1) Data Cleaning:

• What are the common issues you might encounter in a messy dataset?

1) Missing Data.
2) Duplicate.
3) Inconsistent formatting
4) Outliers
5) Incorrect data
6) Data type mismatches

• How would you handle missing values in a dataset?

1) Fill in missing values with statistical methods like the mean, median, or mode for numerical data or the most frequent category for categorical data.
2) If the missing values are limited and don't represent a significant portion of the dataset, removing rows or columns might be acceptable.
3) Create a new feature that flags whether a value was missing or not.

• What is the importance of data type consistency in data analysis?

1) **Accuracy**: When your data types are consistent, you're less likely to make errors in your analysis. For example, if a column is meant to hold numbers but accidentally contains text, your calculations will be off.
2) **Efficiency**: Consistent data types make your analysis faster and smoother. The software can optimize calculations when it knows exactly what type of data it's working with.
3) **Avoiding errors**: Inconsistent data types can cause all kinds of issues. For example, you might try to do a sum on a column that has mixed text and numbers, and it'll throw an error.
4) **Compatibility**: If you're merging or joining datasets, having the right data types ensures that everything fits together as expected. You won't run into problems where two datasets can't be joined because one column has the wrong type of data.

2) SQL Queries:

1) What is the difference between INNER JOIN and LEFT JOIN?

INNER JOIN: returns only the rows where there is a match between the two tables based on the specified join condition. It filters out the rows that don't have a match in both tables.

LEFT JOIN: Return all rows from the left table, even if there are no matches in the right table

2) How would you use the GROUP BY clause to aggregate data?

To use the GROUP BY clause to aggregate data, you first specify the column(s) by which you want to group the rows. Then, you can apply aggregate functions like COUNT (), SUM (), AVG(), MIN(), or MAX() on other columns to calculate summary statistics for each group.

3) What is the purpose of the HAVING clause in SQL?

- HAVING clause is used for applying conditions on group functions
- HAVING clause can filters results

3. Python Analysis

How would you use Pandas to clean a dataset with mixed data types?

- Identify Data Types: Start by checking the data types of each column using (df.dtypes)to understand what's going on.
- Convert Data Types: If some columns have the wrong data type, you can convert them using df['column_name'] = df['column_name'].astype('desired_type'). For example, if a column is showing numbers as strings, you can convert them to integers or floats.
- Handle Categorical Data: If you have categorical data stored as strings, you might want to convert it to a category data type for memory efficiency and faster operations. Use df['column_name'] = df['column_name'].astype('category').
- Fix Inconsistent Formatting: For columns with inconsistent formatting (like dates), you can use pd.to_datetime() to standardize date columns. For string columns, you could use .str.strip() to remove extra spaces or .str.lower() to make everything lowercase.
- Check for Missing Values: Use df.isnull().sum() to check for missing values. You can fill missing values with appropriate methods like df['column_name'].fillna(value) or drop rows with df.dropna().

What are the benefits of using visualizations in data analysis?

1) Makes Data Easier to Understand.
2) Helps with Decision Making.