Customer Segmentation

AUTHOR NAME : Velkumar M
DATE : 26 June, 2025

Project details

Client

UK-based Online Retailer

About the company

UK-based non-store online retailer

Main Customers

Individual retail buyers and wholesale clients

Customer Segmentation

Products sold

Unique, all-occasion gift items such as home décor, kitchenware, candles, toys, and holiday-themed gifts

Date Range

December 1, 2010 – December 9, 2011

Objective

To analyse historical purchase behaviour of customers and design a predictive strategy that anticipates future purchases for new customers, based on the patterns observed from their first transaction.

Problem statement

The challenge is to understand customer purchasing behaviour and to predict what a new customer is likely to buy during the year, based solely on their first transaction.

The client, a UK-based online retailer, has over half a million transactions from ~4,000 customers across one year.

Project Goals:

1) **Analyse Transaction Patterns :** Understand purchase frequency, value, and customer types.

2) **Segment Customers :** Identify distinct customer groups based on buying behaviour (RFM Analysis).

3) **Visualise Key Insights :** Using charts and dashboards to uncover trends in time, geography, and product preferences.

4) **Predict Future Purchases :** Anticipate what a new customer might purchase next based on historical patterns.

5) **Support Business Strategy :** Helps in creating personalised marketing and inventory planning strategies.

Roadmap of the project

Understanding the Dataset

Data Cleaning

Exploratory Data Analysis

Feature Engineering

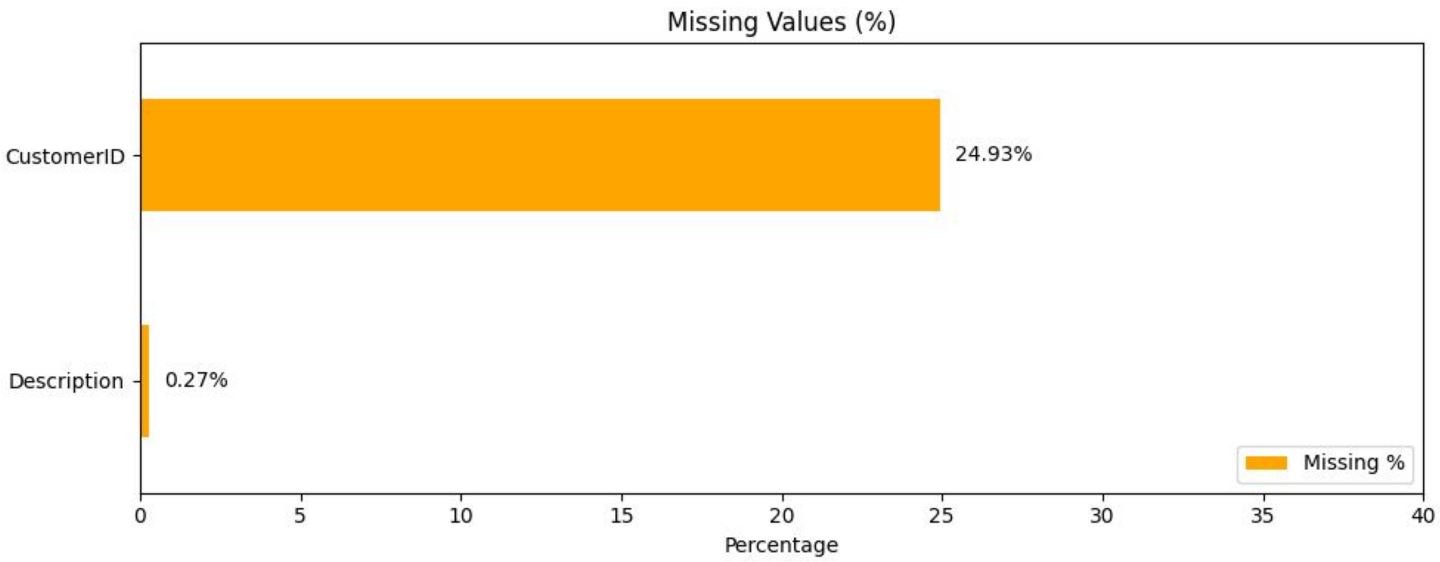
Prediction & Clustering

Data set overview

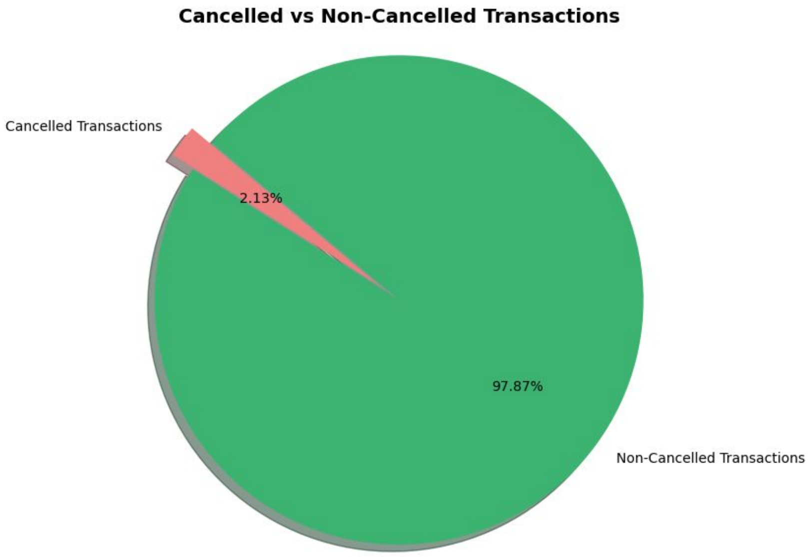
COLUMN NAME	DESCRIPTION
InvoiceNo	Unique invoice number for each transaction (starts with "C" if canceled)
StockCode	Unique code assigned to each product
Description	Name/description of the product
Quantity	Number of units purchased per product per transaction
InvoiceDate	Date and time of purchase
UnitPrice	Price per unit in GBP
CustomerID	Unique identifier for each customer
Country	Country where the customer is located

Total Records : 5,41,910
Total unique Customers : ~4,000
Total Unique Products: ~4,070 (Identified by StockCode)
Geographical Scope: 38 Countries (Majority from the United Kingdom, followed by Germany, France, and others)

Data cleaning



Removed missing CustomerID & Description
→ Incomplete or invalid transactions



Rather than removing cancelled transactions, they are **retained and flagged** for deeper analysis.

Removal of Anomalous Stock Codes :

Some Stock Code values were found to represent non-product or irrelevant entries, such as:

POST, D, C2, M, BANK CHARGES, PADS, DOT, CRUNK

Inferences on Product Descriptions:

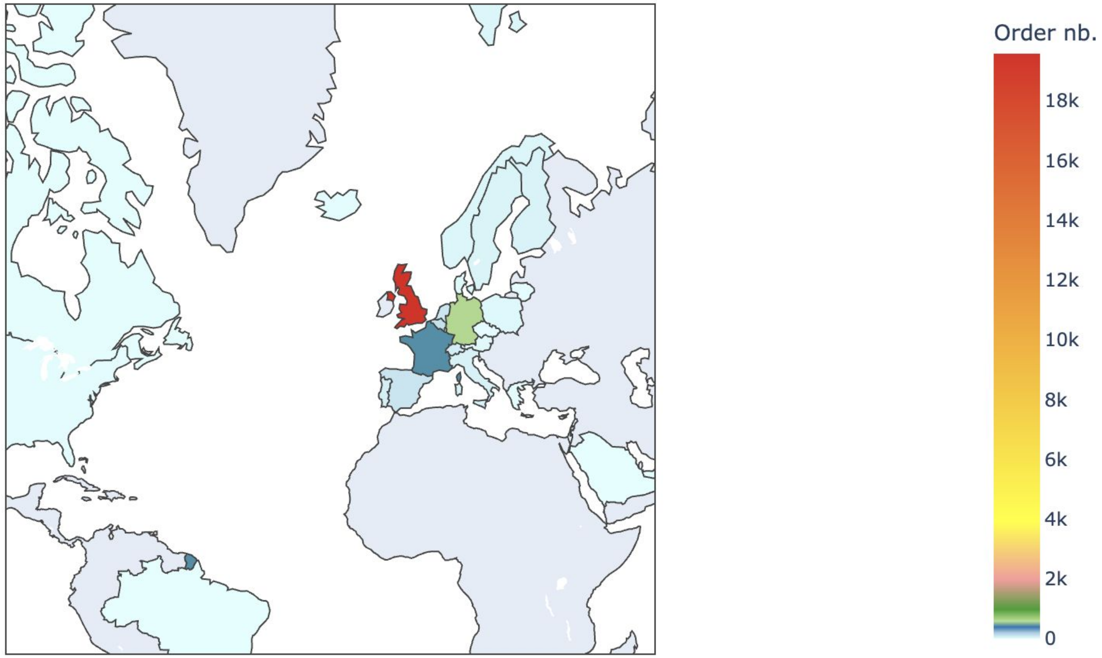
- Most frequent descriptions relate to **kitchenware, lunch bags, and decorative items**.

Identified non-product descriptions like **"Next Day Carriage"** and **"High Resolution Image"**, which were removed.

- Remaining mixed-case entries were standardized to

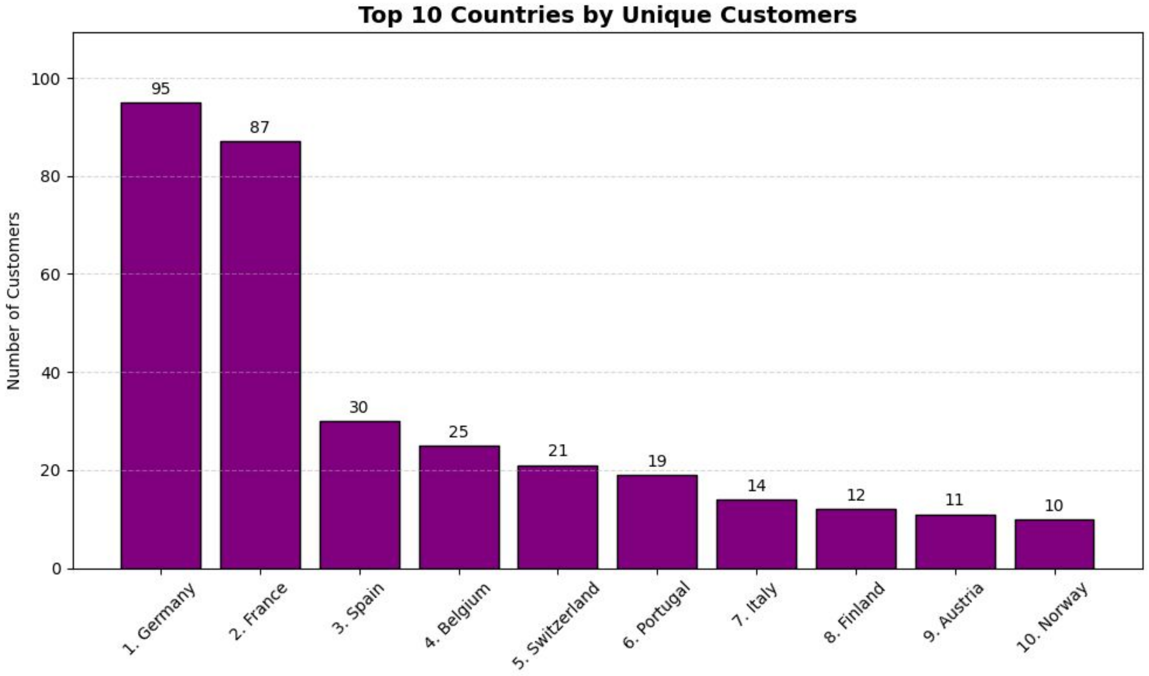
Geographic distribution of customers & orders

Number of orders per Country



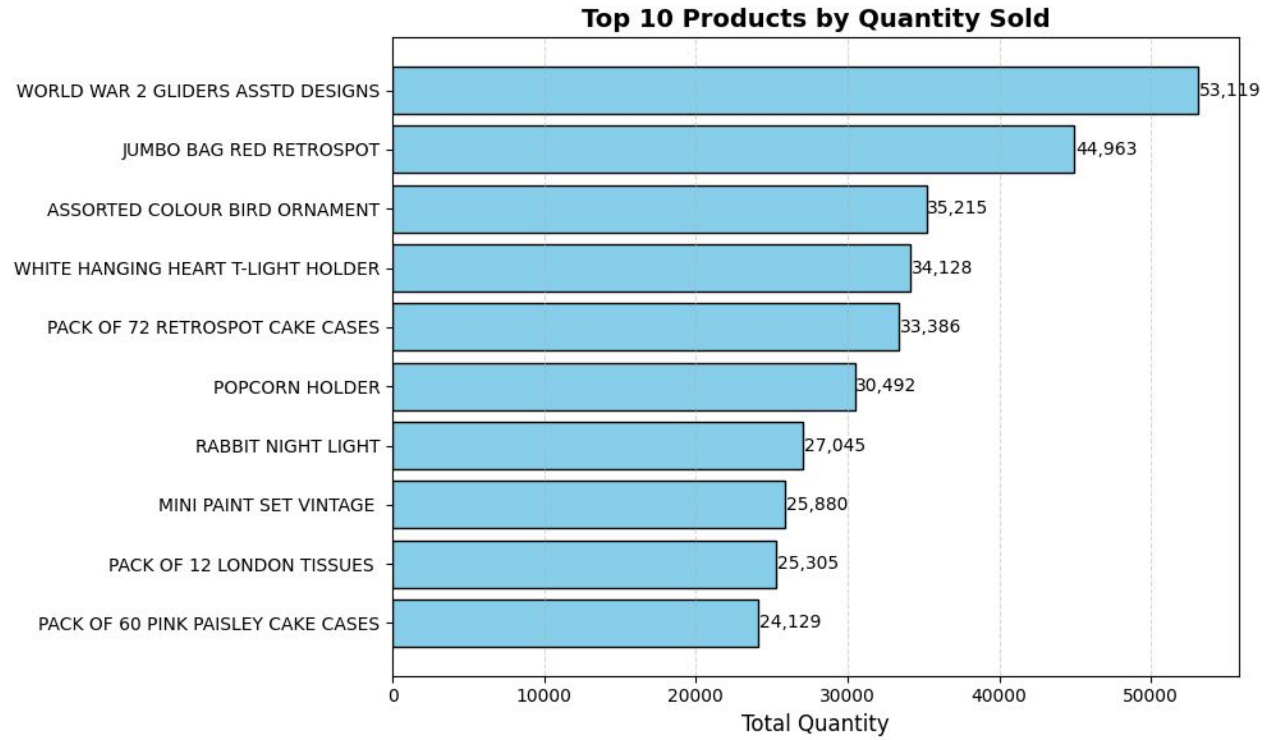
Key Insights:

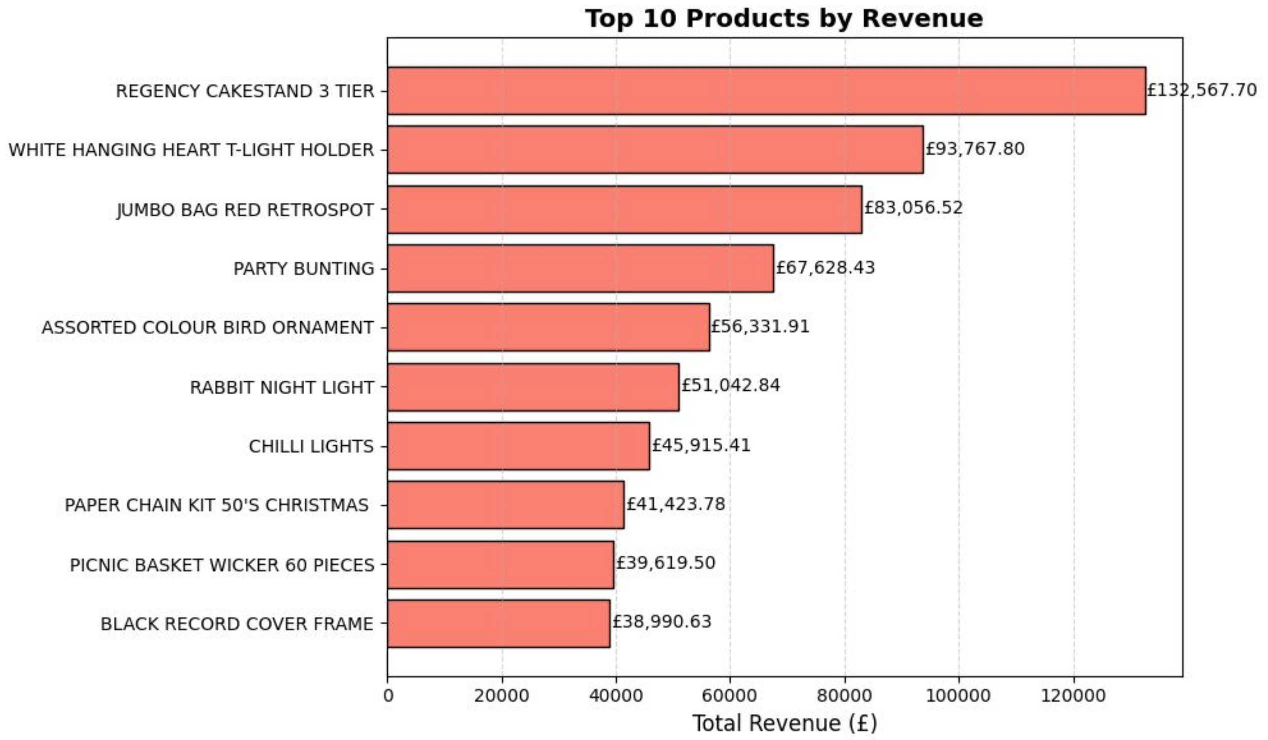
- **UK dominates** customer base and order volume, being the home country of the business — hence excluded from comparative charts.



- **Top 10 countries by unique customers** (excluding UK) are primarily European nations, suggesting strong regional demand.
- **Germany, France** show the highest number of non-UK customers.

Data Insights – Product popularity & performance

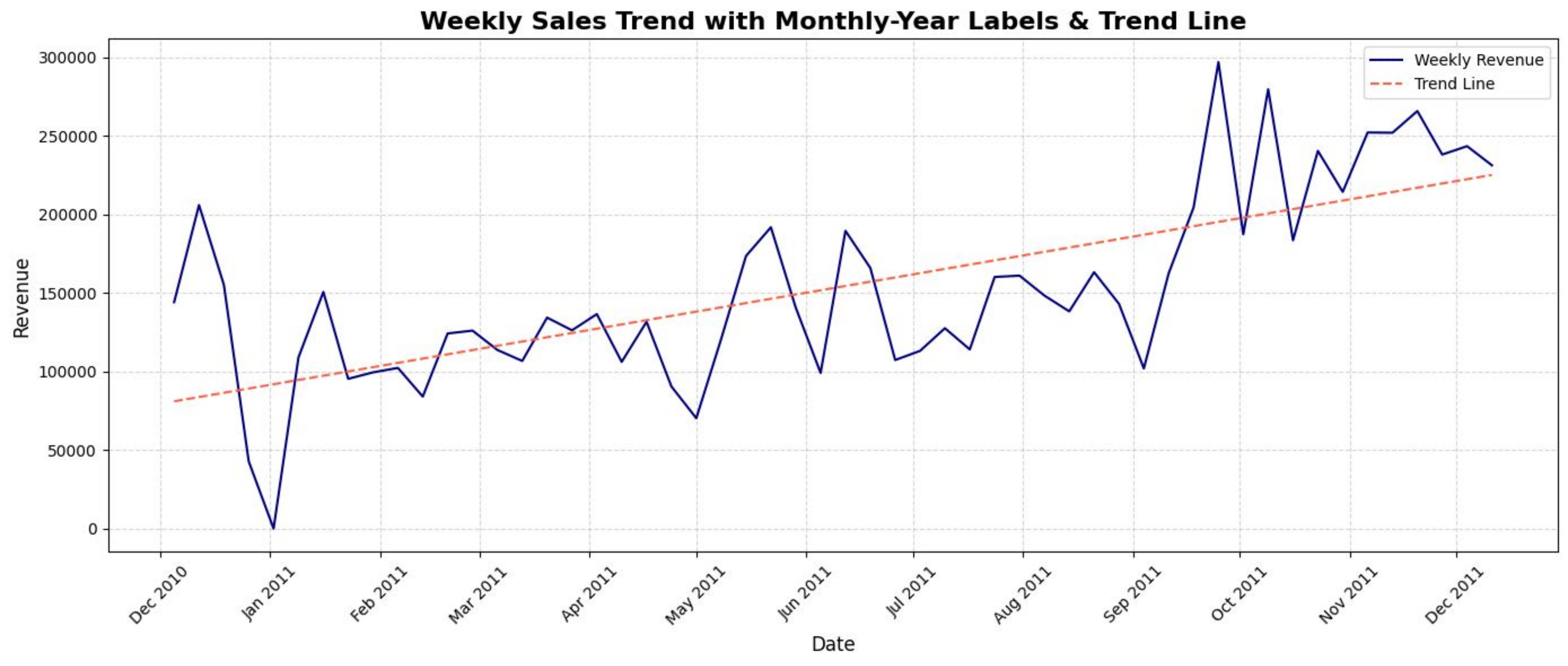




Insights :

1) High-Selling ≠ High-Earning: Products with the highest quantities sold are generally low-cost items, contributing less to overall revenue.
2) Premium priced products are the revenue drivers

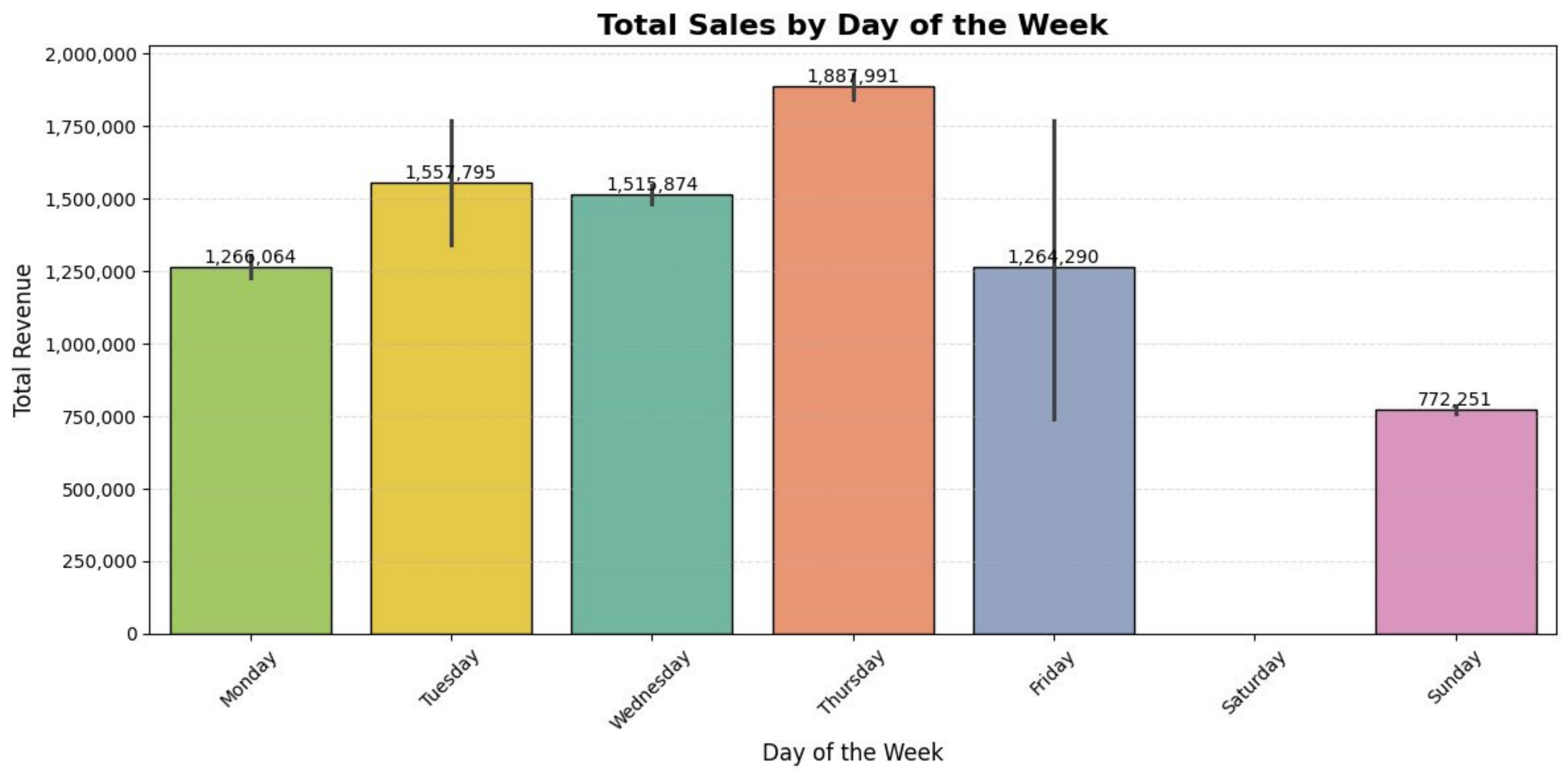
Sales trend analysis



Weekly Sales Trend with Linear Growth Pattern

- A **clear upward trend** can be observed, especially from mid-year onward, indicating **growing customer activity and higher sales volume** over time.
- The **linear trend line** (dashed red) confirms a **steady increase in sales**

Sales analysis

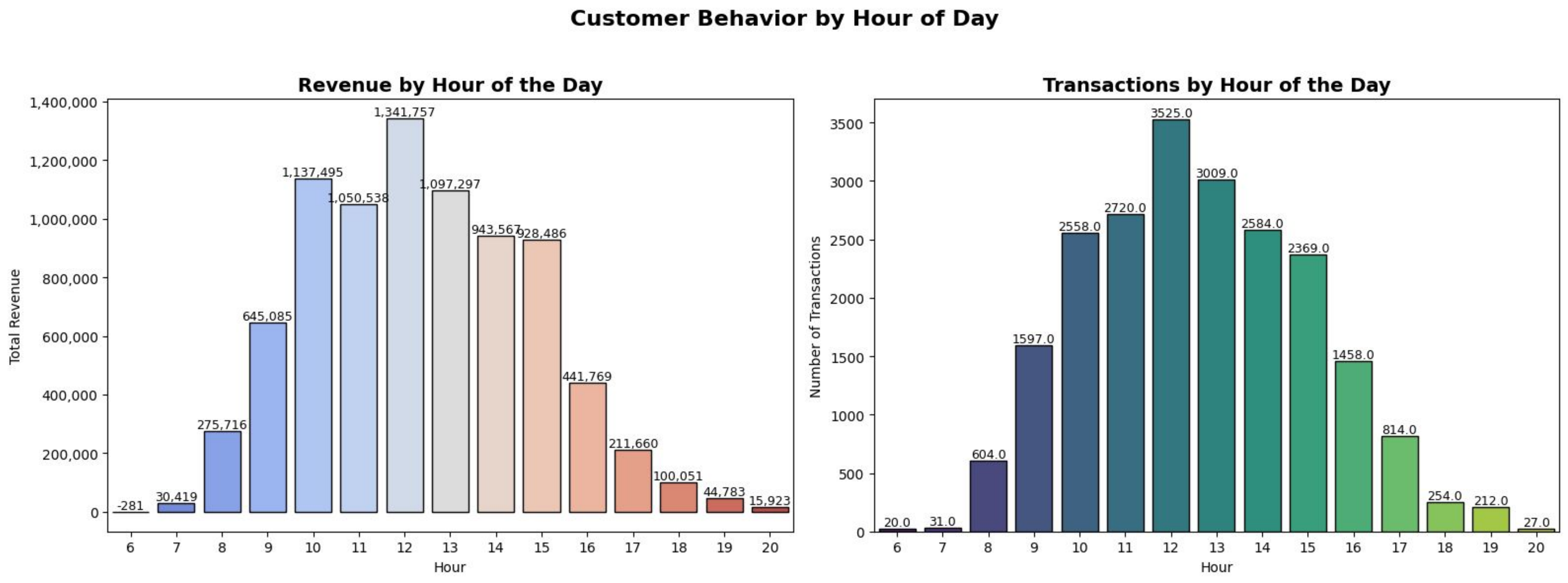


Insights :

1) Thursday records the highest total sales, making it the most active shopping day.

2) Saturday shows zero sales, which could indicate a non-operational day or company holiday.

Customer behaviour



Insights :

1) A higher **number of transactions** is observed during late morning to early afternoon (11 AM – 2 PM).

2) This suggests that customers are **more likely to shop during midday**, possibly during breaks or lunch hours.

Feature engineering

Train-Test Split :

- Total customers: **4362**
- Split into:
  - **3489 (80%)** for training
  - **873 (20%)** for testing
- Splitting was done **before feature engineering** to avoid data leakage.

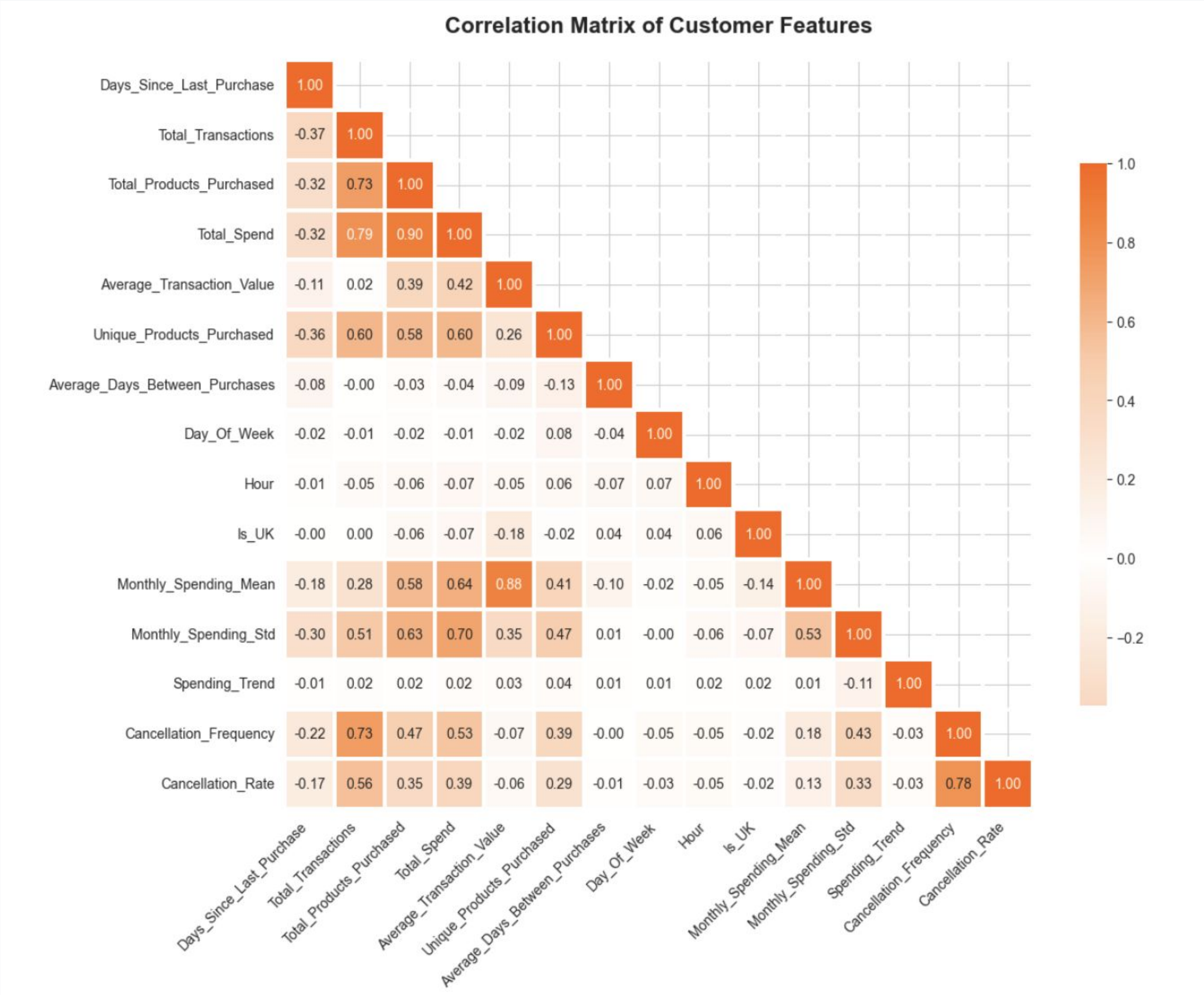
Feature Engineering :

Feature engineering was performed on the training dataset by aggregating transactional-level data to the customer level.In total, **15 customer-level features** were created across dimensions like:

- Recency & frequency of purchases
- Monetary value
- Product diversity
- Time-based trends
- Cancellations & returns
- Geographic info

```
Data columns (total 16 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   CustomerID                     3489 non-null   float64
 1   Days_Since_Last_Purchase       3489 non-null   int64
 2   Total_Transactions             3489 non-null   int64
 3   Total_Products_Purchased       3489 non-null   int64
 4   Total_Spend                    3489 non-null   float64
 5   Average_Transaction_Value      3489 non-null   float64
 6   Unique_Products_Purchased      3489 non-null   int64
 7   Average_Days_Between_Purchases 3424 non-null   float64
 8   Day_Of_Week                    3489 non-null   int32
 9   Hour                           3489 non-null   int32
 10  Is_UK                          3489 non-null   int64
 11  Monthly_Spending_Mean          3489 non-null   float64
 12  Monthly_Spending_Std           3489 non-null   float64
 13  Spending_Trend                 3489 non-null   float64
 14  Cancellation_Frequency         3489 non-null   float64
 15  Cancellation_Rate              3489 non-null   float64
dtypes: float64(9), int32(2), int64(5)
memory usage: 409.0 KB
```

Dimensionality reduction & scaling



Dimensionality Reduction using PCA:

The correlation matrix reveals that several features are highly correlated, introducing multicollinearity.

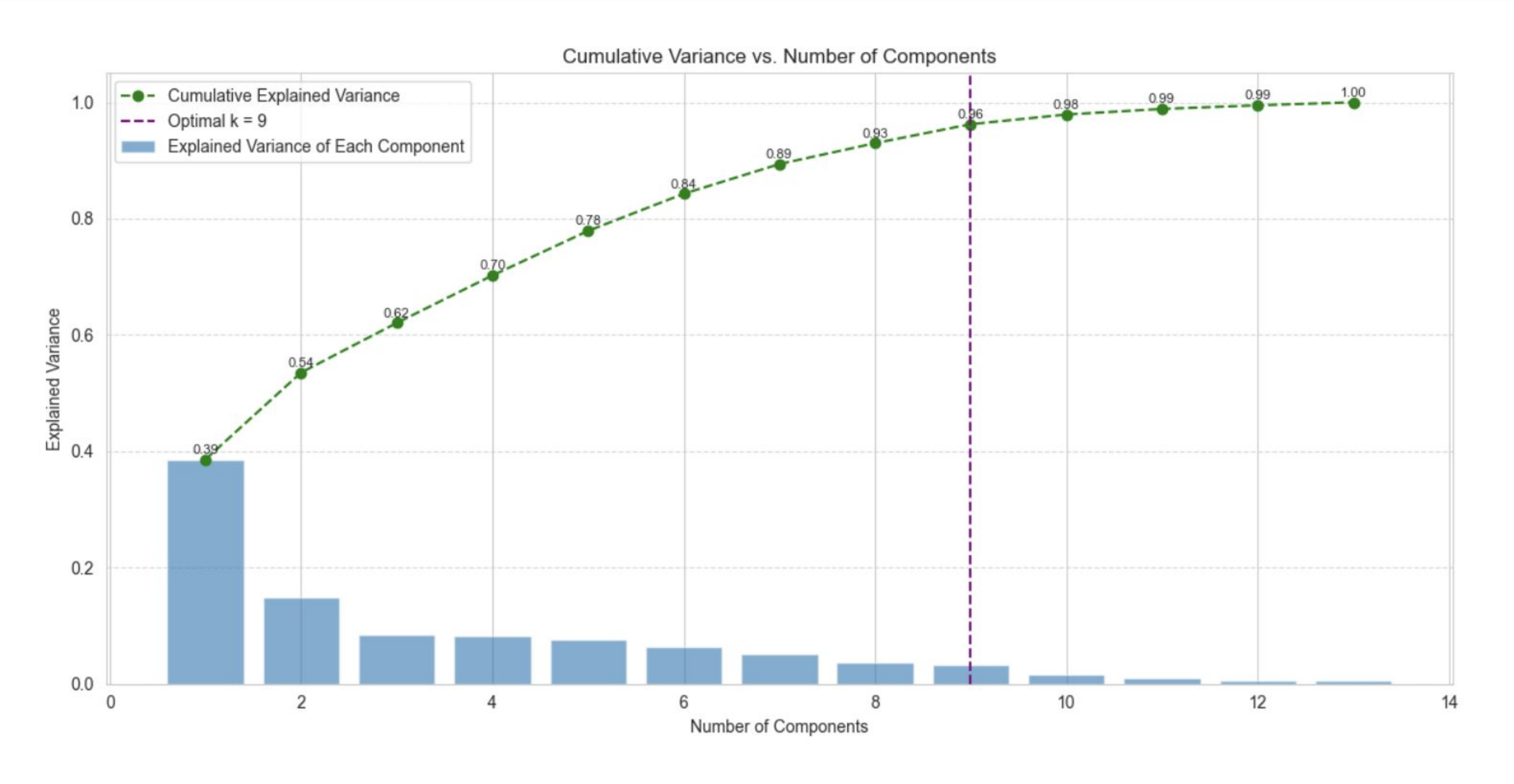
Why this is an issue:

- Correlated features contain redundant information.
- This can distort distance calculations used in clustering.
- It may lead to biased or **overlapping** clusters.

To address this, we applied Principal Component Analysis (PCA) to:

- Remove multicollinearity
- Reduce feature space while preserving variance
- **Improve clustering** quality and visualisation

Principal Component Analysis



We had 15 original features. PCA helped reduce dimensionality by transforming them into 9 principal components which still explain **96% of the variance.** These components are not individual features, but combinations of the originals. So **PCA doesn’t remove features directly** — it compresses them while retaining maximum information

Principal Component Analysis

	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9
0	-2950.061293	-2.379035	-1.733134	-1.589924	-0.077159	-1.224425	1.483082	0.341658	1.099029
1	-2948.061524	0.050618	0.567753	0.421267	-0.912138	0.911629	-2.163709	1.688222	0.691543
2	-2947.062216	1.604882	-2.544528	5.432994	0.733382	-0.460195	0.964361	-0.005842	-2.835178
3	-2944.061511	0.242700	-1.501119	-1.241069	-0.672642	0.587045	-0.472281	0.062493	-0.274631
4	-2943.061277	-2.516858	0.474674	-0.683677	-1.558638	0.254045	-0.623549	1.152407	0.656529

After applying Principal Component Analysis (PCA) to the scaled dataset, we reduced the **15 original features into 9 new components** (Feature_1 to Feature_9).
These are not direct columns from the original data but linear combinations of them that retain maximum variance.

- The table here shows the transformed dataset after PCA.
- Each Feature_i (i = 1 to 9) is a principal component — capturing patterns from all 15 original features.
- This transformation **removes redundancy** caused by multicollinearity and **improves clustering performance.**
- These components explain around 96% of the total variance, ensuring we preserved most of the original information while reducing dimensionality.

Why K-means clustering?

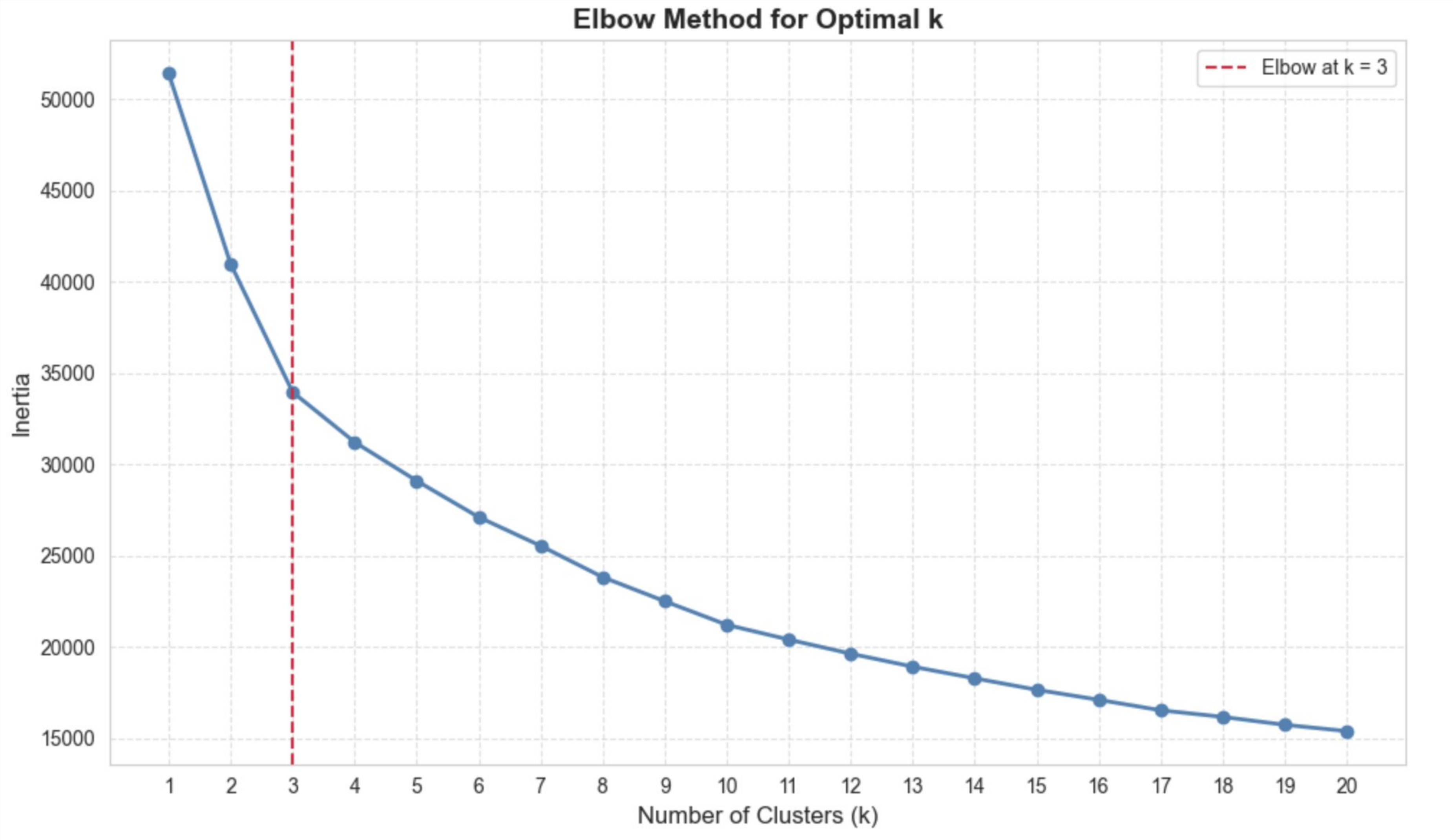
- **Efficient for Large Datasets**
  Hierarchical clustering builds a dendrogram and has high time complexity ($O(n^2)$), making it impractical for datasets with thousands of customers.

- **Well-Separated Clusters**
  K-Means ensures that each data point belongs to exactly one cluster, unlike DBSCAN which allows overlapping clusters — unsuitable for distinct customer groups.

- **Handles Uniform Densities Better**
  DBSCAN struggles when clusters have varying densities, often merging dense and sparse regions incorrectly. K-Means assumes roughly equal density and works well in such scenarios.

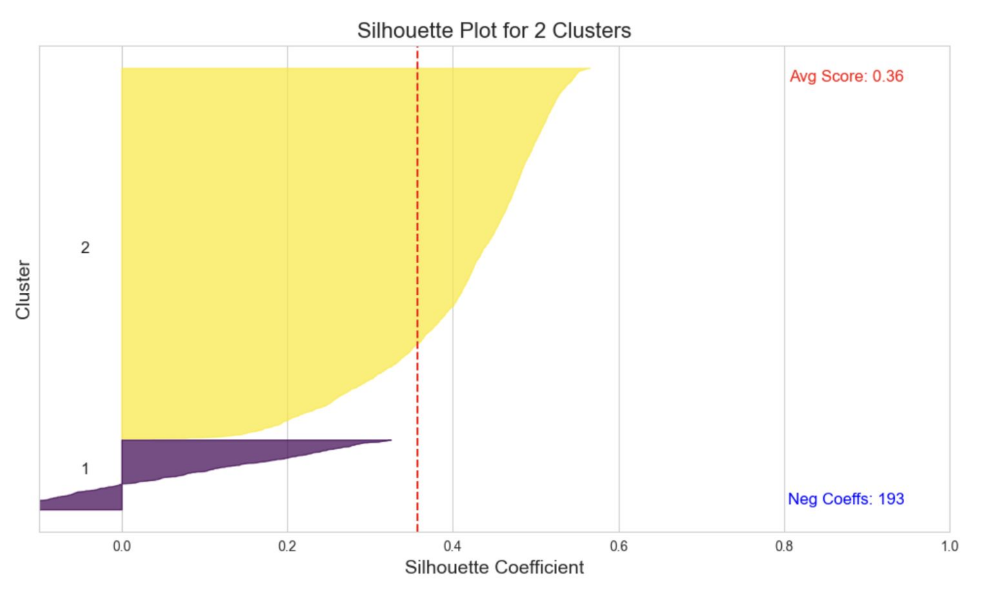
- **Integration with PCA**
  K-Means works effectively in reduced dimensions, making it ideal to combine with PCA for better performance and visualization.

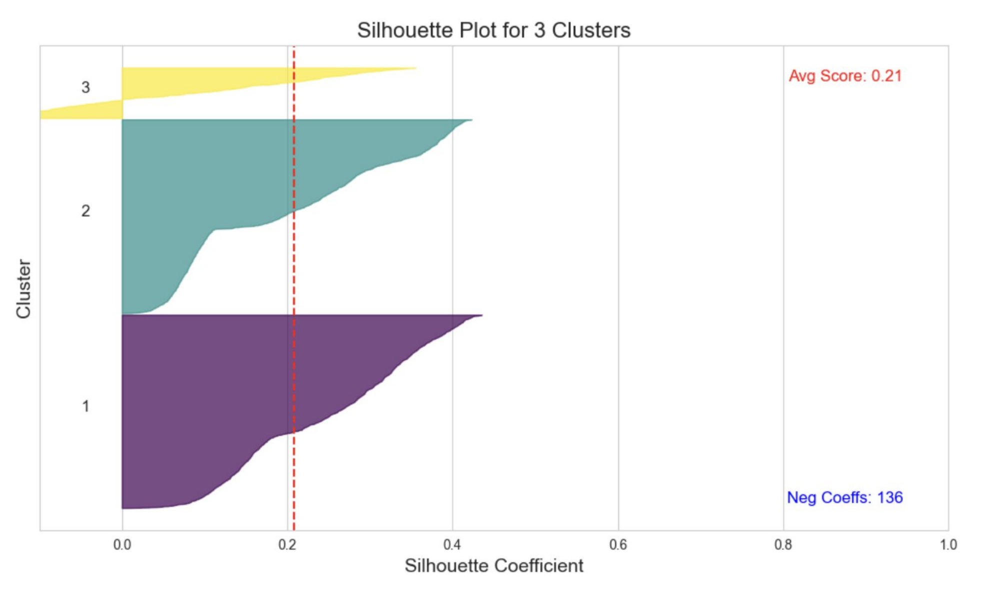
K- Means Clustering(Optimal k)

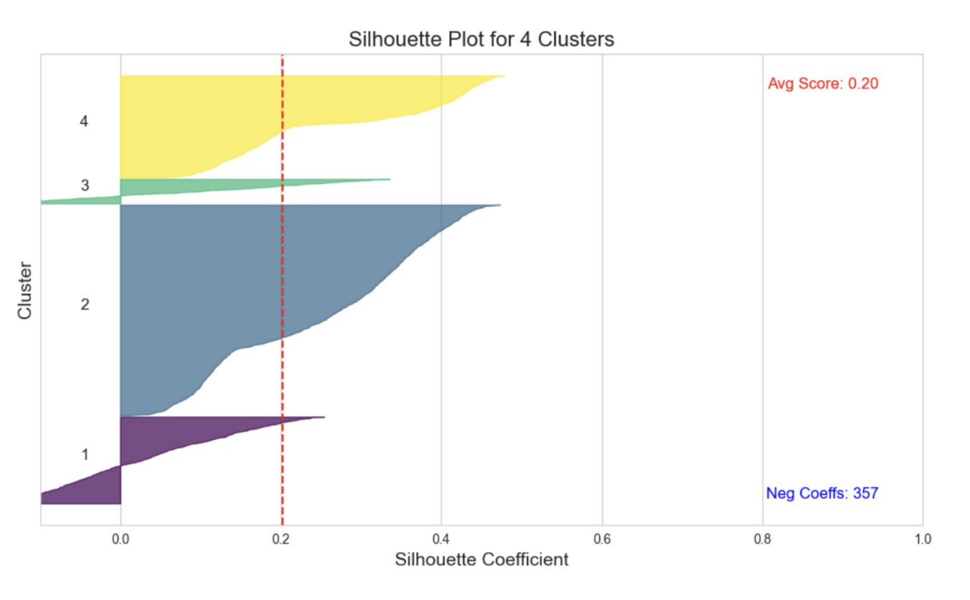


- The Elbow Method shows a noticeable bend at **k = 3**, where inertia sharply drops and **begins to level off**, indicating the most efficient cluster separation with minimal loss of information.

Silhouette Plot







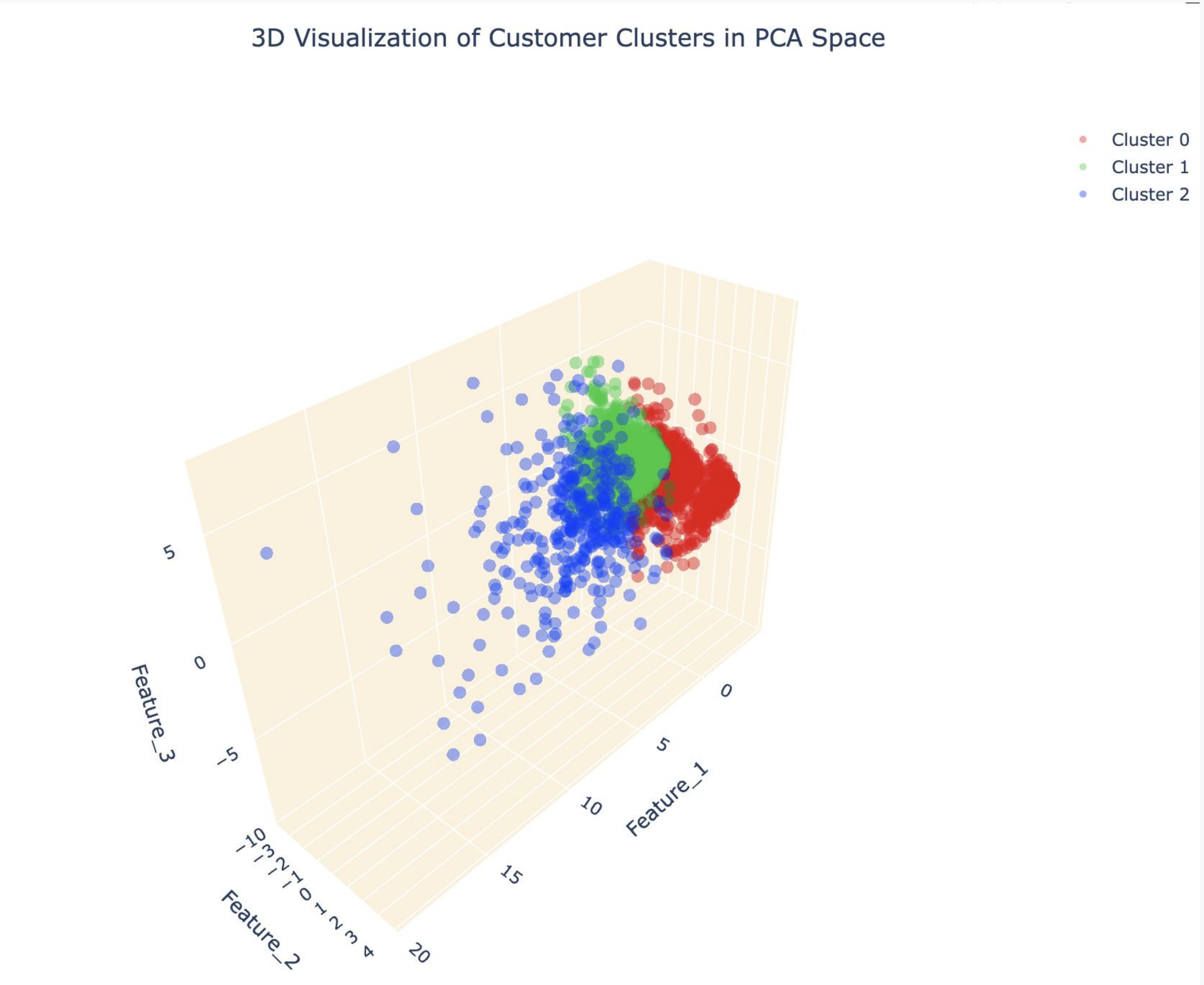
Why K = 3 is Optimal?

- While **K = 2** gives a higher silhouette score (0.36), it has **more negative samples (193)** indicating poor cluster fit.
- **K = 4** significantly increases negative coefficients (**357**), showing high overlap and poor separation.
- **K = 3** provides a **balanced trade-off**:
  - ➢ Fewer poorly clustered samples (**136**)
  - ➢ Distinct and interpretable segments
  - ➢ Aligned with Elbow Method and PCA visualization

Customer Segmentation(k=3)

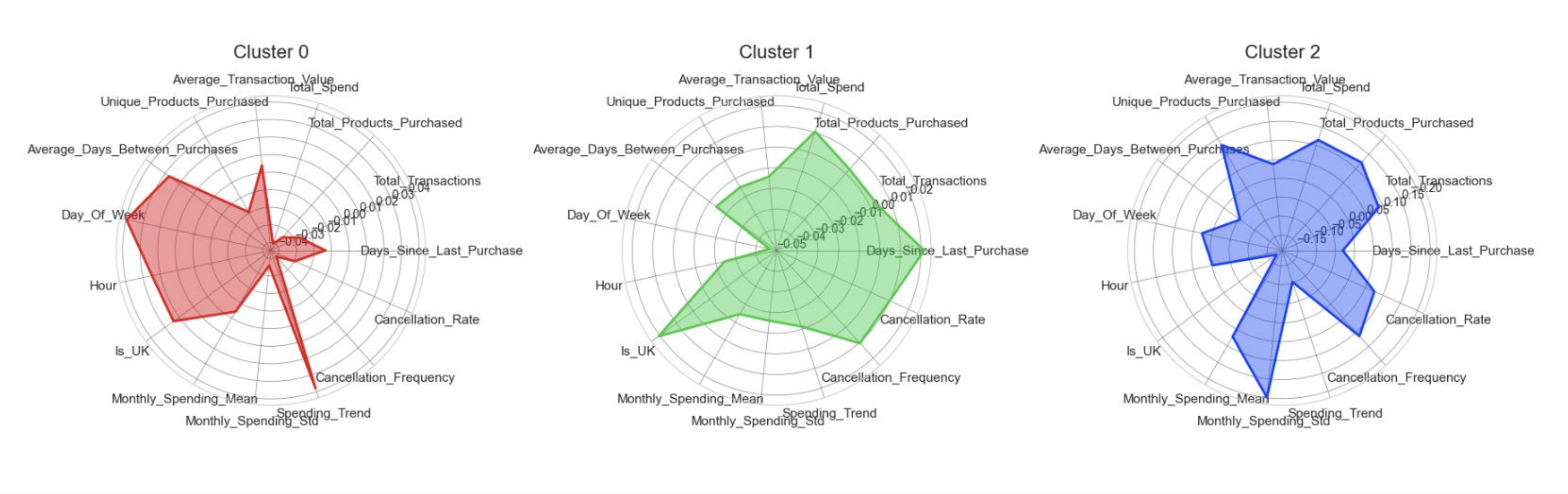
- Using PCA-reduced features, K-Means identified **3 distinct customer groups**.
- Each cluster shows unique behavior patterns that can be used for **targeted marketing and personalized strategies**.

3D Cluster Visualization



- The plot shows 3 customer segments formed using K-Means on PCA components.
 Each color represents a cluster, clearly separating customers based on purchasing patterns.

Customer Segments Profile (Radar Chart)



Cluster 0 :

• Mostly **UK-based customers**

• **High Spending Trend** — consistent increase in monthly purchases

Cluster 1 :

• **High transaction and product volume** but with **irregular purchase intervals**

• Shows **high cancellation rate** — possibly bulk buyers or resellers

Cluster 2

• **Frequent buyers** with stable monthly spending

• Non UK-based customers

Model Evaluation – Clustering Metrics

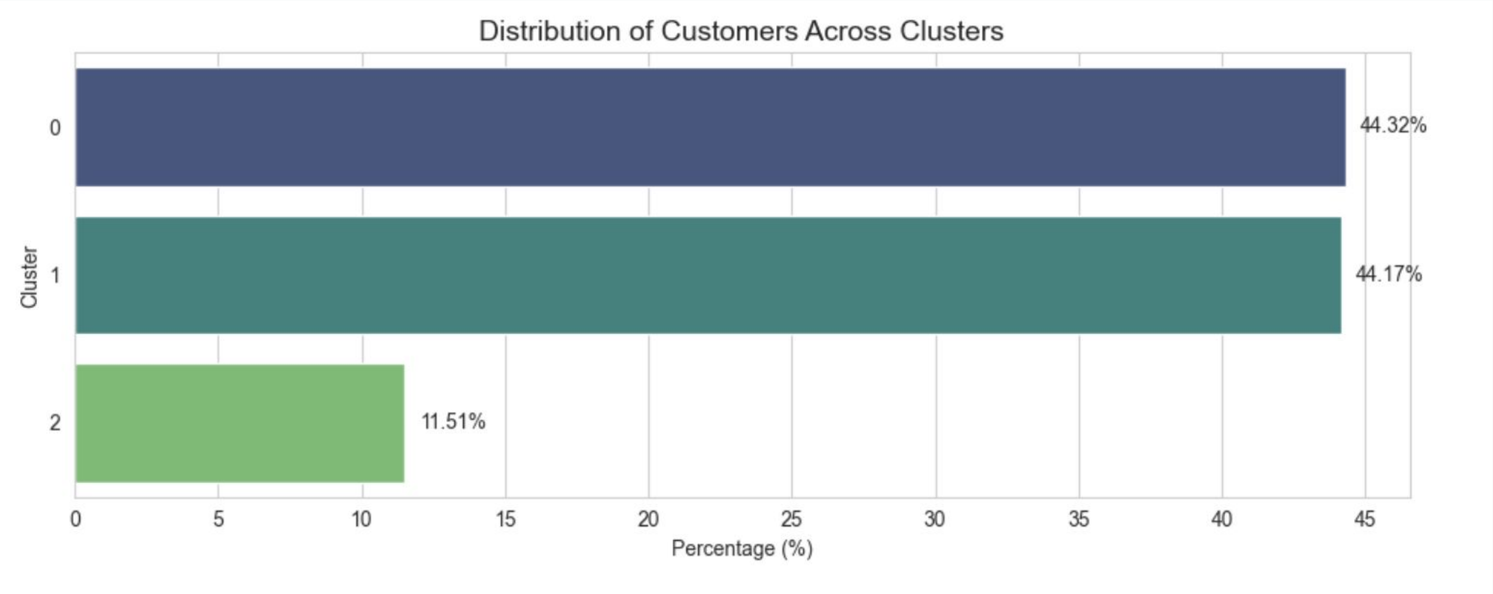
```
from sklearn.metrics import silhouette_score, calinski_harabasz_score, davies_bouldin_score

X = customer_data_clusters.drop(columns=['Cluster', 'CustomerID'])  # Features
labels = customer_data_clusters['Cluster']                          # Cluster labels

print("Silhouette Score:", silhouette_score(X, labels))
print("Calinski-Harabasz Score:", calinski_harabasz_score(X, labels))
print("Davies-Bouldin Score:", davies_bouldin_score(X, labels))
```

✓ 0.1s

```
Silhouette Score: 0.20799285720102292
Calinski-Harabasz Score: 847.7329714895933
Davies-Bouldin Score: 1.515917059970733
```

Customer distribution across clusters

Clustering distribution for train data :



Clustering distribution for test data :

```
customer_data_test_clusters['Cluster'].value_counts()
✓ 0.0s

Cluster
0    448
1    417
2      8
Name: count, dtype: int64
```

Top 5 products & future recommendations by cluster

Cluster	StockCode	Description	Quantity
0.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	6232
0.0	84879	ASSORTED COLOUR BIRD ORNAMENT	4353
0.0	15036	ASSORTED COLOURS SILK FAN	4184
0.0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	3907
0.0	85099B	JUMBO BAG RED RETROSPOT	3581
1.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	5919
1.0	84879	ASSORTED COLOUR BIRD ORNAMENT	5631
1.0	18007	ESSENTIAL BALM 3.5g TIN IN ENVELOPE	5587
1.0	85099B	JUMBO BAG RED RETROSPOT	4946
1.0	85123A	WHITE HANGING HEART T-LIGHT HOLDER	4916
2.0	22616	PACK OF 12 LONDON TISSUES	13641
2.0	85099B	JUMBO BAG RED RETROSPOT	10039
2.0	84879	ASSORTED COLOUR BIRD ORNAMENT	9050
2.0	84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	7782
2.0	22178	VICTORIAN GLASS HANGING T-LIGHT	7757

- Based on historical purchase patterns, we identified the **top 5 frequently purchased products** for each customer cluster.
- These insights can be used to recommend products to new customers assigned to a cluster after their first transaction.
- This enables personalized marketing and targeted product promotion, improving customer experience and sales effectiveness.

Final results/recommendations

CustomerID	Cluster	Rec1_Description	Rec2_Description	Rec3_Description
14299.0	2.0	PACK OF 12 LONDON TISSUES	JUMBO BAG RED RETROSPOT	VICTORIAN GLASS HANGING T-LIGHT
16483.0	1.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ESSENTIAL BALM 3.5g TIN IN ENVELOPE
13740.0	1.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ESSENTIAL BALM 3.5g TIN IN ENVELOPE
17379.0	1.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ESSENTIAL BALM 3.5g TIN IN ENVELOPE
12648.0	0.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ASSORTED COLOURS SILK FAN
16952.0	1.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ESSENTIAL BALM 3.5g TIN IN ENVELOPE
16395.0	0.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ASSORTED COLOURS SILK FAN
15539.0	0.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ASSORTED COLOURS SILK FAN
16842.0	0.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOURS SILK FAN	WHITE HANGING HEART T-LIGHT HOLDER
15866.0	1.0	WORLD WAR 2 GLIDERS ASSTD DESIGNS	ASSORTED COLOUR BIRD ORNAMENT	ESSENTIAL BALM 3.5g TIN IN ENVELOPE

- Using cluster-specific purchase behavior, we derived the **top 3 product recommendations** for each segment.
- These products reflect the **most preferred items** within each cluster and can be offered to new customers shortly after their first purchase.

Conclusion

- Through this customer segmentation project, we successfully **clustered customers** based on behavioral patterns using **K-Means and PCA**.
- The analysis enabled us to uncover **key spending traits**, **seasonal trends**, and **distinct buyer personas**.
- These insights allow the business to deliver **personalized product recommendations** and drive **targeted marketing strategies**.
- Overall, the project supports **data-driven decision-making** that enhances customer engagement and long-term value.

Thank You!

Reimagine your business with data