1️⃣ How do you handle missing values in Pandas?

Answer:

In Pandas, I first identify missing values using isnull() or isnull().sum().
Based on the nature of the data, I either:

- fill missing values using fillna() (mean or median for numerical data, mode for categorical data), or
- remove rows or columns using dropna() if the missing data is insignificant or irrelevant.

2️⃣ Difference between fillna() and dropna()?

Answer:

fillna() is used to **replace missing values** with a specific value like mean, median, or mode, without losing data.
dropna() is used to **remove rows or columns** that contain missing values.

I prefer fillna() when data loss is not acceptable and dropna() when missing values are very few or not useful.

3️⃣ How do you detect duplicates in Pandas?

Answer:

Duplicates are detected using the duplicated() function.
To find how many duplicate rows exist, I use duplicated().sum().
If duplicates are present, I remove them using drop_duplicates() to ensure data accuracy.

4️⃣ Why is datatype conversion important in analytics?

Answer:

Datatype conversion is important because it ensures correct analysis and improves memory efficiency.
For example, categorical data stored as objects should be converted to category type for better performance.
Incorrect data types can lead to wrong calculations and inefficient data processing.

5️⃣ What are the benefits of Python over Excel in data cleaning?

Answer:

Python can handle **large datasets efficiently**, while Excel has size limitations.
Data cleaning in Python is **automated and reproducible**, making it suitable for repeated tasks.
Python also integrates well with data analysis and machine learning libraries, which is not possible in Excel.