Clustering Algorithms Comparison

1. Quantitative Results

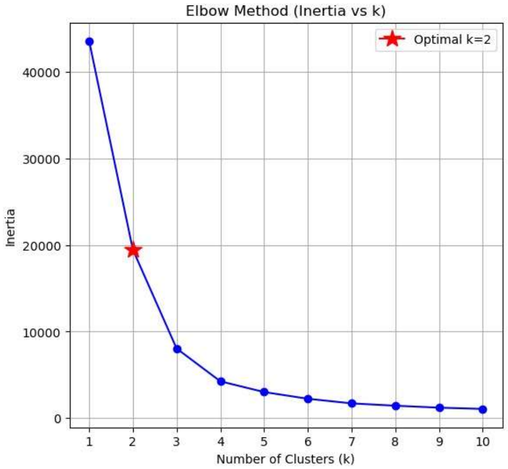
The models were evaluated using the **Silhouette Score**, which measures cluster cohesion and separation (ranges from -1 to 1, where 1 is best)

Aspect	K-Means	DBSCAN
Number of Clusters	2 (Predefined via Elbow Method for Buy/No-Buy)	2 (Determined automatically based on density)
Model Accuracy (Silhouette Score)	**0.74**	**0.76**
Handling Behavioral Patterns	Good (Separates based on average distances)	Excellent (Isolates high-density behavioral groups)

2. Detailed Analysis

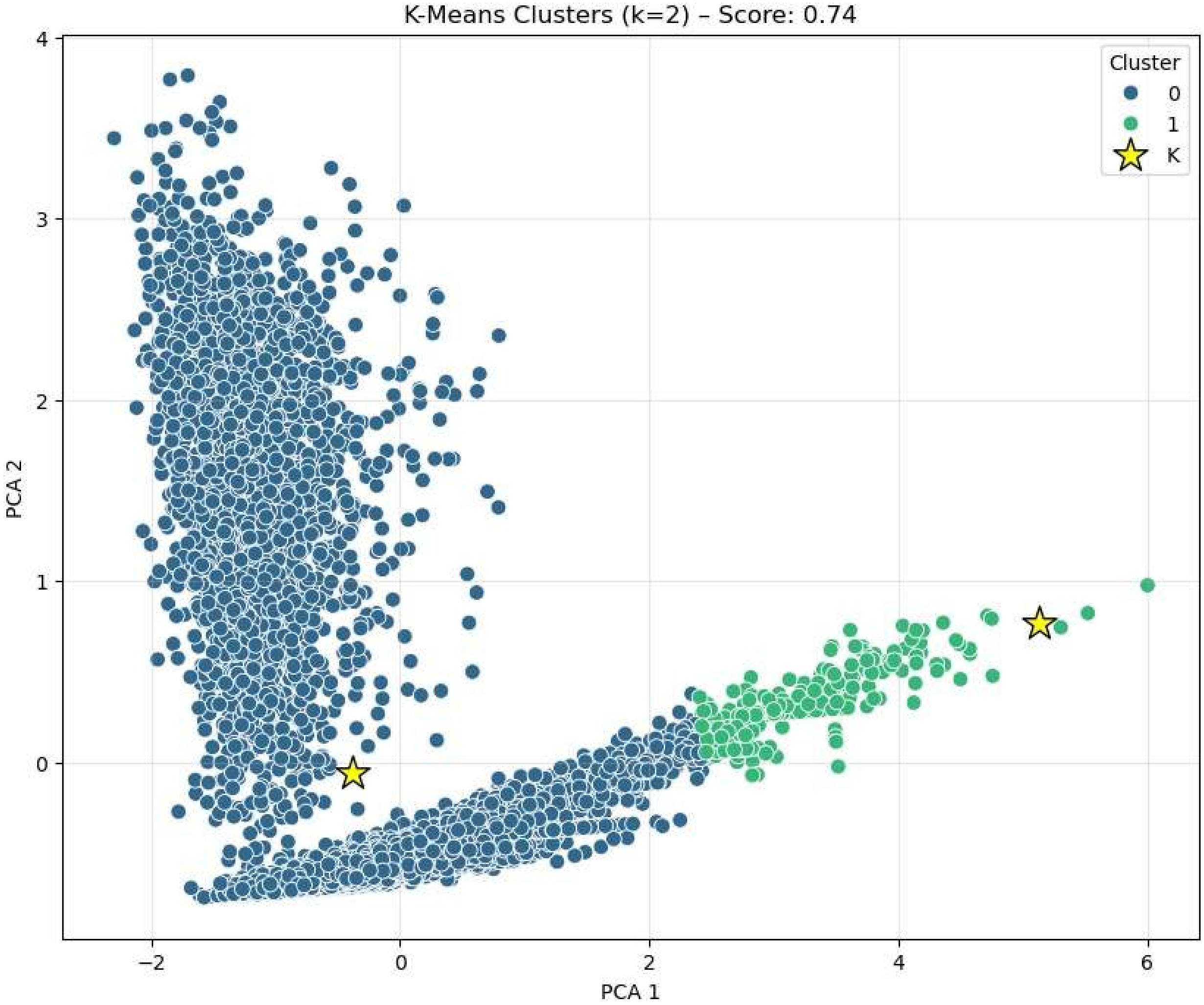
K-Means

- **Performance:** Achieved a high Silhouette Score of **0.74**. The algorithm successfully divided visitors into two main groups: "Potential Buyers" (High Page Values) and "Window Shoppers".



- **Notes:** K-Means forced every data point into one of the two clusters based on distance to the centroid. While accurate, it assumes clusters are spherical and of similar size, which might oversimplify complex visitor behaviors.

- **Visualization:** Shows a clear linear separation between the two groups, largely driven by the 'Page Values' feature.

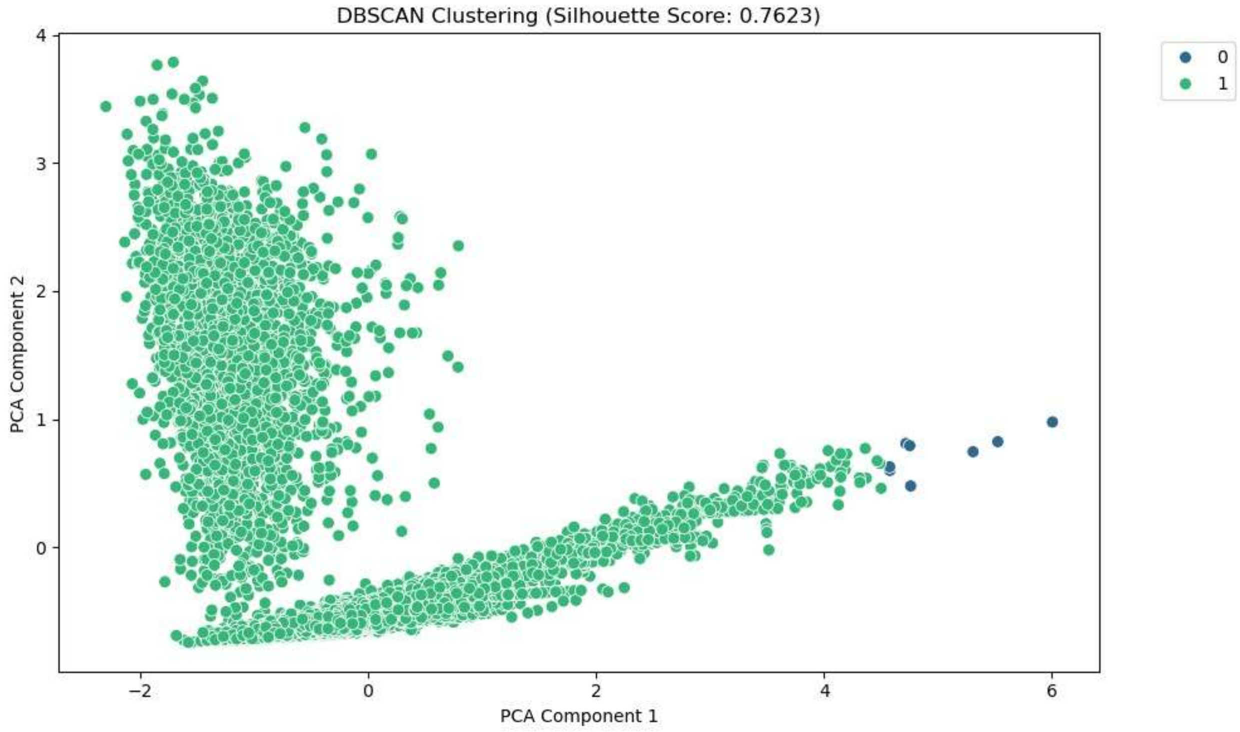


DBSCAN

- **Performance:** Achieved a higher Silhouette Score of **0.76**, slightly outperforming K-Means.
- **Notes:** DBSCAN relied on density. It identified a massive "Core" cluster of general visitors and a distinct, smaller cluster of specific behavior (visitors with extremely

high bounce/exit rates). Unlike K-Means, it didn't just split the data in half; it found the "dense" behavioral patterns naturally.

- **Visualization:** Shows very sharp boundaries between the clusters, effectively isolating the "Quick Exit" visitors from the "Engaged" ones.



3. Conclusion & Recommendation

Recommended Model: DBSCAN

Reasons:

1. **Higher Accuracy:** It achieved the highest Silhouette Score (**0.76** vs 0.74), indicating that the clusters formed are mathematically more distinct and cohesive.
2. **Behavioral Insight:** DBSCAN automatically detected the natural structure of the data (e.g., distinguishing between a large mass of normal traffic and a specific density of bounced visits) without needing us to specify $K=2$ beforehand.
3. **Flexibility:** It is more robust for real-world data where visitor groups might not be circular or equal in size, making it a safer choice for future data updates.