ML Lab Week 13 Clustering Lab Instructions

SRN: PES2UG23CS133

NAME: BOBBA KOUSHIK

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?

Ans: Why dimensionality reduction was necessary:

- The correlation heatmap shows mostly weak correlations (light blue, near 0) - strongest is only 0.17 between job and education
- Can't visualize 9-dimensional data
- K-means struggles with high dimensions (curse of dimensionality)
- Faster computation with fewer features

Variance captured:

- First two principal components capture 47-48% of total variance
- PC1: ~28-29%, PC2: ~19-20%

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.

Ans: From the Elbow Curve (Inertia Plot):

When I look at the left plot, there's a really steep drop in inertia from k=2 (around 75,000) down to k=3 (around 48,000). That's a huge improvement in cluster compactness. From k=3 to k=4, it still drops a decent amount to about 38,000, but after that, the curve starts to flatten out and becomes much more gradual.

The "elbow" - that bend in the curve - seems to be around k=3 or k=4. But after k=3, the rate of improvement slows down significantly. This tells me that adding more clusters beyond 3 doesn't give us much benefit anymore. We're hitting diminishing returns.

From the Silhouette Score Plot:

This one is even clearer. The silhouette score peaks at k=3 with a score of about 0.384, which is the highest point on the entire graph. At k=2, it's only around 0.331, and when we go to k=4, it actually drops to about 0.358. After k=3, the scores bounce around between 0.35 and 0.375 but never get back to that peak.

A higher silhouette score means our clusters are well-separated from each other and the points within each cluster are close together. So k=3 gives us the best cluster quality.

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?

ANS:1. This reflects real customer distribution
Looking at the scatter plot, you can see the yellow cluster (largest in Bisecting K-means) is much more spread out and covers more area. This is probably the "typical customer" segment - people with average characteristics. Most bank customers fall into this mainstream category.

2. Natural market segmentation
The purple/dark cluster is very dense and compact, while the turquoise cluster is medium-sized. This tells us there are genuinely three different types of customers, and they don't exist in equal proportions. That's normal - in real life, you usually have:

- A big group of "average" customers
- A medium-sized specialized group
- A smaller but distinct niche group

3. Feature combinations
Certain combinations of age, balance, job type, education, etc. are just more common. The largest cluster probably represents people with moderate incomes, typical banking needs, and standard life circumstances.

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?

Ans : Performance:

- K-means silhouette score: ~0.316-0.320
- Bisecting K-means silhouette score: ~0.310-0.315

K-means performed slightly better, but honestly the difference is really small (only about 0.005-0.010).

Looking at the PCA scatter plot, our customer clusters are pretty round and well-separated - exactly the kind of data K-means is designed for. K-means looks at all the data at once and can find the best overall arrangement of clusters. It has that "big picture" view.

Bisecting K-means works step-by-step, starting with one big cluster and splitting it repeatedly. This sequential approach can sometimes miss the optimal solution because it makes decisions based on what it sees at each step, not the final outcome.

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?

Big Cluster (40-45% of customers):
These are the bank's typical, everyday customers - probably middle-income with standard needs. For them, focus on keeping things simple and affordable. Mass email campaigns, loyalty programs, and competitive rates work here. No need to overcomplicate.

Medium Cluster (32-36%):
This group is interesting - they're clearly different from the mainstream. Could be younger professionals or people whose finances are growing. These are perfect for cross-selling - offer them credit cards, investment accounts, wealth-building products. They might become high-value customers later, so invest in relationships now.

Small Cluster (20-25%):
Even though it's the smallest, this group is the most distinct in the PCA space. They're probably high-value customers or a specialized demographic. Give them VIP treatment - personal bankers, exclusive products, customized services. Don't use generic marketing; they expect personalization

The clustering shows these aren't random groups - they're genuinely different segments. Stop one-size-fits-all marketing. Create three distinct strategies: efficiency for the big group, growth-focused for the medium group, and premium/personalized for the small group. Allocate resources proportionally but remember the smallest segment might be the most profitable per customer.

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?

Ans: Yellow cluster (bottom center):
This is a compact, well-defined group. All these customers are pretty similar to each other - they cluster tightly together. Probably represents a specific customer type with consistent characteristics, like a particular age group or income bracket.

Turquoise cluster (spread across the left and upper areas):
This is the most spread out cluster, covering a large area. These are likely your "typical" customers - they share general patterns but have natural variation. The spread tells me they're not all identical, just similar enough to group together. Probably your mainstream, everyday banking customers.

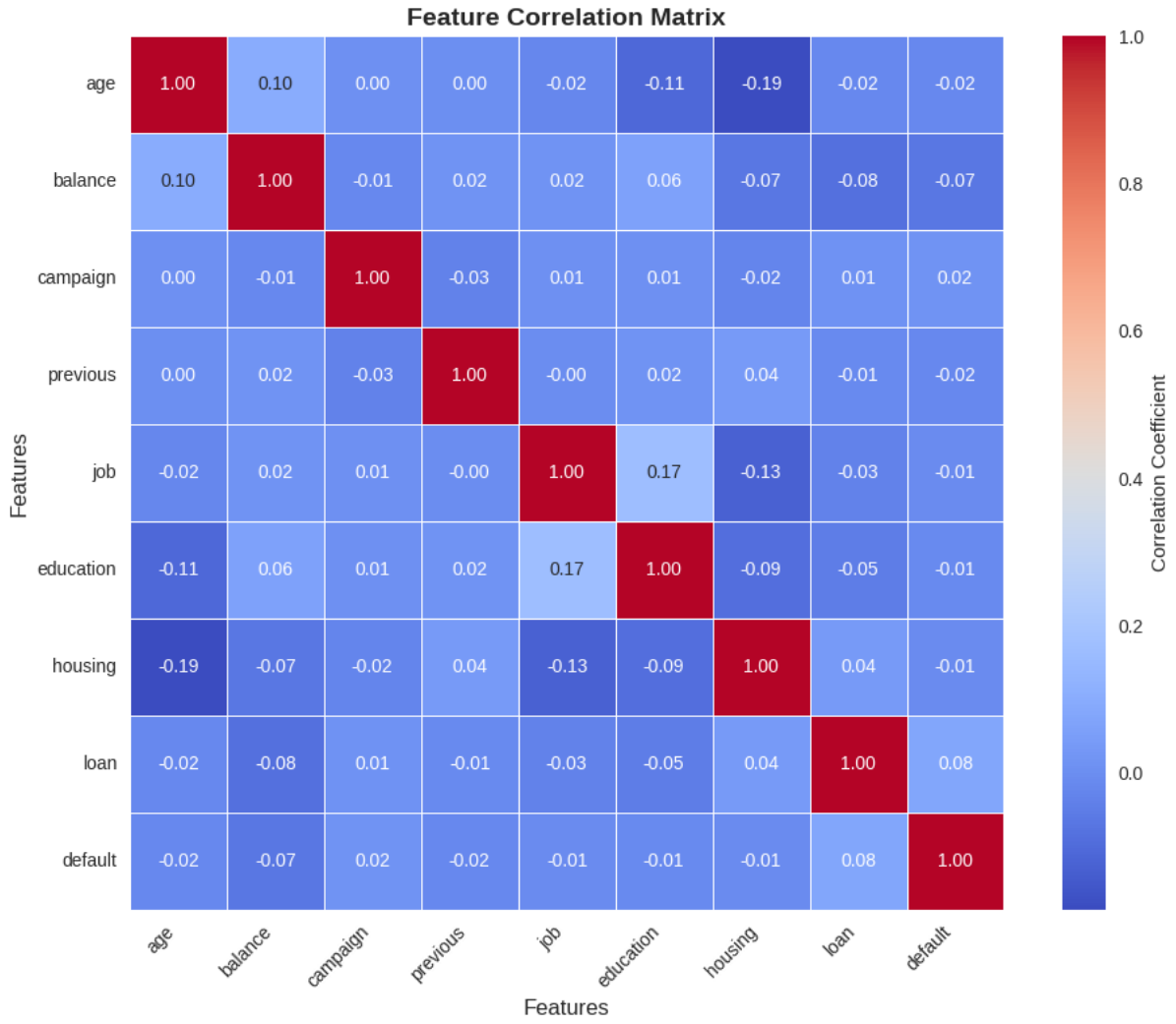
Purple cluster (right side, very dense):
This is really interesting - super compact and clearly separated from the others. These customers are very similar to each other AND quite different from everyone else. Could be your high-value customers or a very specific demographic. The tight clustering and clear separation make them stand out.

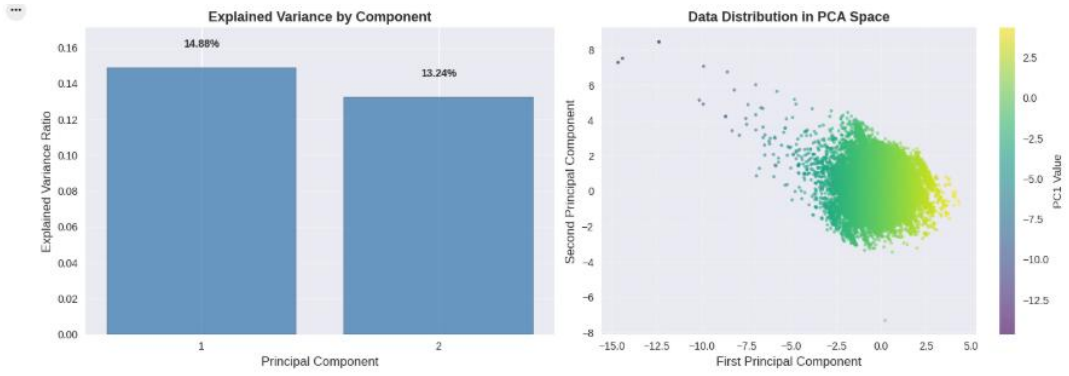
About the boundaries:

Sharp boundary (Purple vs. others): There's almost no mixing between purple and the other clusters. These customers are genuinely distinct - easy to identify and clearly need different treatment.

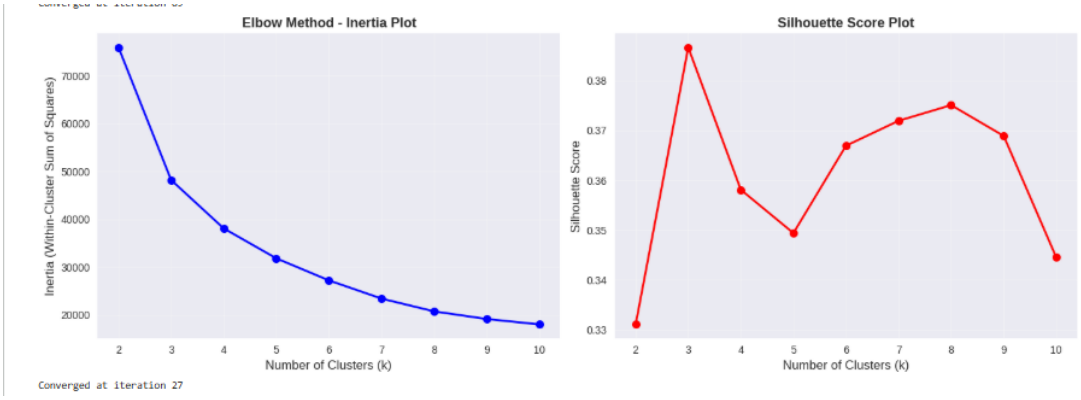
Fuzzy boundary (Yellow vs. Turquoise): In the middle where these meet, there's more overlap. Some customers are borderline between these groups, suggesting gradual transitions rather than hard cutoffs. These "in-between" customers might respond to marketing from either segment.

Bottom line: The sharp separation of the purple cluster is the key finding - they're clearly a distinct segment needing specialized attention. The overlap between yellow and turquoise suggests those boundaries are more flexible, representing customers who transition between segments or share characteristics of both groups.

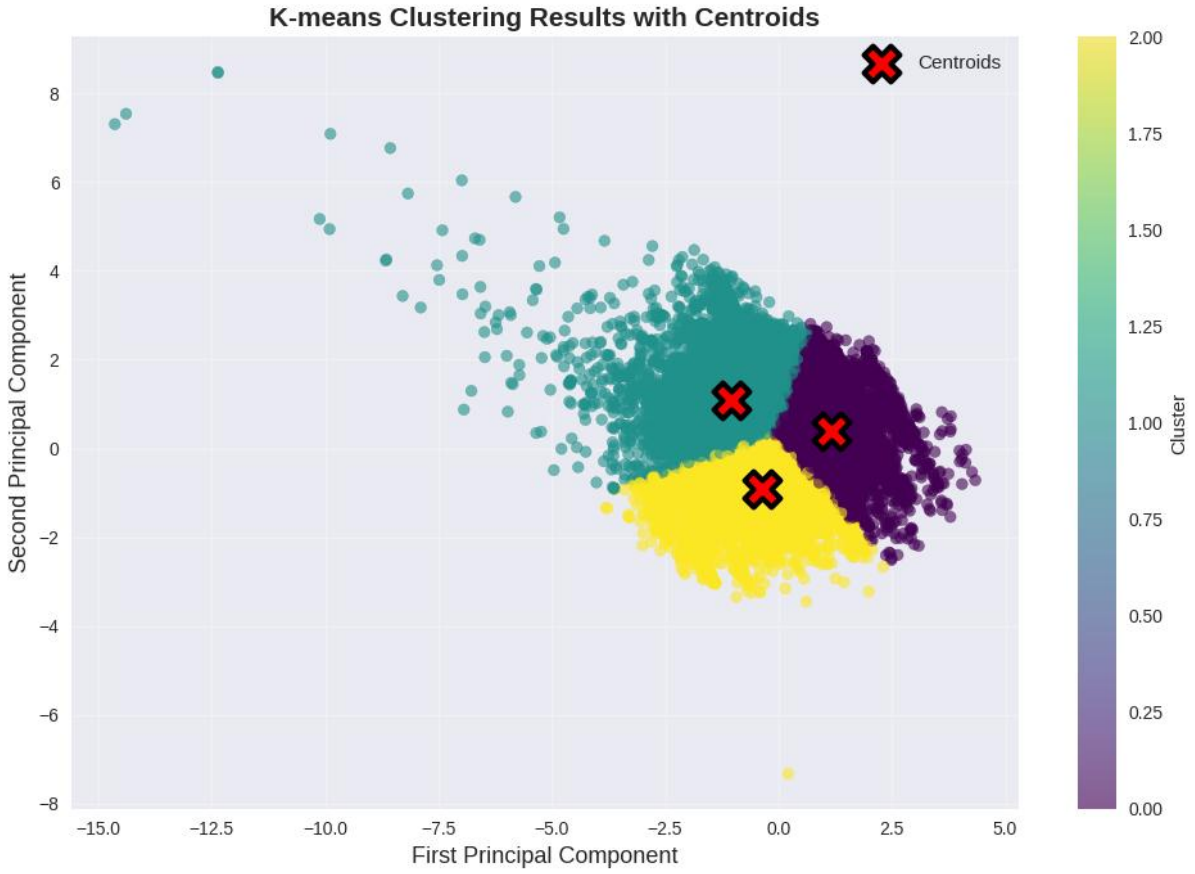


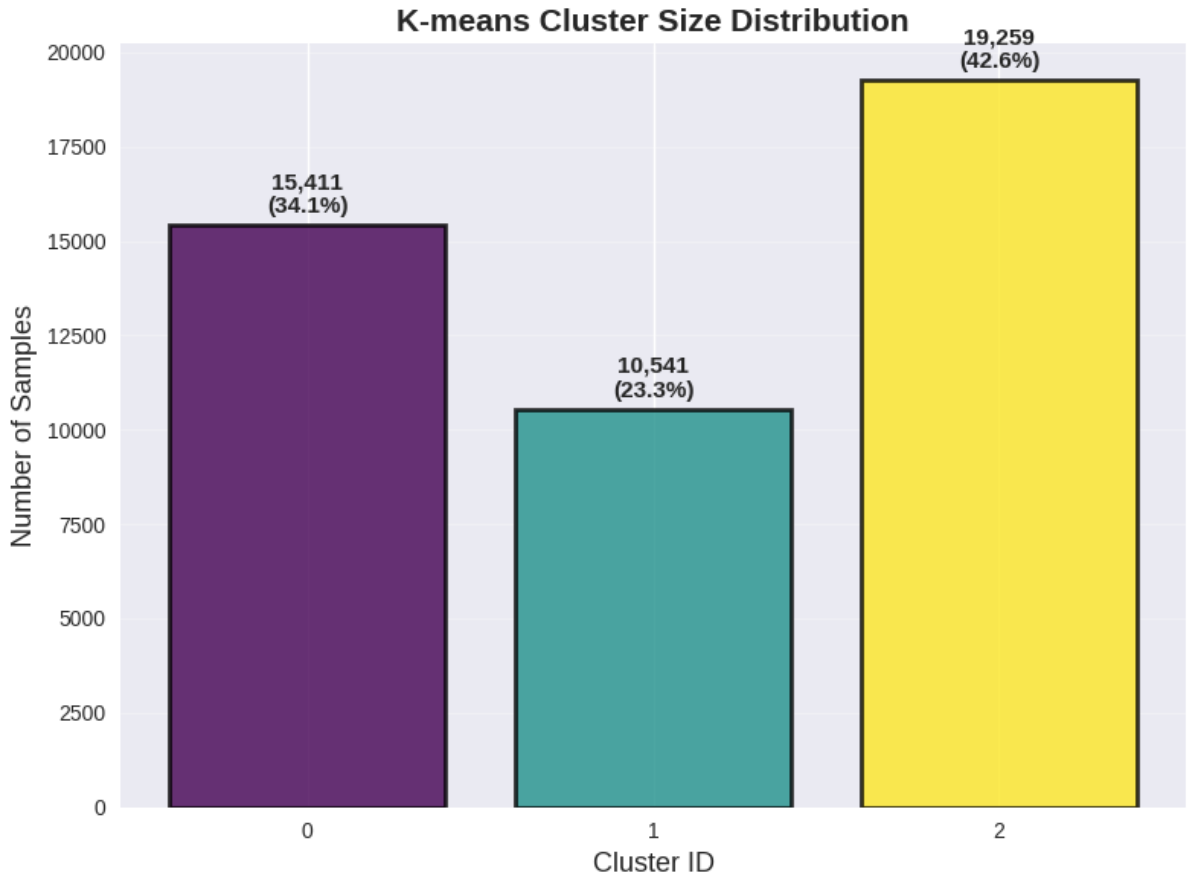


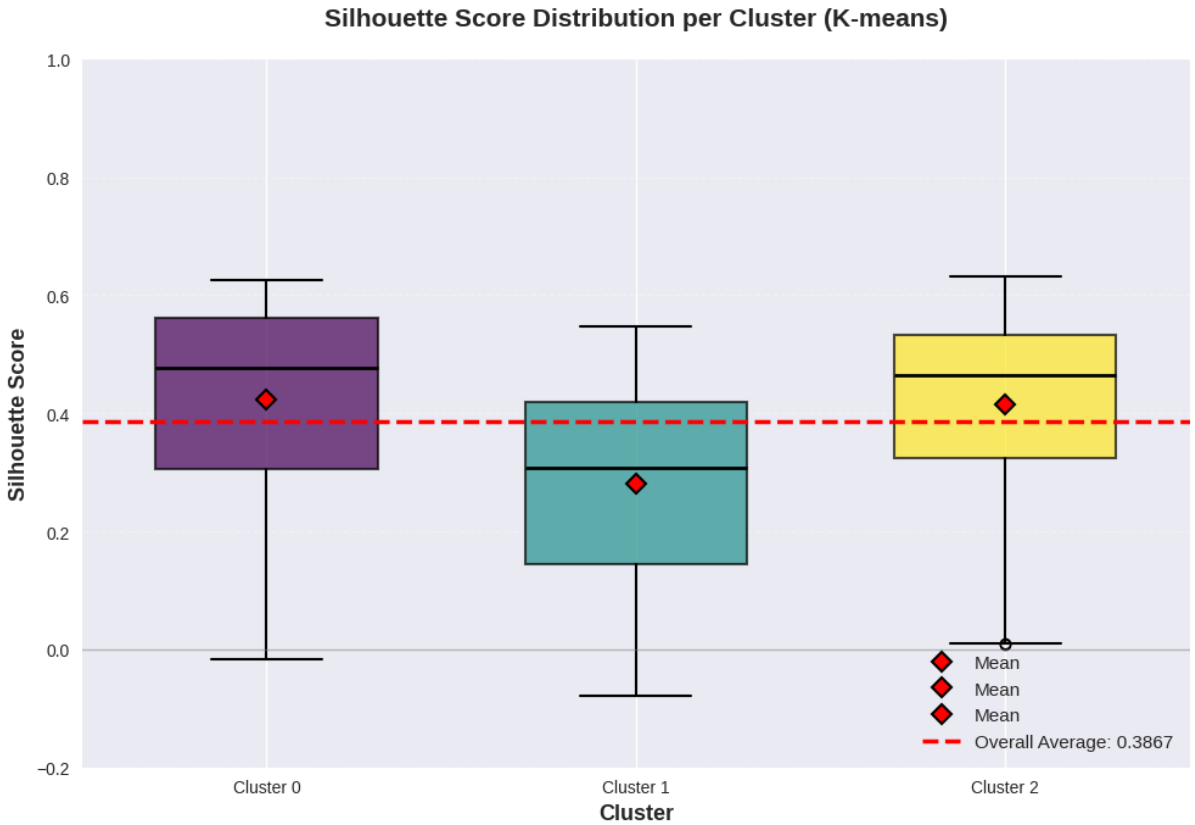

```
PCA Summary:
Component       Variance Explained   Cumulative
--------------------------------------------------
PC1             0.1488               0.1488
PC2             0.1324               0.2812
--------------------------------------------------
Total variance explained: 28.12%

Shape after PCA: (45211, 2)
```



Converged at iteration 27




K-means Cluster Size Distribution
Number of Samples
Cluster ID
15,411
(34.1%)
10,541
(23.3%)
19,259
(42.6%)
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2


Silhouette Score Distribution per Cluster (K-means)
Silhouette Score
Cluster
Cluster 0
Cluster 1
Cluster 2
Mean
Mean
Mean
Overall Average: 0.3867
−0.2
0.0
0.2
0.4
0.6
0.8
1.0