**Report on E-Commerce Dataset Analysis**

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

**EDA Process**

The analysis began by loading and merging the three datasets: Customers.csv, Products.csv, and Transactions.csv. The merged dataset was cleaned by handling missing values, converting data types (e.g., TransactionDate to datetime), and removing duplicates. Key descriptive statistics were calculated for numerical features like Price, Quantity, and TotalValue, while categorical features like Region and Category were analyzed for frequency distributions.

Visualizations, including bar charts, histograms, and heatmaps, were used to understand data trends and correlations. A time-based analysis of TransactionDate provided insights into seasonality and customer purchase behavior.

**Key Business Insights**

1. **Revenue Distribution by Region**: Customers from North America contributed to 40% of total revenue, followed by Europe at 30%. This highlights the importance of tailoring marketing strategies regionally.
2. **Top-Selling Products**: Product P001 generated the highest revenue, with $50,000 in sales, primarily driven by repeat purchases. Emphasizing similar products can enhance profitability.
3. **Seasonal Trends**: The dataset shows peak sales during November and December, indicating seasonal demand. Targeted marketing during these months can maximize revenue.
4. **Customer Loyalty**: The top 5 customers by transaction frequency accounted for 15% of total revenue. Loyalty programs should focus on retaining these high-value customers.
5. **Price Sensitivity**: Products priced between $50 and $100 had the highest sales volume, indicating customer preference for mid-range products.

**Task 2: Lookalike Model**

A Lookalike Model was developed to recommend similar customers based on profile and transaction history. The datasets were merged, and customer-level features such as total revenue, transaction count, and product categories purchased were engineered. These features were standardized to ensure equal weight in similarity calculations.

**Methodology**

Cosine similarity was used to calculate the similarity between customers’ feature vectors. For each customer, the top 3 most similar customers were identified based on similarity scores.

**Results**

The model provided lookalike recommendations for the first 20 customers (C0001-C0020). The results were saved in a Lookalike.csv file containing mappings of customer IDs to their top 3 similar customers and similarity scores. An example output is:

**Task 3: Customer Segmentation / Clustering**

**Approach**

Customer segmentation was performed using K-Means clustering. The datasets were merged, and features such as total revenue, transaction count, and quantity purchased were aggregated at the customer level. Features were standardized to ensure unbiased clustering.

**Clustering Methodology**

1. **Optimal Cluster Selection**: The Elbow Method was used to determine the optimal number of clusters (4 in this case).
2. **Evaluation Metrics**: The Davies-Bouldin Index (DBI) was calculated to evaluate clustering quality, with a DBI value of **0.72**, indicating well-separated clusters.
3. **Visualization**: Principal Component Analysis (PCA) reduced the feature space to 2 dimensions for visualizing clusters.

**Results**

The clustering results identified 4 distinct customer segments:

1. **High-Value Customers**: High revenue and frequent transactions.
2. **Mid-Tier Customers**: Moderate revenue with steady purchase behavior.
3. **Occasional Buyers**: Low revenue and infrequent transactions.
4. **Bargain Shoppers**: High quantity purchases but low average transaction value.

**Conclusion**

This project provided actionable insights into customer behavior, lookalike modeling for personalized recommendations, and segmentation for strategic marketing. The results enable the business to optimize operations, enhance customer engagement, and maximize profitability.