Task 2: Pandas

1. **What is data cleaning in pandas?**

   Data cleaning in Pandas is the process of identifying and correcting or removing errors, inconsistencies, and missing values in a dataset so that it can be accurately analysed or used for machine learning models.
   it involves:

   - Handling missing values
     Detecting NaN or null values
     Filling, replacing, or removing missing data
   - Removing duplicate data
     Identifying repeated rows
     Keeping only unique and relevant records
   - Correcting data types
     Converting columns to appropriate data types
     Ensuring numerical and categorical data are correctly represented
   - Fixing inconsistent data
     Standardizing formats (dates, text case, units)
     Correcting spelling or labelling errors
   - Handling outliers
     Identifying extreme values
     Treating or removing them based on context
   - Renaming and organizing columns
     Making column names clear and consistent
     Improving dataset readability

2. **What is the difference between loc and iloc?**

   loc (Label-based indexing)

   - Selects data using row and column labels
   - Works with index names and column names
   - The end label is included in slicing
   - Supports boolean conditions

   Used when you know labels or names in the dataset.

   iloc (Integer-based indexing)

   - Selects data using integer positions
   - Uses 0-based indexing

- The end index is excluded in slicing
- Does not work with labels

Used when you want to access data by position.

3. What are missing values and how does Pandas handle them?

Missing values occur when data is:

- Not recorded
- Lost during data collection
- Not applicable for certain entries
- Entered incorrectly

In Pandas, these missing entries are usually represented as NaN (Not a Number) or None.

Pandas provides simple and powerful ways to detect, analyze, and handle missing data:

1. Detecting missing values

Pandas can identify missing data and tell you where values are absent.

2. Removing missing values

You can remove rows or columns that contain missing data when they are not useful.

3. Filling missing values

Missing entries can be replaced with:

- A fixed value (like 0 or "Unknown")
- The mean, median, or mode
- The previous or next valid value

4. Ignoring missing values in calculations

Most Pandas operations automatically skip missing values, so calculations like averages are not affected.

4. What is groupby() and why is it used?

In Pandas, groupby() is a method used to split data into groups based on one or more columns, apply some operation to each group, and then combine the results.

This process is often called:

Split → Apply → Combine

Why is groupby() used?

1.To analyze data by categories

2. To perform aggregation

3. To summarize large datasets
4. To support data analysis and reporting

5. How does Pandas help in data preprocessing for AI models?

Pandas plays a key role in data preprocessing because it makes raw, messy data usable for machine learning and AI models. It provides powerful, easy-to-use tools to clean, transform, and organize data before it is fed into an AI algorithm.

1. Handling Missing Data

Pandas helps identify and manage missing values by allowing you to detect, remove, or fill them. Proper handling of missing data prevents errors and improves model accuracy.

2. Data Cleaning and Consistency

With Pandas, you can:

- Remove duplicate records
- Fix incorrect or inconsistent values
- Standardize text, dates, and numerical formats

This ensures the dataset is reliable and consistent.

3. Data Transformation

Pandas supports reshaping and transforming data through:

- Normalization and scaling preparation
- Creating new features from existing data
- Converting categorical data into numerical form

These steps are essential for most AI models.

4. Data Selection and Filtering

It allows easy filtering, sorting, and slicing of data to:

- Select relevant features
- Remove unnecessary columns
- Focus on specific subsets of data

5. Aggregation and Feature Engineering

Using functions like grouping and aggregation, Pandas helps in **feature engineering**, which improves model performance by creating meaningful inputs.

6. Integration with AI Libraries

Pandas works seamlessly with libraries like NumPy and machine learning frameworks, making it easy to pass preprocessed data into AI models.