**IMPORTANT QUESTION BASED ON INTERVIEW**

**1. Regular Expressions (Regex)**

1. How would you extract email IDs and phone numbers from a messy text file using regex?
2. Can you explain greedy vs non-greedy matching in regex with an example?
3. Suppose you have a log file with timestamps. How will you validate whether the timestamps are in YYYY-MM-DD HH:MM:SS format using regex?

**2. NumPy Package**

1. What is the difference between a NumPy array and a Python list in terms of memory and performance?
2. How would you vectorize a Python loop in NumPy for performance optimization? (e.g., element-wise multiplication without a for loop).
3. Explain broadcasting in NumPy with an example where arrays of different shapes can still be added together.
4. Suppose you have a very large dataset that doesn’t fit into memory. How would you handle it using NumPy?

**3. Pandas Package**

1. How would you handle missing values in a Pandas DataFrame? (various strategies).
2. Difference between apply(), map(), and applymap() in Pandas.
3. If you have a DataFrame with millions of rows, how would you improve performance when filtering data?
4. What’s the difference between merge(), join(), and concat() in Pandas?
5. How do you handle categorical data in Pandas efficiently?

**4. Loading Datasets**

1. If you have a **10GB CSV file** but only 16GB RAM, how would you load and process it in Pandas?
2. How do you load data from SQL into Pandas?
3. What is the difference between read\_csv() and read\_table()?
4. How would you handle loading data from an API into a DataFrame?

**5. Accessing Data from Data Frame**

1. What’s the difference between df.loc[], df.iloc[], and df.at[]?
2. How would you select multiple conditions in Pandas (e.g., employees with Age > 30 and Salary < 50,000)?
3. How can you efficiently retrieve only a subset of columns from a large DataFrame?
4. How would you extract the top 5 highest-paid employees from a dataset?