TUTORIAL 5: CLUSTERING

uc3m Universidad Carlos III de Madrid

LUCAS KOHLEY 100497018
JORGE SERRANO 100496297

EXERCISE 1: K - MEANS CLUSTERING

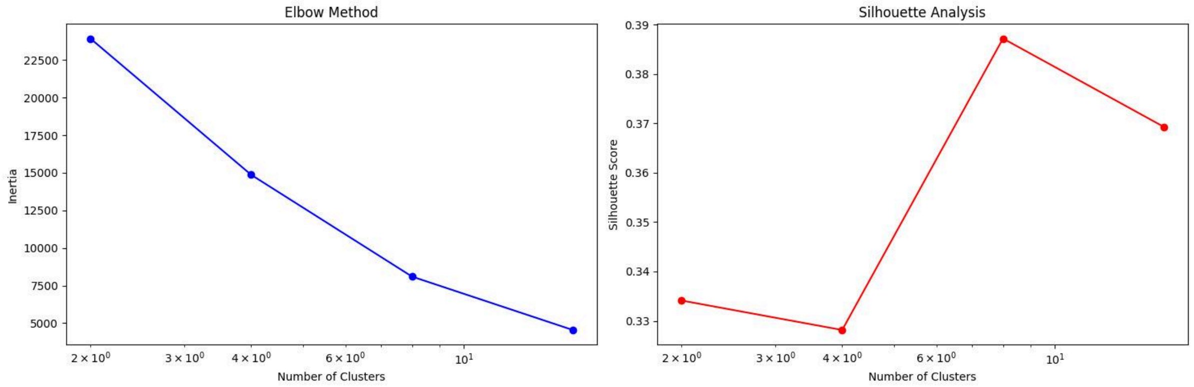
1. **Select a new set of parameters considering the mentioned attributes and perform a training process for the state space. Determine the inertia and silhouette score4 for each number of clusters. Determine the best number of clusters using the elbow method or the silhouette method.**

The new set of parameters chosen are:

- **n_clusters**: which is the parameter that specifies the number of clusters that the K - Means algorithm should find (in our case, tested with 2, 4, 8 and 16).
- **init**: selects initial cluster centers in a smart way to improve convergence.
- **n_init**: number of times the algorithm will run with different initializations. The final model will be the one with the lowest inertia.
- **random_state** : sets a random seed to ensure the results are reproducible.

We tested different numbers of clusters: 2, 4, 8, and 16. The results are as follows:

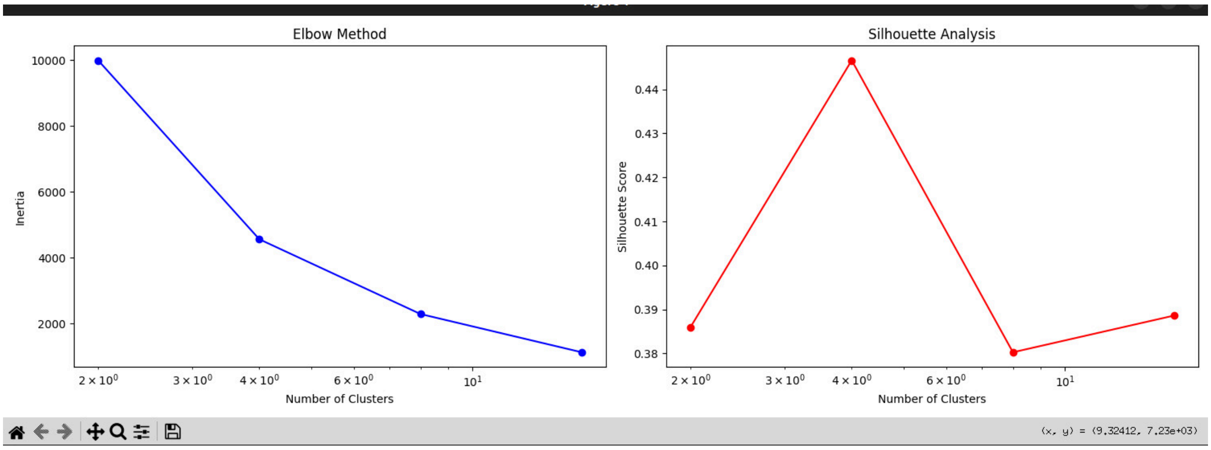
Clusters	**Inertia**	**Silhouette Score**
2	23930.92	0.334
4	14862.31	0.328
8	8095.46	0.387
16	4541.93	0.369



Elbow method: from the inertia graph, we observe a significant decrease as the number of clusters increases. However, the slope becomes less steep after 8 clusters, suggesting a potential "elbow" point. While the "elbow" is not very pronounced, it is sufficient to consider 8 clusters as a reasonable choice.

Silhouette Analysis: the silhouette analysis supports the selection of 8 clusters, as the silhouette score reaches its highest value (0.387). This indicates that 8 clusters provide the best balance between internal cohesion (how close are the points within the same cluster) and separation of the groups (how far are these points from the neighboring clusters).

2. **Repeat the process for the action space. Do the best number of clusters differ from the state space to the action space? Why, or why not?**



Results: the best number of clusters has changed to 4 (silhouette score = 0.447).

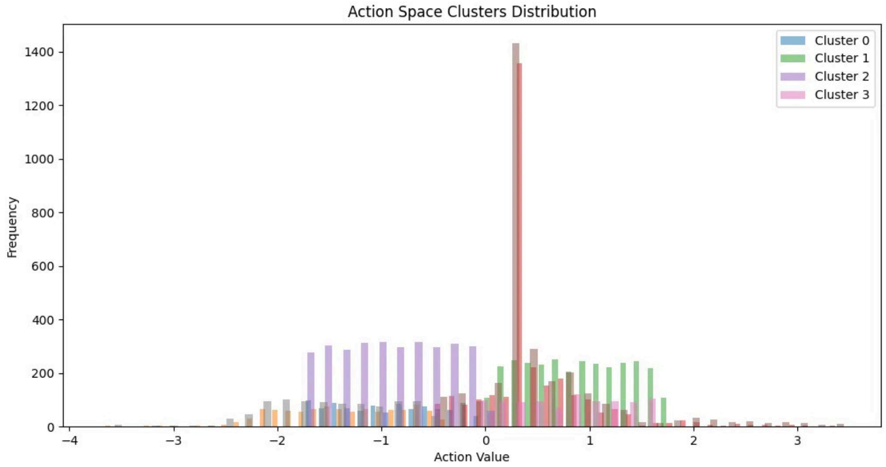
Reasons of this difference:

- The action space is simpler, representing discrete control inputs. This lower dimensionality makes it easier to group actions into fewer, well-defined clusters.
- Actions are typically discrete and naturally fall into distinct categories such as forward, backward… This clear separation means fewer clusters are sufficient.
- Also, from the silhouette scores we can see that the action space clusters are more cohesive and better separated, whereas the state space clusters are less distinct.

In conclusion, the best number of clusters differs because the state space requires more clusters (8) to handle its complexity, while the action space achieves clear and well - separated clusters with fewer groups (4). This reflects the fundamental difference between representing system configurations (states) and discrete control (actions).

3. **Visualize the clusters for the state space and the action space with their respective number of clusters and try to determine why are those clusters generated.**

Action space:

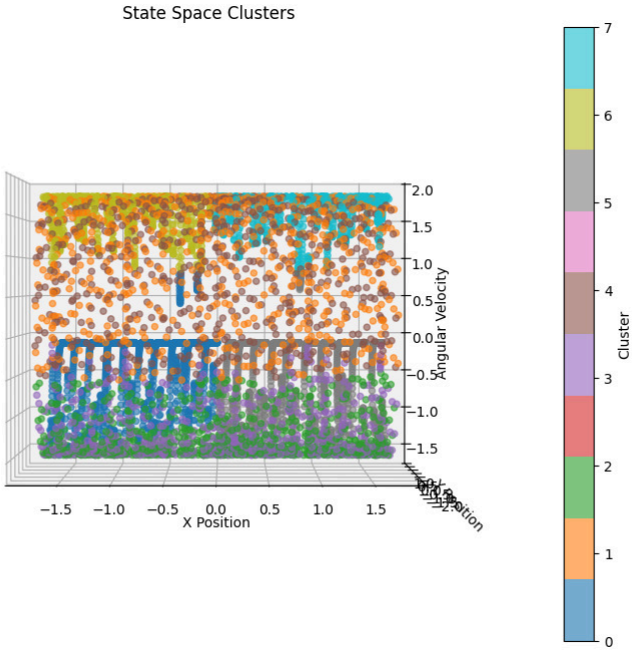


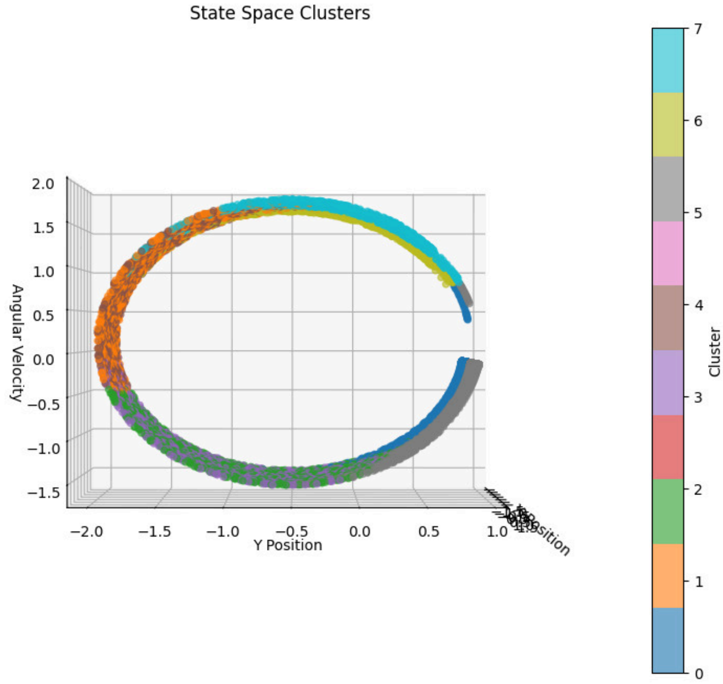
In the action space clusters, a significant number of samples have values close to 0, resulting in the large peak. This indicates that a large portion of the data involves actions or values with minimal magnitude.

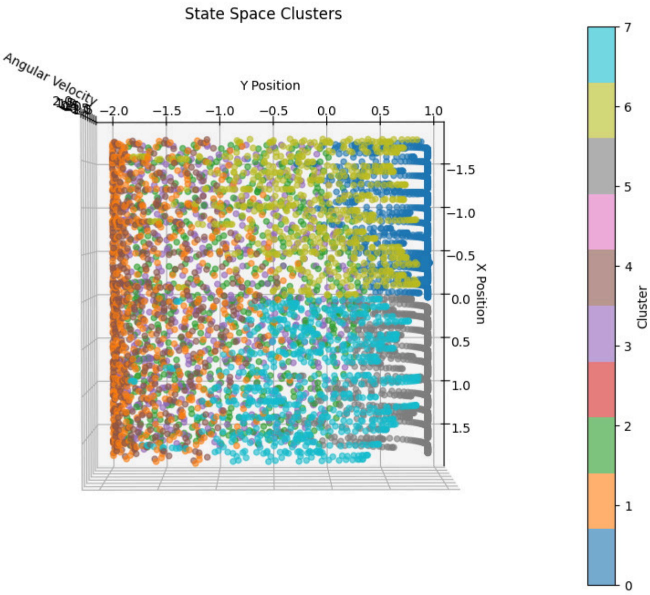
Clusters represent a range of values, from minimal to more significant, allowing the data to be categorized into meaningful groups.

- Cluster 0 and 3 are mainly centered around the zero region, reflecting min values.
- Cluster 1 and 2 expand into positive and larger ranges, showing more significant action values.

State space:



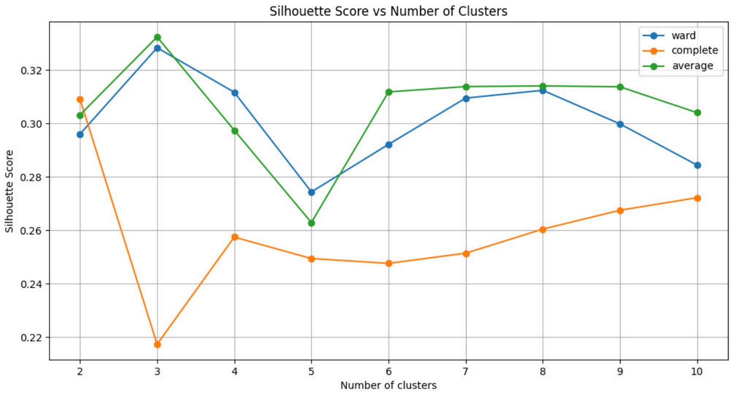


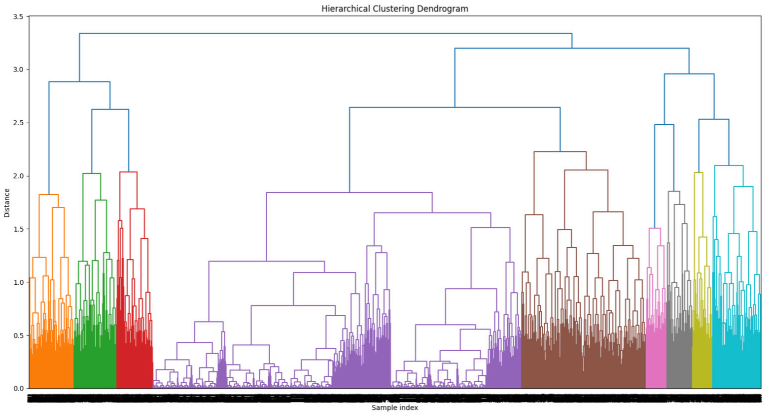


The state space forms 8 distinct clusters, capturing variations in the dataset across multiple dimensions. Each cluster represents a region in the state space where the data points share similar characteristics.
The clusters effectively partition the dataset into regions that share similar characteristics, providing a clear structure to the data.

EXERCISE 2:

1. **Select a set of parameters and perform a training process on the state space data. Determine the best number of clusters using the Silhouette method**





The parameters are:

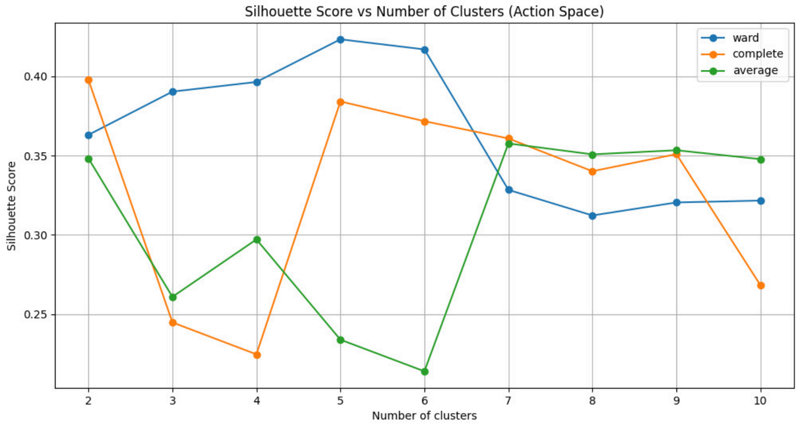
- **Linkage method**: determines how the distance between clusters is calculated.
- **Number of clusters**: specifies the final number of groups formed by cutting the dendrogram.
- **Silhouette score**: a metric to evaluate the clustering quality.

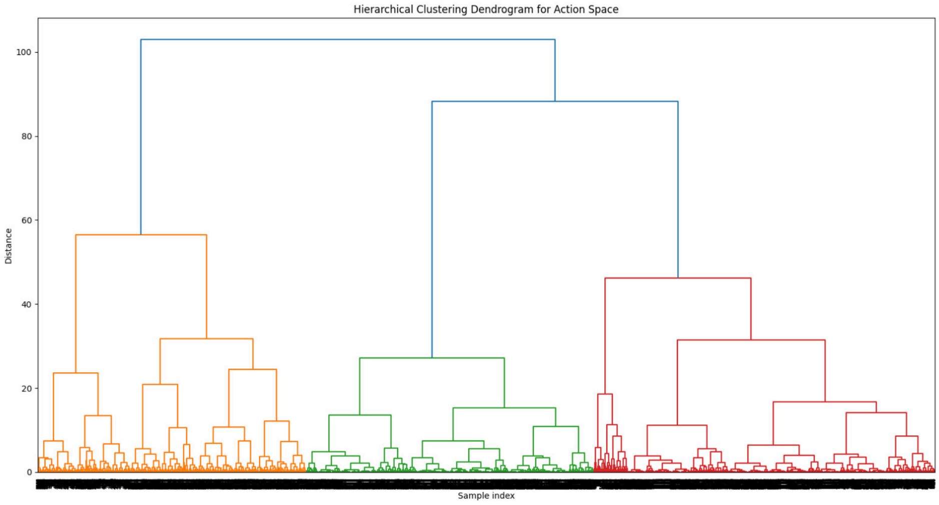
The best parameters for hierarchical clustering on the state space are:

- Linkage method: **average**.
- Number of clusters: **3**.
- Silhouette score: **0.332**

The average linkage method achieved the best results because it balances the distances between clusters, providing a well-defined grouping in the state space. With 3 clusters, the silhouette score indicates moderate separation and cohesion, suitable for capturing key patterns in the state space while avoiding over - segmentation.

2. **Repeat the process for the action space.**



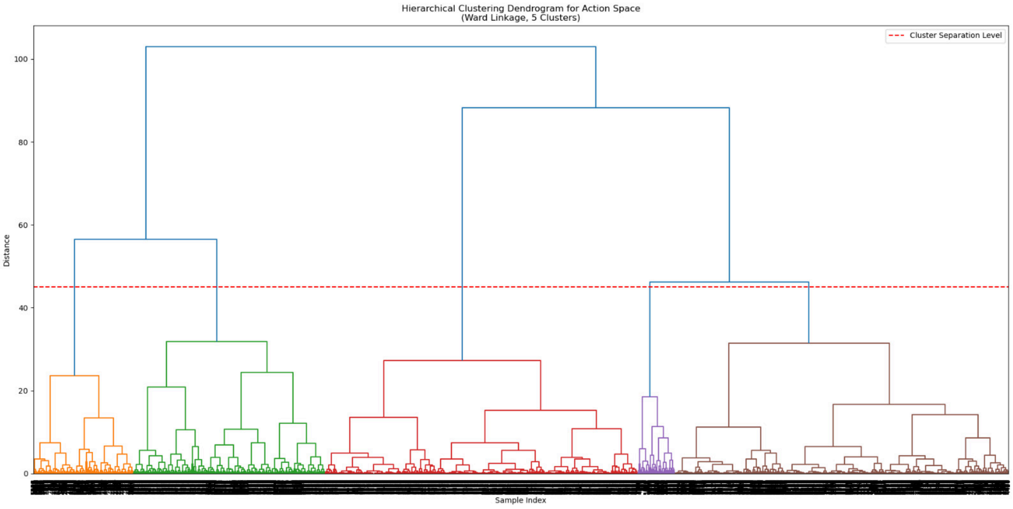


The best parameters for hierarchical clustering on the action space are:

- Linkage method: **ward**.
  The ward method minimizes the variance within clusters.
- Number of clusters: **5**.
  This number provides the best balance, grouping actions into meaningful categories.
- Silhouette score: **0.423**.
  A high silhouette score indicates well - defined and separated clusters, confirming the quality of the grouping.

In conclusion, the best number of clusters for the action space is 5. This number is supported by the highest silhouette score of 0.423, achieved when using the ward linkage method. This indicates that the clustering with 5 clusters provides the best balance between cohesion within clusters and also separation between clusters.

3. **Visualize the dendrograms for the action space with the best number of clusters. Interpret the results and determine why are those clusters selected.**

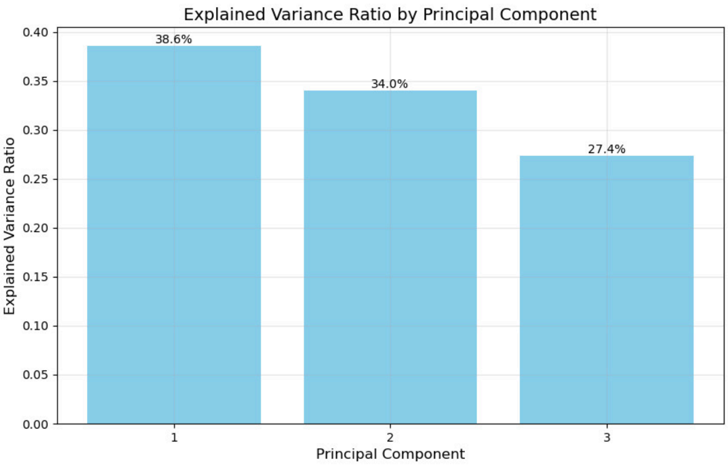


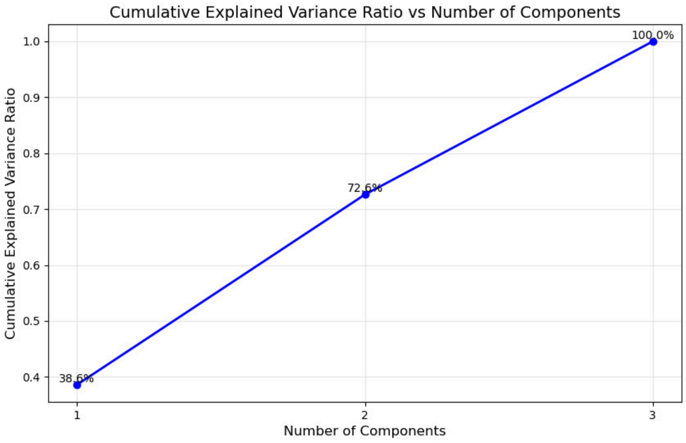
The dendrogram for the action space shows the hierarchical clustering process, which divides the data into 5 clusters, as determined by the ward linkage method.

- Clusters 0 and 2 capture high-value observations, with Cluster 2 being more focused and consistent.
- Cluster 1 represents the mid-range values, forming the largest and most central grouping.
- Meanwhile cluster 3 groups lower values.

EXERCISE 3:

1. **Perform the PCA method on the state space using 3 as the number of components on the state space data. Plot the explained variance ratio for each principal component. How many components are necessary to capture at least 95% of the variance?**





When applying PCA with 3 components to the state space data, we observe the following explained variance ratios:

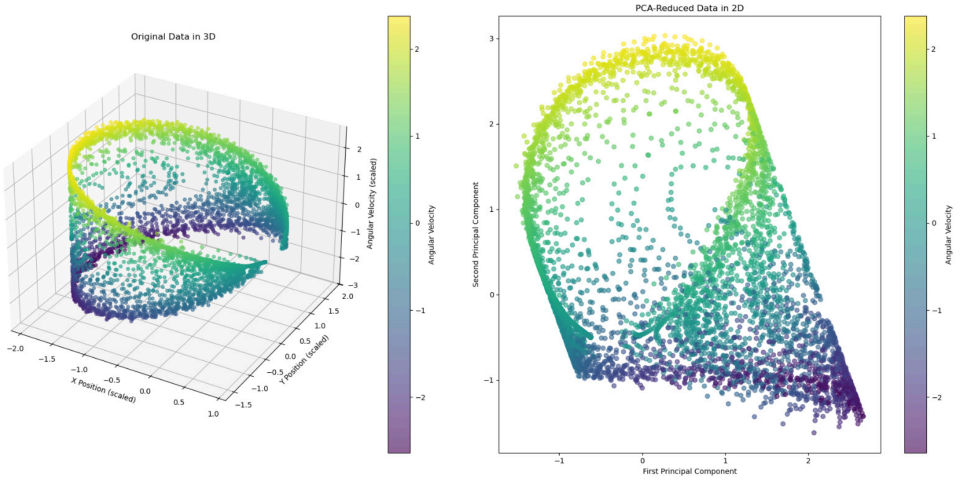
1. First Principal Component: 38,6% of total variance.
2. Second Principal Component: 34,0% of total variance.
3. Third Principal Component: 27.4% of total variance.

Looking at the cumulative explained variance:

- Using 1 component: 38,6%.
- Using 2: 72.6% (sum of PC1 + PC2).
- Using 3: 100%.

So, to capture at least 95% of the variance in the data, we need all three principal components. This can be seen clearly in the cumulative explained variance plot, where only when including the third component we reach 100% variance explained, thus exceeding our 95% requirement.

2. **Perform again the PCA method with 2 as the number of components for the same data. Transform the data to be able to visualize it in 2D. Compare the visualization of the data in 2D and 3D. Do the spatial structures remain after the reduction?**



We observe that in the 2D PCA Reduced visualization, the spiral structure is preserved, though flattened into two dimensions.

Also, the color gradient pattern remains consistent between both visualizations, indicating that the relationship with angular velocity is maintained.

In addition, in the 2D PCA visualization we can see that the First Principal Component is mainly influenced by the spatial position, while the Second Principal Component is strongly influenced by angular position.

Regarding whether the spatial structures remain after reduction we can say that the essential spatial structures are maintained. The main characteristics of the data (the spiral shape or pattern and the distribution of angular velocities) remain clearly visible in the 2D representation.

3. **Discuss how PCA could benefit clustering methods.**

PCA can benefit clustering methods in several important ways, for example:

1. **Reduction of complexity**:
   By reducing the number of dimensions, clustering algorithms can run faster which can be very useful when working with high-dimensional datasets.
2. **Visual interpretation**:
   As we have seen in the graphs, PCA allows visualization of high-dimensional data in 2D or 3D. This facilitates visual interpretation of clusters and validation of clustering results.
3. **Interpretation simplification**:
   PCA can help to better understand why certain points are grouped together in clusters. This means that, for example, in our case: the PC1 mainly captures positional information and the PC2 mainly captures angular velocity info.
   So, when we see points clustered together in our 2D PCA visualization, we understand that points close together horizontally have similar positions in space. The same with PC2 that will have similar angular velocities.

In conclusion, PCA significantly improves clustering methods by reducing data dimensionality while preserving essential information. It not only makes clustering algorithms more efficient but also enables better visualization and interpretation of results.