MACHINE LEARNING

UCI-Online-Retail-II

Customer Purchasing Behavior Analysis

Ainedmbe Denis

Reg. No: 2024-M132-23999

Master's in Information Systems / Intelligent Systems

Introduction - Project Objectives

This project aims to extract meaningful customer segments and identify purchasing patterns for a UK-based online retailer using clustering, deep embedding techniques, and association rule mining in order to generate meaningful customer segments and actionable marketing insights for the UK retailer.

Specific Objectives

1. **Understand Customer Purchasing Behavior:**

   Analyze transaction data to identify patterns in how customers purchase products, including spending habits, transaction frequency, and basket composition.

2. **Segment Customers Based on Spending and Buying Patterns:**

   Apply advanced clustering techniques to group customers into distinct segments based on their purchasing characteristics helping to understand different customer types and their unique needs.

3. **Compare Traditional Clustering with Deep Embedding Clustering:**

   Evaluate the performance of traditional clustering methods (k-Means, DBSCAN) against modern deep learning approaches (Autoencoder embeddings) to determine which method provides better customer segmentation and insights.

4. **Discover Frequently Co-Purchased Product Combinations:**

   Use association rule mining to identify products that are frequently bought together. These insights enable effective cross-selling strategies and product bundling opportunities.

5. **Generate Actionable Marketing Recommendations:**

   Translate analytical findings into concrete, implementable business strategies that can drive revenue growth, improve customer retention, and optimize marketing efforts.

Dataset Overview - Online Retail II UCI

Dataset Source and Context:

The analysis is based on the UCI Online Retail II Dataset, a real-world transactional dataset from a UK-based online retailer. This dataset is publicly available on Kaggle and represents actual business transactions, making it highly valuable for understanding real customer behavior patterns.

Business Context:

The dataset contains transactional data from December 2009 to December 2011, covering a two-year period of business operations. The company operates as a non-store online retailer specializing in unique all-occasion gift-ware. The customer base is diverse, consisting of both individual consumers and wholesale buyers.

Data Attributes:

Each transaction record includes comprehensive information: InvoiceNo (unique invoice identifier), StockCode (product code), Description (product name), Quantity (items purchased), InvoiceDate (transaction timestamp), UnitPrice (price per item), CustomerID (unique customer identifier), and Country (customer location).

Kaggle Link: https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

Real-Life Applications - Applications of Analysis

1

Customer Segmentation & Targeting

This analysis enables retailers to identify distinct customer groups based on their purchasing behavior.

2

Inventory Management

By understanding which customer segments purchase which products, retailers can predict product demand more accurately.

3

Cross-Selling & Product Bundling

Association rule mining reveals products that are frequently purchased together, informing effective product bundling strategies.

4

Customer Retention

The Recency feature helps identify customers who are at risk of churning, allowing for proactive engagement strategies.

5

Business Strategy & Competitive Advantage

Insights generate data-driven decision-making across the organization, leading to strategic advantages.

Analysis Methodology

Part A

Data Cleaning & Clustering (k-Means, DBSCAN)

Part B

Deep Embedding Clustering (Autoencoder)

Part C

Association Rule Mining (FP-Growth)

Part D

Interpretation & Business Recommendations

Loading Dataset & Display Basic information

```
Shape: (1067371, 8)
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
5	489434	22064	PINK DOUGHNUT TRINKET POT	24	2009-12-01 07:45:00	1.65	13085.0	United Kingdom
6	489434	21871	SAVE THE PLANET MUG	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
7	489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	2009-12-01 07:45:00	5.95	13085.0	United Kingdom
8	489435	22350	CAT BOWL	12	2009-12-01 07:46:00	2.55	13085.0	United Kingdom
9	489435	22349	DOG BOWL , CHASING BALL DESIGN	12	2009-12-01 07:46:00	3.75	13085.0	United Kingdom

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  object
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 65.1+ MB
```

	Quantity	Price	Customer ID
count	1.067371e+06	1.067371e+06	824364.000000
mean	9.938898e+00	4.649388e+00	15324.638504
std	1.727058e+02	1.235531e+02	1697.464450
min	-8.099500e+04	-5.359436e+04	12346.000000
25%	1.000000e+00	1.250000e+00	13975.000000
50%	3.000000e+00	2.100000e+00	15255.000000
75%	1.000000e+01	4.150000e+00	16797.000000
max	8.099500e+04	3.897000e+04	18287.000000

Dataset Composition:

The dataset is huge in size, containing **1,067,371** individual transaction records

Data Cleaning

Removal of Missing Product Descriptions

All records with missing product descriptions were removed to maintain data integrity.

Removal of Negative Quantities

Negative quantities (returns) were excluded to ensure the focus on actual purchase decisions.

Removal of Cancelled Invoices

Cancelled invoices were eliminated to provide accurate transaction counts for analysis.

```
Original shape: (1067371, 8)
Cleaned shape: (1042727, 9)
Removed rows: 24644
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	CustomerID	Country	TotalPrice
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom	83.4
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom	81.0
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom	81.0
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom	100.8
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom	30.0
5	489434	22064	PINK DOUGHNUT TRINKET POT	24	2009-12-01 07:45:00	1.65	13085.0	United Kingdom	39.6
6	489434	21871	SAVE THE PLANET MUG	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom	30.0
7	489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	2009-12-01 07:45:00	5.95	13085.0	United Kingdom	59.5
8	489435	22350	CAT BOWL	12	2009-12-01 07:46:00	2.55	13085.0	United Kingdom	30.6
9	489435	22349	DOG BOWL , CHASING BALL DESIGN	12	2009-12-01 07:46:00	3.75	13085.0	United Kingdom	45.0

Customer-Level Features

Customer-level features consisted of monetary value (total spending, transaction count), frequency, Avg. basket composition, and recency of engagement.

	CustomerID	TotalSpending	TransactionCount	TotalQty	AvgBasketSize
0	12346.0	77556.46	12	74285	6190.416667
1	12347.0	5633.32	8	3286	410.750000
2	12348.0	2019.40	5	2714	542.800000
3	12349.0	4428.69	4	1624	406.000000
4	12350.0	334.40	1	197	197.000000
5	12351.0	300.93	1	261	261.000000
6	12352.0	2849.84	10	724	72.400000
7	12353.0	406.76	2	212	106.000000
8	12354.0	1079.40	1	530	530.000000
9	12355.0	947.61	2	543	271.500000

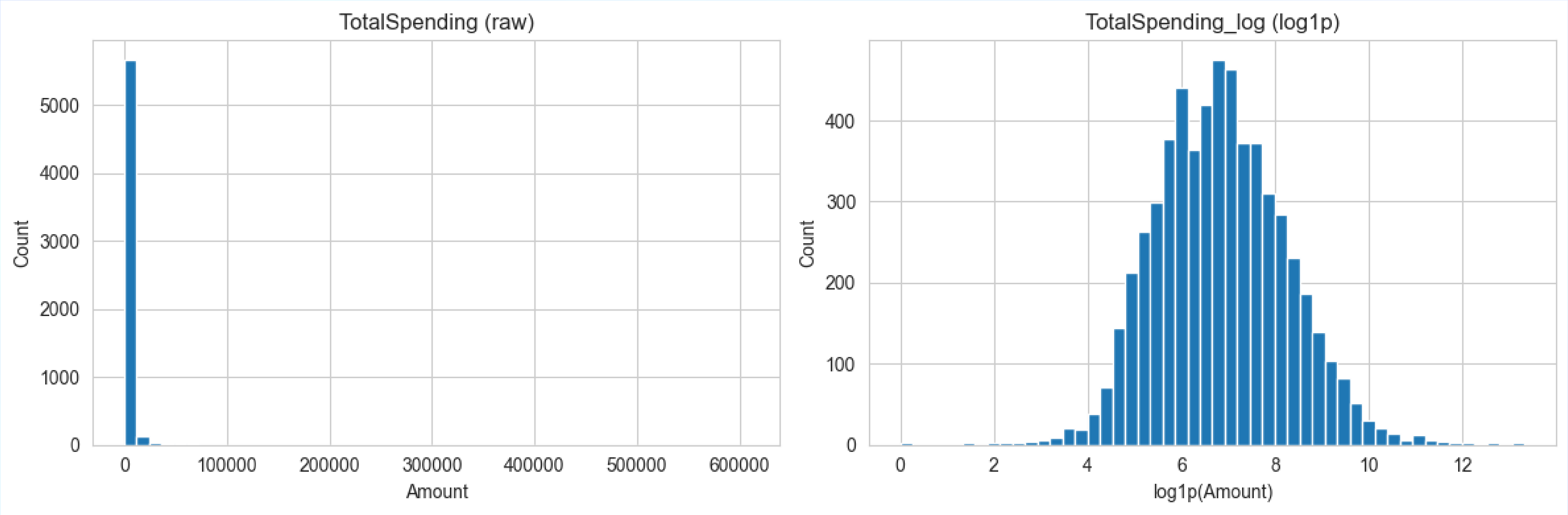
Grouped the data by CustomerID & then calculated aggregates:

- Computed Total Spending, Transaction Count, Total Qty and Avg. Basket Size per customer.
- Compute Average basket size
  = (total items / number of invoices)
- Customers aggregated: 5,881

Final Feature Set

The analysis used four features for clustering: Log-transform was done to reduce skew in money/count columns (TotalSpending & TransactionCount) before clustering

1. TotalSpending_log (log-transformed total spending)
2. TransactionCount_log (log-transformed transaction count)
3. AvgBasketSize (average basket size, original scale)
4. Recency (days since last purchase, original scale)



Final Feature Set —- Cont’ d

The analysis used four features for clustering: Log-transform was done to reduce skew in money/count columns (TotalSpending & TransactionCount) before clustering

1. TotalSpending_log (log-transformed total spending)
2. TransactionCount_log (log-transformed transaction count)
3. AvgBasketSize (average basket size, original scale)
4. Recency (days since last purchase, original scale)

Scaled features preview (first 5 rows):

	CustomerID	TotalSpending_log	TransactionCount_log	AvgBasketSize
0	12346.0	3.170387	1.254938	4.142494
1	12347.0	1.291999	0.800635	0.109969
2	12348.0	0.557317	0.299705	0.202101
3	12349.0	1.119681	0.074457	0.106655
4	12350.0	-0.729063	-1.057568	-0.039166

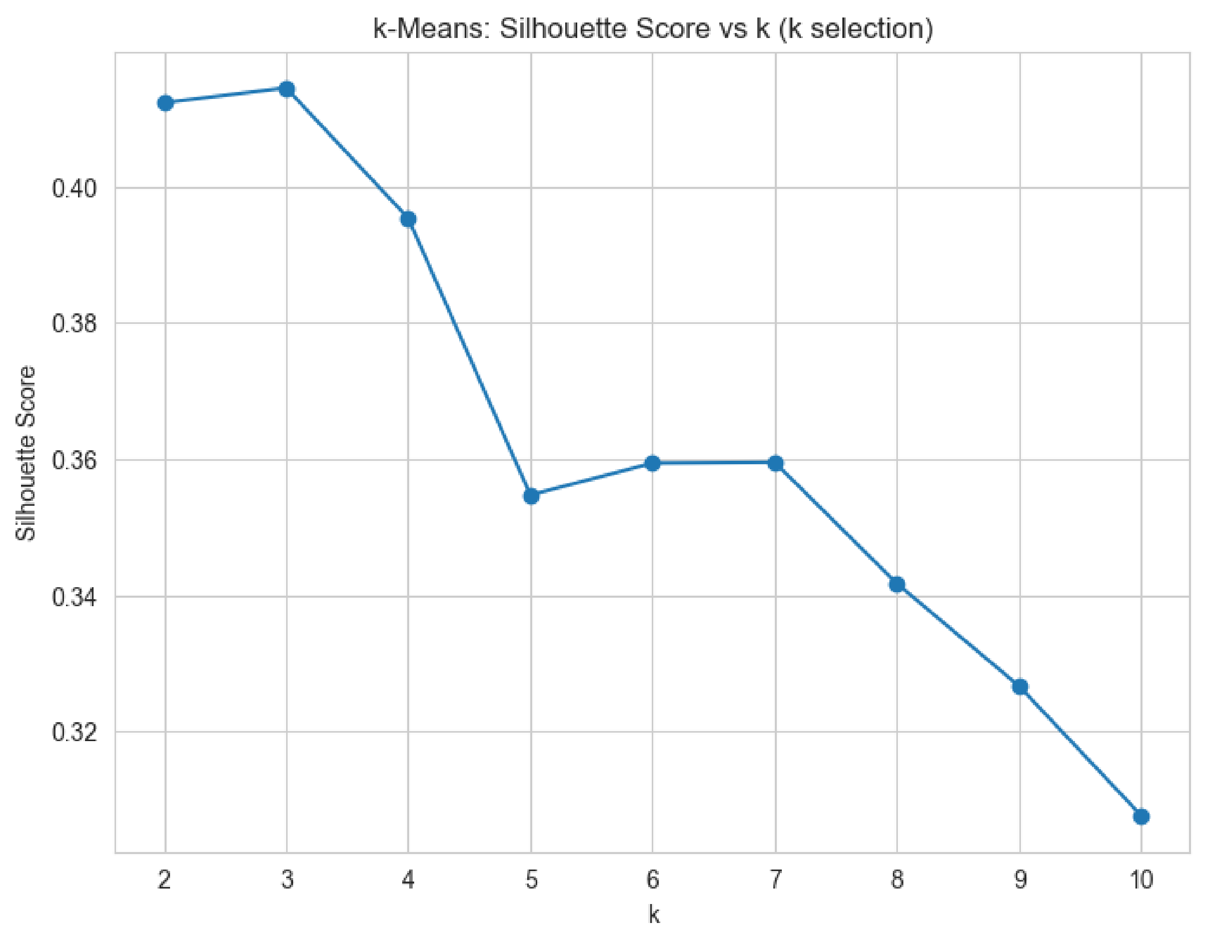
X_scaled shape (customers x features): (5881, 3)

```
Calculating Recency feature...
Recency statistics:
  Min: 0 days
  Max: 738 days
  Mean: 200.5 days
  Median: 95.0 days

Feature preparation for k-Means and DBSCAN:
Features: ['TotalSpending_log', 'TransactionCount_log', 'AvgBasketSize', 'Recency']
X_scaled_kmeans shape: (5881, 4)

First 5 rows of scaled features:
   TotalSpending_log  TransactionCount_log  AvgBasketSize   Recency
0           3.170387              1.254938       4.142494  0.594598
1           1.291999              0.800635       0.109969 -0.952264
2           0.557317              0.299705       0.202101 -0.603743
3           1.119681              0.074457       0.106655 -0.871102
4          -0.729063             -1.057568      -0.039166  0.518209
```

k-Means, DBSCAN clustering & Computing silhouette scores

The lot shows the process of selecting the optimal number of customer segments. I tested k from 2 to 10 and measured clustering quality using the Silhouette Score. The peak at k=3 (score 0.4144) indicates that three segments best capture distinct customer groups. Using fewer or more clusters reduces separation quality, so we proceed with k=3

k-MEANS CLUSTERING - Parameter Tuning




```
Silhouette scores for k=2..10:
  k=2: 0.4123
  k=3: 0.4144 <-- OPTIMAL
  k=4: 0.3953
  k=5: 0.3547
  k=6: 0.3594
  k=7: 0.3595
  k=8: 0.3417
  k=9: 0.3267
  k=10: 0.3077

----------------------------------------------------------
k-MEANS FINAL RESULTS:
  Selected k: 3
  Number of clusters: 3
  Silhouette Score: 0.4144
```

Findings

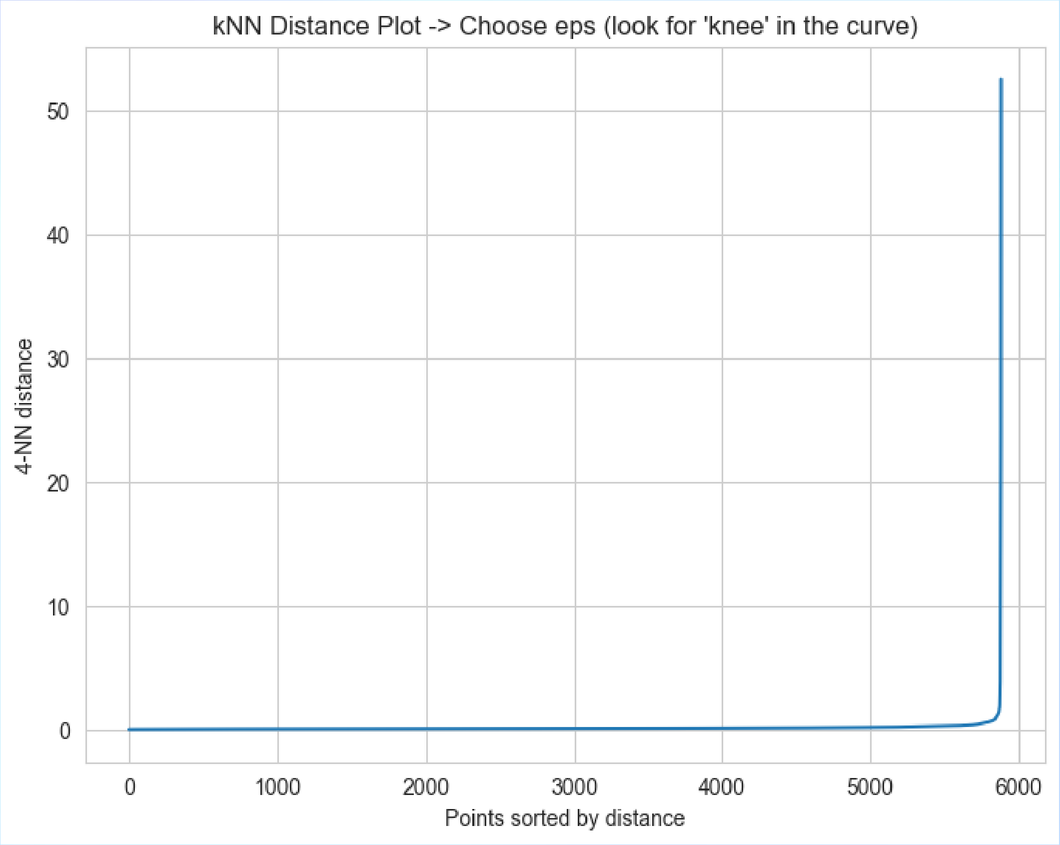
- k=3 provides the best separation for this dataset
- The score of 0.4144 indicates reasonable clustering quality.
- This suggests 3 distinct customer segments

k-Means, DBSCAN clustering & Computing silhouette scores

The kNN distance plot shows the "knee" around point 5800, indicating where dense regions transition to outliers. I tested eps from 0.3 to 1.5; eps=0.5 produced 2 clusters with the best silhouette score (0.3186).

Smaller eps created too many small clusters, while larger values merged everything into one cluster. Final selection: eps=0.5, identifying 2 customer segments and 101 noise points (outliers).

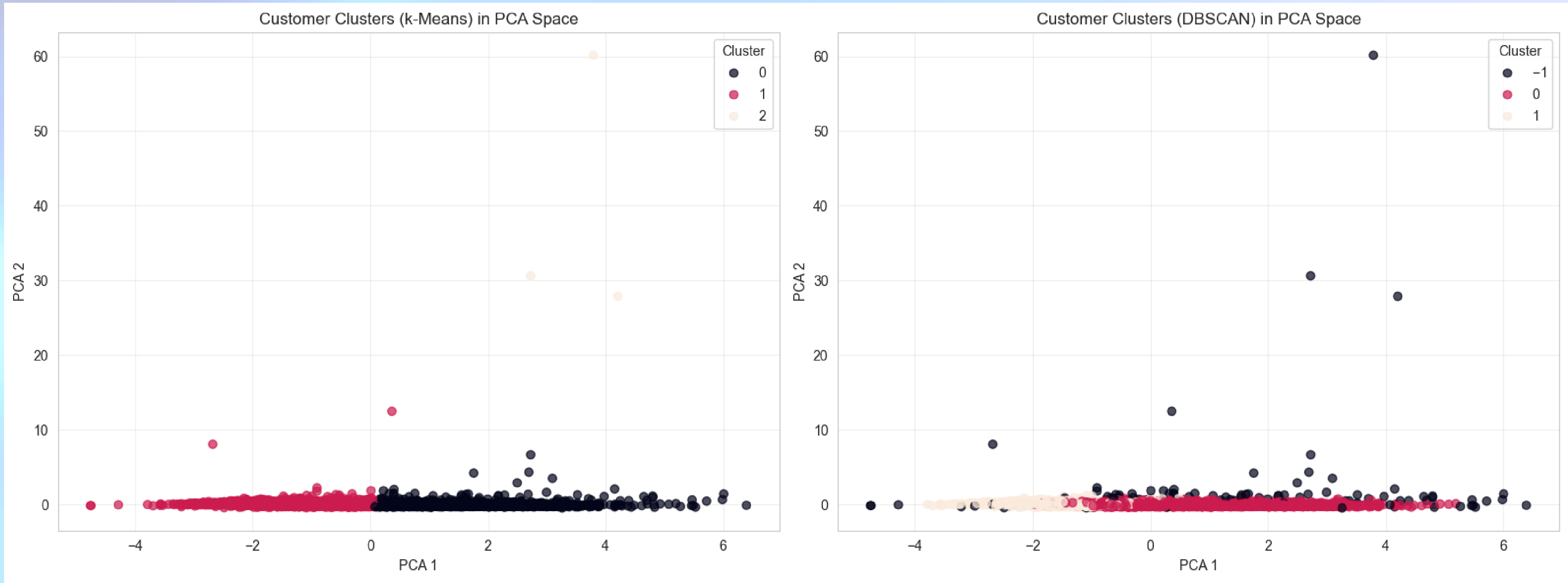
DBSCAN CLUSTERING - Parameter Tuning



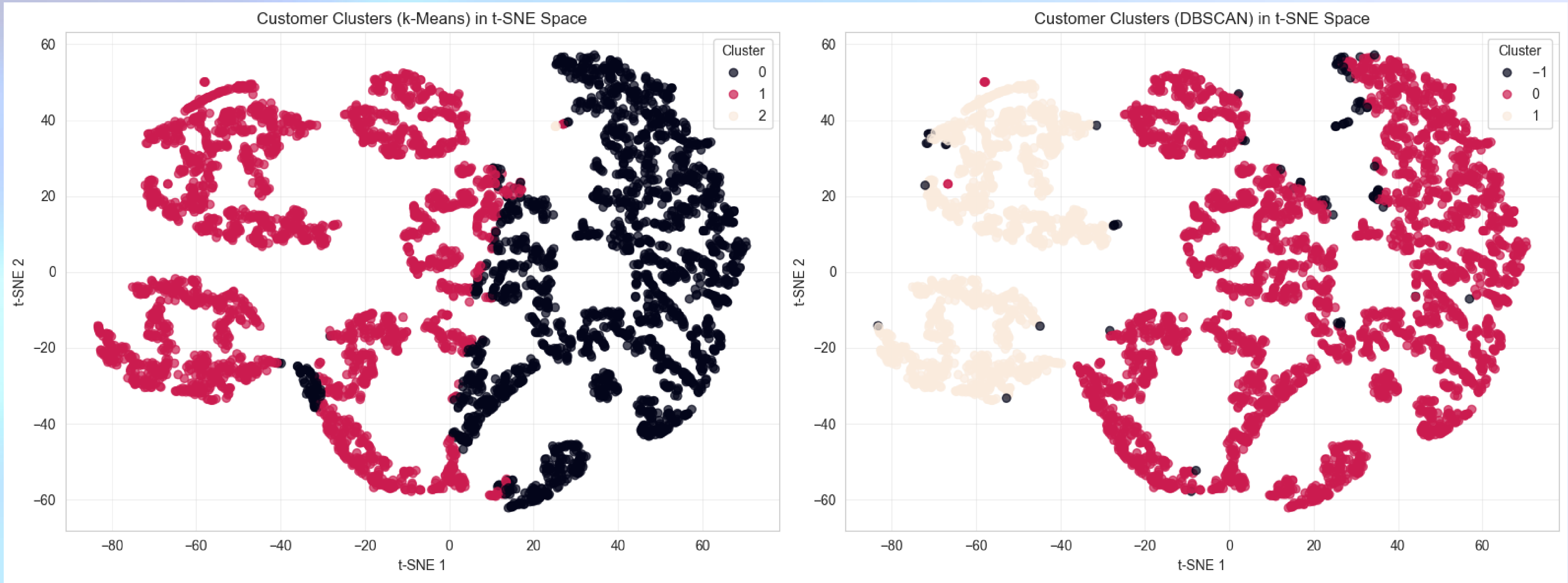

```
Testing different eps values for DBSCAN:
  eps=0.3: 5 clusters, 241 noise points, silhouette=0.0871 <-- OPTIMAL
  eps=0.5: 2 clusters, 101 noise points, silhouette=0.3186 <-- OPTIMAL
  eps=0.7: 1 clusters, 43 noise points, silhouette=not meaningful
  eps=1.0: 1 clusters, 26 noise points, silhouette=not meaningful
  eps=1.5: 1 clusters, 11 noise points, silhouette=not meaningful

----------------------------------------------------------------------
DBSCAN FINAL RESULTS:
  Selected eps: 0.5
  Number of clusters: 2
  Noise points: 101
  Silhouette Score: 0.3186 (excluding noise)
```

Unlike k-Means, DBSCAN automatically flags outliers that don't fit the main segments.

Visualizations: Scatter plots for k-Means & DBSCAN clusters using PCA projections.



Visualizations: Scatter plots for k-Means & DBSCAN clusters using t-SNE projections.



- t-SNE is computationally expensive: it scales quadratically with the number of samples ($O(n^2)$).
- Used Sub sample of 5000 to keep t-SNE fast on large datasets

Clustering Results

k-Means Clustering Results

Cluster 0 - Medium-Value Regular Customers: 2,714 customers, average total spending of £5,848.67.

Cluster 1 - Low-Value Occasional Buyers: 3,164 customers, average total spending of £523.29.

Cluster 2 - Bulk Buyers: 3 customers, average total spending of £71,482.87.

DBSCAN Clustering Results

DBSCAN identified 2 distinct clusters, highlighting the ability to identify outliers.

k-Means vs DBSCAN Comparison

Metric	k-Means	DBSCAN	Interpretation
Number of Clusters	3	2	k-Means provides more granular segmentation
Silhouette Score	0.4144	0.3186	k-Means achieves better cluster separation
Noise Points	None (all assigned)	101 outliers	DBSCAN identifies unusual customers

Deep Embedding Clustering (Autoencoder)

The autoencoder architecture was designed for customer segmentation.

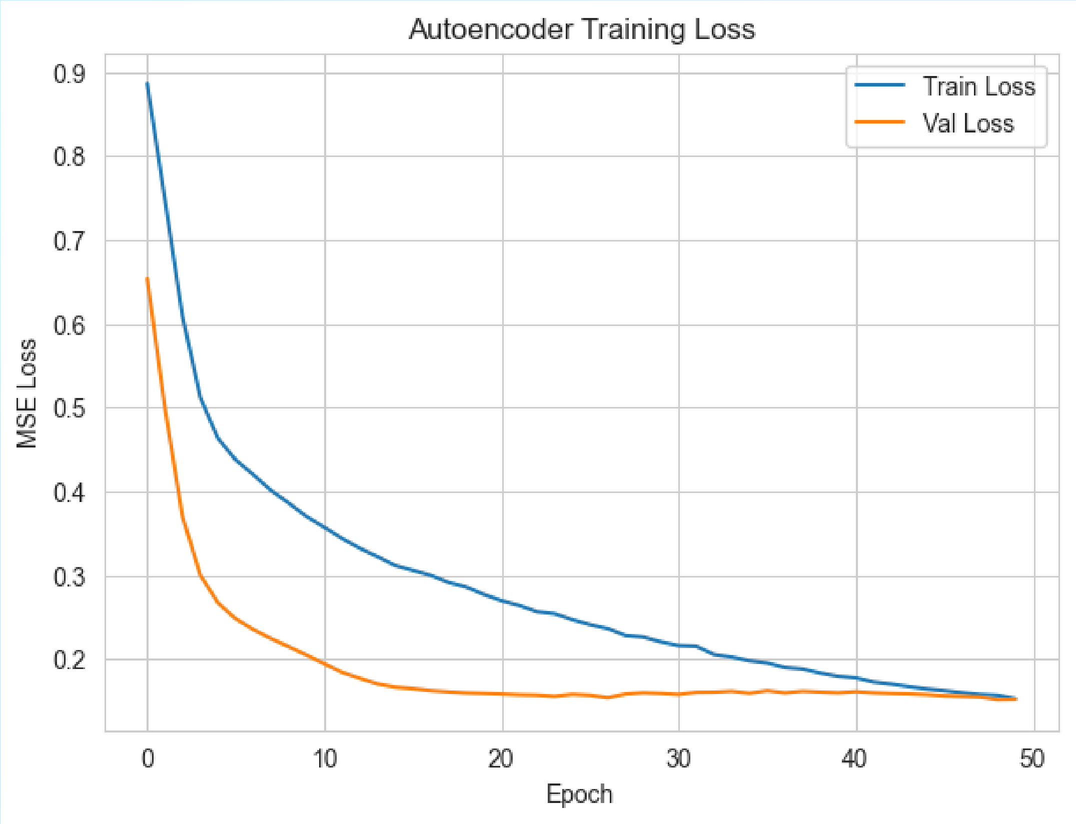
The autoencoder was trained using: **Loss Function Mean Squared Error (MSE)** - measuring how well the network reconstructs the input; **Optimizer: Adam optimizer** - an adaptive learning rate algorithm that adjusts learning rates for each parameter; **Training Objective** - Minimize reconstruction error, forcing the network to learn efficient representations

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 4)	0
dense_3 (Dense)	(None, 8)	40
bottleneck (Dense)	(None, 2)	18
dense_4 (Dense)	(None, 8)	24
dense_5 (Dense)	(None, 4)	36

Total params: 118 (472.00 B)

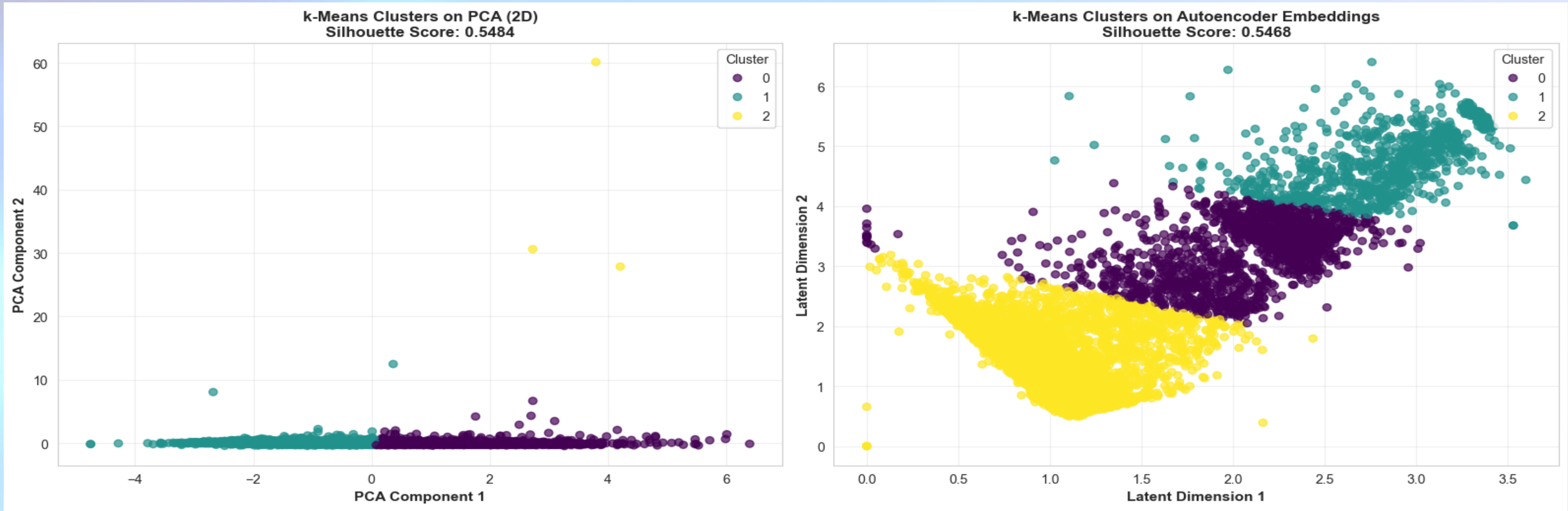
Trainable params: 118 (472.00 B)

Non-trainable params: 0 (0.00 B)



Embeddings and clustering & Compare cluster quality

The side-by-side plots compare k-Means clustering on PCA-reduced data (left) versus Autoencoder embeddings (right)



Despite similar silhouette scores, the Autoencoder embeddings show clearer separation. PCA's score is boosted by the well-separated outliers, while the main clusters overlap.
The Autoencoder learned a latent space that better separates the customer segments, making the clusters more interpretable and actionable for business use.

Association Rule Mining (FP-Growth)

Convert data into basket format: Invoice to a list of Description items [Created basket: rows = Invoice, columns = Description (1/0)], Built binary matrix with Invoice as rows and Description as column

Applied FP-Growth algorithm to find frequent itemsets, then generates association rules and extracted the top 10 rules sorted by lift.

```
BASKET MATRIX VERIFICATION
______________________________________________________________________________
Matrix shape: 40301 invoices x 5469 products

Total purchases (sum of all 1s): 997,007
Total possible entries: 220,406,169
Sparsity: 99.55% (expected for retail data)

Invoices with at least one purchase: 40,301
Products purchased at least once: 5,469

Sample invoices with purchases:
  Invoice 489434: 8 items
    Sample items:  WHITE CHERRY LIGHTS, 15CM CHRISTMAS GLASS BALL 20 LIGHTS...
  Invoice 489435: 4 items
    Sample items: CAT BOWL , DOG BOWL , CHASING BALL DESIGN...
  Invoice 489436: 19 items
    Sample items:  PEACE WOODEN BLOCK LETTERS, AREA PATROLLED METAL SIGN...
  Invoice 489437: 23 items
    Sample items: BLUE PADDED SOFT MOBILE, CHOCOLATE HOT WATER BOTTLE...
  Invoice 489438: 17 items
    Sample items: CARROT CHARLIE+LOLA COASTER SET, CHARLIE & LOLA WASTEPAPER BIN BLUE...
```

```
Number of frequent itemsets: 1056
```

	support	itemsets
0	0.057319	(STRAWBERRY CERAMIC TRINKET BOX)
1	0.019330	(SAVE THE PLANET MUG)
2	0.017220	(PINK DOUGHNUT TRINKET POT)
3	0.013771	(RECORD FRAME 7" SINGLE SIZE)
4	0.012828	(15CM CHRISTMAS GLASS BALL 20 LIGHTS)
5	0.069676	(ASSORTED COLOUR BIRD ORNAMENT)
6	0.051264	(HOME BUILDING BLOCK WORD)
7	0.042083	(LOVE BUILDING BLOCK WORD)
8	0.041910	(SCOTTIE DOG HOT WATER BOTTLE)
9	0.020669	(HEART IVORY TRELLIS LARGE)

Association Rule Mining (FP-Growth) —- Cont’ d

The Top 10 strongest Association rule

TOP 10 STRONGEST ASSOCIATION RULES (by Lift)

Total rules generated: 848

Top 10 rules:

	antecedents	consequents	support	confidence	lift
809	(POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE KITCHEN)	0.010149	0.725177	52.469247
804	(POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE KITCHEN)	(POPPY'S PLAYHOUSE LIVINGROOM)	0.010149	0.734291	52.469247
808	(POPPY'S PLAYHOUSE BEDROOM)	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE LIVINGROOM)	0.010149	0.581792	49.465849
805	(POPPY'S PLAYHOUSE KITCHEN, POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE BEDROOM)	0.010149	0.862869	49.465849
806	(POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE KITCHEN)	0.010149	0.887202	48.187489
807	(POPPY'S PLAYHOUSE KITCHEN)	(POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE LIVINGROOM)	0.010149	0.551213	48.187489
803	(POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE BEDROOM)	0.011439	0.817376	46.857846
802	(POPPY'S PLAYHOUSE BEDROOM)	(POPPY'S PLAYHOUSE LIVINGROOM)	0.011439	0.655761	46.857846
801	(POPPY'S PLAYHOUSE LIVINGROOM)	(POPPY'S PLAYHOUSE KITCHEN)	0.011761	0.840426	45.646886
800	(POPPY'S PLAYHOUSE KITCHEN)	(POPPY'S PLAYHOUSE LIVINGROOM)	0.011761	0.638814	45.646886

FP-Growth identified 1,056 frequent itemsets and generated 848 association rules, leveraging the insights to inform cross-selling and bundling strategies.

Association Rule Mining (FP-Growth) —- Cont’ d

```
4 RANDOMLY SAMPLED ASSOCIATION RULES FOR INTERPRETATION
________________________________________________________________________________________________________________________

Rule 1: IF a customer buys [POPPY'S PLAYHOUSE LIVINGROOM ] THEN they also tend to buy [POPPY'S PLAYHOUSE KITCHEN]
  Support:    0.0118
  Confidence: 0.8404
  Lift:       45.6469
--------------------------------------------------------------------------------------------------------------------
Rule 2: IF a customer buys [POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE KITCHEN] THEN they also tend to buy [POPPY'S PLAYHOUSE LIVINGROOM ]
  Support:    0.0101
  Confidence: 0.7343
  Lift:       52.4692
--------------------------------------------------------------------------------------------------------------------
Rule 3: IF a customer buys [POPPY'S PLAYHOUSE KITCHEN] THEN they also tend to buy [POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE LIVINGROOM ]
  Support:    0.0101
  Confidence: 0.5512
  Lift:       48.1875
--------------------------------------------------------------------------------------------------------------------
Rule 4: IF a customer buys [POPPY'S PLAYHOUSE LIVINGROOM ] THEN they also tend to buy [POPPY'S PLAYHOUSE BEDROOM , POPPY'S PLAYHOUSE KITCHEN]
  Support:    0.0101
  Confidence: 0.7252
  Lift:       52.4692
--------------------------------------------------------------------------------------------------------------------
```

Interpretation

```
HIGH-VALUE SEGMENTS - k-Means Clusters (Sorted by Total Spending):
```

kmeans_cluster	TotalSpending	TransactionCount	AvgBasketSize	Customer_Count
2	71482.87	2.67	57261.83	3
0	5848.67	11.48	269.85	2714
1	523.29	1.84	184.75	3164

```
HIGH-VALUE SEGMENTS - Autoencoder Clusters (Sorted by Total Spending):
```

ae_cluster	TotalSpending	TransactionCount	AvgBasketSize	Customer_Count
2	3478.76	7.85	285.38	3613
0	3023.89	4.59	211.15	1424
1	1029.21	2.48	185.94	844

- k-Means: Cluster 0=Low-value (62.5%), Cluster 1=Bulk buyers (0.05%), Cluster 2=Medium-value regular (37.5%)
- Autoencoder: Cluster 0=Low-value (56.9%), Cluster 1=High-value frequent (12.0%), Cluster 2=Medium-value occasional (31.1%)

Interpretation

```
CLUSTER INTERPRETATIONS (Based on Actual Data)
________________________________________________________________________________

k-MEANS CLUSTERS:
--------------------------------------------------------------------------------
Cluster 0: MEDIUM-VALUE CUSTOMERS (2714 customers, 46.1%)
  - Moderate spending (£5,848.67), occasional transactions (11.5), small-medium baskets (£269.85)

Cluster 1: LOW-VALUE OCCASIONAL BUYERS (3164 customers, 53.8%)
  - Low spending (£523.29), infrequent transactions (1.8), small baskets (£184.75) - largest customer segment

Cluster 2: BULK BUYERS (3 customers, 0.1%)
  - High spending (£71,482.87), infrequent purchases (2.7 transactions), very large basket size (£57,261.83) - likely bulk/wholesale buyers


AUTOENCODER CLUSTERS:
--------------------------------------------------------------------------------
Cluster 0: LOW-VALUE OCCASIONAL BUYERS (1424 customers, 24.2%)
  - Low spending (£3,023.89), infrequent transactions (4.6), small baskets (£211.15) - largest customer segment

Cluster 1: LOW-VALUE OCCASIONAL BUYERS (844 customers, 14.4%)
  - Low spending (£1,029.21), infrequent transactions (2.5), small baskets (£185.94) - largest customer segment

Cluster 2: LOW-VALUE OCCASIONAL BUYERS (3613 customers, 61.4%)
  - Low spending (£3,478.76), infrequent transactions (7.8), small baskets (£285.38) - largest customer segment

--------------------------------------------------------------------------------
```

Interpretation

```
HIGHEST-VALUE CUSTOMER SEGMENTS IDENTIFIED
______________________________________________________________________

k-MEANS CLUSTERING:
----------------------------------------------------------------------
HIGHEST-VALUE SEGMENT: Cluster 2
   - Average Total Spending: £71,482.87
   - Average Transactions: 2.67
   - Average Basket Size: £57,261.83
   - Number of Customers: 3
   - Percentage of Total: 0.05%

Top High-Value Segments (k-Means):
   Highest: Cluster 2 - £71,482.87 avg spending (3 customers)
   Second: Cluster 0 - £5,848.67 avg spending (2714 customers)
   Third: Cluster 1 - £523.29 avg spending (3164 customers)

AUTOENCODER CLUSTERING:
----------------------------------------------------------------------
HIGHEST-VALUE SEGMENT: Cluster 2
   - Average Total Spending: £3,478.76
   - Average Transactions: 7.85
   - Average Basket Size: £285.38
   - Number of Customers: 3613
   - Percentage of Total: 61.44%

Top High-Value Segments (Autoencoder):
   Highest: Cluster 2 - £3,478.76 avg spending (3613 customers)
   Second: Cluster 0 - £3,023.89 avg spending (1424 customers)
   Third: Cluster 1 - £1,029.21 avg spending (844 customers)
```

COMPARISON: PCA vs Deep Embedding Clusters

CLUSTERING QUALITY:

- k-Means on PCA (2D): 0.5484
- k-Means on Autoencoder embeddings: 0.5468
- Difference: 0.0017
- PCA performs better by: 0.30%

Business Recommendations

1: Cross-Selling Strategy

Create product bundles based on these strong associations

Display 'Frequently Bought Together' recommendations on product pages

Offer bundle discounts (5-10% off) to incentivize cross-selling

2: Loyalty Programs

Target high-value customers (VIP) with tiered loyalty programs with tiered rewards:

* Exclusive early access to sales
* Free shipping on all orders
* Birthday discounts and personalized offers
* Points multiplier (2x-3x points per £1 spent)

Focus retention efforts on these segments (highest lifetime value)

3: Targeted Discounts

Develop segment-specific discount strategies to enhance customer re-engagement.

Segment-specific email campaigns with personalized offers

Time-limited promotions to encourage immediate purchases

Conclusion

This project highlighted the power of data-driven strategies in understanding and improving customer purchasing behavior.

- END