Data Science Assignment Zeotap: eCommerce Transactions Dataset

Task 1

Objective

To analyze eCommerce transaction data, understand customer behavior, product performance, and sales trends, and derive actionable insights for improving business performance.

Data Overview

The dataset includes details on customers, products, and transactions. Key attributes explored during the analysis were:

1. Customers.csv

○ CustomerID: Unique identifier for each customer. ○ CustomerName: Name of the customer. ○ Region: Continent where the customer resides. ○ SignupDate: Date when the customer signed up.

2. Products.csv

○ ProductID: Unique identifier for each product. ○ ProductName: Name of the product. ○ Category: Product category. ○ Price: Product price in USD.

3. Transactions.csv

○ TransactionID: Unique identifier for each transaction. ○ CustomerID: ID of the customer who made the transaction. ○ ProductID: ID of the product sold. ○ TransactionDate: Date of the transaction. ○ Quantity: Quantity of the product purchased. ○ TotalValue: Total value of the transaction. ○ Price: Price of the product sold.

Code Explanation

Each step in the code is designed to clean, analyze, and visualize the data, uncovering patterns and insights that can inform business decisions. Visualizations are used to make the findings accessible and actionable for stakeholders.

- Importing Necessary Libraries
  Libraries like pandas, numpy, matplotlib.pyplot, and seaborn are imported to handle data manipulation, analysis, and visualization.

- Loading the Datasets
  The program reads three datasets: **Customers**, **Products**, and **Transactions**, using the pd.read_csv() function. This allows the program to access and analyze the data.

- Previewing the Data
  The first five rows of each dataset are displayed using the head() method to get an initial understanding of the structure and content.

- Checking for Missing Values
  The program identifies missing data in each dataset using the isnull().sum() method, which helps in determining data cleaning requirements.

- Summary Statistics for Numerical Columns
  Descriptive statistics are generated using the describe() function to understand the distributions, means, medians, and ranges of numerical and categorical variables in each dataset.

- Checking Data Types
  The dtypes property is used to check the data types of all columns, ensuring they are correctly formatted for analysis (e.g., numeric, categorical, or datetime).

- Exploratory Analysis - Customers Dataset
  The distribution of customers by region is analyzed using `value_counts()`, followed by a bar chart created with `seaborn` to visualize the results.
  Monthly trends in customer signups are plotted after converting the `SignupDate` column to datetime format.

- Exploratory Analysis - Products Dataset
  Product categories are analyzed using `value_counts()` and visualized using a count plot.
  A boxplot is created to study the price distribution across categories, and outliers are flagged using the Z-score method from `scipy.stats`.

- Exploratory Analysis - Transactions Dataset
  The program converts the `TransactionDate` column to datetime format for time-based analysis.
  Daily total transaction values are calculated using `groupby()` and visualized as a line plot.
  Monthly sales trends are visualized with a bar chart to identify seasonal patterns.

- Analyzing Top Products by Sales
  Products are ranked by total sales using `groupby()` and `sum()`. The `merge()` function integrates product details for the top 10 products.

- Analyzing Top Customers by Spending
  Customers are ranked by their spending using groupby() and sum(). Customer details for the top 10 spenders are added using the merge() function.

- Correlation Analysis for Transaction
  The correlation between variables like quantity, price, and total value is calculated using the corr() method.

A heatmap is created using seaborn to visualize the relationships between these variables.

Business Insights

Customer Insights

1. The regional distribution analysis showed which areas have the highest number of customers. By targeting campaigns in these regions, we can maximize the return on marketing investment as the audience is larger.

2. The trend of customer signups over time revealed certain months with a noticeable drop in activity. Understanding the reasons behind these declines (e.g., seasonal trends or ineffective campaigns) can help you improve signup rates during these periods.

Product Insights

3. The analysis of product categories showed that some categories are underrepresented, which might indicate unmet demand. Diversifying the product lineup in these areas could attract new customers and increase revenue.

4. The boxplot of price distributions highlighted significant variation in certain categories. By adjusting pricing strategies, we can cater to a wider range of customers, from budget-conscious buyers to premium shoppers.

Sales & Transactions

5. Sales trends over time identified periods with consistently high transaction volumes. Leveraging these times with targeted promotions could further boost sales.

6. The same sales trend analysis showed specific days with lower transaction volumes. Identifying the causes (e.g., low traffic or less engaging offers) and addressing them can help improve overall sales performance.

7. The profit margin analysis revealed which products generate the most profit. Prioritizing these high-margin items in marketing efforts or promotions can significantly improve profitability.

Strategic Focus

8. Analysis of customer spending identified the most valuable customers. These individuals contribute a large share of revenue, so personalized services or loyalty programs can ensure their continued engagement and spending.

9. The analysis of best-selling products highlighted items with the highest sales volumes. Keeping these items in stock prevents stockouts and ensures a steady revenue stream.