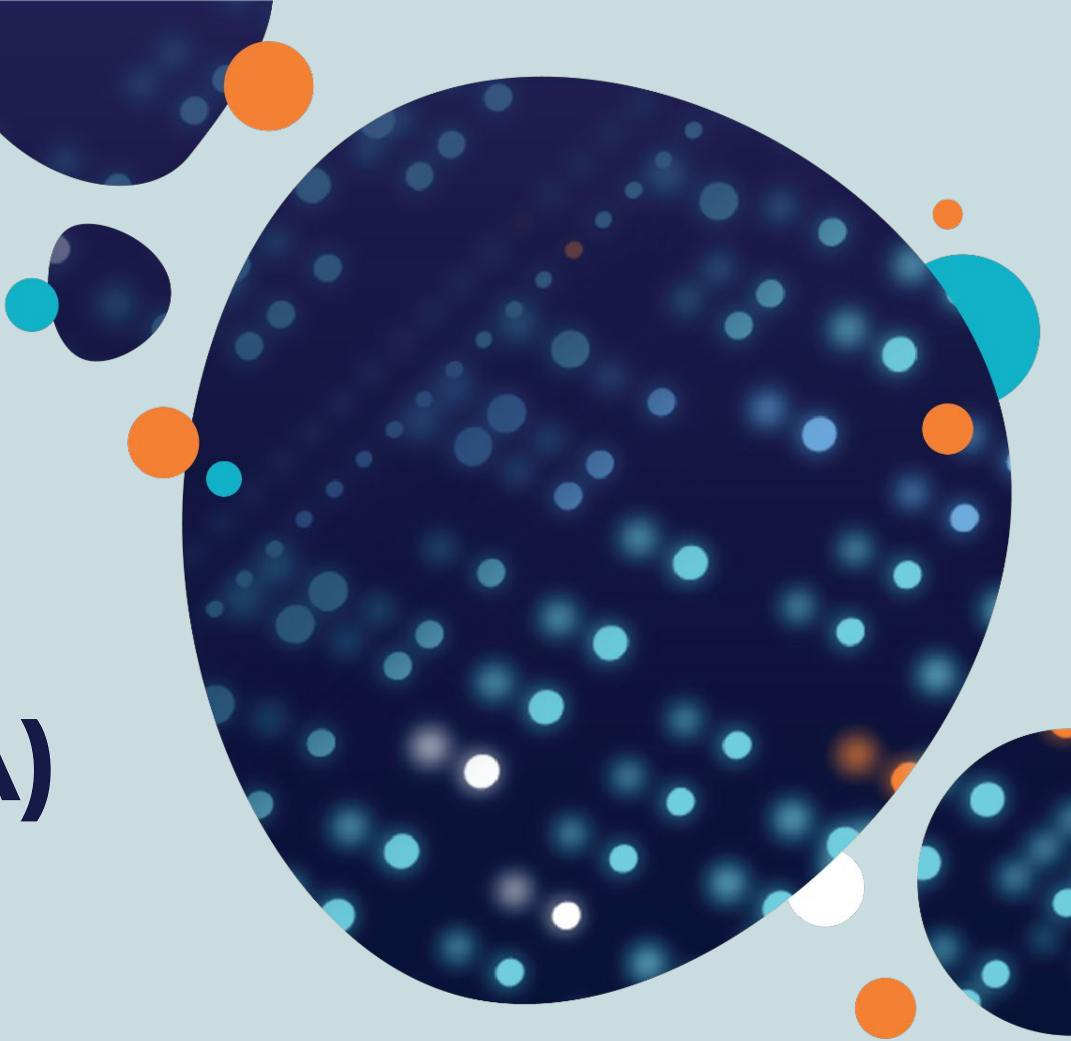
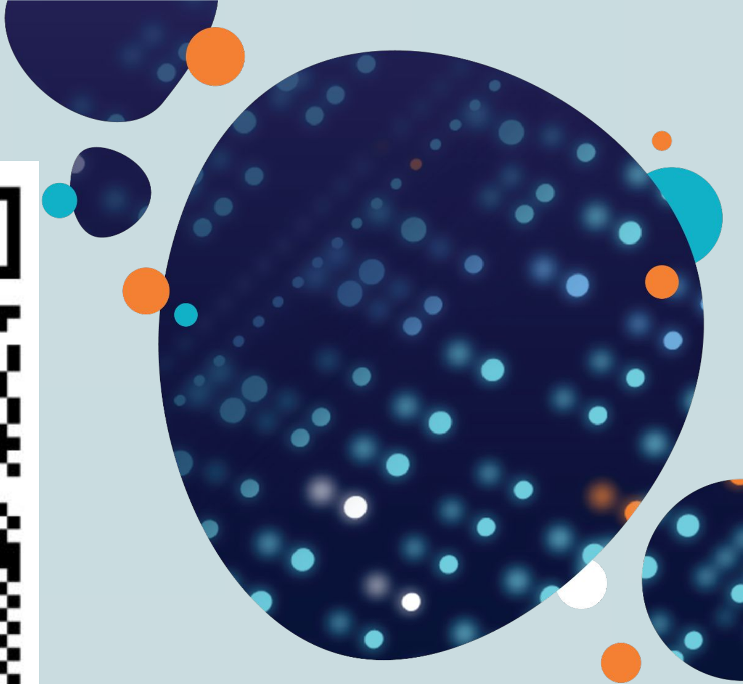
data science
student society

Exploratory Data Analysis (EDA)

November 20, 2024



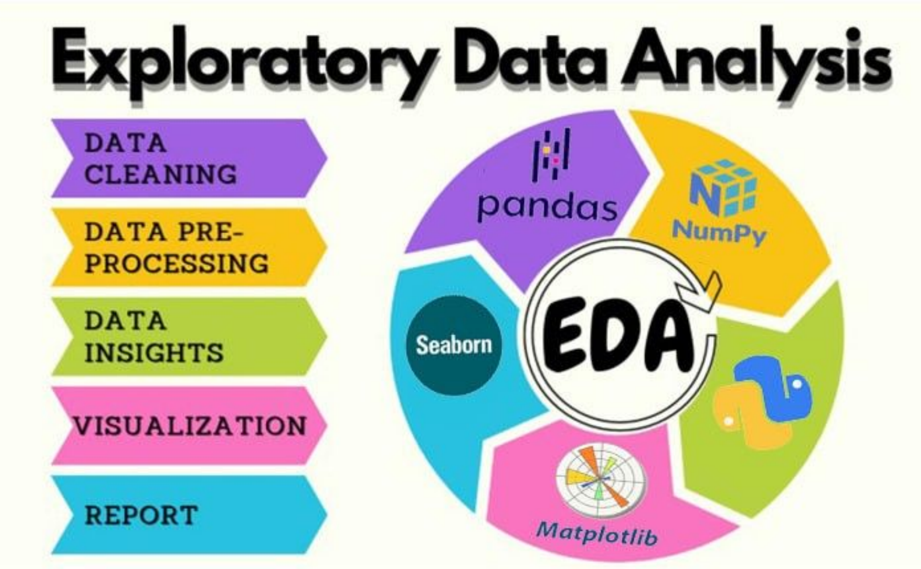
data
science
student
society



Please check in!

What is EDA?

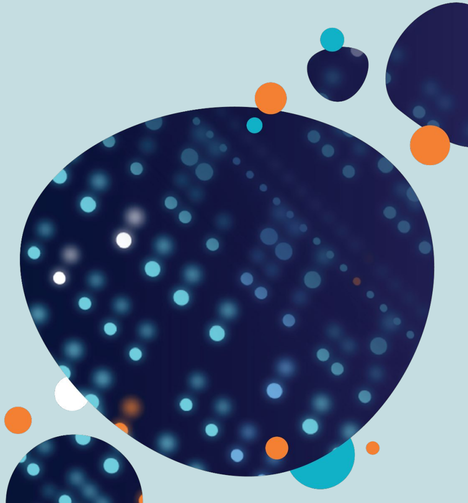
- Exploratory Data Analysis is the process of analysing datasets to summarize their main characteristics, often using visual methods.
- EDA helps you become more familiar with your data, identify patterns, detect anomalies, and check assumptions before moving on to more formal modeling


Exploratory Data Analysis

Why We Use EDA

Key objectives:

- Understanding the structure of the data: shape, size, distribution of data
- Identifying errors: spot missing values, duplicates, or incorrect data types
- Detecting patterns, relationships, or anomalies
- Generating hypotheses: develop insings that guide your modeling decisions
- Communicate findings: visual summaries make your data comprehensible



This Workshop

Goals

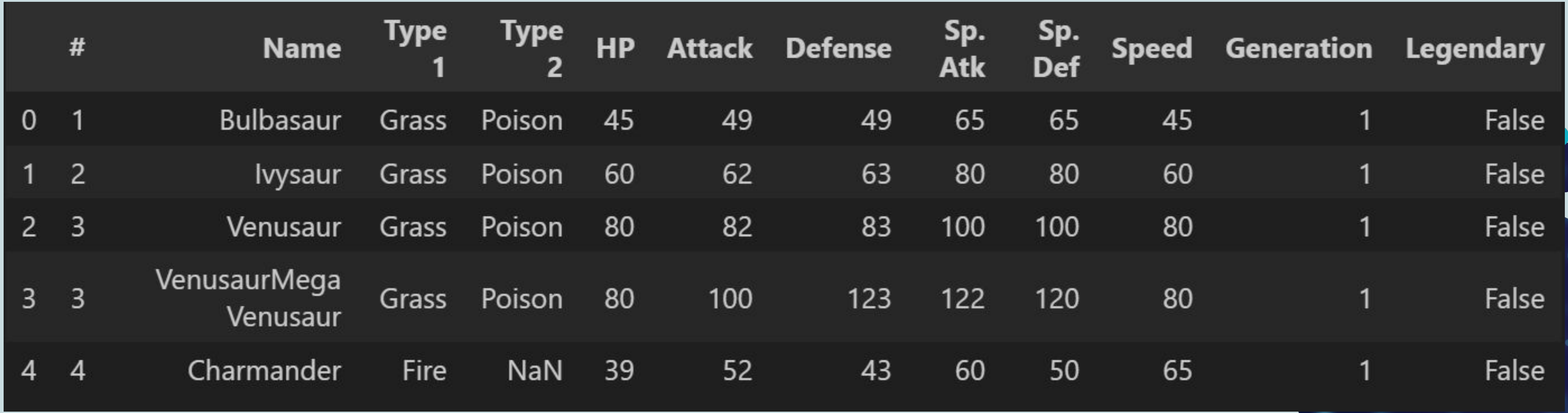
- Introduction to the dataset
- Data cleaning and preparation
- Descriptive and comparative analysis
- Data visualization
- Advanced analysis and feature exploration

https://github.com/mekapur/EDA.git

Notebook and Dataset

Clone the repo here: https://github.com/mekapur/EDA.git

Here is the first 5 rows of the data we'll be looking at:



	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	39	52	43	60	50	65	1	False

Dataset Overview

What to look for:

- Size: how many rows (observations) and columns (features) are there in the dataset?
- Columns: what variables are available? (HP, attack, type, legendary status)
- Data types: are columns numerica, categorical, or text?
- Sample data: look at the first few rows to get an idea of its structure

Data Cleaning

Cleaning data ensures that analyses are accurate and reliable

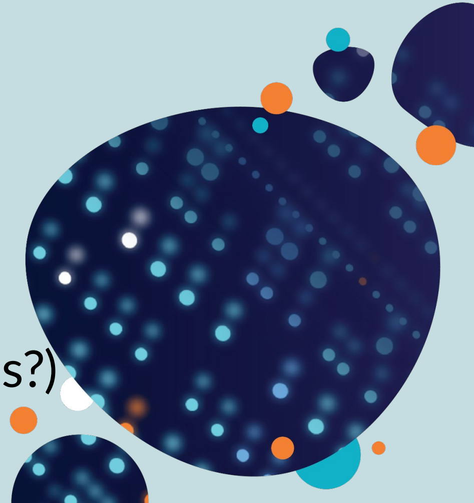
1. Missing values: some data points may be missing due to reasons like errors in data collection
    - Possible solutions: fill in missing values with an appropriate placeholder (e.g. ‘None’) or use mean/mode for numerical data
2. Duplicate entries: identical rows may exist due to repeated data entry
    - Remove duplicates to prevent skewing results
3. Incorrect data types: numeric data may be read as text or vice versa
    - Convert columns to their correct types

Descriptive Analysis

Descriptive statistics help us summarize and interpret the data before diving into deeper analysis

1. Summary statistics: mean, median, min, max, and standard deviation give insights into data distribution
2. Group comparisons: compare groups (e.g. Legendary vs Non-legendary Pokemon)
3. Outliers: identify extreme values that might affect analyses

(What are the average stats for each Pokemon type?
How do legendary Pokemon compare to non-legendary ones?)

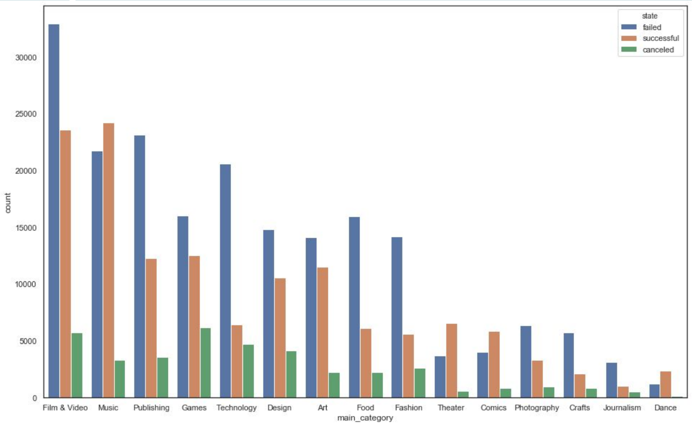


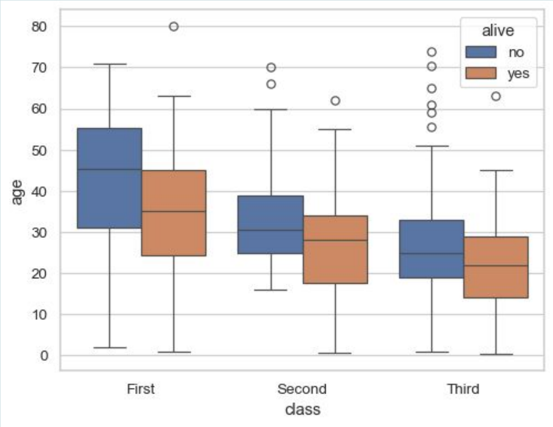
https://github.com/mekapur/EDA.git

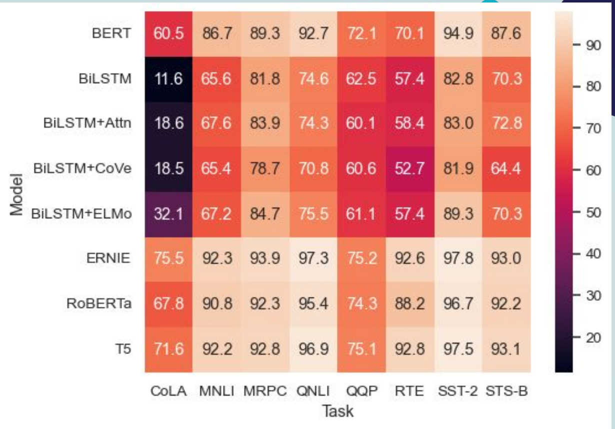
Visualization

Visualizations make patterns and relationships in data more apparent and easier to interpret

1. Count plots: shows distribution of categorical variables
2. Box plots: highlight the spread and detect outliers
3. Heatmaps: reveal correlations between numerical variables





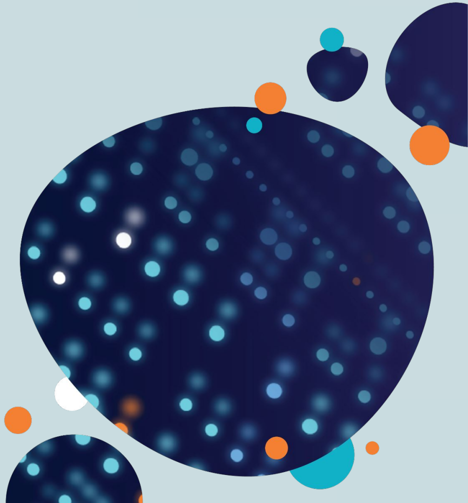


	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B
BERT	60.5	86.7	89.3	92.7	72.1	70.1	94.9	87.6
BiLSTM	11.6	65.6	81.8	74.6	62.5	57.4	82.8	70.3
BiLSTM+Attn	18.6	67.6	83.9	74.3	60.1	58.4	83.0	72.8
BiLSTM+CoVe	18.5	65.4	78.7	70.8	60.6	52.7	81.9	64.4
BiLSTM+ELMo	32.1	67.2	84.7	75.5	61.1	57.4	89.3	70.3
ERNIE	75.5	92.3	93.9	97.3	75.2	92.6	97.8	93.0
RoBERTa	67.8	90.8	92.3	95.4	74.3	88.2	96.7	92.2
T5	71.6	92.2	92.8	96.9	75.1	92.8	97.5	93.1

Advanced Analysis & Feature Exploration

Advanced techniques allow deeper insights and open up new perspectives on the data

- Scikit-learn (sklearn)
- K-Means clustering
- Feature engineering



https://github.com/mekapur/EDA.git

Scikit-Learn

Scikit-Learn (sklearn) is a powerful Python library for machine learning and data analysis

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Model evaluation and selection

Why we use it

- Easy to use API for common machine learning tasks
- Integrates well with other Python libraries ike Pandas and NumPy
- We will use K-Means clustering to group Pokemon based on their stats

K-Means Clustering

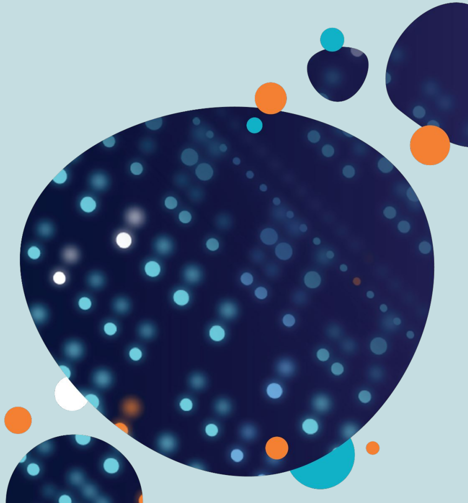
K-Means is an unsupervised ML algorithm that groups data points into a specified number of clusters based on their similarities

1. Choose k (number of clusters)
2. Assign points: each data point assigned to the nearest cluster center
3. Update centers: cluster centers recalculated based on the mean of points in each cluster
4. Repeat until clusters stabilize

Why we use it

- Identify patterns and natural groupings in data
- Useful for exploratory analysis when labels are not available

We will group Pokemon based on stats such as HP, attack, defense, and speed



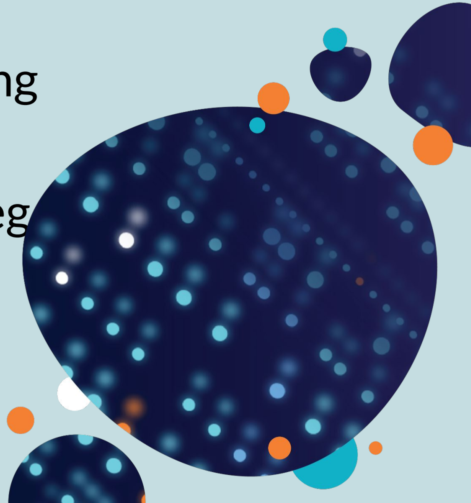
Feature Engineering

Feature engineering involves creating new features or modifying existing ones to improve analysis or model performance

- Helps uncover hidden patterns or relationships
- Provides additional insights by synthesizing data

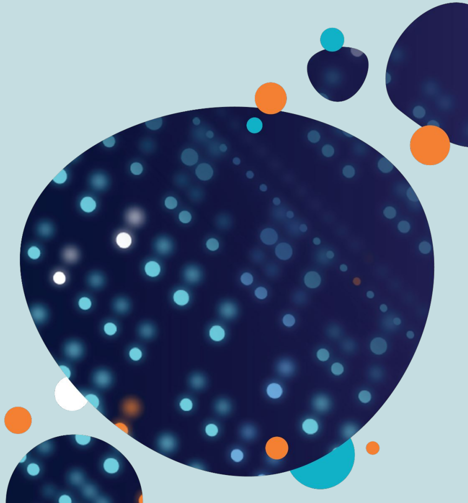
Examples

1. Summing stats: we create a Total Power feature by summing HP, attack, defense, and speed
2. Transformations: scaling features for consistent analysis (eg standardizing data for clustering)

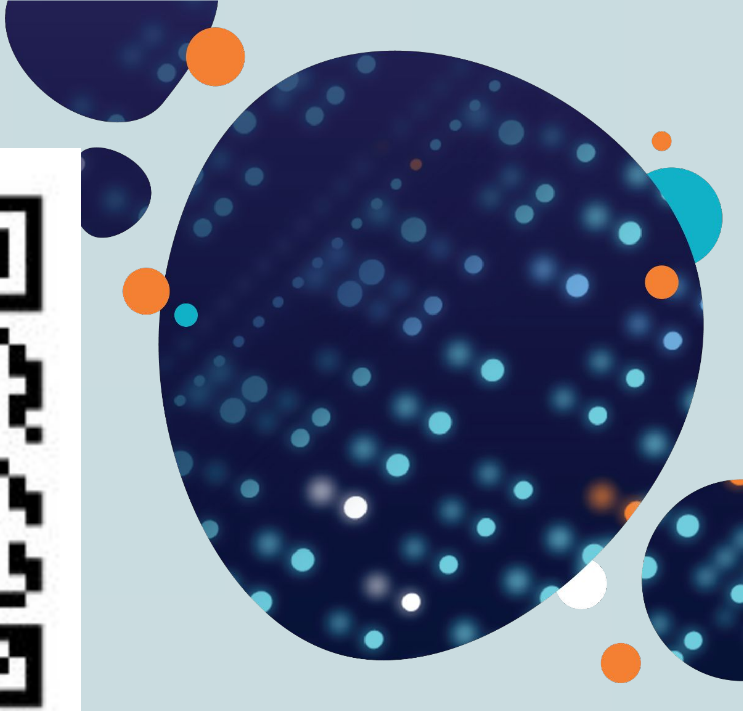


Key Ideas

- EDA is the foundation for any data analysis project
- Clean data leads to accurate and meaningful insights
- Visualizations help uncover hidden patterns and communicate findings effectively



data science student society



Leave your feedback here!