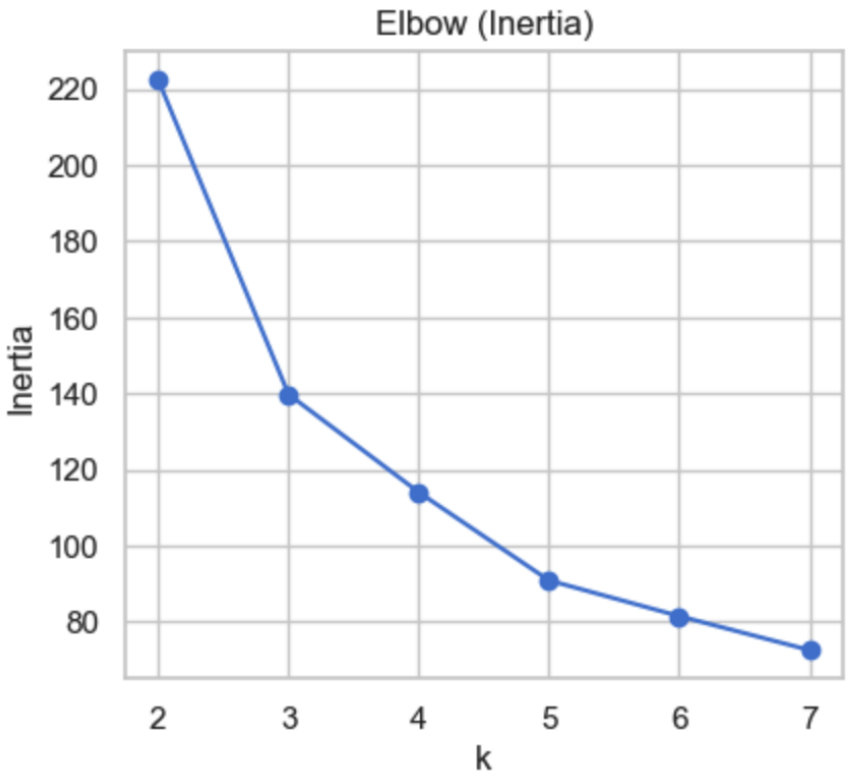
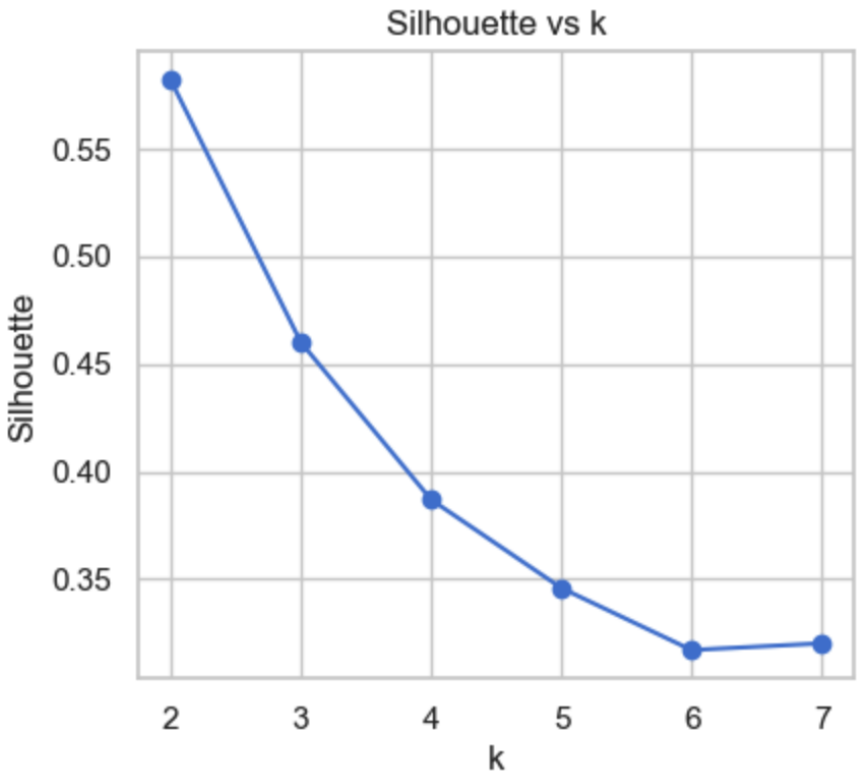
Clustering Iris Flowers with K-Means

Dataset & Preparation

- **Dataset**: 150 samples, 4 numeric features (sepal length/width, petal length/width), 3 true species (hidden during clustering).
- **StandardScaler:** giving each equal weight in distance calculations. This is usefull for K-Means since it’s using Euclidean Distance.

Number of Clusters (k)





Number of Clusters (k)

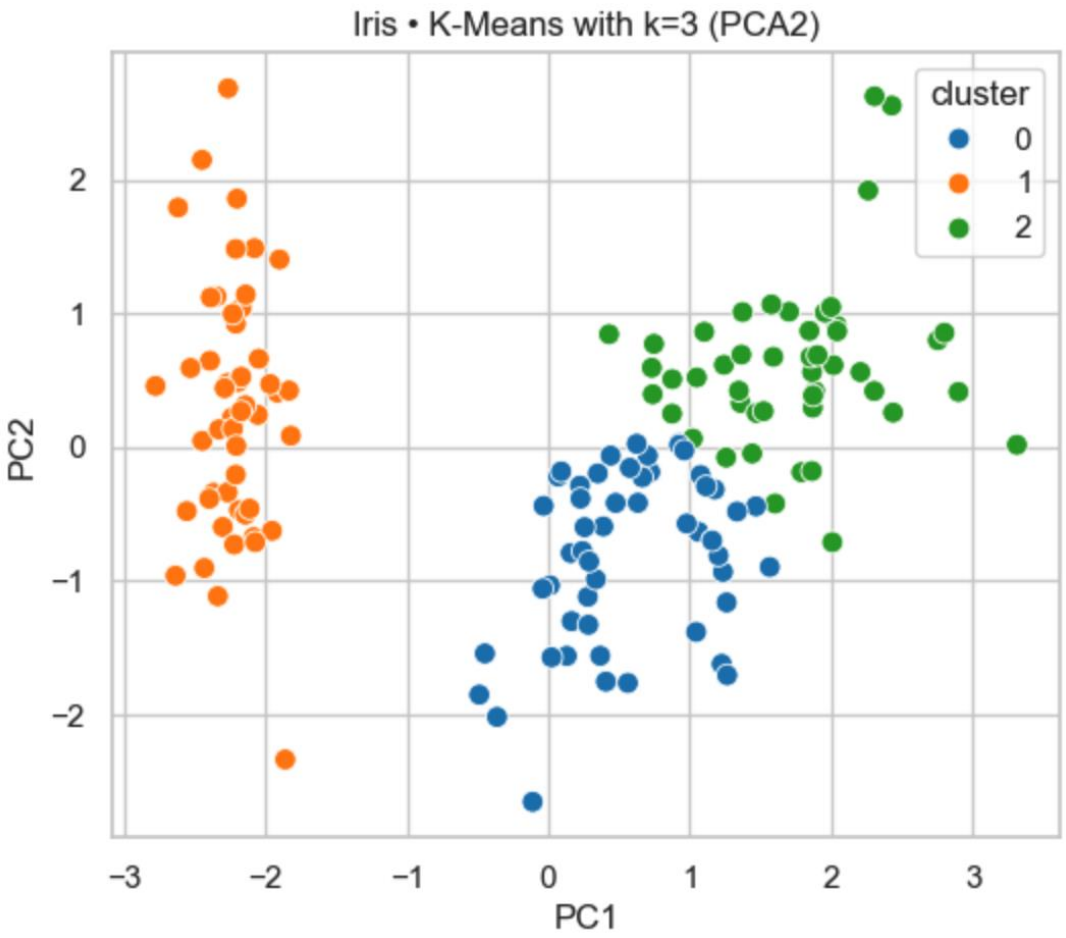
Tested cluster range between **2 to 7** using two methods:

- **Elbow method (inertia)** → drop until k=3, flattens afterwards.
- **Silhouette score** → top score at k=2, decreased at k=3 but still strong, then declined.
- **Decision**: chose **k=3** because it aligns with biological truth (3 species).

Running K-Means

- Ran **KMeans(n_clusters=3, n_init=25, random_state=42)**.
- Multiple initializations ensured stability and reproducibility.
- Labels assigned for each flower.

Visualization with **PCA**



Visualization with **PCA**

- Reduced 4 feature space into 2D PCA (PC1 & PC2), capturing ~95% of variance.
- Scatter plot colored by clusters showed:

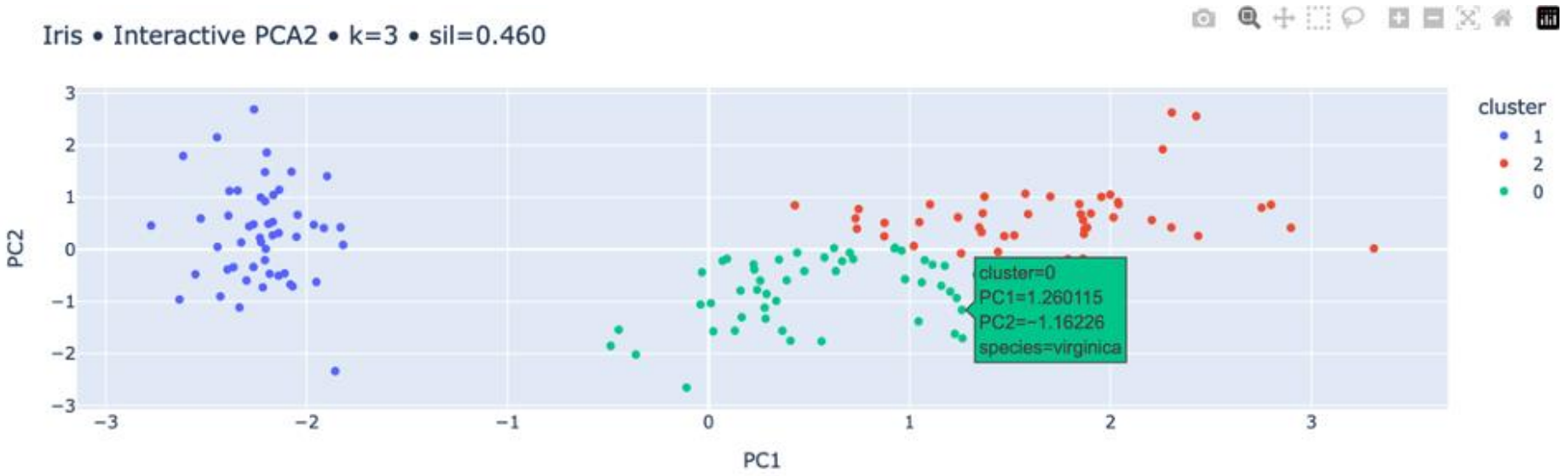
Cluster 1 (Setosa)

Cluster 2 (Versicolor)

Cluster 3 (Virginica) Interpretation

- PCA confirmed Setosa is distinct, while the other two species naturally overlap.

Interactive Visualization



Interactive Visualization

- Used method: **Plotly scatter** for PCA2D.
- Features:

Color = cluster

Hover tooltip = true species

- Makes results explorable & intuitive:
- See misclassifications instantly.
- Compare clusters to species interactively.
- Turns analysis into storytelling tool.

Evaluation

- **Silhouette score**: 0.46
- **Adjusted Rand Index (ARI)**: 0.62
- **Cluster sizes**: balanced (~50 each).

Takeaways

- Using two metrics (elbow, silhouette) to choose k.
- Using Standard Scaling method and PCA dimensionality reduction which are critical for clustering.
- Visualizations (static and interactive) make results clear to any audience.