Exploratory Data Analysis & Cleaning

DSCUBED
UNIVERSITY OF MELBOURNE

Agenda

01

Introduction:

02

Data analysis concepts

03

Hands on practise

04

Q&A

Meet the Speakers

Jordan
Education Director

Stanley
Education Officer

Motivation: Why EDA?

Why do we need EDA?

🔍 **Understand the data**: Know what you're working with — the distributions, types, and quirks of your features.

🧹 **Catch issues early**: Spot missing values, outliers, or data leakage before they hurt your model.

🧠 **Guide your ML workflow**: Use insights from EDA to inform preprocessing, feature engineering, and model choice.

Spend the most time cleaning and preprocessing your data

Garbage in = Garbage Out

Getting to Know Your Data

Dataset Statistics

Distribution

- imbalance
- skewness
- outliers

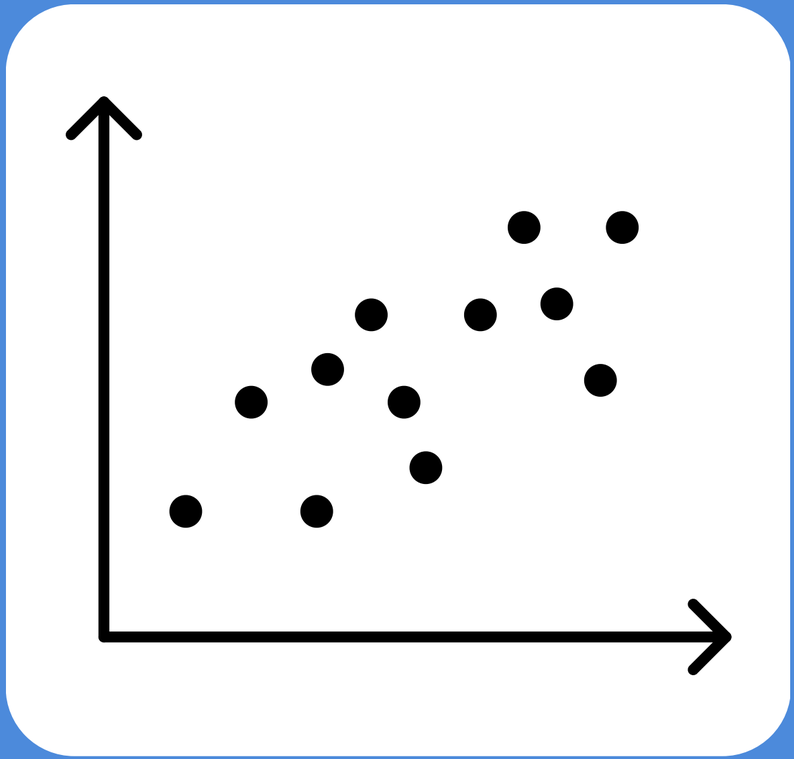
Mean

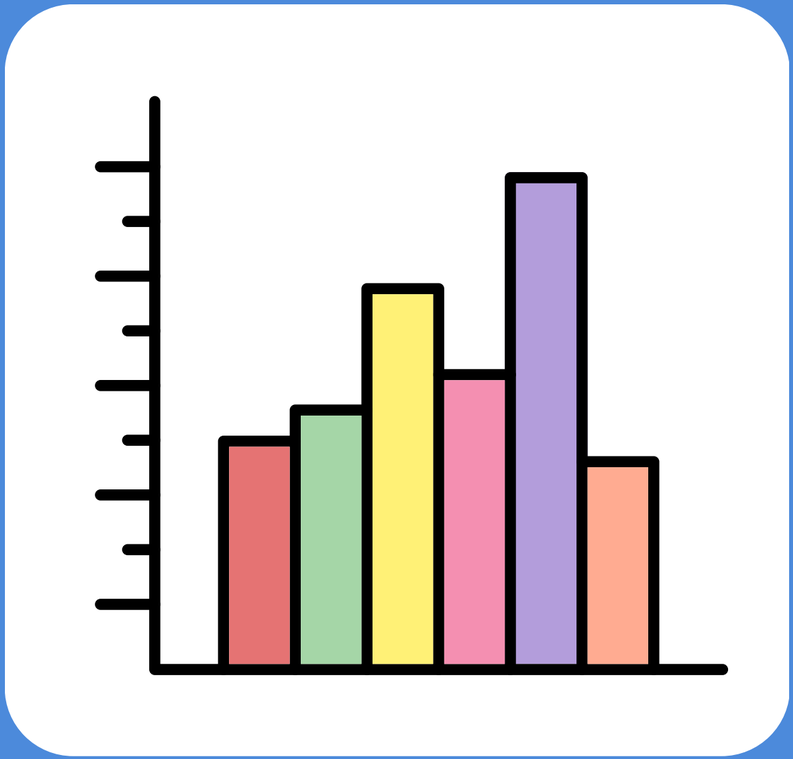
- tendency
- differences across groups

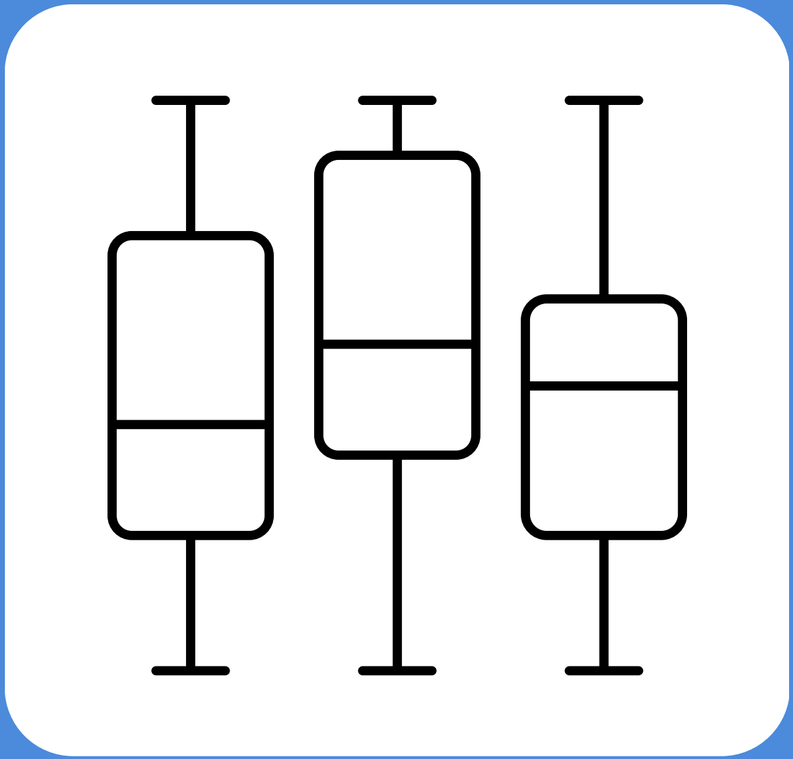
Variance

- feature spread
- normalization?

Visualisations







Dataset Structure

`.head(), .info(),.describe`

	index	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income
8	8	-122.26	37.84	42.0	2555.0	665.0	1206.0	595.0	2.0804
10	10	-122.26	37.85	52.0	2202.0	434.0	910.0	402.0	3.2031
11	11	-122.26	37.85	52.0	3503.0	752.0	1504.0	734.0	3.2705
12	12	-122.26	37.85	52.0	2491.0	474.0	1098.0	468.0	3.0750
13	13	-122.26	37.84	52.0	696.0	191.0	345.0	174.0	2.6736

Feature Correlations

Pearson	best for linear relationships, values between -1 and 1
Mutual information	can handle non-linear & both numerical and categorical data
Entropy	measure of uncertainty, used in DTs for splitting

Cleaning the Dataset

Handling Missing Values

Drop

Impute (mean, median, mode, model based)

Handling Outliers

🗑️ **Remove**

if they're irrelevant and their absence won't bias the findings

🔄 **Transform**

log or sqrt the data to reduce the sparsity

⚖️ **Standardize**

transform the data so that it follows Normal, mean=0, var=1

Data Type Corrections & Encoding

Often we need to change data values to suitable types for analysis
E.g. “42” to 42
yes/no to True/False

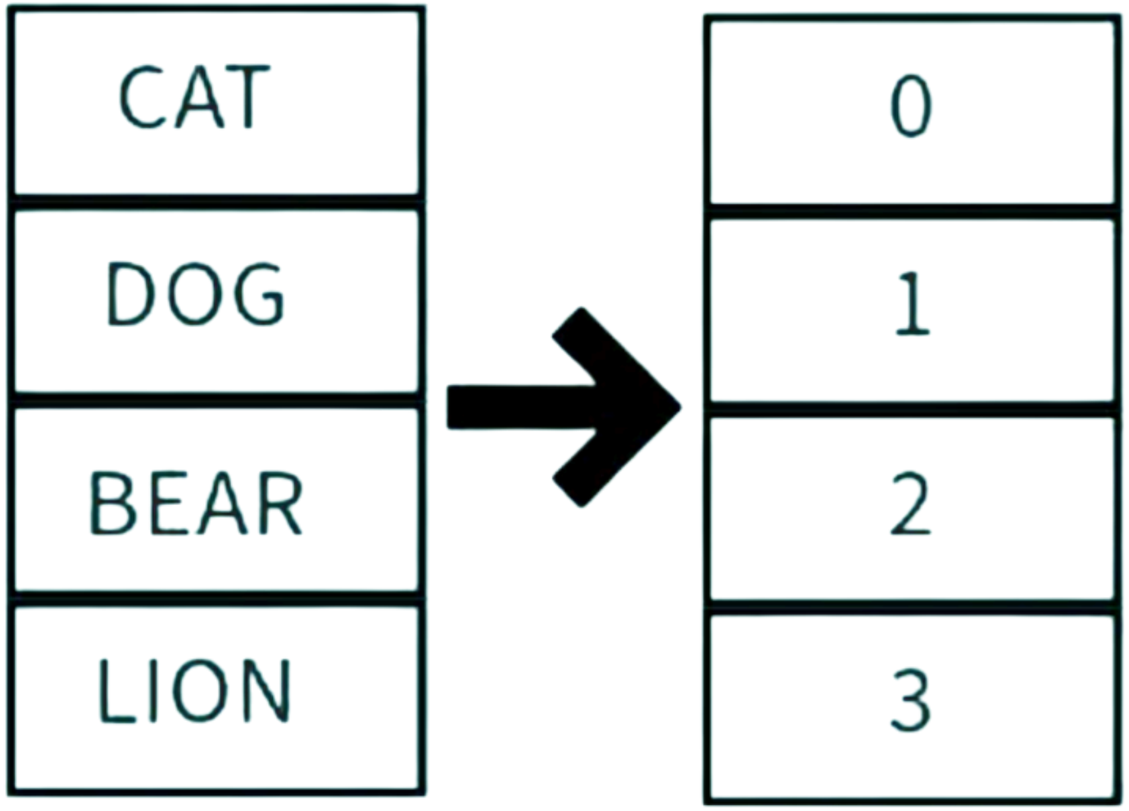
We can use one-hot encoding or label encoding for replacing categorical data with numerical data

One-Hot Encoding

datagy.io

Island	Biscoe	Dream	Torgensen
Biscoe	1	0	0
Torgensen	0	0	1
Dream	0	1	0

Label encoding



Feature Scaling & Transformations

Feature Scaling

bringing data features to the same scale

Transformation

involves modifying features to support the learning algorithm

Hands on Practise (Collab Notebook)

https://shorturl.at/KdPro

Q&A

Thank You!!