21/04/2025

Report Day-01

Task 1: Data Cleaning and Preprocessing.

Objective: The goal of this task was to clean and prepare the raw dataset for further analysis. The main objectives were to handle missing data, remove duplicates, standardize formats, and ensure the data is ready for analysis.

Tools Used: Python (Pandas)

Dataset Overview:

- **Dataset Name**: Sample Sales Data
- **Source**: Kaggle - Sample Sales Data
- **Size**: 2,823 rows × 23 columns
- **Contents**: Order details, sales data, customer information, shipping status, order dates, product IDs, and more.
- **Use Case**: Originally designed for Pentaho DI Kettle, this dataset is ideal for segmentation, customer analytics, clustering, and sales simulation training.

Summary of changes made:

- **Removed unnecessary data :** Columns like territory and addressline2 had excessive null values and provided little to no analytical value. These were dropped from the dataset to streamline the data.

- **Replaced null values:**
  - **state column:** Null entries were replaced with 'unknown' to maintain data integrity.

- **postalcode column:** Null values were replaced with '0000' as a placeholder for unidentified.

- **Standardized Date Formats**
  - The orderdate column had inconsistent formats such as '1/2/2002 0:00:00' and '1-2-2022'.
  - All dates were converted to a consistent format: dd-mm-yyyy (e.g., 01-02-2022) using pd.to_datetime() and .strftime().

- **Standardized the text case and formatting:** Columns like state and country were converted to lowercase to maintain uniformity and prevent issues during grouping or filtering.