Clustering

Making categories for Penguins species using clustering;

- Inertia and silhouette score can be used to find the optimal value of clusters.
- Clusters can find natural groupings in data.

Note: Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.

Details

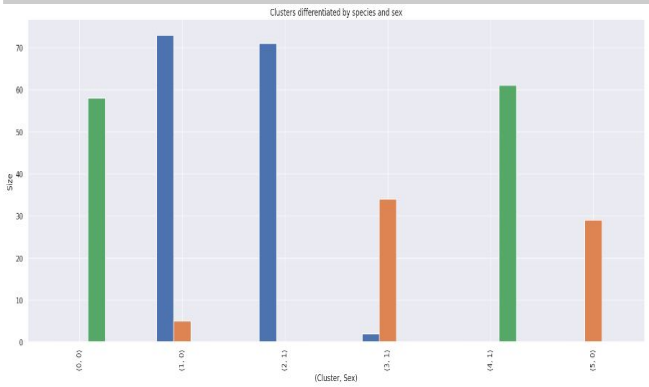
K-means clustering is very effective when segmenting data and attempting to find patterns. You are a consultant for a scientific organization that works to support and sustain penguin colonies. You are tasked with helping other staff members learn more about penguins in order to achieve this mission. The data for this activity is in a spreadsheet that includes data points across a sample size of 345 penguins, such as species, island, and sex. We will use a K-means clustering model to group this data and identify patterns that provide important insights about penguins. The data was already uploaded in `Seaborn` library; There were three types of species. But we determined other numbers to find a relationship in data and explore, whether the algorithm can discover the different species. Having 6 clusters makes sense because the study suggests that there is sexual dimorphism (differences between the sexes) for each of the three species (2 sexes * 3 different species = 6 clusters).

Key Insights

Defining other numbers for `k` is essential, Because,

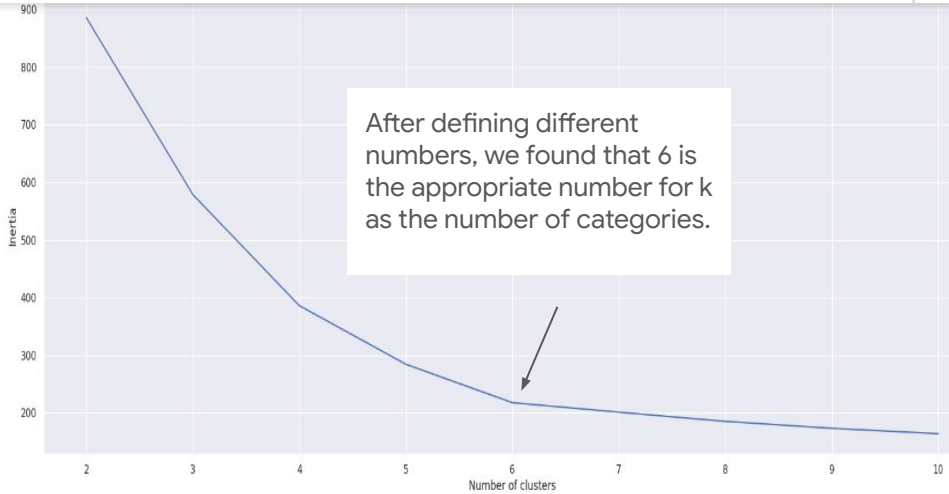
1. The clusters should be clearly identifiable.
2. Within each intercluster, there should be lots of empty space.
3. Within each intracluster, the points should be close to each other.

For an effective clustering model, the clusters should be clearly identifiable. Within each intracluster, the points are close to each other; within each intercluster, there is lots of empty space.



The graph shows that each 'cluster' can be differentiated by 'species' and 'sex_MALE'. Furthermore, each cluster is mostly composed of one sex and one species.

Next Steps



After defining different numbers, we found that 6 is the appropriate number for k as the number of categories.