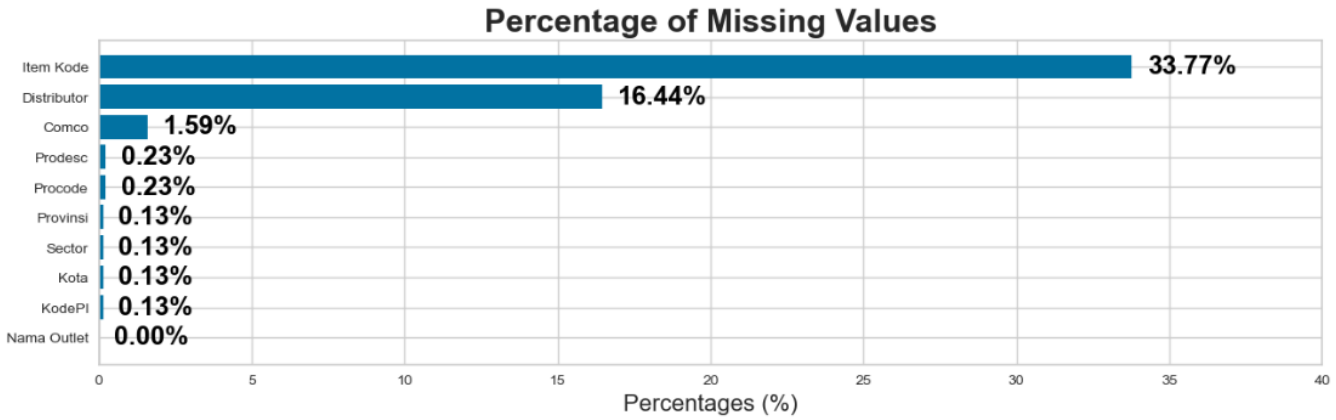
1. Data Cleaning & Transformation



2. Feature Engineering

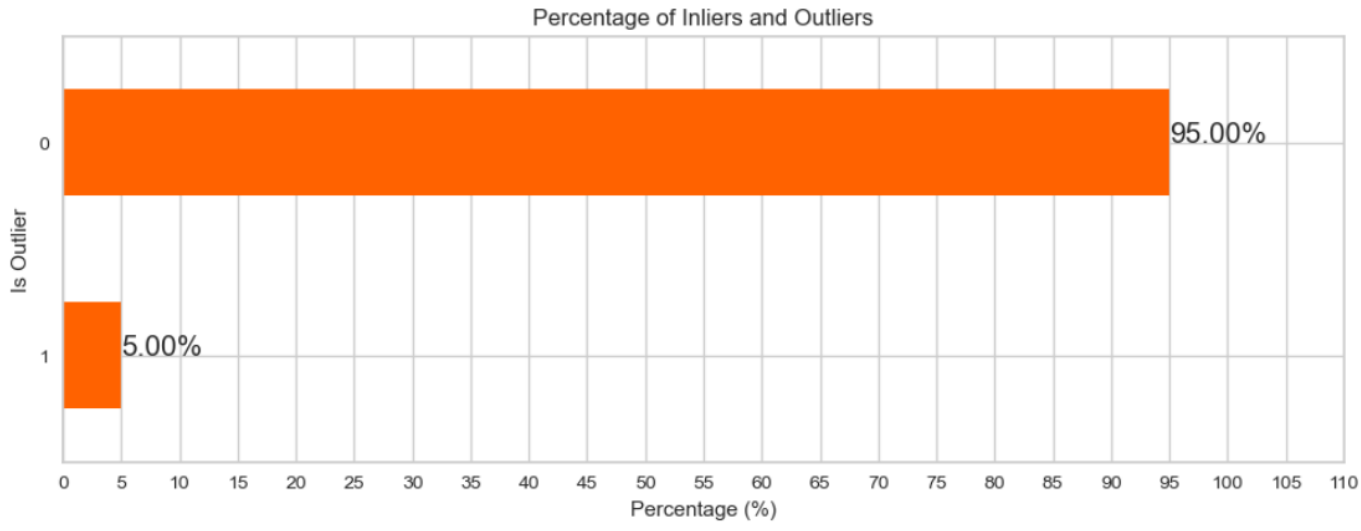
The RFM method is used to group customers based on their transactional behavior. The following is a brief explanation of each RFM dimension:

1. Recency: Reflects how recently the customer made the last transaction. Customers who have recently made a transaction may be more valuable than those who have not transacted in a long time.

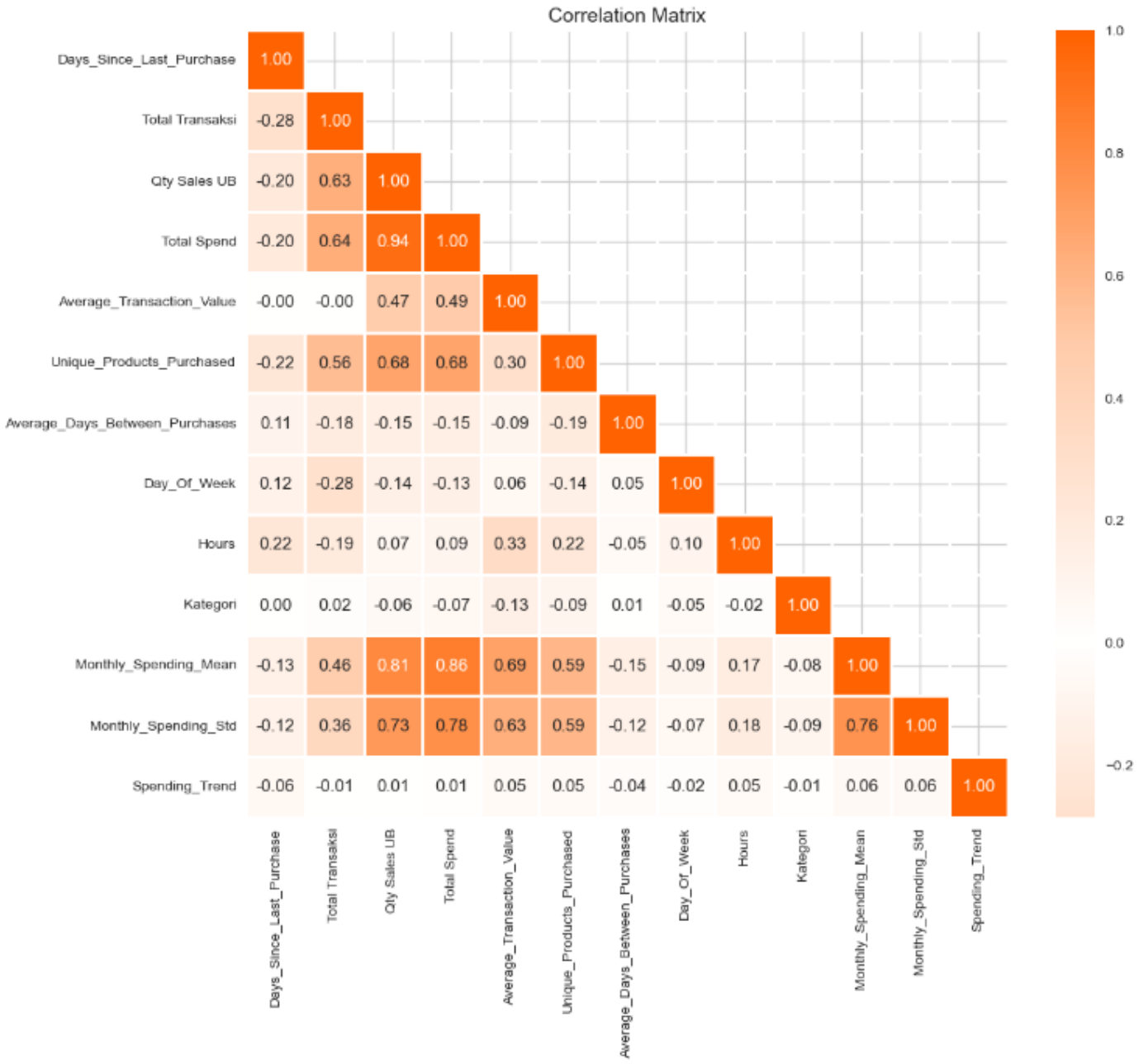
2. Frequency: Measures how often customers make transactions within a certain period of time. Customers who transact frequently may have a higher value to the business.

3. Monetary Value: Represents the total value of transactions made by customers in a period of time. Customers who spend more money on your product or service are usually considered more valuable.

3. Outlier Detection & Treatment



4. Correlation Analysis

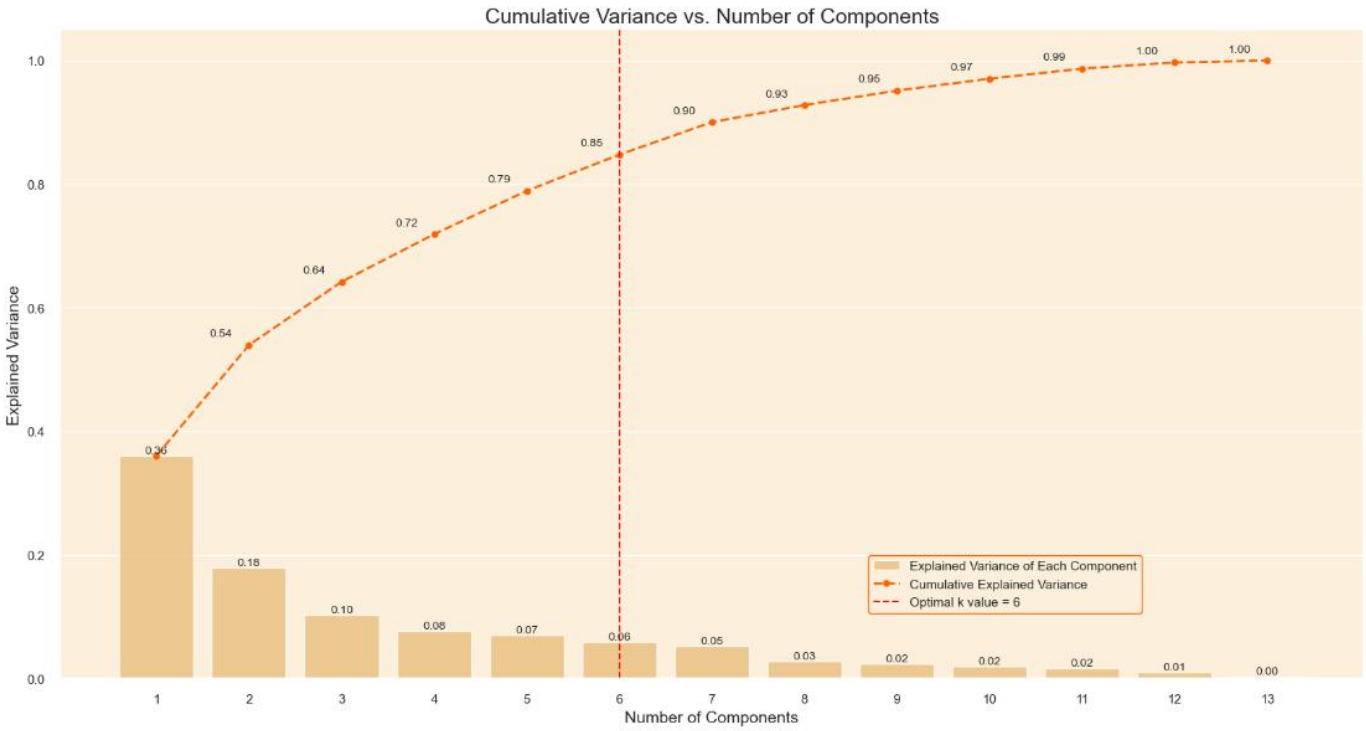


5. Feature Scalling

	Nama Outlet	Days_Since_Last_Purchase	Total Transaksi	Qty Sales UB	Total Spend	Average_Transaction_Value	Unique_Products_Purchased	Average_Days_Between_
0	1000 SEHAT, APT	-0.432787	-0.088772	-0.270220	-0.338466	-0.427385	-0.671345	
1	1001 FARMA, APT	0.560308	-0.754080	-0.588111	-0.582553	-0.572248	-0.728364	

que_Products_Purchased	Average_Days_Between_Purchases	Day_Of_Week	Hours	Kategori	Monthly_Spending_Mean	Monthly_Spending_Std	Spending_Trend
-0.671345	-0.121552	1	-0.745339	1	-0.465846	-0.414626	0.115742
-0.728364	0.466459	2	-0.745339	1	-0.626804	-0.507428	0.282697

6. Dimensionality Reduction



Conclusion:

The plot and the cumulative variance explained value show how much of the total variance in the data set is captured by each principal component, as well as the cumulative variance explained by the first n components.

Here, we can see that:

- The first component explains about 36% of the variance.
- the second component explains about 54% of the variance.
- the third component explains about 64% of the variance, and so on.

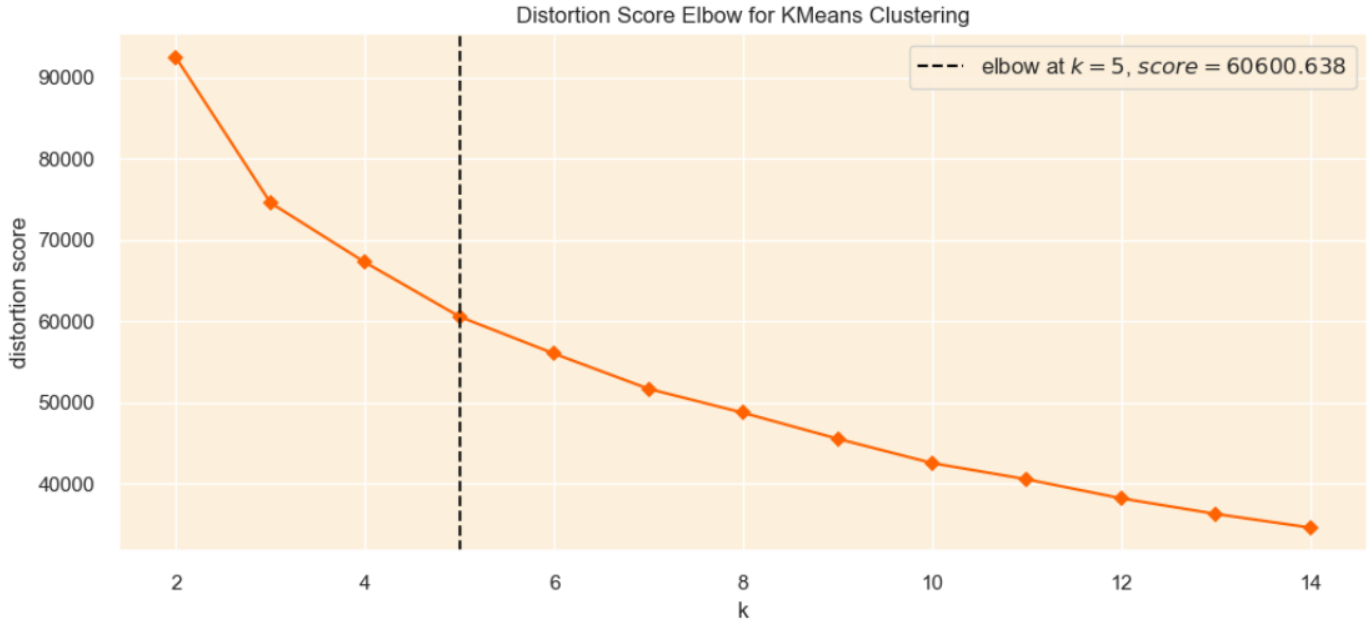
To choose the optimal number of components, we usually look for the point where adding another component does not significantly increase the cumulative variance explained, which is often referred to as the “elbow point” in the curve.

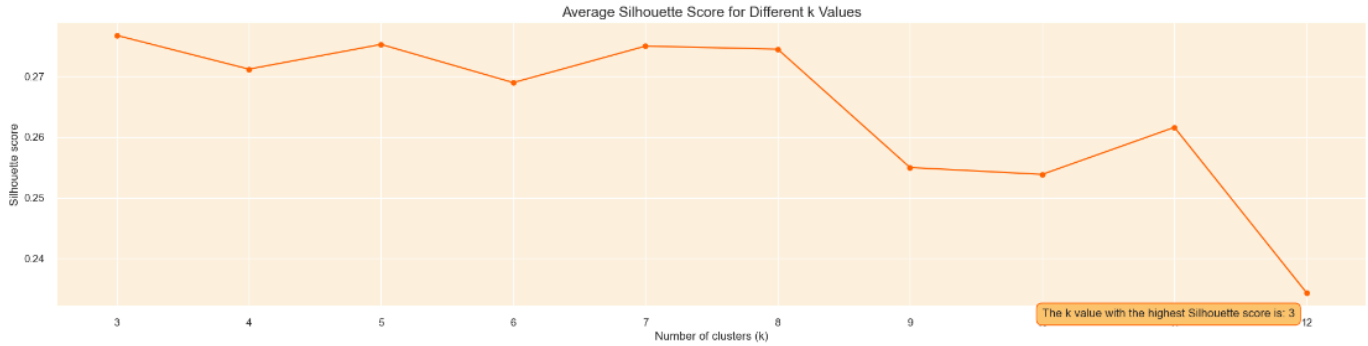
From the plot, we can see that the increase in cumulative variance starts to slow down after the 6th component (which captures about 85% of the total variance).

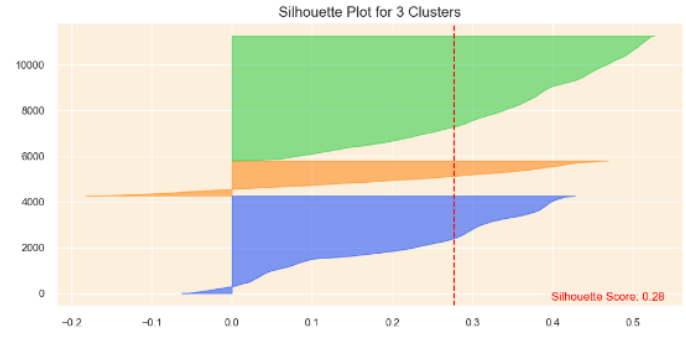
Considering the context of customer segmentation, I wanted to retain a sufficient amount of information to effectively identify different customer groups. Therefore, retaining the first 6 components may be a balanced choice, as they together explain most of the total variance while reducing the dimensionality of the dataset.

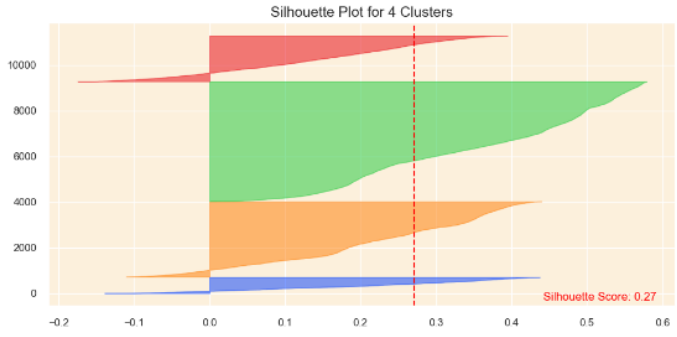
	PC1	PC2	PC3	PC4	PC5	PC6
Days_Since_Last_Purchase	-0.115734	0.148356	0.455757	0.264133	-0.135671	-0.558945
Total Transaksi	0.296739	-0.204077	-0.369101	0.063663	-0.067232	-0.379574
Qty Sales UB	0.415638	0.041663	-0.101834	0.104019	0.034401	-0.090955
Total Spend	0.425208	0.052280	-0.084689	0.105686	0.044621	-0.083576
Average_Transaction_Value	0.264511	0.266675	0.344879	-0.018357	0.089171	0.489152
Unique_Products_Purchased	0.350042	0.012682	-0.074279	-0.065268	-0.117887	-0.319763
Average_Days_Between_Purchases	-0.102770	0.020436	0.098526	0.516790	0.770052	-0.145243
Day_Of_Week	-0.176090	0.875569	-0.435932	-0.011394	-0.006512	-0.098094
Hours	0.067839	0.239636	0.540288	-0.147451	-0.242412	-0.230345
Kategori	-0.021631	-0.040147	-0.015247	0.004847	-0.018872	-0.053966
Monthly_Spending_Mean	0.403506	0.123907	0.078965	0.043108	0.078843	0.116847
Monthly_Spending_Std	0.377319	0.138213	0.100506	0.040392	0.098174	0.121607
Spending_Trend	0.025945	0.008098	0.076316	-0.779190	0.532967	-0.267578

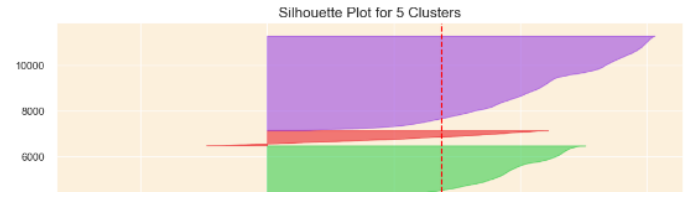
7. K - Means Clustering (Outlet Segmentation)

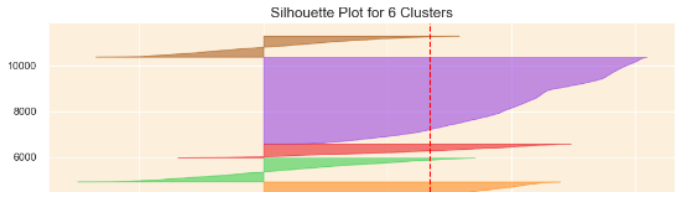


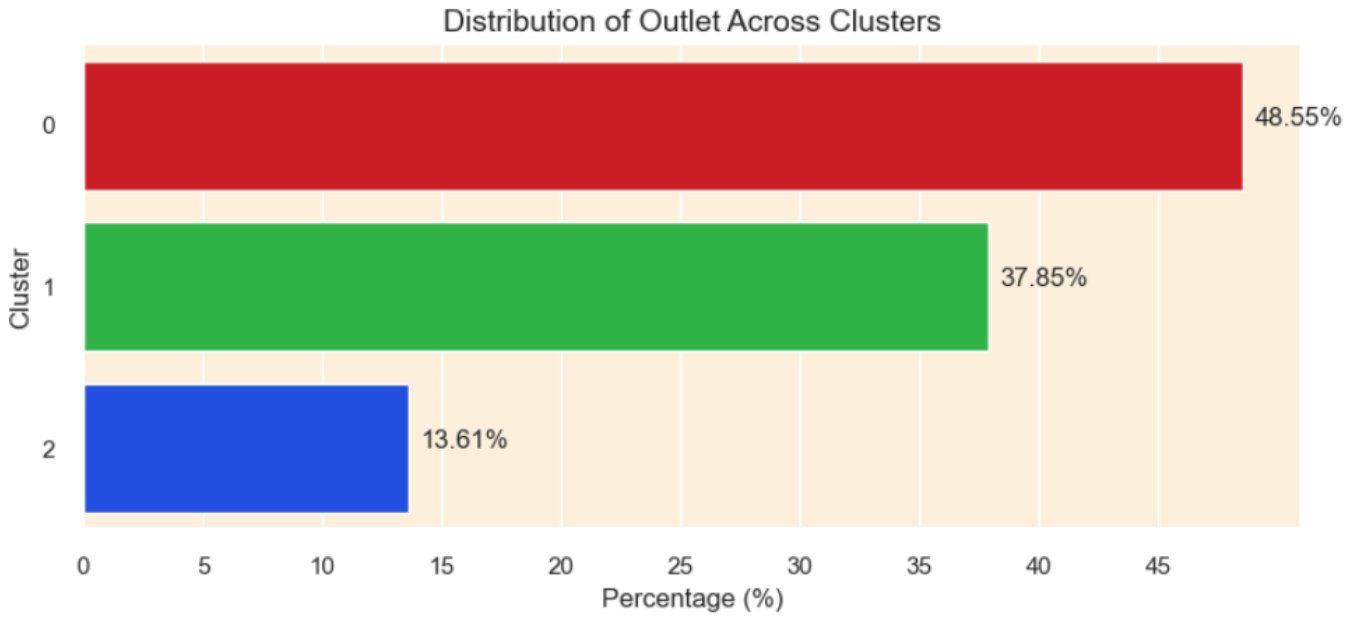












To further examine the quality of the cluster, I will use the following metrics:

- Silhouette Score: A measure to evaluate the separation distance between clusters. A higher value indicates better cluster separation. This value ranges from -1 to 1.
- Calinski Harabasz Score: This score is used to evaluate the dispersion between and within clusters. Higher values indicate better clusters.
- Davies Bouldin Score: This score assesses the average similarity between each cluster and the most similar cluster. Lower values indicate better cluster separation.

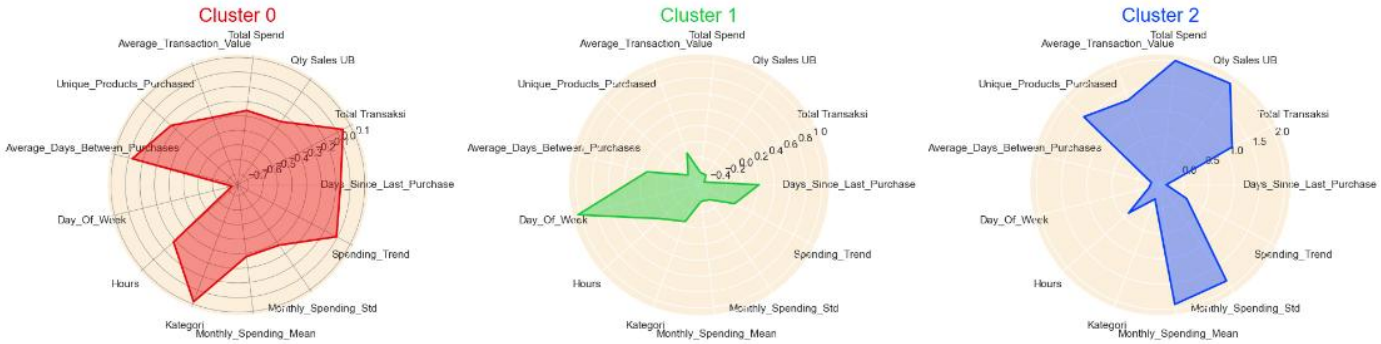
Metric	Value
Number of Observations	11245
Silhouette Score	0.27672537411716236
Calinski Harabasz Score	4187.411511378494
Davies Bouldin Score	1.2869765745957105

Conclusion on clustering quality

- The Silhouette score is around 0.27, although not close to 1, it still indicates a fairly good separation between the clusters. This suggests that the clusters are somewhat distinct, but there may be little overlap between them. In general, a score close to 1 is ideal, indicating more distinct and well-separated clusters.
- The Calinski Harabasz score is 4187.41, which is a very high score, indicating that the clusters are well-defined. Higher scores in this metric generally signify better cluster definition, thus implying that this clustering has successfully discovered substantial structure in the data.
- A Davies Bouldin score of 1.28 is a reasonable score, indicating a moderate level of similarity between each cluster and the most similar cluster. A lower score is generally better as it indicates less similarity between clusters, thus the score here indicates a decent separation between clusters.

Finally, these metrics have shown that the clusters are of very good quality and sufficiently separated. However, there may still be room for further optimization to potentially improve cluster separation and definition by trying other clustering or dimensionality reduction algorithms.

8. Cluster Analysis & Profilling

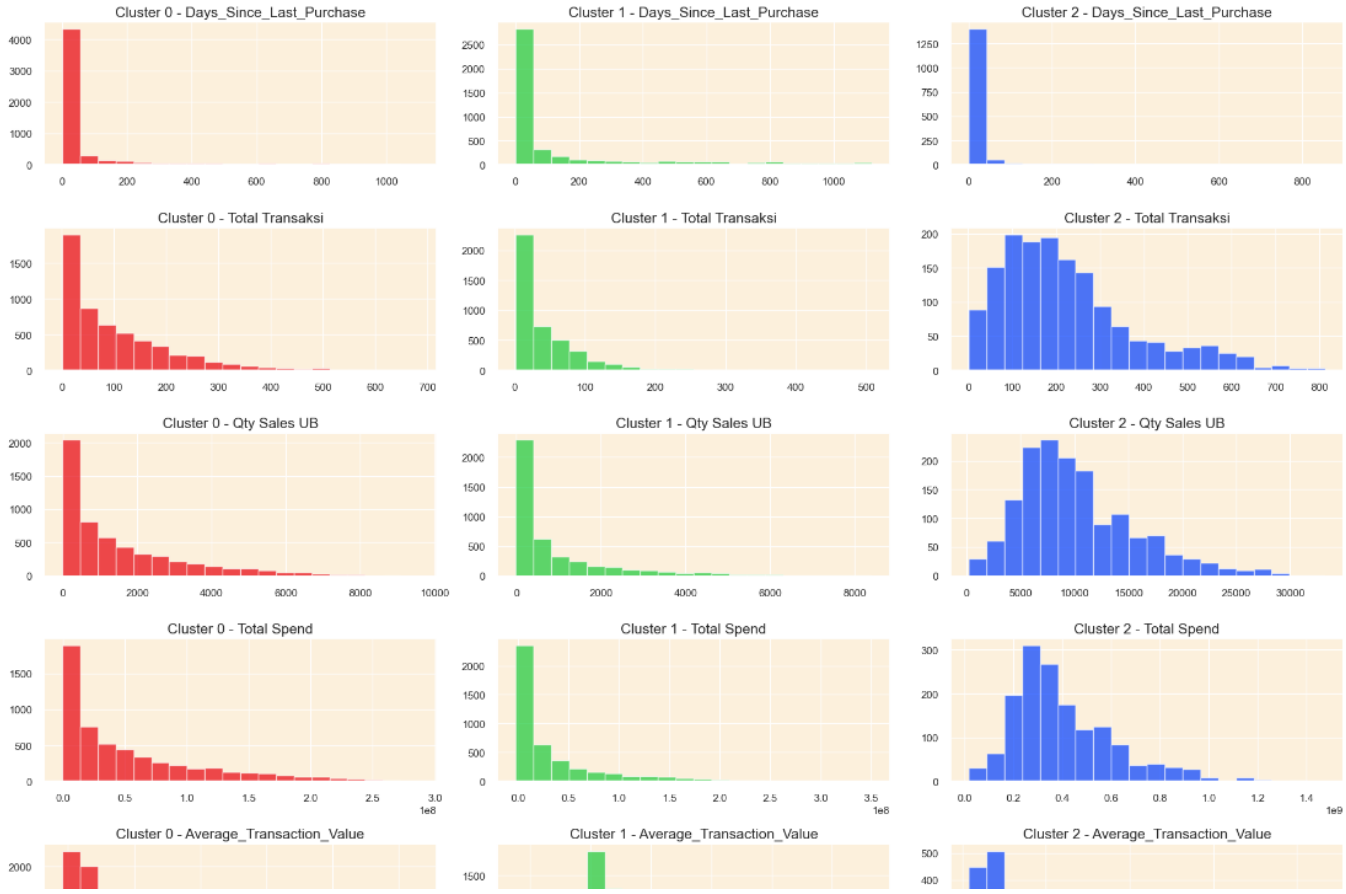


Outlet Profiles Derived from Radar Chart Analysis

1. Cluster 0 (Red Chart): 🎯 **Profile: Active Shopping Nexus**
   - Outlets in this cluster tend to shop fairly large and actively make transactions with a fairly varied range of products purchased.
   - These outlets have a tendency to shop on weekdays, as shown by the very low Day_of_Week value.
   - The outlet's spending trend is relatively very high, and they have a large variation in monthly spending (Monthly_Spending_Std is quite high).
   - Judging from the high association of (average days between) and (days since last), it shows that outlets in this cluster often purchase products with a variety of transactions and varying purchase hours.
   - This outlet has an average transaction with a fairly large quantity, and has a good total spend.

2. Cluster 1 (Green Chart): 🎯 **Profile: Weekend Warriors**
   - Outlets in this cluster tend to be inactive shoppers and make transactions with only a few products that they purchase (less variety).
   - These outlets have a tendency to shop on weekends only, as shown by the very high Day_of_Week value.
   - The outlet's spending trend is relatively low, and they have a low monthly spending variation (Monthly_Spending_Std).
   - Judging from the relationship between (average days between) and (days since last) which is small, this shows that the outlets in this cluster buy our products with transactions that are relatively ordinary, not too varied by making purchases that can be said to be rare.
   - This outlet has a small average transaction with a very small quantity, and has a poor total spend.
   -

3. Cluster 2 (Blue Chart): 🎯 **Profile: Super Shopper Sanctuary**
   - Outlets in this cluster tend to make very large purchases and are quite active in making transactions with many of the products purchased being quite varied.
   - These outlets have a tendency to shop on weekdays, as shown by the very low Day_of_Week value.
   - The outlet's spending trend is relatively small (this may be due to their exponential way of shopping), indicated by transactions that are not high but tend to be normal, and they have a very large monthly spending variation (Monthly_Spending_Std is very high).
   - Judging from the relationship between (average days between) and (days since last) which are very low, this reinforces point 3 that the outlets in this cluster make exponential purchases only a few times.
   - This outlet has an average of large transactions with a very high quantity, and has a very high total spend.



9. Result - Recommendation System Product

	Nama Outlet	cluster	Rec1_Procode	Rec1_Prodesc	Rec2_Procode	Rec2_Prodesc	Rec3_Procode	Rec3_Prodesc
0	1000 SEHAT, APT	0	201630.0	MICROLAX GEL OBAT PENCAHAR 5ML	202935.0	SANMOL PARACETAMOL SIROP 60ML	102055.0	NEBACETIN OINT 5GR
3	168, APT	0	202572.0	PRAXION SUSP 60ML RASA JERUK	108299.0	ALBOTHYL CONCENTRATE 5ML	NaN	None