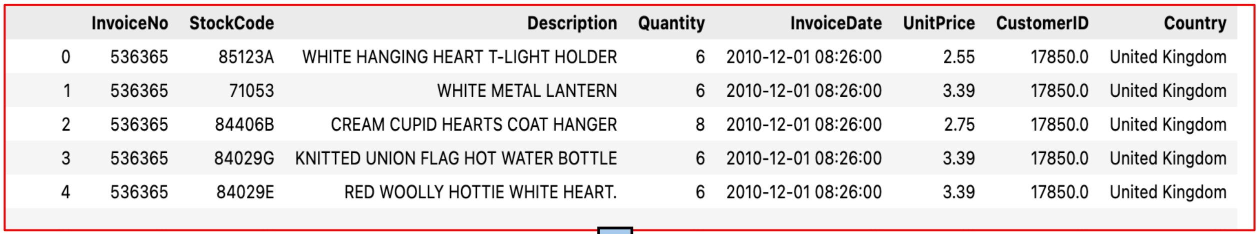
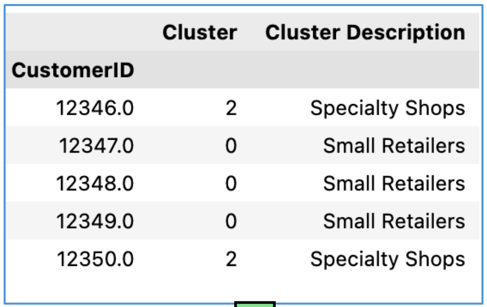
# CUSTOMER SEGMENTATION OF ONLINE RETAIL PURCHASES

## ORIGINAL DATA:



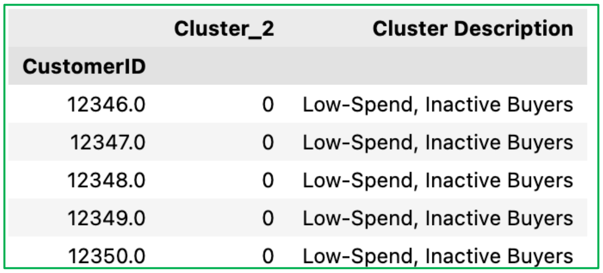
| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|
| 0 | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 2010-12-01 08:26:00 | 2.55 | 17850.0 | United Kingdom |
| 1 | 536365 | 71053 | WHITE METAL LANTERN | 6 | 2010-12-01 08:26:00 | 3.39 | 17850.0 | United Kingdom |
| 2 | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 2010-12-01 08:26:00 | 2.75 | 17850.0 | United Kingdom |
| 3 | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 2010-12-01 08:26:00 | 3.39 | 17850.0 | United Kingdom |
| 4 | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 2010-12-01 08:26:00 | 3.39 | 17850.0 | United Kingdom |

## CLUSTER 1



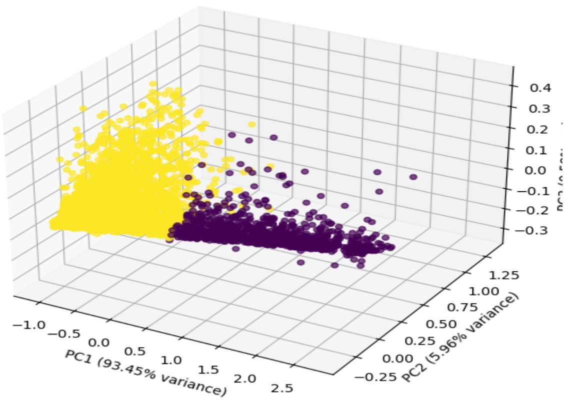
| | Cluster | Cluster Description |
|---|---|---|
| CustomerID | | |
| 12346.0 | 2 | Specialty Shops |
| 12347.0 | 0 | Small Retailers |
| 12348.0 | 0 | Small Retailers |
| 12349.0 | 0 | Small Retailers |
| 12350.0 | 2 | Specialty Shops |

## CLUSTER 2



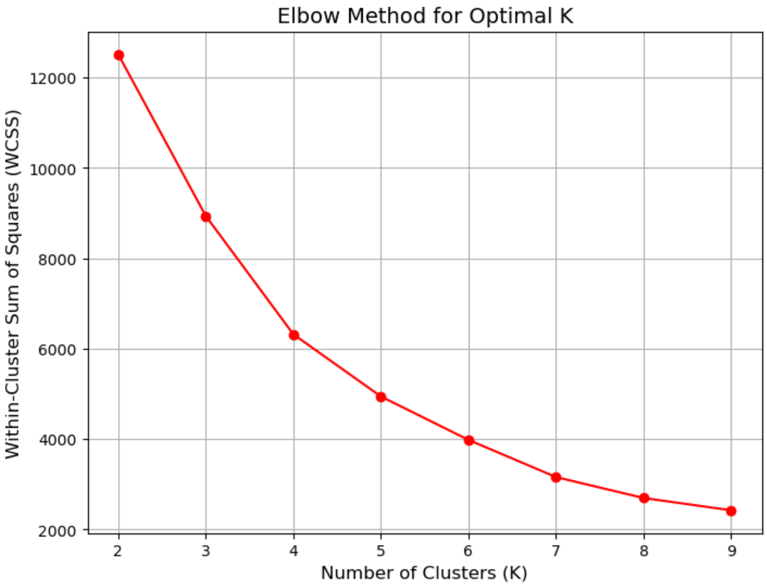
| | Cluster_2 | Cluster Description |
|---|---|---|
| CustomerID | | |
| 12346.0 | 0 | Low-Spend, Inactive Buyers |
| 12347.0 | 0 | Low-Spend, Inactive Buyers |
| 12348.0 | 0 | Low-Spend, Inactive Buyers |
| 12349.0 | 0 | Low-Spend, Inactive Buyers |
| 12350.0 | 0 | Low-Spend, Inactive Buyers |

3D PCA-reduced Data (k=2)

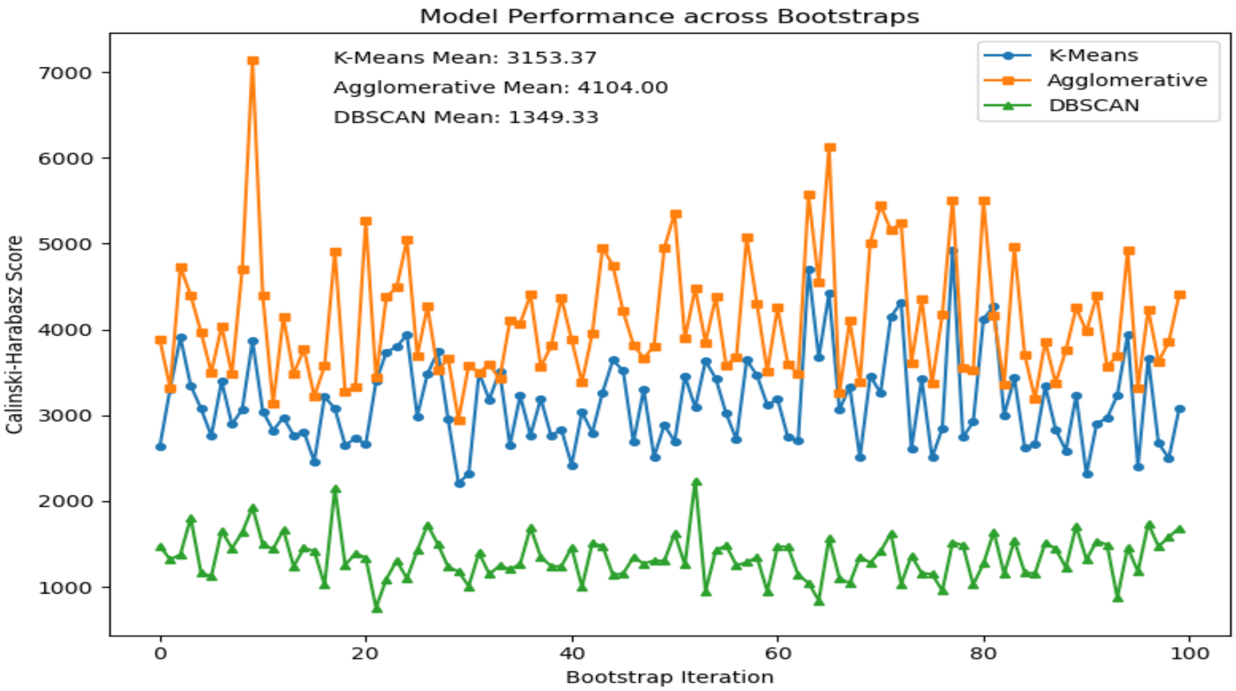


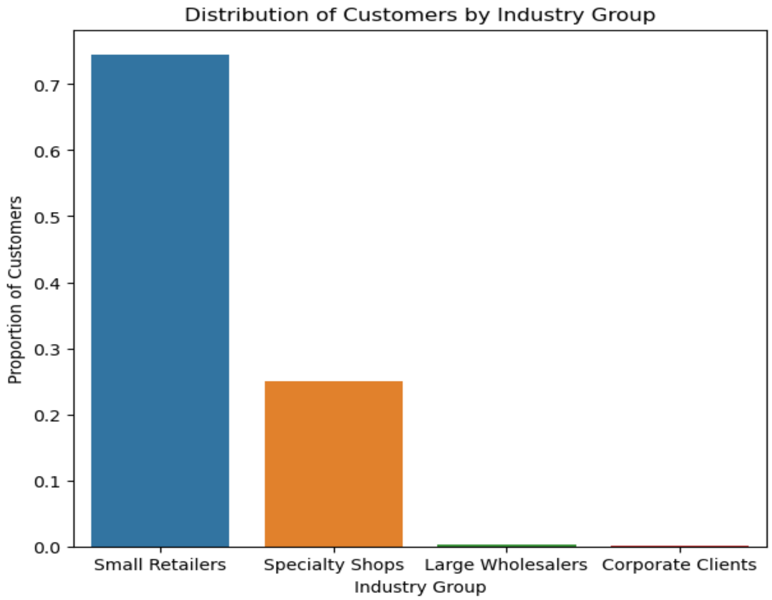
immanueltettehtsu@gmail.com --- https://github.com/EOsamau

The elbow appears to be at K = 4. Up until K=4, the WCSS decreases significantly, but after that, the reduction in WCSS becomes more gradual. An approximate K for the data is 4.



While the Agglomerative algorithm achieved the highest Calinski-Harabasz score, its mean is fairly comparable to K-Means, which uses fewer clusters. Given that K-Means offers better interpretability and simplicity, it may be a preferable despite the slight trade-off in score across 100 Bootstraps.

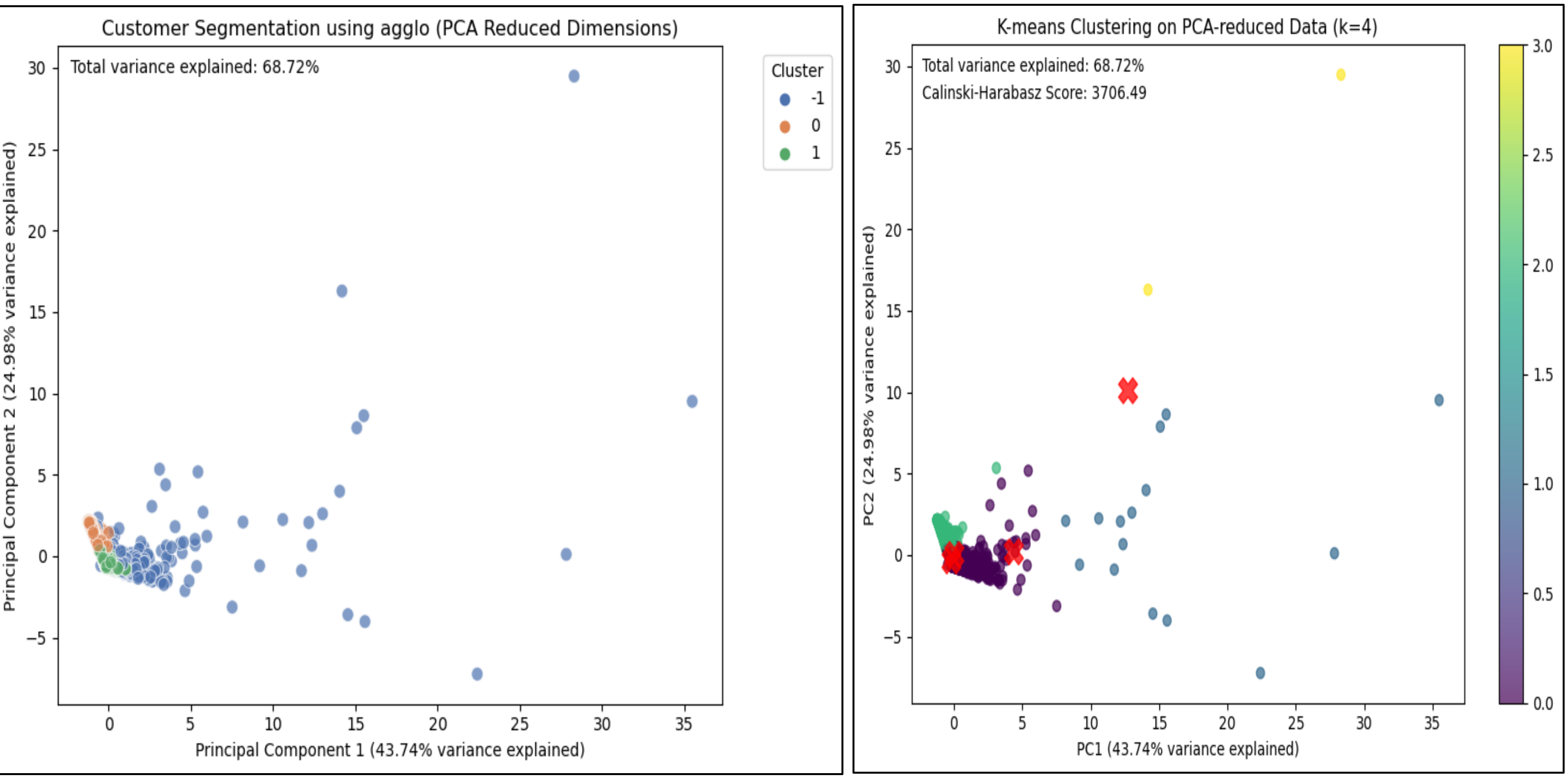




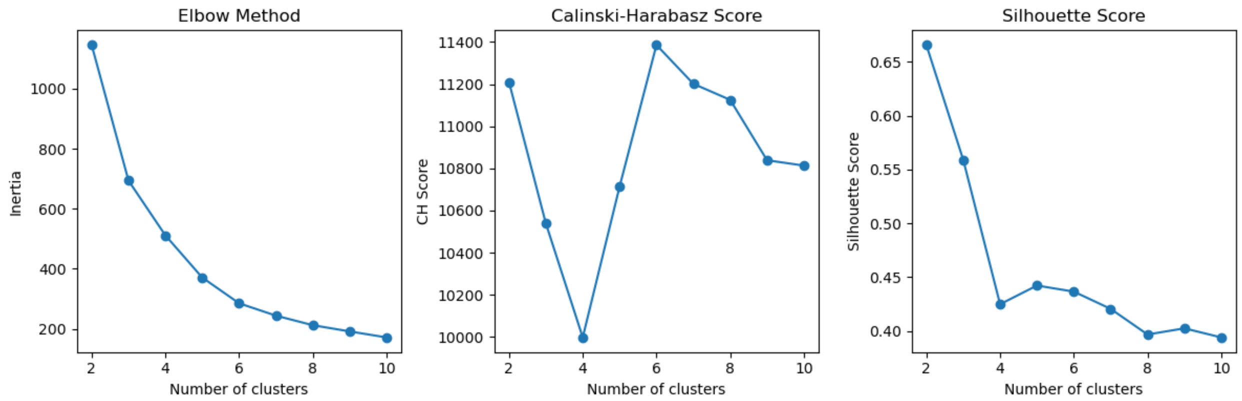
**Cluster 0: Small Retailers:** Customers in this cluster buy a moderate quantity of goods (1123 units on average) at a relatively low unit price (4.09). Their frequency of purchases is high, and they spend a significant amount overall $3,771.

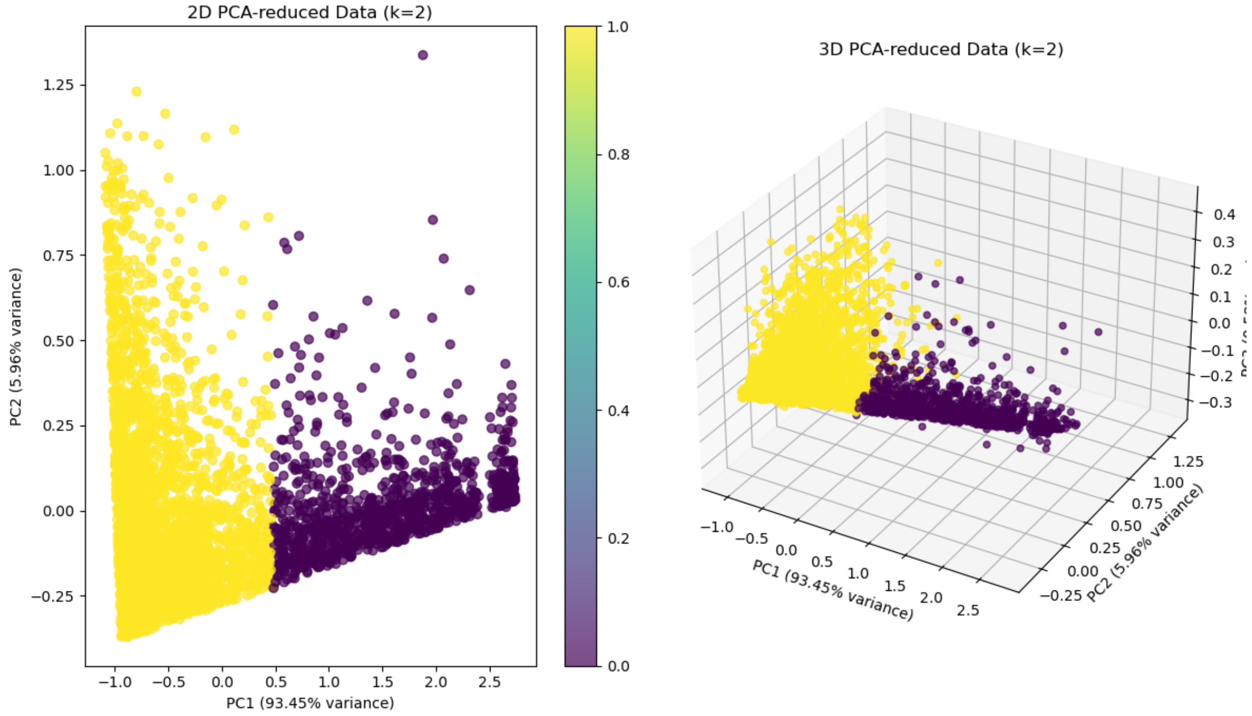
**Cluster 1: Large Wholesalers:** This group has a significantly high total spend (~ $251,969) and makes massive bulk purchases (60,307 units on average), which suggests these are likely large wholesale distributors. They operate with relatively low recency (3.73 days), indicating frequent transactions, typical of large wholesale operations.

**Comparing the Best Algo (Agglomerative – 9 Clusters) and K-Means (K=4) in a 2D PCA:**



The overlap between the purple and green clusters suggests that these might not be truly distinct clusters in the original high-dimensional space. Therefore, we can experiment increasing the dimensions to a 3D-PCA from a 2D-PCA and selecting the number of clusters with across the silhouette and the Calinski HB Scores.





The recent analysis suggests 2 clusters (k=2) as optimal, compared to the previous 4 clusters for K-means. The new 2D and 3D PCA plots show a clearer separation between two main groups with minimal overlap. The total variance explained by the first two components is now 98.81%, compared to 68.72% previously.