Applied AI & Machine Learning CS-333

Dr. Abbas Hussain

PNEC, NUST

Lecture 2

Spring 2026

Exploratory Data Analysis (EDA)

A First Look at the Data

Learning Outcomes

By the end of this lecture, students will be able to:

1. Define EDA and explain why it is the first step before applying AI/ML models.
2. Identify data types in engineering datasets (numerical, categorical, time-series, sensor signals).
3. Perform data quality checks (missing values, duplicates, impossible values, noise).
4. Compute and interpret summary statistics (mean, median, variance, standard deviation, IQR).
5. Analyze data distribution using histograms and interpret skewness/kurtosis.
6. Detect outliers using IQR and Z-score methods and discuss their impact in AI/ML.
7. Evaluate relationships between variables using correlation and understanding multicollinearity in feature sets.

Foundational supervised learning concepts

Supervised machine learning is based on the following core concepts:

- Data
- Model
- Training
- Evaluating
- Inference

Data

Data is the driving force of ML. Data comes in the form of words and numbers stored in tables, or as the values of pixels and waveforms captured in images and audio files.
We store related data in datasets. For example, we might have a dataset of the following:

- Images of cats
- Housing prices
- Weather information

- Datasets are made up of individual examples that contain features and a label.
- Examples that contain both features and a label are called labeled examples.

Data

Features

Label

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23

Example

Features

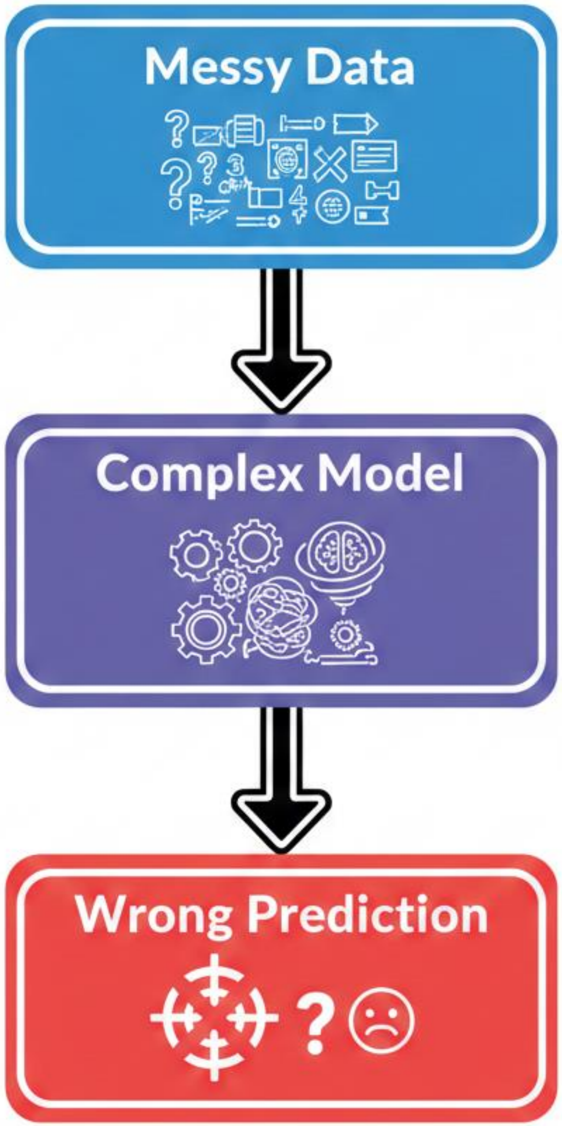
date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure
2021-09-09	49.71N	82.16W	74	20	3	N	18.6
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94

Example

Why Data Matters

"Garbage In ⟶ Garbage Out"

- 80% of an ML Engineer's time is spent cleaning and preparing data.



Types of Data

- **Structured:** Highly organized (Excel, SQL tables).
- **Semi-Structured:** Tags and markers (JSON, XML).
- **Unstructured:** The 'Wild West' (Images, Audio, Video, PDFs).

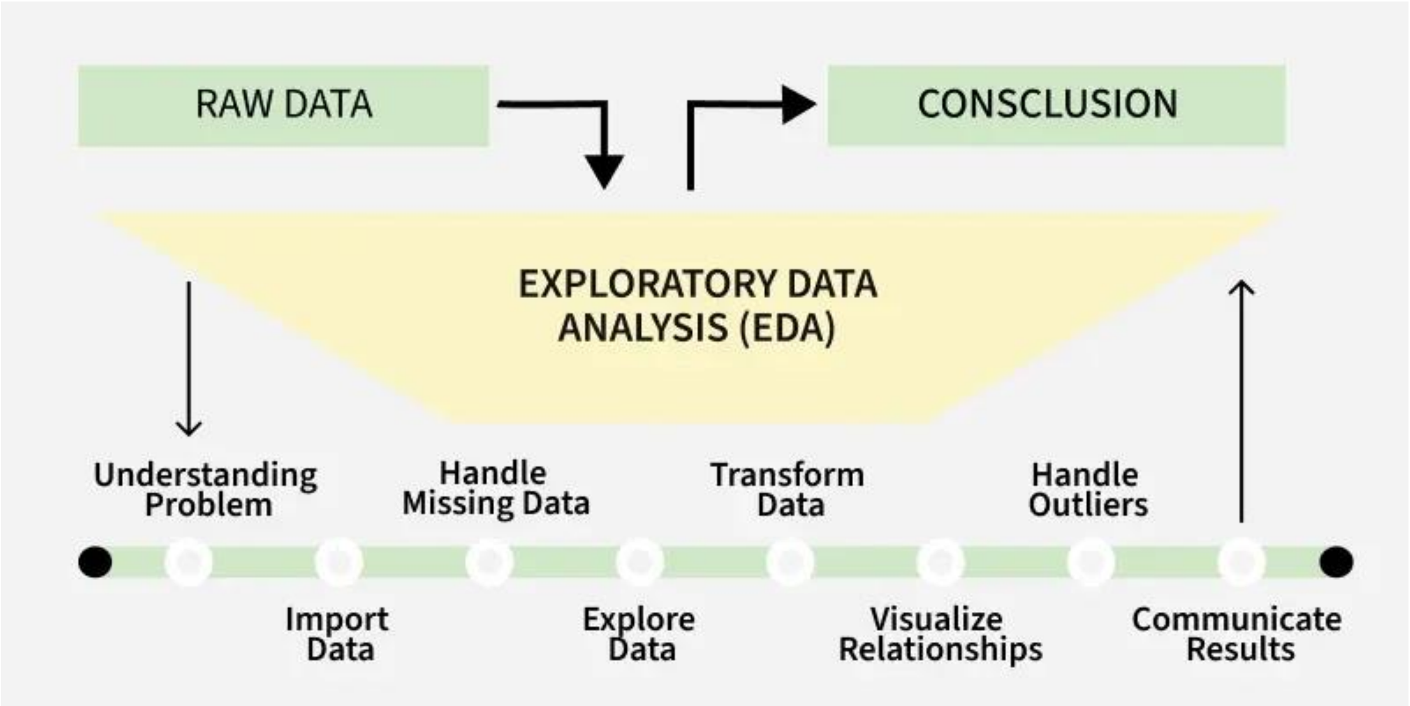
OR

1) Numerical / Quantitative
2) Categorical
3) Time-series

Exploratory Data Analysis (EDA)

•**What is it?** "Interviewing" your data before you use it

- It is the process of understanding and exploring your dataset before building any model
- It visualizes data to understand its **main features**, **find patterns** and discover how different parts of the data are **connected**.



Types of EDA

1) Univariate

- It focuses on analyzing one variable (single feature) at a time.
- It helps to understand the characteristics of that variable.
- It is used to describe the data and identify patterns within a single feature.
- Common summary statistics include:
  Mean, median, mode (to describe central tendency)
  Variance and standard deviation (to describe spread/variability)

2) Bivariate

- Focuses on identifying relationship between two variables to find connections, correlations and dependencies

3) Multivariate

- Identify relationships between two or more variables in the dataset and aims to understand how variables interact with one another

Why use EDA !

- Detect mistakes (wrong entries, missing rows)

  ➡ Data cleaning / Data quality assessment

- Check assumptions

  ➡ Diagnostic analysis

- Preliminary model selection

  ➡ Model screening

- Find relationships between explanatory variables

  ➡ Correlation analysis

- Assess direction/rough size of relationships

  ➡ Association strength analysis

Data cleaning / Data quality assessment

Detect mistakes

Let Consider this data

Spot the Flaws

ID	Age	Salary	City	Purchased?
1	25	$50k	NY	Yes
2	**?**	$1M	LA	No
3	25	$50k	NY	Yes
4	30	-$5k	SF	Yes

Missing values (?), Outliers ($1M), Duplicates (Row 1 & 3), Errors (-$5k),

Summary Statistics

It help us quickly understand the sample distribution of a variable

For a quantitative variable, the main characteristics that need to understand are:

1. Center (typical value)
2. Spread (how much variation exists)
3. Shape (distribution pattern)
4. Outliers (unusual/extreme values)

Remember

Your dataset is only a **sample**, so these statistics describe the **sample distribution**, and we use them to understand the possible **population distribution**

Summary Statistics

1. Center (typical value)

    mean, median, mode

2. Spread (how much variation exists)

    Variance, S.D, Interquartile Range (IQR)

3. Shape (distribution pattern)

    modality (peaks), skewness, kurtosis

4. Outliers (unusual/extreme values)

Summary Statistics

Shape (distribution pattern)

1) Modality

- Unimodal → one peak
- Bimodal → two peaks
- Multimodal → many peaks

For shape visualization:

- *histogram*
- *boxplot*
- *Q-Q plot (Quantile-Normal plot)*

2) Skewness (asymmetry)

- Positive skew (right-skew): long tail on high side
- Negative skew (left-skew): long tail on low side

3) Kurtosis

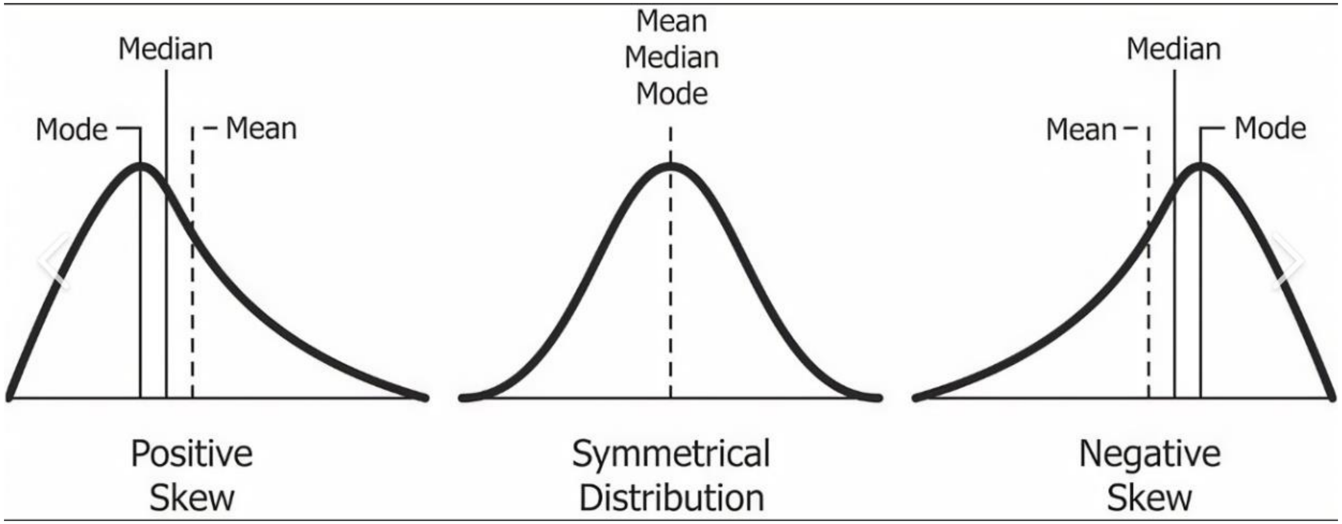
Kurtosis relates to **tails and peaked ness** compared to normal.

- High kurtosis (fat tails) → more extreme events
- Low kurtosis → fewer extremes, more uniform distribution

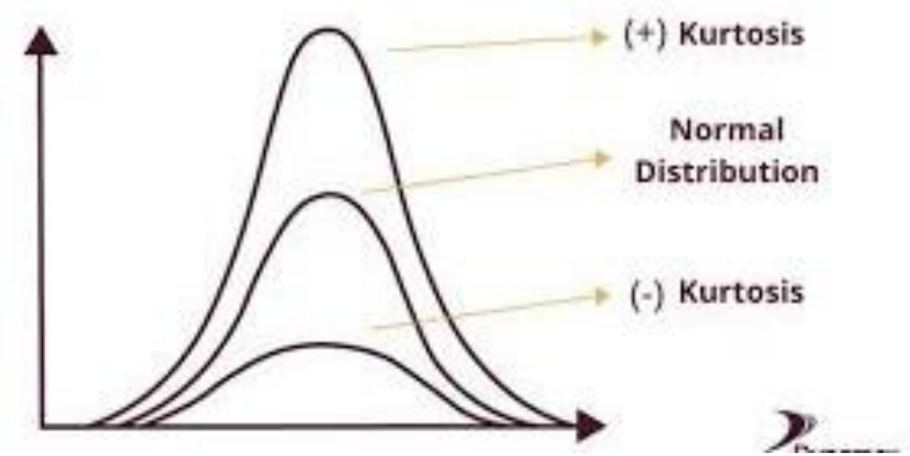
Summary Statistics

Shape (distribution pattern)

Skewness



Kurtosis



Summary Statistics

Outliers (unusual/extreme values)

Why outliers matter in AI/ML?

Outliers can:

- distort mean and standard deviation
- affect correlation
- confuse ML models (especially regression, kNN, SVM)
- create false alarms or wrong classification

Methods

1) IQR
2) Z-Score Method
3) Modified Z-score
4) Mahalanobis

Summary Statistics

Outliers (unusual/extreme values)

1) Interquartile Range (IQR)

- It is a measure of data spread (variation).
- It tells you how wide the middle 50% of your data is.
- Use for Outlier Detection

Formula:

$$IQR = Q3 - Q1$$

Where:

- Q1 (1st quartile) = 25% of data is below this value
- Q3 (3rd quartile) = 75% of data is below this value

the middle 50% of values lie within a range

Tukey Outlier Rule (Boxplot rule)

$$\text{Lower } bound = Q1 - 1.5 \times IQR$$

$$\text{Upper } bound = Q3 + 1.5 \times IQR$$

Where:

$$IQR = Q3 - Q1$$

Summary Statistics

Outliers (unusual/extreme values)

2) Z-Score Method

Mathematical Definition

$$z_i = \frac{x_i - \bar{x}}{s}$$

Where:

- $\bar{x}$ =mean
- s= standard deviation

Outlier Rule

Common thresholds:

- $|z| > 3$→ strong outlier
- $|z| > 2.5$→ possible outlier

3) Modified Z-score

Mathematical Definition

$$z_i^* = 0.6745 \cdot \frac{x_i - Median}{MAD}$$

Where:

- **MAD** = Median Absolute Deviation

$$MAD = median(|x_i - Median|)$$

Outlier Rule

$$|z_i^*| > 3.5 \Rightarrow outlier$$

Limitation

- Mean and std are **sensitive to outliers**
- Poor choice for skewed or heavy-tailed data

- Median-based → **robust**
- Works better for skewed engineering data

Summary Statistics

Outliers (unusual/extreme values)

S.No	Property	Meaning	Common Stats/Tools
1	Center	typical value	mean, median, mode
2	Spread	variability	std, variance, IQR
3	Shape	distribution pattern	histogram, skewness, kurtosis
4	Outliers	unusual values	IQR rule, z-score, boxplot