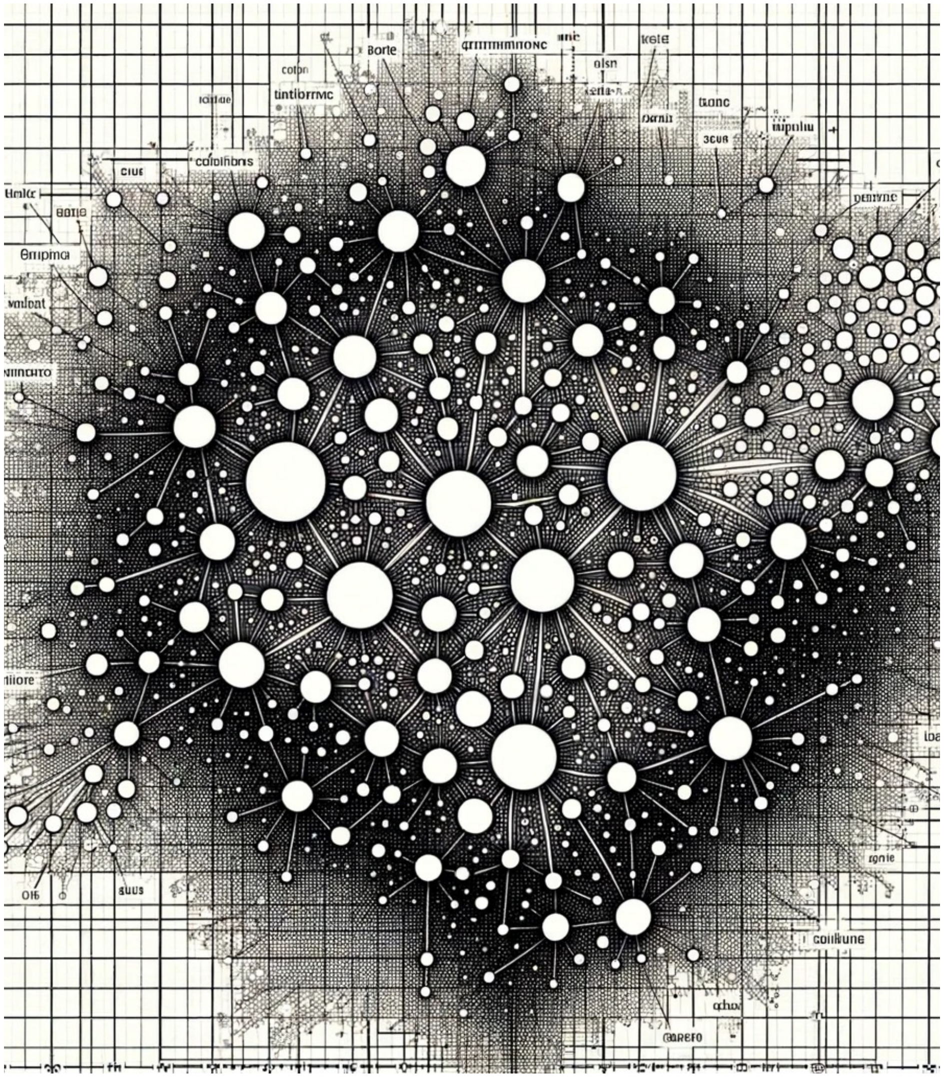


Comparison of Strategies to Inform K-means Clustering

Chris Taylor

DTSA 5510

Frame the Problem

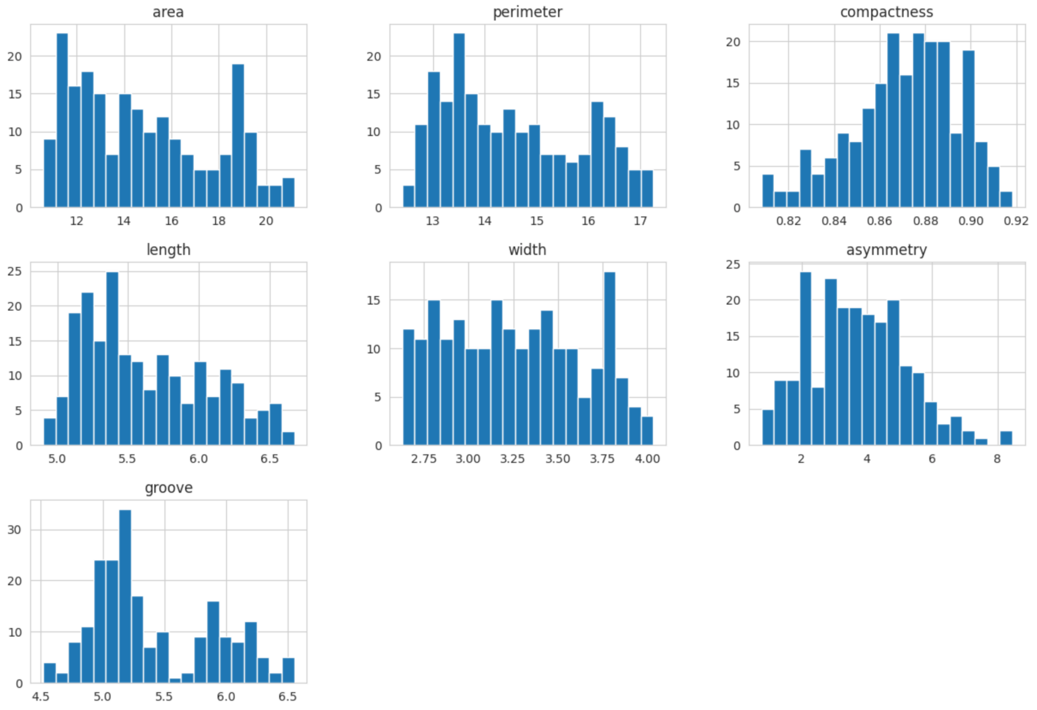
- K-means requires selection of k before implementing the algorithm
- Two popular methods for selecting k
  - Elbow Plot
    - “Coarse” according to Géron (2021)
  - Silhouette Score
- How does the use of each method impact the algorithm?

Exploratory Data Analysis

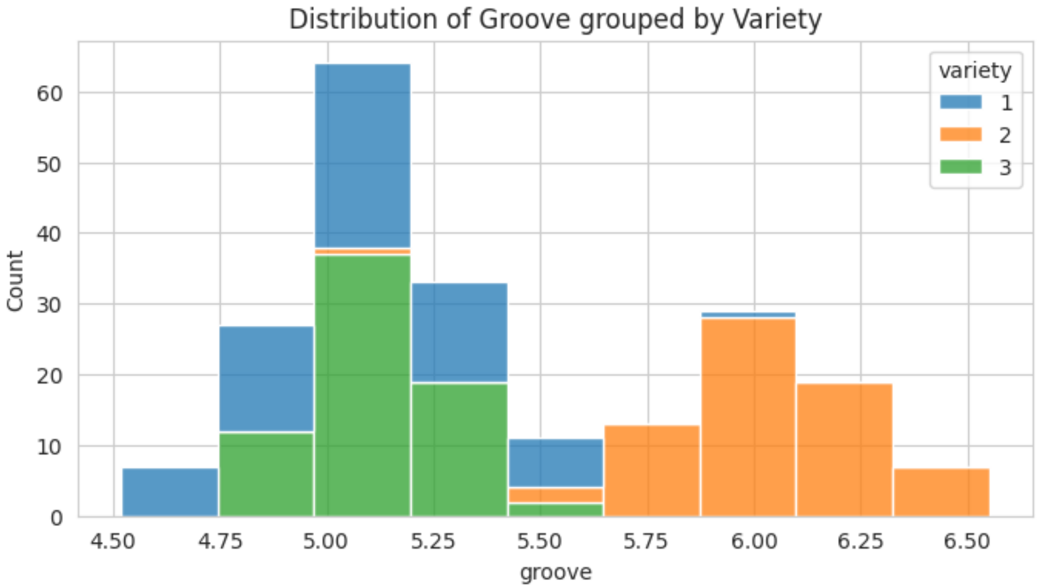
Description of Data

- 210 observations
- Seven geometric properties
  - Area
  - Perimeter
  - Compactness
  - Kernel length
  - Kernel width
  - Asymmetry coefficient
  - Kernel groove length

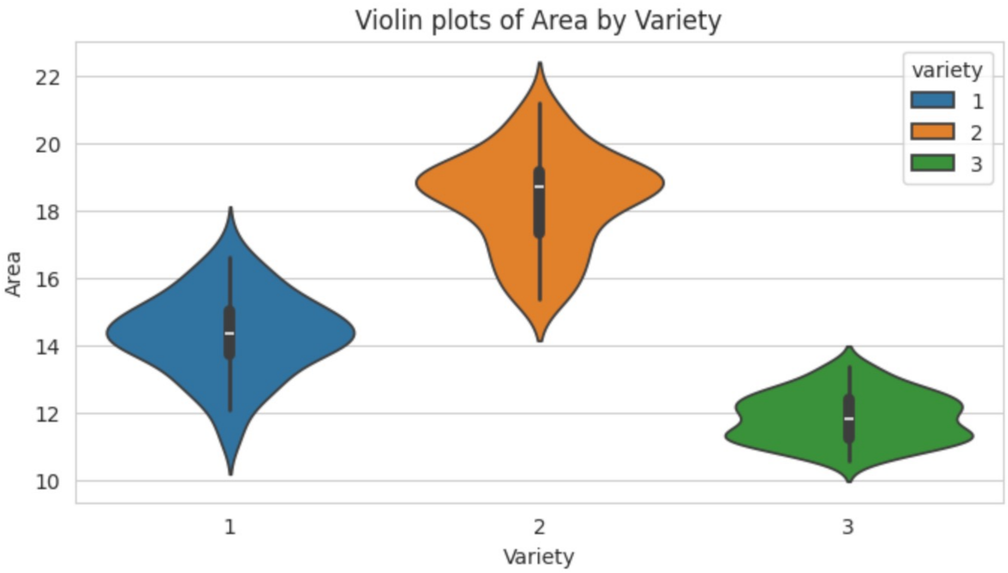
Histogram of Features



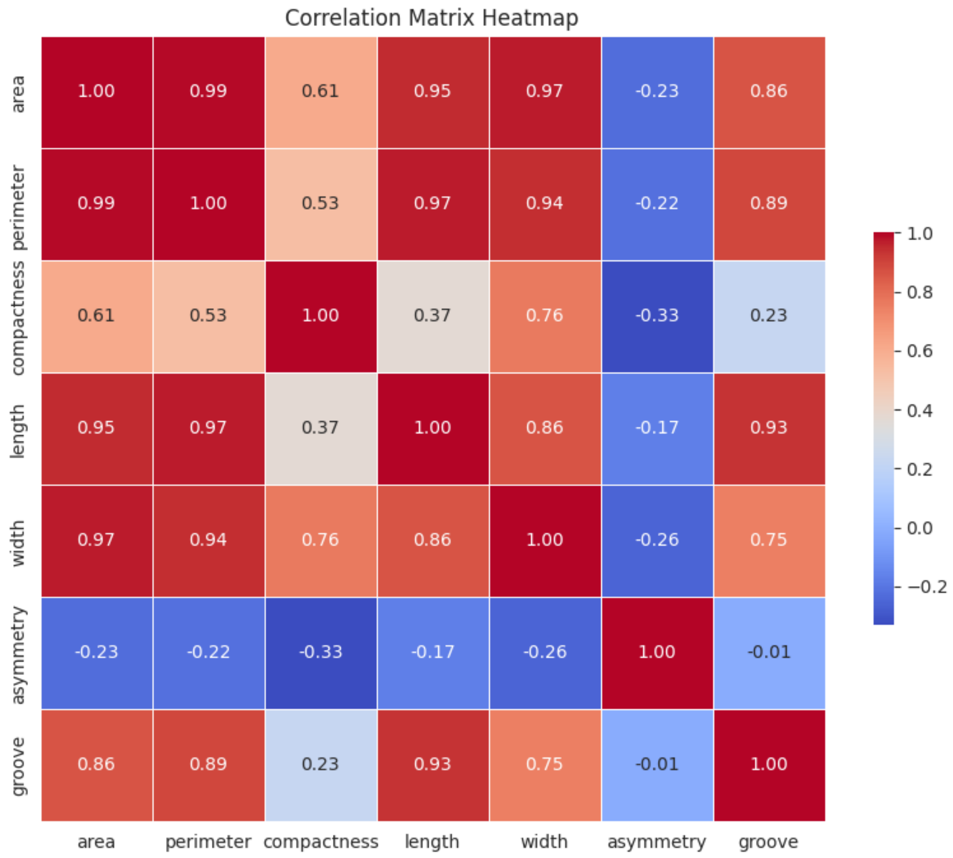
Histogram of Features



Violin Plot of Features



Correlation Between Features



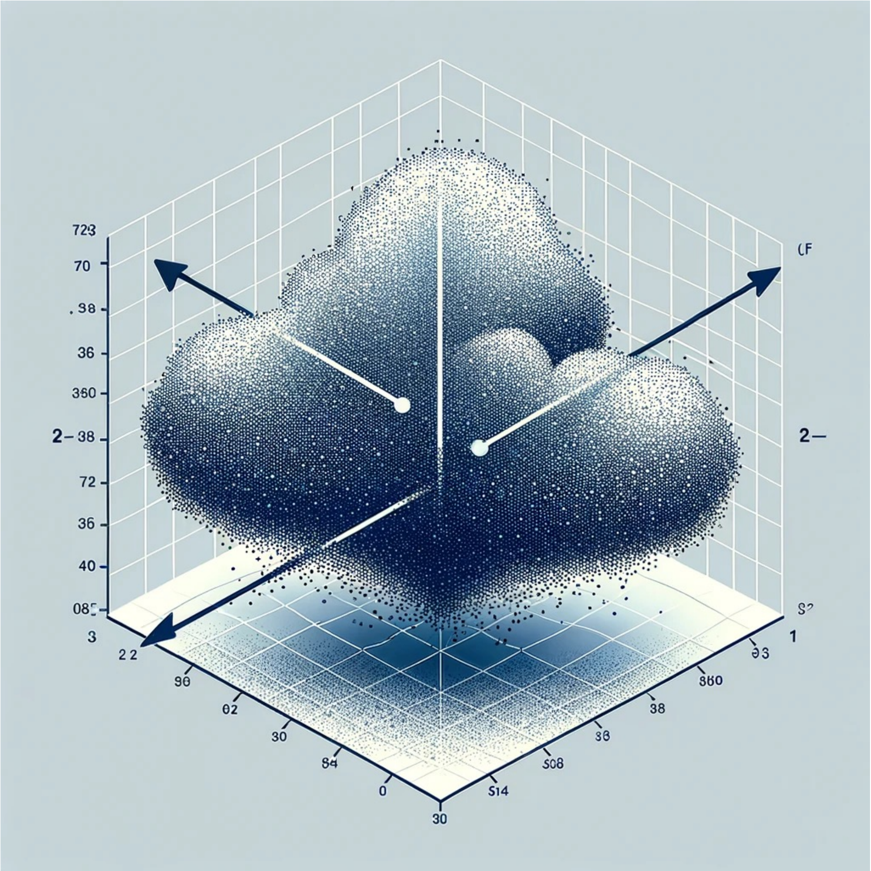
	area	perimeter	compactness	length	width	asymmetry	groove
area	1.00	0.99	0.61	0.95	0.97	-0.23	0.86
perimeter	0.99	1.00	0.53	0.97	0.94	-0.22	0.89
compactness	0.61	0.53	1.00	0.37	0.76	-0.33	0.23
length	0.95	0.97	0.37	1.00	0.86	-0.17	0.93
width	0.97	0.94	0.76	0.86	1.00	-0.26	0.75
asymmetry	-0.23	-0.22	-0.33	-0.17	-0.26	1.00	-0.01
groove	0.86	0.89	0.23	0.93	0.75	-0.01	1.00

Data Transformation

$$z = \frac{(x - \mu)}{\sigma}$$

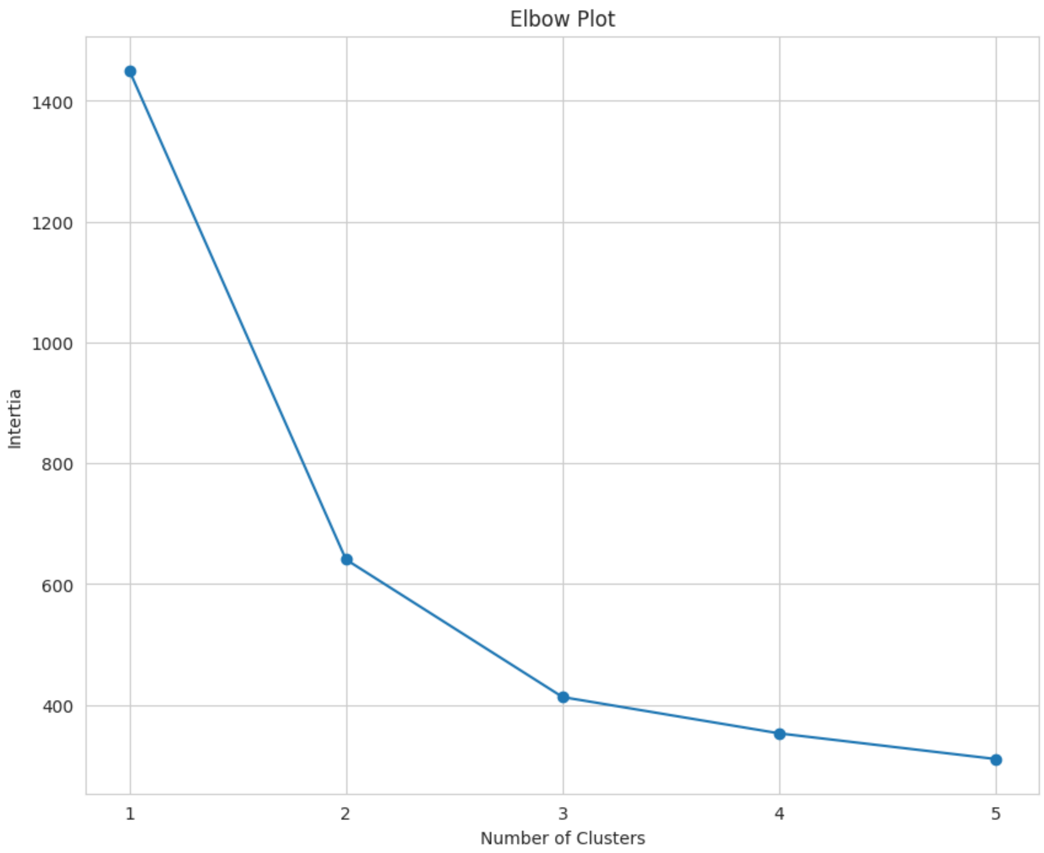
Principle Component Analysis

- Target variance = 0.95
- Reduced dataset used for both implementations of k-means

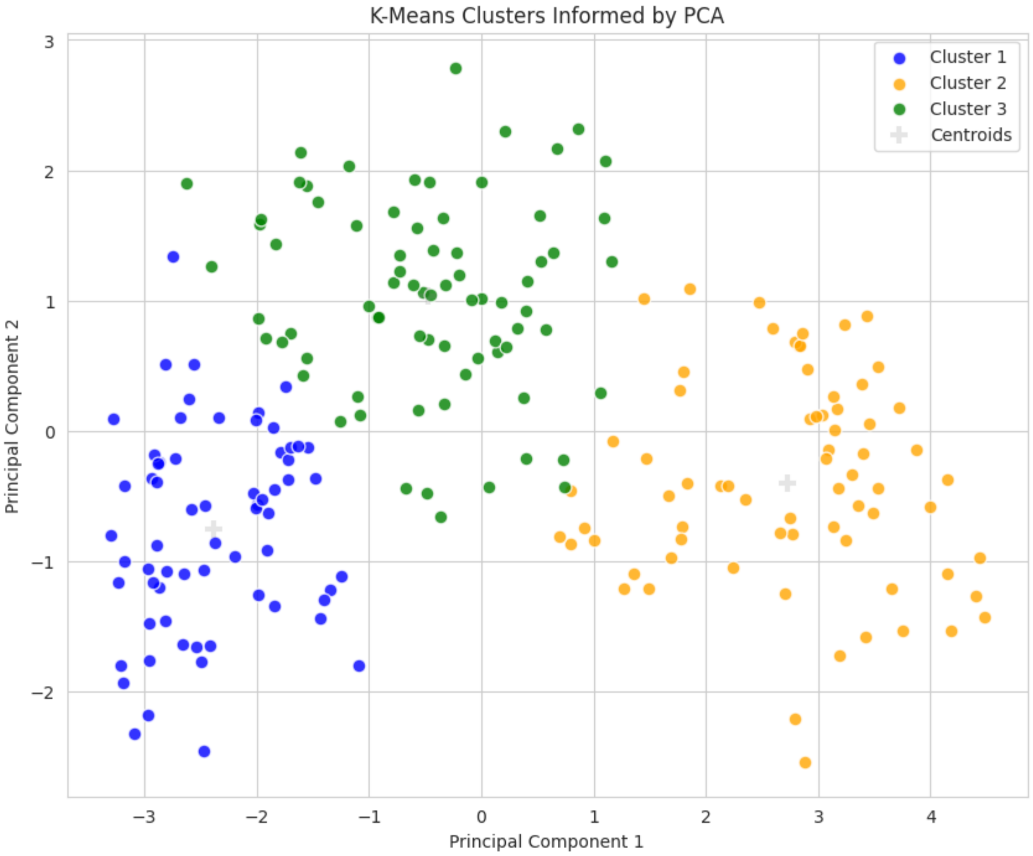


Elbow Plot Informed K-means Clustering

Elbow Plot

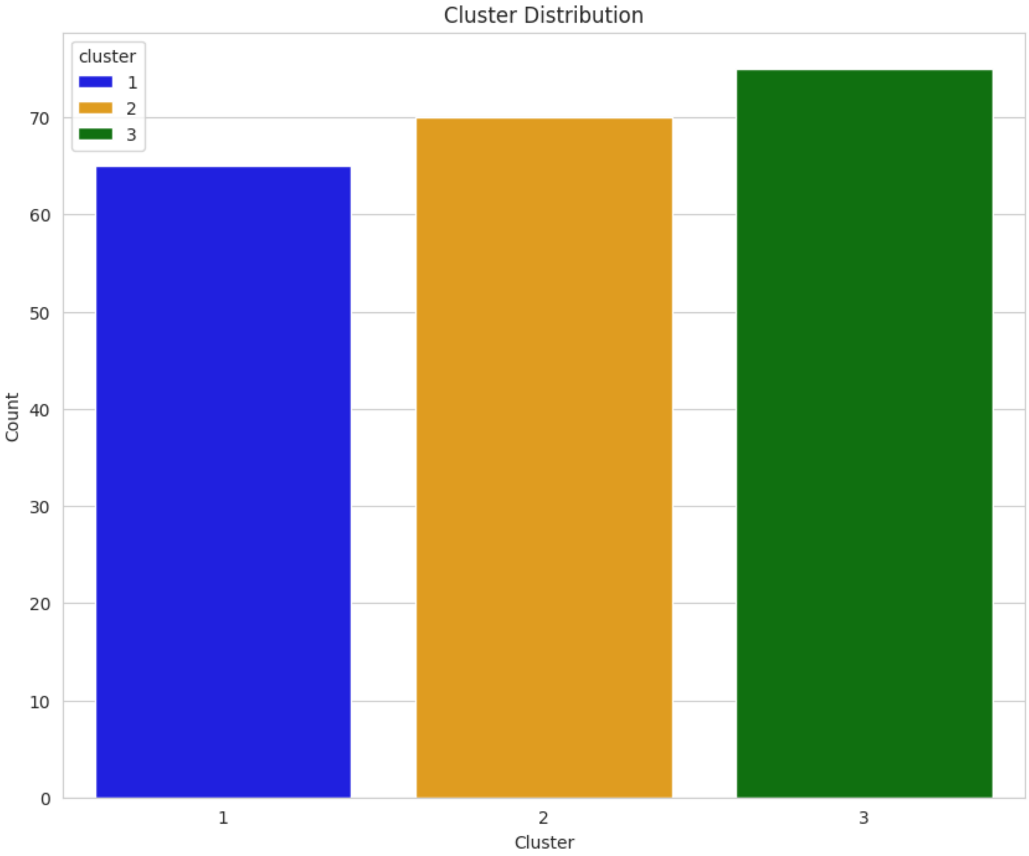

Elbow Plot
1400
1200
1000
800
600
400
Intertia
1
2
3
4
5
Number of Clusters

Plot of Clusters



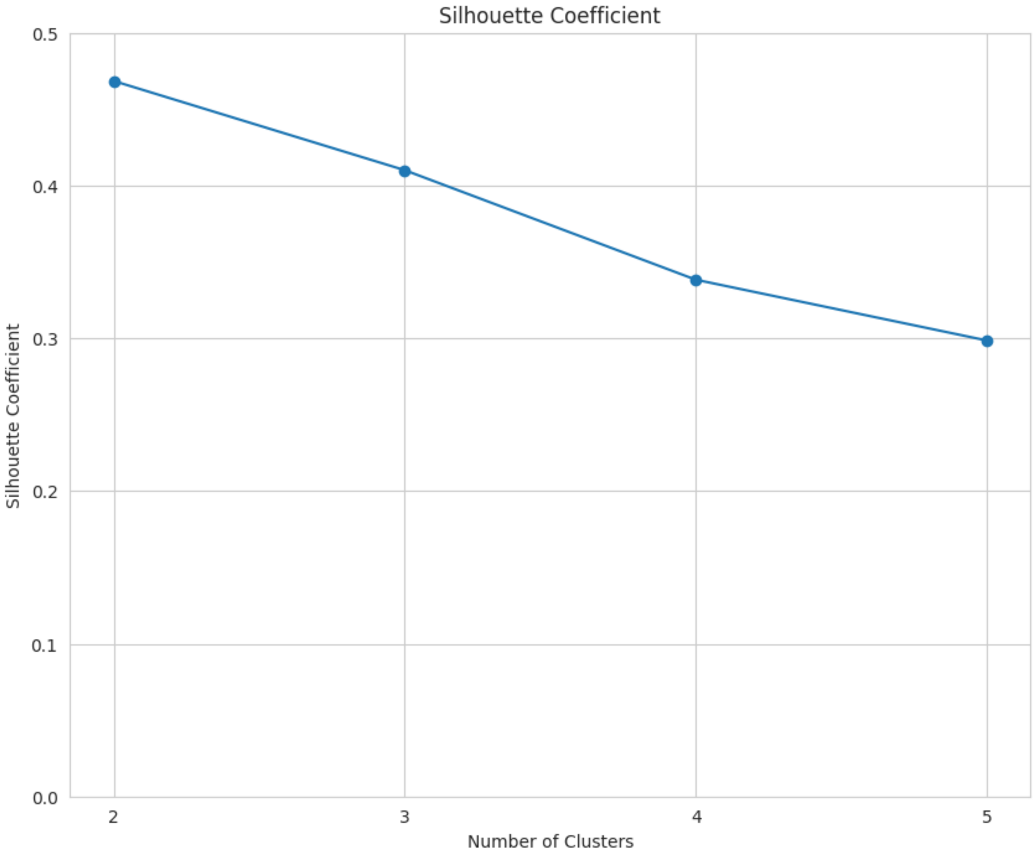
Size of Clusters

- Cluster 1 = 65
- Cluster 2 = 70
- Cluster 3 = 75

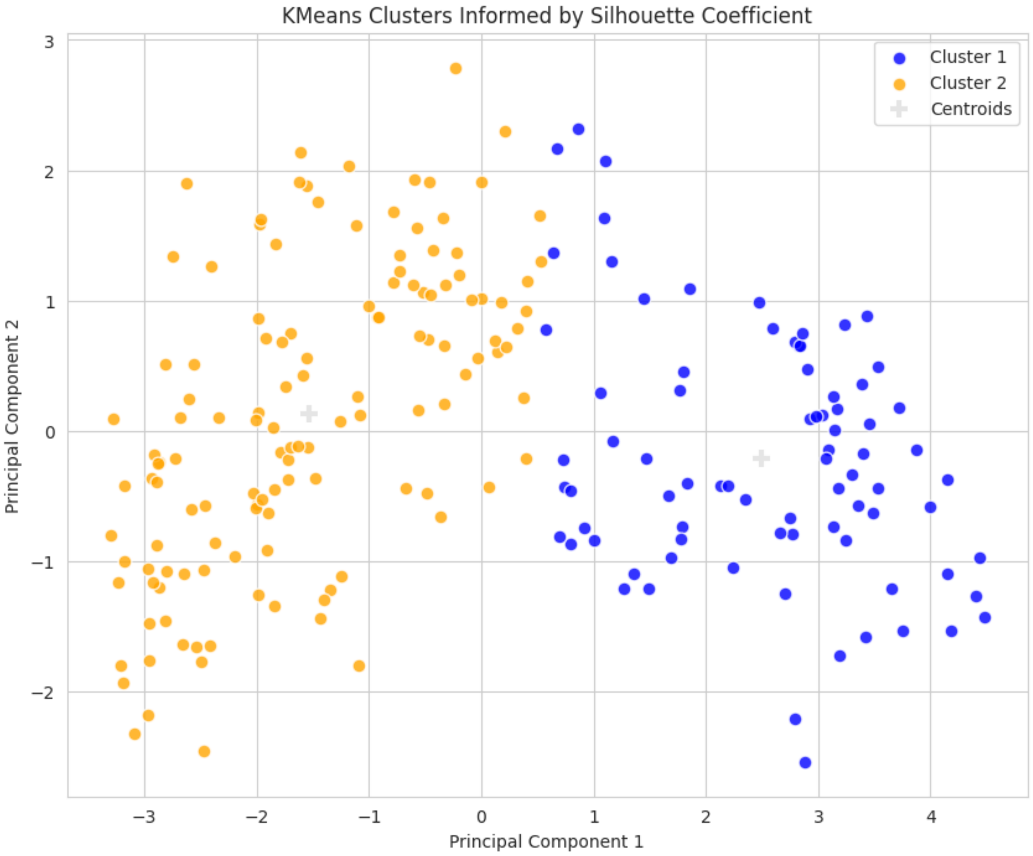


Silhouette Score Informed K-means Clustering

Silhouette Coefficient

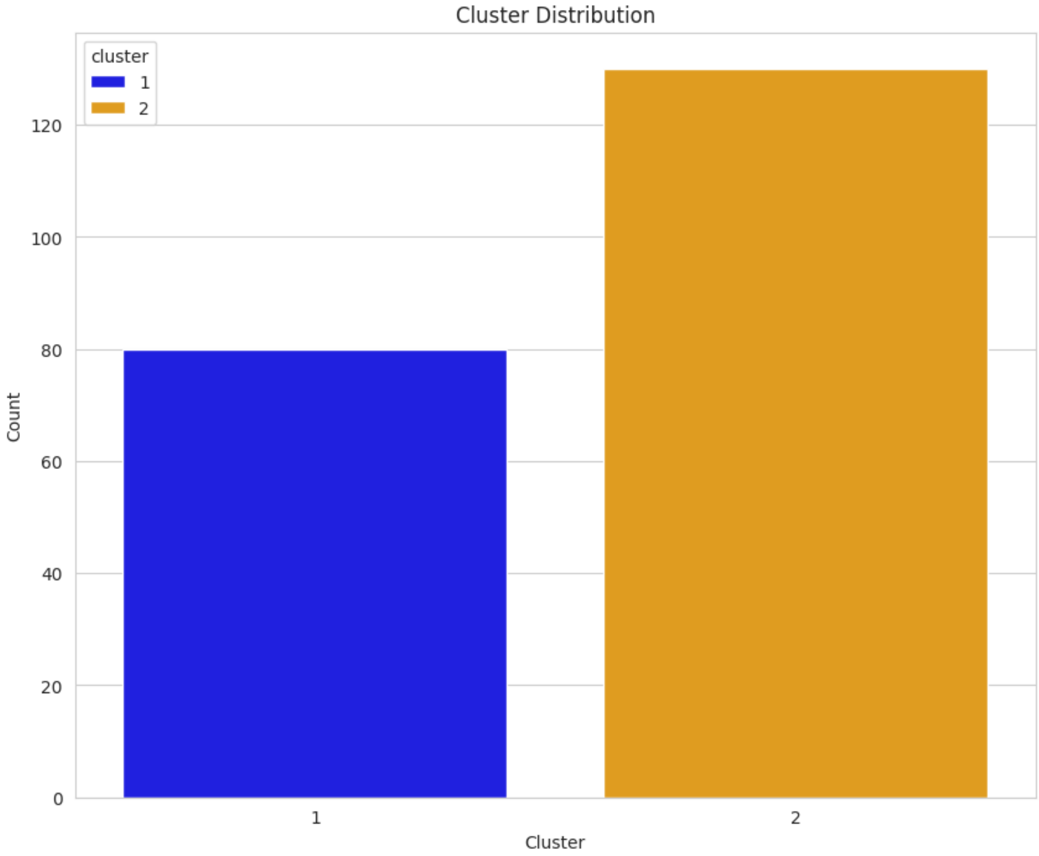


Plot of Clusters



Size of Clusters

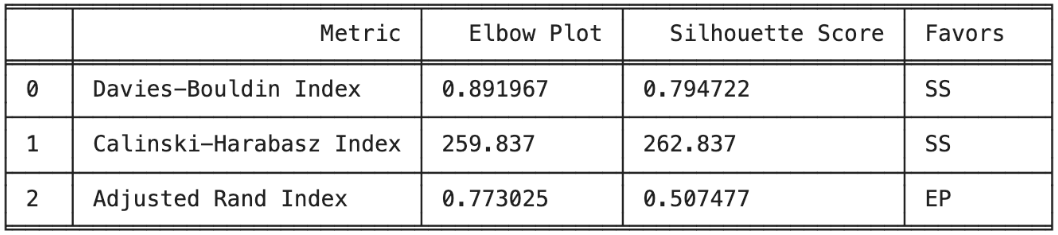
- Cluster 1 = 80
- Cluster 2 = 130



Comparison of K-means Implementations

Evaluation Metrics

- Two of three metrics favor K-means informed by Silhouette Score
- Adjusted Rand Index incorporates ground truth



	Metric	Elbow Plot	Silhouette Score	Favors
0	Davies-Bouldin Index	0.891967	0.794722	SS
1	Calinski-Harabasz Index	259.837	262.837	SS
2	Adjusted Rand Index	0.773025	0.507477	EP

Conclusion

Conclusion

- K-means informed by Elbow Plot performed better
- Best to explore all options instead of relying on general suggestions
- Future iterations of project could incorporate different dimension reduction techniques