**Interview Questions & Answers**

**1. What are the different types of missing data?**

**Types of Missing Data:**

* **MCAR (Missing Completely At Random):**  
  The missingness has no relationship with any feature.  
  *Example: A sensor randomly fails to record temperature.*
* **MAR (Missing At Random):**  
  Missingness is related to **observed** variables.  
  *Example: Older participants are more likely to skip tech-related questions.*
* **MNAR (Missing Not At Random):**  
  Missingness is related to the **unobserved** value itself.  
  *Example: People with very high income may skip the salary question.*

+---------+---------+---------+-------------+

| Age | Gender | Salary | Type |

+---------+---------+---------+-------------+

| 25 | Male | 50,000 | Normal |

| 40 | Female | | MAR |

| 31 | Female | 80,000 | Normal |

| 30 | Male | | MCAR or MNAR|

+---------+---------+---------+-------------+

**2. How do you handle categorical variables?**

You typically encode them into numbers so ML models can understand them.

**Techniques:**

* **Label Encoding** – assigns integers to categories.  
  *Example:* Red = 0, Green = 1, Blue = 2
* **One-Hot Encoding** – creates binary columns.  
  *Example:*

Color → Red | Green | Blue

Red 1 0 0

Green 0 1 0

Blue 0 0 1

**3. What is the difference between normalization and standardization?**

| **Feature** | **Normalization** | **Standardization** |
| --- | --- | --- |
| Scale | [0, 1] | Mean = 0, Std Dev = 1 |
| Formula | (x - min) / (max - min) | (x - mean) / std |
| Use when | Data is not normally distributed | Data is normally distributed |
| Sensitive to outliers | Yes | Less sensitive |

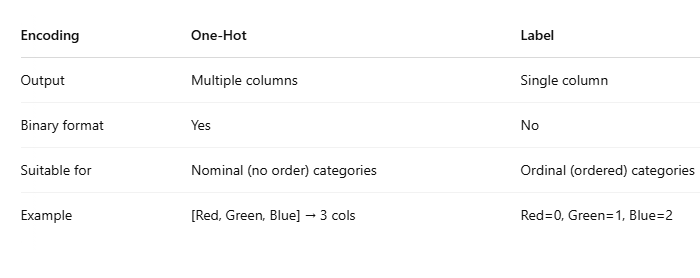
**4. How do you detect outliers?**

**Methods:**

* **Boxplot/IQR Method:**
  + Q1 = 25th percentile, Q3 = 75th percentile
  + IQR = Q3 - Q1
  + Outliers: < Q1 - 1.5×IQR or > Q3 + 1.5×IQR

**5. Why is preprocessing important in ML?**

* Ensures **consistency** in data.
* Removes noise, handles missing data.
* Improves **accuracy and speed** of ML models.
* Helps in **faster convergence** and better generalization.
* **6. What is one-hot encoding vs label encoding?**



**7. How do you handle data imbalance?**

* **Resampling Techniques:**
* **Oversampling** (e.g., SMOTE)
  + - **Undersampling**
* **Use proper metrics** like F1-score instead of accuracy
* **Use class weights** in ML algorithms

**8. Can preprocessing affect model accuracy?**

Yes! Proper preprocessing like handling nulls, encoding, and feature scaling can:

* Improve **model performance**
* Reduce **overfitting**
* Ensure **fair comparison** between features
* Help with **convergence in neural networks**