11/11/2025

MACHINE LEARNING LAB WEEK 13

NAME: CHERUKURI VENKATA KARTIK
SRN: PES2UG23CS148
SECTION: C

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?

- **PC1 ≈ 15%**, **PC2 ≈ 13%**, and
- **Cumulative ≈ 28%** of total variance is captured by the first two components
- The captured variance is average , PCA helps project corelated financial features into a low dimensional plane
- PCA simplified the dataset from 9+ features to 2 dimentions.

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.

When we see the elbow curve , we can see that the inertia decreases drastically after k=3,the silhouette score is highest when k=3. Hence optimal k value is 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?

- k=3 produces one large dense cluster and two smaller clusters
- k=4 creates four clusters forming four clearer and distinct groups

- Larger clusters represent dominant customer base with loan behaviors and bank balances while smaller clusters represents specialised groups
- Uneven cluster size says that the data is skewed towards common banking profiles

---

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?

- Silhouette score for k=3 is 0.39 while silhouette score for k=4 turned out to be 0.33
- K-Means slightly outperforms Bisecting K-Means here, since increasing k to 4 added overlap without improving cohesion

---

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?

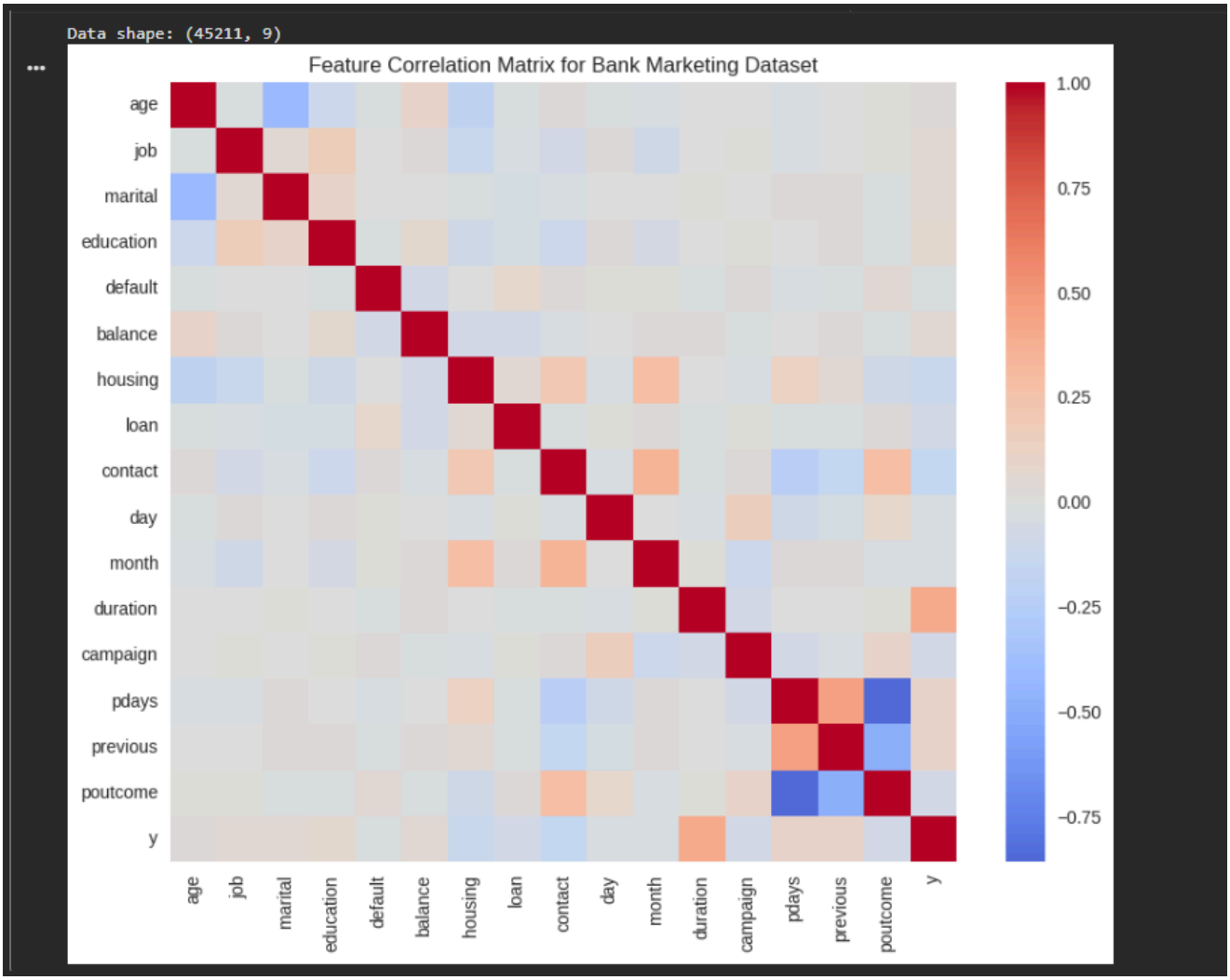
- **Cluster 0:** Majority of customers, middle-income, stable balances, existing housing loans : investment banking , money saving strategies
- **Cluster 1:** High-balance, low-loan group: premium account upgradation
- **Cluster 2:** Low-balance, younger customers : saving plans, credit cards, student loans etc
- **Cluster 3 :** Transitional customers with fluctuating balances: retention campaigns to retain such customers

---

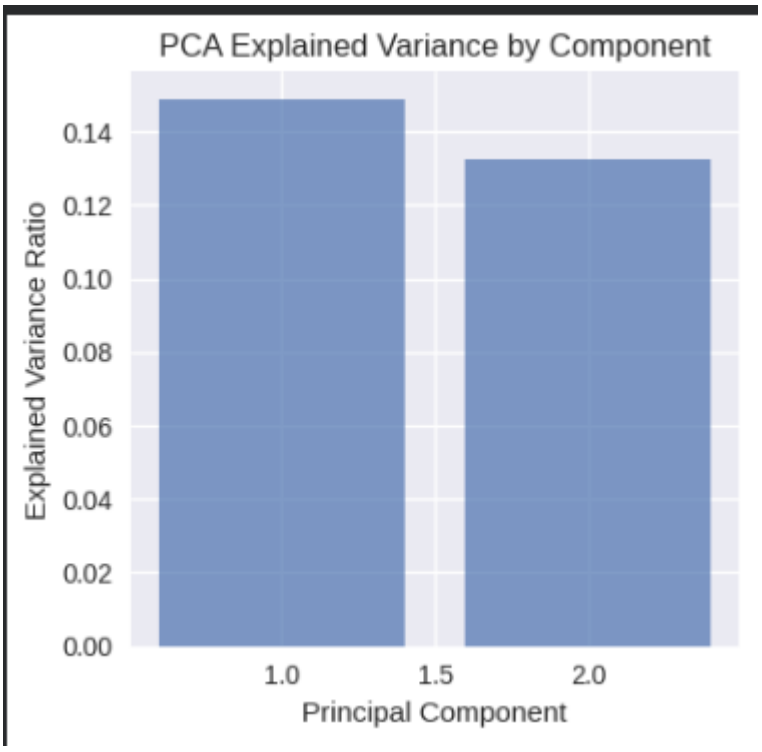
Visual Pattern Recognition: In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

- The PCA scatter plots display three main colored regions (turquoise, yellow, purple).
- Overlap indicates gradual transaction between customer groups with similar financial profile
- Sharp boundaries occur between high and low balance segments showing strong feature influence

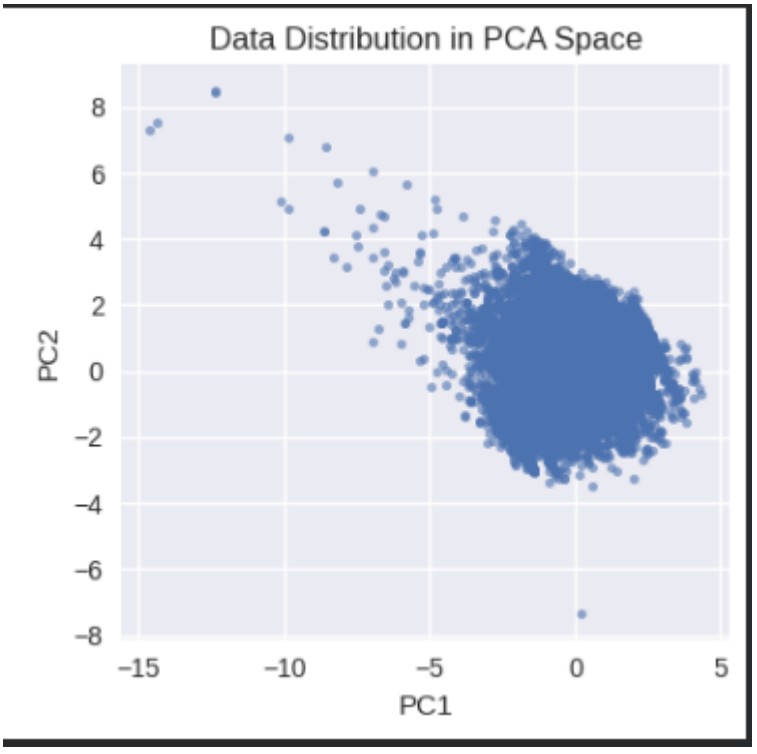
FEATURE COREALTION MATRIX



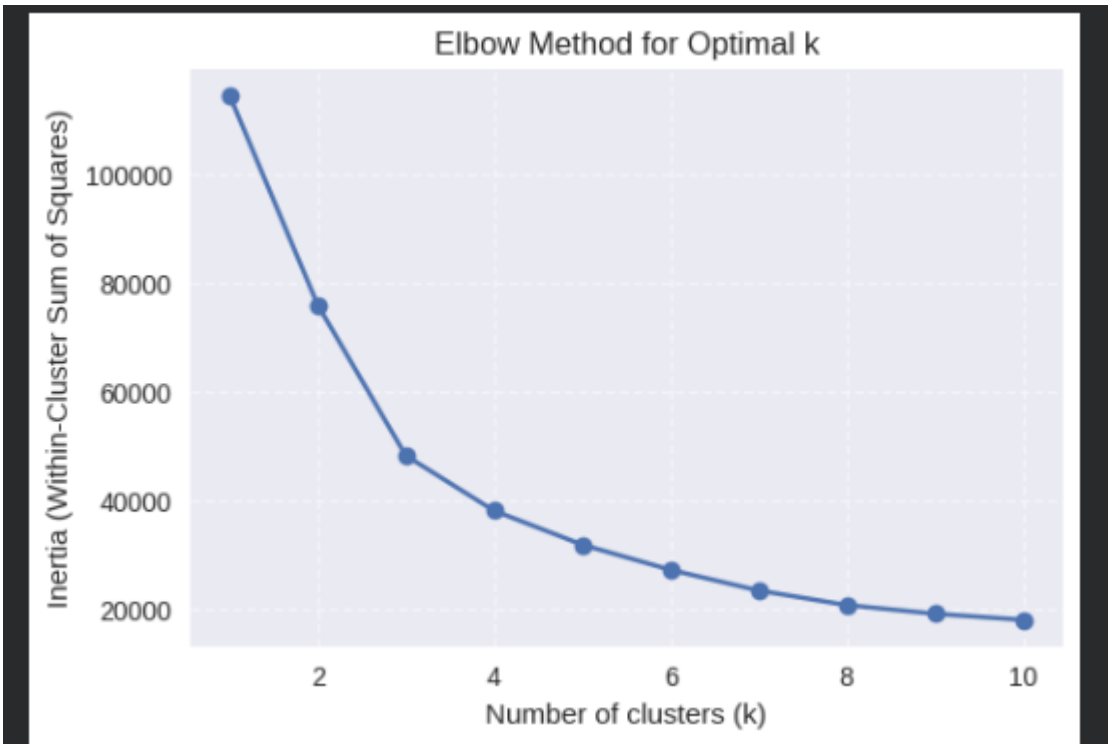
EXPLAINED VARIANCE BY COMPONENT



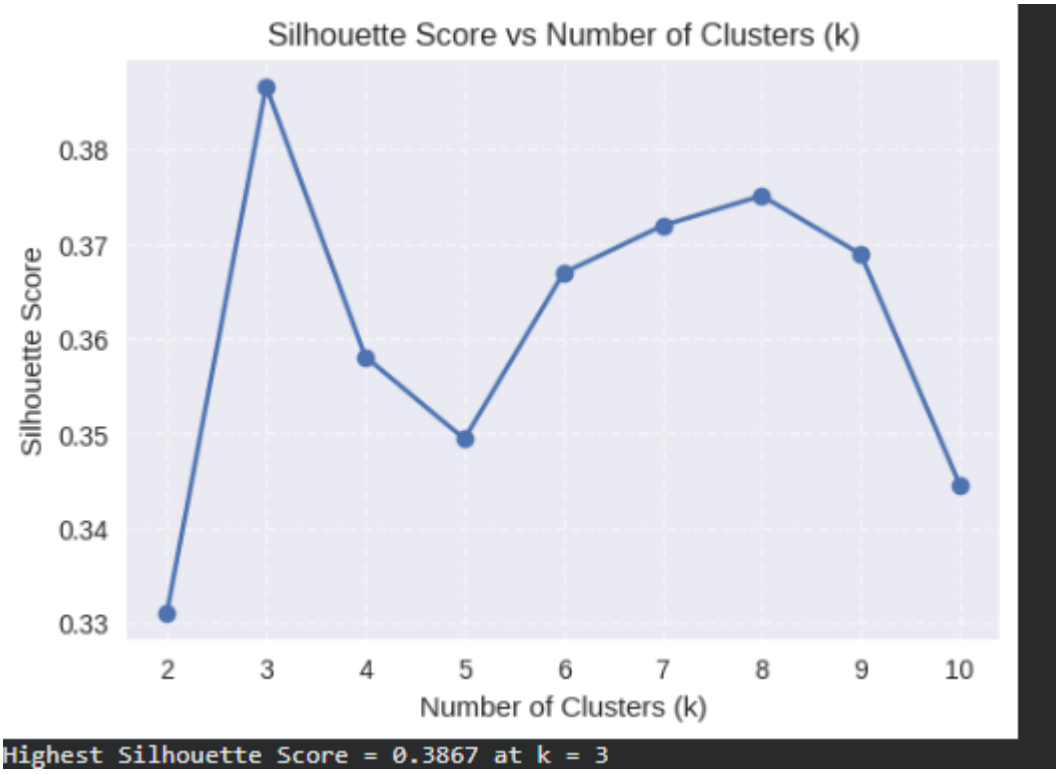
DATA DISTRIBUTION IN PCA SPACE



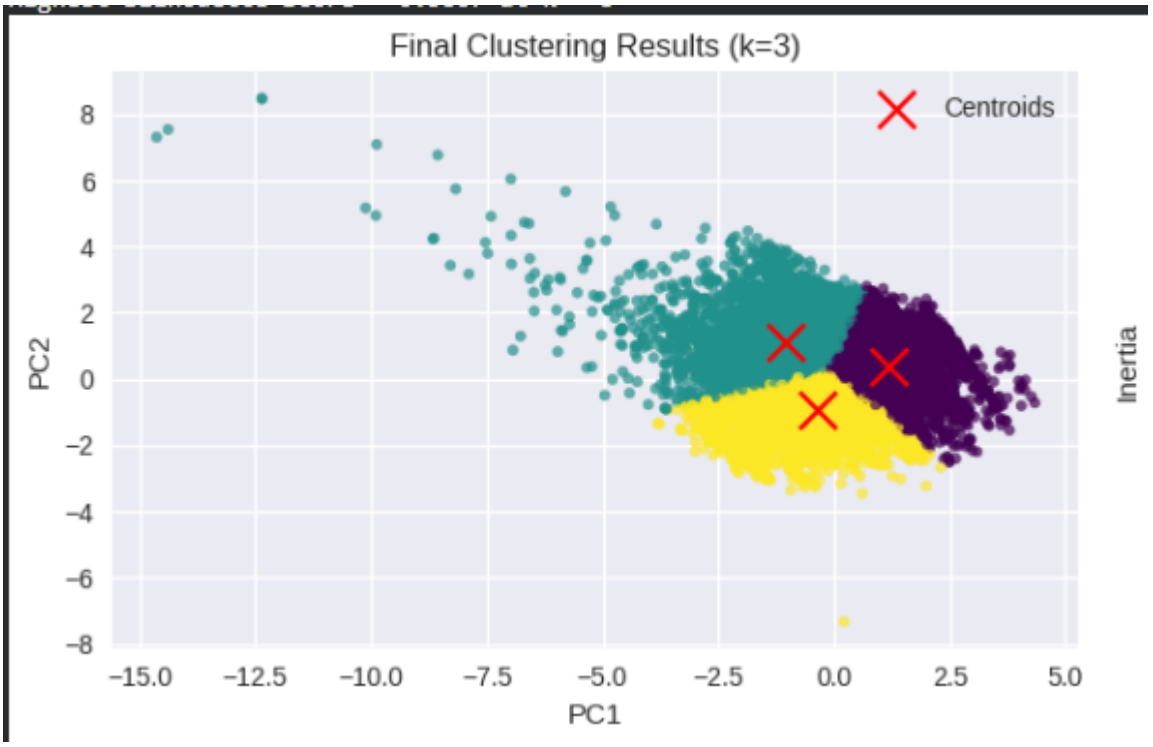
K VALUE USING ELBOW METHOD



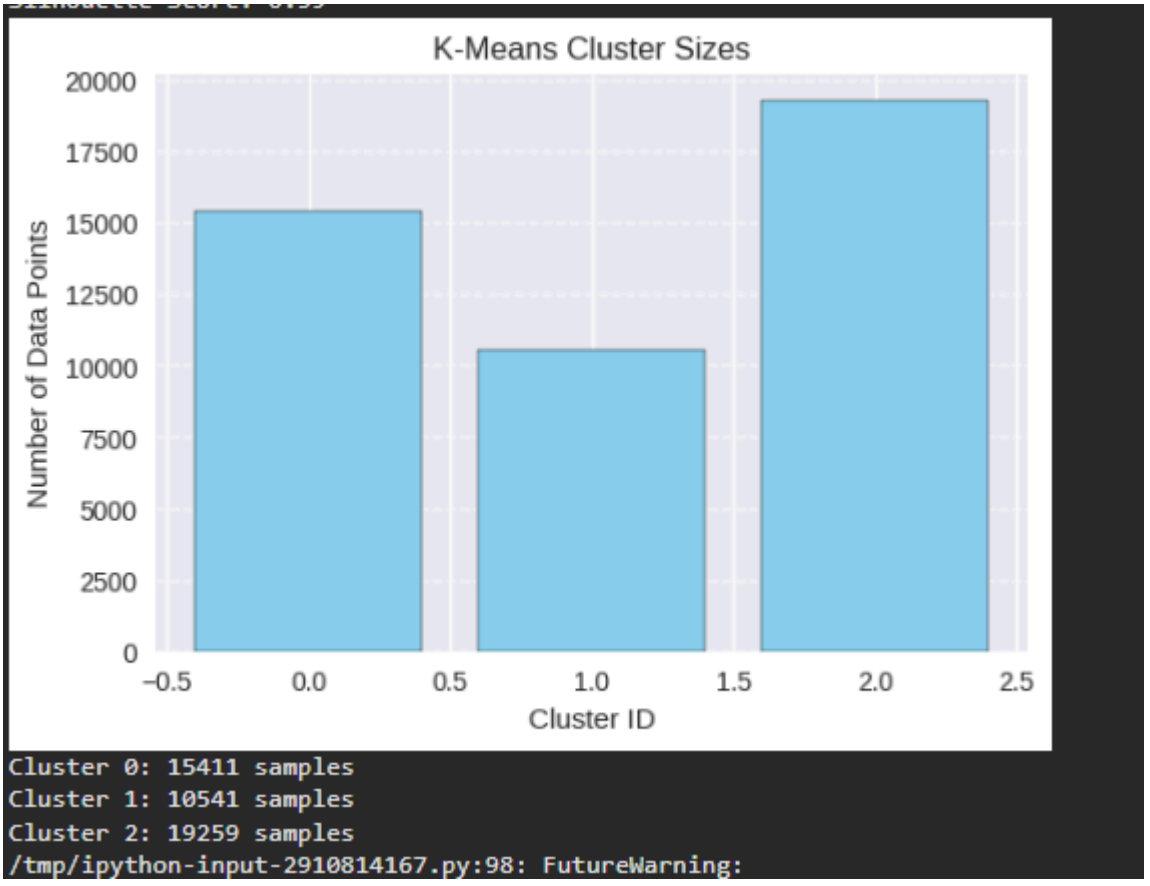
SILHOUETTE SCORE PLOT



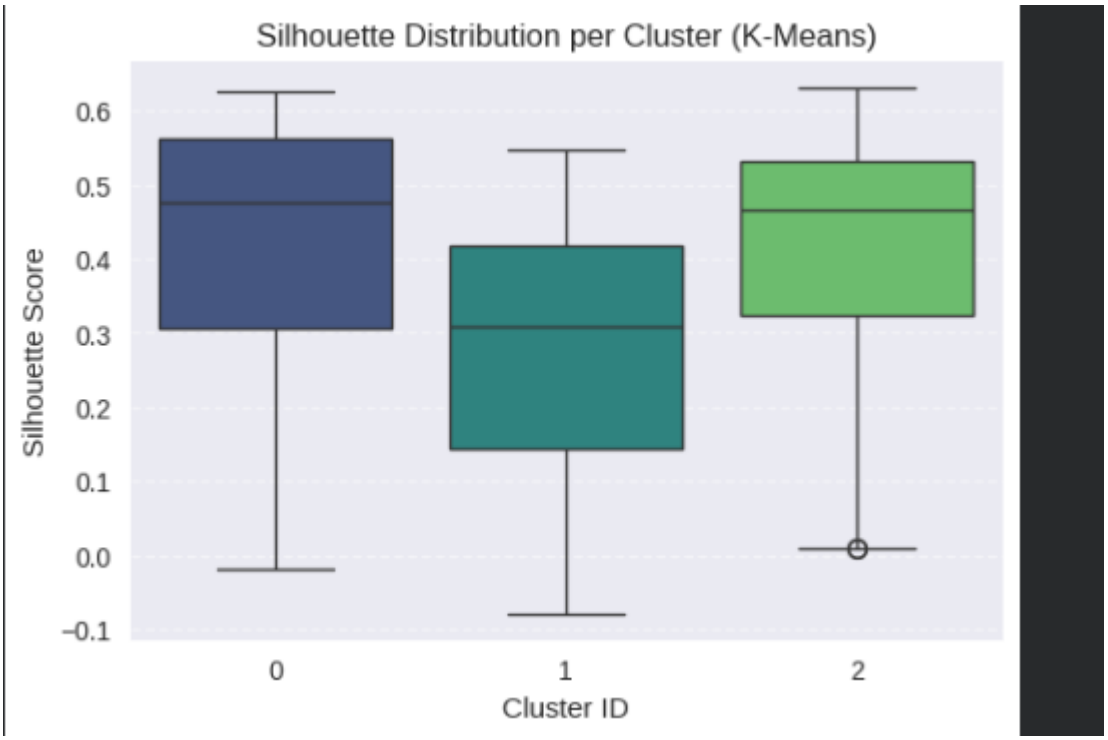
K-MEANS CLUSTERING WITH CENTROIDS



K-MEANS CLUSTER SIZES



SILHOUETTE DISTRIBUTION PER CLUSTER FOR K MEANS


Silhouette Distribution per Cluster (K-Means)
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster ID