DATA SCIENCE INTERN ASSIGNMENT

Submitted by: Smruthika B J

Institution: KVGCE

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

Business Insights from EDA:

1. **Regional Distribution:**

- Customers are predominantly from South America, followed by Europe and North America.
- South America emerges as the primary market for focused marketing strategies.

2. **Top Product Categories:**

- Books, Electronics, and Clothing dominate as the most purchased product categories.
- Home Decor shows strong growth potential for future campaigns.

3. **Sales Trends:**

- Sales reached their peak in February 2023, suggesting this period should be leveraged for future promotions.
- Identifying the drivers behind this trend is crucial for replication in low-sales months.

4. **Data Preparation:**

- Missing and duplicate records were successfully handled, ensuring data accuracy.
- Aggregated metrics provide a clear understanding of target customers and their behaviors.

5. **Market Focus:**

- Strategic emphasis on top regions and categories can streamline marketing and increase ROI.

Task 2: Lookalike Model

Overview of Lookalike Model:

Objective: Recommend three similar customers for each user based on profile and transaction history.

Methodology:

- Customer and product information were integrated for building the similarity model.
- Cosine similarity was used to determine closeness between customer vectors.

Output:

- Lookalike results for the first 20 customers (CustomerIDs C0001-C0020) are provided in the attached "Smruthika_B_J_Lookalike.csv."
- Each entry includes a similarity score for the recommended customers.

Tools Used:

- Python (pandas, scikit-learn).
- Output validated using exploratory checks.

Task 3: Customer Segmentation / Clustering

Clustering Methodology:

1. **Data Preparation:**

- Datasets (Customers.csv and Transactions.csv) merged using CustomerID.
- Derived features such as total transaction value, average transaction value, total quantity, number of transactions, and customer tenure.

2. **Feature Scaling:**

- Standard Scaler applied to normalize the dataset due to sensitivity of K-Means clustering.

3. **Algorithm:**

- K-Means clustering algorithm applied with five clusters (n=5).
- Evaluated using the Davies-Bouldin Index (DB Index).

Results:

- DB Index: 1.68, indicating well-separated clusters.

Cluster Insights:

1. **Cluster 0:**

- High transaction value and frequency.
- Represents the most profitable customers.
- Strategy: Personalized marketing and loyalty programs.

2. **Cluster 1:**

- Moderate transaction value and frequency.
- Strategy: Upselling opportunities with incentives.

3. **Cluster 2:**

- Low transaction value and frequency.
- Strategy: Re-engagement campaigns with promotional offers.

4. **Cluster 3:**

- Long tenure but low spending.
- Strategy: Retention campaigns to improve frequency.

5. **Cluster 4:**

- New customers with moderate activity.
- Strategy: Onboarding programs to build trust and loyalty.

Visual Representation:

- PCA was used to reduce dimensionality and create a 2D scatter plot for cluster visualization.
- The scatter plot clearly depicts distinct cluster boundaries.

GitHub Link: https://github.com/smruthiiibommetty/ecommerce_analysis