Machine Learning Engineer

Creative Machine Learning Test Case

Test Case 3 : Data Exploration and Innovation with Pandas

This document provides detailed documentation for the Python script "Task 3: Data Exploration and Innovation with Pandas.ipynb," which focuses on analysing sales data. The script utilises the Pandas library to perform various data exploration and innovative techniques. This documentation aims to provide comprehensive insight into the script's functionality, data processing methods, and the underlying analytical approach employed.

Purpose

This script analyses sales data from a CSV file named "Data.csv". It performs various tasks including:

- Data cleaning and preprocessing
- Exploratory data analysis (EDA)
- Daily sales time series visualisation
- Customer spending analysis
- Product category analysis
- Daily sales clustering
- Anomaly detection in daily sales

Libraries Used

- **Pandas :** Data manipulation and analysis library
- **Ydata_profiling :** Generates a detailed data profile report
- **Matplotlib.pyplot :** Creates visualisations

- **sklearn.cluster.KMeans :** Performs K-Means clustering
- **Pyod.models.abod :** Detects anomalies in data

Script Structure

1. **Import Libraries :** Necessary libraries are imported.
2. **Data Loading :** The **" Data.csv "** file is loaded into a pandas DataFrame named **`data`**.
3. **Data Cleaning :**
    - A copy of the **DataFrame** is created **(`df`)**.
    - The **"Date"** column is converted to **datetime format** with year, month, and day extracted as separate columns.
    - Commas in **"Price"** and **"Total"** columns are replaced with periods and converted to float data type.
    - Typos in product names are corrected using string replacements.
4. **Data Exploration :**
    - Missing values are inspected using **`df.isnull().sum()`**.
5. **Daily Sales Analysis :**
    - Daily total sales are calculated and plotted as a time series.
6. **Customer Spending Analysis :**
    - Customer IDs are formatted for better visualisation.
    - Total spending per customer is calculated and plotted as a time series.
7. **Product Category Analysis :**
    - Daily sales by product category are grouped and visualised using a stacked area chart.
    - Total sales by product category are calculated, sorted, and visualised as a bar chart with average sales displayed for each category.
8. **Daily Sales Clustering :**
    - **K-Means** clustering is applied to group daily sales data into three clusters.

- Daily sales time series with clusters are visualised.
- Clusters are visualised on a scatter plot.
- Average sales for each cluster are analysed.

9. Anomaly Detection :

- Anomaly Detection **(ABOD)** is applied to identify anomalous days in daily sales data.
- Daily sales with anomaly labels are displayed.
- Daily sales time series with anomalies highlighted are visualised.
- Average sales on normal and anomalous days are compared.
- The number of anomalies detected is reported.
- The distribution of sales on anomalous days is visualised using a histogram.