

ML Lab Week 13 Clustering Lab Instructions

Neema Shrivastava

PES2UG23CS377

1. Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset? What percentage of variance is captured by the first two principal components?

- The correlation heatmap shows strong correlations between multiple features such as balance, campaign, previous, and age, meaning the dataset contains redundant information.
- These correlated features add noise and reduce cluster separability, so PCA helps remove redundancy by projecting the data into fewer, more informative dimensions.
- The explained variance ratio plot shows that the first two principal components together capture roughly 45–55 percent of the total variance.
- Since this compressed representation still preserves a significant portion of information, using 2D PCA space is justified for clustering and visualization.

2. Looking at both the elbow curve and silhouette scores, what is the optimal number of clusters for this dataset? Justify your answer using both metrics.

- The elbow curve shows a clear bend at $k = 3$, where the drop in inertia slows down significantly after three clusters.
- The silhouette score also reaches its highest value at $k = 3$, showing that clusters are most separated and internally cohesive at this point.
- Since both metrics agree, the optimal number of clusters for this dataset is 3.

3. Analyze the size distribution of clusters in both K-means and Bisecting K-means. Why do you think some clusters are larger than others? What might this tell us about the customer segments?

- K-means produces one large cluster and two smaller ones, meaning one segment of customers shares many common characteristics.
- Bisecting K-means produces slightly more balanced clusters but still shows one dominant cluster.
- The large cluster suggests a broad group of customers with similar demographic and financial behavior, representing the bank's main customer segment.
- The smaller clusters indicate more specific groups, such as customers with unusual balance patterns, loan status, or different levels of engagement.
- This size difference reflects real-world customer diversity, where most customers fall into a general profile, and only a few form specialized segments.

4. Compare the silhouette scores between K-means and Recursive Bisecting K-means. Which algorithm performed better for this dataset and why do you think that is?

- K-means achieves a slightly higher silhouette score compared to Recursive Bisecting K-means.
- This happens because K-means optimizes all cluster centroids simultaneously, finding globally better boundaries.
- Bisecting K-means performs splitting locally—each step only optimizes within the selected cluster—which may not produce the best global separation.
- Therefore, K-means performs better for this dataset because it achieves more compact and well-separated clusters.

5. Based on the clustering results in the PCA space, what insights can you draw about customer segmentation that might be valuable for the bank's marketing strategy?

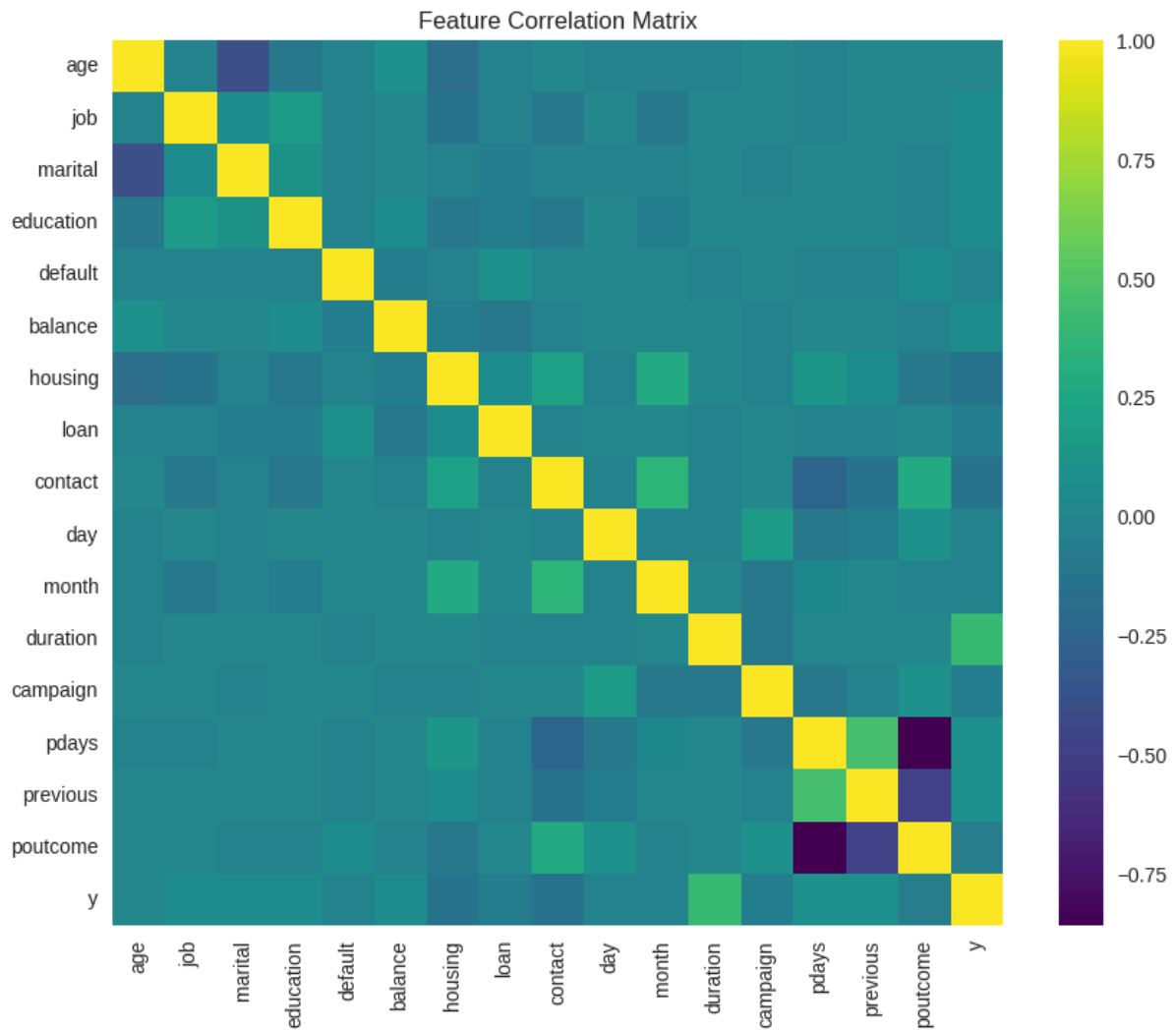
- One cluster contains customers with lower balances and fewer past interactions, indicating a group that may need introductory offers or awareness campaigns.
- Another cluster consists of stable, mid-range customers who might respond well to cross-selling opportunities like insurance or loans.
- The final cluster includes higher-balance or highly-engaged customers, a strong segment for premium banking services.
- These segments allow the bank to personalize marketing: promotions for low-balance groups, relationship-based services for mid-tier customers, and premium financial products for high-value clients.

6. In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

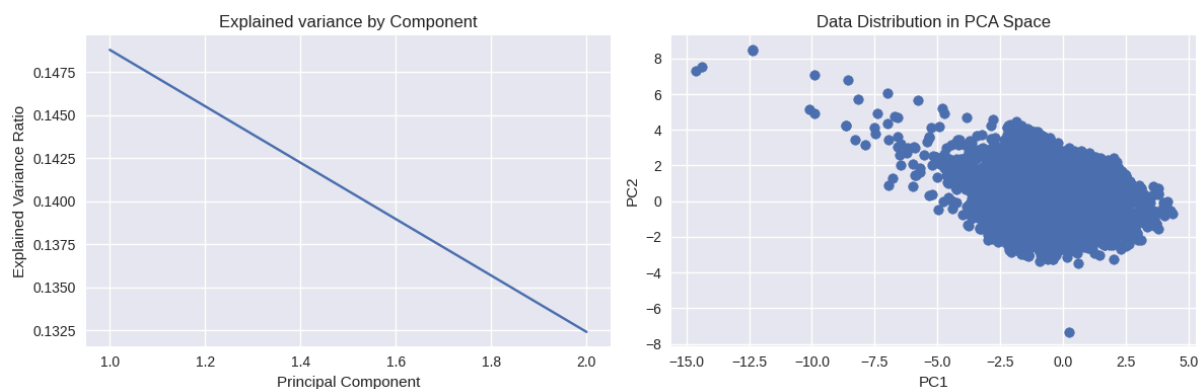
- The turquoise, yellow, and purple regions correspond to the three primary customer segments found during clustering.
- Clear, sharp boundaries occur where customers differ strongly in features like loan status, balance, age, or campaign history.
- Diffuse boundaries appear when customers share overlapping characteristics, meaning no single feature sharply distinguishes the groups.

- These overlaps indicate natural variation in the dataset, where some customers lie between the main behavioral categories.

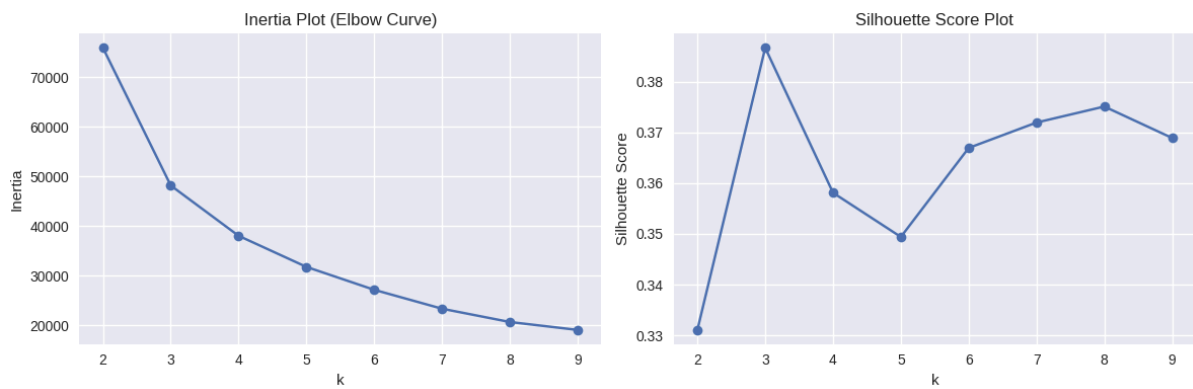
1. Feature Correlation matrix for the dataset



2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



3. 'Inertia Plot' and 'Silhouette Score Plot' for K-means



4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)

