ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA
Department of Information Technology
B. Com Semester VI
Business Data Analytics Practical under NEP

Unit 1: Introduction to Python for Business Data Analytics

B. Python for Data Analysis

Assignment 4: Importing Python Library Pandas for Business Data Analytics

Topics Covered: Pandas for Data Manipulation and Analysis, Handling Missing Data, Data Aggregation and Grouping

The objective of this assignment is to equip with practical skills in using Pandas for data analysis. Students will learn to efficiently load, inspect, and explore datasets while ensuring data quality through cleaning and validation. They will develop the ability to manipulate data by selecting, filtering, renaming, and reshaping it according to the needs. Through aggregation and analysis techniques, students will extract insights via grouping, summarization, and computations. They will master data transformation methods to prepare datasets for advanced analysis and modelling. This assignment aims to provide students with a comprehensive understanding of the data analysis pipeline using Pandas.

Data Loading and Inspection: Load data from various sources (CSV, Excel etc.) into a Pandas DataFrame and inspect its structure, size, and data types. The goal is to understand the overall shape of the dataset, detect missing values, identify column data types, and preview the data for further analysis.

Data Exploration and Cleaning: Explore and clean raw data to ensure its quality and consistency. This includes handling missing values, removing duplicates, detecting outliers, and validating data types. The aim is to prepare a clean, reliable dataset for further analysis.

Data Manipulation: Modify and reshape the DataFrame to extract useful information. This includes renaming columns, filtering rows, selecting specific columns, creating new columns, updating existing values, and deleting unnecessary data.

Data Transformation: Transform and reformat the data to enable analysis and modelling.

Data Aggregation: Summarize, group, and analyze data to extract insights. This involves grouping by categories, calculating totals, counts.

Students are required to create a folder as FourDigitsRollno_Roomno_DeptCode_BDA (For Example, 0012_14_COMA_BDA) on their desktop. All the files related to the assignment should be saved within this folder. Students are required to use either Spyder or Jupyter Notebook to complete the assignment. The final working code should be saved as AssignmentNo_Rollno.py

Creating and Storing Business Data

1. **Sales Data Entry and Storage:** A retail company wants to digitally store daily sales transaction details. Accept sales details from the user for at least 5 transactions, including: Customer ID, Order Date, Region, Product Name, Product Category, Price, Quantity, Profit, Customer Age. Store the entered data in a Pandas DataFrame. Save the DataFrame to a CSV file named sales_data1.csv.

Pandas for Data Manipulation and Analysis (Series & DataFrame)

2. Data Loading and Inspection

Create a simple dataset (sales_data.csv) according to the following structure.

Cust_ID	Order_Date	Region	Prod_Name	Prod_Category	Price	Qty	Profit	Cust_Age
10052	2025-01-01	West	Laptop	Gadgets	3716.80	3	150	34
10093	2025-01-02	West	Smartwatch	Gadgets	3500.23	1	-200	30
10015	2025-01-03	East	Headphones	Electronics	3527.30	1	100	37
10072	2025-01-04	North	Smartwatch	Gadgets	1829.48	8	50	42
10061	2025-01-05	North	Laptop	Gadgets	1503.28	2	200	21
10015	2025-01-03	East	Headphones	Electronics	3527.30	1	100	37

a. Load the dataset into a Pandas DataFrame.
b. Create a Pandas Series from the "Price" column and display its first three values.
c. Display the first two rows and last two rows of the DataFrame.
d. Display a summary of the DataFrame showing column names, non-null counts, and data types.
e. Display the data type of each column.

f. Display the shape, size, and number of dimensions of the DataFrame.

Handling Missing Data and Data Cleaning

3. Data Quality Assessment

a. Display the unique values in the "Region" column and count the total number of unique regions.
b. Check for missing values in the DataFrame and display the count of missing values per column.
c. Identify any duplicate rows.
d. Remove all duplicate rows and display the updated version.

Data Manipulation using Pandas

4. Business Data Transformation

a. Rename the column "Cust_ID" to "CustomerCode" and display the updated column list.
b. Display:

- All rows where "Price" is greater than 1000 using loc()
- The 2nd to 4th columns of the first 4 rows using iloc()

c. Replace negative values in the "Profit" column with 0.
d. Remove the "Cust_Age" column from the DataFrame.

Data Processing

5. Preparing Data for Analysis

a. Convert the "Order_Date" column into datetime format.
b. Sort the DataFrame based on "Price" in ascending order.

Data Aggregation and Grouping

6. Business Insight Generation

a. Group the data by "Region" and calculate the total sales value for each region.
b. Count the frequency of each product category.
c. Display the top three highest and lowest prices.
d. Extract all records where "Region" is West and "Price" is greater than 1000.

****