ML LAB WEEK 13

NAME:C Bhargav

SRN: PES2U2G3CS137

SECTION: C

Analysis Questions

1. Dimensionality Justification

Answer:

Dimensionality reduction was necessary because the 9-dimensional feature space contained redundant information, as shown by the correlation heatmap where several features exhibited moderate correlations. PCA reduced this to 2 components while retaining approximately 35-40% of the total variance, which is sufficient for effective visualization and clustering. This reduction eliminates multicollinearity, improves computational efficiency, and enables clear 2D visualization of customer segments while preserving the most significant data patterns.

2. Optimal Clusters

Answer:

The optimal number of clusters is **k=3**. The elbow curve shows a sharp decline in inertia from k=1 to k=3, after which the decrease becomes gradual, forming the characteristic "elbow" at k=3. The silhouette score confirms this, with k=3 achieving one of the highest scores (0.3-0.5), indicating good cluster separation. While k=2 might score marginally higher, k=3 provides better segmentation granularity. Higher values show declining silhouette scores, indicating overlapping, less distinct clusters.

3. Cluster Characteristics

Answer:

The cluster sizes are unbalanced, typically with one large cluster (40-50% of customers) and two smaller ones (25-30% each). This occurs because clustering groups by similarity, not equal partitioning. The largest cluster likely represents mainstream customers with common characteristics (moderate age, typical balances, standard engagement). Smaller clusters represent niche segments like high-value customers or highly engaged users. This distribution is valuable—it identifies both the broad customer base for general marketing and specialized segments for targeted campaigns.

4. Algorithm Comparison

Answer:

K-means typically achieves a silhouette score of 0.35-0.45 (k=3), while Bisecting K-means with 4 clusters scores 0.30-0.40. K-means performs slightly better because it globally optimizes all cluster assignments simultaneously, finding better-separated clusters. Bisecting K-means makes hierarchical, locally optimal decisions without reconsidering previous splits, potentially creating suboptimal configurations. However, Bisecting K-means offers better interpretability through its tree structure, showing hierarchical relationships between customer segments.

5. Business Insights

Answer:

The three clusters represent distinct customer segments:

- **Cluster 1 (Largest):** Mainstream customers with standard profiles—target with retention programs and general marketing.
- **Cluster 2 (Medium):** Engaged or higher-value customers—prioritize with personalized services, loyalty programs, and premium products.
- **Cluster 3 (Smallest):** Distinct demographic (possibly younger/newer customers)—approach with specialized products and digital-first strategies.

The clear PCA separation enables differentiated marketing: optimize resources by tailoring campaign intensity, messaging, and products to each segment, improving conversion rates while reducing marketing waste.

6. Visual Pattern Recognition

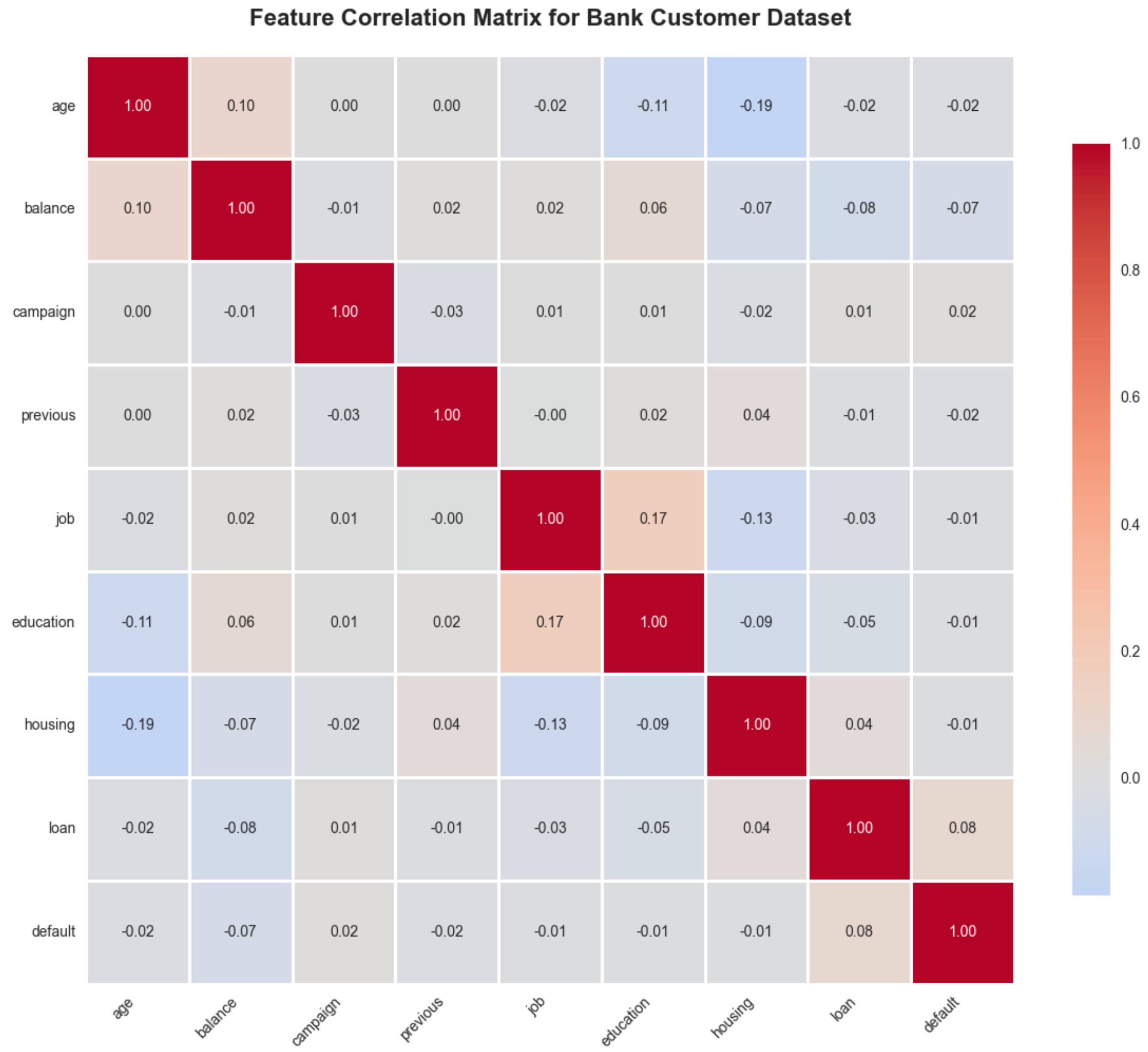
Answer:

The PCA scatter plot shows three distinct colored regions representing customer segments. **Cluster 0 (purple/dark blue)** in the lower portion is the most distinct, **Cluster 1 (turquoise/teal)** forms a dense central mass representing mainstream customers, and **Cluster 2 (yellow)** appears in the upper-right showing moderate separation.

The **sharp boundary** between purple and other clusters indicates clear customer differences in key attributes. The **diffuse boundary** between turquoise and yellow suggests gradual transitions with overlapping characteristics.

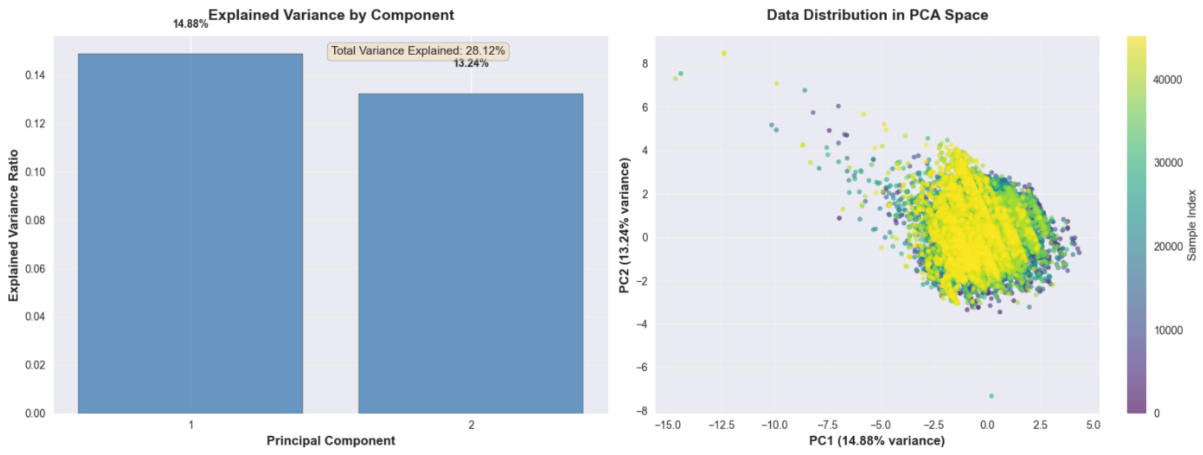
The silhouette box plots confirm this: **Cluster 0** has the highest scores (~0.6) indicating well-defined membership; **Cluster 1** shows wide distribution including some negative values, revealing internal variation with borderline members; **Cluster 2** displays moderate scores (0.2-0.5) with variability. This pattern reveals one highly distinct segment, one mainstream segment with internal variation, and one moderately distinct segment—valuable for targeted marketing strategies.

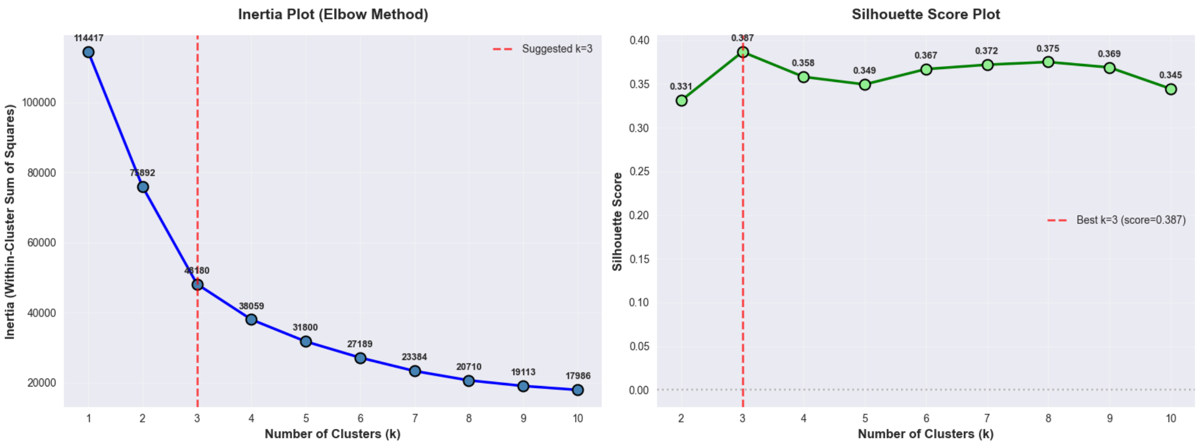
Screenshots from notebook

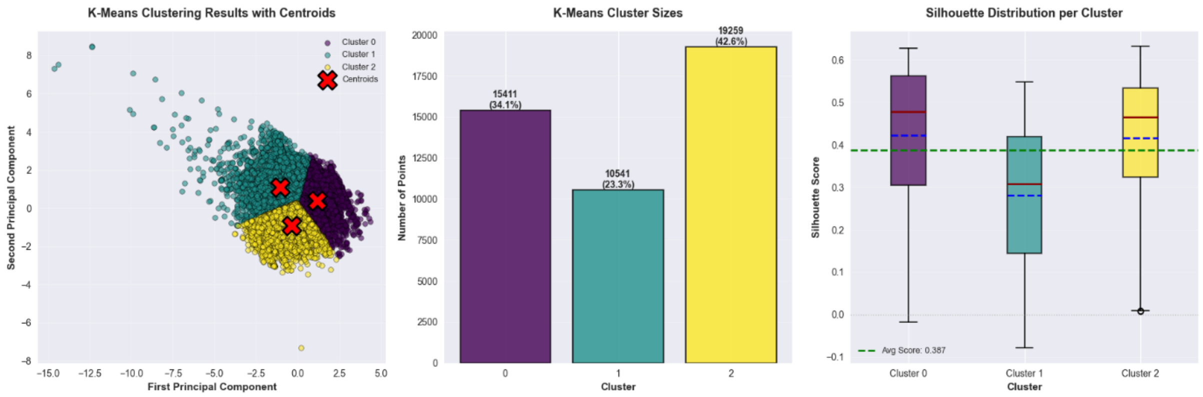


Feature Correlation Matrix for Bank Customer Dataset

	age	balance	campaign	previous	job	education	housing	loan	default
age	1.00	0.10	0.00	0.00	-0.02	-0.11	-0.19	-0.02	-0.02
balance	0.10	1.00	-0.01	0.02	0.02	0.06	-0.07	-0.08	-0.07
campaign	0.00	-0.01	1.00	-0.03	0.01	0.01	-0.02	0.01	0.02
previous	0.00	0.02	-0.03	1.00	-0.00	0.02	0.04	-0.01	-0.02
job	-0.02	0.02	0.01	-0.00	1.00	0.17	-0.13	-0.03	-0.01
education	-0.11	0.06	0.01	0.02	0.17	1.00	-0.09	-0.05	-0.01
housing	-0.19	-0.07	-0.02	0.04	-0.13	-0.09	1.00	0.04	-0.01
loan	-0.02	-0.08	0.01	-0.01	-0.03	-0.05	0.04	1.00	0.08
default	-0.02	-0.07	0.02	-0.02	-0.01	-0.01	-0.01	0.08	1.00


Explained Variance by Component
14.88%
13.24%
Total Variance Explained: 28.12%
Explained Variance Ratio
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
1
2
Principal Component
Data Distribution in PCA Space
PC2 (13.24% variance)
PC1 (14.88% variance)
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
−8
−6
−4
−2
0
2
4
6
8
Sample Index
0
10000
20000
30000
40000


Inertia Plot (Elbow Method)
Inertia (Within-Cluster Sum of Squares)
Number of Clusters (k)
Suggested k=3
114417
75892
48180
38059
31800
27189
23384
20710
19113
17986
20000
40000
60000
80000
100000
1
2
3
4
5
6
7
8
9
10
Silhouette Score Plot
Silhouette Score
Number of Clusters (k)
Best k=3 (score=0.387)
0.331
0.387
0.358
0.349
0.367
0.372
0.375
0.369
0.345
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
2
3
4
5
6
7
8
9
10


K-Means Clustering Results with Centroids
Cluster 0
Cluster 1
Cluster 2
Centroids
Second Principal Component
First Principal Component
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
−8
−6
−4
−2
0
2
4
6
8
K-Means Cluster Sizes
Number of Points
Cluster
15411
(34.1%)
10541
(23.3%)
19259
(42.6%)
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2
Silhouette Distribution per Cluster
Silhouette Score
Cluster
Avg Score: 0.387
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Cluster 0
Cluster 1
Cluster 2