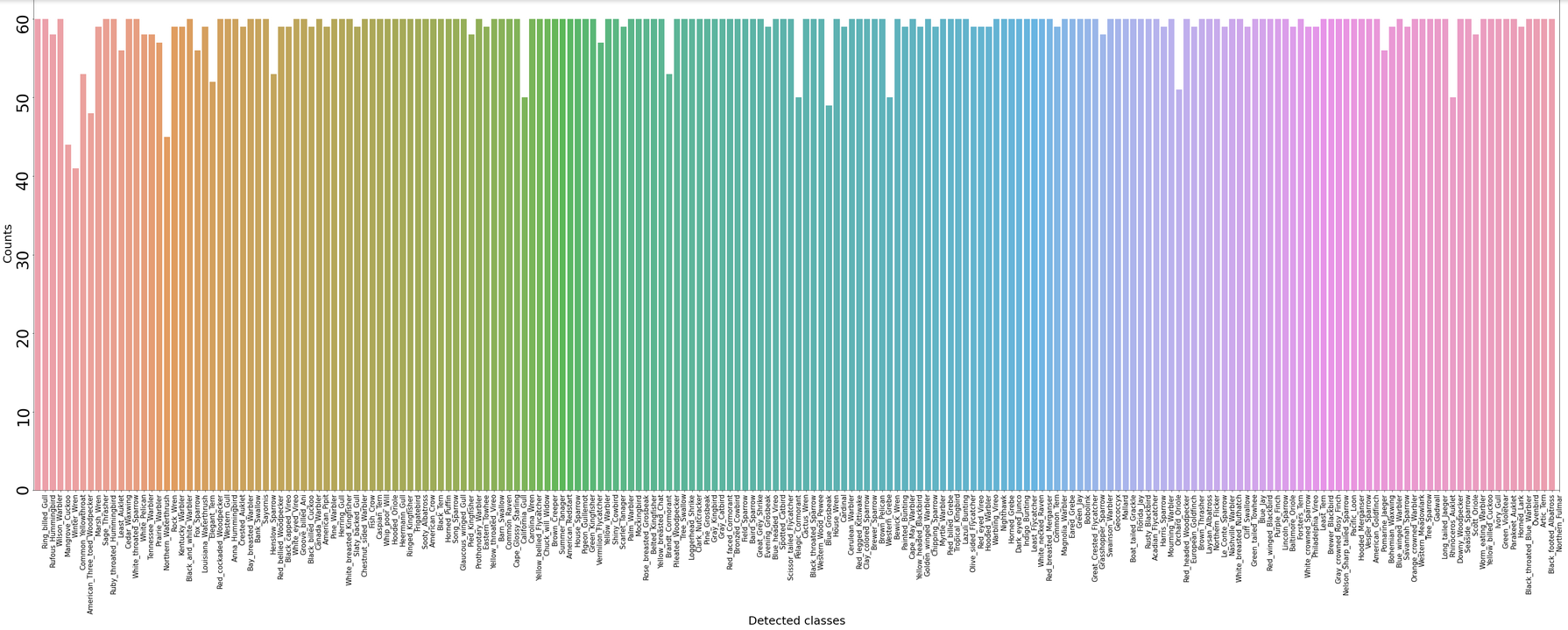
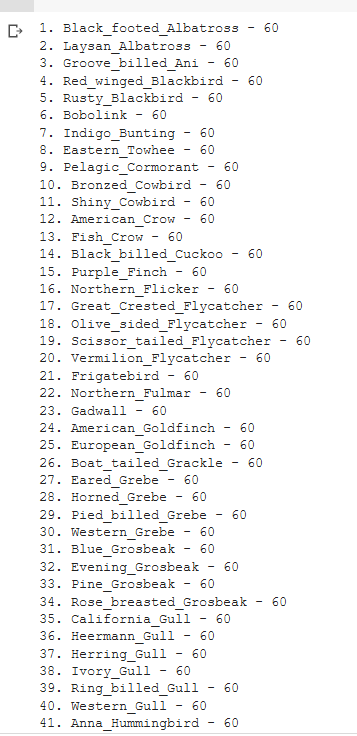
Graph showing class Vs number of images.



The total number of images are **11788** which are divided into **200** classes (200 types of birds)



There are many classes with the highest number of images, the max number of images for a class is **60**.



The lowest number of images for a class is **41** and the class name is Least\_Auklet



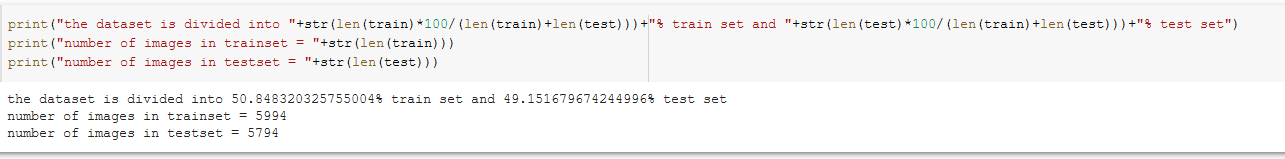
The average number of images for a class is **59**



**the dataset is divided into 50.85% train set and 49.15% test set**

number of images in trainset = 5994

number of images in testset = 5794

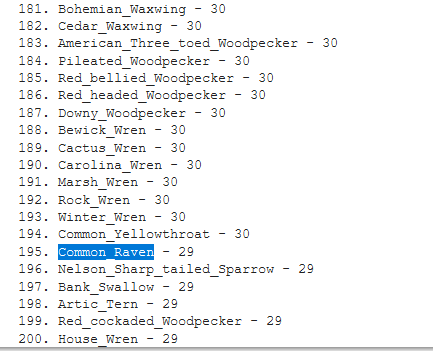


The max number of images for a class in **training data** is 30

The min number of images for a class in **training data** is 29 and the class names are:

House\_Wren, Red\_cockaded\_Woodpecker, Artic\_Tern, Bank\_Swallow, Nelson\_Sharp\_tailed\_Sparrow, Common\_Raven.

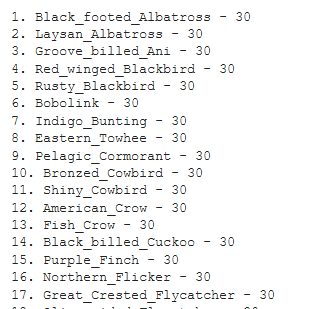
All the other classes in the training data have 30 images



The max, min number of images for a class in **testing data** are 30, 11 respectively

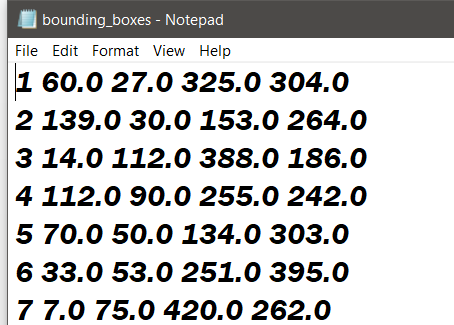
The 11 images are present for the class **Least\_Auklet** while there are many classes with 30 images in the testing set





We are also given bounding box for every image, the bounding box contains the **image ID x,y co-ordinates** and **width and height** of the image. So instead of using the whole image for training we can use only the part of this bounding box.

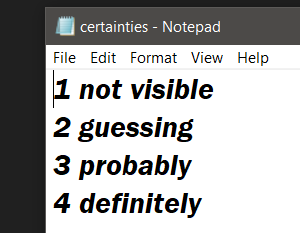
Every image has different bounding box.



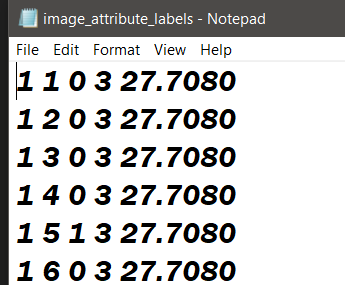
Apart from this we are also given 312 attributes, each image contains atleast one of this attribute but it can also contain all the attributes, this is given in the format **attribute\_ID , attribute\_name**



We are also given how visible the attribute is present in a picture, the given format is **certainity\_ID, certainity\_name**



And here we are given the **image id, attribute id, is\_present** (1- present 0-not present), **certainity id, Time**



Using this table we can identify which attributes are present in an image and extract them. The time denotes the time taken my **Mturker** to identify a particular attribute in an image. Since we are also given the bounding box for each image we can try to find this attribute in that bounding box rather than the whole image.

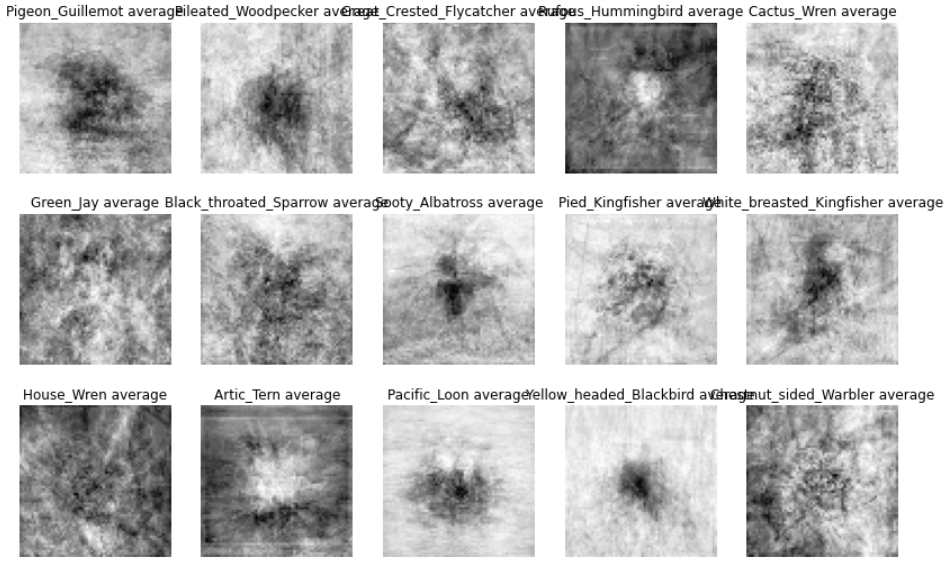
The following is a random selection of training images that have been taken from the dataset.



Images of birds in flight may be a point of confusion for the classifier, as it appears that there are many more images of birds sitting still than there are of those in flight. There is no attribute describing if a bird is in flight, so with a seemingly low number of training examples for birds in flight, this may cause errors.

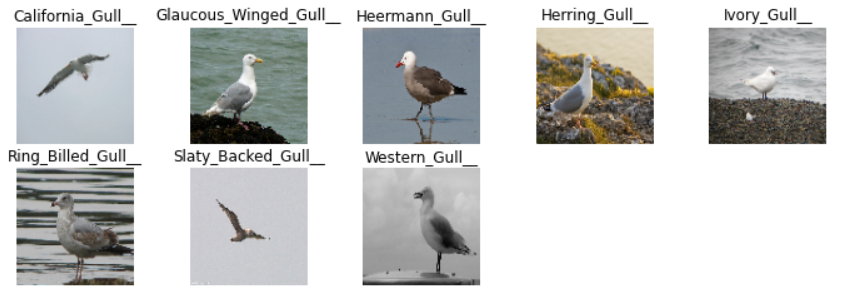
The picture of the Yellow Billed Cuckoo above is obscured by a number of branches. This isn’t ideal for a training image, as the important part of the image for this situation, the bird, is slightly obscured. If too many training images like this appear for one bird, then classification errors may arise.

By converting images into a matrix representation using the image class from tensorflow.keras.preprocessing, finding what the average image for each species was possible. The following are images representing an average image for some random species. The images are converted into black and white before find the average image.

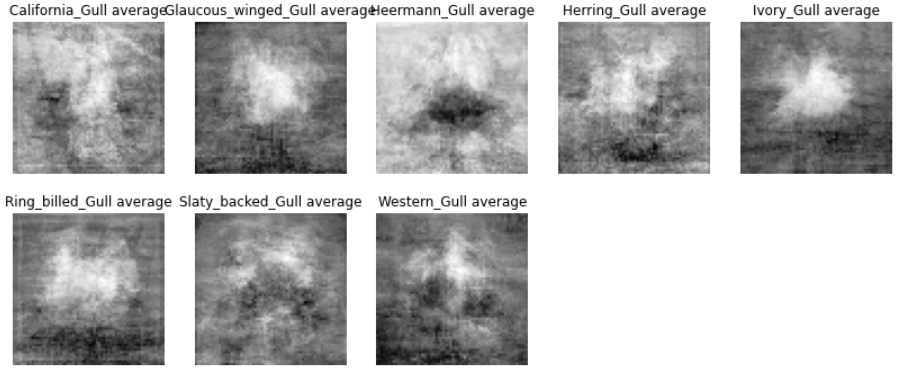


Some of these averages look very similar to each other, which may indicate problem points. The kingfisher example is interesting however, as the darker area in the image has an appearance not too far from the shape of a bird. It will be interesting to see how well that specific species of Kingfisher is classified.

The classifier may struggle on deciding between birds of similar species. Many of the Gulls for example look incredibly similar. They are all relatively similar colours and shapes, and this could lead to errors in classification. The following is one training image of every type of Gull in the dataset.



There are few noticeable differences between many of these species, and the attributes provided may not be enough to classify between them. The following is the average of the training images for the 8 different types of gull in the dataset:



Some of the averages look very similar, which could indicate that the classifier may struggle to differentiate between them. For example, the Glaucous Winged Gull, the Ivory Gull and the Ring Billed Gull average all look quite similar. The Heermann gull average is also interesting, as it is almost opposite to all the others, having a dark patch with white surroundings. Seeing how well the classifier is able to classify Heermann gulls from all the other gulls will be interesting. The dataset has provided bounding boxes for each image, so the classifier will use cropped versions of the images shown above to train. However, this does give an insight into the data that will be trained upon.