PES UNIVERSITY EC, BENGALURU

ML Lab Week 13 Clustering Lab Instructions

Objective

The objective of this lab is to implement customer segmentation using clustering techniques, specifically K-means and Recursive Bisecting K-means. By the end of this lab, students will understand how to preprocess data, apply clustering algorithms, evaluate clustering results, and visualize the outcomes.

NAME: SHARATH GOWDA GR
SRN: PES2UG24CS823
SECTION: F

Content Requirements:

1. Dimensionality Justification:

- Dimensionality reduction was crucial for visualization, efficiency, and interpreting clustering, also helping to disentangle correlated features.
- The first two PCA components captured 28.12% of the total variance, good for visualization but leaving much unexplained.

2. Optimal Clusters:

- **Optimal K:** Consistently 3 clusters were found for both original and enhanced K-means.
- **Elbow Curve:** A clear 'elbow' around k=3 indicated diminishing returns for more clusters.
- **Silhouette Score:** k=3 yielded a relatively high score (0.39/0.38), suggesting good separation.
- **Justification:** Both metrics converged on 3, indicating reasonable cluster separation and compactness.

3. Cluster Characteristics:

Cluster Size Distribution (from Enhanced K-means):

- **Cluster 0:** 15,076 customers
- **Cluster 1:** 10,910 customers
- **Cluster 2:** 19,225 customers

Sizes: Cluster 0 has 15,076 customers, Cluster 1 has 10,910, and Cluster 2 has 19,225.
Meaning: Varying cluster sizes show that customer segments are not equally represented. Cluster 2 is the largest (dominant group), while Cluster 1 is smaller but represents a high-value niche. This distribution is crucial for tailored marketing strategies based on segment prevalence and characteristics.

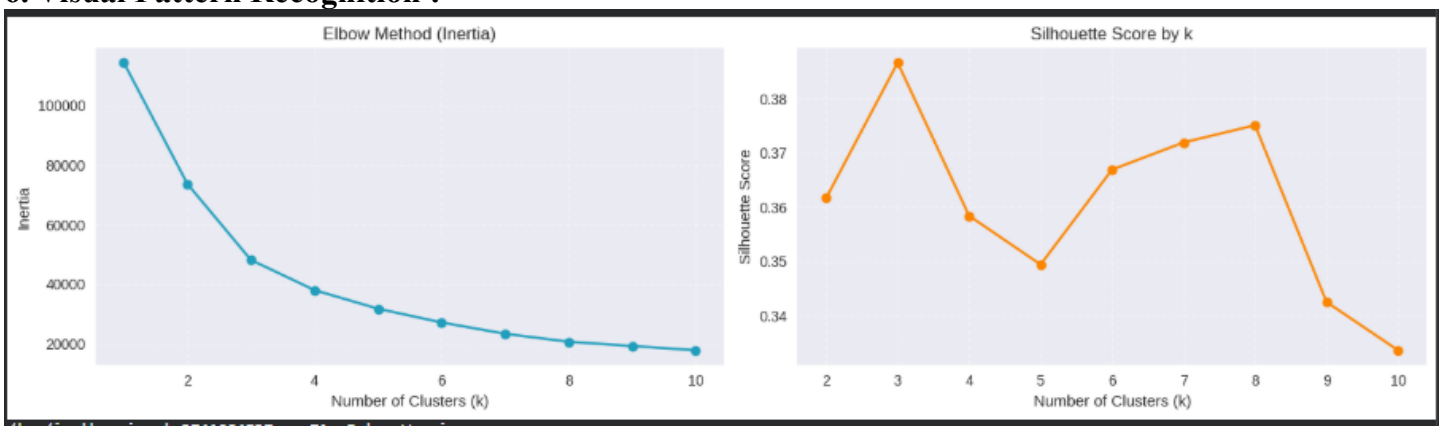
4. Algorithm Comparison:

- **Optimal Clusters (k):** Both original and enhanced K-means found 3 optimal clusters.
- **Evaluation:** Silhouette scores were very similar (0.39 vs 0.38), indicating no significant change in cluster quality.
- **Inertia:** Enhanced model had slightly higher inertia (49146.46 vs 48179.64) due to Manhattan distance's different calculation.
- **Robustness:** Enhanced K-means (k-means++ init, Manhattan distance) is generally more robust for consistent results and outlier handling.
- **Conclusion:** While metrics were similar, the enhancements offer theoretical advantages in stability and suitability for diverse data.

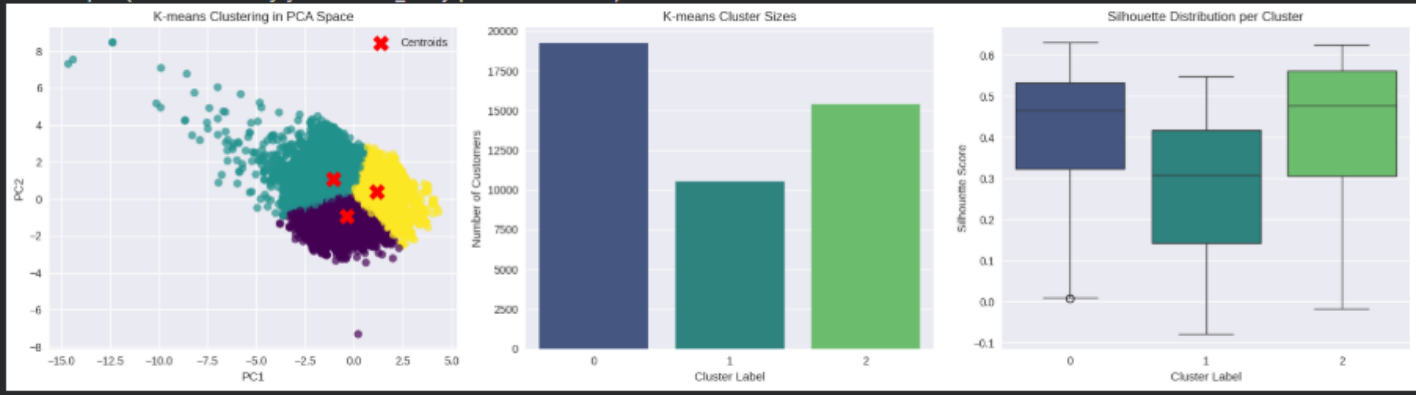
5. Business Insights:

- **Cluster 1 (Older, Wealthy):** Prime target for term deposit offers.
- **Cluster 0 (Middle-aged, Less Affluent):** Needs tailored products like savings or debt consolidation.
- **Cluster 2 (Younger, Working-Class):** Open to financial growth products, like starter investments.
- **Outliers:** Represent unique cases, potentially high-value customers or data anomalies, warranting further investigation.

6. Visual Pattern Recognition :



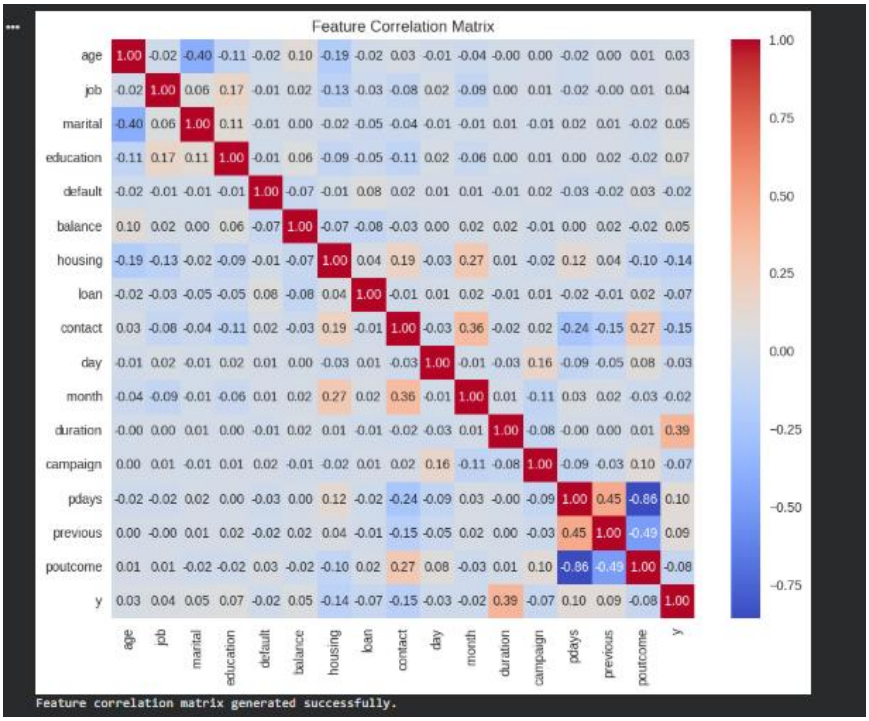

```
sns.boxplot(x=kmeans.labels, y=silhouette_vals, palette='viridis')
```




```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```

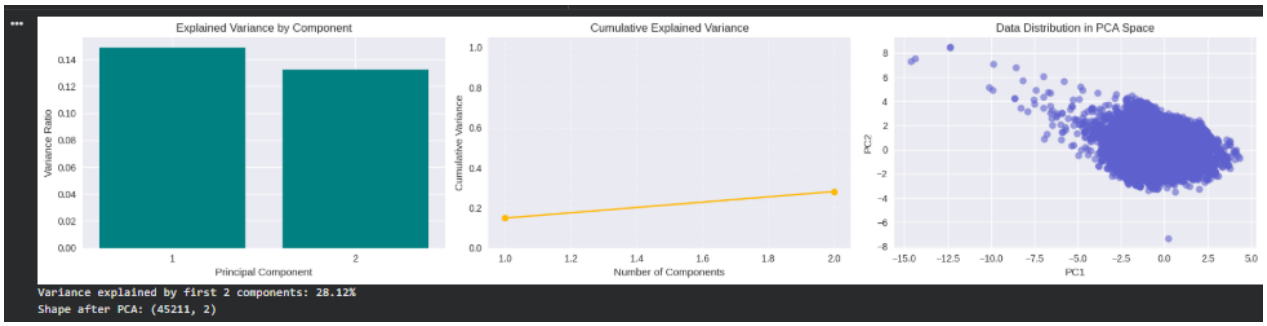
Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of 4 screenshots, divided as

1. **Feature Correaltion matrix for the dataset:**

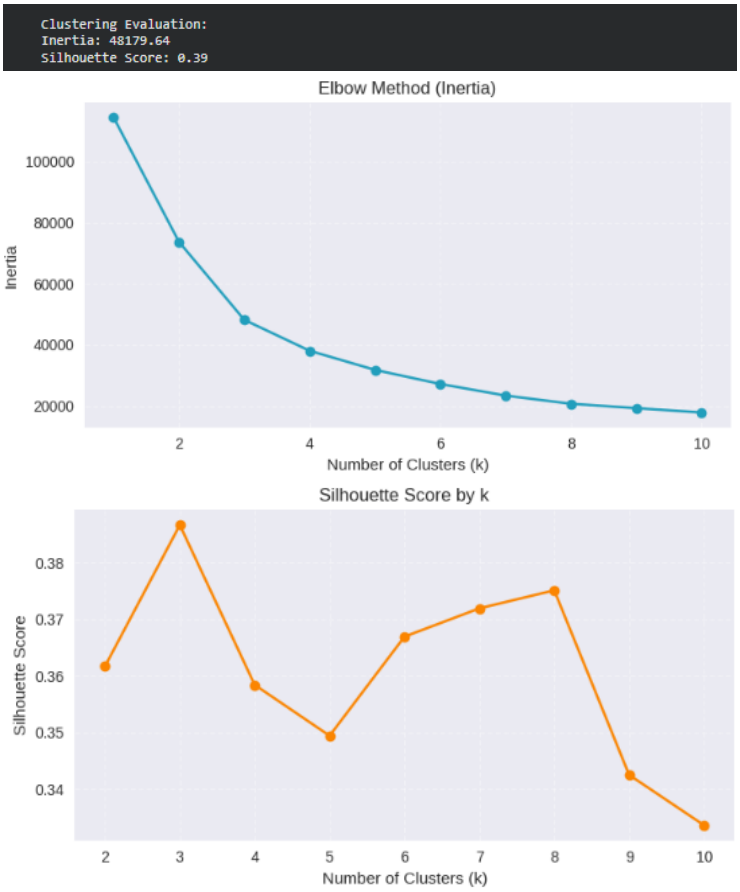


2. **‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA :**

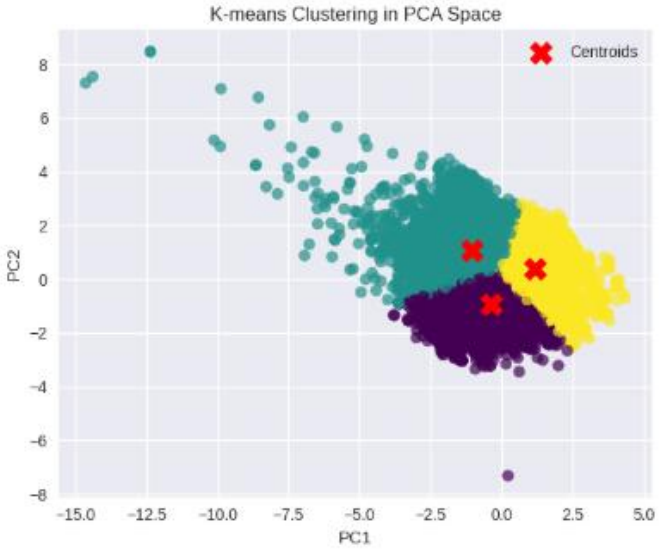
- **Explained Variance by Component:** This bar chart shows how much variance each principal component (PC) individually captures from the original data.
- **Data Distribution in PCA Space:** This scatter plot visualizes your dataset projected onto the first two principal components (PC1 and PC2).
- It helps to see the data's inherent structure and potential clusters in a reduced, 2D view.
- The first two components captured 28.12% of the total variance, indicating their importance for visualization.
- This visualization is crucial for understanding the effect of dimensionality reduction on data representation.



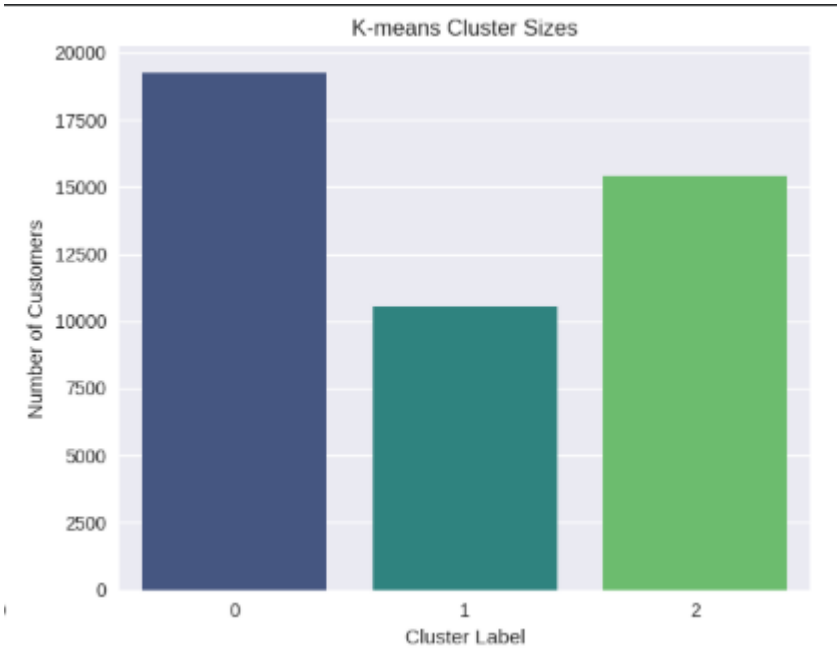
3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means



4. K-means Clustering Results with Centroids Visible (Scatter Plot) :



5. **K-means Cluster Sizes (Bar Plot):**



6. **Silhouette distribution per cluster for K-means (Box Plot) :**