📘 Silhouette Analysis (Silhouette Clustering)

1️⃣ WHY SILHOUETTE EXISTS

After clustering, we always ask:

> **“Are these clusters actually good?”**

Problems:

- K-Means needs **k** → how to choose it?
- DBSCAN gives clusters → how to **evaluate** them?
- Labels exist, but **no ground truth**

👉 **Silhouette score measures how well each point fits in its cluster compared to others.**

2️⃣ CORE IDEA (ONE LINE)

> **A point is good if it is close to its own cluster and far from other clusters.**

Silhouette captures exactly this.

3️⃣ TWO DISTANCES YOU MUST UNDERSTAND

For **each data point i**:

🔹 a(i) — *Intra-cluster distance*

- Average distance from point **i** to all other points in **its own cluster**
- Measures **compactness**

👉 Smaller = better

🔹 **b(i) — *Nearest-cluster distance***

- Average distance from point **i** to points in the **nearest neighboring cluster**
- Measures **separation**

👉 Larger = better

4️⃣ SILHOUETTE FORMULA (VERY IMPORTANT)

✅ Silhouette value for a single point (PLAIN TEXT)

```
s(i) = ( b(i) - a(i) ) / max( a(i), b(i) )
```

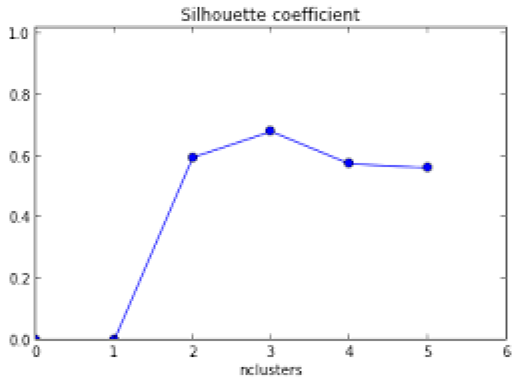
✅ Range (PLAIN TEXT)

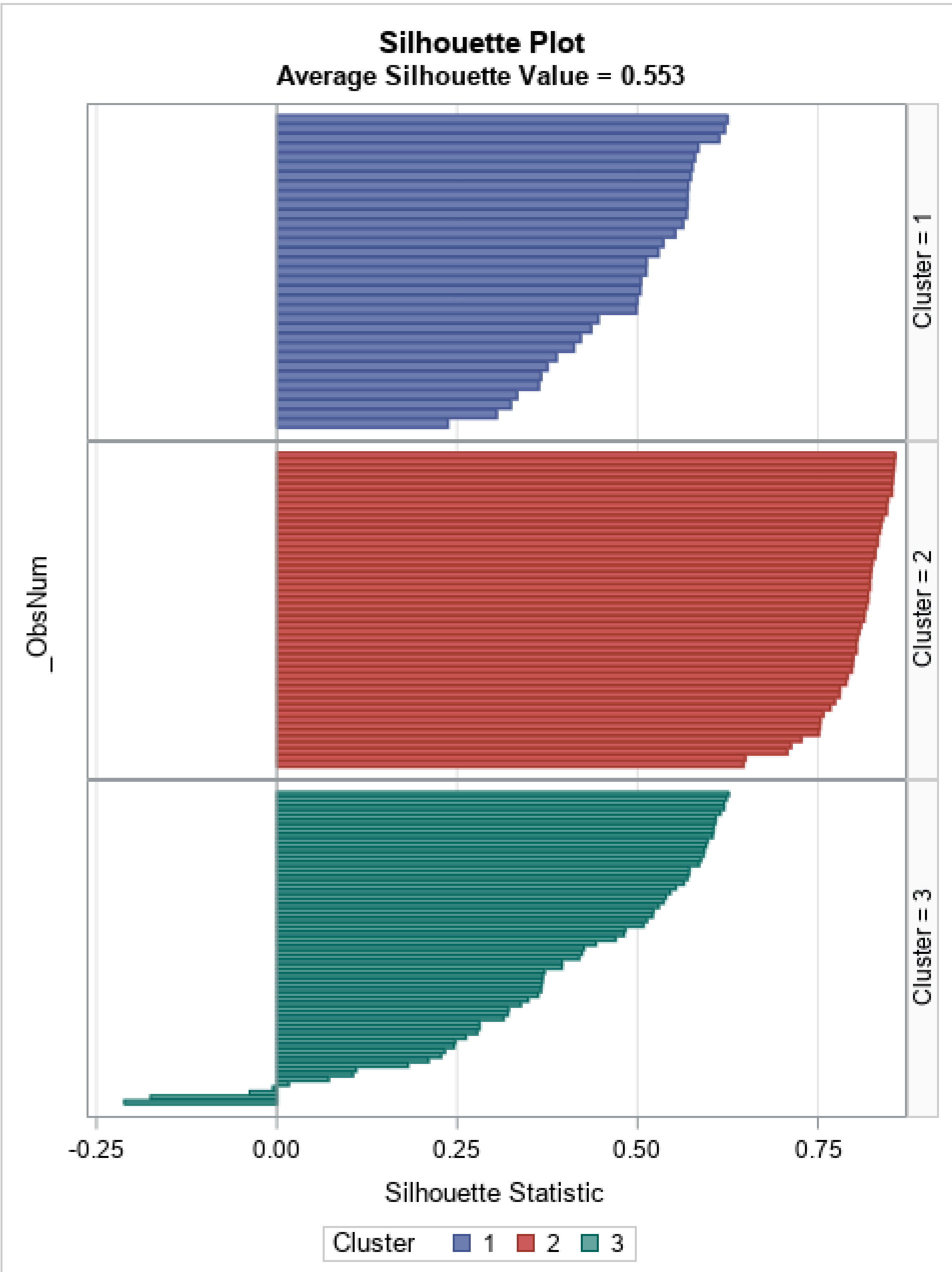
```
-1 <= s(i) <= 1
```

5️⃣ HOW TO INTERPRET THE VALUE

Silhouette value	Meaning
≈ +1	Perfectly clustered
≈ 0	On decision boundary
< 0	Probably misclustered

Visual intuition 👇


Silhouette coefficient
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
nclusters


Silhouette Plot
Average Silhouette Value = 0.553
_ObsNum
Cluster = 1
Cluster = 2
Cluster = 3
-0.25
0.00
0.25
0.50
0.75
Silhouette Statistic
Cluster
1
2
3

6️⃣ OVERALL SILHOUETTE SCORE

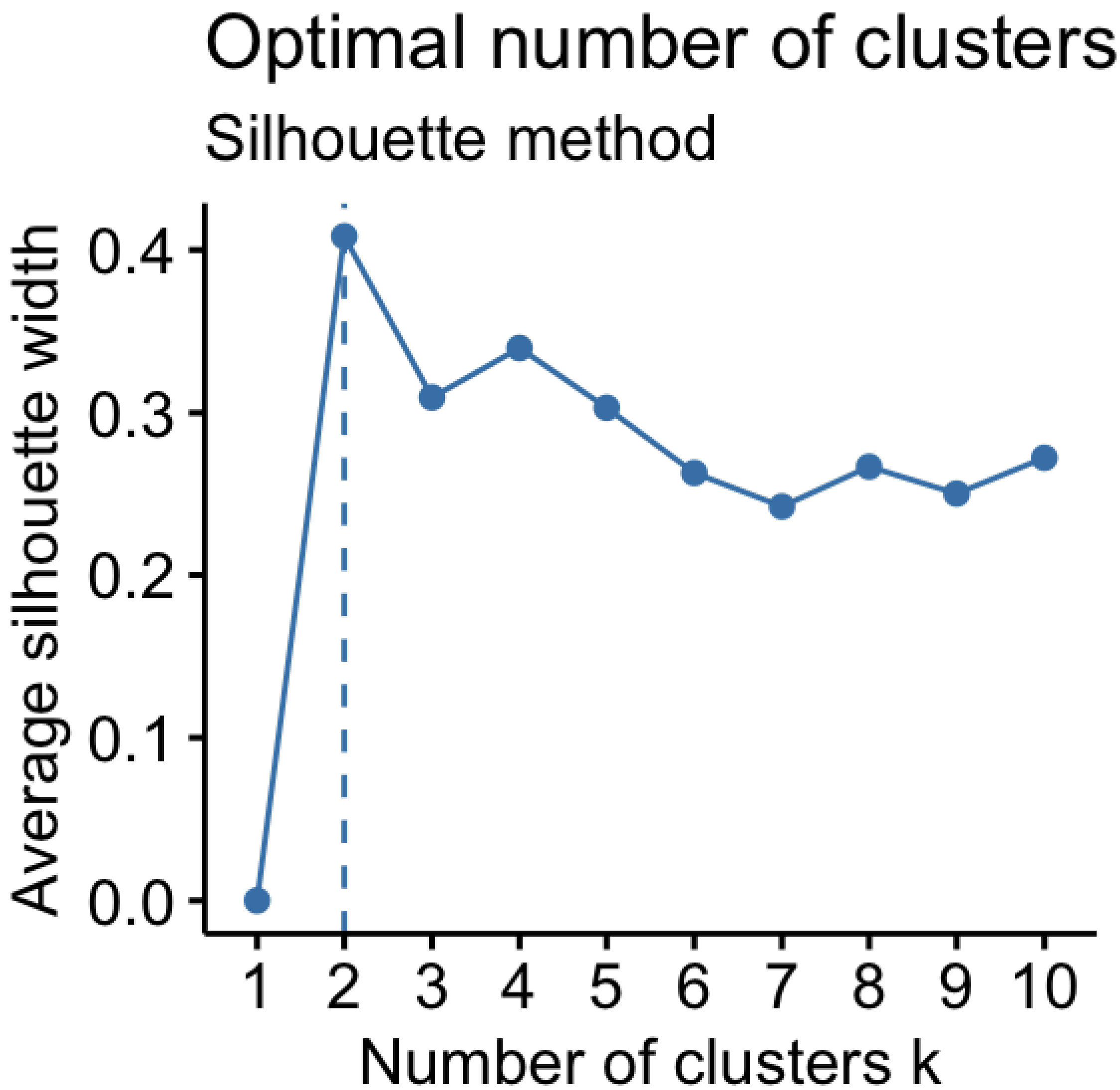
- Compute **s(i)** for every point
- Take the **mean**

Silhouette Score=1n∑s(i)\text{Silhouette Score} = \frac{1}{n}\sum s(i)Silhouette Score=n1∑s(i)

📌 Used to:

- Compare different **k**
- Compare different clustering algorithms

7️⃣ USING SILHOUETTE TO CHOOSE k (VERY COMMON)



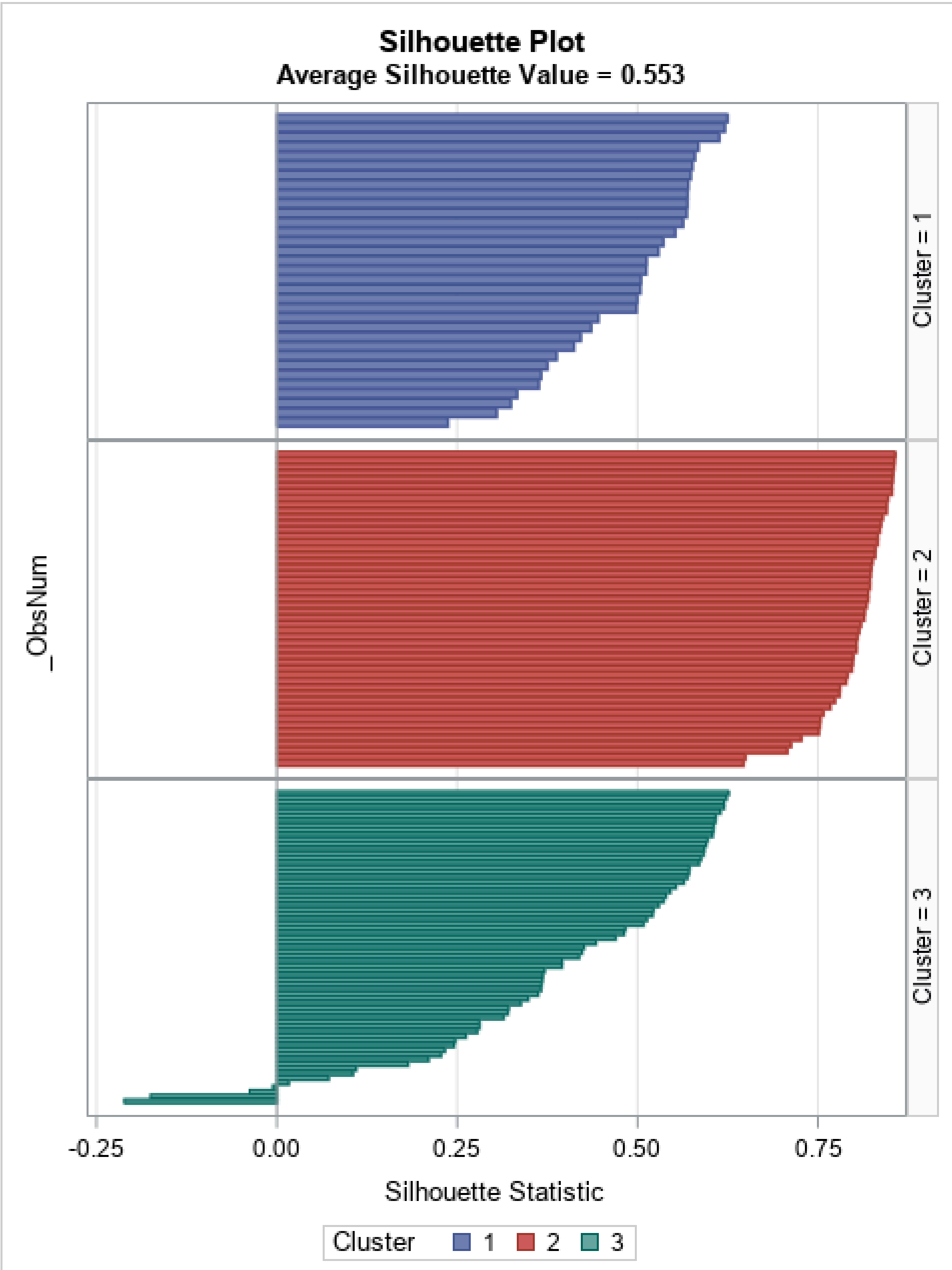
Steps:

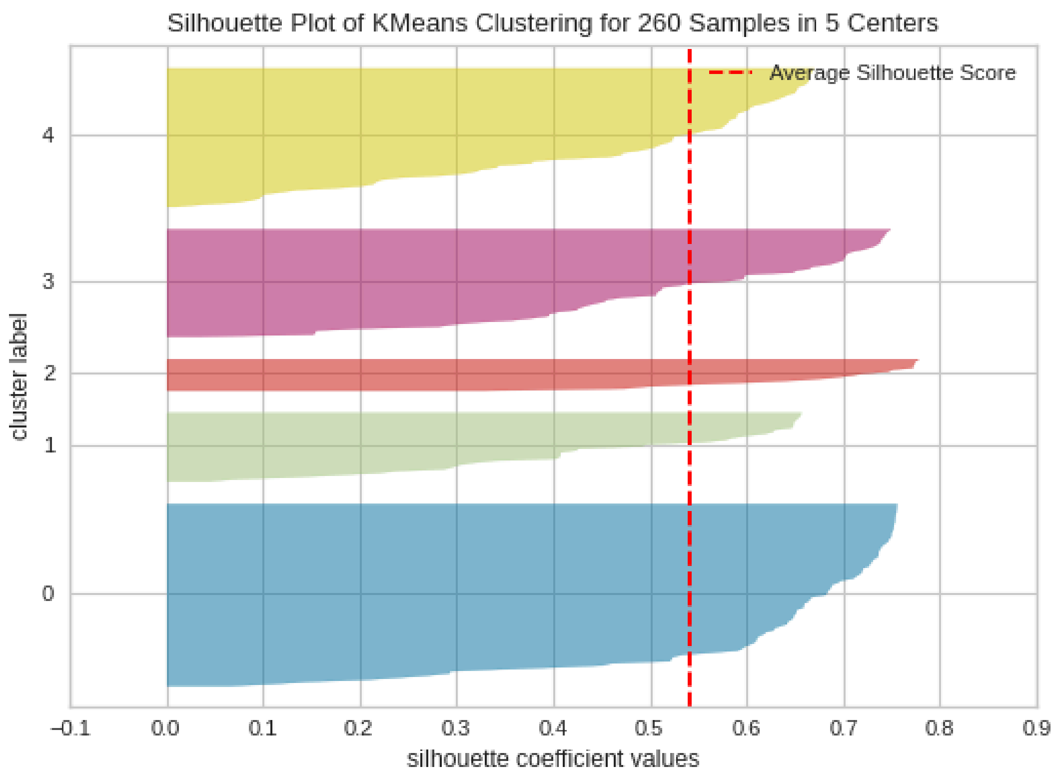
1. Try k = 2, 3, 4, ...
2. Compute silhouette score for each
3. Choose **k with highest score**

👉 Unlike elbow method, silhouette has a **clear numeric meaning**

---

8 SILHOUETTE PLOT (INTERVIEW FAVORITE)





What it shows:

- Each bar = one point
- Width = silhouette value
- Grouped by cluster

Good plot:

- Mostly positive bars
- Similar heights across clusters

9️⃣ WHEN SILHOUETTE WORKS BEST

✅ Convex / well-separated clusters
✅ K-Means, Agglomerative clustering
✅ Low–medium dimensional data

🔟 WHEN SILHOUETTE FAILS (IMPORTANT)

❌ Non-convex clusters (DBSCAN shapes)
❌ Varying densities
❌ High-dimensional data (distance meaningless)
❌ Single cluster (k = 1 → undefined)

1️⃣1️⃣ SILHOUETTE WITH DBSCAN

⚠️ Careful here:

- DBSCAN labels **noise as -1**
- Silhouette **cannot handle noise directly**

Correct approach:

```
mask = labels != -1
silhouette_score(X[mask], labels[mask])
```

📌 Even then:

- DBSCAN clusters are **shape-based**
- Silhouette may **underestimate quality**

1️⃣2️⃣ SILHOUETTE vs ELBOW (VERY COMMON)

Feature	Silhouette	Elbow
Metric meaning	Clear (−1 to 1)	Heuristic

Works without k	❌	❌
Cluster separation	✅	❌
Interpretability	High	Medium

1️⃣3️⃣ SKLEARN CODE (MINIMAL)

```
from sklearn.metrics import silhouette_score

score = silhouette_score(X, labels)
print(score)
```

For plotting:

```
from sklearn.metrics import silhouette_samples
```

1️⃣4️⃣ EXAM / INTERVIEW ANSWER (PERFECT)

> Silhouette score evaluates clustering quality by comparing intra-cluster compactness and inter-cluster separation. Values close to +1 indicate well-separated clusters, while negative values suggest misclassification.

1️⃣5️⃣ ONE-LINE MEMORY TRICK 🧠

> **Silhouette asks: "Am I closer to my own group than to others?"**

🔑 FINAL SUMMARY

- Uses **distance**
- No ground truth needed

- Range **−1 to +1**
- Best for **choosing k**
- Weak for **DBSCAN & complex shapes**