Online Retail Customer Segmentation & Insights – Final Project Summary

---

Project Objective:

1. Derive **useful insights** from customer purchasing history to benefit the online retailer.
2. **Segment customers** based on their purchasing behavior using unsupervised learning.

---

1. Data Preprocessing

- **Loaded dataset** with proper encoding (unicode_escape) to handle special characters.
- **Removed rows with missing CustomerID** to ensure valid customer-level analysis.
- Converted InvoiceDate to datetime format.
- Created a new column:
  TotalAmount = Quantity × UnitPrice to represent monetary value of each invoice line.
- Cleaned and validated all necessary columns for accurate computation.

---

2. Feature Engineering

Grouped data at the CustomerID level to generate a **customer summary** table with the following metrics:

- NumPurchases: Number of unique invoices
- TotalQuantity: Sum of quantity purchased
- TotalSpend: Total money spent
- AvgUnitPrice: Average price per item

This aggregated table (customer_summary) was used for clustering.

---

3. Feature Scaling

- Used **StandardScaler** from sklearn.preprocessing to standardize numerical features.
- Scaling was necessary to bring all features to a similar range for effective distance-based clustering.

---

4. Customer Segmentation using K-Means

- Applied the **Elbow Method** to determine the optimal number of clusters (k=4).
- Ran **KMeans Clustering** on the standardized data.
- Assigned cluster labels to each customer (Cluster column).

---

5. Visualization with PCA (Principal Component Analysis)

- Reduced the dimensions to **2D** using PCA for visualization.
- Plotted the clusters on a scatter plot using PCA1 and PCA2.
- While PCA helped with visualization, the actual interpretation was driven by cluster-wise metrics.

---

6. Cluster Interpretation

Used groupby('Cluster').mean() to analyze the characteristics of each cluster and mapped them to real-world segments:

Cluster	Segment Name	Characteristics
0	Average Shoppers	Medium purchase frequency and spend
1	High-Value Customers	High spenders with frequent purchases
2	Regular Customers	Consistent buyers with balanced quantity and spending
3	Refund-Prone Customers	Negative TotalSpend, low quantity, high average unit price

---

7. Additional Insights

- **Monthly sales** trends were analyzed using InvoiceDate; highest sales observed during holiday months.
- **Bar chart** was created to show the number of customers in each segment.
- **Refund-Prone Customers** identified using negative spend values, helping the business detect possible fraud or return abuse.

Business Value

- Enables **targeted marketing** based on customer value.
- Identifies **refund-prone customers** for fraud prevention or stricter return policies.

- Provides a **data-driven basis for customer loyalty programs** and personalized campaigns.
- Supports **operational planning** by identifying purchase peaks and customer trends.

Tools & Techniques Used

- Python, Pandas, NumPy
- Matplotlib & Seaborn (for visualization)
- Scikit-learn: KMeans, PCA, StandardScaler
- Groupby, feature creation, and filtering techniques in pandas

---

Conclusion

This project effectively combines **data preprocessing**, **feature engineering**, **clustering**, and **visual interpretation** to derive **actionable insights**.
It supports better business decisions through **customer segmentation** and highlights hidden patterns like refund behavior.
