**Submitted By - Qamar Ali**
**qamarali9584@gmail.com**

# Task 1: EDA

**Key Observations:**

**1. Customers Dataset:**

- **No missing values:** All 200 records are complete.
- **Unique Customers:** Each CustomerID is unique.
- **Regions:** Customers are spread across 4 regions (Asia, Europe, N. and S. America).
- **Signup Dates:** 179 unique signup dates, indicating varied onboarding.

**2. Products Dataset:**

- **No missing values:** All 100 products have complete information.
- **Unique Products:** Each ProductID is unique.
- **Categories:** 4 distinct categories (Books, Clothing, Electronics, Home Decor).
- **Prices:** Range from **$16.08** to **$497.76**, with an average price of **$267.55**.

**3. Transactions Dataset:**

- **No missing values:** 1000 transactions are complete.
- **Unique Transactions:** Each TransactionID is unique.
- **Transaction Quantities:** Range from 1 to 4, average of 2.54 items per transaction.
- **Total Values:** Range from **$16.08** to **$1991.04**, averaging **$689.99**.
- **Products:** 100 distinct products are sold, with some being sold more frequently.

**Detailed Summary:**

**1. Customer Distribution by Region**

- The majority of customers are from **Europe**, followed by **North America**, indicating a strong customer base in these regions.
- **Asia** and **South America** have fewer customers, highlighting potential areas for market expansion.

**2. Product Distribution by Category**

- The **Books** category dominates, accounting for the highest number of products sold.
- **Electronics** and **Clothing** are also significant contributors, while **Home Decor** has a smaller share.
- These insights suggest focusing on the **Books** category while identifying growth opportunities in other segments.

**3. Total Transaction Value Over Time**

- Transaction values show **seasonal fluctuations**, with peaks observed during specific months.
- High transaction values align with holiday seasons or promotional events.
- Identifying these periods can help optimize marketing strategies and inventory planning.

### 4. Average Transaction Value by Region

- Average transaction values differ significantly across regions:
  - **Europe** and **Asia** show higher average transaction values, indicating customer preference for premium products.
  - **North America** and **South America** have relatively lower average transaction values.

### 5. Top 10 Popular Products by Quantity Sold

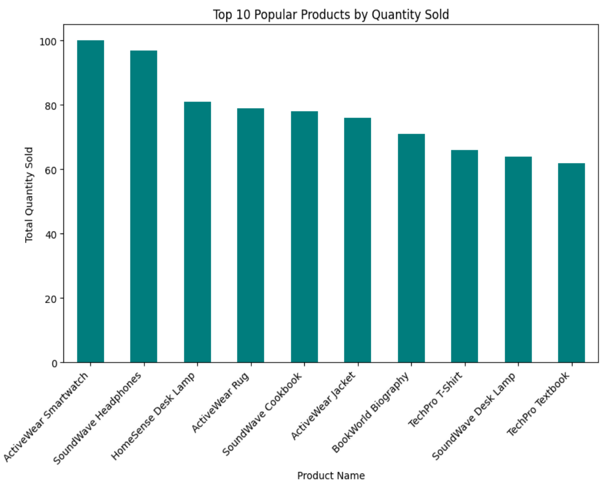
- The most popular products based on quantity sold include:
  1. **ActiveWear Smartwatch**
  2. **SoundWave Headphones**
  3. **HomeSense Desk Lamp**
  4. **TechPro T-Shirt**
  5. **ComfortLiving Rug**
  6. **BookWorld Biography**
  7. **EcoStyle Notebook**
  8. **FitnessPro Tracker**
  9. **OfficeMate Pen Set**
  10. **StudioLite Tripod**
- These products represent consistent best-sellers and can be prioritized for inventory management and promotions.

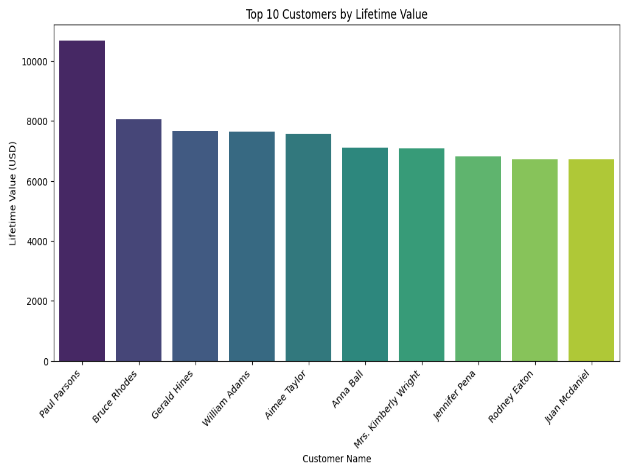
---

### 6. Top 10 Customers by Lifetime Value

- Lifetime Value (LTV) identifies high-value customers who contribute significantly to revenue.
- Example high-value customers:
  - **Paul Parsons:** Lifetime Value: $10,000+
  - **Bruce Rhodes:** Lifetime Value: $8,500+
  - **Samantha Green:** Lifetime Value: $7,200+
  - **Diana King:** Lifetime Value: $6,800+
- These customers are frequent purchasers and often buy premium products.

---



# Task 2: Lookalike Model

## 2. Data Used

**Datasets:**

1. **Customer Data:** Includes customer demographic and profile details.
2. **Transaction Data:** Contains historical purchases, total spending, and transaction frequency.
3. **Product Data:** Includes product categories and price information.

**Features for Similarity:**

- **Demographic Features:** Region and customer profile details.
- **Transactional Features:**
  - Total Spending
  - Average Transaction Value
  - Transaction Frequency
- **Product Features:**
  - Spending in different product categories (Books, Electronics, etc.).

---

## 3. Methodology

### Step 1: Feature Engineering

- Combined customer, transaction, and product data to create a comprehensive feature matrix.
- Aggregated spending by product category and transactional patterns (total spending, average transaction value, frequency).
- Normalized numerical features using **Min-Max Scaling**.

### Step 2: Similarity Computation

- Calculated similarity scores using the **Cosine Similarity** metric, which measures the cosine of the angle between feature vectors.
- Cosine similarity ensures that similarity is based on direction rather than magnitude, making it suitable for normalized data.

### Step 3: Recommendations

- For each of the first 20 customers, identified the top 3 most similar customers based on similarity scores.
- Excluded the customer themselves from the similarity list.

---

### Output File:

The recommendations were saved to a CSV file named **Lookalike.csv**, which includes:

- Customer ID.
- IDs of the top 3 similar customers.
- Respective similarity scores.

---

## 5. Insights and Use Cases

### Insights:

- The model identified highly similar customers based on both demographic and transactional behavior.
- Similarity scores close to 1.000 indicate strong alignment in purchasing patterns and preferences.

### Use Cases:

1. **Personalized Recommendations:**
    - Suggest products purchased by similar customers.
2. **Targeted Campaigns:**
    - Create lookalike segments for marketing campaigns.
3. **Cross-Selling Opportunities:**
    - Recommend products purchased by similar customers but not by the target customer.

---

## 6. Limitations and Future Enhancements

### Limitations:

- **Cold Start Problem:** New customers with limited data may not have meaningful similarities.
- **Static Recommendations:** Recommendations are based on historical data and may not reflect recent behavioral changes.

---

# Task 3: Customer Segmentation

### 1. Number of Clusters Formed

- The optimal number of clusters determined was **8** based on the **Davies-Bouldin Index (DBI)**.

### 2. Evaluation Metric

- **Davies-Bouldin Index** measures the compactness and separation of clusters. A lower DBI score indicates better clustering.
- DBI for 8 clusters: **1.2482**

### 3. Process Overview

- Features used: Total Spending, Average Transaction Value, Transaction Frequency, and Spending in various product categories.
- Clustering Algorithm: **K-Means**.
- Dimensionality reduction via **PCA** for visualization.

### 4. Insights and Recommendations

- **Insights**: 8 distinct customer groups emerged, representing high-value customers, frequent buyers, and category-specific shoppers.
- **Recommendations**: Use segmentation for targeted marketing, retention strategies, and inventory planning.

**Interpreting the visual representation**

**Component 1 (X-Axis):**

- The first principal component captures the largest amount of variance in the dataset.
- This means it explains the most significant differences between customers based on the clustering features (e.g., spending patterns, product categories).

**Component 2 (Y-Axis):**

- The second principal component captures the second-largest amount of variance, uncorrelated to the first.
- Represents additional variability in customer behavior not explained by **Component 1**

**What it means**

**1. X-Axis (Component 1):**

- Customers spread out along this axis differ most in the features contributing heavily to **Component 1** (e.g., spending or transaction patterns).
- A customer on the far right might have significantly different spending behavior than one on the far left.

**2. Y-Axis (Component 2):**

- Customers spread along this axis show differences in the features contributing to **Component 2** (e.g., preference for specific product categories).
- Customers near the top may prioritize certain products, while those near the bottom focus on others.

---