esade

AI2 Assignment 2

Online Retail Business Case

Darren Darius Tan
MIBA 24/25

Do Good. Do Better.

01

Goals Definition

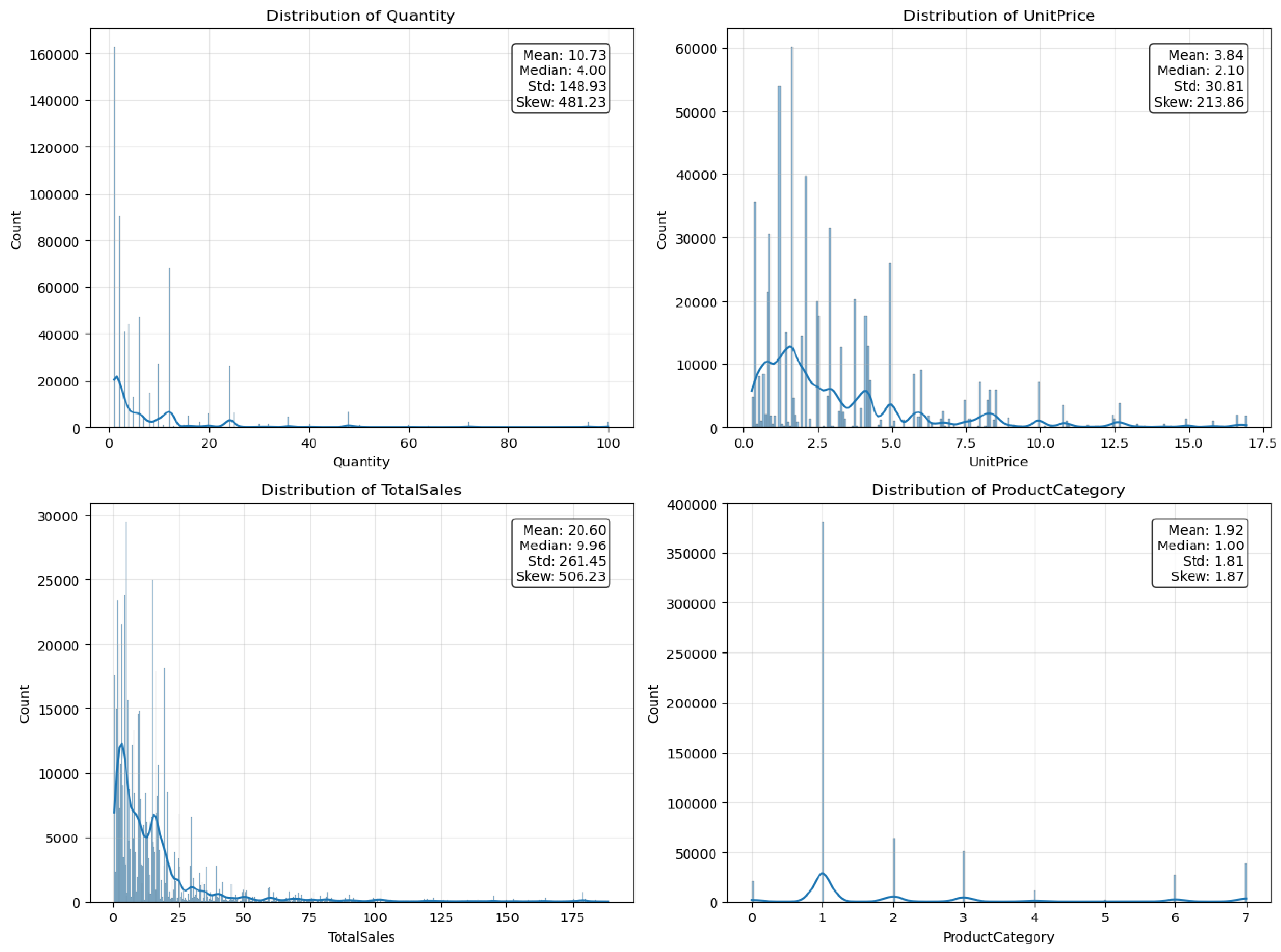
1.1: Initial Exploration of Data

1. **There are 8 features, and 541,909 rows, and can be categorized below:**
   a) Payment (Invoice) Features: InvoiceDate, InvoiceNo
   b) Product Features: StockCode, Description
   c) Customer Features: CustomerID, Country
   d) Financial Features: Quantity, Unit Price

2. **The above groupings give the following insights:**
   a) Payment, Product and Customer Features are categorical features
   b) The only numerical features are Quantity and Unit Price
   c) Amongst the categorical features, there is an order of hierarchy:
   d) Data can be aggregated by unique product, invoice or customer numbers
   e) Data can also be aggregated from a dataframe with unique product OR invoice numbers to one of unique customer numbers
   f) However, the reverse is not true, where a table with unique customers cannot be aggregated to unique invoice or product names.

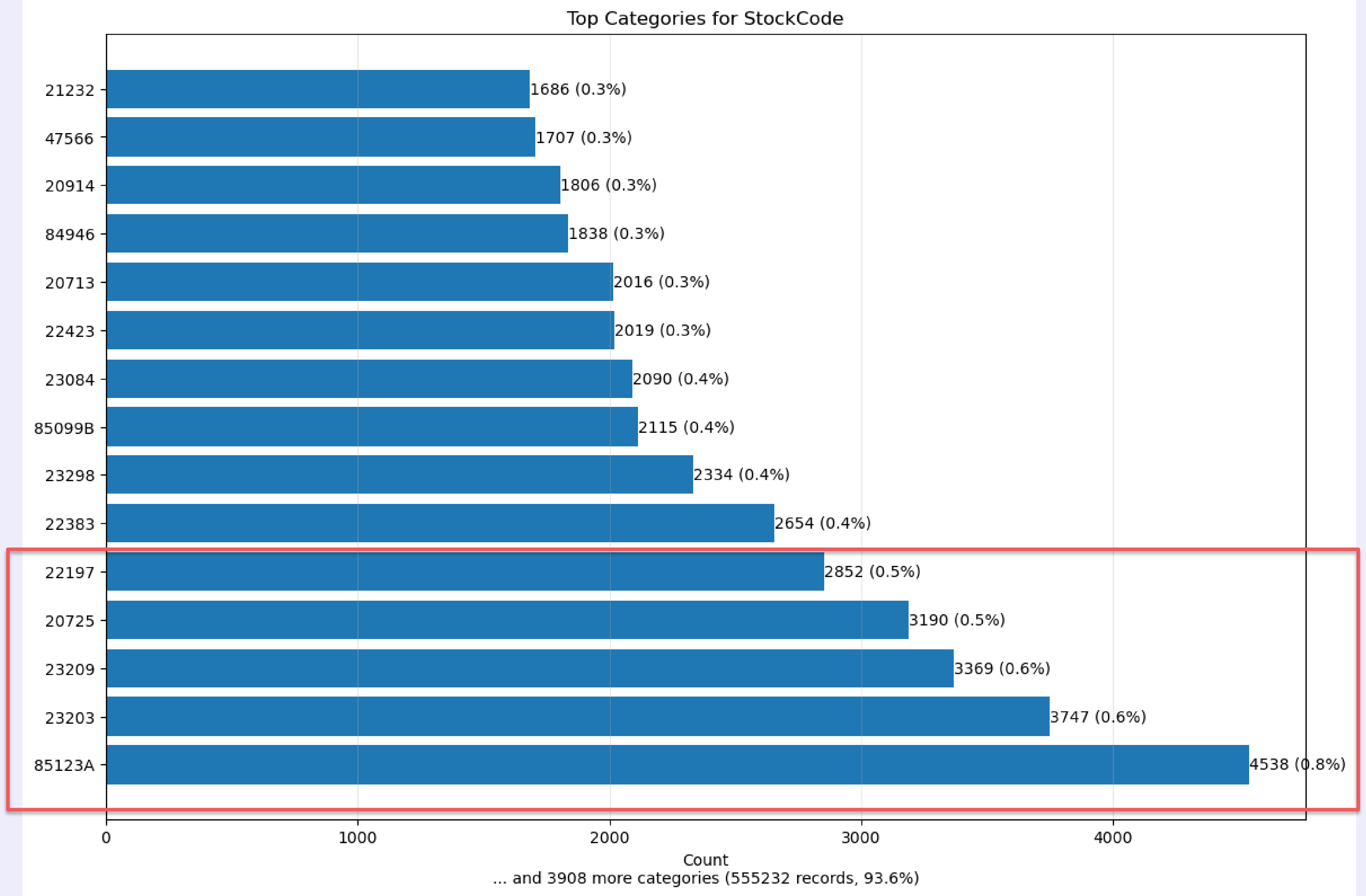
3. **Next slides, we show how the descriptive statistics of each feature in a pre-cleaned dataframe would look like:**

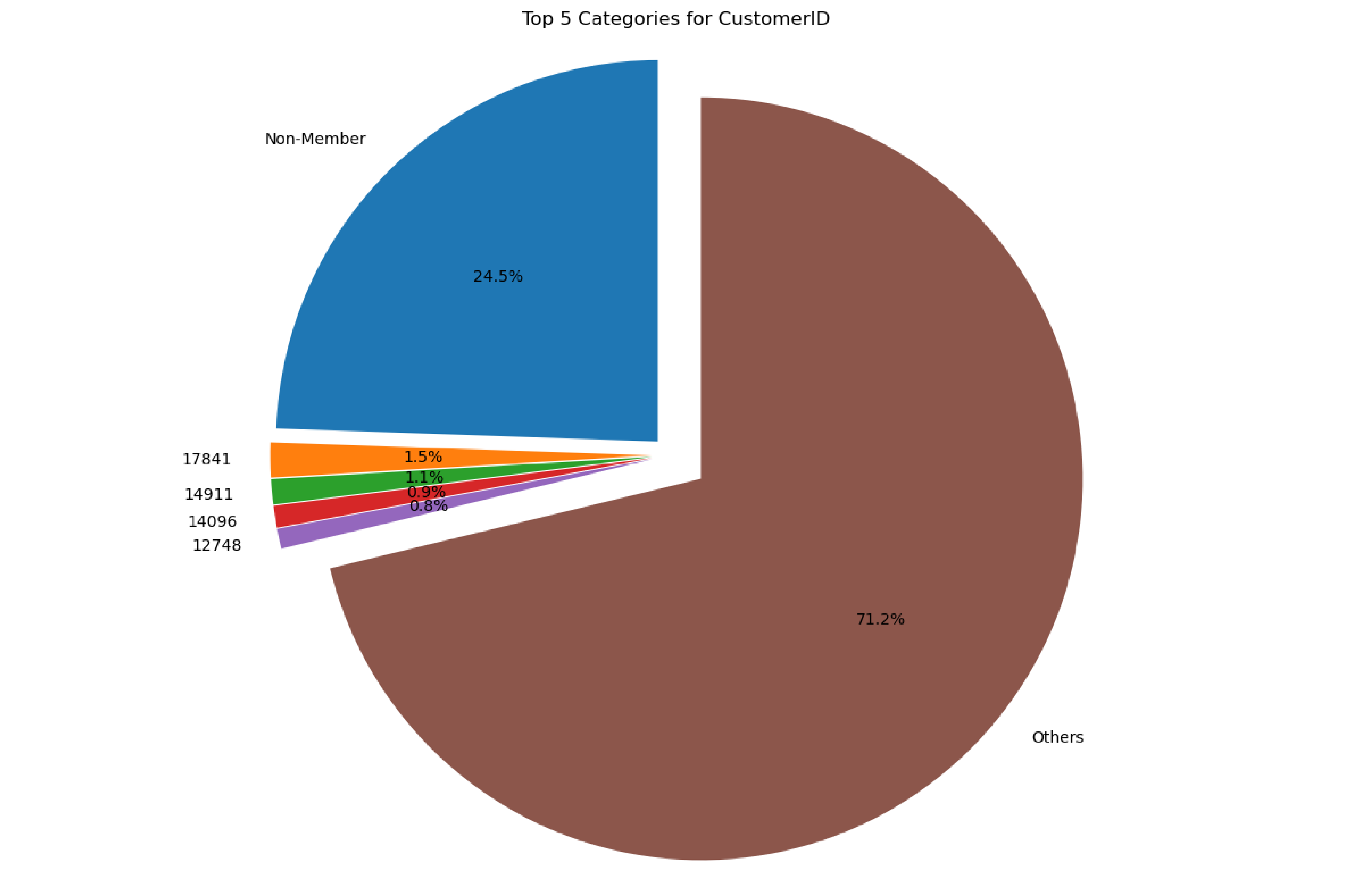
1.1: Initial Exploration of Data (Numeric)



- After removing top and bottom 1% of outliers, the first glance of out descriptive statistics is attained as the starting point to visualize the numerical data

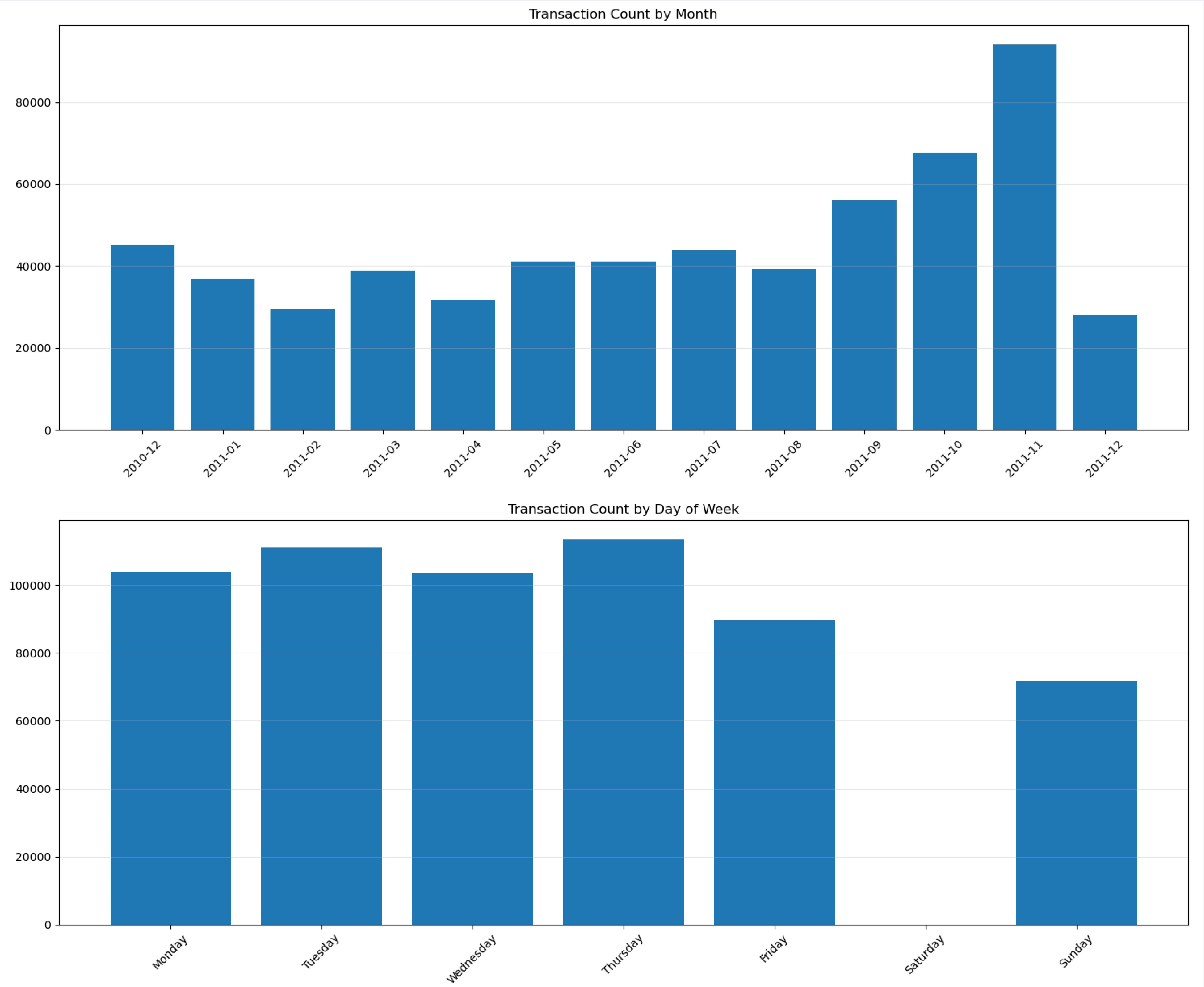
1.1: Initial Exploration of Data (Categorical)





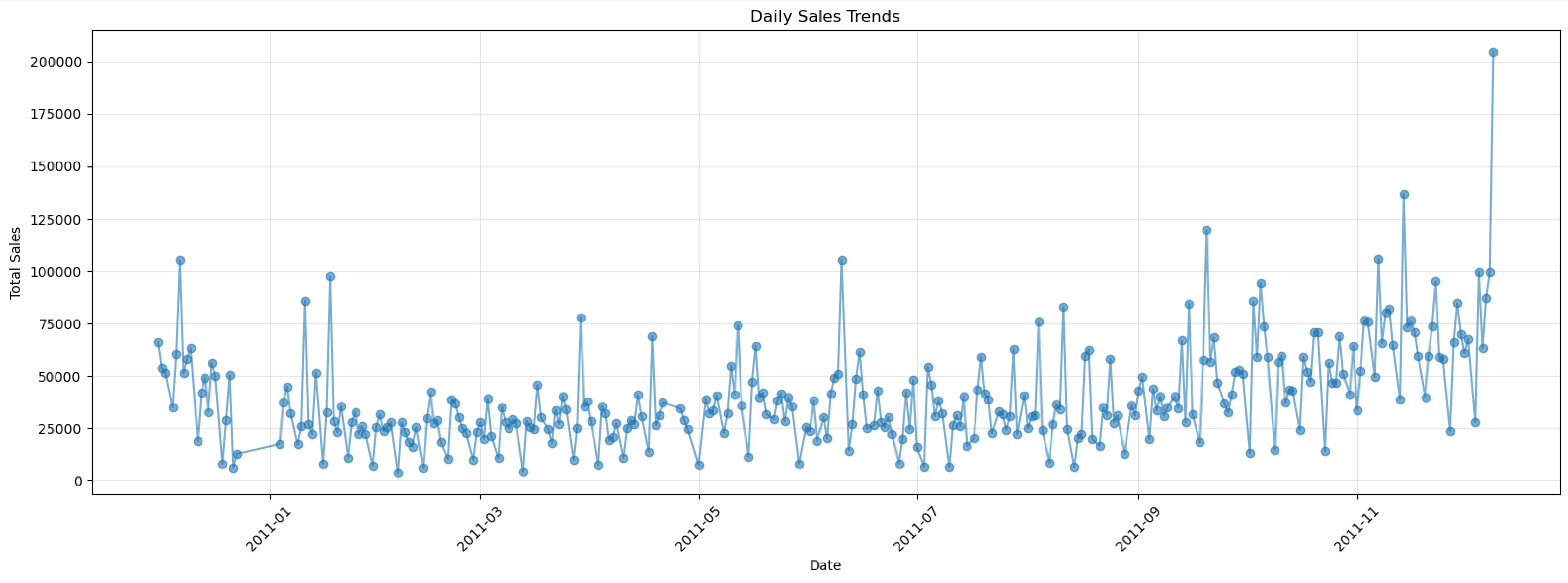
- Identifying “Hero Products” in Marketing → Top 5 StockCodes are visualized above on the left for Marketing Recruitment Activities
- Identifying the Best Customers based on line items spend in total invoices (17841, 14911, 14096, 12748) → Loyalty Rewards program could be considered
- 24.5% of line item spend are from non-members → Potential Marketing Recruitment Communications for eventual profitability

1.1: Initial Exploration of Data (Categorical)



- Over the 13 months of data, November is the best time of the month for sales
- Saturday has 0 transactions despite being an online store

1.1: Initial Exploration of Data (TimeSeries)



- Total Gross Merchandise Value has been stable
- Peaked at end of November 2011

1.2: Defining Goals from Available Data

Problem: All customers receive same promotions and communications

Breakdown	Business Goal 1: Optimize Marketing Strategies	Business Goal 2: Optimize Retention Strategies
Broad Business Objectives	1. Customer Segmentation 2. Targeting of each Segment 3. Positioning of Communications via Product (Purchase Behaviour)	1. Recommending next item to purchase for profitable identified CustomerIDs
Available Useful Data	1. CustomerID 2. Unit Price (by invoice by stockcode) 3. Quantity (by invoice by stockcode) 4. InvoiceDate	
Data to Engineer	1. Aggregation by Unique Products 2. Aggregation by Unique Customer 3. Creation of Features for RFM Analysis using InvoiceDate, CustomerID and their total spend per product (TotalSpend)	1. Clustering of StockCode into Product Categories – to be used for simplified recommendations for business stakeholders 2. Creation of CLV Estimate for identified CustomerIDs
Refined 6 Feasible Objectives **(Revisit in Section 5)**	1. Create a Non-Member base strategy → Recruitment Marketing Strategy of those without CustomerID) 2. Identify Customer Segments via Cluster Methods **[Main Targets]** 3. Identify Anomalous Customer Segments → Create Specific Marketing Strategies for Exceptional Handling (with Heuristics)	For **Main Targets**: For Business Focus 1. Justifying each segment/cluster's CLV through a supervised learning regression method 2. Create Product Recommendations for Hyper-personalized below-the-line marketing communications to specific CustomerIDs 3. Create Product Category Recommendations for Above-the-line marketing communications to Customer Segments

02

Data Preparation

- **Refer to RUN_FIRST file, which gives a df_clean parquet to work with (a clean dataset that fits our business needs)**
- **From there we refer to RUN_SECOND file.**

All important code are then in the RUN_SECOND file.

2.1: Data Cleaning Methods in 1.1

***Note: This cleaning is not final until before we proceed to obtain df_customer_labeled (final usable dataframe)**

S/N	Category of Variables	Variables Affected	Description of Issue	Planned Fix
1	Payment Data	InvoiceNo	• Contains irregular alphanumerics	1. InvoiceDate should be Datetime object 2. InvoiceDate to extract only the Date and remove timestring. 3. Then InvoiceDate check if 1-1 match between InvoiceNo and Date. 4. Last, generate new column for InvoiceDate to map to the number of days since 1st day
2		InvoiceDate	• 43 InvoiceNo has 2 Datetime values • InvoiceNo also contains timestamp (giving rise to 2 datetime values)	
3	Product Data	StockCode		

4		Description	• Missing Descriptions
			• Descriptions that are not Product Descriptions
				1. If description has missing value:
				  1. If corresponding stockcode has >= 1 mapped description, impute by mode
				  2. Else, impute as 'No Description' and remove these
				2. If contains gibberish – remove these

5	Customer Data	Customer ID	• Convert CustomerID from float to object • Missing CustomerIDs	1. Change to object variable 2. Impute missing values as Non-Member (and evaluate if to remove later on)
6		Country		
7	Financial Data	Quantity	• Some values are less than 0, not possible	• Remove all of these rows
8		Price	• Some values are less than 0, not possible	• Remove all of these rows

2.2: Feature Engineering of New Features

S/N	Category of Variables	Variables Created	Purpose / Use	Engineering Method
1	Product Related Data	-	Create product summary dataframe for: 1. Ensuring 1-1 product description mapping to StockCode (as part of further cleaning) 2. Assigning later the StockCodes into Product Categories via Clustering	• Creating df_product_summary that ensures the selection of the first Description that appears for the StockCode, if > 1 available
2		[ProductCategory]	1. Cluster stockcode into product categories for easier Supervised ML later, as we are not encoding all stockcodes	• Perform KMeans with TF-IDF Vectorize to obtain optimal clusters with elbow method
3	Customer Related Data	Aggregations based on available data after pre-cleaning in 2.1	• Obtaining useful data that is by **PER CUSTOMER**, rather than at the product level or invoice level which is not helpful for marketing in general	1. [TotalSpend] = sum of TotalSales 2. [TotalQuantity] = sum of Quantity 3. [FirstPurchaseDate] = min(Day_Number) 4. [LastPurchaseDate] 5. [UniqueProducts] = count of distinct StockCode 6. [FavoriteProduct] = product purchased most often by quantity 7. [FavouriteDescription] 8. PurchaseIntervalAvg = average gap between consecutive purchases
5	Financial Related Data	[Recency]	• Obtaining measurable metrics at the customer level for analysing customer behaviour, justifying predictions and solution recommendations	• = Today’s date – last purchase
6		[Frequency]		• = count of distinct invoices
7		[Monetary]		• = sum of total spend (TotalSpend)
8		[CLV]		• CLV = (avg purchase value) * (purchase frequency per year) * (retention time in years)

2.3.1: Data Transformation (Product Categories)

1. Clustering Products into Product Categories:

a) To optimize marketing communications at the end to various customer segments:

   i. Customer Segments have to be identified via Cluster Analysis

   ii. Selected Products would have to be included in the marketing communications to each customer segment

b) Issue:

   i. However, every product would have a different set of association rules:

   ii. It may be challenging for businesses to hyper-personalize their Below-The-Line messages to target every customer and recommending ‘consequent’ products to buy based on their already bought ‘antecedent’ products

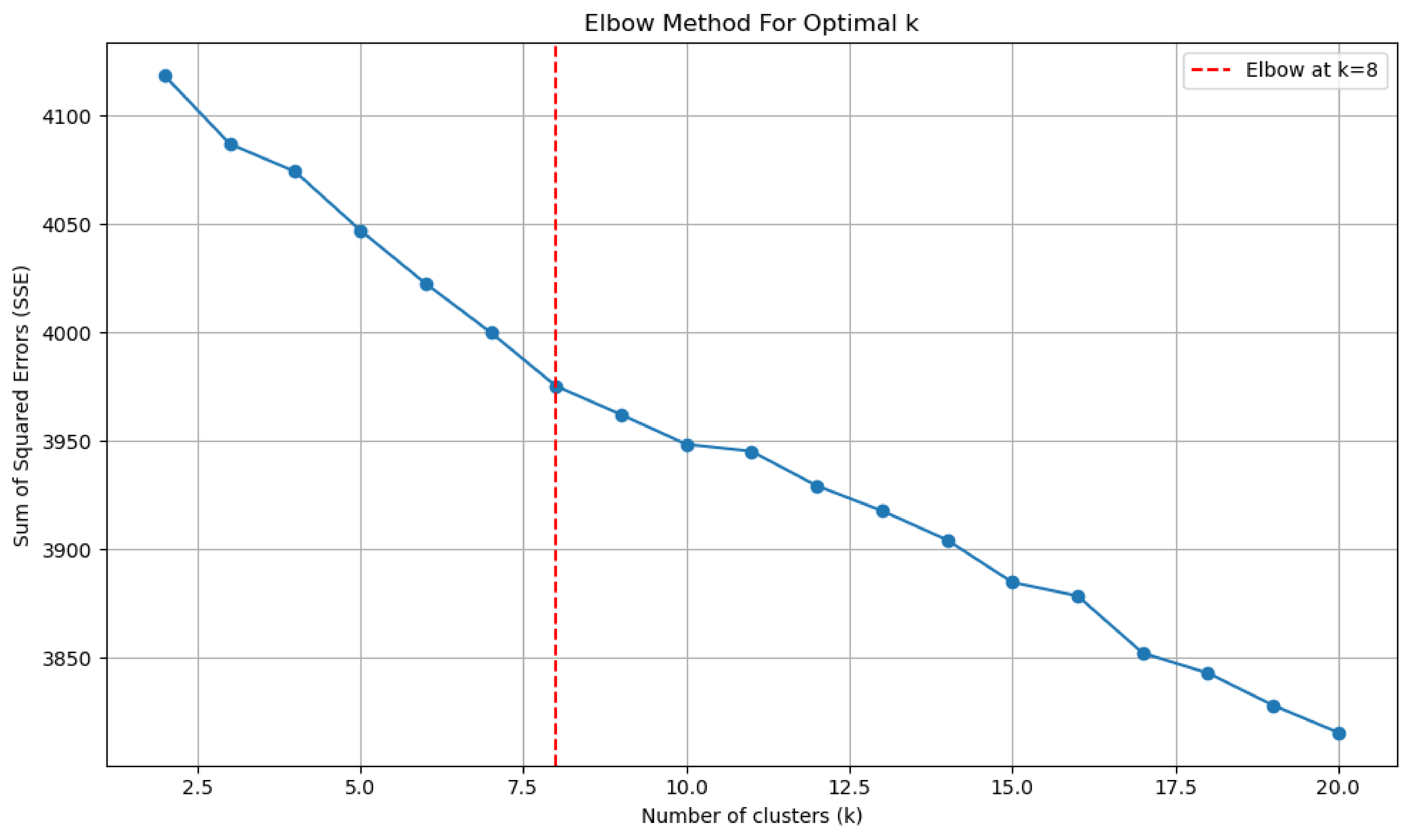
c) For implementation simplicity as a consideration:

   a) It may be easier to see which are the categories identified from an antecedent set of products which are then recommended to the consequent set of products.

   b) From business perspective, above-the-line communications to target customers and target set of products (as 1 category) would also help.

   c) In product carousels in ecommerce, product categorization by clustering may also help in Online Store Visual Merchandising to help customers understand which products are grouped together

2. K-Means Clustering with TF-IDF Vectorizer is deployed:



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Day_Number	TotalSales	ProductCategory
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01	2.55	17850	United Kingdom	1	15.30	3
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01	2.55	17850	United Kingdom	1	15.30	3
2	536365	71053	WHITE METAL LANTERN	6	2010-12-01	3.39	17850	United Kingdom	1	20.34	1
3	536365	71053	WHITE METAL LANTERN	6	2010-12-01	3.39	17850	United Kingdom	1	20.34	1
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01	2.75	17850	United Kingdom	1	22.00	1

2.3.2: Data Transformation (Feature Scaling)

1. Feature Scaling is required for Clustering Algorithms:

a) Distance-based methods (like K-Means, Hierarchical) are sensitive to feature magnitudes — unscaled features dominate the distance calculations

b) Scaling ensures equal contribution from all features

2. Features to scale from df_customer_all:

a) Feature engineered Numeric features:

- a) 'Recency', 'Frequency', 'TotalSpend', 'TotalQuantity',
- b) 'UniqueProducts', 'NumberOfInvoices',
- c) 'PurchaseIntervalAvg', 'AvgPurchaseValue',
- d) 'DaysActive', 'AvgPurchaseFrequency', 'PurchasesPerYear', 'CLVEstimate'

b) Skipped Categorical Features:

- a) Only 'Country' is the original column from df that is meaningful to scale for clustering later on
- b) We skip the one-hot-encoding of 'Countries' which by right should be done to ensure that distances contribute to the cluster, because most of the values of 'Countries' come from the UK.
- c) Even if we one-hot encode them, it causes the country columns to make the dataframe large, and not ideal due to the signal-to-noise ratio.

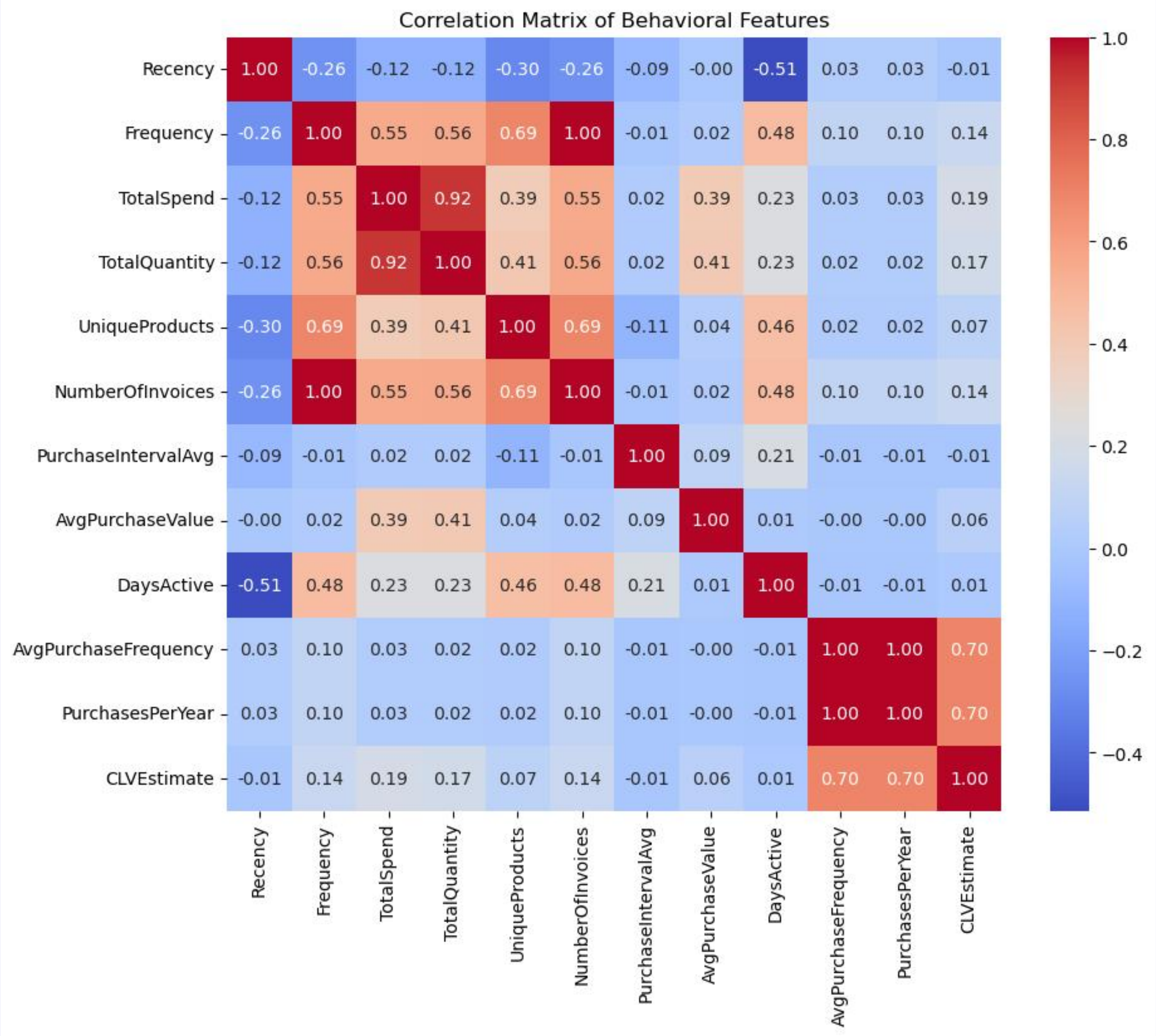
```python
from sklearn.preprocessing import StandardScaler

# Select relevant behavioral numerical features
features_to_scale = [
    'Recency', 'Frequency', 'TotalSpend', 'TotalQuantity',
    'UniqueProducts', 'NumberOfInvoices',
    'PurchaseIntervalAvg', 'AvgPurchaseValue',
    'DaysActive', 'AvgPurchaseFrequency', 'PurchasesPerYear', 'CLVEstimate'
]

# Standardize
scaler = StandardScaler()
scaled_values = scaler.fit_transform(df_customer_all[features_to_scale])

# Create new DataFrame
df_customer_scaled = df_customer_all.copy()
df_customer_scaled[features_to_scale] = scaled_values

df_customer_scaled.head()
```

0.0s

	CustomerID	Recency	Frequency	TotalSpend	TotalQuantity	FirstPurchaseDate	LastPurchaseDate	UniqueProducts	NumberOfInvoices
0	12346	2.329673	-0.424675	8.359634	14.445201	2011-01-18	2011-01-18	-0.708687	-0.424675
1	12347	-0.900449	0.354080	0.251046	0.250006	2010-12-07	2011-12-07	0.486336	0.354080
2	12348	-0.170421	-0.035297	-0.028546	0.226861	2010-12-16	2011-09-25	-0.462653	-0.035297
3	12349	-0.740443	-0.424675	-0.032963	-0.111417	2011-11-21	2011-11-21	0.134858	-0.424675
4	12350	2.179667	-0.424675	-0.191315	-0.197272	2011-02-02	2011-02-02	-0.521232	-0.424675

2.4: Correlation Matrix for Independent Features

1. **Correlation Matrix Visualization to the right:**
2. **To select independent features for clustering to work well:**
   a) Remove highly correlated features (as they duplicate information and distort distance metrics → leading to biased clusters
   b) Remove conceptually similar variables due to arithmetic derivations → resultant clusters have higher dimensions (generally disadvantageous) and lead to interpretability issues for business stakeholders.
3. **Final Selection of features for clustering:**
   a) 'Recency'
   b) 'Frequency'
   c) 'TotalSpend'
   d) 'UniqueProducts'
   e) 'PurchaseIntervalAvg'
   f) 'DaysActive'



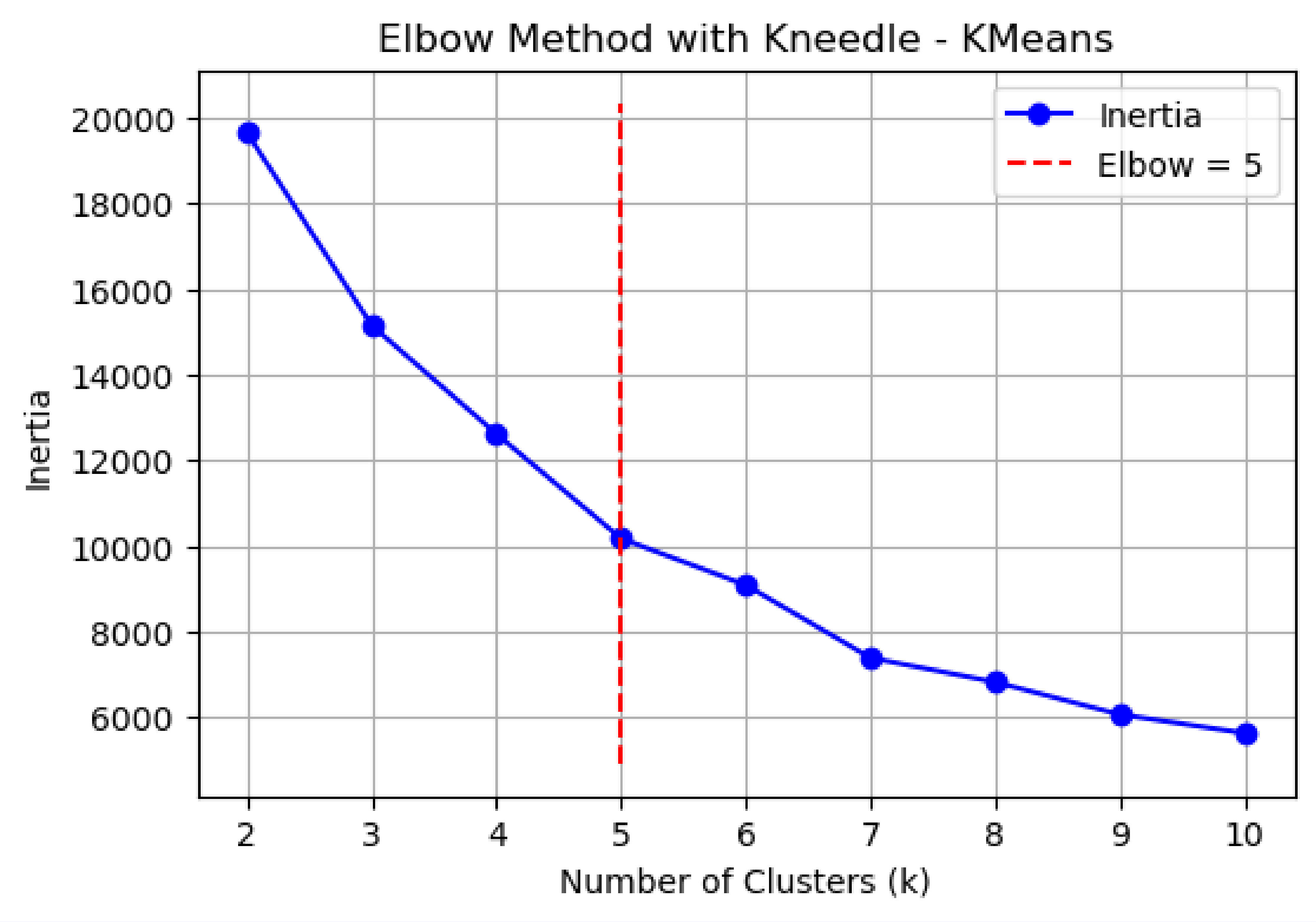
03

Cluster Model Creation

- To cluster customers into customer segments for marketing communications targeting primarily, Kmeans Clustering and Hierarchical Clustering are chosen for the clustering algorithm as required by the assignment.
- DBScan is the least preferred because (i) it performs the worst on large datasets, (2) it is very sensitive to outliers like what is seen during data exploration in earlier slides.

3.1: K-Means Clustering (First Choice)

- Using silhouette score as the evaluation metric comparing Kmeans and Hierarchical Clustering, silhouette score (kmeans) == 0.383 while silhouette score (hierarchical) == 0.253.
- Hence we fit predict the Kmeans model back to the df_customer_labeled to assign clusters optimally to each CustomerID.
- In Kmeans here, the best number of clusters is obtained with the elbow method automatically using kneedle → Optimal clusters == 5




```
# Analyze cluster characteristics (mean per feature)
cluster_kmeans = df_customer_labeled.groupby('Cluster')[features_for_clustering].mean().round(2)
cluster_kmeans
```

✓ 0.0s

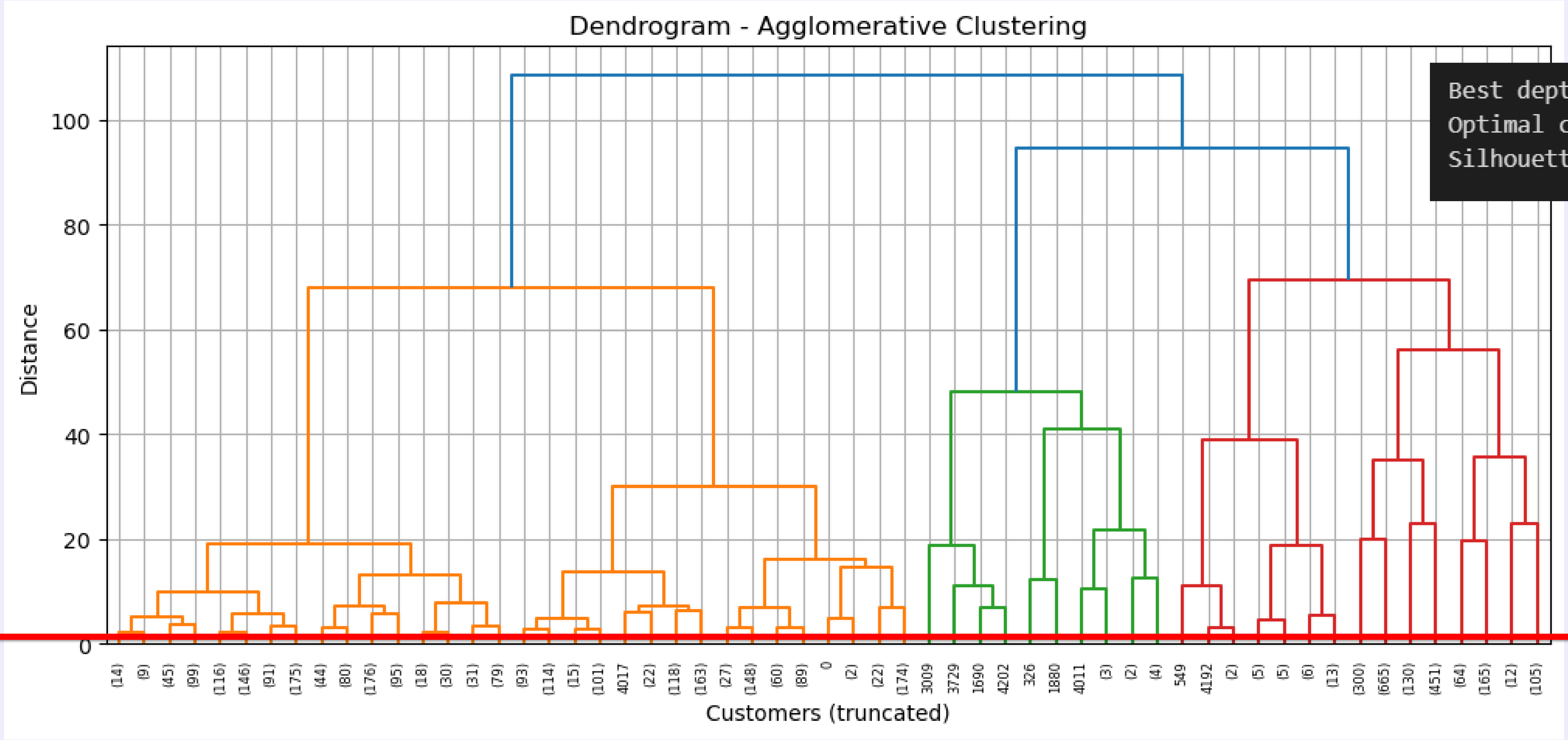
Cluster	Recency	Frequency	TotalSpend	UniqueProducts	PurchaseIntervalAvg	DaysActive
0	60.90	2.70	1137.34	2.05	126.02	250.05
1	58.18	2.08	742.31	36.91	2.29	59.01
2	7.00	75.24	108781.56	700.12	6.39	337.06
3	30.45	7.95	3379.47	109.57	4.15	288.20
4	258.34	1.49	620.73	23.35	0.98	20.34

```
# compute silhouette score for kmeans to compare with hierarchical clustering later
silhouette_kmeans = silhouette_score(X, labels)
print("Silhouette Score for KMeans:", silhouette_kmeans)
```

✓ 0.2s

```
Silhouette Score for KMeans: 0.3837168660216574
```

3.2: Hierarchical Clustering (Second Choice)



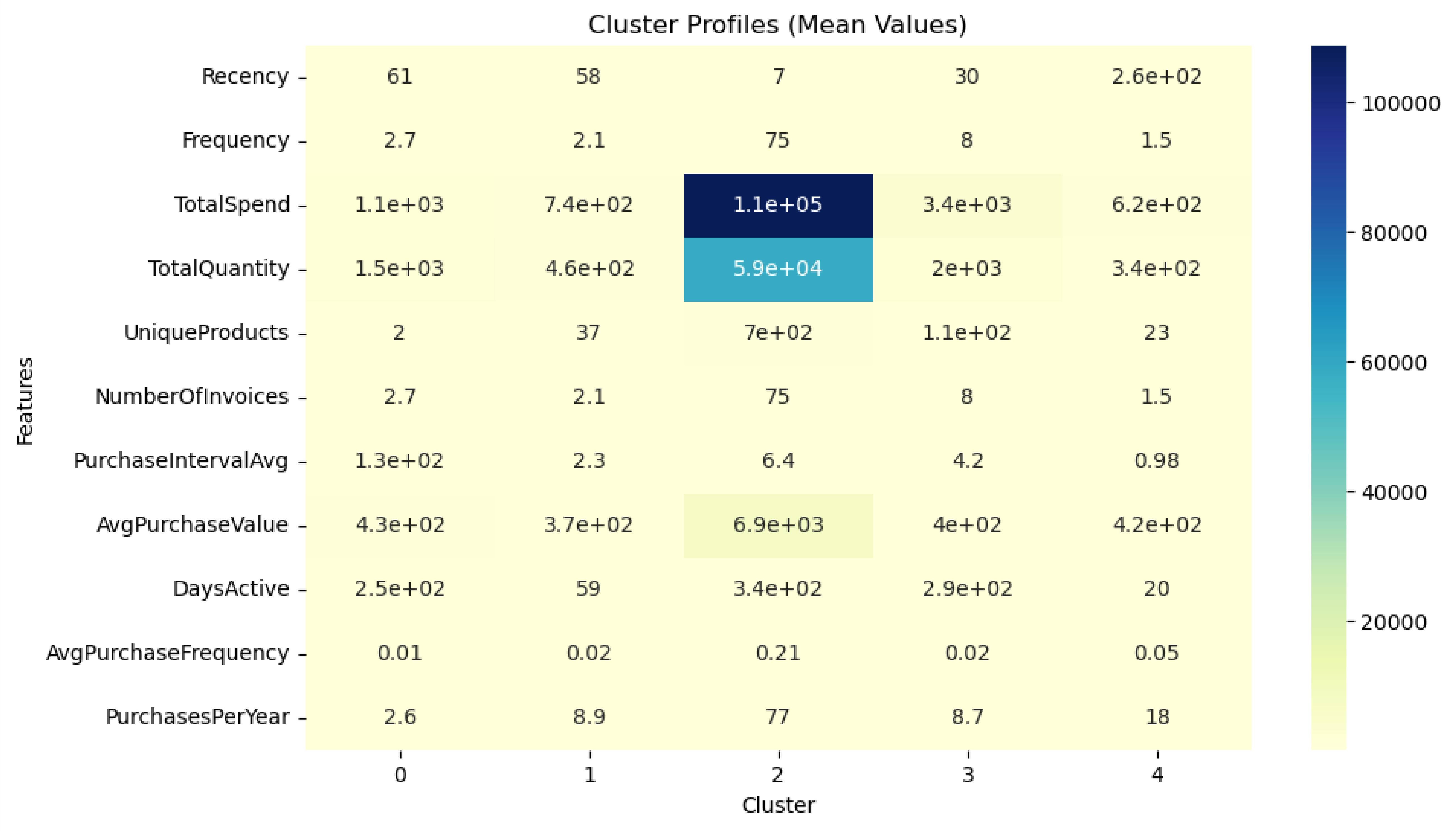

```
Best depth: 3
Optimal clusters: 1676
Silhouette score: 0.253
```

- Applied Agglomerative Clustering
- Best depth is bottom up and we have 1676 clusters → too much and will reject

04

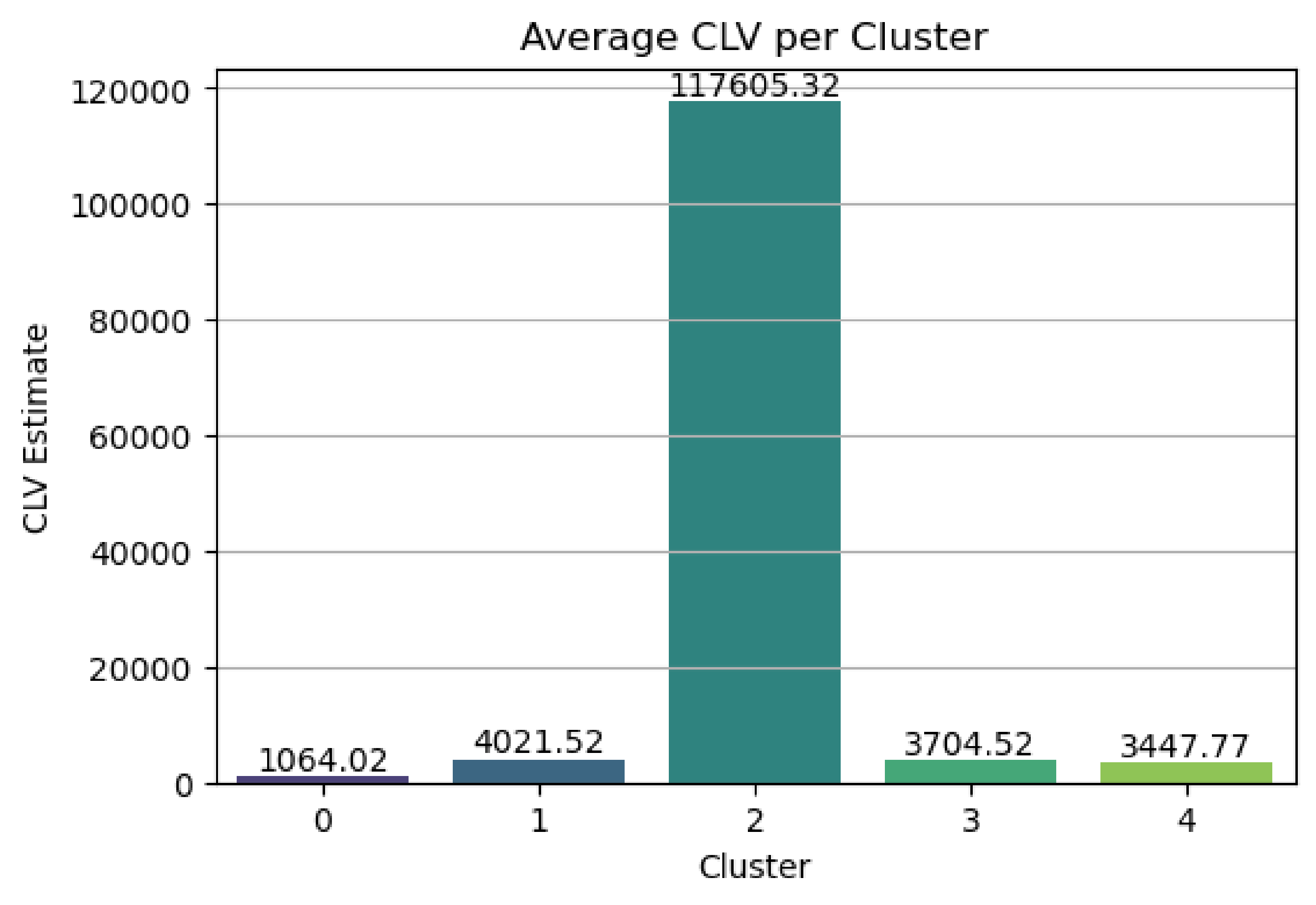
Model Interpretation

4.1: Customer Segmentation Insights



Category Spending by Cluster:

Cluster	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	0.00	998.47	39.90	82.62	0.46	0.00	0.00	74.97
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
2	3082.14	79953.23	12379.31	12055.01	1126.56	7.83	3785.70	10866.18
3	104.89	2528.94	403.95	370.22	66.25	2.56	145.98	290.16
4	15.68	531.10	56.03	70.53	8.50	0.69	21.39	31.44



Cluster Sizes and CLV:

Cluster	Count	Percentage	Avg CLV
0	20	0.5	1064.02
1	1831	42.2	4021.52
2	17	0.4	117605.32
3	1503	34.6	3704.52
4	968	22.3	3447.77

4.2: Category Spend Per Customer

df_customer_category
✓ 0.0s

	CustomerID	ProductCategory	Spend
0	12346	1	77183.60
1	12347	0	80.40
2	12347	1	3247.87
3	12347	2	469.15
4	12347	3	277.98
...	...	...	...
21174	Non-Member	3	131327.84
21175	Non-Member	4	44002.42
21176	Non-Member	5	5002.23
21177	Non-Member	6	71143.31
21178	Non-Member	7	133444.20

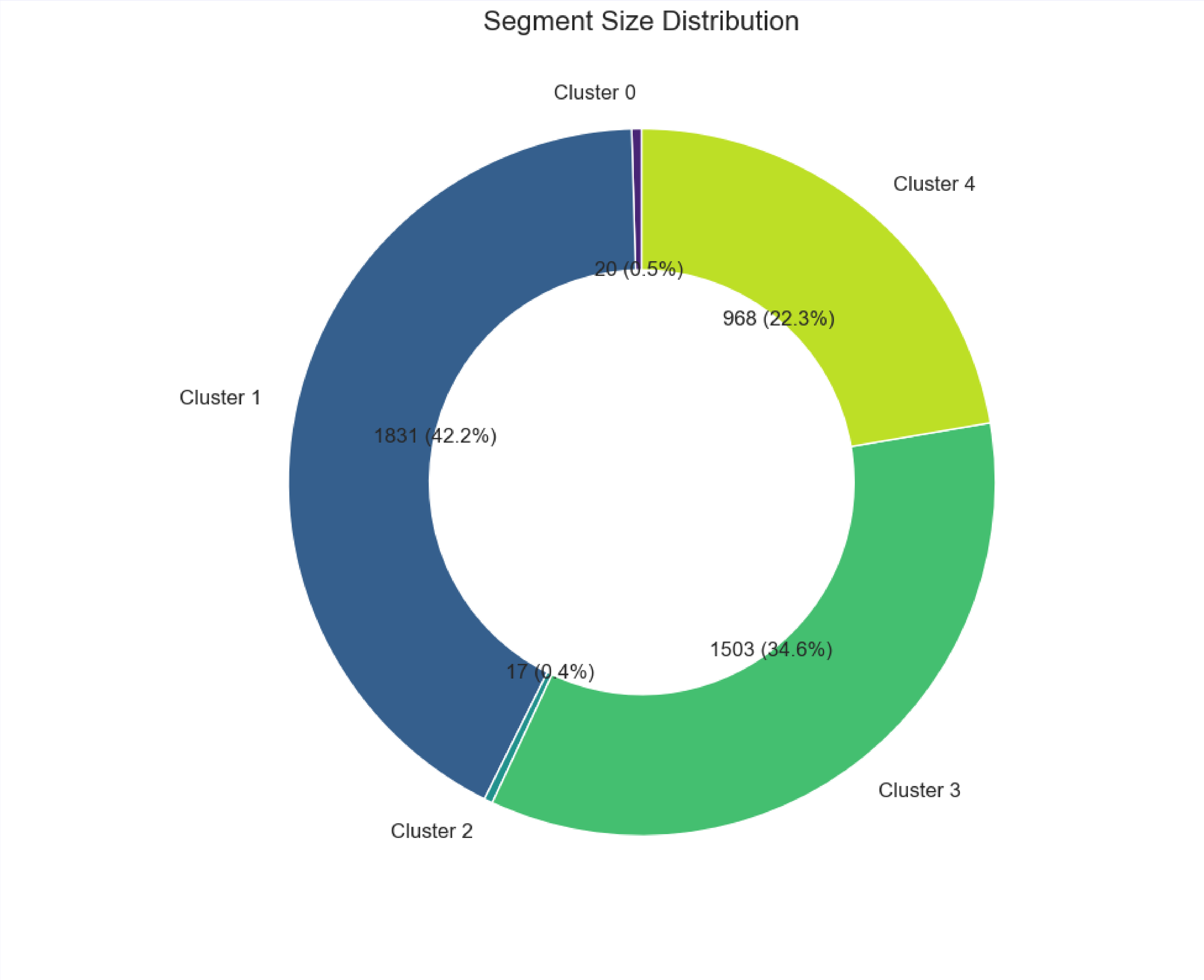
21179 rows × 3 columns

	CustomerID	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	12346	0.00	77183.60	0.00	0.00	0.00	0.00	0.00	0.00
1	12347	80.40	3247.87	469.15	277.98	26.56	0.00	278.00	226.10
2	12348	41.76	1522.72	191.00	0.00	0.00	0.00	63.60	41.76
3	12349	17.85	1209.85	399.02	139.65	0.00	0.00	19.80	256.60
4	12350	40.20	271.10	0.00	0.00	0.00	0.00	17.70	20.40
...	...	...	...	...	...	...	...	...	...
4335	18281	0.00	79.92	33.90	0.00	0.00	0.00	0.00	0.00
4336	18282	0.00	171.91	6.64	0.00	0.00	0.00	13.00	0.00
4337	18283	40.68	1920.38	197.13	104.25	0.00	2.55	104.96	151.22
4338	18287	30.30	673.12	535.24	62.64	380.88	0.00	129.30	40.80
4339	Non-Member	49555.81	1353842.56	156222.36	131327.84	44002.42	5002.23	71143.31	133444.20

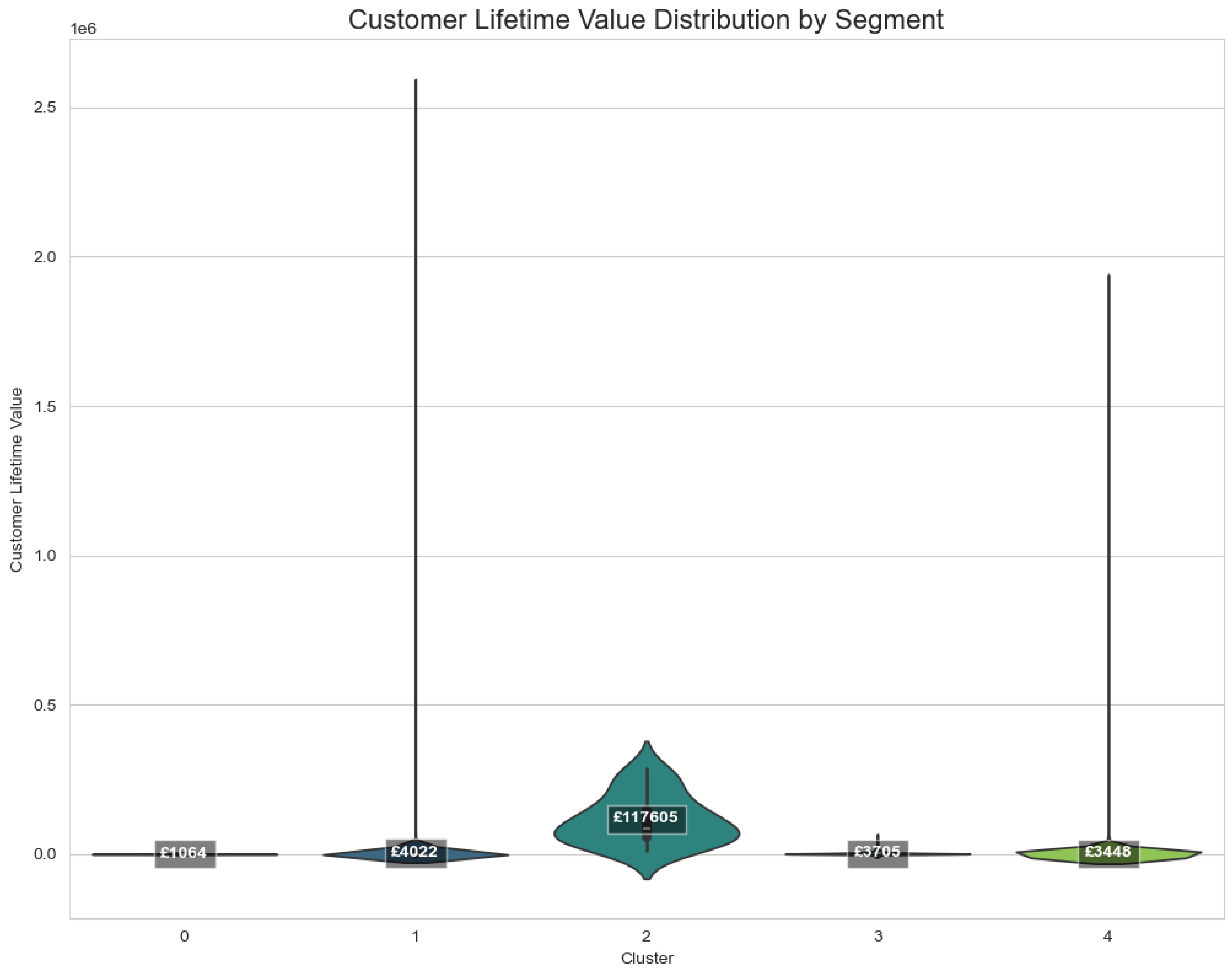
4340 rows × 9 columns

```
# Get all category spending columns
cat_spend_cols = [col for col in df.columns if
col.startswith('CatSpend_')]
```

4.3: Detecting Anomalous Clusters



4.4: Customer Segments before Removing Anomalies



4.5: Customer Segments after Removing Anomalies

Cluster	0	1	2	3	4	5	6	7
0	0.00	548.90	57.00	118.03	0.00	0.00	0.00	107.10
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
3	94.84	2265.26	366.49	312.58	60.05	2.56	125.29	246.05
4	15.71	483.78	55.81	49.42	8.52	0.69	21.44	31.29

05

Model Implementation For Business Insights

For Parts 04, we refer only to AI2_Assignment1_open_loans.ipynb exclusively

5.1: Applying Association Rules to Recommend Products

1. Create df_basket:

a) Converted all quantities in df_customer_product_matrix that are > 1 to become == 1, else 0 (binary coding)

b) Extracted and merged CustomerID and Cluster onto df_basket

2. Create cluster_basket per cluster for recommending products based on certain parameters:

Cluster #	Item_counts parameter	Max_len parameter	Reason
0	1	2	Max_len == 1 returns no recommendations
1	1	2	
2	5	2	Item_counts more than certain arbitrary threshold to improve runtime
3	20	2	
4	1	2	Following the default standard of Cluster 0 and 1

Recommendation: For Below-The-Line Messaging to each cluster, business team can hyper-personalize message to each customerID with the recommended products at the customer level, as every customerID has a different set of consequent products for recommendation

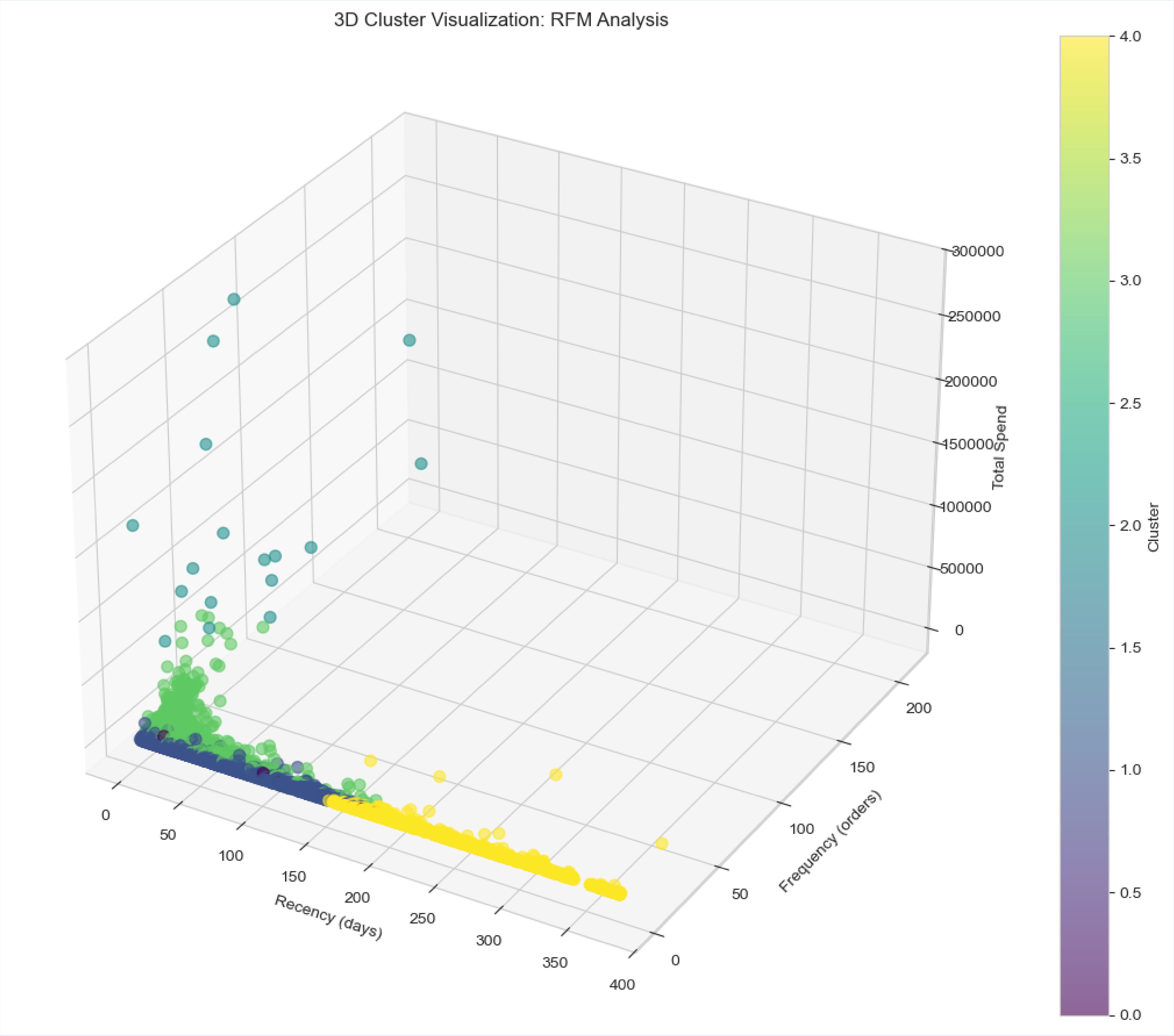
3. Next in Section 5.2, the following is performed to obtain recommended insights for each cluster:

a) Merge descriptions to recommendations of StockCode per cluster (with build_recommendation_df)

b) Identifying antecedent and consequent product categories per cluster (with build_category_recommendation_df)

c) Making a matrix visualization for business level stakeholders to see on an approximate level per cluster, the likelihood of product recommendations messaging to send to each cluster above-the-line

5.2: Customer Segment Profiling and Positioning Strategy



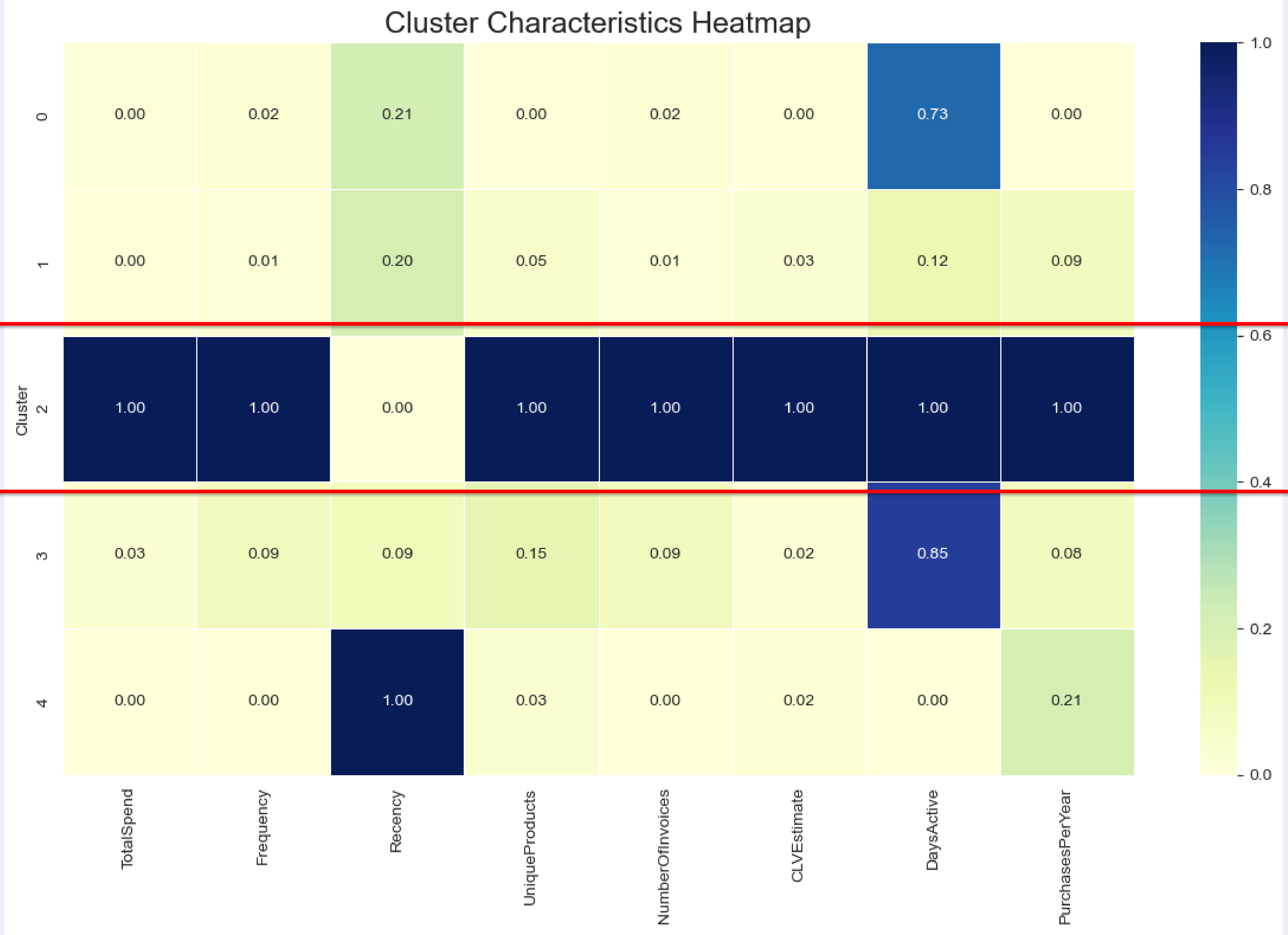
Customer Distribution Across Clusters:

Cluster	Customer_Count	Customer_Percentage
0	20	0.46
1	1831	42.20
2	17	0.39
3	1503	34.64
4	968	22.31

From 3D cluster chart, we can rank the 5 segments by heuristics (we get specific evaluations in next few slides):

1. **Cluster 2: Champions**
2. **Cluster 3: Potential Loyalists (Better)**
3. **Cluster 0: Potential Loyalists (Worse)**
4. **Cluster 1: Low-value Customers**
5. **Custer 4: Lost Customers**

5.2: Positioning Strategy to Targeted Customer Segments



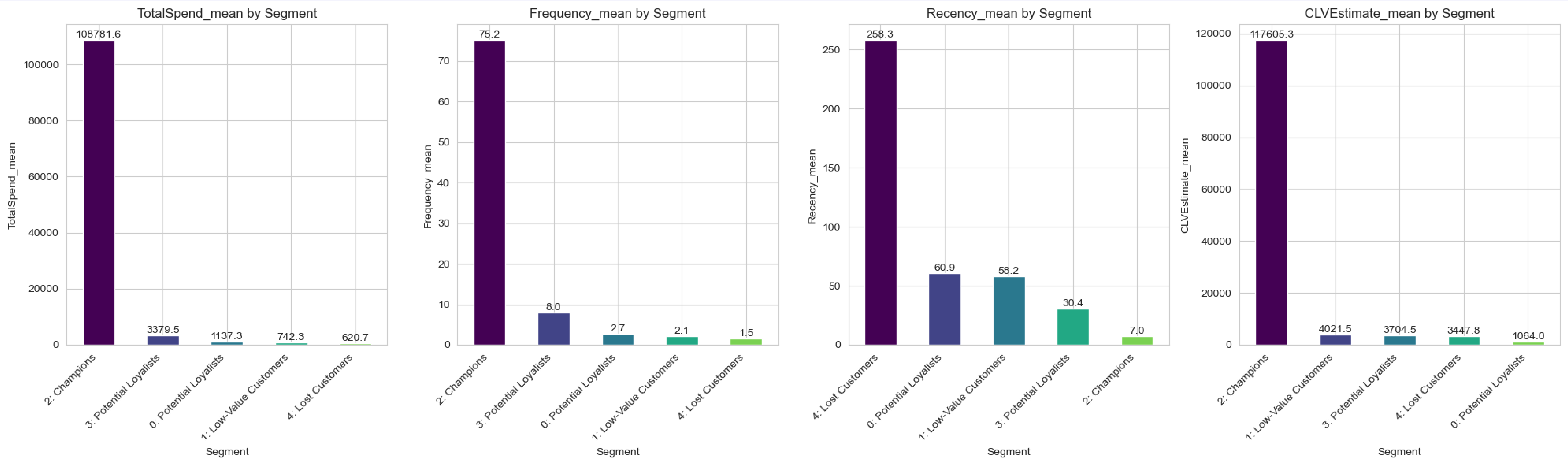
Visualizing the best cluster and its associated 17 CustomerIDs who deserved to be rewarded the most (if a Loyalty Program exists)

```
Cluster 2 - Unknown (Total: 17 customers)
Sample CustomerIDs:
```

	CustomerID
55	12415
326	12748
562	13089
997	13694
1290	14096
1334	14156
1435	14298
1662	14606
1690	14646
1880	14911
2177	15311
2703	16029
3009	16446
3729	17450
3772	17511
4011	17841
4202	18102

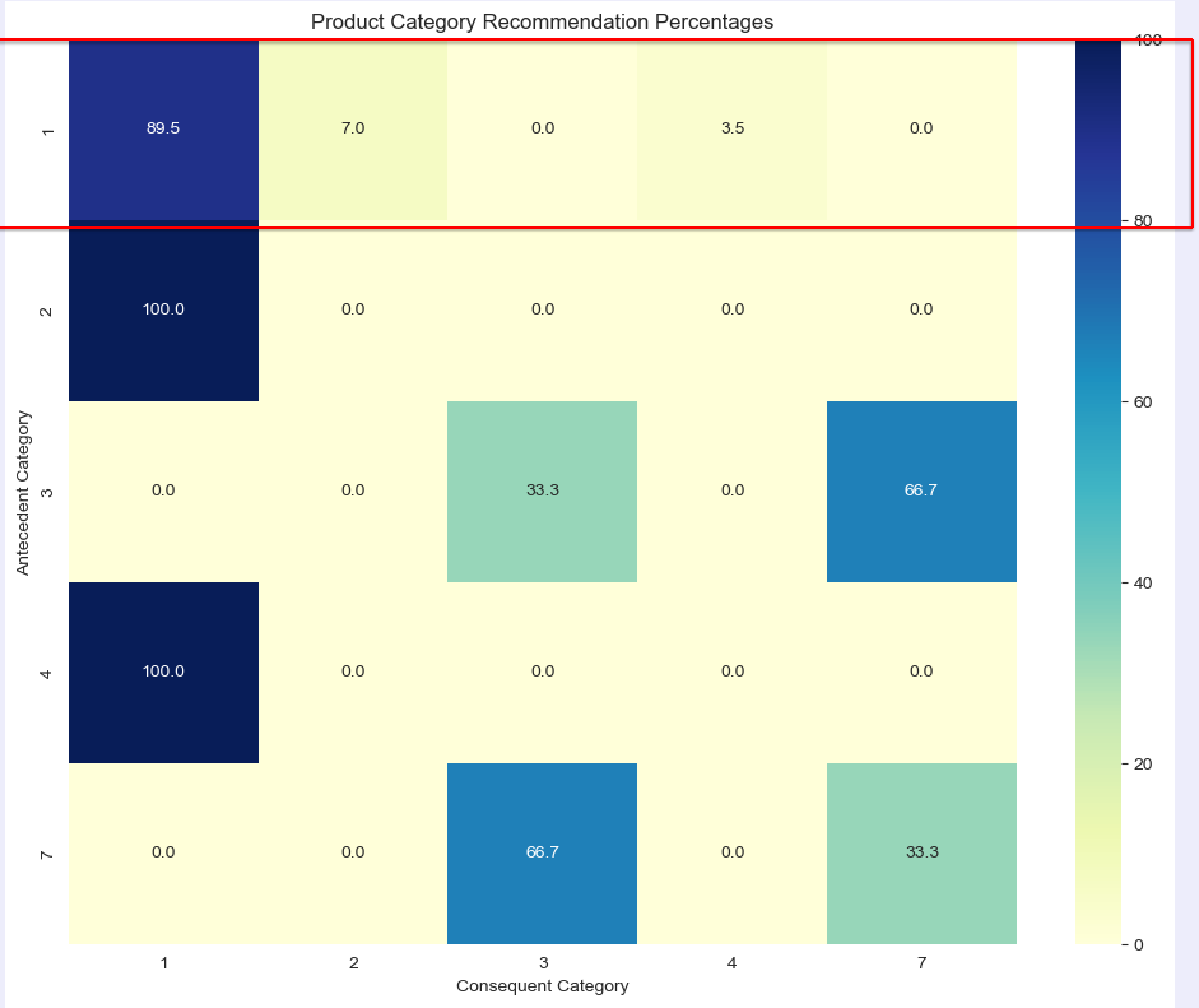
5.2: Positioning Strategy to Targeted Customer Segments

Visualizing the RFM metrics and estimated Customer Lifetime Value for an idea of Return of Investment (ROI) when budget is used onto marketing communications for each segment:



esade

5.3: Product Recommendation Targeting Customer Segments



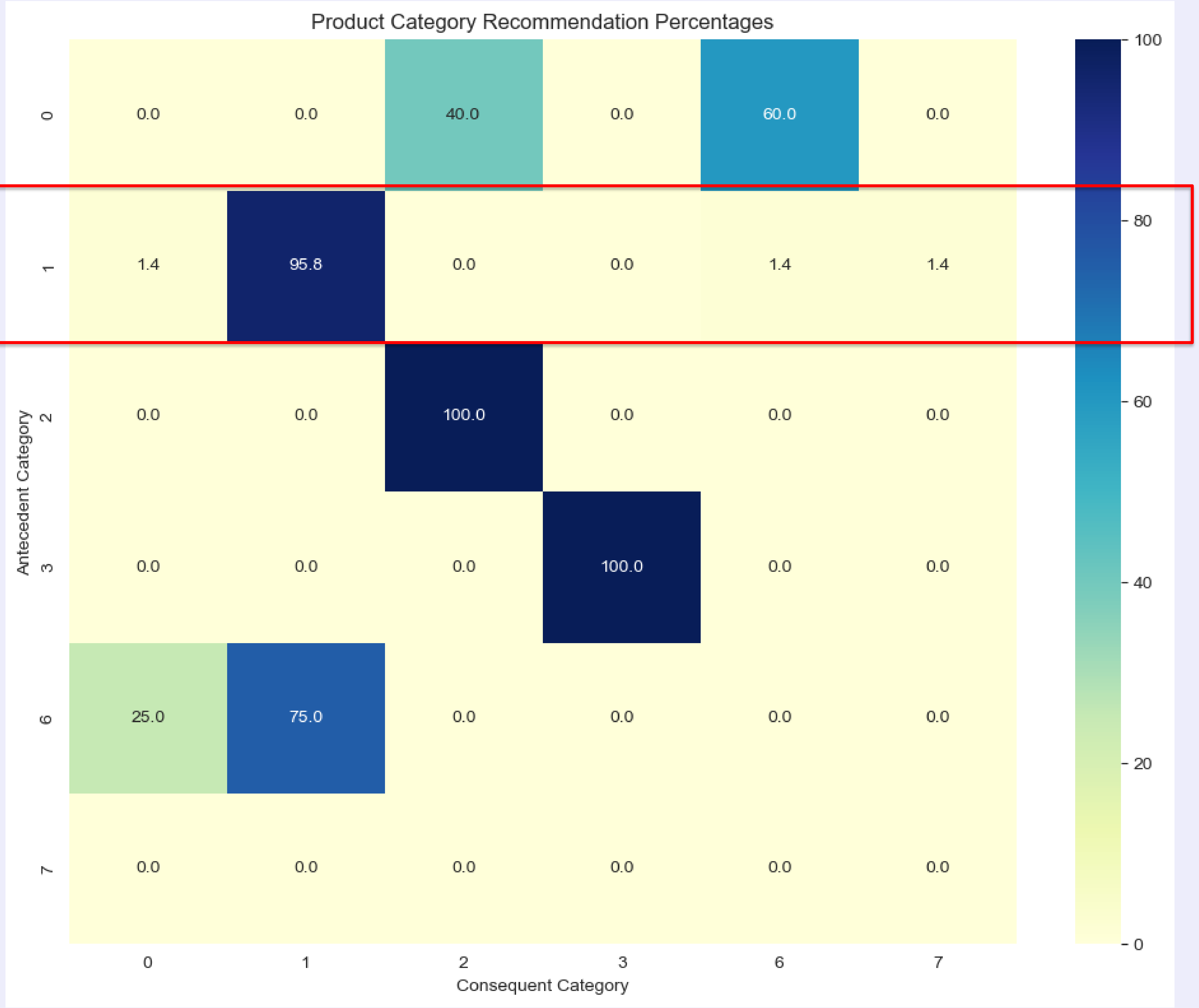
For Cluster 0: Potential Loyalists

Category Spending by Cluster:

Cluster	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	0.00	998.47	39.90	82.62	0.46	0.00	0.00	74.97
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
2	3082.14	79953.23	12379.31	12055.01	1126.56	7.83	3785.70	10866.18
3	104.89	2528.94	403.95	370.22	66.25	2.56	145.98	290.16
4	15.68	531.10	56.03	70.53	8.50	0.69	21.39	31.44

1. **Profile: Moderate shoppers with growth potential**
2. **Size: 20 customers (0.46% of total)**
3. **Average Metrics:**
    1. 60.9 days,
    2. Frequency 2.7 orders,
    3. Total Spend $1137.34
4. **Recommended Strategies:**
    1. Develop targeted offers to encourage category exploration
    2. Create loyalty program incentives that reward increased engagement
    3. Implement personalized product recommendations based on browsing behavior
    4. Use targeted communications highlighting benefits of your most popular products

5.3: Product Recommendation Targeting Customer Segments



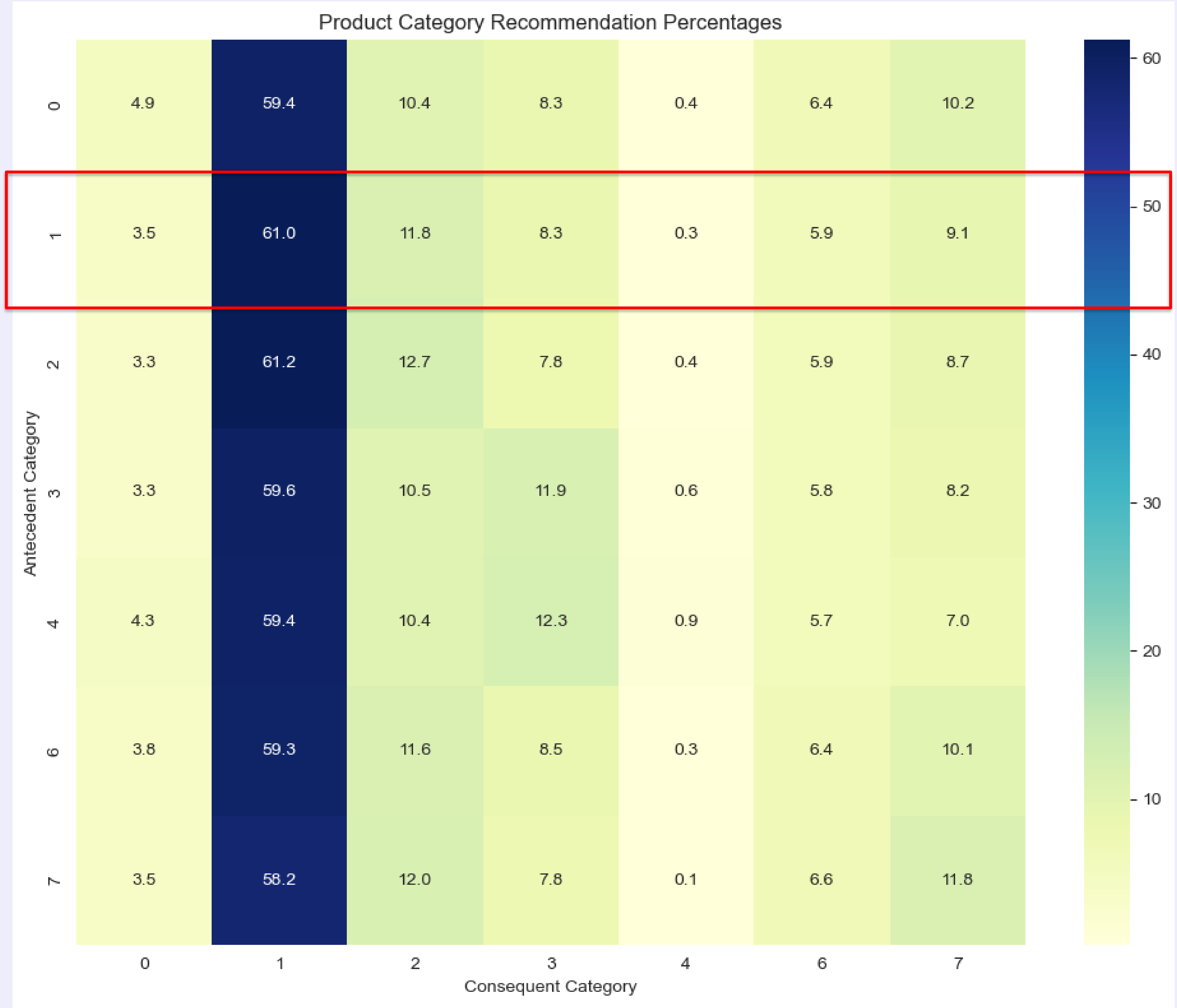
For Cluster 1: Low Value Customers

Category Spending by Cluster:

Cluster	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	0.00	998.47	39.90	82.62	0.46	0.00	0.00	74.97
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
2	3082.14	79953.23	12379.31	12055.01	1126.56	7.83	3785.70	10866.18
3	104.89	2528.94	403.95	370.22	66.25	2.56	145.98	290.16
4	15.68	531.10	56.03	70.53	8.50	0.69	21.39	31.44

1. **Profile: Infrequent shoppers with low spending**
2. **Size: 1831 customers (42.2% of total)**
3. **Average Metrics:**
    1. Recency 58.2 days,
    2. Frequency 2.1 orders,
    3. Total Spend $742.31
4. **Recommended Strategies:**
    1. Focus on increasing transaction value with bundle offers
    2. Create educational content to increase product engagement
    3. Implement limited promotions to encourage more frequent purchases
    4. Analyze if customer acquisition costs justify retention efforts

5.3: Product Recommendation Targeting Customer Segments



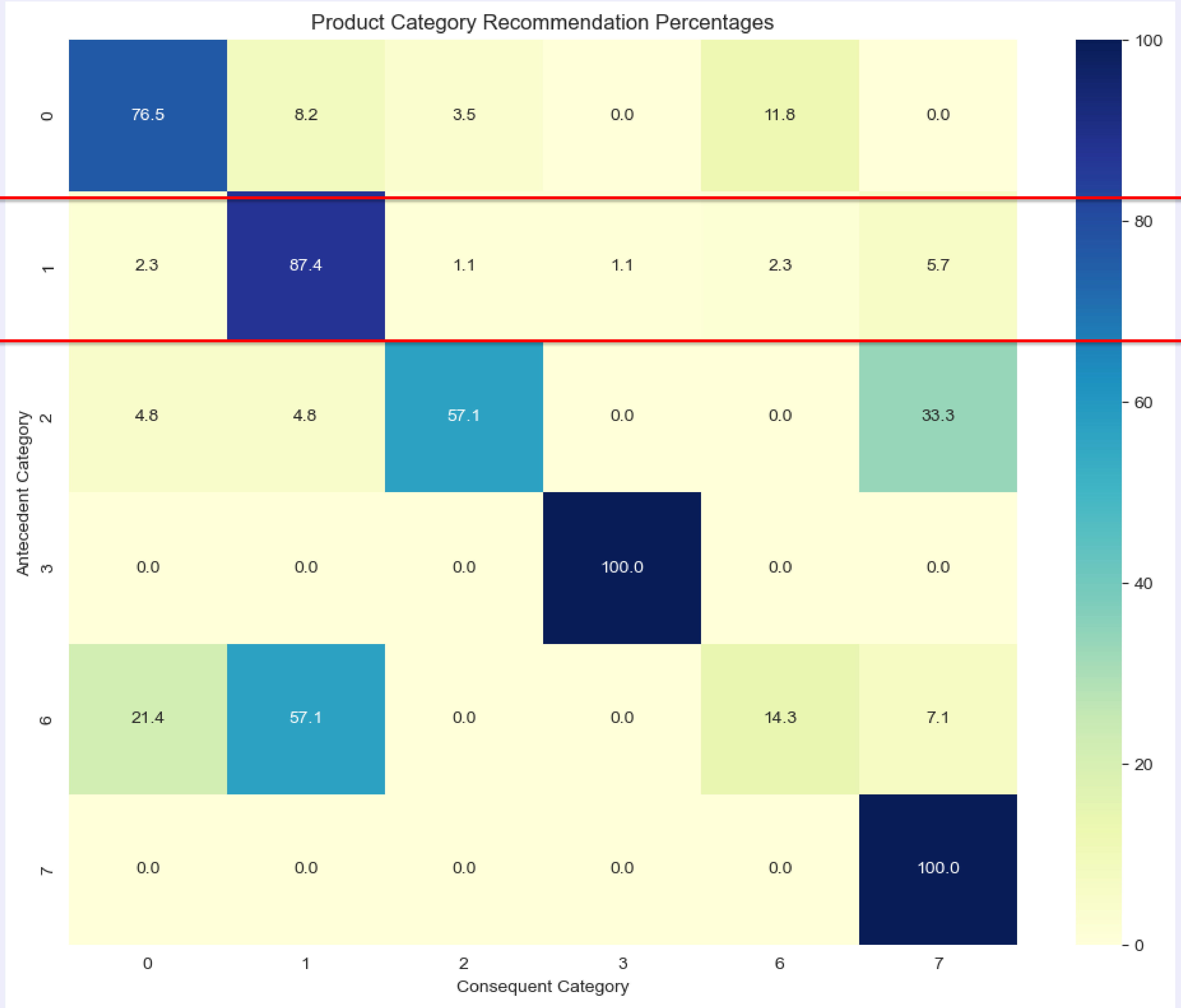
For Cluster 2: Champions

Category Spending by Cluster:

Cluster	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	0.00	998.47	39.90	82.62	0.46	0.00	0.00	74.97
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
2	3082.14	79953.23	12379.31	12055.01	1126.56	7.83	3785.70	10866.18
3	104.89	2528.94	403.95	370.22	66.25	2.56	145.98	290.16
4	15.68	531.10	56.03	70.53	8.50	0.69	21.39	31.44

1. **Profile: High-value active customers who purchase frequently**
2. **Size: 17 customers (0.39% of total)**
3. **Average Metrics:**
   1. Recency 7.0 days,
   2. Frequency 75.2 orders,
   3. Total Spend $108781.56
4. **Recommended Strategies:**
   1. Implement a VIP loyalty program with exclusive benefits
   2. Create early access programs for new products and collections
   3. Personalize communications with thank you messages and special recognition
   4. Develop a referral program to leverage their network

5.3: Product Recommendation Targeting Customer Segments



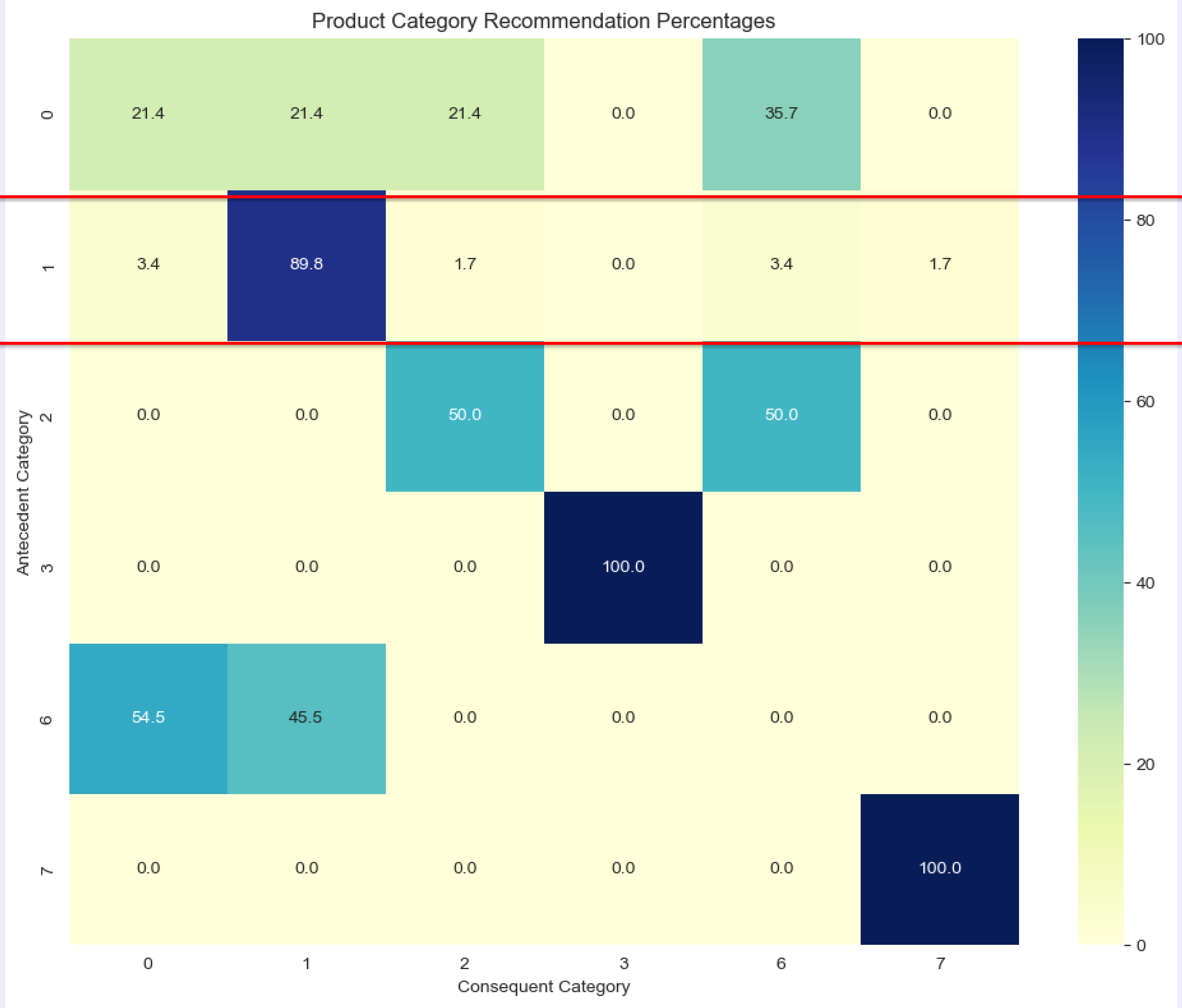
For Cluster 3: Potential Loyalists

Category Spending by Cluster:

Cluster	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	0.00	998.47	39.90	82.62	0.46	0.00	0.00	74.97
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
2	3082.14	79953.23	12379.31	12055.01	1126.56	7.83	3785.70	10866.18
3	104.89	2528.94	403.95	370.22	66.25	2.56	145.98	290.16
4	15.68	531.10	56.03	70.53	8.50	0.69	21.39	31.44

1. **Profile: Moderate shoppers with growth potential**
2. **Size: 1503 customers (34.64% of total)**
3. **Average Metrics:**
   1. Recency 30.4 days,
   2. Frequency 8.0 orders,
   3. Total Spend $3379.47
4. **Recommended Strategies:**
   1. Develop targeted offers to encourage category exploration
   2. Create loyalty program incentives that reward increased engagement
   3. Implement personalized product recommendations based on browsing behavior
   4. Use targeted communications highlighting benefits of your most popular products

5.3: Product Recommendation Targeting Customer Segments



For Cluster 4: Lost Customers

Category Spending by Cluster:

Cluster	CatSpend_0	CatSpend_1	CatSpend_2	CatSpend_3	CatSpend_4	CatSpend_5	CatSpend_6	CatSpend_7
0	0.00	998.47	39.90	82.62	0.46	0.00	0.00	74.97
1	19.65	556.91	102.52	78.69	21.31	0.85	23.71	49.03
2	3082.14	79953.23	12379.31	12055.01	1126.56	7.83	3785.70	10866.18
3	104.89	2528.94	403.95	370.22	66.25	2.56	145.98	290.16
4	15.68	531.10	56.03	70.53	8.50	0.69	21.39	31.44

1. **Profile: Formerly active customers who haven't returned in a long time**
2. **Size: 968 customers (22.31% of total)**
3. **Average Metrics:**
    1. Recency 258.3 days,
    2. Frequency 1.5 orders,
    3. Total Spend $620.73
4. **Recommended Strategies:**
    1. Implement last-chance re-engagement campaigns
    2. Conduct exit surveys to understand why they left
    3. Consider acquisition-cost strategies to win them back
    4. Use their feedback to improve customer retention overall

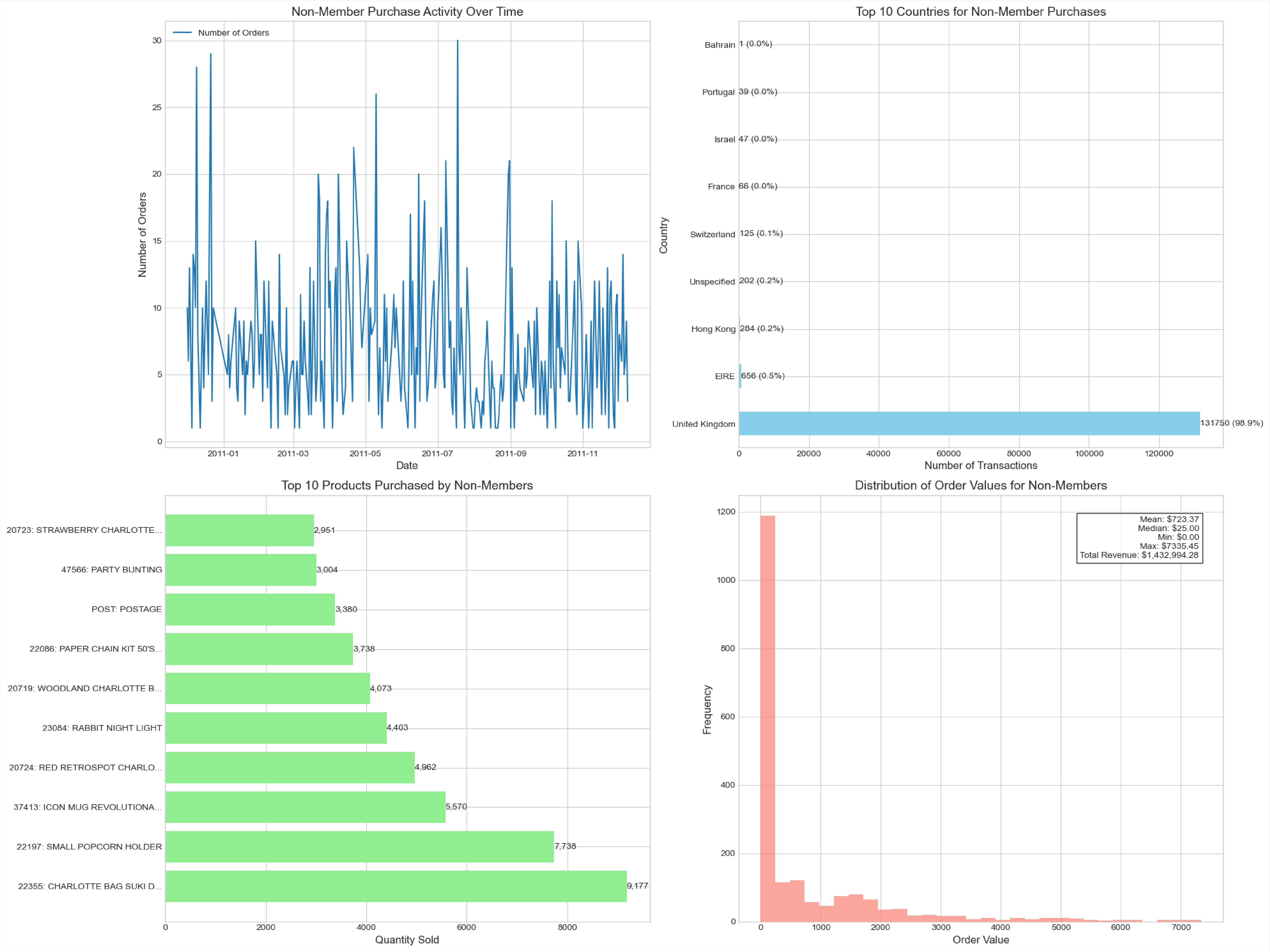
5.3: Product Recommendation Targeting Customer Segments

```
1 (Total: 2625 products)
Sample Products:
```

	StockCode	Description
655	22902	TOTE BAG I LOVE LONDON
85206	79062D	RETRO TIN ASHTRAY,REVOLUTIONARY
9701	22859	EASTER TIN BUNNY BOUQUET
1291	22164	STRING OF STARS CARD HOLDER
247115	23292	SPACEBOY CHILDRENS CUP
2605	90214M	LETTER "M" BLING KEY RING
340561	23433	HANGING QUILTED PATCHWORK APPLES
47	22544	MINI JIGSAW SPACEBOY
150	22646	CERAMIC STRAWBERRY CAKE MONEY BANK
2028	20840	FRENCH FLORAL CUSHION COVER

- From the previous 5 slides showing the targeting strategies/tactics for each segment, it is noticed that Category 1 products (that were previously clustered) were also consistently chosen by all 5 segments
- Below is a sample of Category 1 StockCode and descriptions → For more information, do refer to the attached csv file named **‘product_categories.csv’.**

5.4: Recruitment Strategy for Non-Members



Lastly for non-members, communications could be improved to get them to become customers:

1. Target United Kingdom market
2. For communication tactics to non-members, the top 10 products as per the bottom-left graph boxed in red can be used in the messaging – probably the ‘hero’ products of the online store.
3. This is the best naïve guess as there is no further information about the customers who are not members (no CustomerID)