Assignment Questions

1. What is Pandas? What are the important features of Pandas?

Pandas is an open-source Python library used for data analysis and data manipulation.
It is mainly used for working with structured data such as tables, spreadsheets, and databases.
Pandas provides two main data structures: Series and DataFrame.

Important features of Pandas include:

- Easy handling of missing data.
- Powerful data alignment and indexing.
- Built-in functions for grouping, filtering, and aggregation.
- Support for reading and writing data from different formats like CSV, Excel, SQL, and JSON.
- High performance and efficient data handling.
- Integration with NumPy and other scientific libraries.

2. How do you handle large datasets efficiently in Pandas?

Large datasets can be handled efficiently in Pandas by using techniques such as:

- Loading data in chunks using the chunksize parameter.
- Selecting only required columns while reading files.
- Using appropriate data types to reduce memory usage.
- Avoiding unnecessary copies of data.
- Using vectorized operations instead of loops.
- Using indexing properly for faster searching.

3. What are the different methods available in Series and their purpose?

A Series in Pandas is a one-dimensional labeled array. Some important methods include:

- head() – Displays the first few rows.
- tail() – Displays the last few rows.
- describe() – Provides summary statistics.
- value_counts() – Counts unique values.
- sort_values() – Sorts the data.
- isnull() – Checks for missing values.
- fillna() – Fills missing values.
- unique() – Returns unique values.

These methods help in analyzing and cleaning the data.

4. How do you optimize memory usage in Pandas?

Memory usage can be optimized by:

- Converting data types to smaller types like int32 instead of int64.
- Using category type for repeated string values.
- Dropping unnecessary columns.
- Reading only required columns from files.

- Using chunk processing for large files.

These techniques reduce RAM usage and improve performance.

5. Explain about groupby method in Pandas?

The groupby() method is used to split the data into groups based on some criteria.
After grouping, we can apply aggregate functions like sum(), mean(), count(), min(), and max().
It follows the split-apply-combine concept:
Split – Divide data into groups.
Apply – Perform operations on each group.
Combine – Combine the results into a final output.
It is very useful for analyzing department-wise salary, product-wise sales, etc.

6. What is the difference between Series and DataFrame?

A Series is a one-dimensional data structure that contains a single column of data.
A DataFrame is a two-dimensional data structure that contains multiple rows and columns.

Series:
- Single column
- Labeled index
- Used for single variable data

DataFrame:
- Multiple columns
- Labeled rows and columns
- Used for tabular data

7. Connect SSMS server to Pandas to display any one dataset:

To connect SSMS (SQL Server Management Studio) database with Pandas, we use pyodbc or sqlalchemy.

Example:

```
import pandas as pd
import pyodbc

conn = pyodbc.connect(
    "Driver={SQL Server};"
    "Server=YOUR_SERVER_NAME;"
    "Database=YOUR_DATABASE_NAME;"
    "Trusted_Connection=yes;"
)

query = "SELECT TOP 10 * FROM YourTableName"
df = pd.read_sql(query, conn)
```

```
print(df)
```

This code connects SQL Server and displays dataset in Pandas DataFrame.

8. Take any one dataset from Kaggle and connect to Pandas:

To use a Kaggle dataset, first download the CSV file. Then load it using Pandas.

Example:

```
import pandas as pd

df = pd.read_csv("train.csv")
print(df.head())
```

This reads the Kaggle dataset into a DataFrame for analysis.

9. Create one Excel file and connect to Pandas:

First create an Excel file manually or using Python. Then read it using Pandas.

Example:

```
import pandas as pd

df = pd.read_excel("sample.xlsx")
print(df.head())
```

Pandas provides read_excel() method to import Excel data into DataFrame.