ICT @ Mzuni

Exploratory Data Analysis and the Data Science Process

Basic tools (plots, graphs and summary statistics) of EDA

Philosophy of EDA

The Data Science Process

Introduction to EDA

- Exploratory Data Analysis (EDA) is the process of analysing and visualizing data to understand its main characteristics before applying any machine learning algorithms or models.
- The goal of EDA is to make sense of the data, uncover patterns, spot anomalies, and check assumptions.
- In simple terms, the analysis of datasets based on various numerical methods and graphical tools.

Importance of EDA in Data Science

- By exploring the data, you get a sense of its distribution, relationships between features, and possible patterns that help decide which machine learning model is most suitable.
- It allows you to create visualizations (charts, graphs) to clearly see trends, correlations, or distributions, which can be much easier to understand than raw numbers.

Exploratory vs Confirmatory Data Analysis

EDA	CDA
What is in the data?" Goal: **Explore**, discover patterns, outliers, and trends. • No hypothesis — just curiosity. • No hypothesis at first • Generate hypothesis • Uses graphical methods (mostly)	*Is what I believe true about the data?"* Goal: **Test** a specific hypothesis. • Seeks **evidence** to support or reject a claim. • Start with hypothesis • Test the null hypothesis • Uses statistical models

The Data Science Process

A structured way to turn data into insights.

1. Problem Understanding – What are you trying to solve?
2. Data Collection – Gather relevant data.
3. Data Cleaning & Preparation – Fix errors, handle missing values.
4. Exploratory Data Analysis (EDA) – Understand patterns and relati
5. Modelling – Build algorithms to make predictions or classificatio
6. Evaluation – Test how well your model works.
7. Deployment – Use the model in real-world applications.
8. Monitoring – Track model performance over time.

Philosophy of EDA (Exploratory Data Analysis)

- Let the data speak for itself.
- EDA is about exploring data before modeling.
- It helps you understand patterns, spot errors, and find relationships.
- You use visuals, summaries, and plots (like histograms, boxplots, correlations).
- Goal: Know your data well before jumping into predictions.

Classification of EDA Methods

- EDA methods can be classified in two main ways:
  - Graphical vs. Non-Graphical
  - Univariate vs. Multivariate

Graphical vs. Non-Graphical Methods

- Graphical methods use visual representations to summarize data.
- These may be histograms, box plots, scatter plots, heatmaps, etc.
- Non-graphical methods rely on summary statistics and numerical measures. Examples include mean, median, mode, standard deviation, variance, correlation coefficient, etc.

Univariate vs. Multivariate Methods

- Univariate methods analyse one variable at a time. Example: Analysing the distribution of a single column, such as the average income of a population.
- Multivariate methods generally analyse many variables simultaneously.
  - Bivariate analysis is a subset of multivariate analysis that focuses on two variables at a time. Example: Exploring the relationship between income and education level.
  - Multivariate analysis examines three or more variables together. For instance, analysing how income, education level, and age interact to predict j

Univariate vs. Multivariate Methods

- It is almost always a good idea to perform univariate EDA on each of the components before performing multivariate EDA.
- This is important because:
  - Univariate EDA helps identify outliers, missing values, and data distribution.
  - Provides a foundation for understanding variables before exploring relationships.

Preliminary Steps

- Critical steps in EDA that should be performed before diving into summary statistics or exploring relationships between variables or generally understanding the data's structure, patterns, and potential insights are:
  - Outlier detection: Spotting unusual data points that can skew results
  - Missing value analysis: Understanding where data is incomplete

Summary Techniques

- Summary techniques focus on using statistics to give a quick snapshot of the data.
- These numbers help you understand what's happening in your dataset at a high level.
- Common summary techniques include:
  - Mean
  - Median
  - Mode
  - Standard Deviation
  - Percentiles and Q

Summary Techniques: Percentiles and Quartiles

- These show how the data is divided into parts.
- For example, the 25th percentile is the value below which 25% of the data falls.
- A percentile divides a dataset into 100 equal parts, while a quartile divides a dataset into only four equal parts;

Identifying Outliers

- Outliers are data points that differ significantly from the rest of the data in a dataset.
- These values are much higher or lower than the majority of the other data points and can potentially skew or mislead statistical analyses.
- Outliers can result from variability in the data, errors, or rare occurrences, and they may need to be investigated or handled appropriately in data analysis.

Identifying Outliers

- If you have extreme values, they might affect your analysis or model.
- For example, if most people earn between MK30,000 and MK70,000 per year, but one person earns MK1,000,000, this could distort your findings.
- Outliers can be detected using statistical methods, visualization tools (like box plots), or algorithms designed to identify unusual data points.
- Common methods for detecting outliers are:
  - IQR method (may be visualised using box plots)
  - Z-score (standard deviations from mean)

Z-Score Method

- The Z-Score method is a statistical technique used to identify outliers by measuring how many standard deviations a data point is from the mean of the dataset. It is useful when the data follows a normal distribution.

Formula:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X = Data point
- μ = Mean of the dataset
- σ = Standard devi

Bivariate EDA

- Exploring relationships between two variables using summary statistics.
- Common methods:
  - Correlation
  - Contingency Tables (Cross-tabulations
  - Covariance

Correlation

A measure of how two **numbers** move together.

- **Example:**
  If kids who study more **tend to score higher**, there's a **positive correlation** between study time and scores.
- Positive correlation: both go up (e.g., height and weight)
- Negative correlation: one goes up, the other goes down (e.g., exercise and weight)
- 0 correlation: no relationship
- It is used check whether one thing increases/decreases when another one does.

Contingency Table (Cross-tabulation)

- A table that shows how **categories** relate to each other.
- This helps you see how preferences vary by group.
- **Used for:**
  Comparing counts across categories (e.g., gender vs. preferen
- **Example:**
  You survey 100 people on gender and if they like coffee:

Gender	Like Coffee	Don't Like Coffee
Male	30	20
Female	40	10

Covariance

- Tells whether two numbers change together (like correlation),

Example:
If temperature and ice cream sales go up together, their covariance is positive.

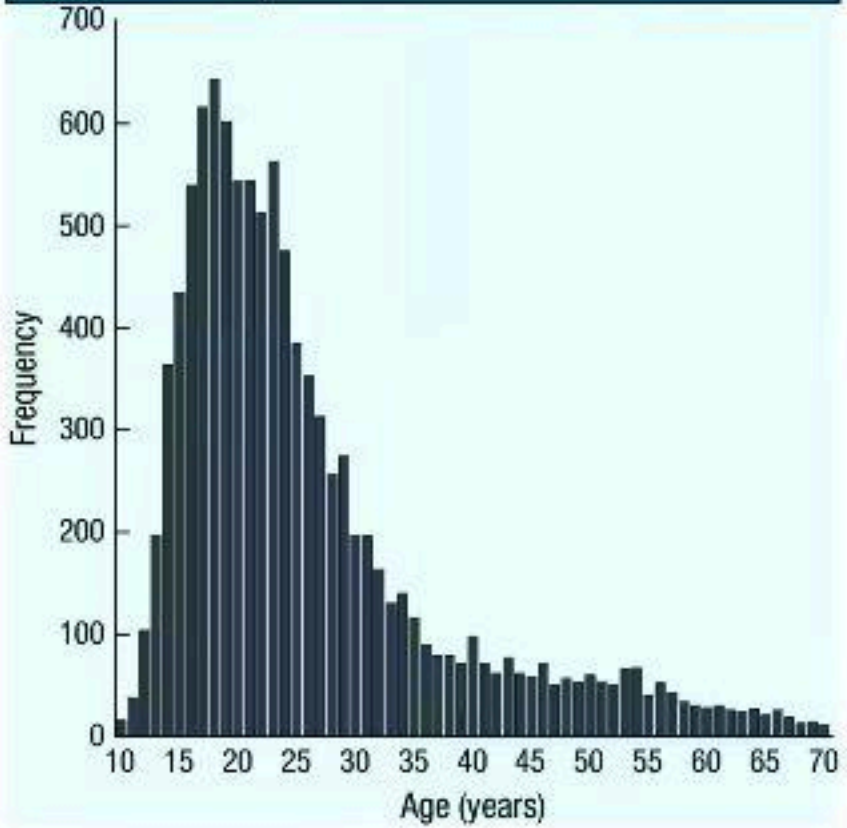
- Positive: they rise/fall together
- Negative: one rises while the other falls
- 0: no consistent movement
- It is used in understanding the direction of relationship between two numeric variables, but

Height & Weight (0.95): Very strong positive relationship → taller people weigh more.
Age & Score (0.20): Weak positive relationship → age doesn't explain much of the score.
Score & Height (0.60): Moderate positive → maybe taller students scored a bit higher.

	Height	Weight	Age	Score
Height	1.00	0.95	0.45	0.60
Weight	0.95	1.00	0.50	0.55
Age	0.45	0.50	1.00	0.20
Score	0.60			1.00

Understanding Data Distributions : Histograms

Histogram: Shows the frequency distribution of a single continuous variable.



ICT @ Mzuni

Understanding Data Distributions : Histograms

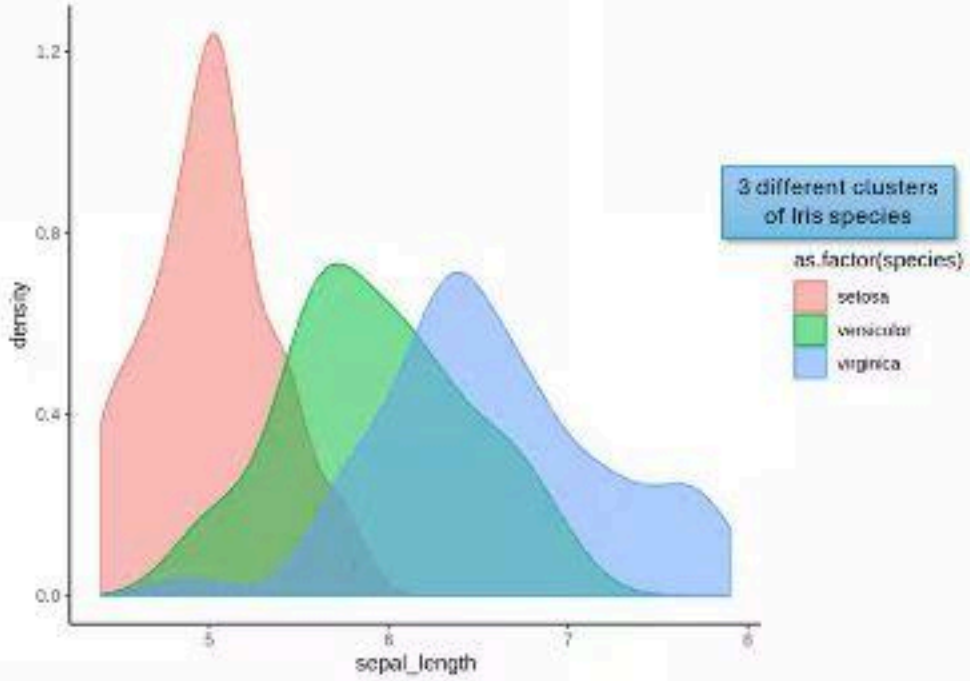
- Why Use a Histogram?
  - To understand the distribution of data (e.g., normal, skewed, uniform).
  - To detect skewness, peaks, and gaps in the dataset.To identify outliers or unusual patterns.
- Common Histogram Shapes:
  - Normal (Bell-shaped): Symmetric around the center.
  - Right-skewed (Positively skewed): Tail extends to the right (e.g., income distribution).
  - Left-skewed (Negatively skewed): Tail extends to the left.
  - Uniform: Bars are roughly the same height, indicating an even distribution.

Understanding Data Distributions :Density Plots

- A density plot is a smoothed version of a histogram, where instead of showing the count of data points in discrete bins, it shows a continuous probability density function (PDF).
- Density plots are useful for understanding the overall shape and spread of a dataset.
- Peaks indicate the regions where the data is concentrated (i.e., the most frequent values), and valleys represent areas where there are fewer data points.
- An advantage with density plots over histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram).

Understanding Data Distributions :Density plot

The figure below shows a density plot of sepal length grouped by the flower species.

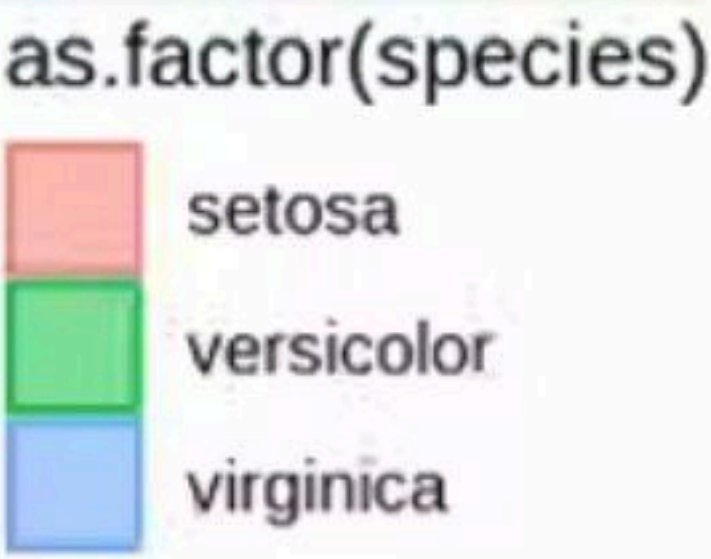


ICT @ Mzuni

tions .Density p

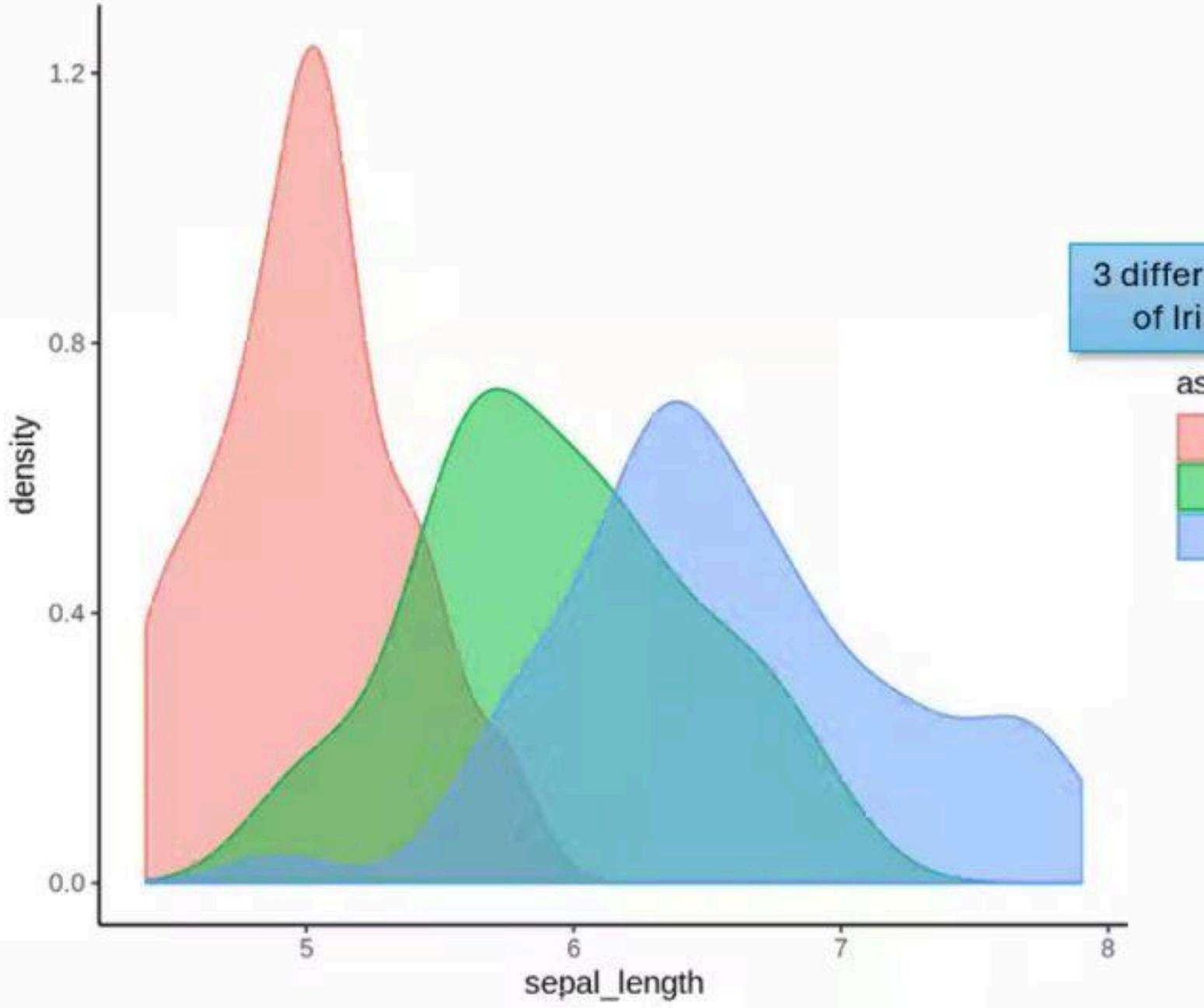
t of sepal length grou

3 different clusters of Iris species



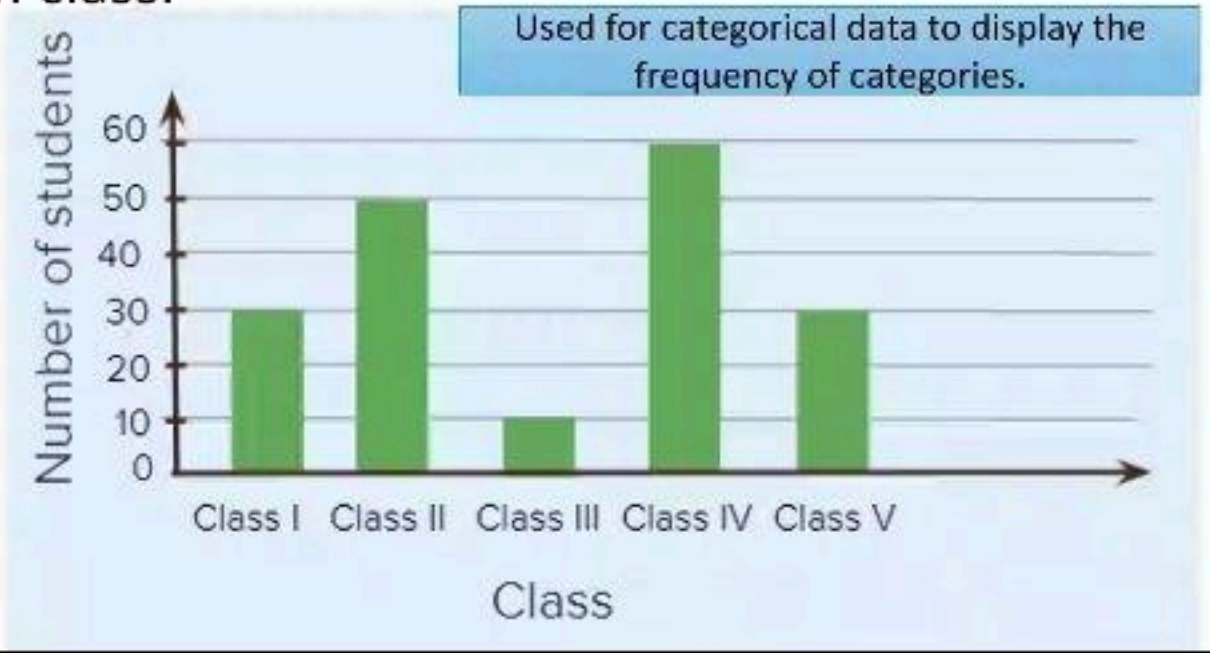
derstanding Data Distributions

figure below shows a density plot of sep
he flower species.



Relationships and Comparisons: Bar Charts

- Bar charts show comparisons between different categories.
- For example, a bar chart might show the number of students in each class.



Relationships and Comparisons: Scatter Plots

- Scatter plots show the relationship between two numerical variables.
- For example, you could see if there's a connection between the hours someone studied and their test scores.