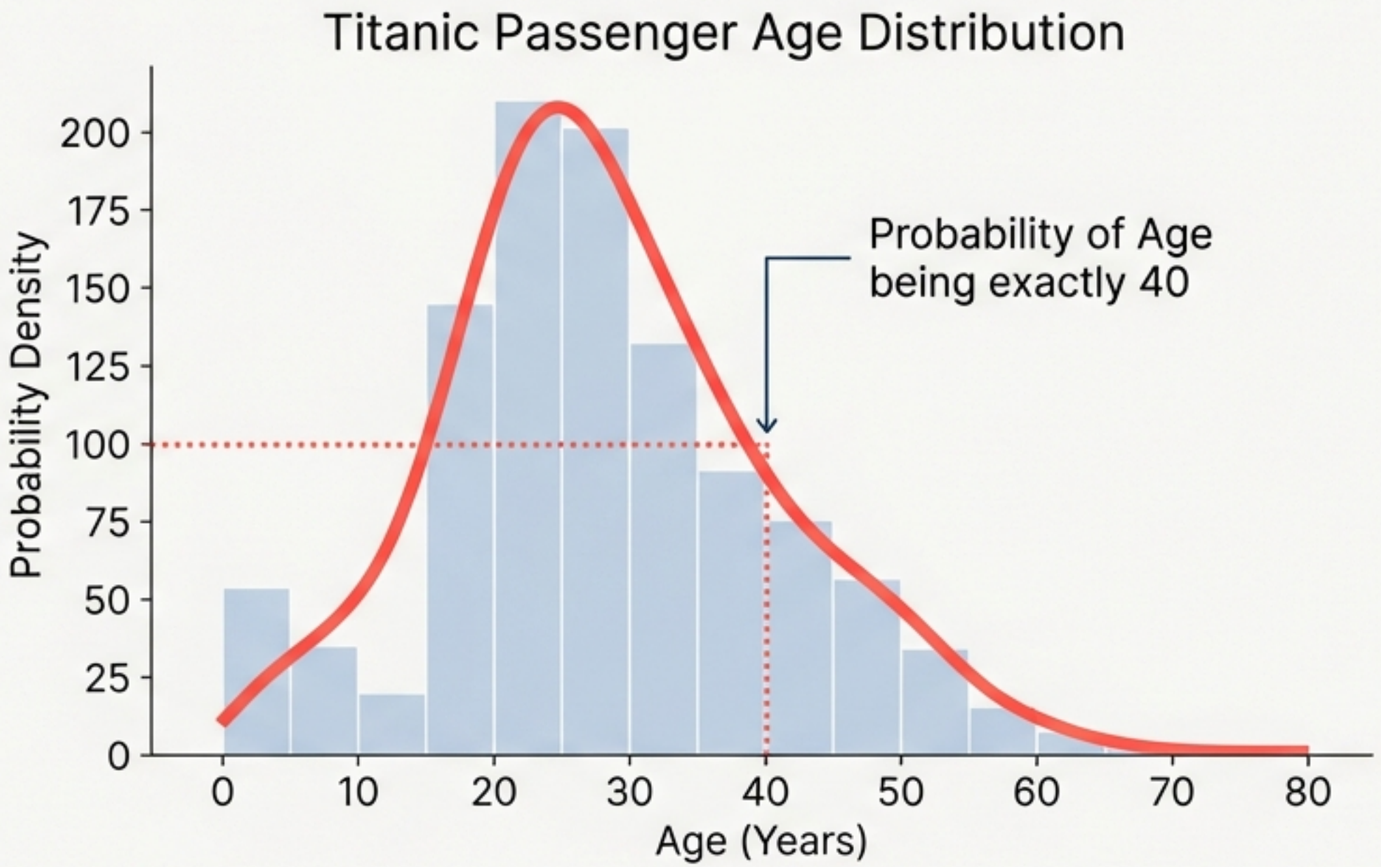


The Distplot and Probability Density

The Tool: Distplot / KDE (Kernel Density Estimation).

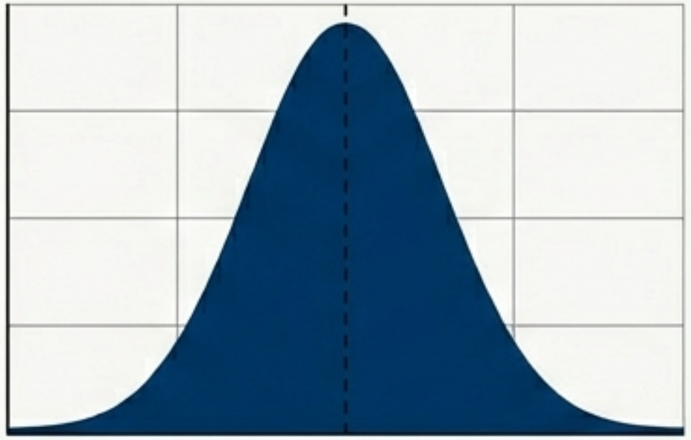
The Upgrade: Adds a smoothed curve (PDF) over the histogram.

The Shift: Y-Axis represents **Probability** of a specific value, not raw count.



Diagnosing Data Shape: Skewness

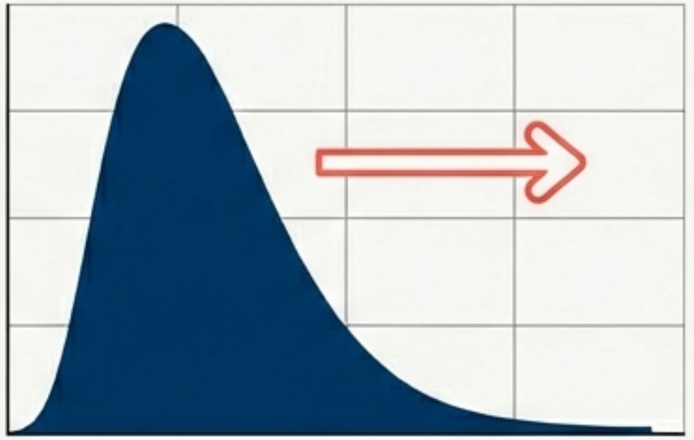
Normal Distribution



Symmetrical. Skewness ≈ 0.

Human Height.

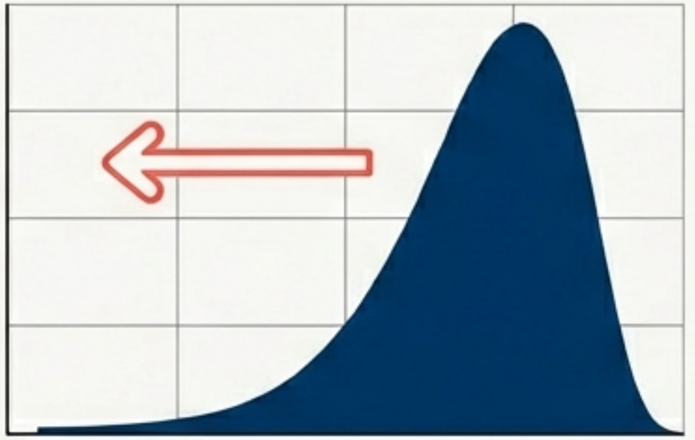
Positively Skewed



Right Skew. Skewness > 0.

Salaries (Many low, few high).

Negatively Skewed



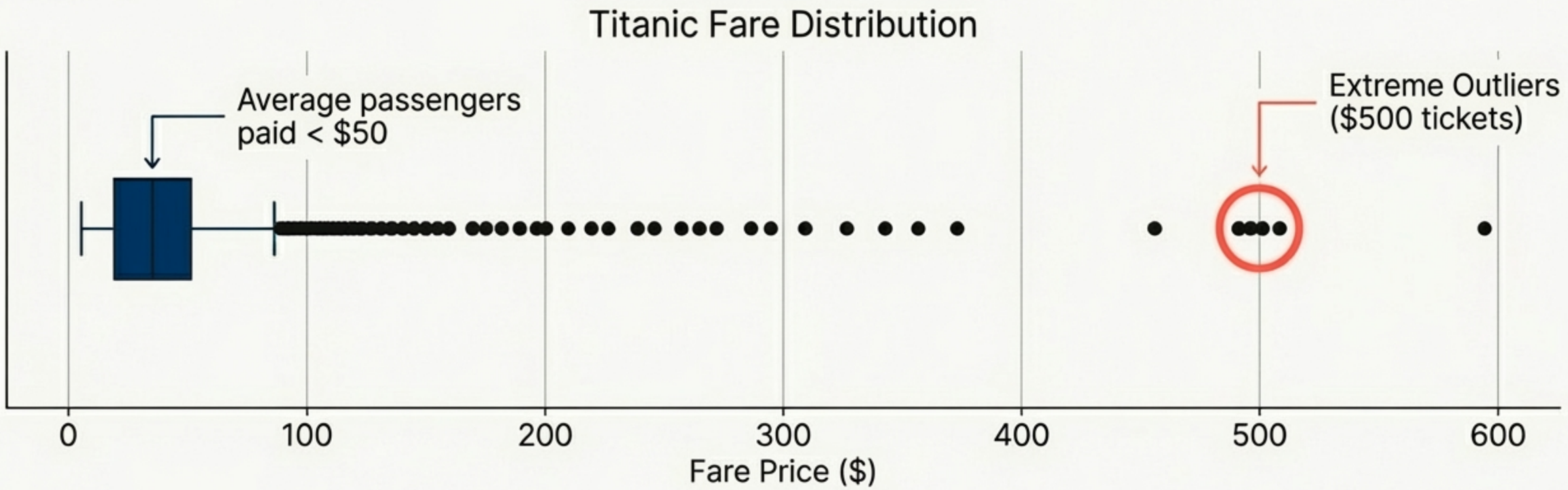
Left Skew. Skewness < 0.

Easy Exam Marks (Many high, few low).

Code Check: Use `df['column'].skew()` to find the value.

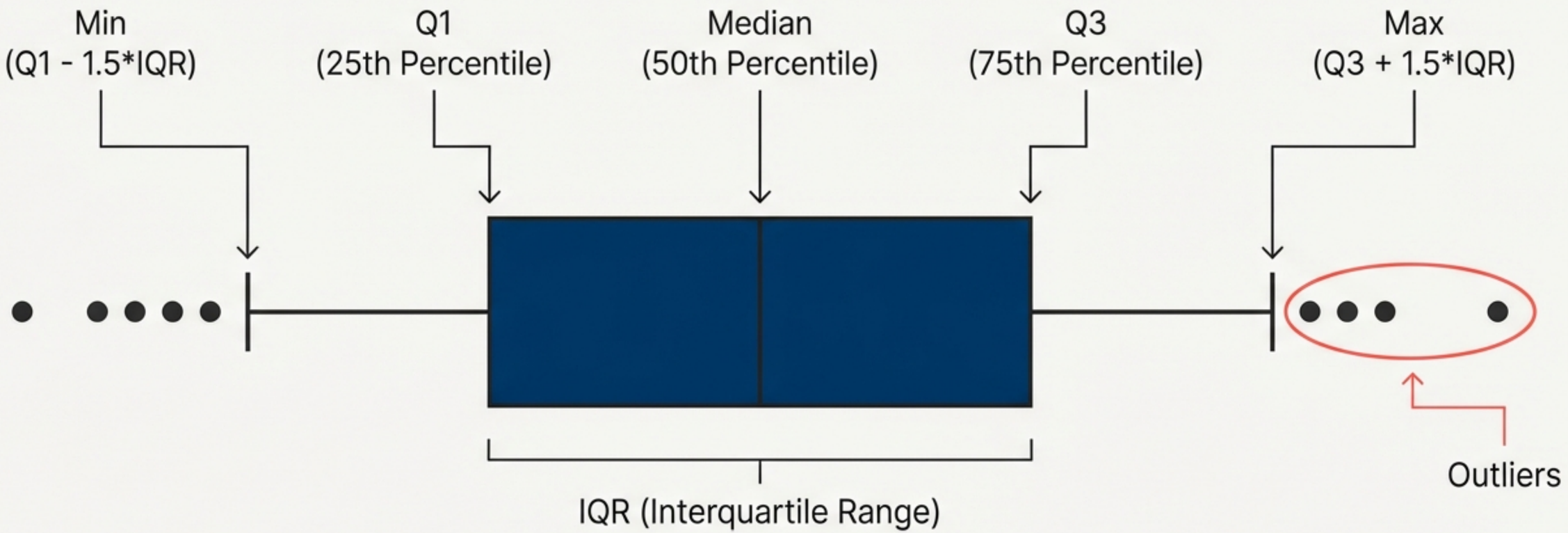
The Boxplot: Hunting for Outliers

Histograms can hide extreme values. The
The Boxplot is designed to flag anomalies and
"noise".



Swiss International Data Editorial

Anatomy of a Boxplot (The 5-Number Summary)



Any data point falling outside the Min/Max whiskers is considered an Outlier.

Descriptive Statistics

Inter: Validating visual insights with raw numbers.

min() / max()	Absolute boundaries of the data. (Titanic Age: Min **0.42**, Max **80**)
mean()	Arithmetic average. (Titanic Age: **~29.7** years)
std()	Standard Deviation. Measures spread/volatility.

Code Snippet

```
df['Age'].describe()

count, mean, std, min,
25%, 50%, 75%, max
```

The Univariate Toolkit Summary

Categorical Data

Goal: Frequency **A**

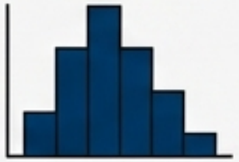
Tool: Countplot

Goal: Market Share % **B**

Tool: Pie Chart

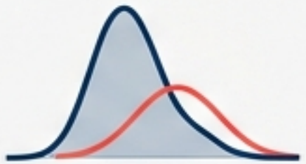
Numerical Data

Goal: Distribution/Shape **C**



Tool: Histogram

Goal: Probability/Skew **D**



Tool: Distplot (KDE)

Goal: Outliers **E**

Tool: Boxplot

The Practical Workflow

1 Import

Load data (pandas) and plotting libraries (seaborn/matplotlib).

2 Identify

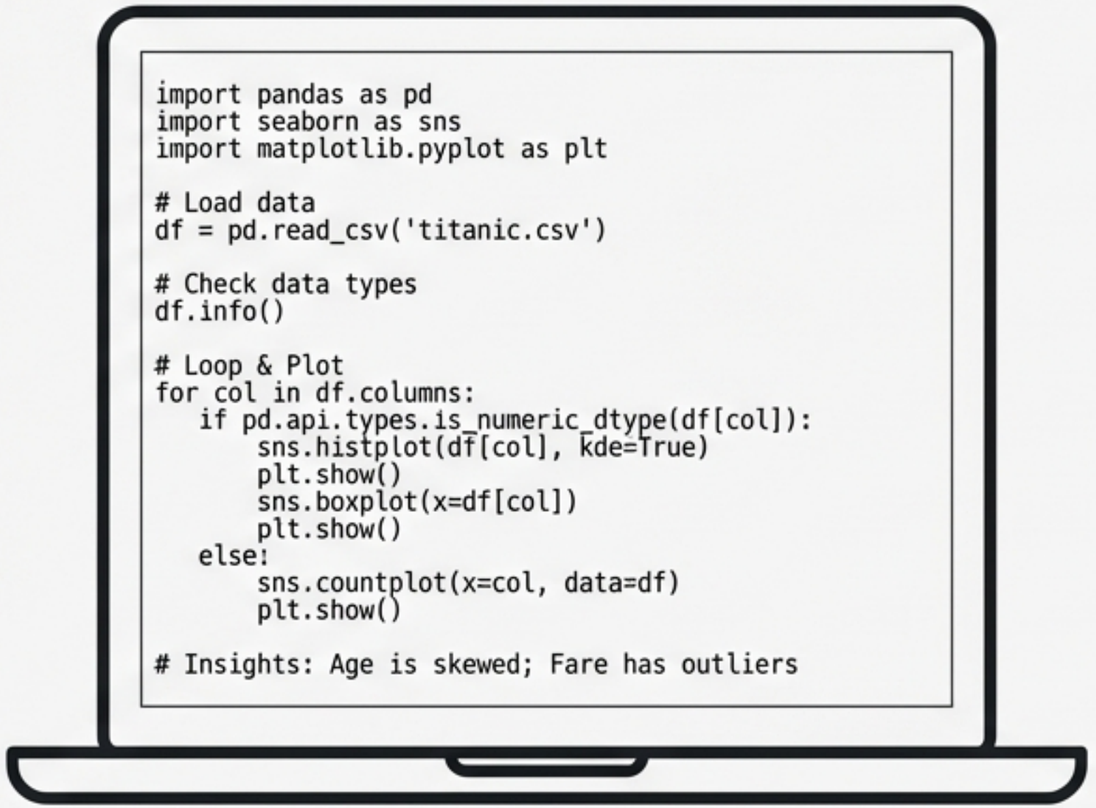
Check data types using `df.info()`. Is it Numerical or Categorical?

3 Loop & Plot

Iterate through columns. If Categorical -> Countplot. If Numerical -> Distplot & Boxplot.

4 Note Insights

Write down observations immediately (e.g., 'Age is skewed', 'Fare has outliers').



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

# Load data
df = pd.read_csv('titanic.csv')

# Check data types
df.info()

# Loop & Plot
for col in df.columns:
    if pd.api.types.is_numeric_dtype(df[col]):
        sns.histplot(df[col], kde=True)
        plt.show()
        sns.boxplot(x=df[col])
        plt.show()
    else:
        sns.countplot(x=col, data=df)
        plt.show()

# Insights: Age is skewed; Fare has outliers
```

Start Your Investigation

You have now inspected the witnesses (columns) individually. You understand the data quality, outliers, and basic structure.

Next Step: Bivariate Analysis (Identifying relationships between columns).

Pick a dataset that interests you—Cricket, Finance, or Movies—and perform a full Univariate Analysis.

"You cannot solve the mystery if you don't know the characters."