Module-2.3 Exploratory Data Analysis (EDA)

Before we throw models like Random Forest or Neural Networks on our data, we need to **talk to the data**.

EDA is that conversation.

We ask: *What does my data look like? What is normal? What is weird? What relates to what?*
If you skip this step, you are basically doing *‘andha ML’* — blind machine learning.

1. What is EDA & Why It Matters

It is Systematic process of **summarizing** and **visualizing** data to:

- Understand distributions (are features skewed? heavy-tailed?)
- Spot data quality issues (outliers, inconsistent categories, etc.,)
- Discover relationships between features and target
- Generate hypotheses about which features / models might work

If you don’t understand the data:

- You might pick completely wrong models (e.g., linear model when relationship is clearly non-linear).
- You might miss important interactions (e.g., “age effect is different for men and women”).

2. Basic Stats & Distributions

We start with:

- Measures of **central tendency**: mean, median, mode
- Measures of **spread**: variance, standard deviation, IQR
- Distribution shape: symmetric, skewed, heavy-tailed

2.1 Central Tendency

For a numeric feature x1,x2,…,xn:

- **Mean**:

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

- **Median**: middle value when sorted
- **Mode**: most frequent value

- Mean: good if distribution is roughly symmetric and no extreme outliers.
- Median: robust to outliers (if one salary = 1 crore, median still stable).
- Mode: more relevant for categorical or discrete data.

	age	income	area_sqft	bedrooms	distance_km	house_price
count	500.000000	500.000000	500.000000	500.000000	500.000000	5.000000e+02
mean	35.052000	50168.834000	1499.271654	2.498000	12.877999	4.582259e+06
std	7.873836	49385.606864	582.191714	1.135153	6.792282	2.060355e+06
min	9.000000	232.000000	509.879962	1.000000	1.037563	1.057020e+06
25%	29.000000	13480.750000	974.750267	1.000000	7.246110	2.876980e+06
50%	35.000000	35090.000000	1533.307672	2.000000	12.952818	4.476452e+06
75%	40.000000	71093.250000	1967.870541	4.000000	18.502172	6.008516e+06
max	66.000000	309110.000000	2498.827452	4.000000	24.960340	1.171026e+07

2.2 Spread & Variability

For feature x:

- **Variance**:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- **Standard Deviation**:

$$\sigma = \sqrt{\sigma^2}$$

- Two features can have the same mean but different variability.

- Imagine you know the average of something (mean), but you also want to know: “How much do individual values usually deviate from that average?”

- Example:

  - Dataset A: {2, 4, 6}, Dataset B: {1, 4, 7}

- mean is the same but variance is different.

- Models like k-NN and clustering care a lot about scale and spread; features with larger variance can dominate distance.

- Actually, in high variance dataset it is hard to predict something, in low variance you can predict very close to actual value because of low spread of data.

2.3 Distribution Shape & Skewness

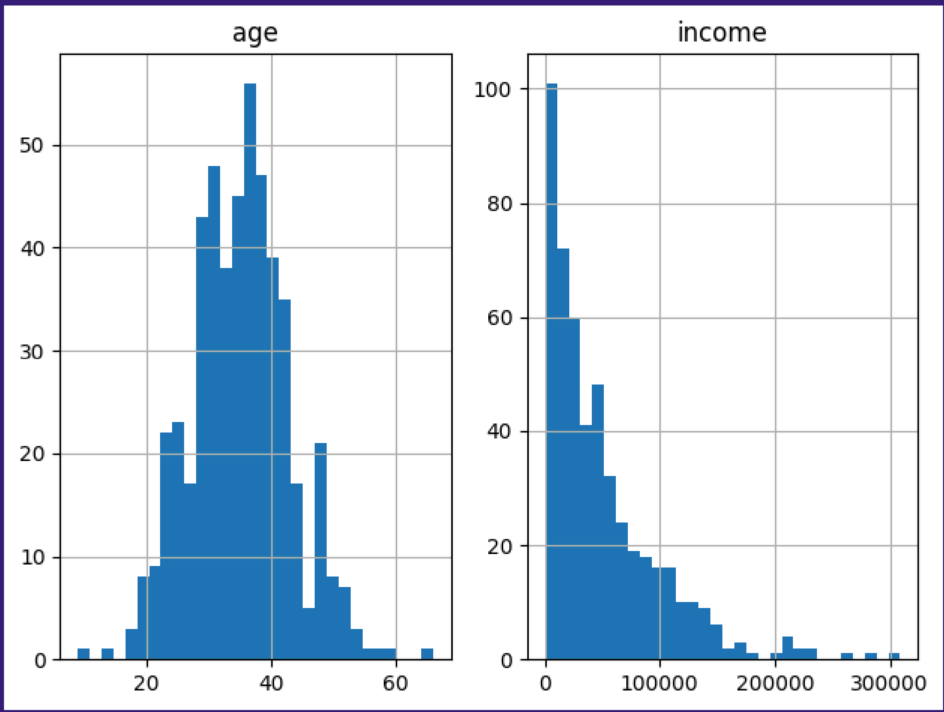
- **Symmetric** vs **skewed** distributions
- Many real-world features (income, house price, demand spikes) are **right-skewed**: long tail on the right.

Right-skewed: mean > median (few very large values).
Left-skewed: mean < median.

- Skewness suggests whether we might need transformations like **log**, **sqrt**, or **Box–Cox**.
- Many algorithms (especially linear models) work better when features are closer to normally distributed.

2.4 Visualizing Distributions:

Histogram



- Histogram shows how often values fall into different ranges.
- You can see skewness, multi-modality (multiple peaks), gaps.
- If you see two peaks → maybe two different markets (e.g., “flats” vs “villas”).

3. Correlations

We want to know: **How are features related to each other and to the target?**

3.1 Covariance

For two variables X and Y:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

Sample version:

$$\widehat{\mathrm{Cov}}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})$$

- Positive covariance: when X is above its mean, Y tends to be above its mean.
- Negative covariance: opposite pattern.

Problem: units are weird; not standardized.

3.2 Pearson Correlation Coefficient

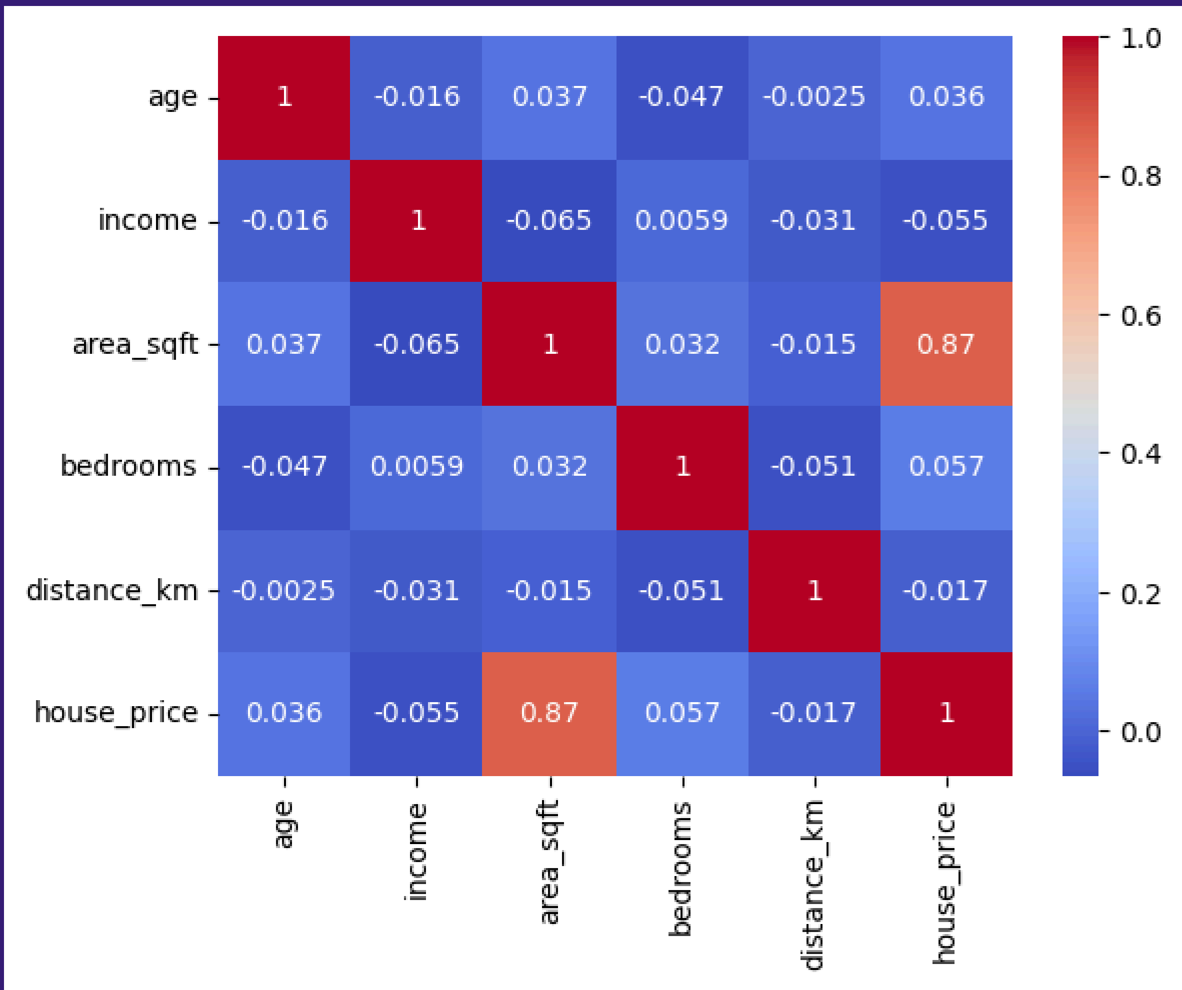
$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

- $\rho \in [-1, 1]$
- 1 = perfect positive linear relationship
- -1 = perfect negative linear relationship
- 0 = no **linear** relationship (but there might still be non-linear!)

```
house_price    1.000000
area_sqft      0.869095
bedrooms       0.056728
age            0.036184
distance_km   -0.016830
income        -0.055391
Name: house_price, dtype: float64
```

- Correlation helps you quickly see which features are **potentially useful** for predicting targets.
- "**Correlation is not causation.** Just because price and area correlate doesn't mean area causes price alone; many other confounders exist."

3.3 Correlation Heatmap



- Visually spot:
    - Which features are strongly correlated to the target.
    - Which features are strongly correlated with each other (multicollinearity).
- If two features are highly correlated (e.g., area in sq ft and area in sq meter), they provide almost redundant info; later, when doing linear models, this can cause instability.

4. Univariate, Bivariate, Multivariate EDA

Now we structure EDA into three levels.

4.1 Univariate EDA

Goal: Understand each feature individually.

Typical actions:

- For numeric:
  - Summary: `.describe()`
  - Histograms
  - Box plots to detect outliers
- For categorical:
  - Value counts
  - Bar plot

- Univariate EDA tells you:
  - Is the feature usable as-is?
  - Does it need transformation?
  - Are there extreme values or weird categories?
- You can also detect data collection problems (e.g., negative age, invalid categories).

4.2 Bivariate EDA

Goal: Understand relationships between **two variables** at a time.

Cases:

1. Numeric vs Numeric → scatter plot, correlation
2. Numeric vs Categorical → box plots, violin plots
3. Categorical vs Categorical → crosstab

4.3 Multivariate EDA

Goal: Look at **3 or more variables together** to see interactions.

Examples:

- Pair plots (scatter matrix).
- Grouped statistics (e.g., mean price by (city, number_of_bedrooms)).

6. Using EDA to Form Hypotheses (Feature & Model Ideas)

EDA is not just "plots for decoration".
We must **convert observations** → **hypotheses**.

6.1 Example: House Price Dataset

After EDA, we might conclude:

1. **price** is right-skewed ⇒
   - Hypothesis: apply log-transform to stabilize variance.
   - Use `log_price = log(price)` for modeling.
2. **price** has strong positive correlation with `area` ($r \approx 0.8$) ⇒
   - Hypothesis: area is a key feature. Linear regression may capture most variance using area.
3. **city** changes the relationship between area and price ⇒
   - Hypothesis: add interaction features (`area × city`), or use a tree-based model that handles interactions.
4. **bedrooms** has weak but non-zero correlation with price ⇒

    - Hypothesis: might help but not as much as area. Keep but don’t expect miracles.
5. **Some cities have very few data points ⇒**
    - Hypothesis: model might be unstable for those cities; maybe combine them into “Other”.

6.2 Typical “EDA → Model choice” connections

EDA Observation	Hypothesis / Model Idea
Strong linear patterns	Try linear/regularized models first
Curvy/threshold-like patterns	Try tree-based models or add polynomial features
Heavy skew / long tails	Apply log/Box–Cox transforms
Strong multicollinearity	Use regularization (Ridge/Lasso), or drop redundant features
Different pattern per group (city, segment)	Add interaction terms or build segment-specific models

Key line you can say:

“EDA doesn’t give answers. It gives **better questions**—good hypotheses.

Then models + validation tell us which of those hypotheses are true.”

EDA PLOT CHEAT SHEET

Plot	Why We Use It
Histogram	Understand distribution shape & skewness
Box Plot	Detect outliers & compare spread
KDE	Smooth distribution comparison
Bar / Count	Category frequency & imbalance
Scatter	Relationship between two numbers
Reg Plot	Average trend & linearity
Violin	Full distribution per category
Crosstab	Categorical interaction
Heatmap	Quick correlation scanning
Pair Plot	One-shot multivariate view
Line Plot	Trend & seasonality (time data)
Rolling Mean	Smooth noisy trends

KEY INTERPRETATION RULES

- Mean > Median → Right skew
- Mean < Median → Left skew
- Large std → high variability
- Strong correlation → potential predictor
- Non-linear scatter → tree / polynomial models
- Imbalanced target → accuracy is misleading