More Exploratory Analysis: Clustering

4 min

Unsupervised machine learning techniques, such as **clustering** algorithms, are useful tools for exploratory analysis. These techniques "learn" patterns from untagged data, or data that do not have classifications already attached to them, and they help us see relationships between many

Preview: Docs Loading link description

variables

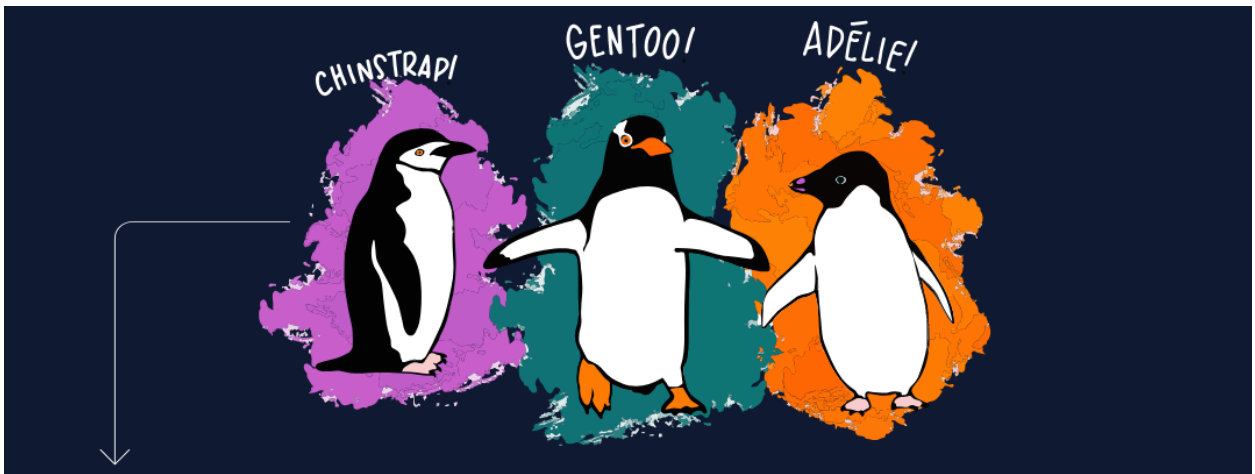
at once.

Let's see how cluster analysis works with penguin data that scientists collected in Antarctica! Take a look at the data in the learning environment, which comes from the palmer penguins dataset.

- We have four measurements (weight, beak length, beak depth, and flipper length) from Gentoo, Adelie, and Chinstrap penguins living on three islands (penguin artwork by @allison_horst). It's pretty difficult to see relationships between the measurements looking at the raw data.
- Ideally, we want to see how all the measurements relate to each other, but plotting them all would be a lot to look at. Instead, we can perform a **Principal Component Analysis** or PCA, which compresses the variables into principal components that can be plotted against each other. After PCA, we can use **k-means clustering** to look for trends in the data. We see that the penguins fall into three distinct clusters in the PCA plot!
- It's cool that our analysis shows three clusters because we know that our data include three penguin species living on three islands. We can use the Rand statistic to see how well species and islands match the k-means clusters. This is similar to the R-squared statistic we use to check the fit of a trend line. With a Rand score of 0.97 out of 1.0, species have a stronger relationship to clustering!

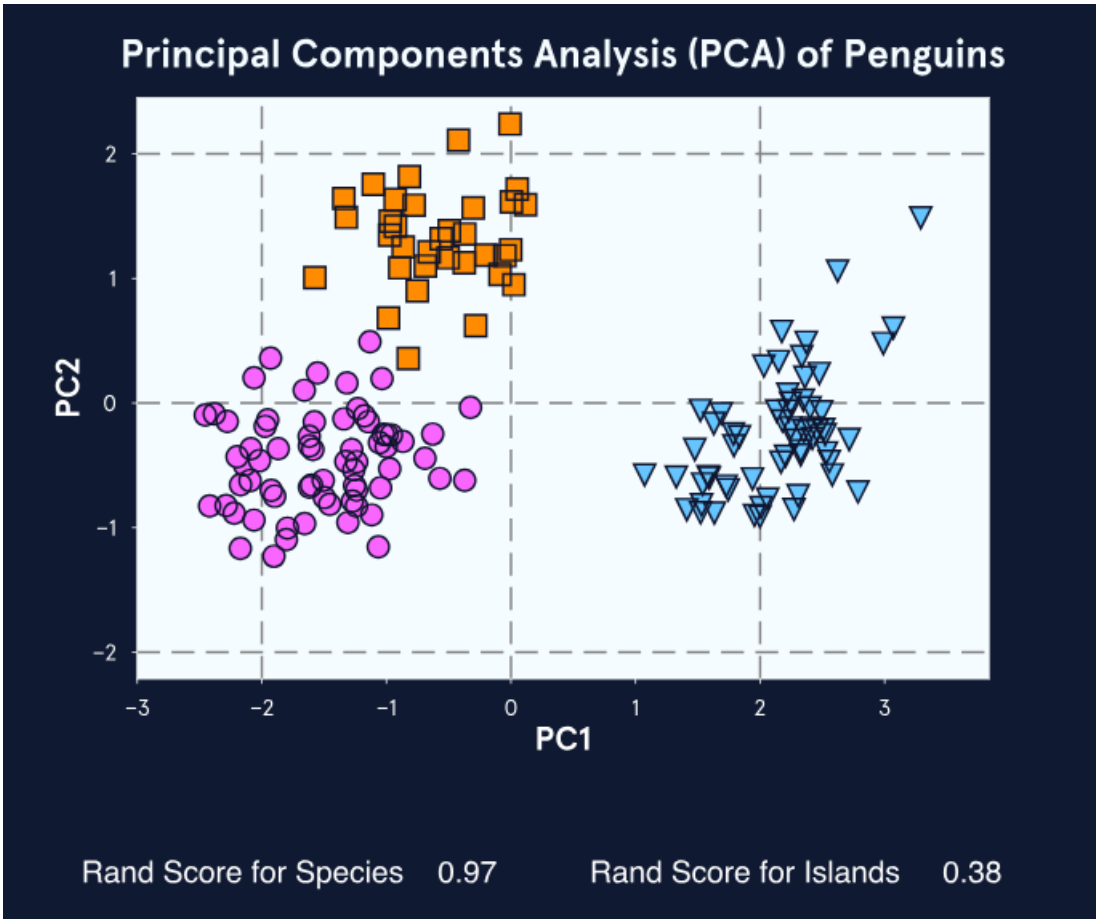
Next up, we'll check out inferential analysis!

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Adelie	Torgersen	41.5	18.3	195.0	4300.0
Adelie	Biscoe	43.2	19.0	197.0	4775.0
Gentoo	Biscoe	49.6	16.0	225.0	5700.0
Adelie	Biscoe	38.2	18.1	185.0	3950.0
Chinstrap	Dream	52.0	19.0	197.0	4150.0
Chinstrap	Dream	51.3	18.2	197.0	3750.0



species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Adelie	Torgersen	41.5	18.3	195.0	4300.0
Adelie	Biscoe	43.2	19.0	197.0	4775.0
Gentoo	Biscoe	49.6	16.0	225.0	5700.0
Adelie	Biscoe	38.2	18.1	185.0	3950.0
Chinstrap	Dream	52.0	19.0	197.0	4150.0
Chinstrap	Dream	51.3	18.2	197.0	3750.0

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Adelie	Torgersen	41.5	18.3	195.0	4300.0
Adelie	Biscoe	43.2	19.0	197.0	4775.0
Gentoo	Biscoe	49.6	16.0	225.0	5700.0
Adelie	Biscoe	38.2	18.1	185.0	3950.0
Chinstrap	Dream	52.0	19.0	197.0	4150.0
Chinstrap	Dream	51.3	18.2	197.0	3750.0


Principal Components Analysis (PCA) of Penguins
PC2
PC1
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Rand Score for Species 0.97
Rand Score for Islands 0.38