Data Analysis and Visualization

CentraleDigitalLab@Nice

Greta Damo - PhD - greta.damo@univ-cotedazur.fr
MARIANNE, Université Côte d'Azur, CNRS, INRIA, I3S

Basics of Data Analysis

What is a Random Variable?

A random variable (r.v.) X is a function $X: \Omega \rightarrow R$ where Ω is the state space and R is the set of values that the variable can take called Range.

Basics of Data Analysis

What is a Random Variable?

A random variable (r.v.) X is a function $X: \Omega \rightarrow R$ where Ω is the state space and R is the set of values that the variable can take called Range.

Intuitively, a r.v. is equivalent to a column of your dataset after applying zero or more fi lters.

Basics of Data Analysis

What is a Random Variable?

A random variable (r.v.) X is a function X: Ω ⇾ R where Ω is the state space and R is the set of values that the variable can take called Range.

Intuitively, a r.v. is equivalent to a column of your dataset after applying zero or more fi lters.

A random variable can be of di ff erent types: numerical or categorical.

Basics of Data Analysis

What is a Random Variable?

Numerical:

- **Continuous** : Can take on any value within a range, including decimals and fractions. E.g., the height of students in a school (150.2 cm, 165.8 cm …).
- **Discrete** : Can take on speci fi c, separate values and is countable. E.g., the number of cars passing a toll booth in a day (0, 1, 2, 3 …).
- **Finite Set** : The number of siblings a person has (e.g., 0, 1, 2, 3... up to a reasonable maximum).
- **Infinite Set** : The number of times you need to roll a dice until you get a six (potentially in fi nite but countable).

Basics of Data Analysis

What is a Random Variable?

Categorical: Variables that represent distinct groups or categories.

- **Nominal:** no inherit order. E.g., eye color of individuals (Blue, Brown, Green).
- **Ordinal** : Variables with a meaningful order or ranking. E.g., rating of a restaurant on a scale from 1 to 5 (Poor, Fair, Good, Very Good, Excellent).

Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Columns

(Random Variables)

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Rows

(Elements of Ω)

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Basics of Data Analysis

What is a Random Variable? function X: Ω → R

Age	Height	Degree's level
25	172	Master
26	167	Master
22	170	Bachelor
23	160	Bachelor

Set of values of a r.v.

(Range R)

Data types mini-quiz

Match each variable with its type:

- Temperature —> ?
- Marital Status —> ?
- Age Group —> ?

A. Numerical continuous
B. Numerical discrete
C. Categorical nominal
D. Categorical ordinal

Exploratory Data Analysis (EDA)

The importance of EDA

EDA is a method of analyzing and examining data sets to uncover patterns, identify relationships, and gain insights. It aims to develop a deep understanding of data sets and identify potential problems early.

Why is EDA important?

- Provides a deep understanding of data
- Detects errors, outliers, and patterns before formal modelling
- Forms the foundation for e ff ective decision-making

Exploratory Data Analysis (EDA)

Key processes of EDA

- **Data Cleaning and Preparation** : Ensures data quality

Exploratory Data Analysis (EDA)

Data Cleaning and Preparation

Why is important to clean the data?

- Raw data often contains inconsistencies, missing values, or errors
- Clean data ensures reliable analysis

Exploratory Data Analysis (EDA)

Data Cleaning and Preparation

Why is important to clean the data?

- Raw data often contains inconsistencies, missing values, or errors
- Clean data ensures reliable analysis

Steps in Data Cleaning:

1. Handling missing values
2. Removing duplicates
3. Resolving inconsistencies in formats and categories
4. Formatting and standardising data

Exploratory Data Analysis (EDA)

Key processes of EDA

- **Data Cleaning and Preparation** : Ensures data quality
- **Descriptive Statistics** : Summarises key metrics

Exploratory Data Analysis (EDA)

Descriptive statistics

Descriptive statistics basically provides a snapshot of data characteristics

Key metrics:

- **Central Tendency:** mean, median, mode
- **Variability:** range, variance, standard deviation
- **Distribution shape:** Skewness, kurtosis

Exploratory Data Analysis (EDA)

Descriptive statistics: Central Tendency

Definition: Measures that summarise a dataset within a single value

Key metrics:

- **Mean:** average of all points

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median:** Middle value when the data is sorted
- **Mode:** Most frequently occurring value

When to use?

- **Mean** for symmetric distributions
- **Median** for skewed distributions
- **Mode** for categorical data

Exploratory Data Analysis (EDA)

Descriptive statistics: Measures of Variability

Definition: Understand how spread out the data is

Key metrics:

- **Range:** Difference between max and min values. $\text{Range} = x\ \max - x\ \min$
- **Variance:** Average squared deviation from the mean. $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- **Standard Deviation:** Square root of variance. $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
- **Interquartile Range (IQR):** Range between Q1 (median of the lower half of the data) and Q3 (median of the upper half of the data).

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

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR}$$

Example: Univariate Data Summary

Data:
Exam Scores = [50, 55, 75, 78, 79, 80, 82, 90, 95, 100]

Mean: ___

Median: ___

Mode: ___

Standard Deviation: ___

Fill in the blanks or discuss with a partner.

When to use Mean vs Median

Dataset A: 40, 45, 47, 50

Dataset B: 10, 12, 50, 150

Which summary is better? Why?

When to use Mean vs Median

Dataset A: 40, 45, 47, 50

Dataset B: 10, 12, 50, 150

Which summary is better? Why?

Mean: for symmetric data

Median: for skewed/outlier data

Exploratory Data Analysis (EDA)

Key processes of EDA

- **Data Cleaning and Preparation** : Ensures data quality
- **Descriptive Statistics** : Summarises key metrics
- **Data Visualization** : Makes data insights visible

Exploratory Data Analysis (EDA)

Data Visualization

Make patterns and insights more accessibile

Benefits:

- Highlights trends and anomalies
- Simplify complex datasets

Types of Visualisations:

- Univariate, Bivariate, Multivariate

Exploratory Data Analysis (EDA)

Data Visualization: Univariate visualisations

Focus on single variables

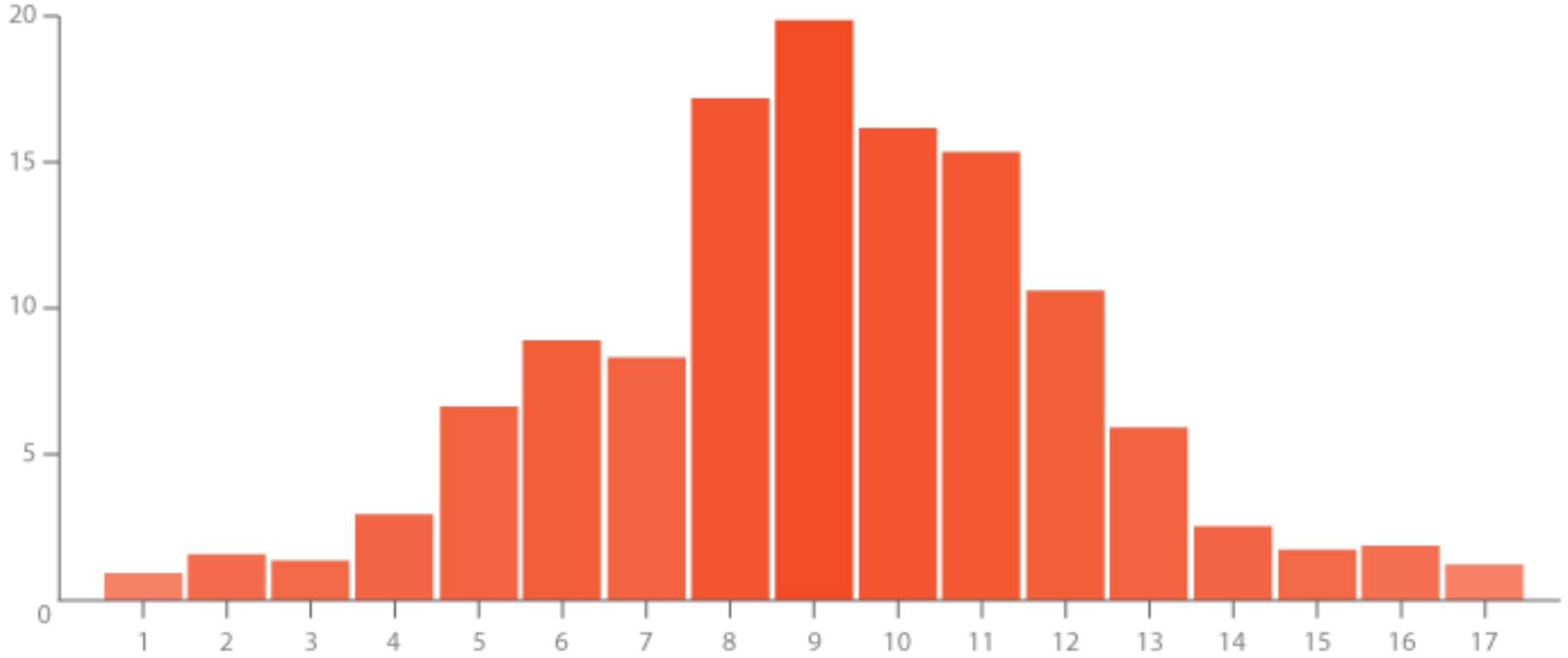
Examples:

- **Histograms** : Show frequency distribution
- **Box Plots** : Highlight spread and outliers
- **Density Plots** : Visualise the distribution

Use Case: Understanding the range and distribution of a variable.

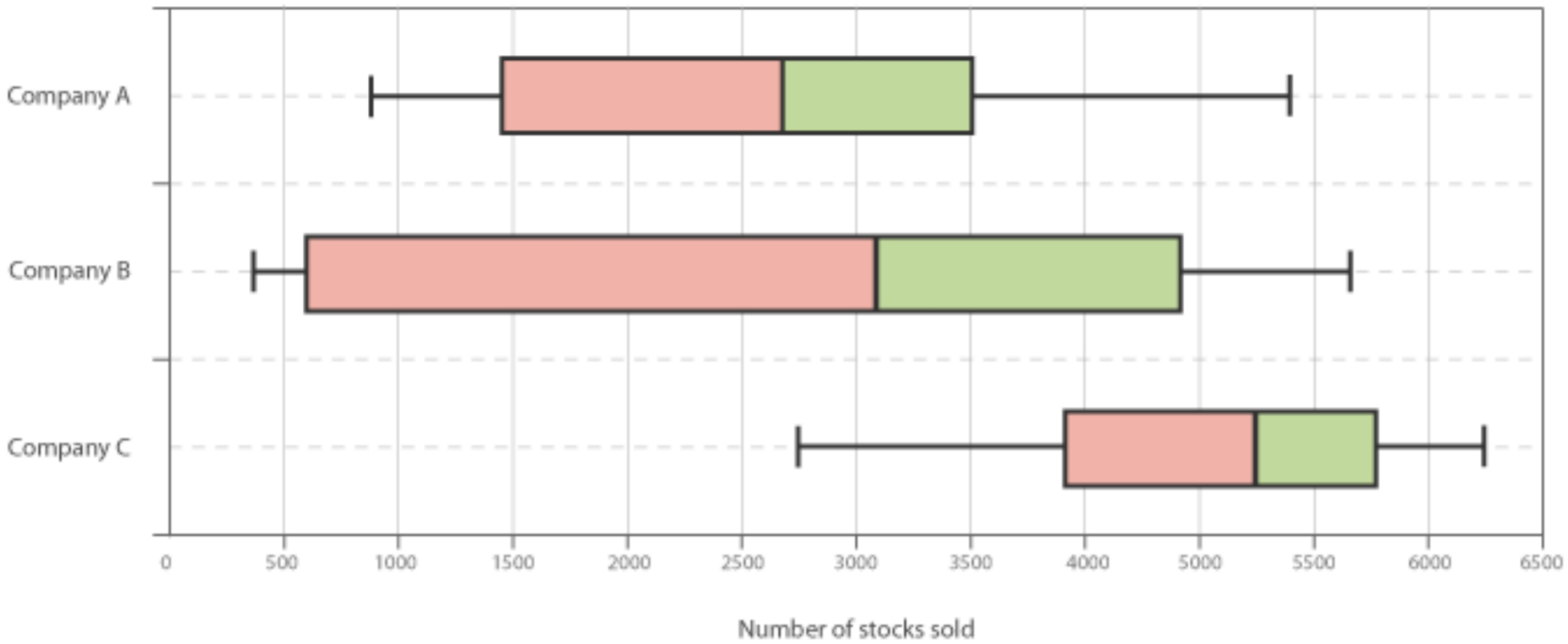
Exploratory Data Analysis (EDA)

Data Visualization: Univariate visualisations → Histograms



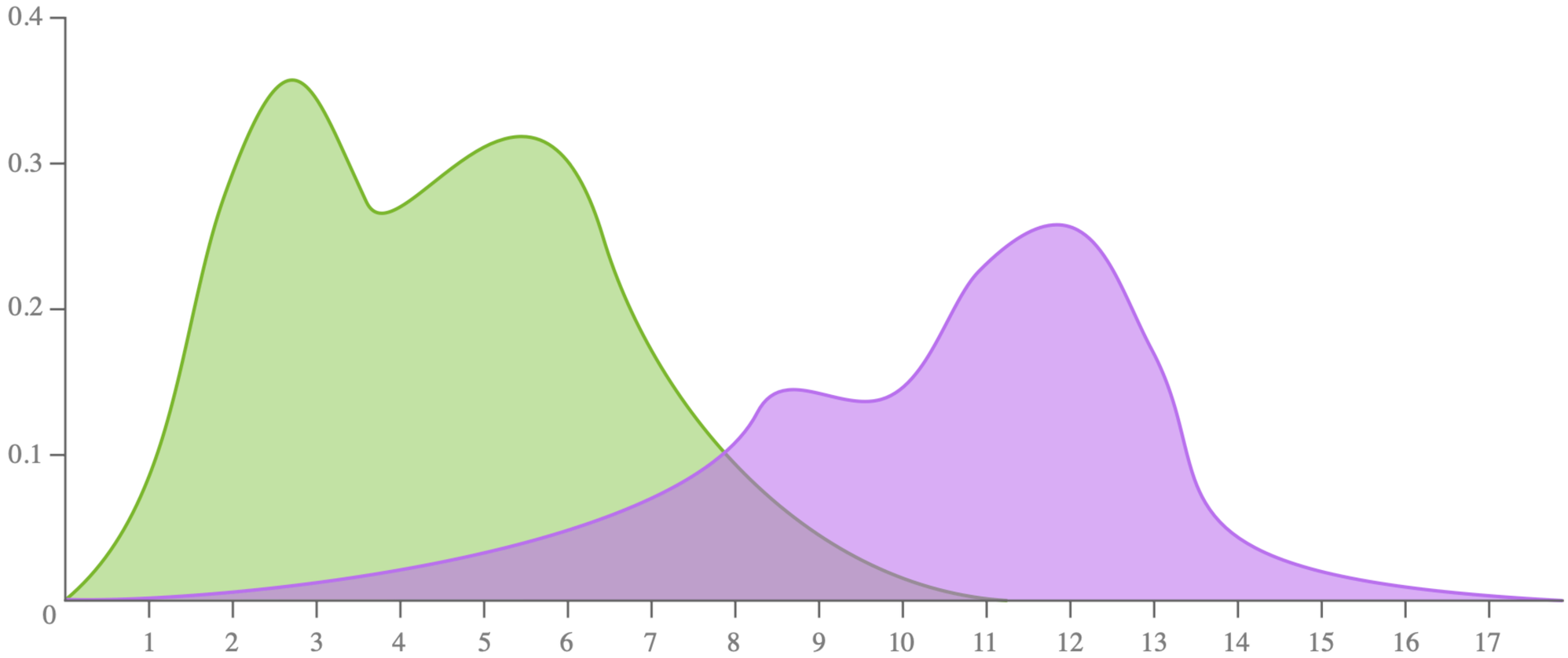
Exploratory Data Analysis (EDA)

Data Visualization: Univariate visualisations → Box Plot



Exploratory Data Analysis (EDA)

Data Visualization: Univariate visualisations → Density Plot



Exploratory Data Analysis (EDA)

Data Visualization: Bivariate Visualisations

Explore Relationships Between Two Variables

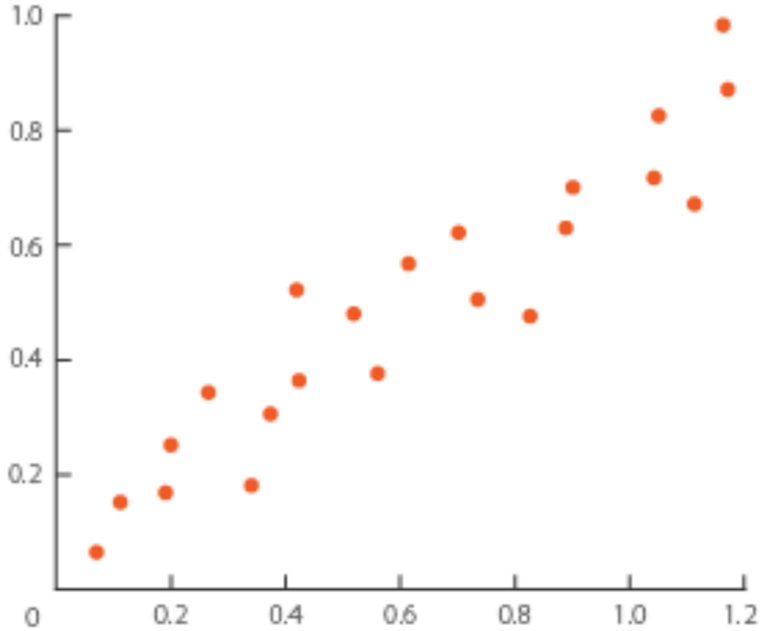
Examples:

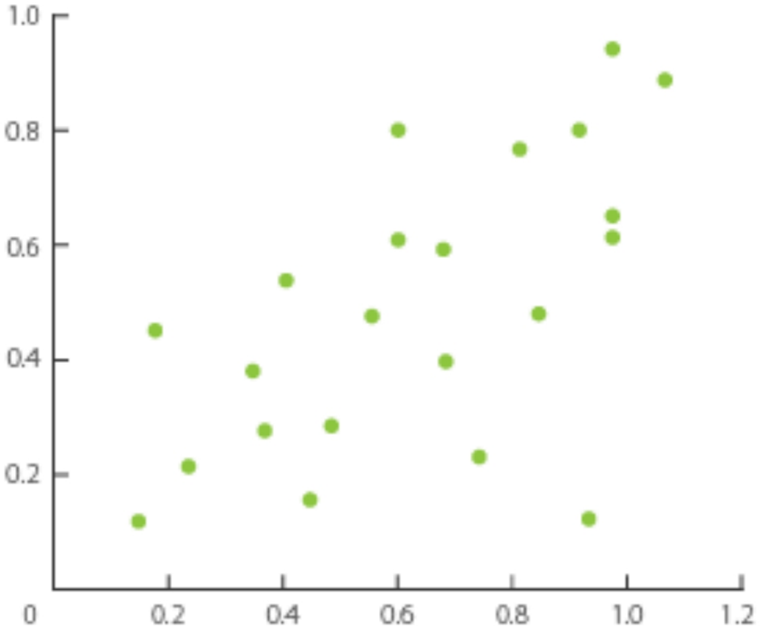
- **Scatter Plots:** Show correlation (positive, negative, or none)
- **Line Graphs:** Display trends over time
- **Heatmaps:** Visualize correlations in a matrix format

Use Case: Identifying correlations

Exploratory Data Analysis (EDA)

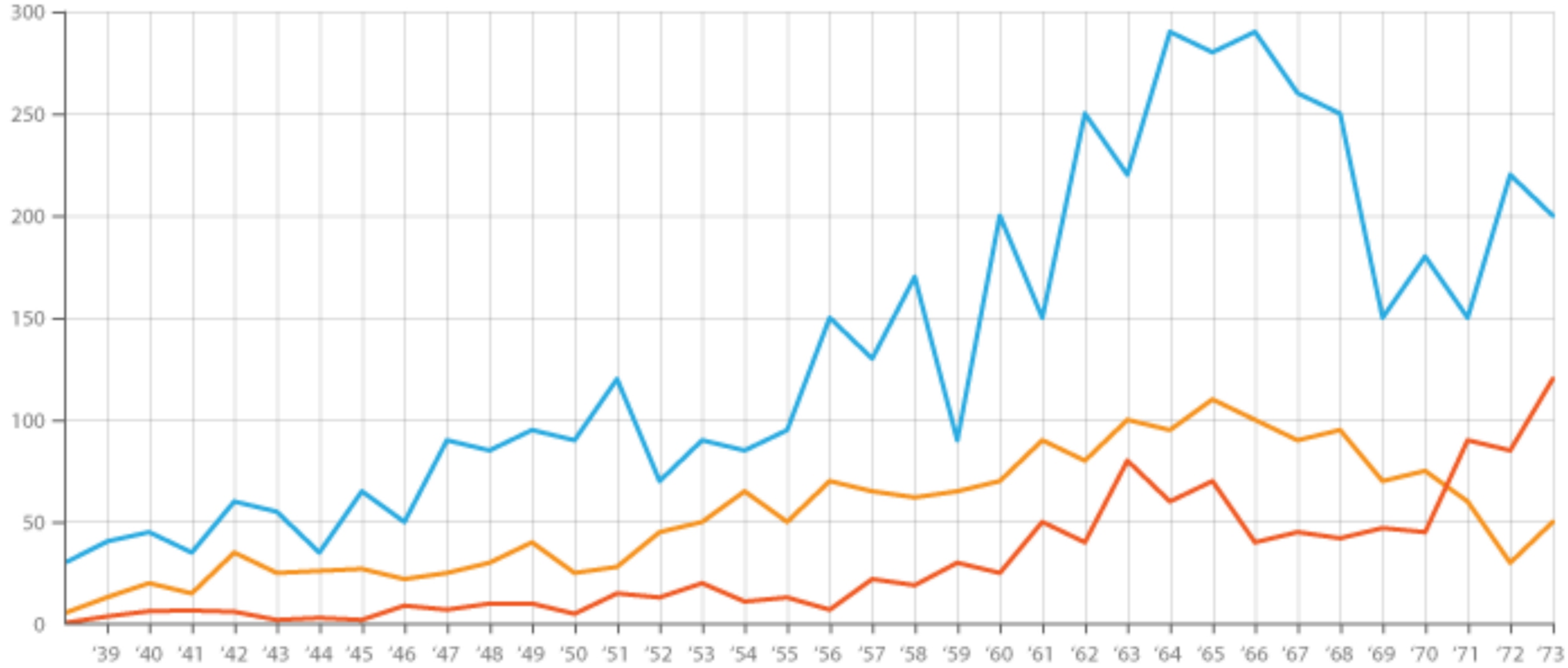
Data Visualization: Bivariate Visualisations → Scatter Plots





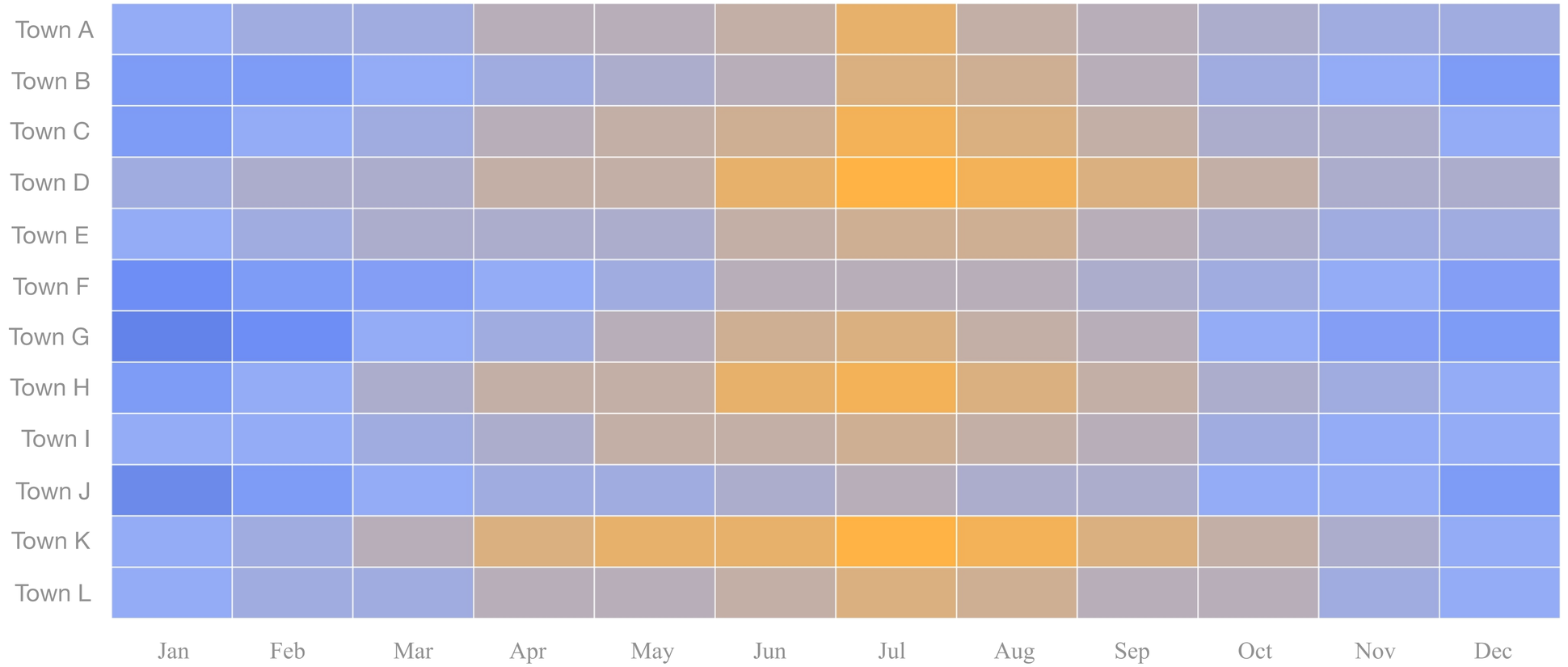
Exploratory Data Analysis (EDA)

Data Visualization: Bivariate Visualisations → Line graphs



Exploratory Data Analysis (EDA)

Data Visualization: Bivariate Visualisations → Heatmaps



Exploratory Data Analysis (EDA)

Data Visualization: Multivariate Visualisations

Visualize Multiple Variables

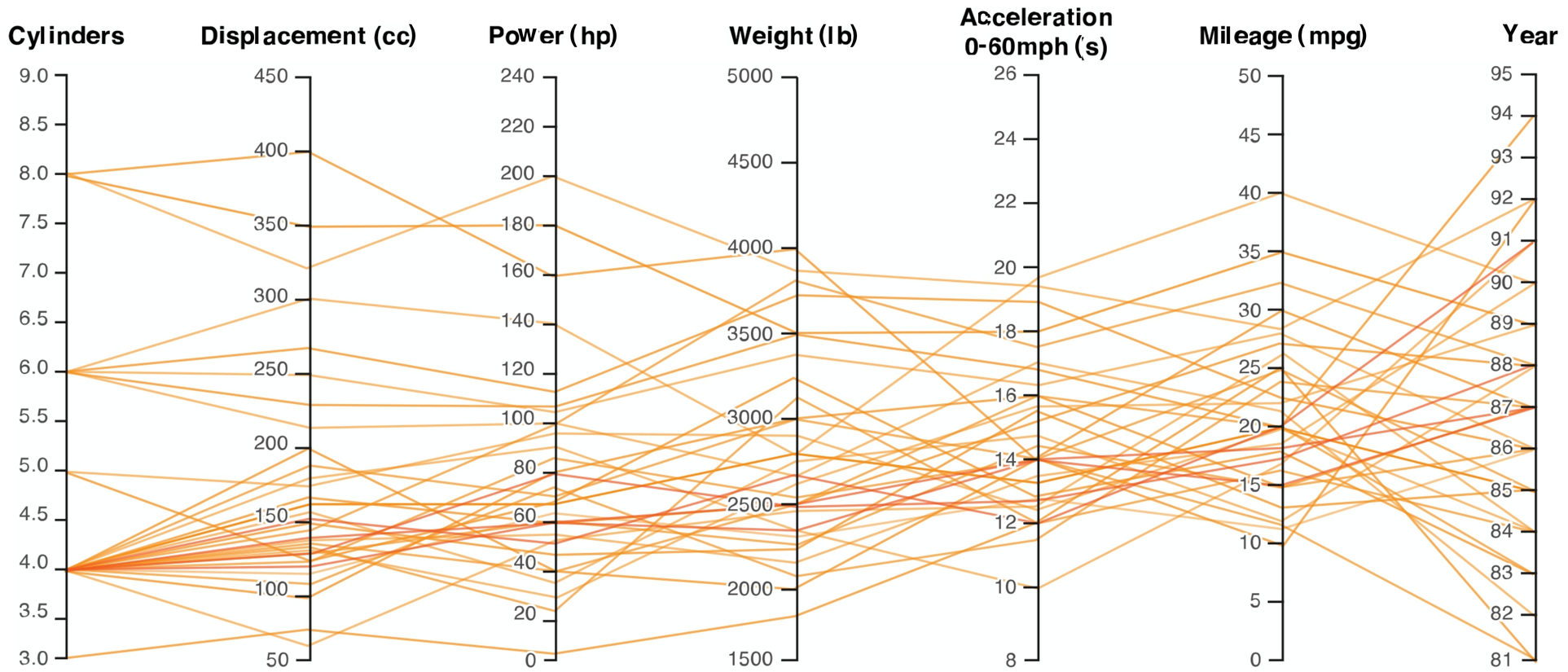
Examples:

- **Pair Plots:** Matrix of scatter plots
- **3D Scatter Plots:** Add depth to bivariate data
- **Parallel Coordinates:** Compare variables across observations

Use Case: Represent phenomenons in the dataset

Exploratory Data Analysis (EDA)

Data Visualization: Multivariate Visualisations → Parallel Coordinates



Exploratory Data Analysis (EDA)

Key processes of EDA

- **Data Cleaning and Preparation** : Ensures data quality
- **Descriptive Statistics** : Summarises key metrics
- **Data Visualization** : Makes data insights visible
- **Pattern Recognition** : Identi fi es trends and anomalies

Exploratory Data Analysis (EDA)

Pattern Recognition

Discover trends, clusters, and relationships in data

Key Techniques:

- Clustering
- Dimensionality Reduction

Examples:

Customer segmentation, detecting seasonal patterns

Exploratory Data Analysis (EDA)

Key processes of EDA

- **Data Cleaning and Preparation** : Ensures data quality
- **Descriptive Statistics** : Summarises key metrics
- **Data Visualization** : Makes data insights visible
- **Pattern Recognition** : Identifies trends and anomalies
- **Hypothesis Formation and Testing :** Building and validation assumptions

Exploratory Data Analysis (EDA)

Hypothesis Formation

Frame questions for deeper analysis

Examples:

- "Does advertising expenditure affect sales?”
- "Is there a signi fi cant di ff erence in test scores across schools?"

Hypothesis formation directs the focus of the analysis!

Exploratory Data Analysis (EDA)

Hypothesis Testing

Validate assumptions using statistical tests

Concepts:

- Null Hypothesis (H0): No e ff ect or relationship
- Alternative Hypothesis (H1): Proposed effect or relationship exists

Steps:

1. Define hypothesis
2. Choose a test (e.g. t-test)
3. Analyze results (e.g. p-value < 0.05)

Hypothesis Formation Scenarios

Which of these are testable hypotheses?

A. Older users spend more on average
B. Blue is a nicer color
C. Salaries are higher in Paris than Nice

Which of these would guide deeper analysis?

Exploratory Data Analysis (EDA)

Challenges in EDA

- Messy data requires extensive cleaning
- Large datasets can cause computational challenges
- Risk of biased interpretation from visual patterns

Exploratory Data Analysis (EDA)

Best practices

A. Understand the data’s context and domain

B. Clean and preprocess data meticulously

C. Validate insights through statistical methods

D. Document the EDA process for reproducibility

Let’s do some exercises!

Spot the issues in this dataset

Name	Age	Score	Country
Alice	25	78	USA
Bob		90	usa
Claire	27	105	France
Alice	25	78	USA

Scores range is between 0 and 100

What’s wrong here?

Spot the issues in this dataset

Name	Age	Score	Country
Alice	25	78	USA
Bob		90	usa
Claire	27	105	France
Alice	25	78	USA

Scores range is between 0 and 100

Missing data, inconsistent case, impossible value, duplicate row.

Fixing missing values

Small dataset: Missing prices in product list

Which method is best?

A. Drop rows
B. Fill with mean
C. Fill with median
D. Predict from other features

Which plot to use?

Match each case to the right chart:

- Comparing sales trends over 12 months
- Showing distribution of exam scores
- Detecting relationship between weight & height

Options: Line plot, Histogram, Scatter plot

Is this a valid hypothesis?

1. 'There’s no difference in math scores between School A and School B.'
2. 'Bananas are better than apples.'

Question: Which is testable using statistics? Why?

Final recap quiz

Q1: Which statistic is least affected by outliers?

Q2: Which plot is best to check skewness?

Q3: Name one benefit of EDA before modeling.

Demo with Notebook_EDA.ipybn

Useful links:

- https://datavizcatalogue.com/index.html
- https://python-graph-gallery.com/