PENGUINS IN ANTARCTICA

MACHINE LEARNING-REPORT

FINAL PROJECT – SHASHIDHAR REDDY

12/16/2022

INTRODUCTION:

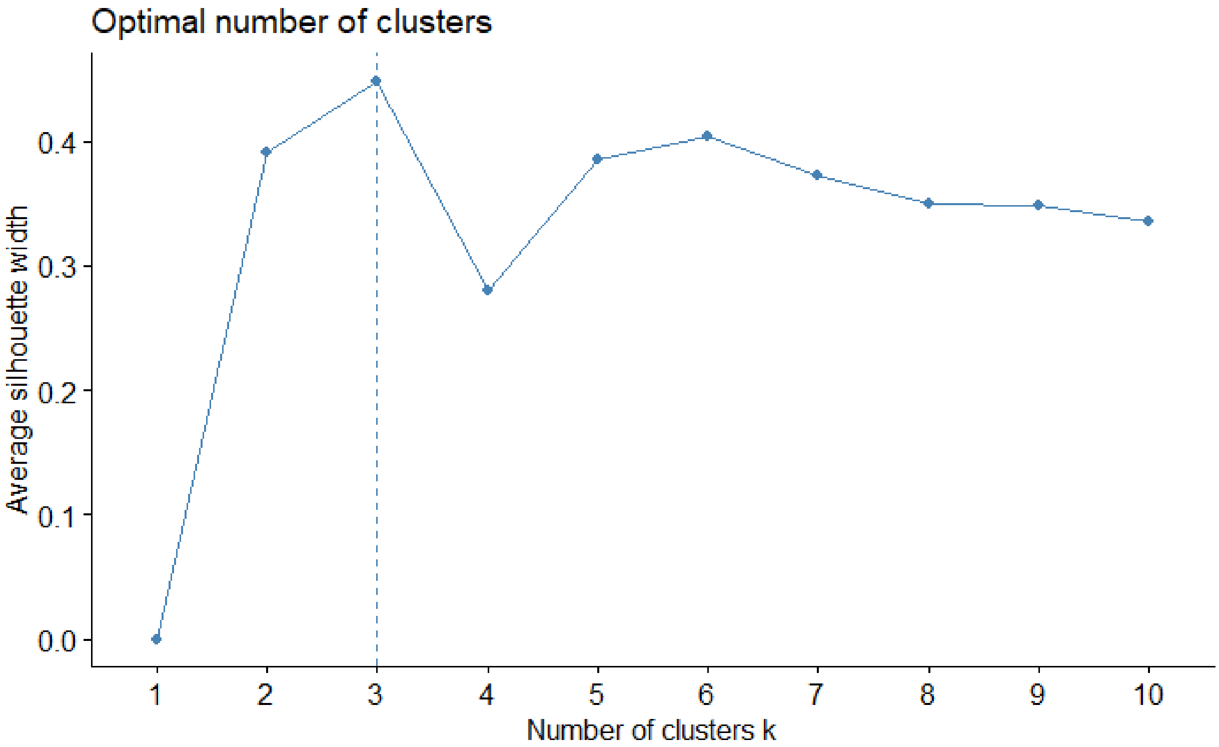
We are looking at data collected about penguins in the Palmer Archipelago for this study (Antarctica). Dr. Kristen Gorman and the Palmer Station, Antarctica LTER obtained this data through Kaggle and made it available (Long Term Ecological Research Network).

The issue with this study is that while data on penguins in Antarctica is obtained every year, it is impossible to extract conclusions from the raw data and compare it across time. The goal of this study is to use the data collected at the Palmer Station to cluster the population of penguins sampled and make it easier to identify clusters and characteristics of these penguins from the data so that insights can be made more efficiently as we continue to sample these penguins in the future. These parameters include species, the number of such species on each of the three islands studied, and physical measures.

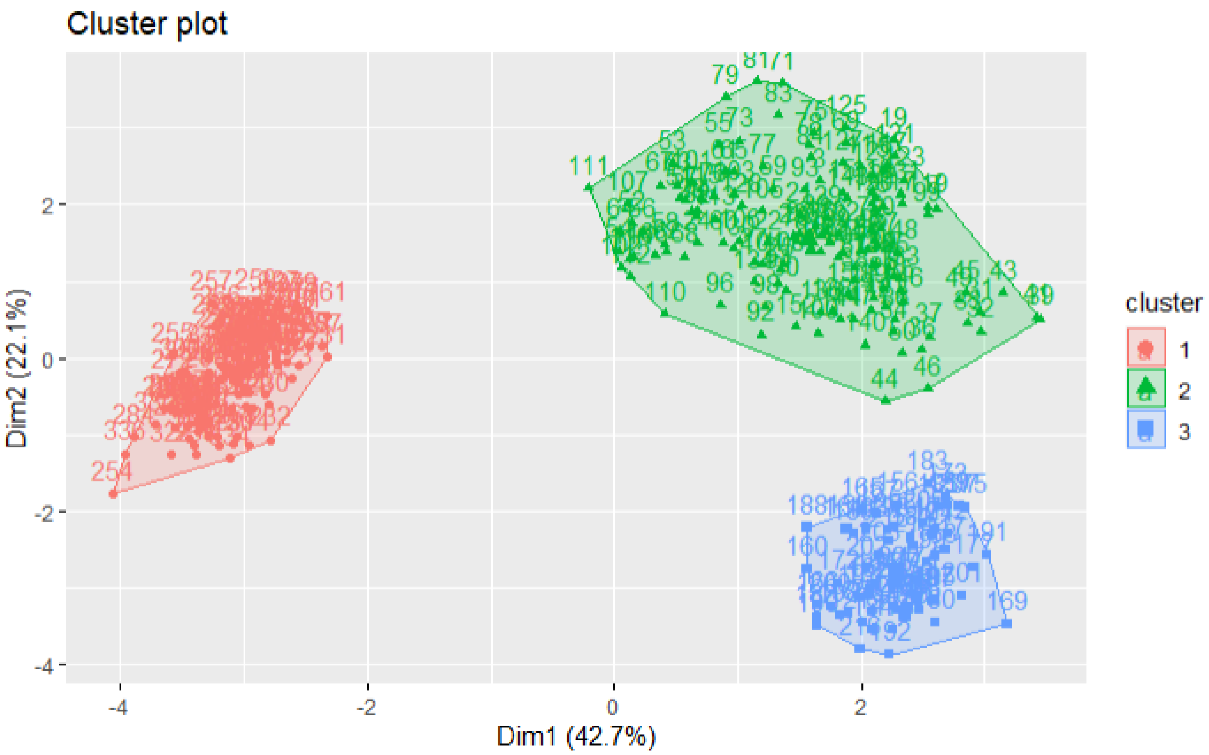
ABSTRACT: Our technique use both K-means and Hierarchical Clustering to achieve the objective of making it simpler to find penguin groupings and traits. This method produces a cluster plot (derived from K-means) as well as cluster attributes (from Hierarchical clustering). A

cluster diagram may clearly show the three separate clusters while doing K-means clustering. This plot enables us to choose individual penguins from the data and track which cluster they belong to, as well as tie them to the cluster characteristics produced in Hierarchical clustering.

APPROACH: Using the silhouette method, we can plainly see that this dataset has three clusters. Three clusters produced the maximum average silhouette width, as seen below, making it the most ideal number of clusters.

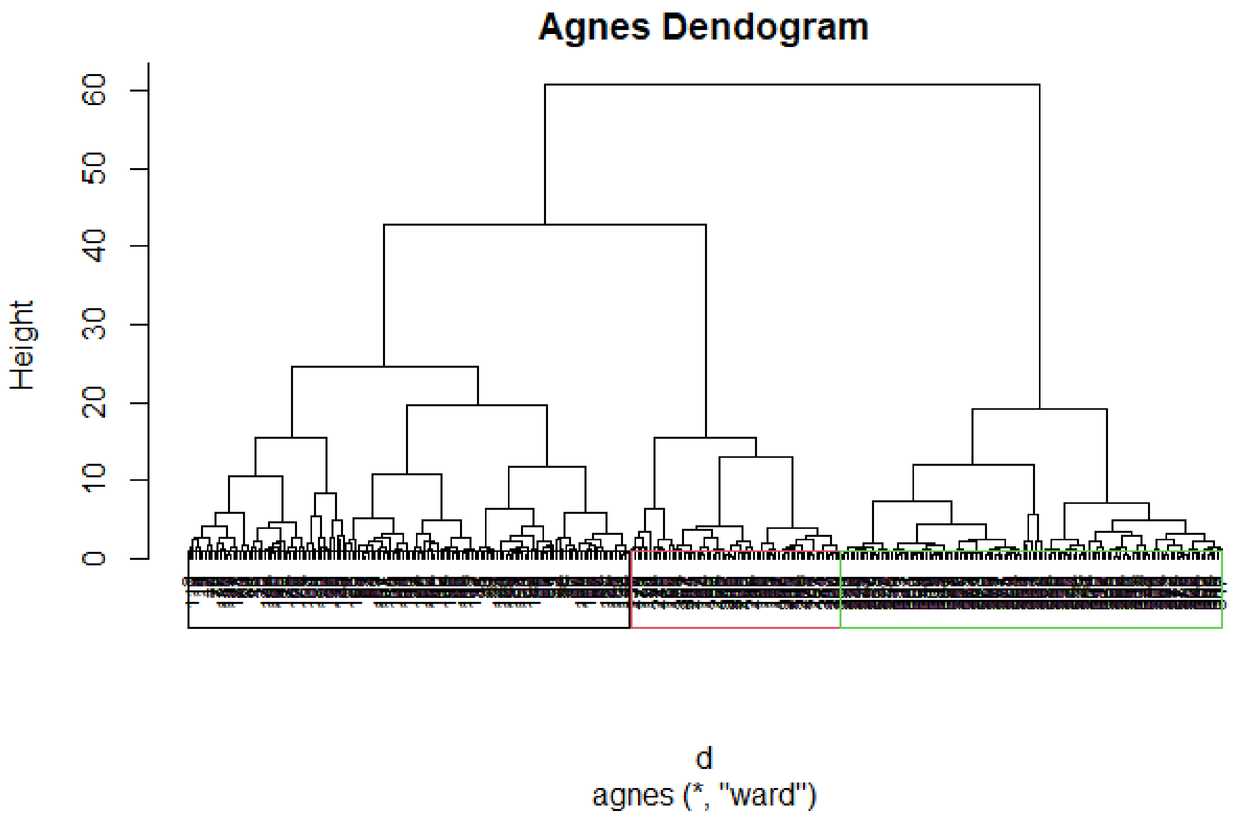


Now that we've found the appropriate number of clusters, we can use K-means to construct a cluster plot to better demonstrate the grouping. The graph below depicts our three groupings as determined using K-means clustering.

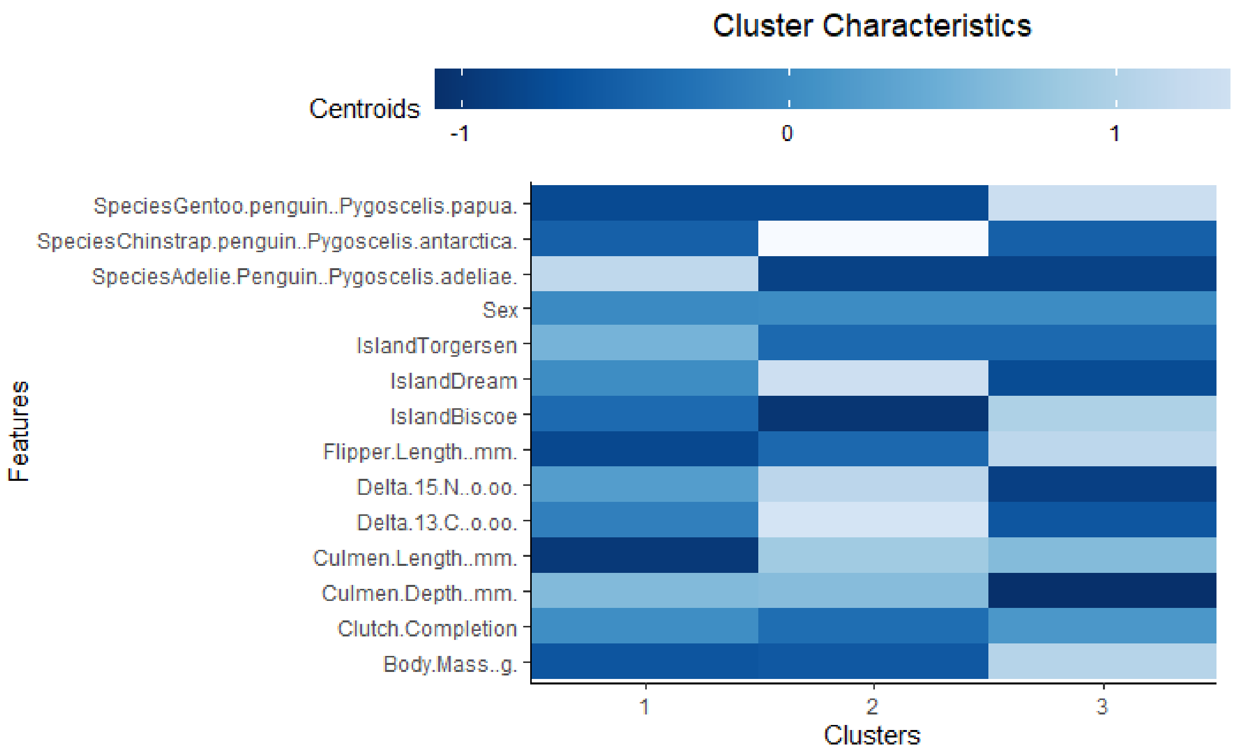


Analysis: The three groups above divide the penguins into our three sampled penguin species. These clusters were formed using all of the numerical data from the dataset, and it is evident that each of these three species has highly unique statistical differences. The cluster characteristics obtained using Hierarchical clustering contain the features that differentiate the species into these three unique groupings.

We wish to revalidate our ideal number of groups for the new clustering algorithm because we are utilizing a different technique of clustering to highlight attributes. This was accomplished by developing an Agnes dendogram that depicts our clusters in a tree structure. Our Agnes dendogram, shown below, shows the highest spike in groups at a height of 30, suggesting three clusters, confirming our previous findings.



We can now create our features after confirming that three clusters would be optimum for Hierarchical clustering. The graph below depicts the values of the centroids of each cluster for each value on which they were assessed.



We may acquire a much better understanding of the three penguin clusters from the figure above. We may conclude that Cluster 1 has the greatest number of Adelie penguins. The Chinstrap penguin has a significant population in Cluster 2. Gentoo penguins are found in Cluster 3. An important fact to note from this data is that cluster two includes relatively few of the other two species, but the other two clusters have a modest gradient since the minority species' populations are denser.

Insights and Conclusions:

Using this data, we can get insights into the clusters' and species' features. Cluster two features a high density of Chinstrap penguins, as well as the biggest Culmen (beak) length and depth, with Dream Island having the highest population. This can tell us about the chinstrap penguin's qualities, such as the fact that they dwell predominantly on Dream Island. Dream Island's culmens are long and deep. Gentoo on Biscoe Island has the largest flippers and body mass, according to the graph. This approach enables us to simply cluster data from our dataset and show it in an understandable manner. This allows us to have a better understanding of the penguins' life throughout time. As more data is acquired over time, we can compare the graphs generated then to the graphs made today to discover how these penguins have changed.

REFERENCE : https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data