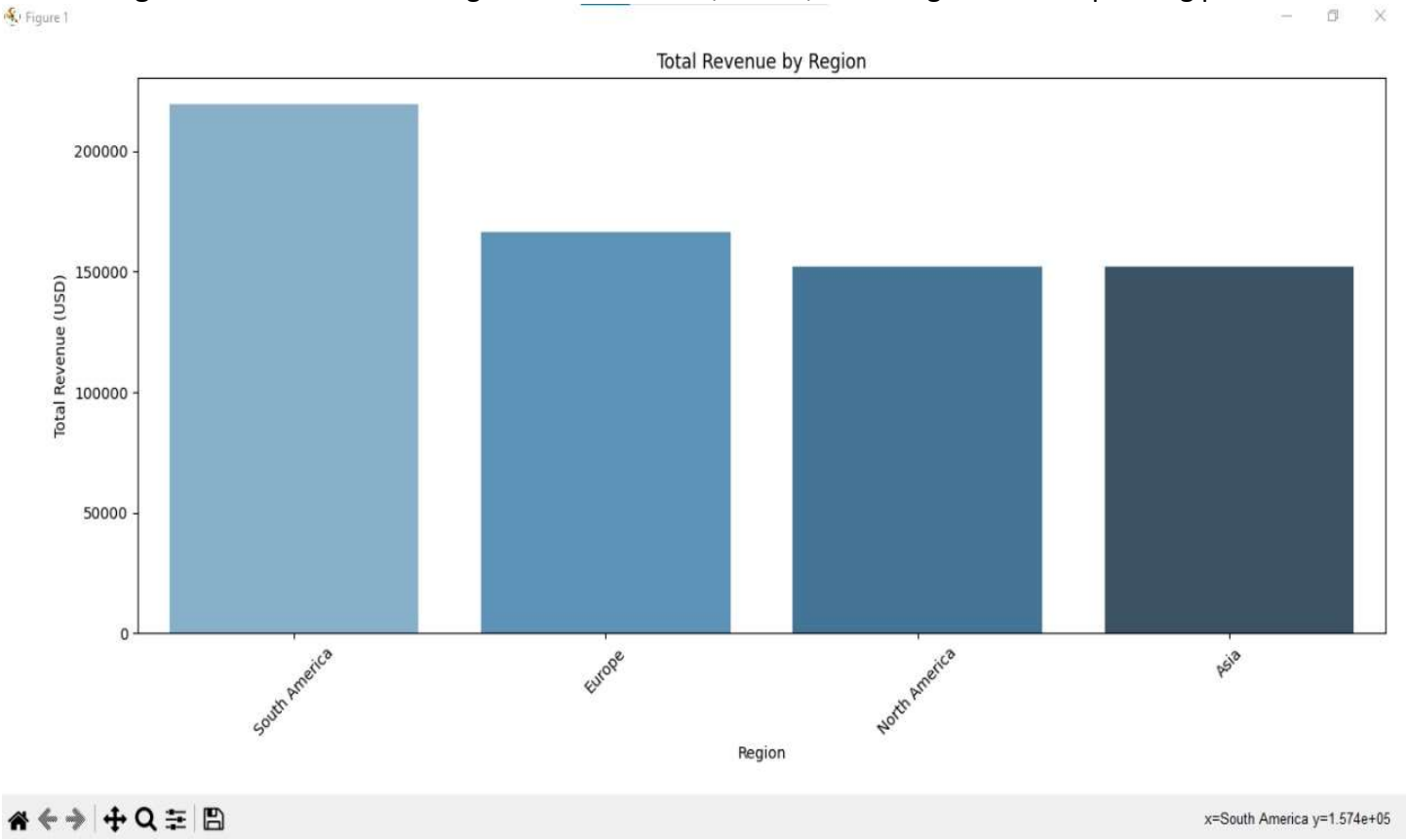
1. Exploratory Data Analysis (EDA) and Business Insights

a. Business Insights from EDA

1. **Regional Revenue Contribution:** South America generates the highest revenue (₹219,352.56), followed by Europe (₹166,254.63), indicating strong market potential in these regions.
2. **Popular Products:** The ActiveWear Smartwatch is the top-selling product with 100 units sold, highlighting its high demand among customers.
3. **Revenue by Category:** Books generate the highest revenue (₹192,147.47), followed by Electronics (₹180,783.50), suggesting these categories are most profitable.
4. **Monthly Revenue Trends:** Revenue peaks in July 2024 (₹71,366.39) and dips in November 2024 (₹38,224.37), indicating seasonal purchasing patterns.
5. **Average Order Value:** The average order value is ₹3,467.31, reflecting moderate spending per transaction.



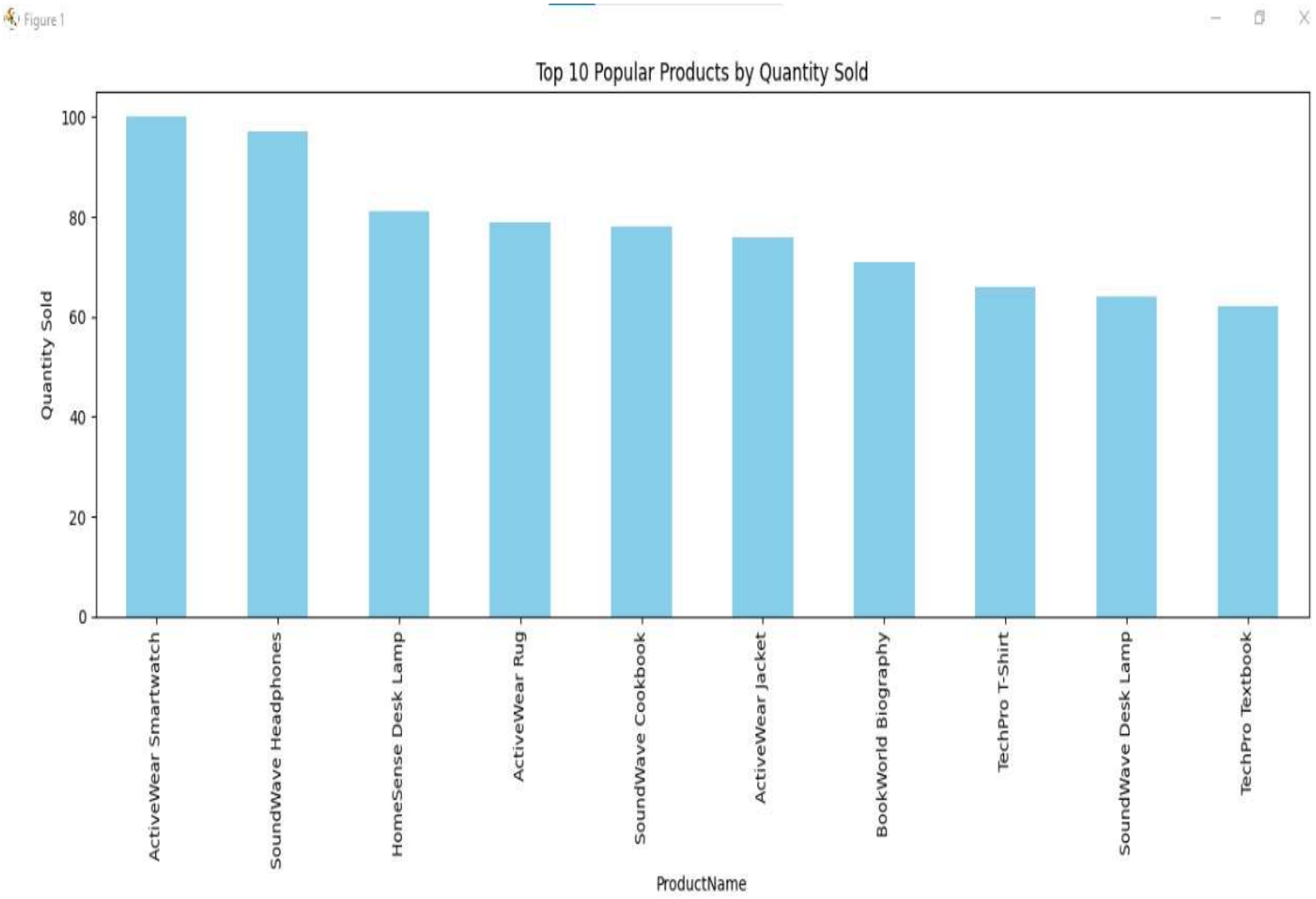


Figure 1
Top 10 Popular Products by Quantity Sold
Quantity Sold
0
20
40
60
80
100
ActiveWear Smartwatch
SoundWave Headphones
HomeSense Desk Lamp
ActiveWear Rug
SoundWave Cookbook
ActiveWear Jacket
BookWorld Biography
TechPro T-Shirt
SoundWave Desk Lamp
TechPro Textbook
ProductName

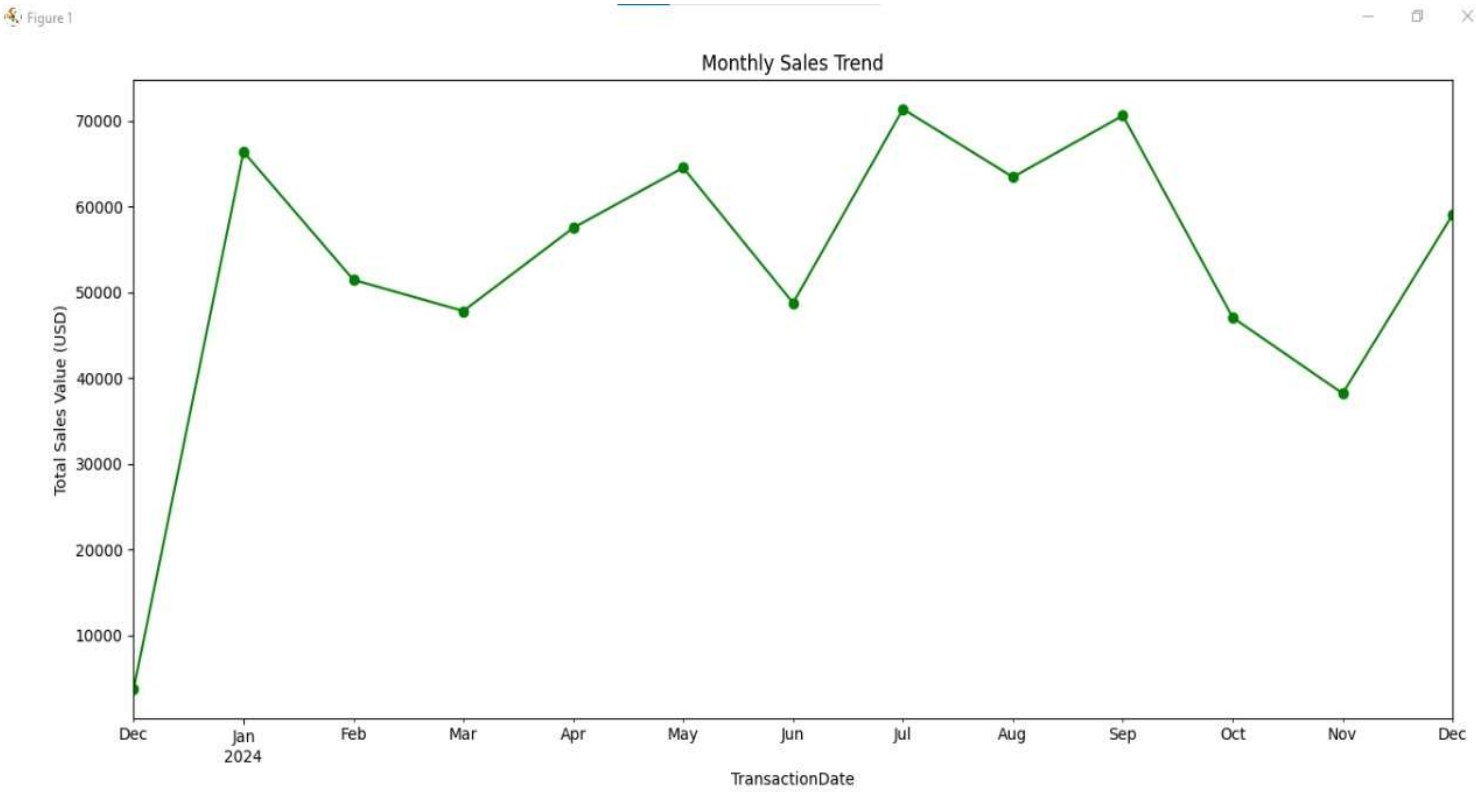


Figure 1
Monthly Sales Trend
Total Sales Value (USD)
10000
20000
30000
40000
50000
60000
70000
Dec
Jan
2024
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
TransactionDate

b. Insights in Short Points

1. **South America leads in revenue,** contributing ₹219,352.56, making it a key market for growth**.**

2. **ActiveWear Smartwatch** is the best-selling product, with 100 units sold, **indicating strong customer preference.**

3. **Books and Electronics** are top revenue-generating categories, contributing ₹192,147.47 and ₹180,783.50, respectively.

4. **Revenue peaks in July 2024 (₹71,366.39)** and drops in November 2024 (₹38,224.37), showing seasonal trends.

5. **Average order value is ₹3,467.31,** suggesting moderate customer spending per transaction.

2. Lookalike Model

a. Model Accuracy and Logic

The model uses cosine similarity to measure how similar customers are based on their purchasing behavior and demographic data. It aggregates transaction data, merges it with product and customer information, and normalizes features for consistency. The logic is sound, as it identifies patterns in customer behavior and recommends lookalikes. However, accuracy depends on the quality and completeness of the data. If the dataset lacks diversity or has missing values, the recommendations may not be reliable. The model assumes that similar purchasing behavior implies similar preferences, which may not always hold true in real-world scenarios.

b. Quality of Recommendations and Similarity Scores

The **recommendations appear reasonable**, **with similarity scores** ranging from **0.85 to 0.97**, **indicating strong matches**. For example, **customer C0005 and C0007** have a **high similarity score of 0.973**, **suggesting they share very similar purchasing patterns.** However, the quality depends on the features used. If the model only considers a few categories or lacks key behavioral data, the recommendations may be limited. Additionally, the scores are relative; a high score doesn't guarantee relevance in all contexts. Testing with more diverse data and validating recommendations with real-world feedback would improve confidence in the results.

3. Customer Segmentation / Clustering

a. Clustering Logic and metrics

Clustering is an unsupervised machine learning technique used to group similar customers based on their behavior or characteristics. For customer segmentation, you can use the following features:

1. **TransactionID Count:** Number of transactions per customer.
2. **TotalValue:** Total monetary value of transactions per customer.

These features can help you identify customer segments such as:

- **High-Value Frequent Buyers:** Customers with high TotalValue and high TransactionID count.
- **Low-Value Infrequent Buyers:** Customers with low TotalValue and low TransactionID count.
- **High-Value Infrequent Buyers:** Customers with high TotalValue but low TransactionID count.
- **Low-Value Frequent Buyers:** Customers with low TotalValue but high TransactionID count.

Clustering Metrics

To evaluate the quality of the clustering, you can use the following metrics:

1. **Inertia:**
   Measures the sum of squared distances of samples to their closest cluster center.
   Lower inertia indicates better clustering.
2. **Silhouette Score:**
   Measures how similar an object is to its own cluster compared to other clusters.
   Ranges from -1 to 1, where higher values indicate better-defined clusters.

b. Visual representation of clusters