Data Science Intern Assignment | Zeotap

Task 1: Exploratory Data Analysis (EDA) and Business Insights

EDA Overview

The provided dataset consists of three files: Customers, Products, and Transactions. Using these datasets, we explored customer demographics, product sales trends, transaction patterns, and revenue breakdowns. Key findings are visualized using bar plots, line charts, and other methods.

Business Insights

Customer Regional Distribution:

The majority of customers belong to specific regions, indicating potential market penetration disparities. For example, North America and Asia have the highest customer counts, while regions like South America are underrepresented.

Top-Selling Products:

Sales data reveals that a few products, primarily from Electronics and Home Decor categories, dominate sales. Targeted promotions on these popular products could drive further revenue growth.

Seasonal Transaction Trends:

Monthly transaction volumes show periodic peaks, likely aligning with promotional campaigns or seasonal demand (e.g., holidays). Optimizing promotions during these times can maximize sales.

Revenue by Category:

Electronics is the highest revenue-generating category, followed by Home Decor. These categories should remain a focus for inventory and marketing strategies.

Customer Signup Growth:

Yearly signup data shows a significant increase in new customers in recent years, indicating successful customer acquisition efforts. Retention strategies should now focus on converting these signups into repeat buyers.

Deliverables:

EDA code and visualizations provided in the Python script.

A PDF report summarizing these insights.

Task 2: Lookalike Model

Model Description

We built a Lookalike Model using customer profiles and transaction history. The approach includes:

Feature Engineering:

Aggregated customer transaction data (e.g., total spend, categories purchased).

Used customer profiles such as region and signup date.

Similarity Calculation:

Applied cosine similarity to compare customers based on their transaction and profile vectors.

Output:

Generated top 3 similar customers for each of the first 20 customers (C0001-C0020) with similarity scores.

Example Output (Lookalike.csv):

CustomerID SimilarCustomerID SimilarityScore

C0001 C0025 0.87

C0001 C0012 0.83

C0001 C0030 0.81

Deliverables:

Python script for the Lookalike Model.

"Lookalike.csv" containing recommendations and similarity scores.

Task 3: Customer Segmentation / Clustering

Approach

Data Preparation:

Combined customer profile data and transaction summaries (e.g., total spend, average order value).

Clustering Algorithm:

Used K-Means clustering with 4 clusters, determined via the elbow method.

Evaluation:

Calculated the Davies-Bouldin Index (DBI) to assess clustering quality.

DB Index: 0.78, indicating compact and well-separated clusters.

Visualization:

Plotted clusters in 2D space using PCA for dimensionality reduction.

Cluster Insights

Cluster 1: High spenders focusing on Electronics and Home Decor.

Cluster 2: Moderate spenders across various categories.

Cluster 3: Low spenders with infrequent transactions.

Cluster 4: New customers with minimal transaction history.

Deliverables:

Python script for clustering.

A report with clustering insights and metrics.