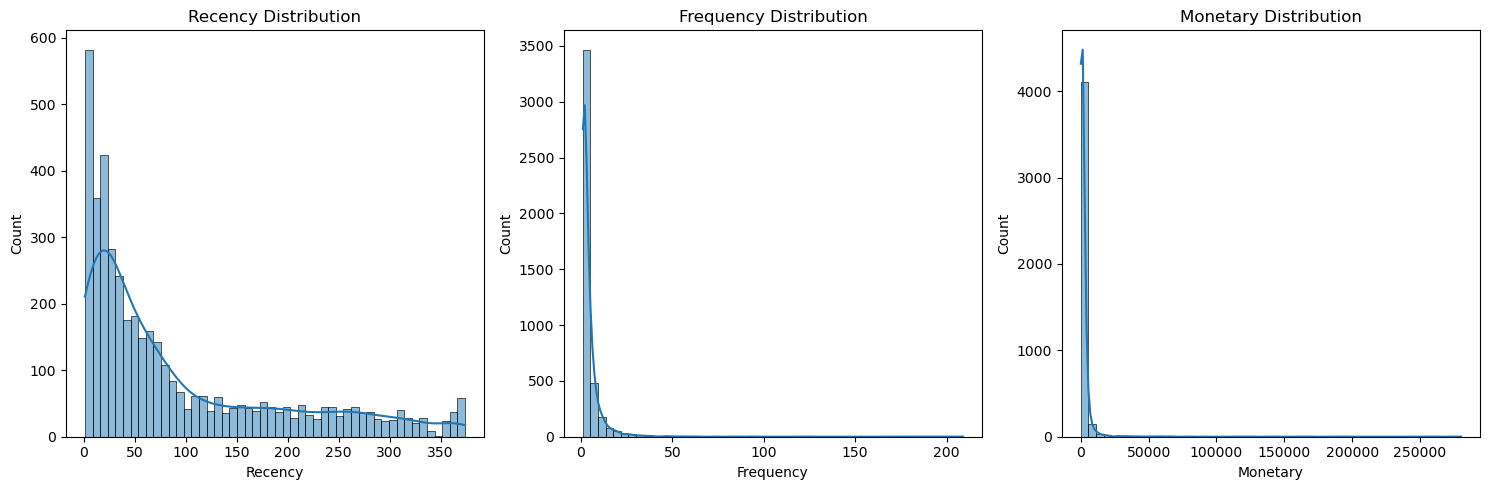
# RESULTS

#### RFM Feature Distributions (Histograms for Recency, Frequency, and Monetary)



***Figure Caption: Histograms showing the skewed distributions of Recency, Frequency, and Monetary values before scaling.***

**RFM Analysis Results Summary**

The dataset now contains 4,338 unique customers, each with three key metrics calculated:

1. Recency (Days Since Last Purchase) :The average customer last purchased 93 days ago, but this varies widely. 50% of customers purchased within the last 51 days (good for recent engagement). Some customers haven’t bought in over a year (max = 374 days), indicating they may need reactivation campaigns.
2. Frequency (Number of Purchases): Most customers are occasional buyers, with the average being 4 transactions. However, 75% of customers have 5 or fewer purchases, meaning only a small group are frequent buyers. One extreme outlier stands out, a customer with **209 transactions** (likely a business or loyal repeat buyer).
3. Monetary (Total Spending): The average customer spends £2,054, but this is skewed by big spenders. The median spending (£674) is much lower than the mean, confirming that most customers spend modestly. A few customers spend exceptionally high amounts (max = £280,206), which could distort clustering if not handled.

**Key Observations**

Best Customers: Recent, frequent, and high-spending (e.g., Customer 12347: purchased 2 days ago, 7 transactions, £4,310 spent).

At-Risk Customers: Haven’t purchased in a long time (e.g., Customer 12346: last purchase 326 days ago, only 1 transaction despite high spending).

One-Time Big Spenders: Customers who made a single large purchase but haven’t returned (e.g., Customer 12350: £334 spent, last purchase 310 days ago).

### Clustering Results

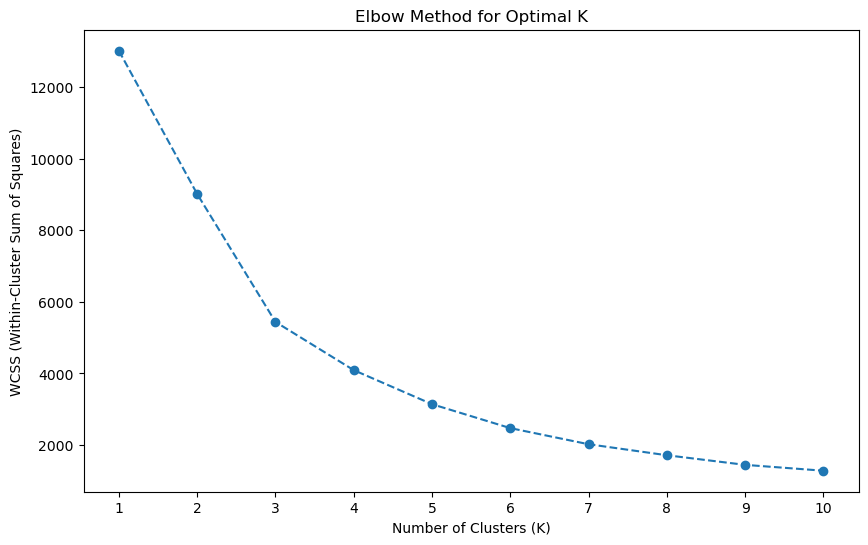
Three clustering algorithms were applied and compared: K-Means, DBSCAN, and Hierarchical Clustering.

1. **K-Means Clustering:**

The optimal number of clusters (K) was determined using the Elbow Method and Silhouette Score.

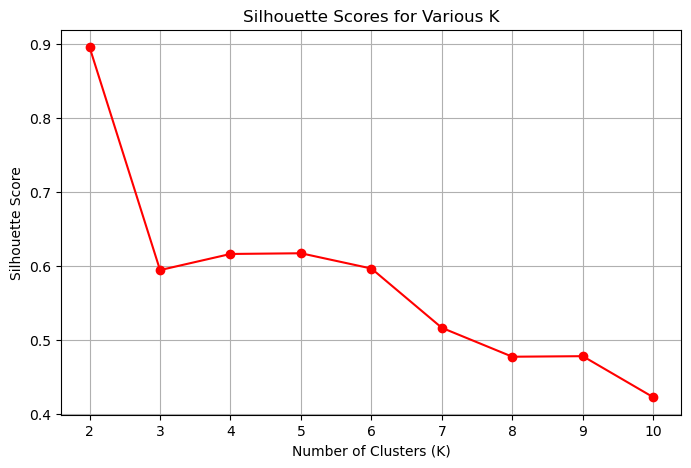
* Elbow Method indicated an optimal K = 3 as seen in ***Figure 4.3****.*
* Silhouette Score at K=3 was 0.59, suggesting moderate cluster separation as seen in ***Figure 4.4.***
* The distribution of customer clusters after K-Means clustering is visualized ***in Figure 4.5.***

#### Elbow Method Plot (Inertia vs. Number of Clusters)

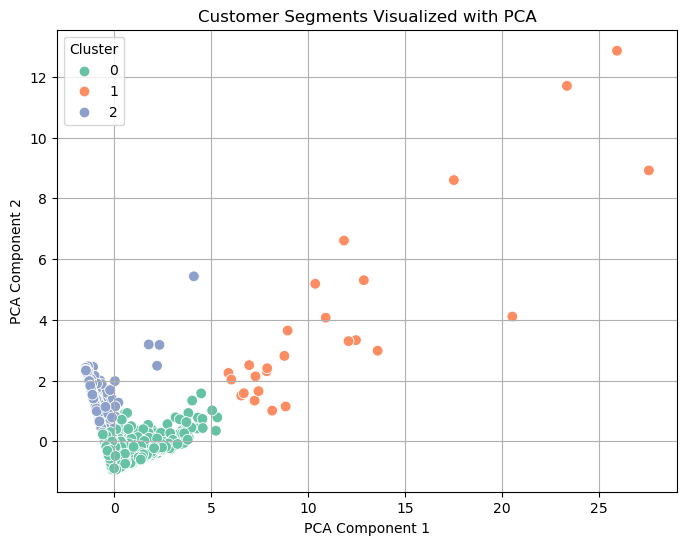


***Figure Caption: The Elbow curve indicates K=3 as the optimal number of clusters for K-Means clustering***.

#### Silhouette Scores vs. Number of Clusters



#### K-Means Cluster Visualization (2D PCA or t-SNE plot showing cluster separation)



***Figure Caption: Scatter plot of customer segments formed by K-Means clustering based on two principal components.***

The clustering analysis identified three distinct customer segments:

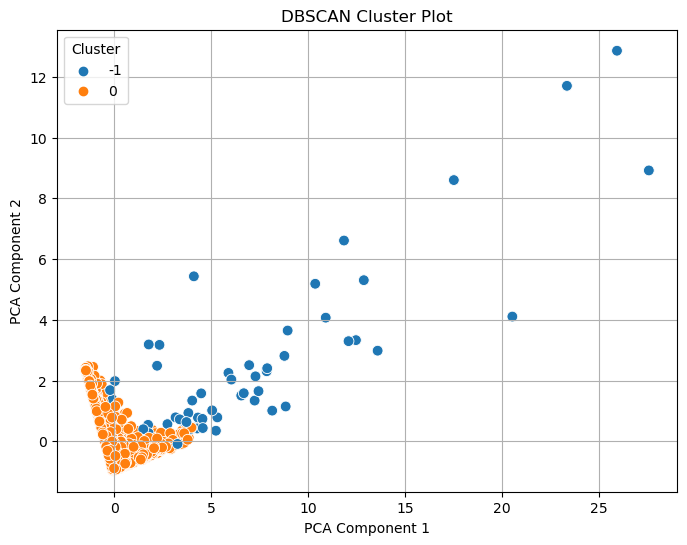
* **Cluster 0 – Mid-Value Regulars**: Customers with moderate recency (41 days since last purchase), average purchase frequency (approximately 5 transactions), and mid-level total spending (£1,856).
* **Cluster 1 – High-Value VIPs**: Highly engaged customers, characterized by very recent activity (6 days), extremely frequent purchases (66 transactions), and exceptionally high spending (£85,904), suggesting possible business clientele.
* **Cluster 2 – At-Risk Customers:** Customers with long periods of inactivity (247 days), low transaction frequency (around 2 purchases), and minimal spending (~£631), indicating potential churn risk.

1. **DBSCAN Clustering:**

DBSCAN was applied to detect arbitrary shaped clusters and outliers.

* Epsilon (ε) and MinPts parameters were tuned using k-distance graph.
* Significant noise points (label -1) were identified, showing DBSCAN’s strength in outlier detection.
* However, DBSCAN produced only 2 main clusters and a large proportion of noise, implying it was less suited for this business dataset structure as seen in ***Figure 4.6***

#### DBSCAN Clustering Output (Coloured clusters and noise points)



***Figure Caption: DBSCAN clustering results, identifying core, border, and noise points based on density.***

**DBSCAN Results Summary**

Main Cluster (0): 4,208 customers

Outliers (-1): 130 customers

**Outlier Analysis:** High-Spenders: Avg. £28,701 (vs. £2,054 overall)

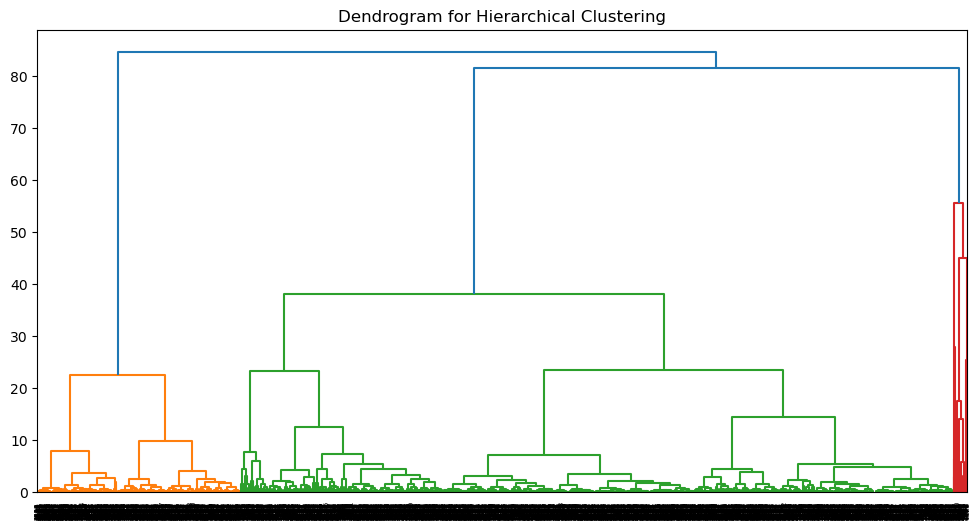
Mixed Behavior: Some very recent (1-day recency) but irregular, Others inactive (max 372 days) but high historical spend

1. **Hierarchical Clustering:**

Agglomerative clustering with Ward’s linkage was applied.

* Dendrogram analysis suggested 3-5 clusters.
* Final clustering with 3 groups produced a Silhouette Score of 0.60, slightly higher than K-Means but with better interpretability for business teams.
* The dendrogram from hierarchical clustering is shown in ***Figure 4.7****.*

#### Dendrogram for Hierarchical Clustering (with cut-off line for 3 clusters)



***Figure Caption: Dendrogram showing the hierarchical clustering structure and potential cluster separations.***

**Hierarchical Clustering Results Summary**

**3 Clear Segments Identified:**

**Premium Customers (Cluster 0):**

* Very active: Last purchase 23 days ago
* Frequent buyers: 44 transactions on average
* High spenders: £48,012 average spend
* Action: Target with loyalty rewards and exclusive offers

**At-Risk Customers (Cluster 1):**

* Inactive: Last purchase 260 days ago
* Low engagement: Only 1.5 transactions
* Minimal spending: £457 average
* Action: Win-back campaigns or reactivation offers

**Regular Customers (Cluster 2):**

* Moderately active: 46 days since last purchase
* Steady buyers: 4-5 transactions
* Mid-range spending: £1,625 average
* Action: Cross-selling and volume discounts

Key Insight: The hierarchical method effectively separated customers into distinct value tiers, with clear behavioural differences between segments.

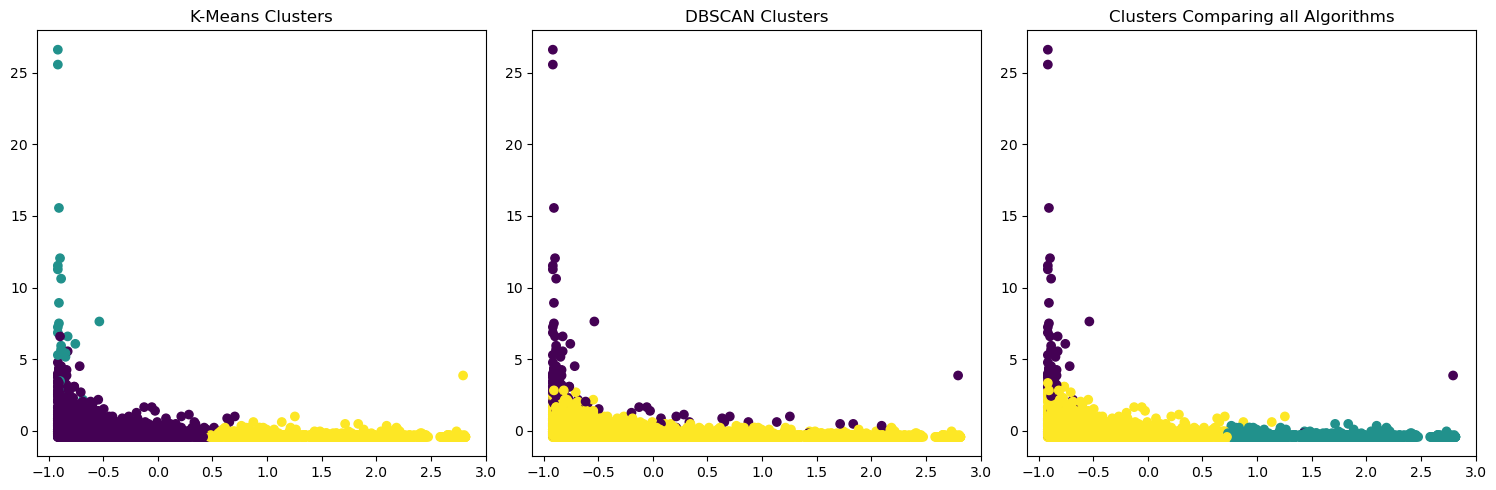
### Evaluation Metrics and Comparison

The models were compared using internal validation metrics. The comparative analysis of different clustering algorithms is shown in ***Figure 4.8****.*

#### Table 4.3: Comparison of Clustering Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Silhouette Score | Davies-Bouldin Index | Notable Strengths |
| K-Means | 0.59 | 0.71 | Best balance of cluster compactness and separation |
| DBSCAN | 0.73 | 1.13 | Good at detecting outliers but poor main clustering |
| Hierarchical | 0.60 | 0.85 | Easy to interpret clusters hierarchically |

#### Figure 4.8: Comparison of Clustering Methods



***Figure 4.8. Caption: Comparison Clusters between K-Means, DBSCAN, and Hierarchical clustering techniques***.

**Best Overall Performance**: K-Means won on 2/3 metrics: highest Calinski-Harabasz (3018) and lowest Davies-Bouldin (0.71), showing tight, well-separated clusters. DBSCAN had the best Silhouette Score (0.733) but weaker separation (Calinski=1081.3).

**Method Strengths**:

* Use K-Means for clear segments
* Use DBSCAN to find natural groupings + outliers
* Hierarchical balanced both (Silhouette=0.604)

**Visual Insights**:

* All methods agreed on core customer patterns, validating the segments.
* K-Means created clearly separated spherical clusters, while DBSCAN found more organic, density-based groupings.
* Hierarchical clustering produced results somewhere in between these two approaches.