Can you discover any interesting relationships between species and locations? e.g. do specific species have a preference for certain habitats.

Data Exploration

First I printed size, length , shape and data type of each array to explore what kind of data we have. Then, depending on data type I checked for any nan values in ‘train\_locs’, any invalid ids (less than 0) in ‘train\_ids’ and ‘taxon\_ids’ and any empty variables in ‘taxon\_names’.

To explore how species are spread out I took 5 random species (we can change it later depending on findings from my groupmates) and plotted two violin plots showing latitudes and longitudes.

A diagram of a graph

Description automatically generated with medium confidence

A white background with black and purple shapes

Description automatically generated with medium confidence

Clustering

To explore sub-habitats of a given specie I performed clustering using the k-means algorithm from the sklearn library. First, I used elbow and silhouette analysis to find an optimal number of clusters. Then I plotted them along with their centroids for better visualisation.

A graph with a line

Description automatically generatedA graph with a line

Description automatically generatedA screen shot of a diagram

Description automatically generated

To explore sub-habitats of multiple species I performed clustering on all species present in Australia (I chose Australia just as an example, we might switch to a different habitat later). Similarly to what I did before, first I identified the optimal number of clusters and then plotted them with the centroids. Then I created a stacked bar plot to explore proportions in all clusters of 5 most commonly occurring species.

A graph with a line

Description automatically generatedA graph with blue lines

Description automatically generated

A diagram of a clustering model

Description automatically generated with medium confidenceA graph of different colored squares

Description automatically generated