Course Name:	Data Analysis Laboratory (216H03L501)	Semester:	V
Date of Performance:	14/7/2025	DIV/ Batch No:	D2
Student Name:	Shreyans Tatiya	Roll No:	16010123325

Experiment No: 1

Title: Studies on Pandas library of python.

Objectives of the Experiment:

1. To understand and apply the fundamental functionalities of the pandas library for data analysis.
2. To manipulate and transform datasets using filtering, sorting, and column operations.
3. To analyze data using grouping and aggregation techniques to derive meaningful insights.

COs to be achieved:

CO1: Understand basic concepts of data analytics to solve real-world problems

Books/ Journals/ Websites referred:

1. https://developers.google.com/workspace/sheets/api/guides/pivot-tables#:~:text=This%20guide%20describes%20how%20and,West
2. https://www.freecodecamp.org/news/pandas-dataframe-groupby-method/#:~:text=If%20you're%20familiar%20with,up%20of%20rows%20and%20columns.
3. https://pandas.pydata.org/docs/reference/frame.html

Theory:

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. When working with tabular data, such as data stored in spreadsheets or databases, pandas is the right tool for you. pandas will help you to explore, clean, and process your data. In pandas, a data table is called a DataFrame.

Pandas supports the integration with many file formats or data sources out of the box (csv, excel, sql, json, parquet,…). Importing data from each of these data sources is provided by function with the prefix read_*. Similarly, the to_* methods are used to store data.

Problem statement/ Tasks

Task 1: Import Required Libraries and Dataset

- Import pandas and load a real-world CSV dataset (e.g., Titanic, Student Performance, COVID-19).
- Display the first and last 5 records using head() and tail().

Task 2: Basic Exploration of the Dataset

- Display the dataset shape using .shape, column names using .columns, and data types using .dtypes.
- Generate summary statistics using .describe() and data info using .info().

Task 3: Identify Missing and Duplicate Data

- Detect missing values using .isnull().sum().
- Remove or fill missing values using .dropna() or .fillna().
- Check for and remove duplicate rows using .duplicated() and .drop_duplicates().

Task 4: Filtering Records

- Extract rows based on specific conditions (e.g., students who scored more than 80%, passengers who survived).

Task 5: Sorting the Dataset

- Sort the dataset based on one or more columns using .sort_values().
  - Example: Sort by age or total score.

Task 6: Creating or Modifying Columns

- Create new columns from existing ones (e.g., Total Marks = Math + Science + English).
- Drop unnecessary columns using .drop().
- Rename columns using .rename().

Task 7: Grouping and Aggregation

- Use .groupby() to find average, count, or sum based on a categorical column.
- Example:
  - Average marks by gender: df.groupby('Gender')['Marks'].mean()
  - Survival rate by class: df.groupby('Pclass')['Survived'].mean()

Task 8: Pivot Tables or Multi-Level Grouping (Optional for advanced students)

- Create pivot tables using .pivot_table() to summarize complex data.
  - Example: Average score by gender and class.

Task 9: Insight Generation

- Write 3-5 key insights based on the group-by and aggregated data.
- Example:
  - "Female students have higher average marks in English."
  - "Survival rate is highest for first-class passengers."

Code :

https://colab.research.google.com/drive/16Do-oAdL9D9-WwEsqWmijYU7Tq2CUckU?usp=sharing

Output:

Based on the pivoted table, we observe that among the various categories, **BTC-INR shows the highest average value**, indicating stronger market interest or performance in that segment. Conversely, the category with the **lowest average BTC-INR** may suggest relatively lower demand or volatility. This comparison helps identify which segments are more valuable or stable in terms of BTC-INR pricing.

Post Lab Subjective/Objective type Questions:

1. What is the difference between .info() and .describe() in pandas?

   `.info()` tells you about the *structure* and *completeness* (non-null counts) of your data.

   `.describe()` tells you about the *statistical distribution* of your data.
2. How does pandas handle missing data? Mention at least two functions used for this purpose.

   Pandas handles missing data using special values like `NaN` (Not a Number) for numerical data and `NaT` (Not a Time) for datetime data.
3. What is a pivot table in pandas, and how is it useful in summarizing data?

   A pivot table is a data summarization tool used to analyze large datasets by grouping and aggregating data based on different criteria. It allows users to quickly query, summarize, and explore data by pivoting rows and columns to reveal patterns and insights.
4. What were the key insights you discovered from the dataset during your analysis?

   From this, we can see that hardtop cars have the highest average price, while hatchback cars have the lowest average price.

Conclusion:

This experiment effectively utilized the **Pandas library** to perform fundamental **data analysis**. We successfully loaded, cleaned, transformed, and summarized datasets, applying key techniques like **handling missing data**, **filtering**, **grouping**, and **pivot tables** to derive meaningful insights.