Clustering Report on Customer Data

The objective of this clustering analysis was to segment customers into distinct groups based on their purchasing behaviors and demographics. The dataset, which combines customer transaction details, was analyzed using K-Means clustering, and the optimal number of clusters was determined based on the Davies-Bouldin (DB) Index.

Number of Clusters Formed:

The analysis was conducted for a range of cluster sizes, from 2 to 10, using the K-Means algorithm. The optimal number of clusters was found to be 3, as determined by the lowest Davies-Bouldin Index score.

Davies-Bouldin Index:

The Davies-Bouldin Index, a measure of clustering quality where lower values indicate better clustering, was calculated for each number of clusters. The DB Index values for cluster sizes ranging from 2 to 10 were:

```
Davies-Bouldin Index for each k: [1.2339267409837922,
1.1940216453410988, 1.273772372822729, 1.2252455816464418,
1.2510383370363571, 1.3434001240152813, 1.3375262993835484,
1.364194852227647, 1.2861113329279972]
```

- DB Index for 2 clusters: 1.23
- DB Index for 3 clusters: 1.19 (Optimal)
- DB Index for 4 clusters: 1.27
- DB Index for 5 clusters: 1.23
- DB Index for 6 clusters: 1.25
- DB Index for 7 clusters: 1.34
- DB Index for 8 clusters: 1.34
- DB Index for 9 clusters: 1.36
- DB Index for 10 clusters: 1.29

The lowest DB score was achieved for 3 clusters, indicating that this configuration provides the best separation of customer groups.

Other Clustering Metrics:

1. **Cluster Distribution**: After applying K-Means with 3 clusters, the dataset was segmented into three customer groups. The distribution of customers across

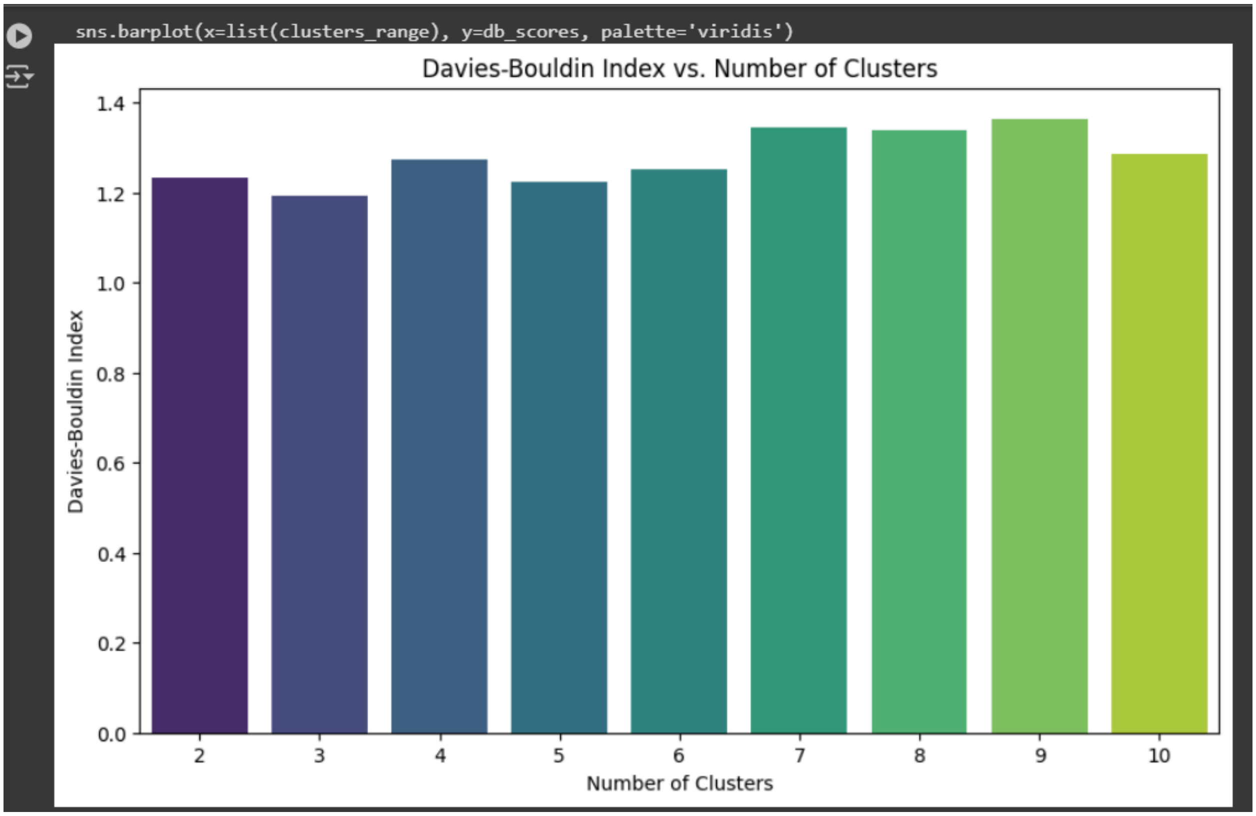
clusters was visualized, revealing the relative sizes of the groups. The distribution of customers per cluster is shown below:

   - Cluster 0: 66
   - Cluster 1: 73
   - Cluster 2: 60

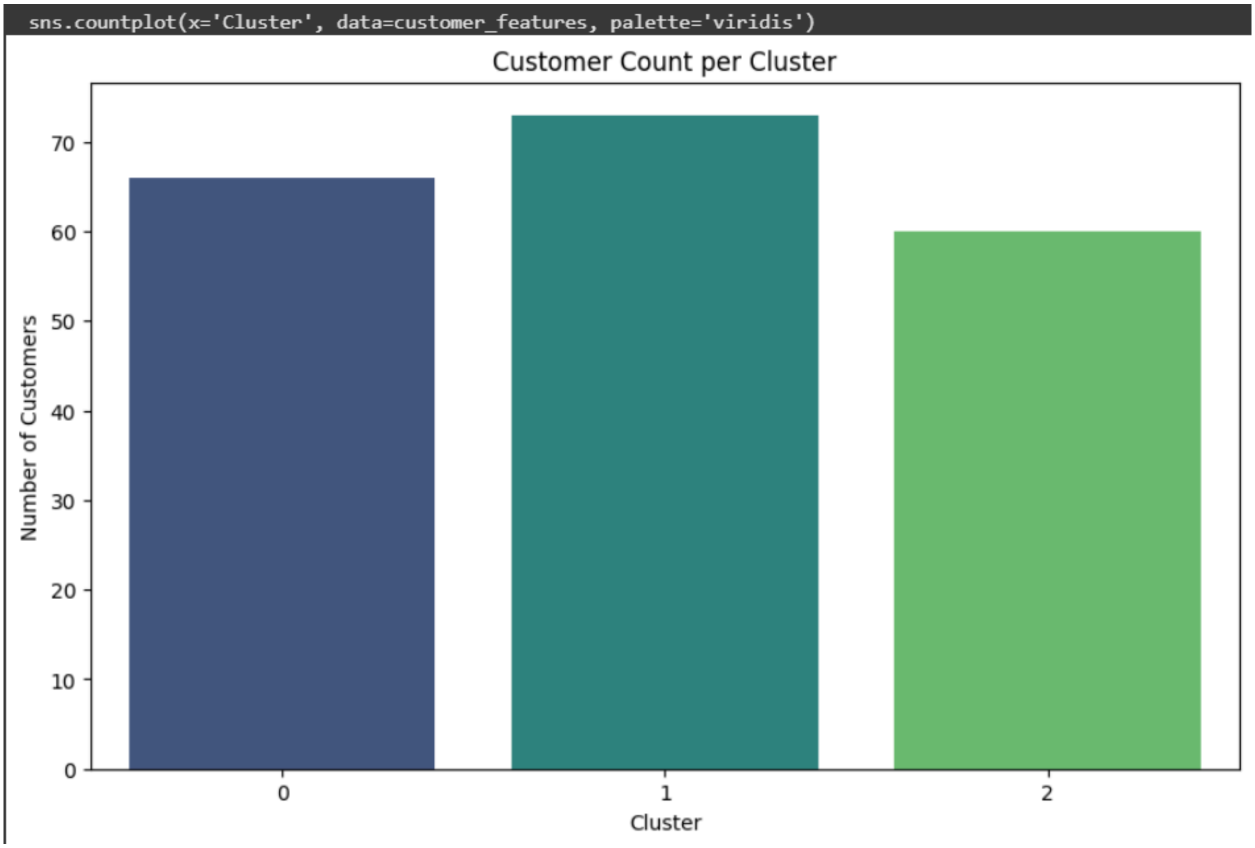
2. **PCA Visualization**: To better understand the customer segmentation, a 2D visualization of the clusters was created using Principal Component Analysis (PCA). This dimensionality reduction technique helped visualize the clusters in two principal components, showing distinct separation between the customer groups.
3. **Feature Analysis**: The clustering model was based on various customer features, including total spending, total quantity purchased, and days since account signup. Categorical variables, such as the customer's region, were one-hot encoded to ensure the model used all available features effectively.

Conclusion:

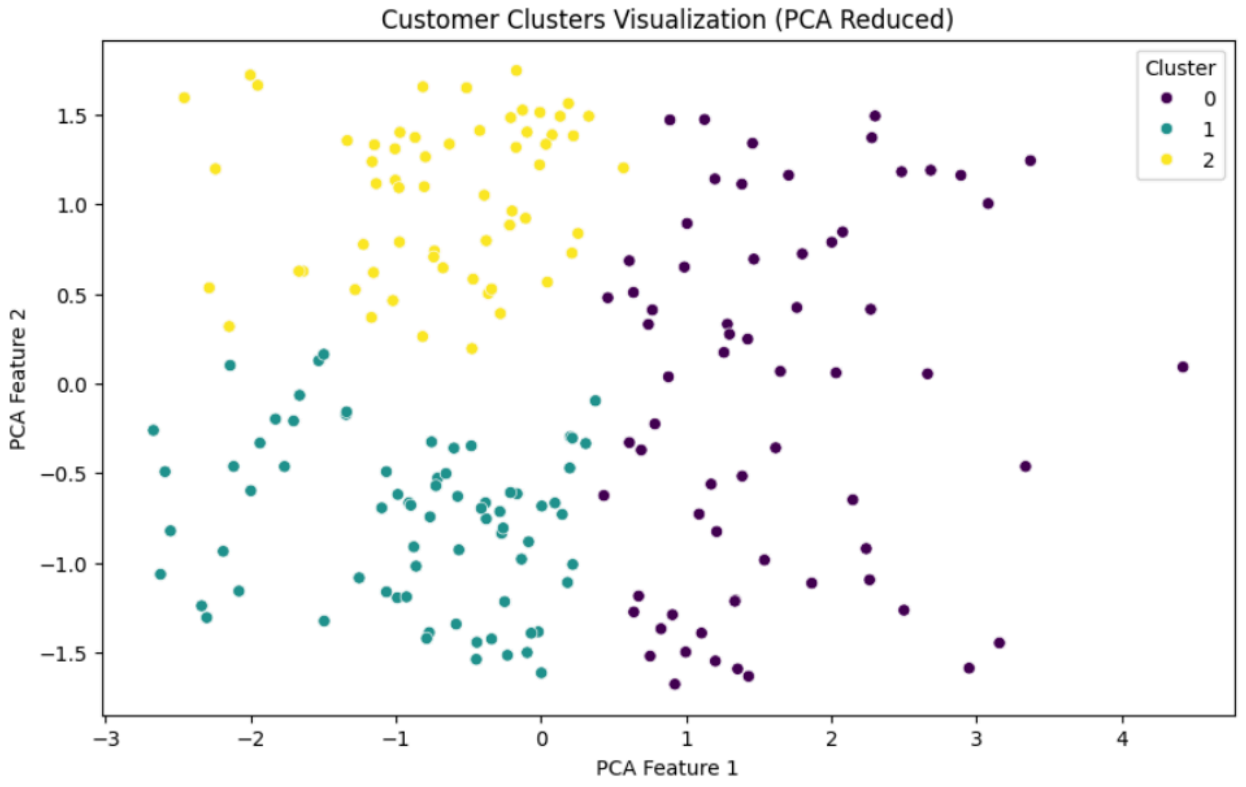
The optimal segmentation of customers, based on the Davies-Bouldin Index and the K-Means algorithm, yielded 3 clusters. This clustering offers meaningful insights into customer behavior, enabling businesses to tailor strategies for different customer segments. The use of PCA and other visualizations also aids in interpreting these clusters in a more actionable way.



```
sns.barplot(x=list(clusters_range), y=db_scores, palette='viridis')
sns.countplot(x='Cluster', data=customer_features, palette='viridis')
```



Customer Count per Cluster



Customer Clusters Visualization (PCA Reduced)