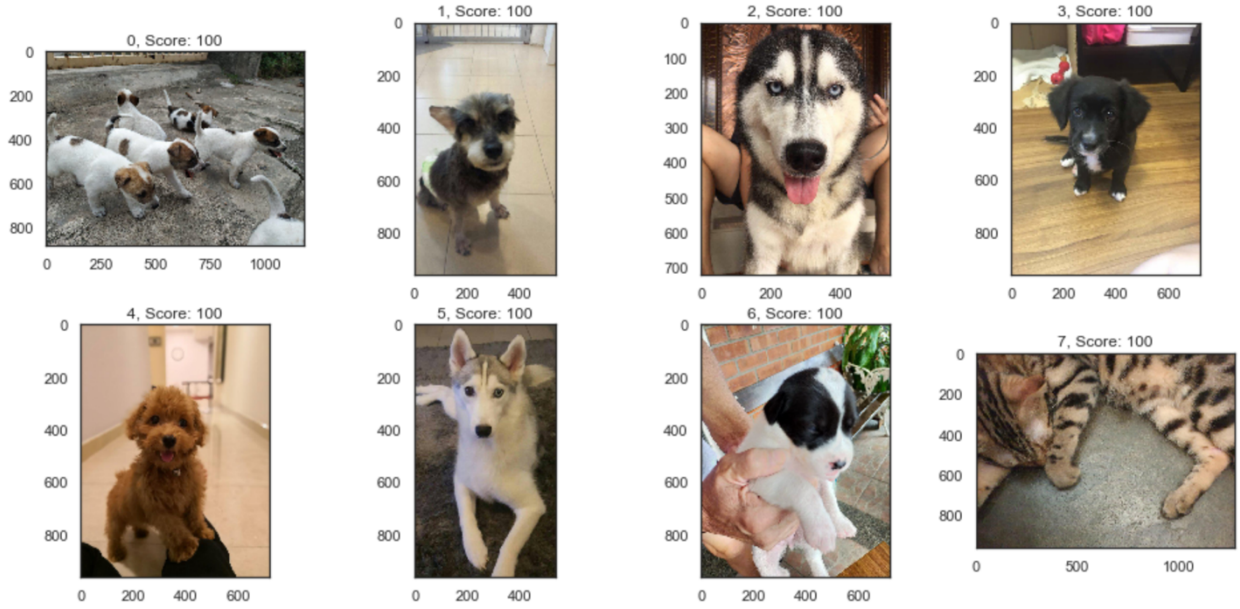
The dataset has 9912 images with 1.06GB in size. They are JPG images and stored in the training folder, the file names are unique ids.

There are three interesting findings from image data.

1. It is hard to predict the pawpularity score by looking at the images.

   Here are 8 images with lowest score:

   As comparison, here are 8 images with highest score:



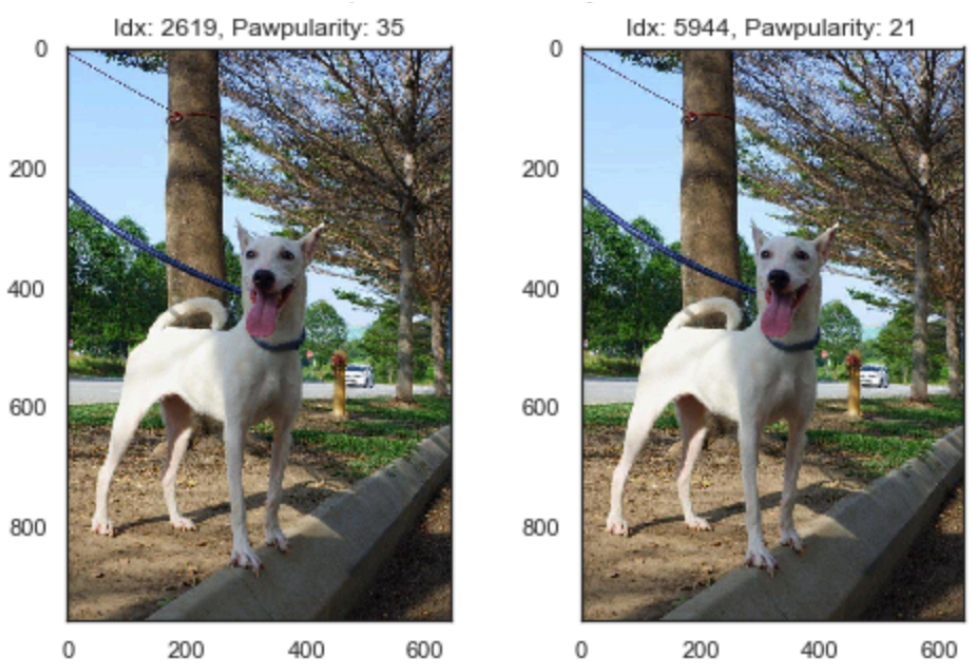
Personally, there is no obvious difference between the low and high scoring images. For example, low scoring images 2 and 3 seem quite cute, whereas high scoring image 2, 3

and 7 do not seem like good photos. I could not accurately predict pawpularity based on these images, which makes me think a machine learning algorithm is also going to have a hard time predicting pawpularity.

2. There are duplicated images but with different scores

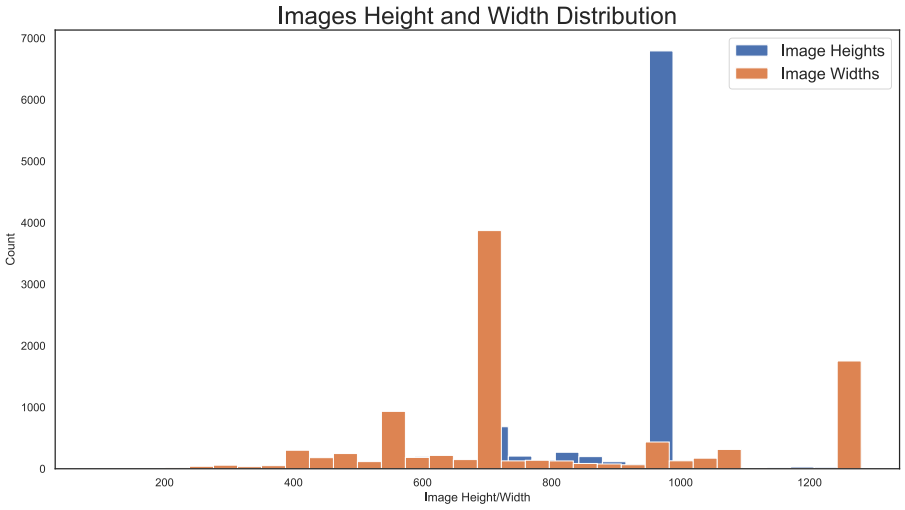
There are about 27 pairs of duplicate images in train dataset with different scores, which would definitely affect the performance of prediction. So, it would be better to remove these images from the dataset.

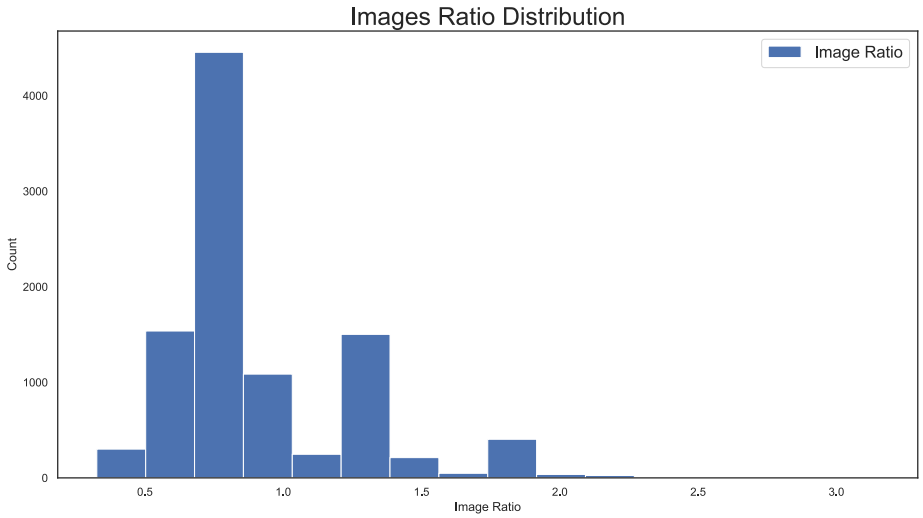
Here are several examples:



3. Image with different resolution

By looking at the distribution of the image width, height and ratio, Images tend to be large for an image classification task, the most common image size is 960*720. This tells us that we need to reshape or resize the images when we end up building our models. The image width to height ratio have a mean below zero and a peak on 0.75, pictures thus tend to be taken vertically, not horizontally.




Images Ratio Distribution
Image Ratio
Count
4000
3000
2000
1000
0
0.5
1.0
1.5
2.0
2.5
3.0
Image Ratio