ML Lab Week 13 Clustering Lab Instructions

DETAILS:

NAME: **PREKSHA KAMALESH**
SRN:**PES2UG23CS902**
SECTION:**F**

Content Requirements:

1. **Analysis Questions:** Provide clear and concise answers to all 8 analysis questions from the notebook. The questions are divided into three sections:

---

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:
   Dimensionality reduction was necessary for two main reasons. First, the **correlation heatmap** shows that the variance is spread across many of the original 9 dimensions, with no single feature being dominant or highly redundant (most correlations are low). Second, the **explained variance plot** confirms this, showing that the first two principal components capture only **28.12%** of the total variance. This makes it impossible to visualize the 9-dimensional data structure. PCA was needed to compress the features into a 2D space for visualization, even though this compression results in a significant loss of variance..

   A total of **28.12%** of the variance is captured by the first two components. (PC1: 14.88%, PC2: 13.24%)

---

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.

   The optimal number of clusters is **3**.
   Inertia Plot (Elbow Method): The plot shows a clear "elbow" at k=3. The inertia (within-cluster sum of squares) drops sharply from k=2 (75892.03) to

k=3 (48179.64). After k=3, the rate of decrease flattens, indicating diminishing returns for adding more clusters.
Silhouette Score Plot: This is strongly confirmed by the silhouette plot , which shows a distinct peak at **k=3** with a score of **0.3867**. This is the highest score in the range, indicating that 3 clusters provides the best balance of cluster density (cohesion) and separation.

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?

Both algorithms produced three clusters of unequal size.
K-means (k=3): The "Cluster Sizes" plot shows that Cluster 2 is the largest (approx. 25,000 samples), followed by Cluster 0 (approx. 11,000), and Cluster 1 (approx. 9,000).

Bisecting K-means (k=3): The "Final Cluster Sizes" plot shows a similar, but not identical, distribution. Cluster 1 is the largest (approx. 20,000), followed by Cluster 2 (approx. 13,700), and Cluster 0 (approx. 11,300).

The unequal sizes suggest that the customer segments are not evenly distributed. The largest cluster (Cluster 2 in K-means, Cluster 1 in Bisecting K-means) likely represents the most common or "mainstream" customer profile for the bank. The two smaller clusters represent more specific, niche segments with different financial characteristics. This is valuable as it tells the bank that a single, general marketing strategy will not be effective for all customers.

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?

K-means (k=3) achieved an average silhouette score of **0.3867**.

Recursive Bisecting K-means (k=3) achieved an average silhouette score of **0.3379**.

K-means performed better for this dataset. Its higher silhouette score suggests it found clusters that were, on average, more dense and better separated. This is likely because K-means is a global optimization method that tries to find the best positions for all 3 cluster centroids simultaneously. In contrast, Bisecting K-means is a "greedy" algorithm; it makes a series of binary (k=2) splits, and an early split (like the one separating C0 and C2 from C1) might not be optimal for the final 3-cluster solution.

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?

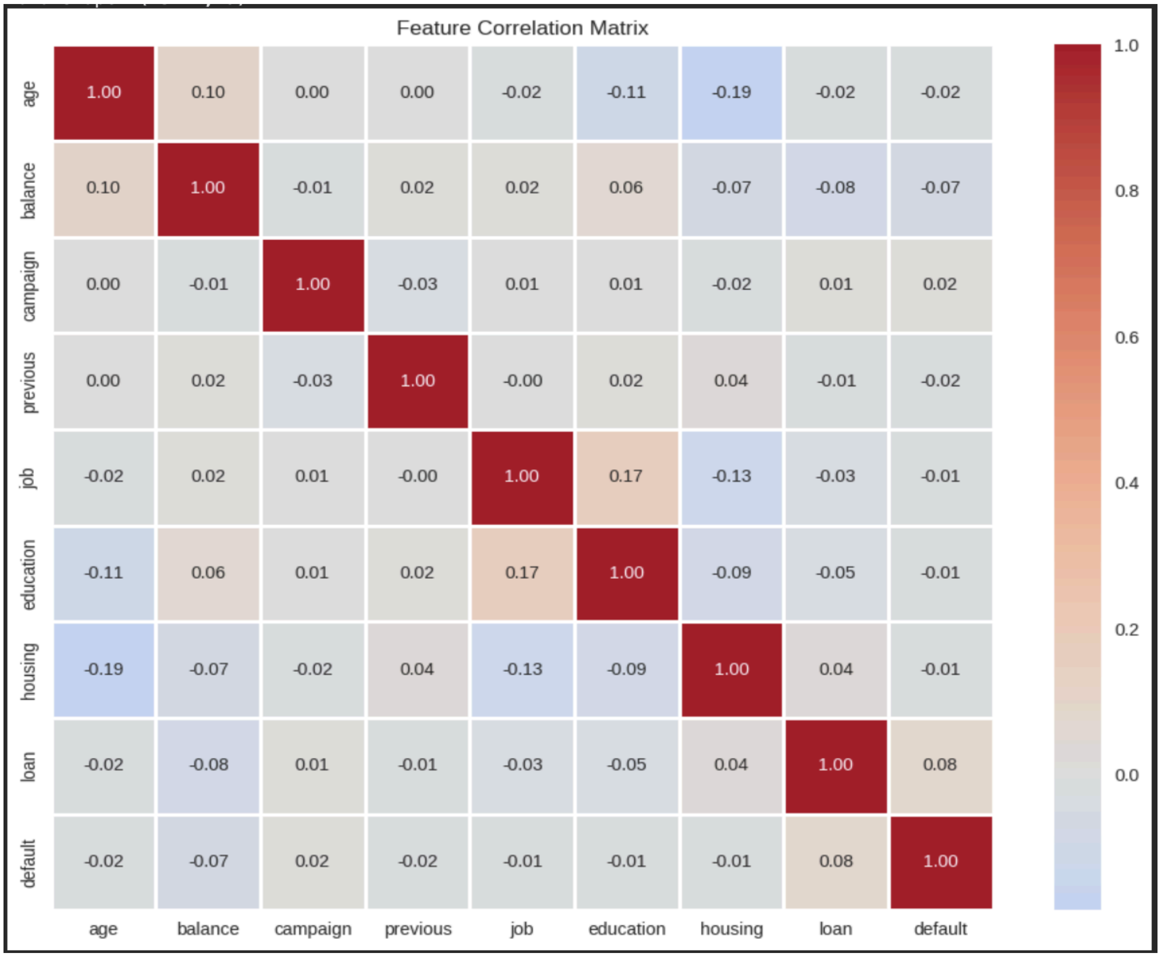
   The K-means scatter plot shows three clear customer segments. One segment (Cluster 2, purple) is very large, dense, and centrally located. The other two segments (Cluster 0, yellow, and Cluster 1, teal) are smaller and more distinct.
   This provides a clear action plan for the bank's marketing:
   Develop a broad, general marketing campaign for the large, "mainstream" segment (Cluster 2).
   Analyze the original features (age, balance, job, etc.) of the two smaller, niche clusters (0 and 1) to understand what makes them different.
   Create two separate, highly targeted campaigns to address the specific needs, behaviors, or demographics of these smaller segments.

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?

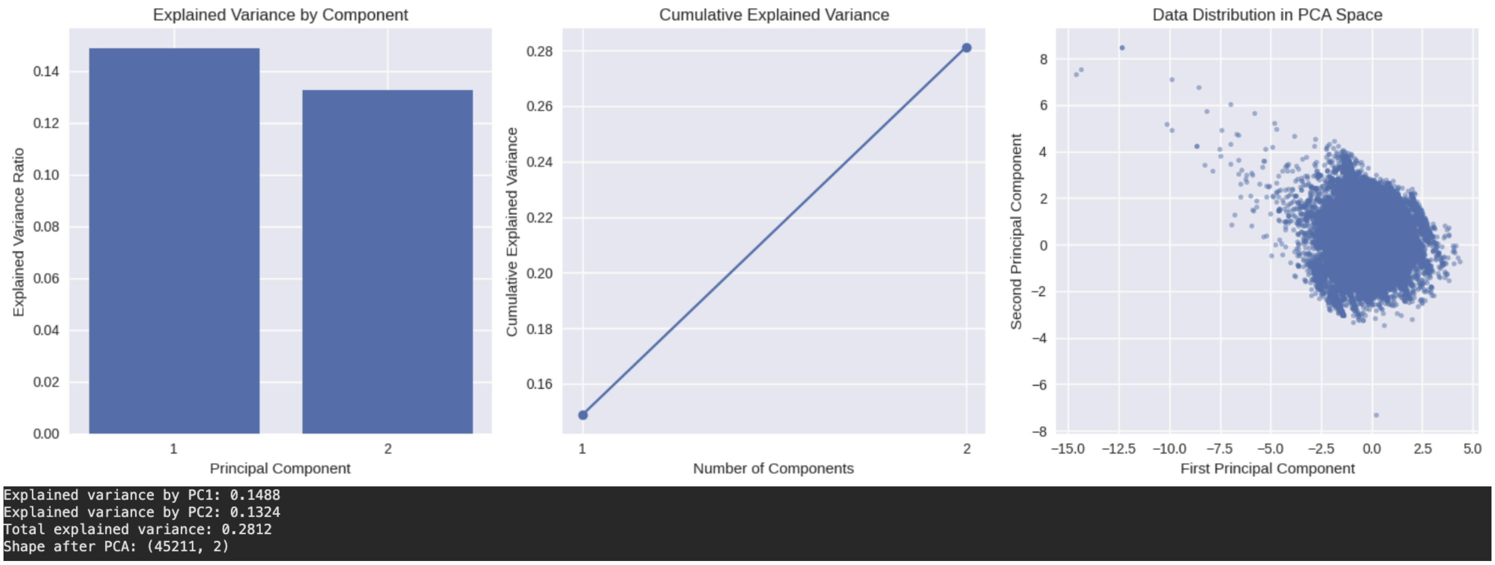
   The colored regions in the PCA scatter plot represent the three different customer segments identified by the K-means algorithm.
   The boundaries between the clusters are **diffuse** (blurry) rather than sharp. This is especially clear between the large purple cluster (2) and the yellow cluster (0). This "fuzziness" indicates that the customer characteristics are not mutually exclusive; there is significant overlap between the segments. Customers located on the boundary between two clusters likely share traits from both segments, making a hard separation impossible. This is also reflected in the box plot, which shows that clusters 0 and 1 have some points with negative silhouette scores, meaning they are close to (or perhaps belong in) another cluster.

2. **Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of** 4 screenshots, divided as

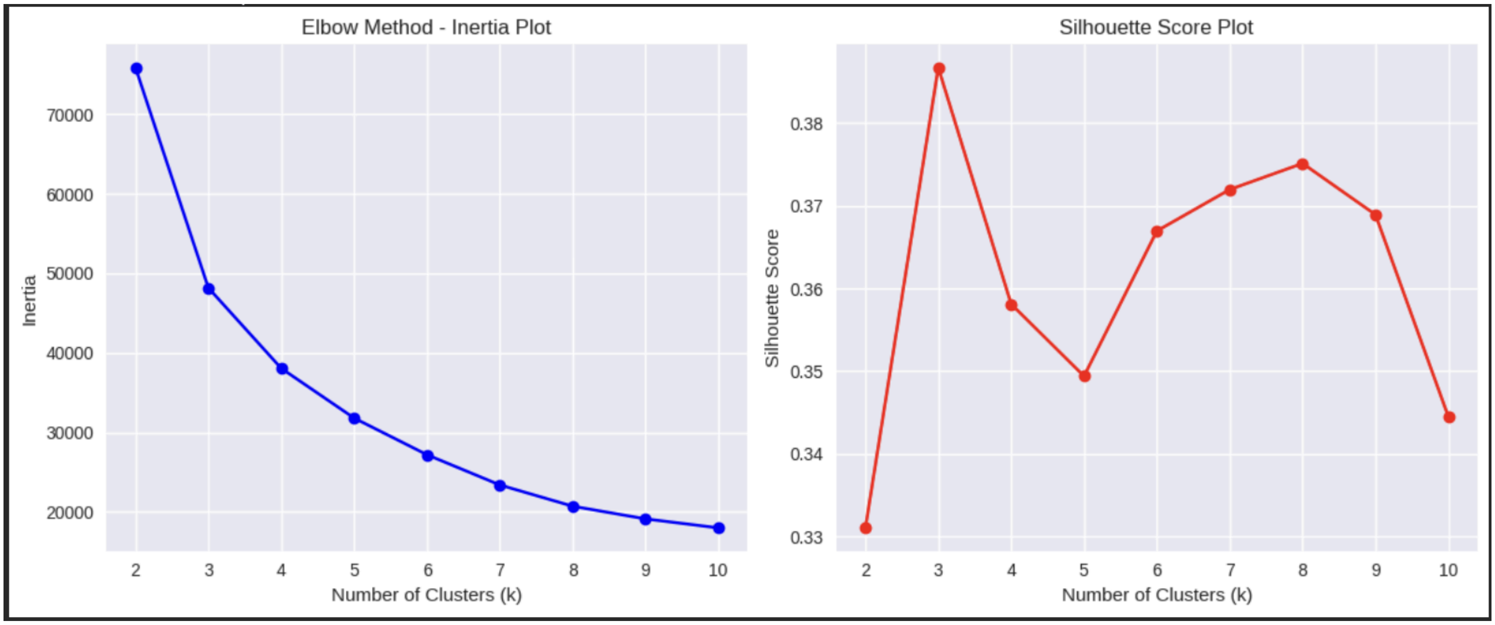
Feature Correaltion matrix for the dataset,‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA,‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means,K-means Clustering Results with Centroids Visible (Scatter Plot),K-means Cluster Sizes (Bar Plot)
Silhouette distribution per cluster for K-means (Box Plot)



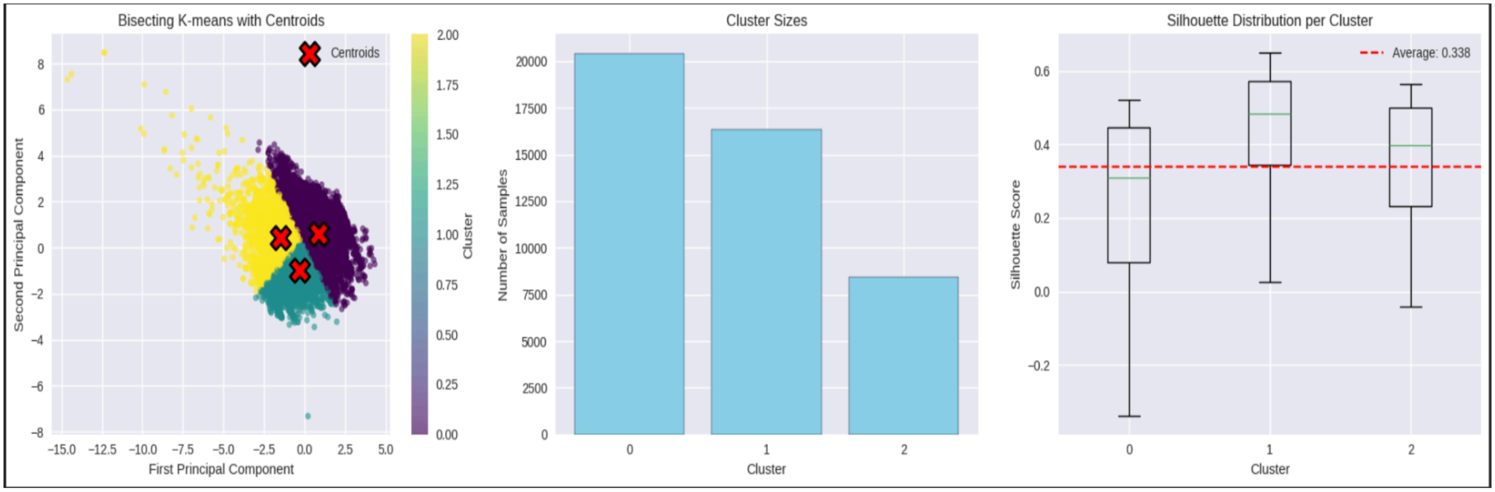
	age	balance	campaign	previous	job	education	housing	loan	default
age	1.00	0.10	0.00	0.00	-0.02	-0.11	-0.19	-0.02	-0.02
balance	0.10	1.00	-0.01	0.02	0.02	0.06	-0.07	-0.08	-0.07
campaign	0.00	-0.01	1.00	-0.03	0.01	0.01	-0.02	0.01	0.02
previous	0.00	0.02	-0.03	1.00	-0.00	0.02	0.04	-0.01	-0.02
job	-0.02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

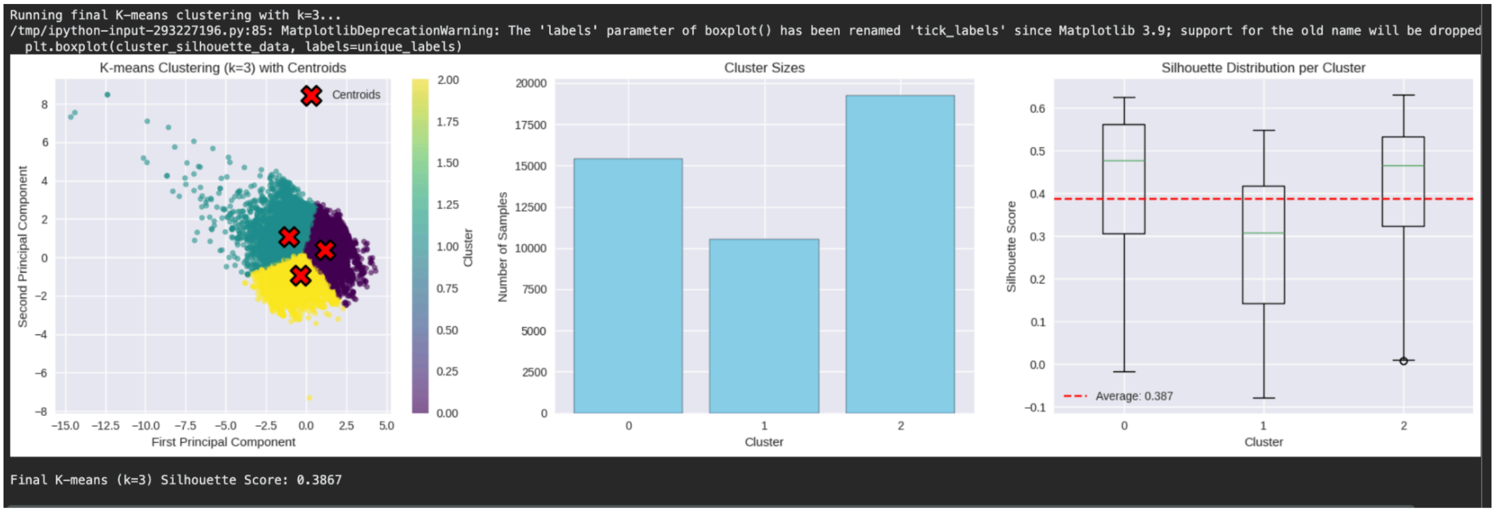


```
Explained variance by PC1: 0.1488
Explained variance by PC2: 0.1324
Total explained variance: 0.2812
Shape after PCA: (45211, 2)
```



bisecting





```
Running final K-means clustering with k=3...
/tmp/ipython-input-293227196.py:85: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick_labels' since Matplotlib 3.9; support for the old name will be dropped
  plt.boxplot(cluster_silhouette_data, labels=unique_labels)
```

```
Final K-means (k=3) Silhouette Score: 0.3867
```