🔎 What is EDA?

Exploratory Data Analysis (EDA) is the process of **exploring, understanding, and summarizing a dataset** before applying any machine learning or statistical models.
It’s like meeting your dataset for the first time — you ask questions, look for hidden patterns, and clean up the mess before moving forward.

🧩 Why is EDA important?

1. **Understand structure** → number of rows, columns, datatypes.
2. **Detect data quality issues** → missing values, duplicates, outliers.
3. **Reveal patterns & relationships** → correlations, distributions, clusters.
4. **Generate hypotheses** → "Does income affect spending?"
5. **Guide feature engineering & modeling** → helps in choosing the right approach.

🛠️ Key Steps in EDA

1. **Data Collection & Loading**
   - Bring the dataset into your environment (CSV, Excel, SQL, APIs).
2. **Data Cleaning**
   - Handle missing values.
   - Fix inconsistencies (datatypes, formatting).
   - Remove duplicates.
3. **Data Profiling**
   - Shape of dataset (rows × columns).
   - Data types of each feature.
   - Summary statistics (mean, median, min, max, quartiles).
4. **Univariate Analysis** (one variable at a time)
   - For categorical: value counts, bar plots.
   - For numerical: histograms, boxplots, distribution curves.
5. **Bivariate / Multivariate Analysis** (relationships between variables)
   - Scatter plots, correlation heatmaps.
   - Grouping and pivot tables.
6. **Outlier Detection**
   - Boxplots, z-scores, IQR method.

7. **Feature Relationships**
    - Correlation (Pearson, Spearman).
    - Categorical vs numerical (ANOVA, chi-square).
8. **Visualization**
    - Tells the story with graphs (seaborn, matplotlib, plotly).

🎯 Goal of EDA

By the end of EDA, you should:

- Know the **quality** of your data.
- Have **insights** about distributions & relationships.
- Be ready to **engineer features** or transform data.
- Decide which models or statistical tests could make sense.

🔢 Data Handling & Cleaning

- **pandas** → backbone of EDA; used for loading, cleaning, and manipulating tabular data.
- **numpy** → efficient numerical computations, array handling, and math operations.

📊 Visualization

- **matplotlib** → base plotting library (low-level, very flexible).
- **seaborn** → built on matplotlib; makes statistical plots easier (correlation heatmaps, boxplots, distributions).
- **plotly** → interactive visualizations, great for dashboards & web apps.
- **missingno** → quick visualization of missing data patterns.

📈 Statistical Analysis

- **scipy.stats** → hypothesis testing, correlations, ANOVA, etc.
- **statsmodels** → deeper statistical modeling and summaries.

⚡ Automated/Advanced EDA Tools

- **pandas-profiling** (now called **ydata-profiling**) → generates full EDA reports automatically.

- **sweetviz** → visual, story-like EDA reports comparing datasets.
- **dtale** → interactive web-based pandas viewer.