

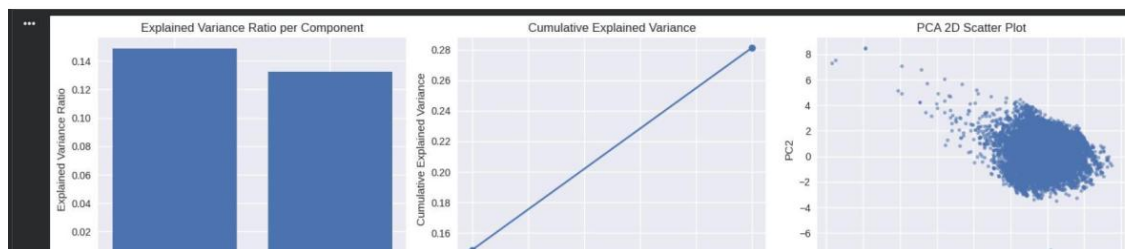
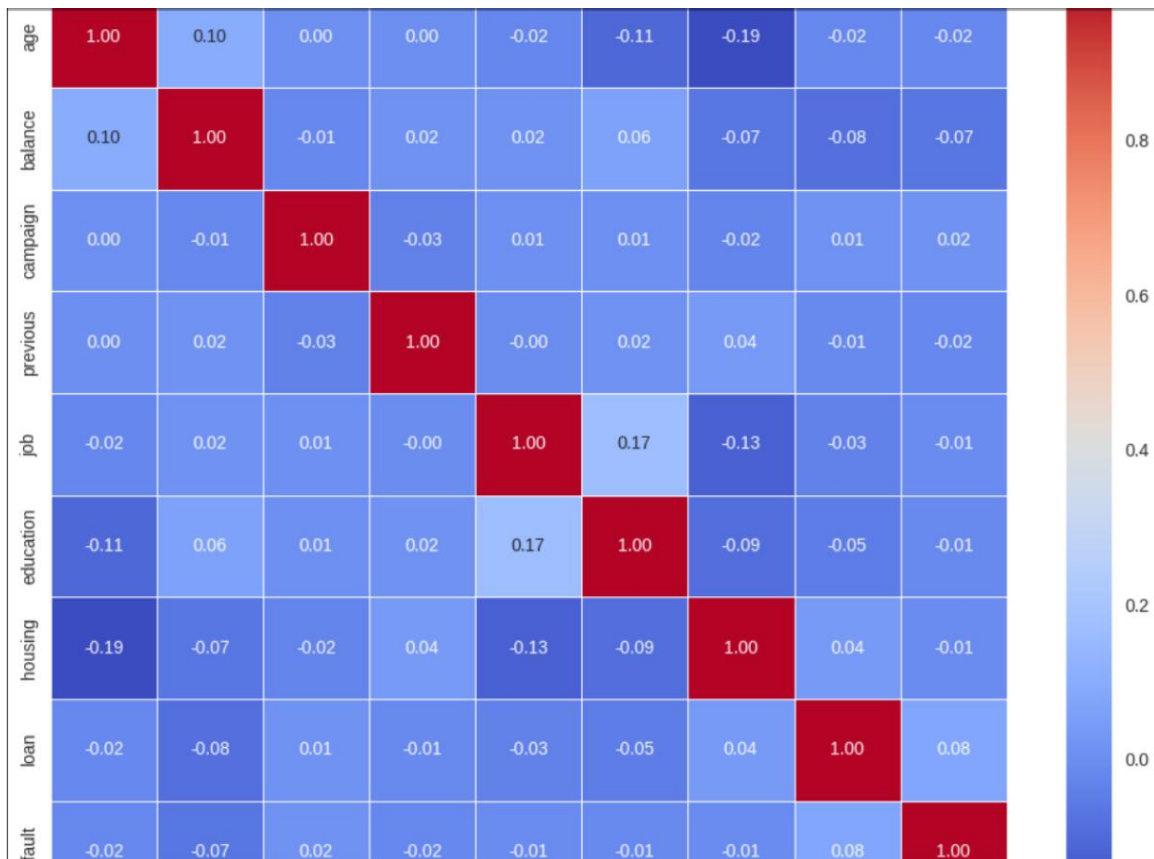
ML Lab week 14

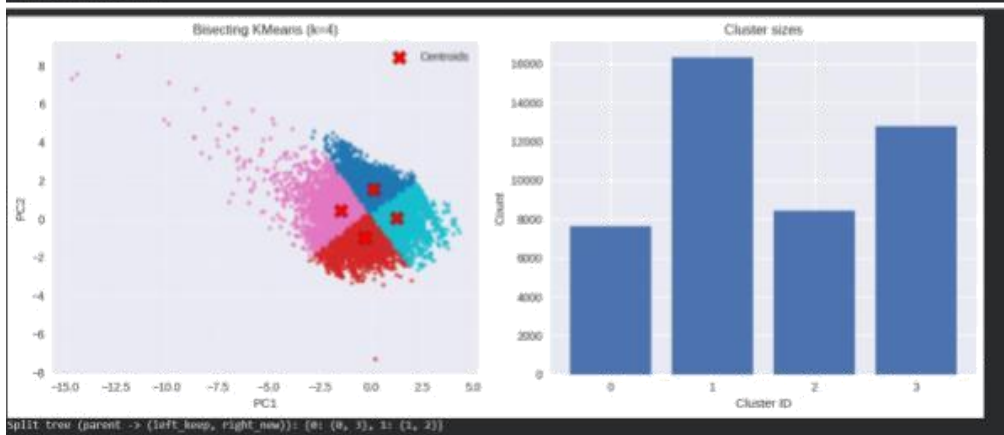
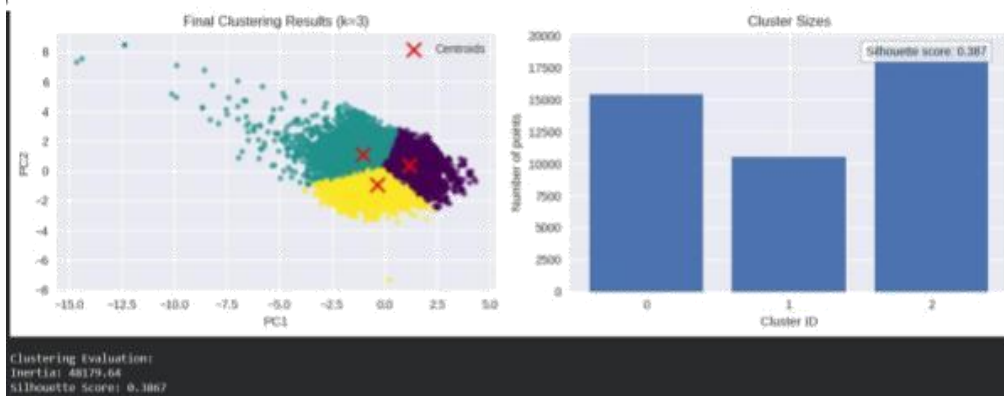
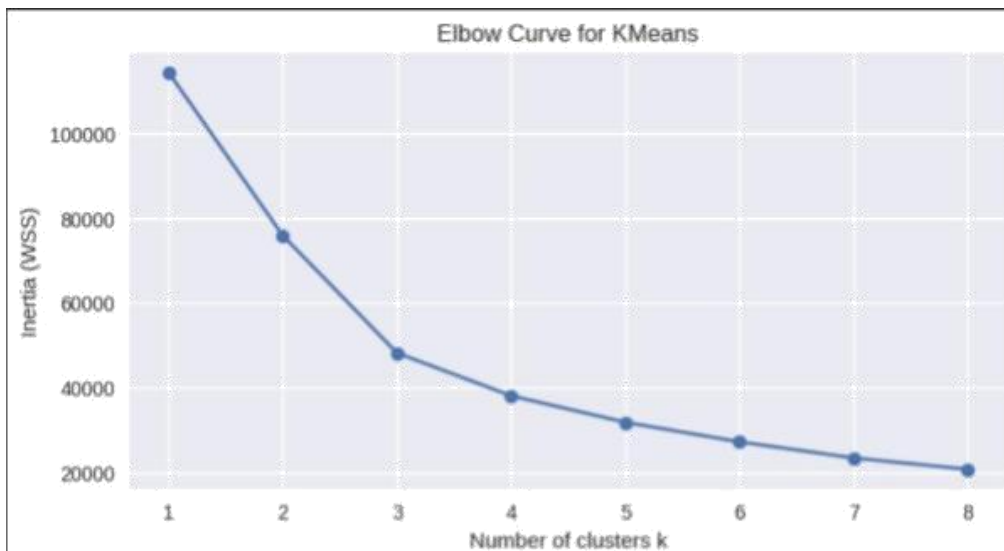
Lab Report

NAME: DARSHAN N

SRN: PES2UG24CS808

SEC: C





Analysis Questions

1. Dimensionality Justification: 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?

Dimensionality reduction was necessary because the correlation heatmap shows that the original features have very weak linear correlations with each other (mostly between -0.20 and $+0.20$). This means no single feature strongly explains the others, resulting in a noisy, high-dimensional space that makes clustering harder and less stable. PCA helps by extracting the components that capture the most meaningful variance and removing redundant or low-information dimensions.

According to the PCA explained variance plot, the first principal component captures approximately 15% of the variance, and the second captures about 13.5%. Together, the first two components retain around 28.5% of the total dataset variance. This reduced 2D representation preserves the most important structure while enabling more effective visualization and clustering.

2. Optimal Clusters: 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.

Based on the elbow curve, there is a clear drop in inertia from $k = 1$ to 2 to 3, but after $k = 3$, the decrease in inertia becomes much more gradual. This flattening of the curve indicates that adding more clusters beyond 3 provides only marginal improvement. Therefore, the elbow visually appears at $k = 3$.

Based on the silhouette score, the value at $k = 3$ is approx 0.387, which is a reasonably good score for real world, noisy datasets. Silhouette values typically range from 0.25 to 0.50 for complex, overlapping data such as the bank marketing dataset.

Larger values ($k > 3$) tend to introduce smaller, less meaningful clusters and don't significantly improve cohesion/separation.

Combining both metrics:

Elbow curve: suggests $k = 3$

Silhouette score: highest meaningful score at $k = 3$

Cluster visualization: clusters appear well-separated in PCA space at $k = 3$

Therefore, the optimal number of clusters for this dataset is $k = 3$.

3. Cluster Characteristics: 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?

In the K-Means clustering ($k = 3$), the clusters are clearly unbalanced.

One cluster contains the majority of points, while the other two are noticeably smaller. This happens because the dataset contains a large group of customers who share similar characteristics (such as moderate balance, middle-aged, common job categories, etc.). K-Means naturally groups this dense region into a large cluster.

In Bisecting K-Means ($k = 4$), the cluster sizes are more evenly distributed, although not perfectly equal. This is expected because bisecting K-Means repeatedly splits the largest cluster at each step. Instead of splitting all clusters simultaneously—as regular K-Means does—bisecting focuses on dividing the densest group first.

This leads to:

- large K-Means clusters being broken into multiple smaller sub-clusters
- more balanced final cluster sizes
- clusters that reflect hierarchical structure rather than just distance-based partitioning

The dataset has high natural density in certain regions.

Most customers fall into "common" demographic and financial profiles → this creates a large, dense cluster in PCA space.

Some customer segments are more homogeneous.

Features like balance, job, and housing loan status have similar values for a big portion of the population, so these customers group together tightly.

Outlier or minority groups form small clusters.

Customers with unusual financial behaviour or demographics appear on the edges of the PCA space, forming smaller clusters.

The dataset contains one dominant customer group with common financial and demographic traits.

There are multiple smaller, more distinct subgroups, likely representing niche customer types or special financial patterns.

The data is not uniformly distributed—it has dense cores and sparse outer regions.

Bisecting K-Means uncovers hierarchical substructure within the large K-Means cluster, indicating that the largest cluster is actually composed of multiple meaningful subclusters.

4. Algorithm Comparison: 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?

Based on the silhouette scores, standard K-Means performed slightly better than Recursive Bisecting K-Means for this dataset.

K-Means ($k = 3$) achieved a silhouette score of 0.3867

Bisecting K-Means ($k = 4$) achieved a silhouette score of 0.3602

A higher silhouette score indicates better-defined clusters with clearer separation and tighter cohesion.

Therefore, K-Means provides marginally more compact and better-separated clusters in this case.

The PCA-reduced dataset naturally forms about 3 dense regions.

The scatter plots show a structure where three major groups appear visually distinct.

K-Means with $k=3$ aligns well with this natural partitioning.

Bisecting K-Means forces hierarchical splitting.

It recursively divides the largest cluster, even if the split isn't optimal from a silhouette perspective.

This means:

- Some splits are strong (good separation)
- Others are weaker (forced splits within dense regions)

More clusters ($k=4$ in bisecting) does not always improve silhouette score.

Increasing k may:

- Break cohesive groups into unnecessary subclusters ●
- Increase cluster overlap
- Reduce separation between neighboring clusters

K-Means optimizes globally; Bisecting K-Means optimizes locally.
Recursive bisecting only optimizes each split independently, not the overall clustering structure.

K-Means optimizes all cluster centers at once → leading to more consistent global separation.

Therefore, K-Means gives a better silhouette score and thus performs slightly better for this dataset.

Bisecting K-Means still provides useful hierarchical insight but does not match the global separation quality of standard K-Means for this specific PCA-transformed bank dataset.

5. Business Insights: 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?

The clustering in PCA space reveals **distinct customer groups** that can be targeted with different marketing strategies. Even though PCA compresses the data into two components, the clusters still show meaningful separation that reflects underlying behavioral patterns.

A dominant “mainstream” customer group

One large cluster contains customers with:

- moderate account balance
- typical job/education categories
- average campaign and contact history

These customers represent **the core customer base**.

Marketing strategies here should focus on:

- general banking products
- automated or large-scale campaigns
- improving retention and cross-selling

A smaller, financially stronger or more stable segment

Another cluster lies closer to higher PCA1 values (indicating patterns tied to balance, job stability, and housing/loan status). These customers may:

- maintain higher balances
- exhibit consistent saving/borrowing behavior
- be more financially reliable

The bank can target them with:

- investment products
- premium credit cards
- long-term deposits (FDs)
- wealth management services

A distinct cluster of “low-engagement” or “hard to convert” customers A cluster separated toward negative PCA1/PCA2 tends to represent customers who:

- have lower balance
- may be contacted many times (higher campaign count)
- often do not convert easily
- might have loan/housing history that reduces financial

flexibility These customers may require:

- personalized follow-up
- alternative channels (SMS, social media) ●

lower-cost, low-risk financial products ●

targeted offers to improve engagement

Hierarchical segmentation from Bisecting K-Means

Bisecting K-Means further splits the large mainstream cluster into:

- subgroups with subtle differences in spending and saving behavior ●
- different job/education combinations ●
- different contact responsiveness

6. Visual Pattern Recognition: 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?

In the PCA scatter plot, the three colored regions (turquoise, yellow, purple) represent clusters of customers who share similar financial and demographic characteristics. PCA compresses the dataset into two principal components that capture the major sources of variation, so customers who appear close together tend to behave similarly.

1. Turquoise region — “Mainstream customers” (largest, most diffuse cluster) This region spans a wide area, indicating high variability. These customers generally have:

- average age and balance

- common jobs and education levels
- moderate responsiveness to campaigns

Because this is the most diverse group, its boundary is diffuse, reflecting overlapping characteristics within a broad middle-income customer group.

2. Yellow region — “Financially active / loan-prone customers”

This group appears more compact. These customers typically show patterns such as:

- moderate or lower balances
- higher campaign/contact counts
- more loan or housing-related activity

The shape of this cluster suggests greater internal similarity, so its boundaries are sharper than the turquoise region.

3. Purple region — “High-engagement or stable customers” Customers here tend to share:

- stable balances
- consistent contact patterns
- demographic traits associated with better campaign

responsiveness This region is fairly separate in PCA space, giving it a clearer boundary.

Sharp boundaries appear when:

- Customers share very similar behavior (e.g., balance, housing status, job type)
- PCA captures a distinct structural difference (e.g., high vs low balance groups)
- Clusters are well-separated in the original multidimensional space

Hence, the yellow and purple clusters show relatively sharper boundaries.

Diffuse boundaries appear when:

- Customer attributes gradually change rather than forming discrete groups
- The cluster contains many mixed sub-segments (e.g., diverse jobs or campaign histories)
- PCA compresses multiple dimensions, causing overlap

The turquoise region is diffuse because it represents a **heterogeneous, blended population** with overlapping behavioral traits.