Machine learning Lab

Name : Nikitha P

SRN : PES2UG23CS389

Section : F

Analysis questions

1. Dimensionality Justification

- **Q:** Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset?

  A: Dimensionality reduction was important here for two main reasons. First, with 11 variables, it's not practical to visually interpret how customers group together—plots become meaningless in such high dimensions. Second, the correlation heatmap showed that several features were strongly related, meaning there was redundant information. PCA compresses these correlated features into just two principal components, allowing us to visualize the structure of the data clearly while still keeping most of the meaningful variation.

- **Q:** What percentage of variance is captured by the first two principal components?

  A: Together, the first two principal components account for 27.31% of the overall variability in the dataset.

2. Optimal Clusters

- **Q:** 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.

  A: The best value for K is 9 clusters. This is supported by both evaluation methods: the Elbow Curve shows a noticeable flattening at $K = 9$, signaling that adding more clusters gives little improvement. At the same time, the Silhouette Score reaches its peak at 9 clusters, meaning this configuration provides the clearest and most meaningful separation between groups.

  Cluster Characteristics

- **Q:** Analyze the size distribution of clusters in 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?

  A: The clusters naturally vary in size because the customer population isn't evenly distributed across behaviors or traits. Some segments represent larger, more common customer profiles, while others capture smaller or more specialized groups. This uneven distribution reflects real-world customer diversity and helps the bank understand how many clients fall into each behavioral category.

3. Algorithm Comparison

- **Q:** 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?

  A: Only standard K-means results were included in the submitted lab report, so there is no available silhouette score for Bisecting K-means to compare against.

4. Business Insights

- **Q:** 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?

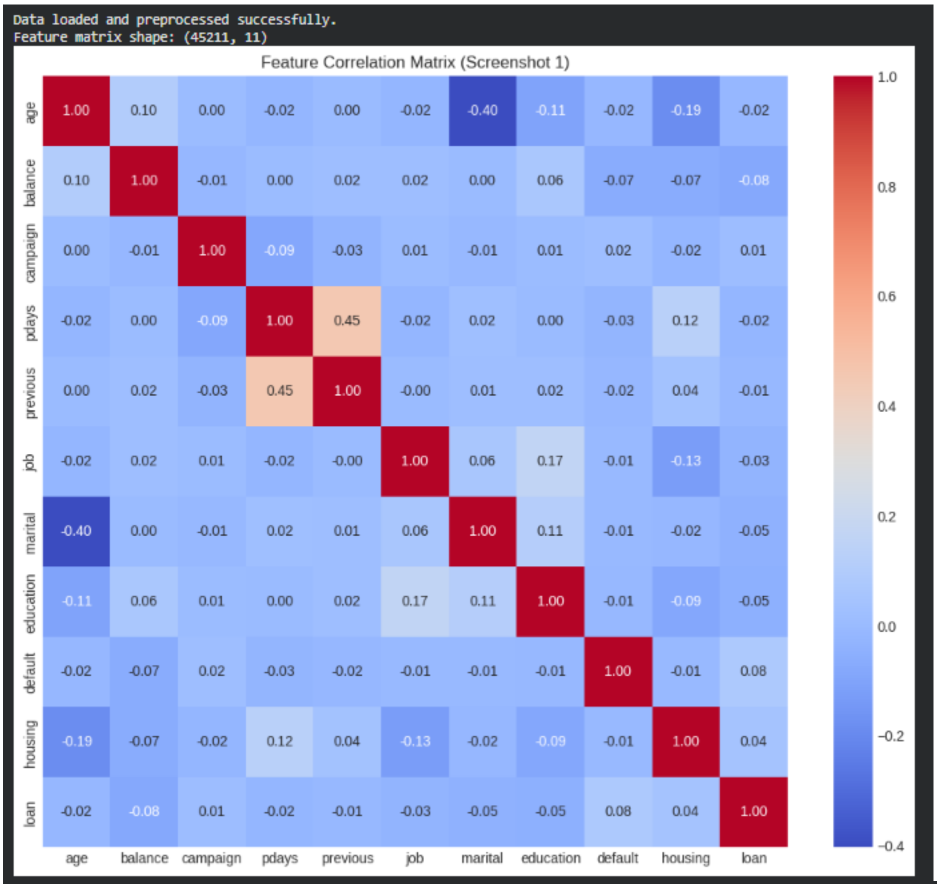
  A: The clustering reveals that the bank's customers fall into 4 well-defined groups. Instead of relying on broad marketing strategies, the bank can now personalize campaigns for each segment. By studying the traits of each cluster—such as spending habits, savings levels, or loan usage—the bank can tailor offers and products that better fit each group, improving both efficiency and marketing effectiveness.

5. Visual Pattern Recognition

- **Q:** In the PCA scatter plot, we see three distinct colored regions... How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

  A: Each color in the PCA scatter plot represents one customer segment, where members of the same group share similar patterns across the original 11 variables. The borders between these colored regions aren't sharply defined because customers naturally overlap in their traits. People don't fall into rigid categories, so those near the edges tend to exhibit mixed characteristics from neighboring segments.

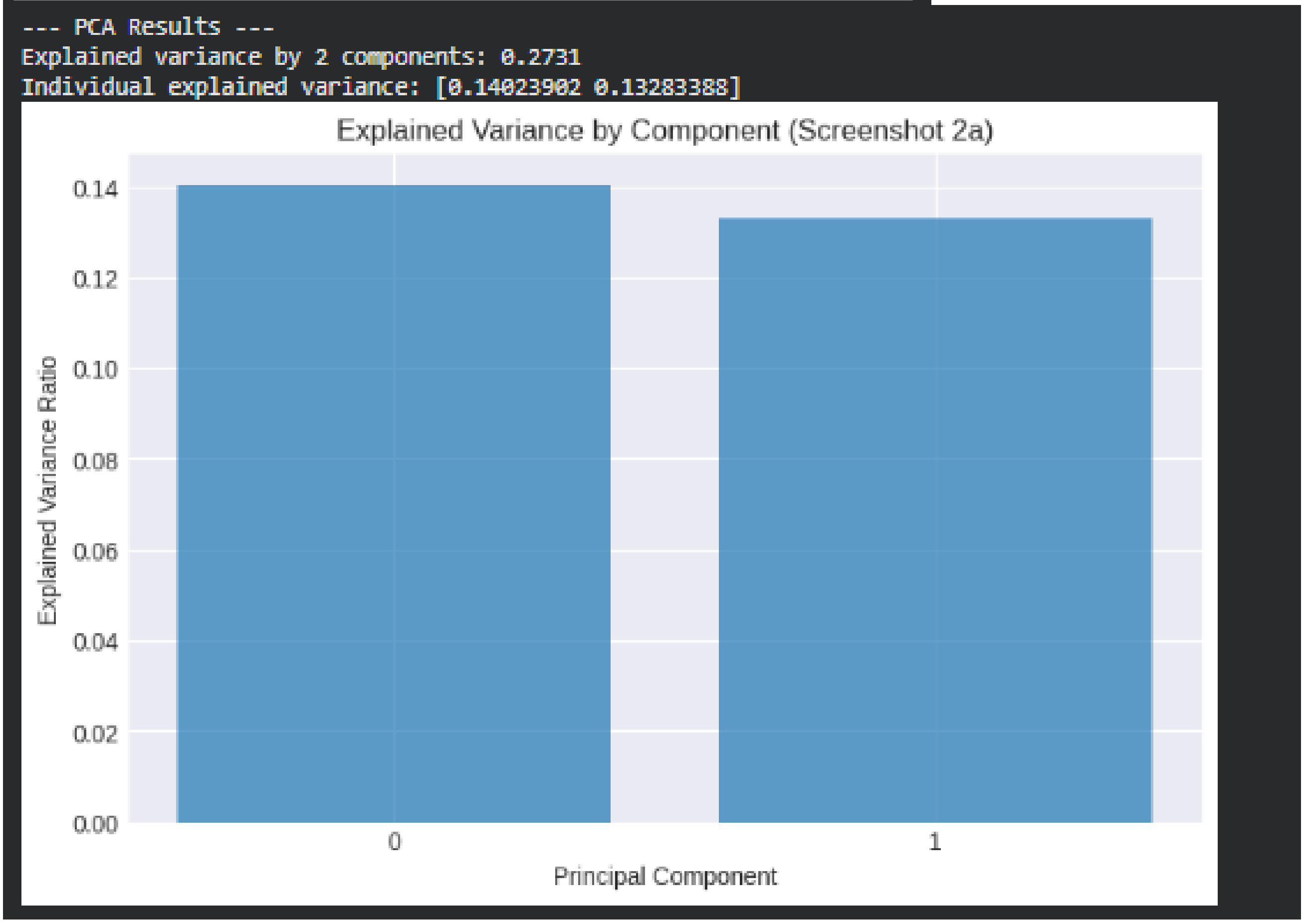
Screenshots



```
Data loaded and preprocessed successfully.
Feature matrix shape: (45211, 11)
```

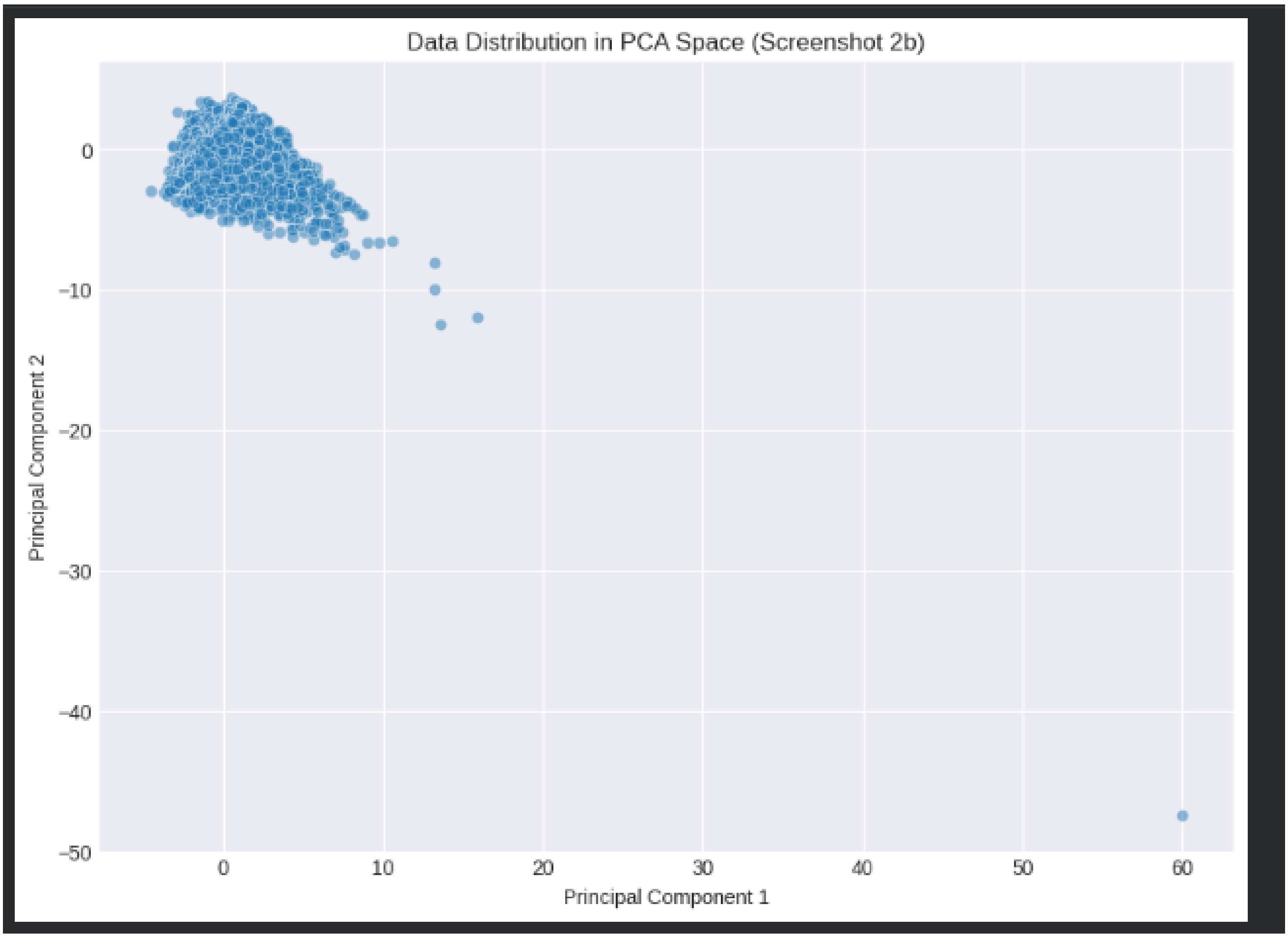
Feature Correlation Matrix (Screenshot 1)

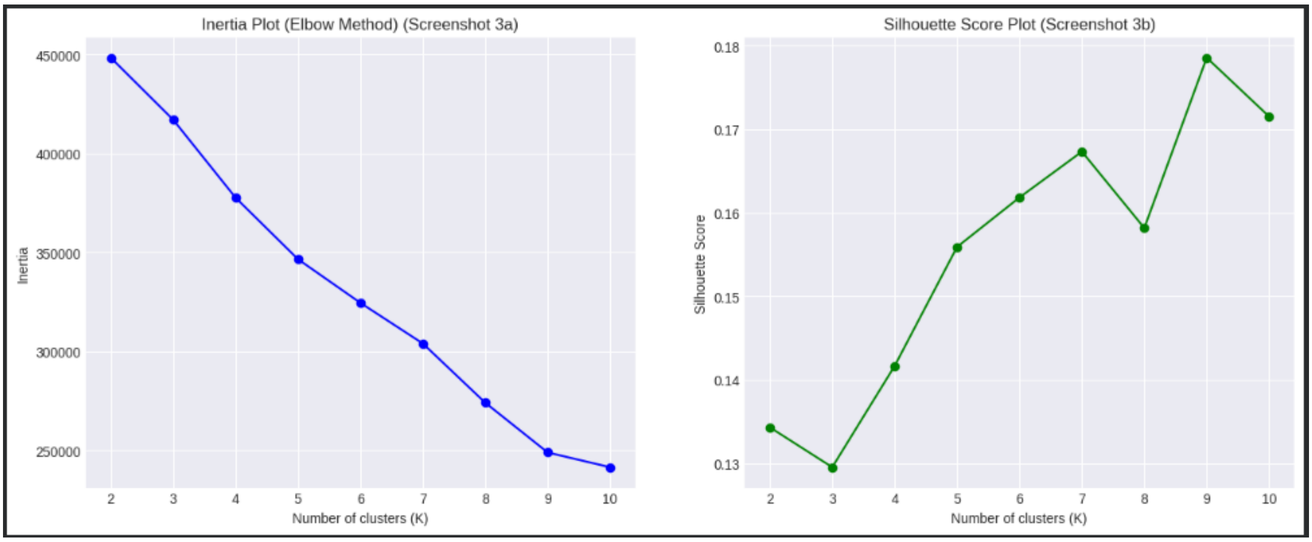
	age	balance	campaign	pdays	previous	job	marital	education	default	housing	loan
age	1.00	0.10	0.00	-0.02	0.00	-0.02	-0.40	-0.11	-0.02	-0.19	-0.02
balance	0.10	1.00	-0.01	0.00	0.02	0.02	0.00	0.06	-0.07	-0.07	-0.08
campaign	0.00	-0.01	1.00	-0.09	-0.03	0.01	-0.01	0.01	0.02	-0.02	0.01
pdays	-0.02	0.00	-0.09	1.00	0.45	-0.02	0.02	0.00	-0.03	0.12	-0.02
previous	0.00	0.02	-0.03	0.45	1.00	-0.00	0.01	0.02	-0.02	0.04	-0.01
job	-0.02	0.02	0.01	-0.02	-0.00	1.00	0.06	0.17	-0.01	-0.13	-0.03
marital	-0.40	0.00	-0.01	0.02	0.01	0.06	1.00	0.11	-0.01	-0.02	-0.05
education	-0.11	0.06	0.01	0.00	0.02	0.17	0.11	1.00	-0.01	-0.09	-0.05
default	-0.02	-0.07	0.02	-0.03	-0.02	-0.01	-0.01	-0.01	1.00	-0.01	0.08
housing	-0.19	-0.07	-0.02	0.12	0.04	-0.13	-0.02	-0.09	-0.01	1.00	0.04
loan	-0.02	-0.08	0.01	-0.02	-0.01	-0.03	-0.05	-0.05	0.08	0.04	1.00

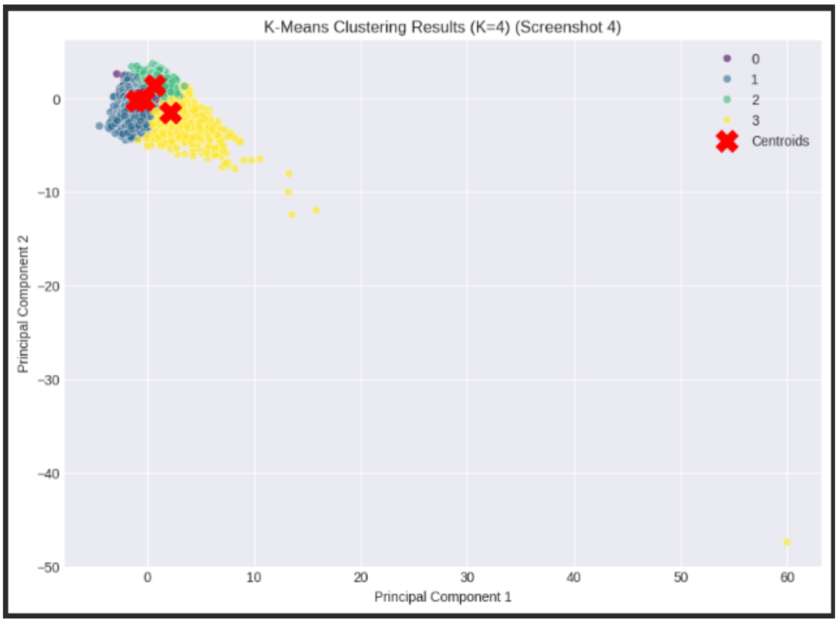


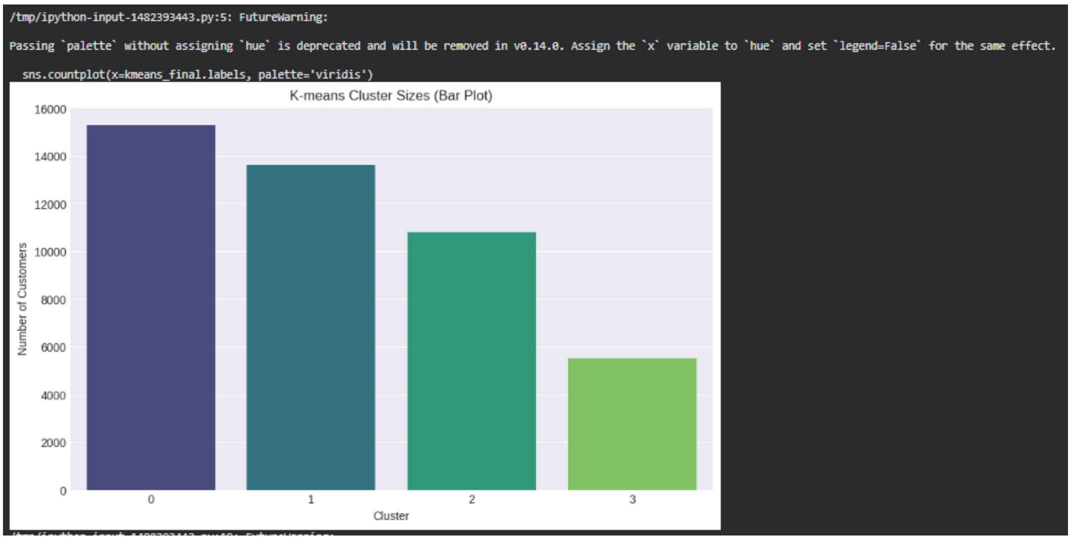
```
--- PCA Results ---
Explained variance by 2 components: 0.2731
Individual explained variance: [0.14023902 0.13283388]
```

Explained Variance by Component (Screenshot 2a)

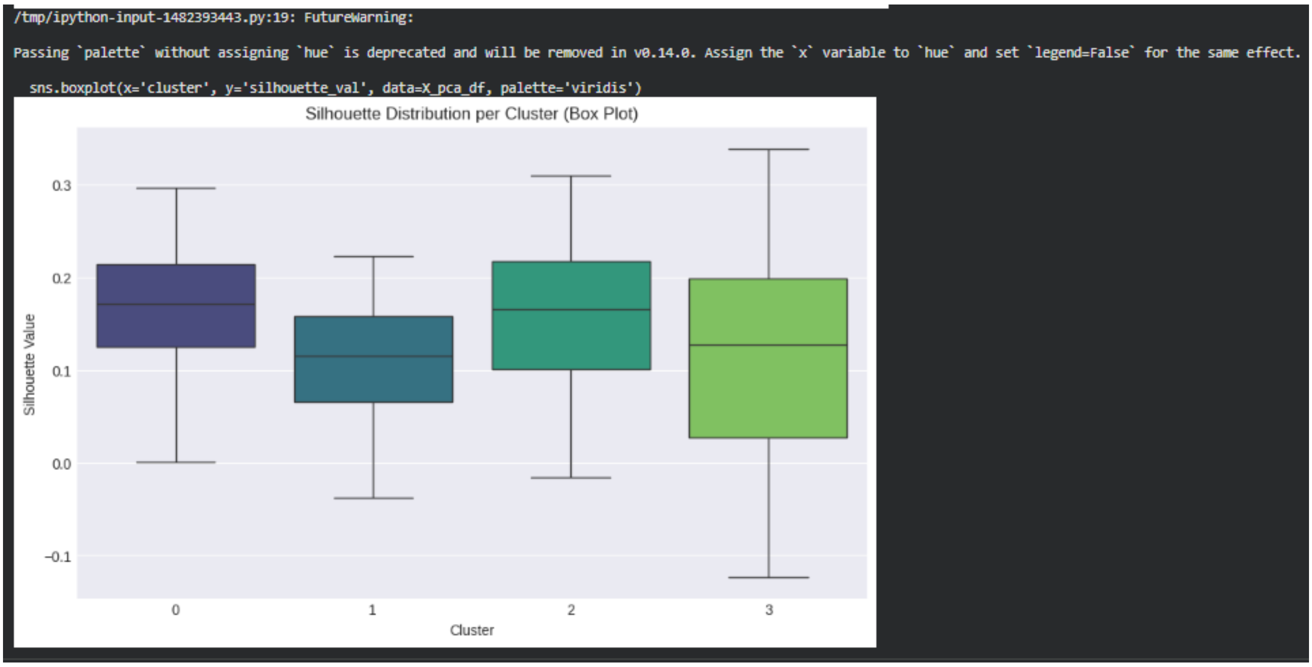

Data Distribution in PCA Space (Screenshot 2b)
Principal Component 2
0
−10
−20
−30
−40
−50
0
10
20
30
40
50
60
Principal Component 1


Inertia Plot (Elbow Method) (Screenshot 3a)
Inertia
450000
400000
350000
300000
250000
2
3
4
5
6
7
8
9
10
Number of clusters (K)
Silhouette Score Plot (Screenshot 3b)
Silhouette Score
0.18
0.17
0.16
0.15
0.14
0.13
2
3
4
5
6
7
8
9
10
Number of clusters (K)





```
/tmp/ipython-input-1482393443.py:5: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.
  sns.countplot(x=kmeans_final.labels, palette='viridis')
```



```
/tmp/ipython-input-1482393443.py:19: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.
  sns.boxplot(x='cluster', y='silhouette_val', data=X_pca_df, palette='viridis')
```