edquest

MLOps Training

MACHINE LEARNING OPERATIONS

edquest

Day 2 Agenda

Exploratory Data Analysis (EDA)

- Day 1 Recap
- Day 1 Workshop Quiz
- Understanding datasets: nulls, outliers, distributions
- Visualization tools (Pandas, Seaborn)
- Feature engineering and transformation
- Day 2 Quiz
- Day 2 Workshop

Day 1 Recap

OVERVIEW OF MACHINE LEARNING

edquest

Q1: Fill in the blank to correctly import the Linear Regression model in Python:

```
import pandas as pd
from sklearn.linear_model import ____________________
```

Choose the correct option

A: LinearRegression

B: linear_regression

C: regression

D: lin_reg

edquest

Q1: Fill in the blank to correctly import the Linear Regression model in Python:

```
import pandas as pd
from sklearn.linear_model import ____________________
```

Choose the correct option

A: LinearRegression

B: linear_regression

C: regression

D: lin_reg

edquest

Q2: Fill in the blank to correctly import the Logistic Regression model in Python:

```
import pandas as pd
from sklearn.linear_model import ____________________
```

Choose the correct option

A: Regression

B: logistic_regression

C: LogisticRegression

D: log_reg

edquest

Q2: Fill in the blank to correctly import the Logistic Regression model in Python:

```
import pandas as pd
from sklearn.linear_model import ____________________
```

Choose the correct option

A: Regression

B: logistic_regression

C: LogisticRegression

D: log_reg

edquest

Q3:Which module should you import train_test_split from in Python?

```
import pandas as pd
from _____________________________ import train_test_split
```

Choose the correct option

A. sklearn.preprocessing

B. sklearn.model_selection

C. sklearn.metrics

D. sklearn.linear_model

edquest

Q3:Which module should you import train_test_split from in Python?

```
import pandas as pd
from _____________________________ import train_test_split
```

Choose the correct option

A. sklearn.preprocessing

B. sklearn.model_selection

C. sklearn.metrics

D. sklearn.linear_model

edquest

Q4: Which of the following is the correct function used to load a .csv file into a pandas DataFrame in Python?

```
import pandas as pd
df = pd._____("ice_cream.csv")
```

Choose the correct option

A. pandas.load_csv('filename.csv')

B. pandas.open_csv('filename.csv')

C. pandas.import_csv('filename.csv')

D. pd.read_csv('filename.csv')

edquest

Q4: Which of the following is the correct function used to load a .csv file into a pandas DataFrame in Python?

```
import pandas as pd
df = pd._____("ice_cream.csv")
```

Choose the correct option

A. pandas.load_csv('filename.csv')

B. pandas.open_csv('filename.csv')

C. pandas.import_csv('filename.csv')

D. pd.read_csv('filename.csv')

edquest

Q5: What does the following line do in Python?

```
joblib.dump(model_lin_reg, 'linear_regression_model.pkl')
```

Choose the correct option

A. It saves the model to a file

B. It loads a model from a file

C. It evaluates the model and prints results

D. It deletes the model from memory

edquest

Q5: What does the following line do in Python?

```
joblib.dump(model_lin_reg, 'linear_regression_model.pkl')
```

Choose the correct option

A. It saves the model to a file

B. It loads a model from a file

C. It evaluates the model and prints results

D. It deletes the model from memory

edquest

Q6: What does test_size=0.2 indicate in the train_test_split() function?

```
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.2,random_state=1)
```

Choose the correct option

A. The data will not be split

B. 20% of the data will be used for testing

C. 80% of the data will be used for testing

D. 20% of the data will be used for training

edquest

Q6: What does test_size=0.2 indicate in the train_test_split() function?

```
X_train,X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.2,random_state=1)
```

Choose the correct option

A. The data will not be split

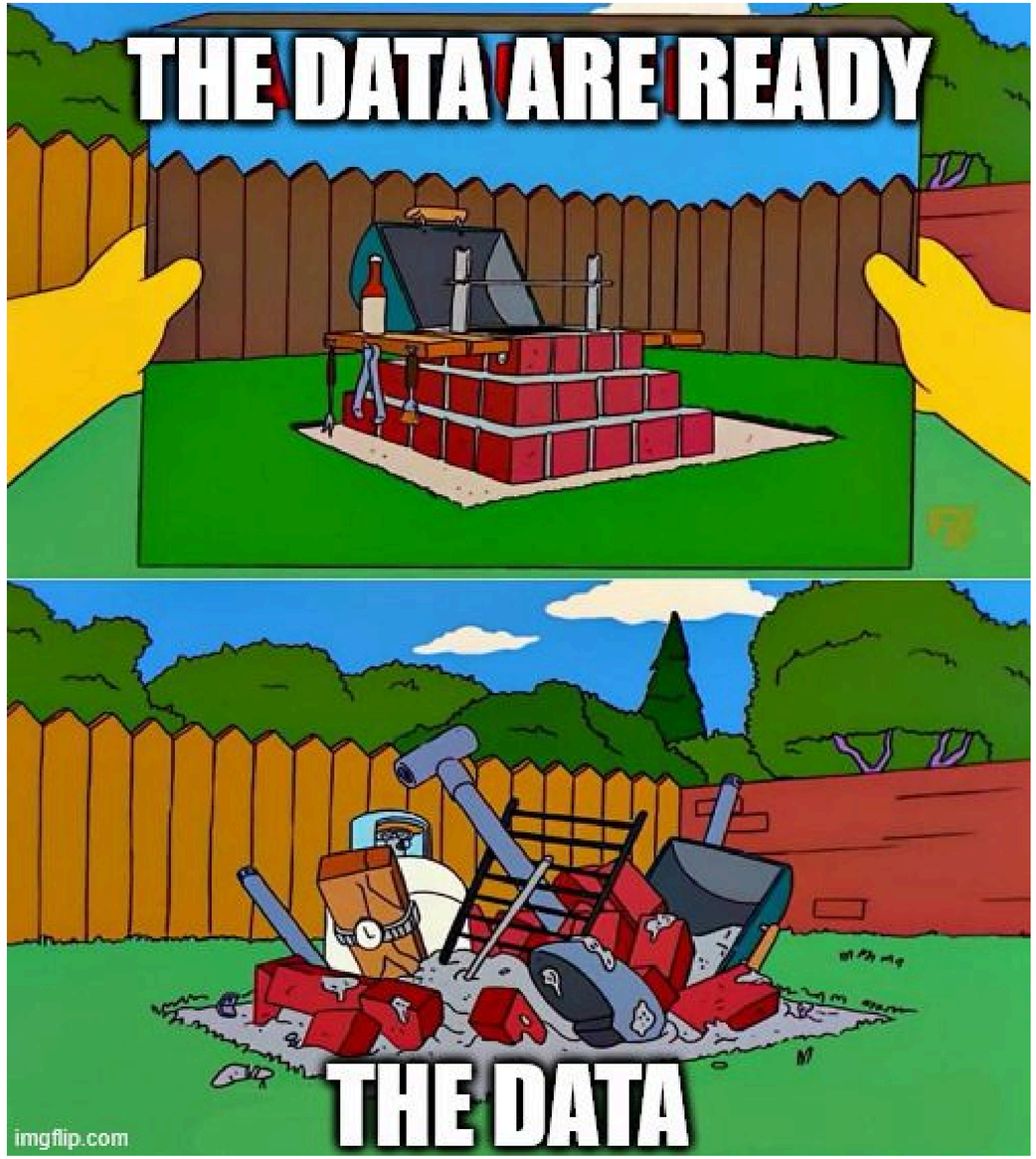
B. 20% of the data will be used for testing

C. 80% of the data will be used for testing

D. 20% of the data will be used for training

What Is EDA?

What Is EDA?

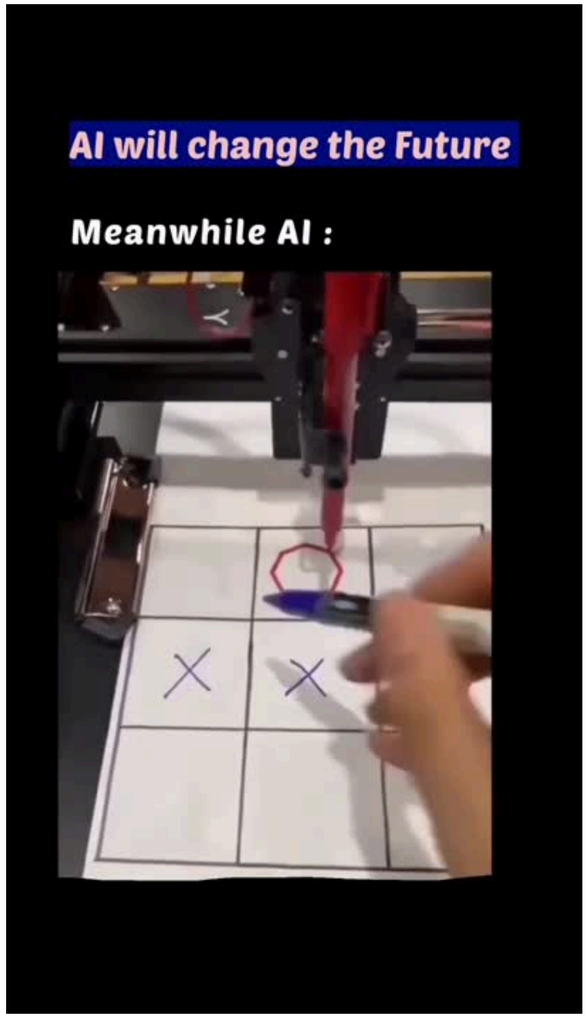


edquest

What Is EDA?


Data Scientist:
Our model is at 99% accuracy
Data Engineer:
The data has 17 null columns, 5 duplicate entries, and timestamps in 3 formats.

Before EDA



After EDA

edquest

What Is EDA?

EDA helps us explore and understand the structure, patterns, and issues in data before modeling.

- It’s the diagnostic phase of data science.
- Helps detect missing values, outliers, and shape of distributions.
- Guides feature engineering and cleaning decisions

Can we split EDA?

- **EDA Level 0 – Initial Exploration of Raw Data**
  - *Gaining a basic understanding of the unprocessed dataset.*
- **EDA Level 1 – Data Cleaning and Transformation**
  - *Applying preprocessing techniques to make the data usable.*
- **EDA Level 2 – Analysis of Processed Data**
  - *Exploring patterns, trends, and relationships in the cleaned dataset.*

Why is EDA Important?

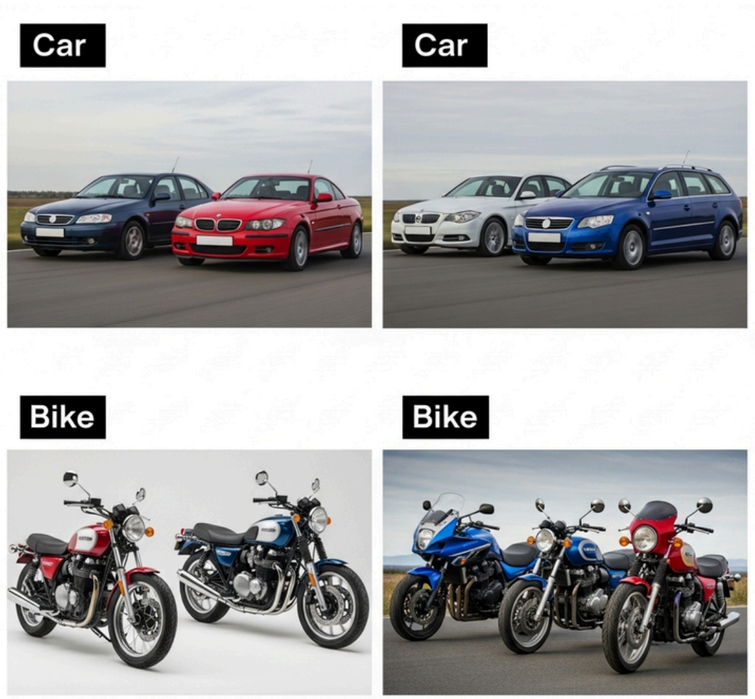
Before you build models, you must understand the data. A model is only as good as the data you feed it. If your dataset contains inconsistencies, missing values, or outliers, your model may underperform or give incorrect results.

EDA enables you to:

- Detect data quality issues (nulls, duplicates, wrong types)
- Understand feature distributions and relationships
- Uncover hidden trends and patterns
- Guide feature selection and transformation

Labeled Data

Labeled data consists of data points, such as images, text, or sensor readings, that have been explicitly tagged with meaningful labels or categories.



Unlabeled Data

Unlabeled data is raw data that doesn't have any pre-defined labels or categories. It's essentially the raw, unprocessed data.

Structured Data

Data that conforms to a well-defined, pre-defined format, typically organized in rows and columns (like in a spreadsheet or database).

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	Shed
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN
7	8	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	Shed
8	9	50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN
9	10	190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN

10 rows × 81 columns

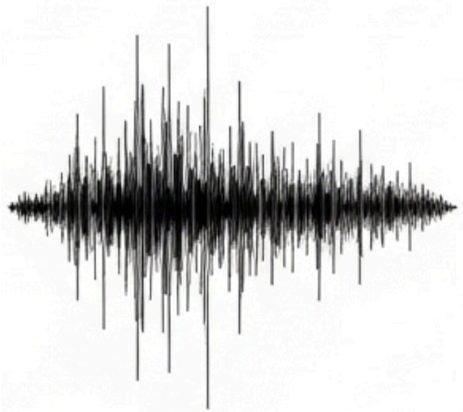
Unstructured Data

Data that does not conform to a predefined format and lacks a clear structure.

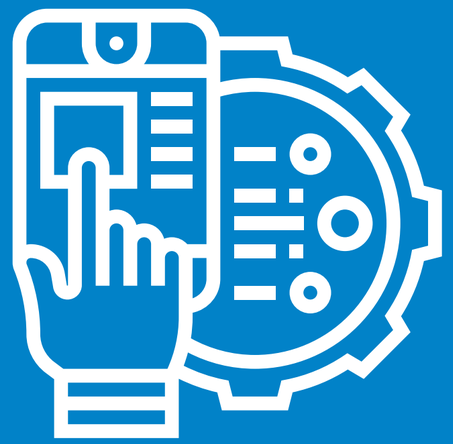
Example :

.txt

.jpg



.mp3



edquest

Understanding Datasets

edquest

What is the Dataset?

Definition

- A Dataset is a set of data grouped into a collection with which engineers can work to meet their goals.

- In a structured dataset, the rows represent the number of data points and the columns represent the features of the Dataset.

Why are datasets used?

Datasets are used to train and test AI models, analyze trends, and gain insights from data. They provide the raw material for computers to learn patterns and make predictions.

Dataset Types

- Numerical Dataset
- Categorical Dataset
- Web Dataset
- Time series Dataset
- Image Dataset

edquest

Dataset Properties

- Center of data
- Skewness of data
- Correlation among the data
- Presence of outliers

edquest

Dataset Features

The features of a dataset are the most critical aspect of the dataset, as based on the features of each available data point, will there be any possibility of deploying models to find the output to predict the features of any new data point that may be added to the dataset.

Numerical Features:

Categorical Features:

Metadata:

Size of the Data:

Examples

edquest

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	I
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	
5	6	50	RL	85.0	14115	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	
6	7	20	RL	75.0	10084	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	
7	8	60	RL	NaN	10382	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	
8	9	50	RM	51.0	6120	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	
9	10	190	RL	50.0	7420	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	

10 rows × 81 columns

Dataset Understanding

This includes checking the number of rows and columns, data types, and summary statistics.

- Viewing top rows using .head()
- Understanding structure using .info() and .describe()
- Checking data types (numerical, categorical, datetime)
- Identifying duplicates or anomalies

Handling Missing Values

Detect → Drop or Impute → Validate

- Use isnull().sum()
- Imputation: Mean, Median, Mode, Domain logic
- Sometimes missing ≠ bad — depends on context

Handling Missing Values

Detect → Drop or Impute → Validate

- Use isnull().sum()
- Imputation: Mean, Median, Mode, Domain logic
- Sometimes missing ≠ bad — depends on context

Dealing with Outliers

Outliers are extreme values that differ significantly from other observations. They can skew analysis and impact model performance.

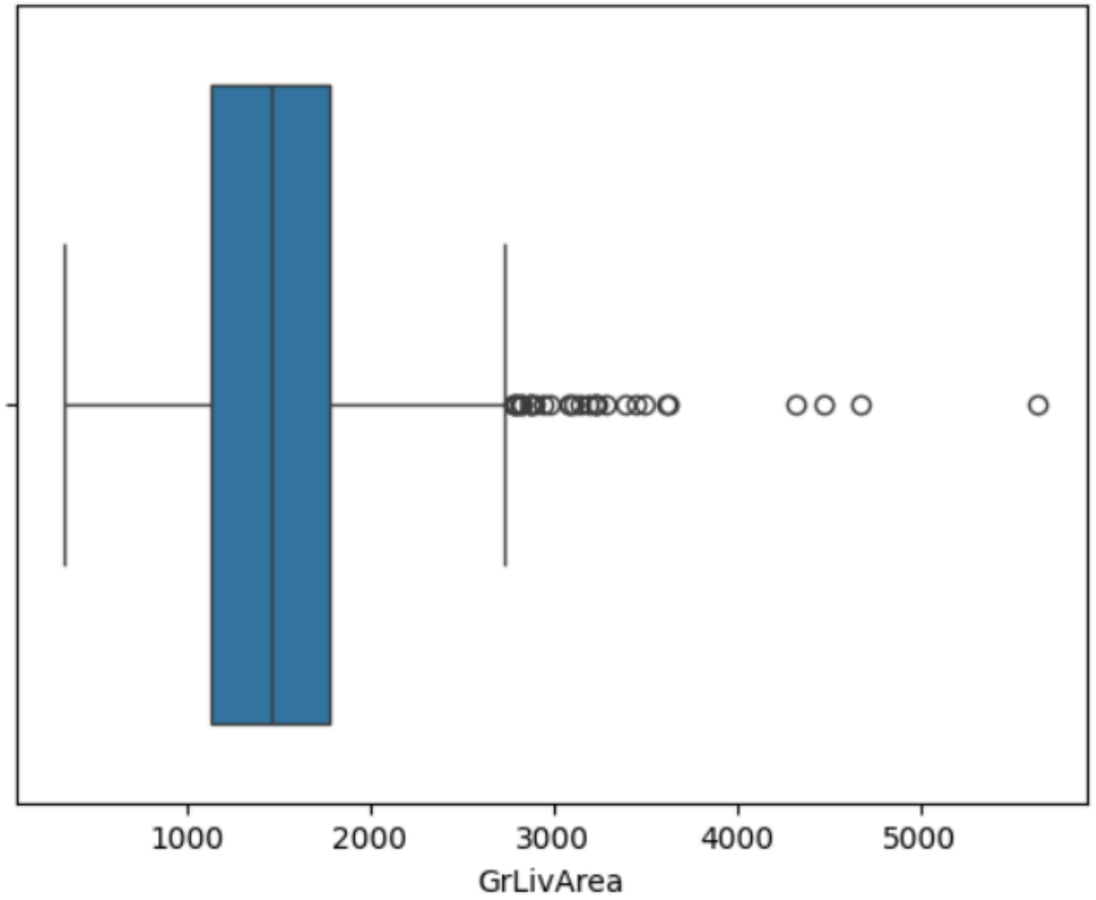
Ways to detect them:

- Using boxplots, histograms

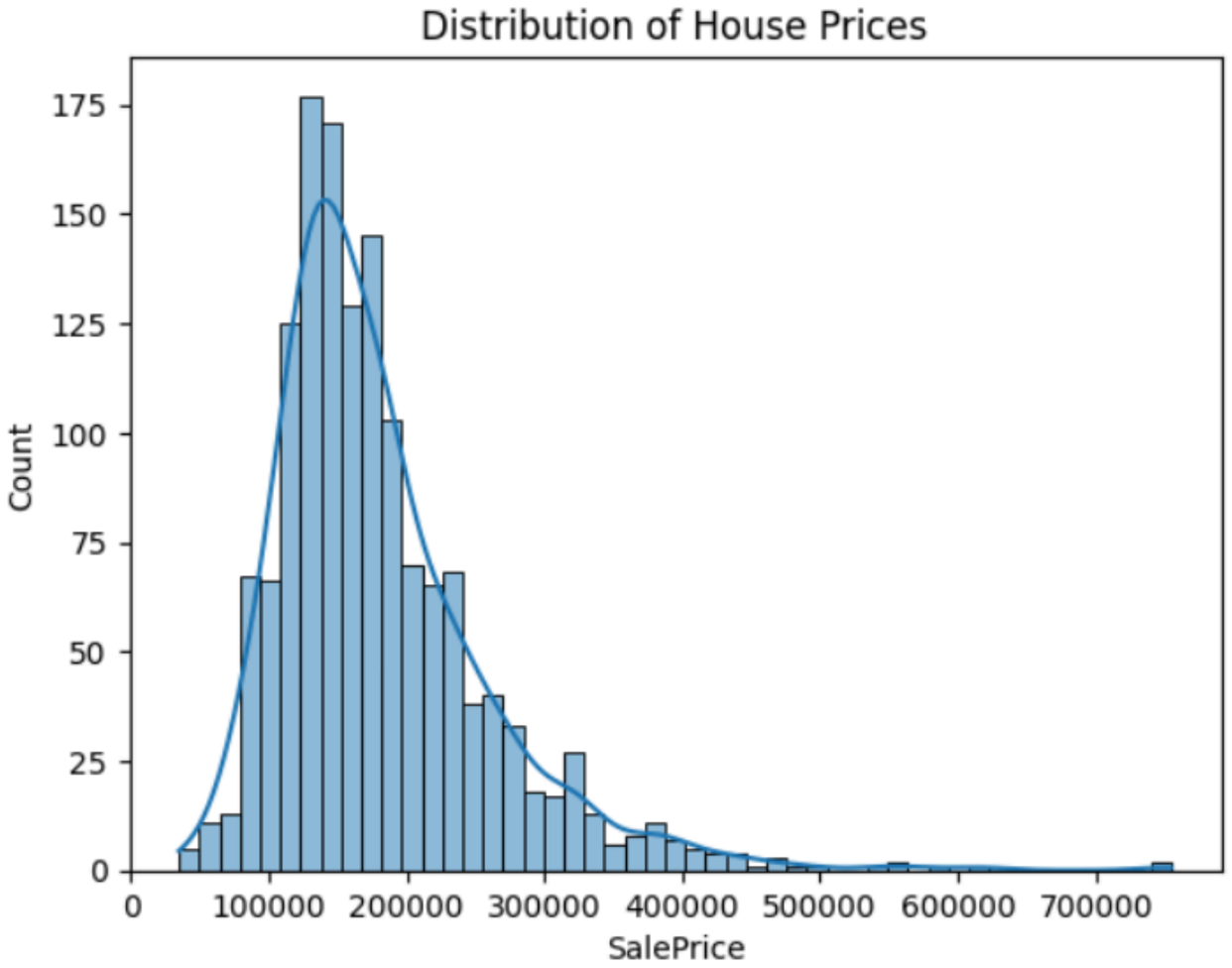
Once detected:

- You can drop them if they’re errors
- Cap them if they are rare but legitimate
- Or transform the variable (e.g., apply log)

Boxplot Outliers



Hisplot Outliers



Understanding Distributions

A distribution shows how data is spread. Know your shape: Normal, Skewed, Peaked.

Fix with log/sqrt transforms if needed.

Why It Matters

- Affects model assumptions
- May require transformation

Fixes

- Use log() or sqrt() on skewed features
- Tools: histplot, KDE, df.skew(), df.kurtosis()

Understanding Distributions

edquest

A distribution shows how data is spread. Knowing the shape helps in choosing transformations and ML algorithms.

Common Types:

- **Normal:** Bell-shaped; good for linear models
- **Skewed:** Left or right-tailed; may require log/sqrt transformation
- **Bimodal:** Two peaks; may indicate subgroups

edquest

Visualization Tools

Visualization Tools

Visualization is a core part of Exploratory Data Analysis. It helps you understand patterns, spot anomalies, and uncover relationships between features that are hard to detect from raw numbers alone.

Python offers two widely used libraries for data visualization in EDA:

- Pandas
- Seaborn