

# Interview Questions:

## 1. What is the purpose of EDA?

**Exploratory Data Analysis (EDA)** is the **first and most important step** in any data science or machine learning project. It helps you understand the structure, quality, and patterns of your dataset **before building models**.

---

### Main Purposes of EDA:

#### 1. Understand the Dataset Structure

- See how many rows and columns are there
- Check types of data (numerical, categorical, etc.)
- Identify the target variable and features

*Example:* In the Titanic dataset, you learn that features include Age, Sex, Pclass, and target is Survived.

---

#### 2. Detect Missing or Incorrect Data

- Find columns with missing values
- Identify incorrect formats or wrong values

*Example:* Many passengers have missing Age values. The Cabin column has many nulls.

---

#### 3. Discover Patterns and Relationships

- Find which variables impact the outcome
- Analyze correlation between features

*Example:* Females and first-class passengers had higher survival rates.

---

#### 4. Detect Outliers and Anomalies

- Use boxplots or z-scores to spot extreme values

*Example:* Some passengers paid extremely high fares — possibly outliers.

---

#### 5. Understand Feature Distributions

- Use histograms or KDE plots to understand the shape (normal, skewed, etc.)

*Example:* The Fare column is right-skewed — most people paid low fares.

---

## 6. Prepare for Feature Engineering

- Based on insights, you can:
  - Drop useless columns
  - Create new features
  - Decide on transformations

*Example:* You may create a new feature:  $\text{FamilySize} = \text{SibSp} + \text{Parch} + 1$ .

---

### Why EDA is Important:

 Jumping into modeling without EDA is like trying to build a house without knowing the land.

---

## 2. How do boxplots help in understanding a dataset?

A **boxplot** (also called a box-and-whisker plot) is a powerful **visual tool** in EDA that helps you understand the **distribution**, **central tendency**, and **variability** of numerical data, as well as detect **outliers**.

---

### What a Boxplot Shows:

|-----|=====|-----|  
min Q1 median Q3 max

It shows **five summary statistics**:

1. **Minimum** – smallest value (excluding outliers)
2. **Q1 (First Quartile)** – 25% of the data is below this point
3. **Median (Q2)** – middle value of the data
4. **Q3 (Third Quartile)** – 75% of the data is below this point
5. **Maximum** – largest value (excluding outliers)

✓ **Box** = Interquartile Range ( $\text{IQR} = \text{Q3} - \text{Q1}$ )

✓ **Whiskers** = Extend from Q1 to min and Q3 to max

✓ **Dots** = Outliers (values beyond  $1.5 \times \text{IQR}$ )

---

## What You Can Learn from a Boxplot:

Insight	How Boxplot Helps
Central Tendency	The <b>line inside the box</b> shows the <b>median</b> , which is more robust than the mean for skewed data.
Spread of Data	The <b>width of the box</b> (IQR) shows how spread out the middle 50% of the data is.
Skewness	If the <b>median is off-center</b> , or one whisker is longer than the other, the data is <b>skewed</b> .
Outliers	Easily identify outliers as <b>dots</b> beyond the whiskers.
Compare Distributions	Side-by-side boxplots help <b>compare features</b> or <b>categories</b> (e.g., male vs female age).

## Example:

Suppose you have a boxplot of Age in the Titanic dataset:

- Median age: 29
- IQR: 22 to 38
- Outliers: Some passengers aged above 65
- Right-skewed distribution (older people are fewer)

---

## Use Case:

**Compare survival based on fare:**

```
sns.boxplot(x='Survived', y='Fare', data=df)
```

This will show if people who paid more had a higher chance of survival.

---

### 3.What is correlation and why is it useful?

#### ✓ 3. What is Correlation and Why Is It Useful?

##### What is Correlation?

**Correlation** is a **statistical measure** that shows the **strength and direction of the relationship between two variables**.

- It answers: "**When one variable changes, does the other change too?**"

##### Types of Correlation:

Type	Description	Example
Positive	Both variables increase or decrease together	Hours studied ↔ Marks
Negative	One increases, the other decreases	Age ↔ Time on Instagram
Zero (None)	No relationship between variables	Shoe size ↔ IQ

##### Correlation Coefficient (r)

Usually measured using **Pearson's correlation coefficient (r)**:


r value	Interpretation
+1	Perfect positive correlation
0	No correlation
-1	Perfect negative correlation




$$r = \text{Cov}(X, Y) / (\sigma_X * \sigma_Y)$$

Where:

- $\text{Cov}(X, Y)$ : Covariance of X and Y
- $\sigma_X, \sigma_Y$ : Standard deviations of X and Y

##### Why is Correlation Useful in ML and EDA?

Benefit	Use
 <b>Feature Selection</b>	Helps identify which features are related (so you can drop redundant ones)

 <b>Multicollinearity Check</b>	High correlation between features can confuse models like linear regression
 <b>Data Understanding</b>	See how variables influence each other (e.g., Fare vs. Survived)
 <b>Model Insight</b>	Understand which variables matter most for predictions

### Example from Titanic Dataset

```
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
```

You might see:

- Fare and Pclass = negative correlation → lower class, lower fare
- SibSp and Parch = slightly positive → families travel together

### Important Notes:

- **Correlation ≠ Causation:** Just because two variables are correlated doesn't mean one causes the other.
- Use **scatter plots** to visually verify correlations.

## 4.How do you detect skewness in data?

### What is Skewness?

**Skewness** is a measure of the **asymmetry** in the distribution of data.

- If data is **perfectly symmetrical**, it has **zero skewness**.
- If data leans **left or right**, it's **skewed**.

### Types of Skewness:

Type	Description	Example
<b>Symmetric</b>	Mean $\approx$ Median $\approx$ Mode	Normal distribution (bell curve)
<b>Right Skewed</b>	Tail on right; Mean $>$ Median	Income, house prices
<b>Left Skewed</b>	Tail on left; Mean $<$ Median	Age of retirement, exam failures

## How to Detect Skewness?

### 1 Mathematically using .skew()

- import pandas as pd
  - df['Age'].skew()
  - **Skewness > 0**: Right-skewed
  - **Skewness < 0**: Left-skewed
  - **Skewness ≈ 0**: Symmetric
- 

### 2 Using Histogram

- import seaborn as sns
  - import matplotlib.pyplot as plt
  - sns.histplot(df['Age'], kde=True)
  - plt.title("Histogram of Age")
  - plt.show()
  - **Right tail → right-skewed**
  - **Left tail → left-skewed**
- 

### 3 Using Boxplot

- sns.boxplot(x=df['Fare'])
  - If the **median line is not centered** and whiskers are uneven → skewed
- 

## Real Example

Feature	Skewness Value	Interpretation
Fare	4.7	Highly right-skewed
Age	0.4	Slightly right-skewed
Pclass	-0.6	Left-skewed

## Why Skewness Matters in ML?

Problem	Solution
Skewed features affect model	Apply log, sqrt, or Box-Cox transform
Outliers cause skew	Consider capping or transformation
Linear models assume normal	Normalize or transform the feature

## 5.What is multico linearity?

### 5. What is Multicollinearity?

#### Definition:

**Multicollinearity** occurs when **two or more independent (predictor) variables** in a regression model are **highly correlated**, meaning one can be linearly predicted from the others with a high degree of accuracy.

#### Why is it a problem?

When multicollinearity exists:

- The model **cannot distinguish** which variable is actually influencing the output.
- It causes **unstable and unreliable coefficients**.
- It **inflates standard errors**, leading to **insignificant p-values** even for important predictors.

#### Example:

Imagine a model predicting house price using:

- Size in square feet
- Number of rooms

These two are likely correlated. Including both may introduce multicollinearity.

#### Effects of Multicollinearity:

Consequence	Explanation
Coefficients become unstable	Change drastically with small data changes

Consequence	Explanation
Hard to interpret feature effects	Coefficients lose meaning
Low model significance	High p-values for predictors
Reduces model generalization	Overfitting risk

## How to Detect Multicollinearity:

### 1 Correlation Matrix

```
import seaborn as sns
```

```
import matplotlib.pyplot as plt
```

```
corr = df.corr()
```

```
sns.heatmap(corr, annot=True, cmap='coolwarm')
```

- Check for values **> 0.8** or **< -0.8** → strong correlation

### 2 Variance Inflation Factor (VIF)

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

```
from statsmodels.tools.tools import add_constant
```

```
X = add_constant(df[['feature1', 'feature2', 'feature3']])
```

```
vif = pd.DataFrame()
```

```
vif['features'] = X.columns
```

```
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
```

- **VIF > 5 or 10** → strong multicollinearity

## How to Handle It?

Solution	Description
Remove one of the correlated variables	Drop redundant predictors
Combine variables	Use PCA or create new feature



Solution	Description
Regularization	Use Ridge or Lasso regression to penalize
Collect more data	Sometimes helps reduce correlation impact

### ✅ Summary:

Aspect	Details
What is it?	High correlation between predictors
Why it's bad?	Unstable, unreliable coefficients
Detection tools	Correlation matrix, VIF
Fixes	Drop, combine features, use Ridge/Lasso

Let me know if you want code to calculate VIF or visualize multicollinearity.

## 6.What tools do you use for EDA?

Exploratory Data Analysis is the **first and most crucial step** in any Machine Learning or Data Science project. A variety of tools help you understand data structure, spot anomalies, and extract insights.

### 🧰 Top Tools Used for EDA:

#### 🐼 1. Pandas (Python Library)

- **Use:** Data loading, cleaning, transformation, basic statistics.
- **Functions:**
  - `.head()`, `.info()`, `.describe()`
  - `.isnull()`, `.value_counts()`, `.groupby()`

#### 📊 2. Matplotlib (Python Library)

- **Use:** Creating basic visualizations.
- **Plots:**
  - Line, bar, histogram, pie, scatter, etc.

- **Command example:**

➤ `plt.hist(df['age'])`

### 3. Seaborn (Built on Matplotlib)

- **Use:** Statistical visualizations with fewer lines of code.
- **Visuals:**
  - boxplot, heatmap, pairplot, countplot, violinplot
  - Easier correlation plots and styling
  - Example:
    - `sns.heatmap(df.corr(), annot=True)`

### 4. Plotly (Interactive Charts)

- **Use:** Interactive dashboards and visuals
- **Advantage:** Great for presentations or web apps
- **Charts:** Line, scatter, histogram, 3D, etc.

### 5. NumPy

- **Use:** Basic numerical operations
- **Functions:**
  - `np.mean()`, `np.median()`, `np.std()`, etc.

### 6. Statsmodels

- **Use:** Statistical testing and diagnostics
- **Example:**
  - `import statsmodels.api as sm`
  - `sm.qqplot(df['age'], line='s')`

### 7. Sweetviz, pandas-profiling, Autoviz (Automated EDA tools)

- **Use:** Auto-generate reports with data profiling
- **Quick Insight:**
  - `pip install sweetviz`

## 7.Can you explain a time when EDA helped you find a problem?

### Example Scenario: Titanic Dataset — Predicting Survival

#### Situation:

While working on the **Titanic dataset** to build a machine learning model that predicts whether a passenger survived or not, the first step was performing **Exploratory Data Analysis (EDA)**.

#### Key EDA Findings (Problems Detected):

##### **1** Missing Values:

- Columns like Age, Cabin, and Embarked had missing data.
- Using `df.isnull().sum()` revealed:
  - Age → 177 missing
  - Cabin → 687 missing
  - Embarked → 2 missing

#### Solution:

- Imputed Age with median, dropped Cabin, filled Embarked with mode.

##### **2** Outliers in Fare:

- Using `sns.boxplot(df['Fare'])`, noticed very high fares that skewed the data.
- Example: One passenger paid **over 500 units**, while most paid under 100.

#### Solution:

- Applied log transformation: `df['Fare_log'] = np.log1p(df['Fare'])`

##### **3** Data Skewness:

- Skewness in Fare and Age detected using:
  - `df['Fare'].skew(), df['Age'].skew()`

#### Solution:

- Used normalization or binning for better model performance.

##### **4** Strong Correlation (Redundancy):

- High correlation between Pclass and Fare via `sns.heatmap()`.

#### Solution:

- Avoided feeding highly correlated features directly into the model.

## 8.What is the role of visualization in ML?

Visualization plays a **critical role** throughout the entire machine learning pipeline — from understanding raw data to communicating model results.

### Why is Visualization Important in ML?

Phase	Role of Visualization
1. Data Understanding (EDA)	Helps uncover patterns, outliers, distributions, and relationships between features.
2. Feature Selection	Correlation heatmaps, pairplots, and bar plots help identify redundant or irrelevant features.
3. Data Preprocessing	Visuals like boxplots or histograms reveal skewness, missing values, or data imbalance.
4. Model Evaluation	Plots such as confusion matrix, ROC curves, precision-recall curves allow deeper understanding of model performance.
5. Communication	Makes it easier to explain results and insights to stakeholders or non-technical audiences.

### Common Visualization Types & Their Use in ML

Plot Type	Purpose
Histogram	Understand data distribution and detect skewness.
Boxplot	Detect outliers and compare distributions.
Pairplot	Examine relationships between multiple variables.
Heatmap (Correlation)	Detect multicollinearity between features.
Confusion Matrix	Evaluate classification performance.
ROC/AUC Curve	Measure how well the model distinguishes classes.
Feature Importance Plot	Show which features contributed most to model decisions.

### Example Use Case:

While working on a classification task using the Titanic dataset, a **boxplot of 'Fare' vs 'Survived'** revealed that passengers who paid higher fares were more likely to survive. This insight led to prioritizing the Fare feature for the model.