DataGradients™

DET RF100 wine-labels

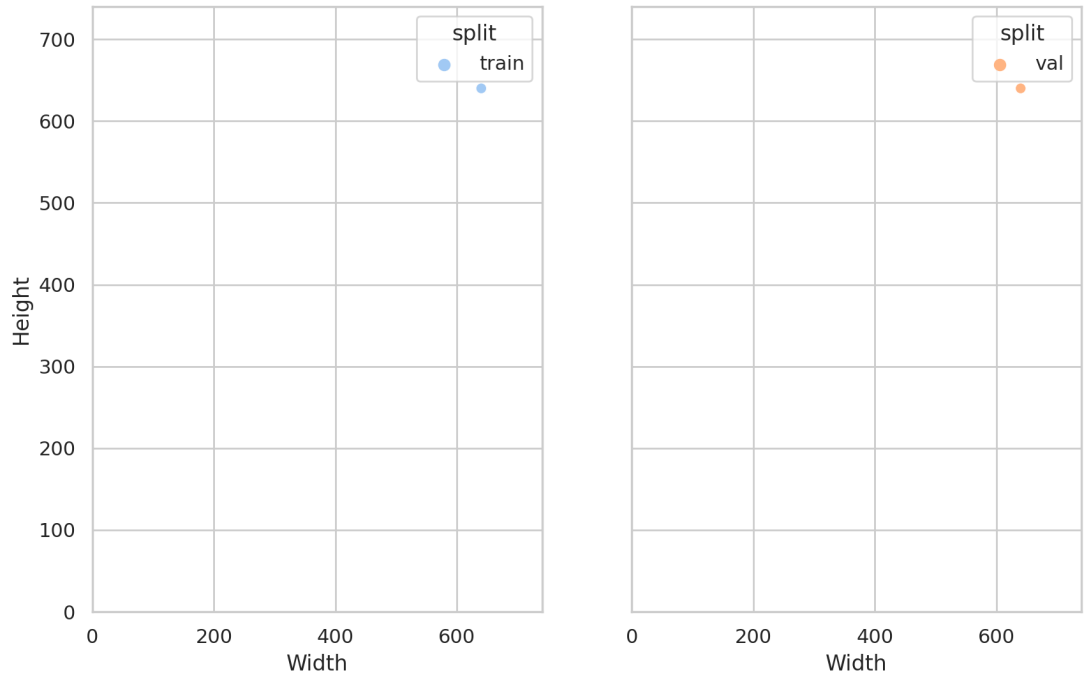
07:13 July 25, 2023

1. Selected features

1.1. General Statistics

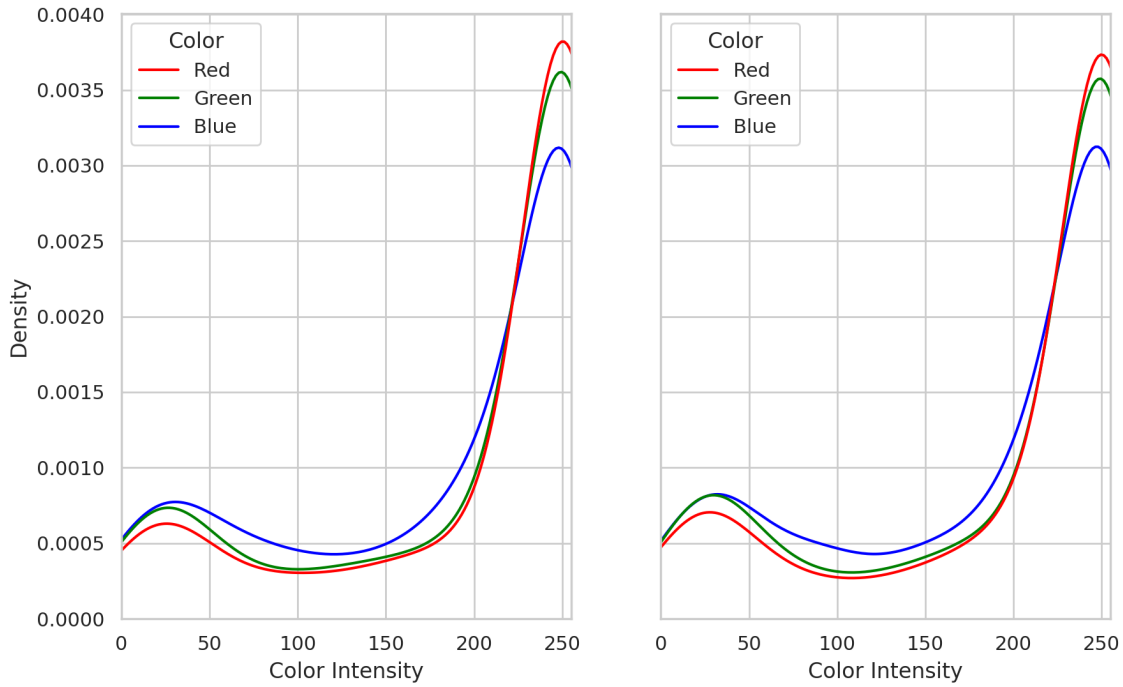
	Train	Validation
Images	3172	841
Classes	13	13
Classes in use	12	12
Annotations	17029	4613
Annotations per images	5.37	5.49
Images with no annotations	1	0
Median image resolution	640x640	640x640
Smallest annotation	4	135
Largest annotation	407044	408321
Most annotations in an image	12	14
Least annotations in an image	0	1

1.2. Image Width and Height Distribution



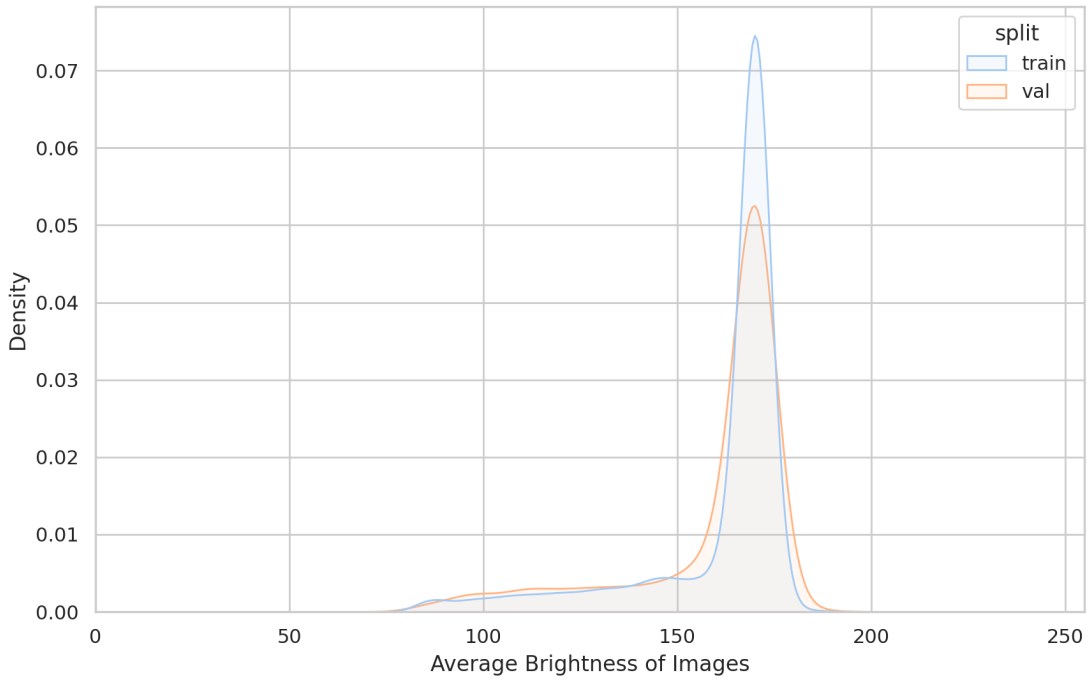
These histograms depict the distributions of image height and width. It's important to note that if certain images have been rescaled or padded, the histograms will represent the size after these operations.

1.3. Color Distribution



Here's a comparison of RGB or grayscale intensity intensity (0-255) distributions across the entire dataset, assuming RGB channel ordering.
It can reveal discrepancies in the image characteristics between the two datasets, as well as potential flaws in the augmentation process.
E.g., a notable difference in the mean value of a specific color between the two datasets may indicate an issue with the augmentation process.

1.4. Image Brightness Distribution



This graph shows the distribution of the brightness levels across all images.
This may for instance uncover differences between the training and validation sets, such as the presence of exclusively daytime images in the training set and nighttime images in the validation set.

1.5. Image Duplicates

Train duplicated images:
There are 41 duplicated images appearing 100 times across the dataset.

Validation duplicated images:
There are 5 duplicated images appearing 10 times across the dataset.

There are 18 duplicates between train and validation appearing 30 times in the train image directory, and 19 times in the validation image directory.

1.6. Visualization of Samples

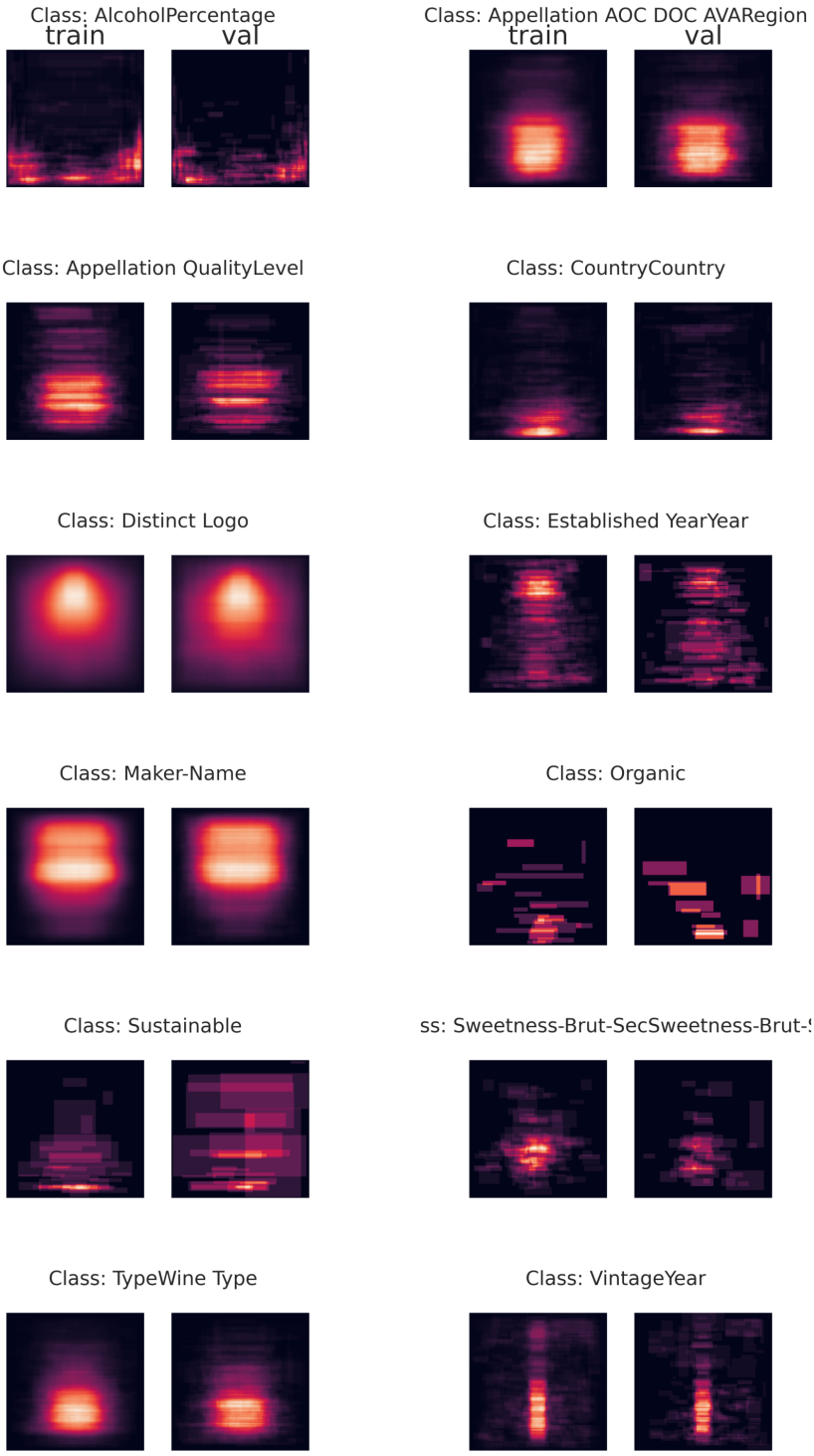
train

val

The sample visualization feature provides a visual representation of images and labels. This visualization aids in understanding of the composition of the dataset.

Notice: Only 12 random samples are shown.
You can increase the number of images by changing `n_cols` and `n_rows` in the configuration file.

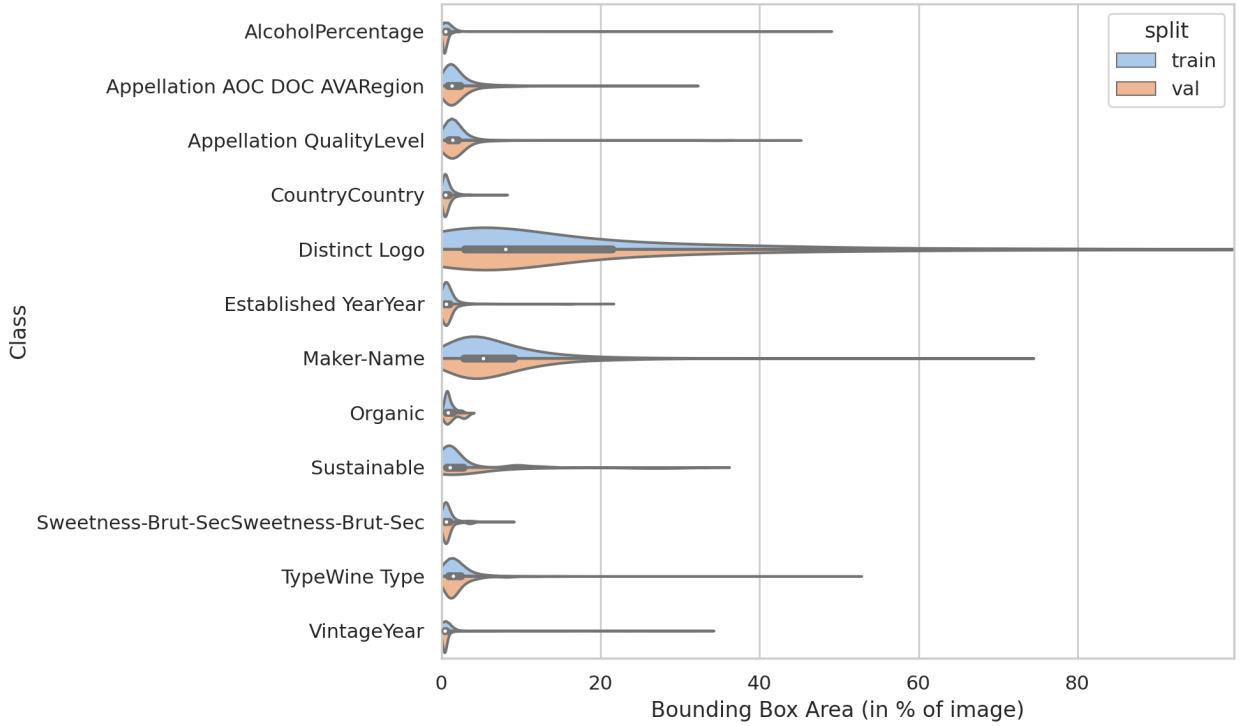
1.7. Bounding Box Density



The heatmap represents areas of high object density within the images, providing insights into the spatial distribution of objects. By examining the heatmap, you can quickly detect whether objects are predominantly concentrated in specific regions or if they are evenly distributed throughout the scene. This information can serve as a heuristic to assess if the objects are positioned appropriately within the expected areas of interest.

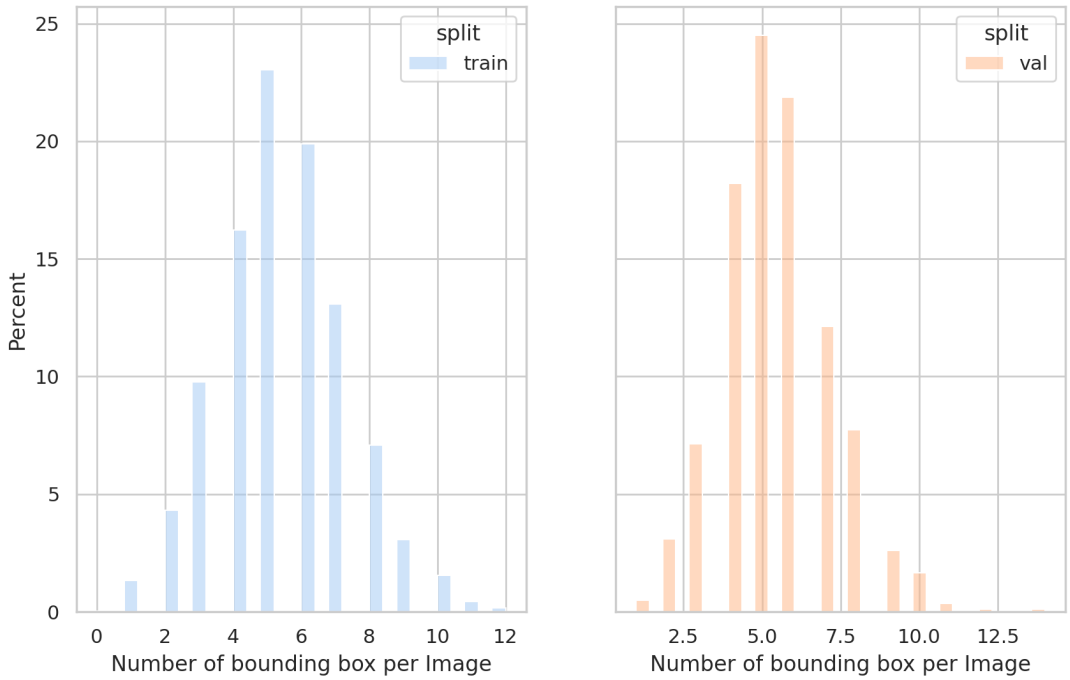
Notice: Only the 12 classes with highest density are shown.
You can increase the number of classes by changing `n_cols` and `n_rows` in the configuration file.

1.8. Distribution of Bounding Box Area



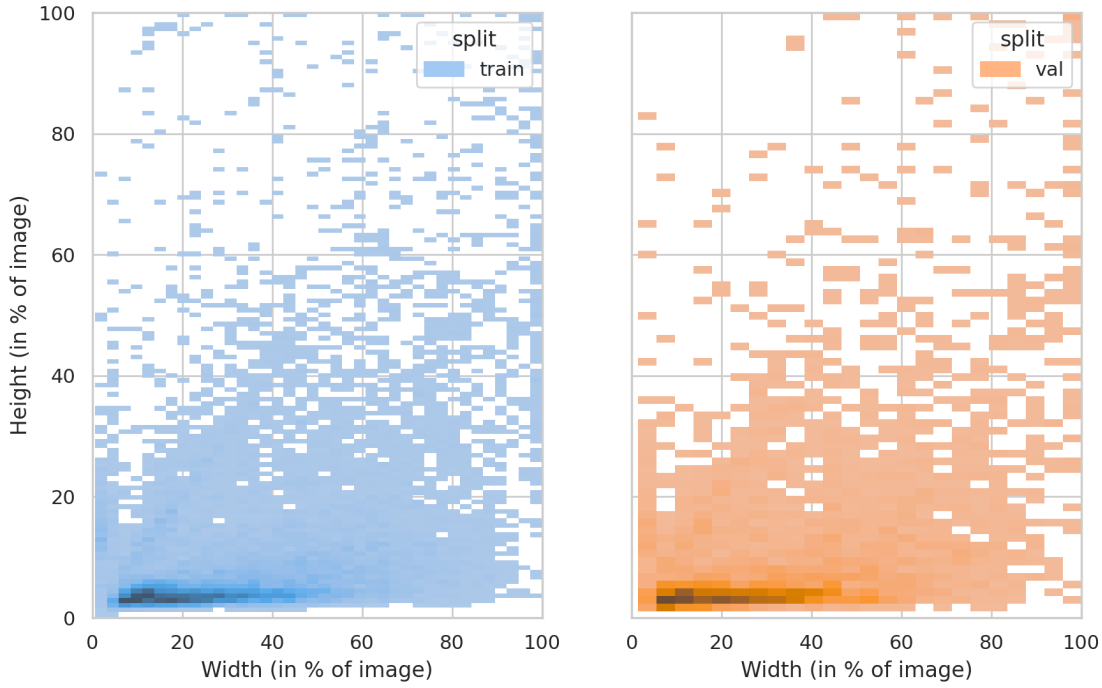
This graph shows the frequency of each class's appearance in the dataset. This can highlight distribution gap in object size between the training and validation splits, which can harm the model's performance.
Another thing to keep in mind is that having too many very small objects may indicate that your are downsizing your original image to a low resolution that is not appropriate for your objects.

1.9. Distribution of Bounding Box per image



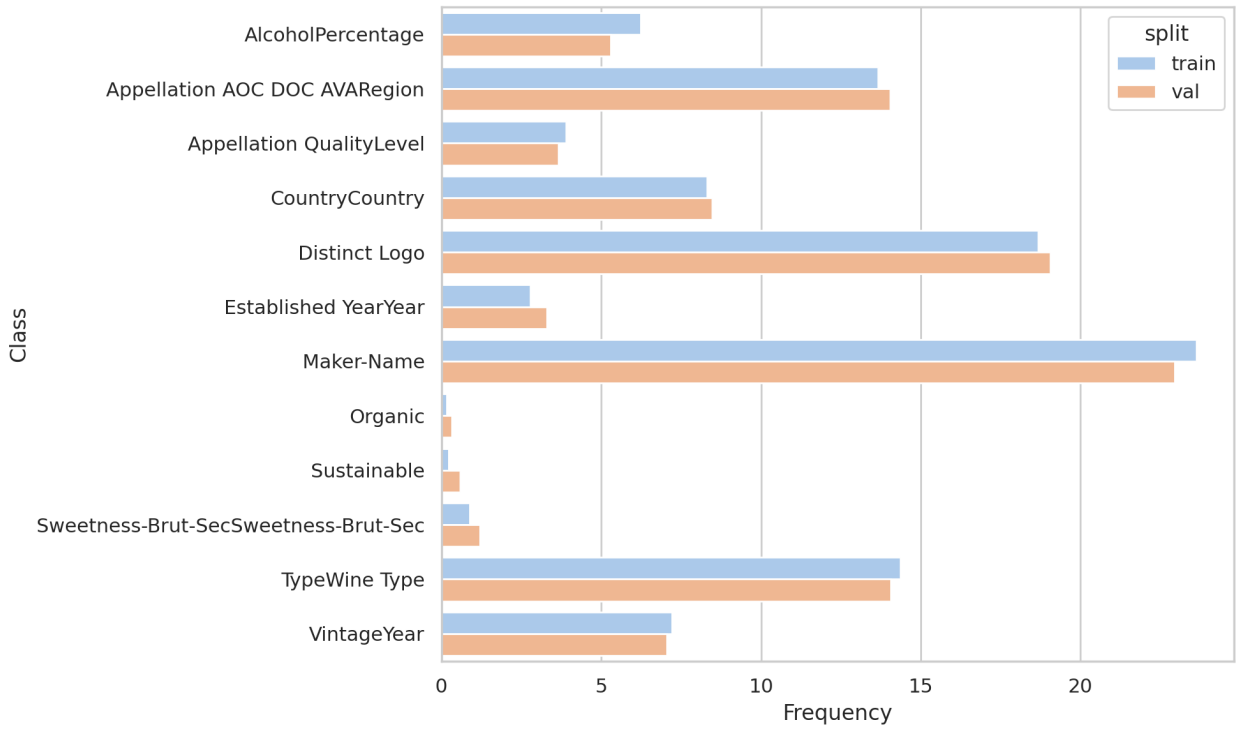
These graphs shows how many bounding boxes appear in images.
This can typically be valuable to know when you observe a very high number of bounding boxes per image, as some models include a parameter to filter the top k results.

1.10. Distribution of Bounding Box Width and Height



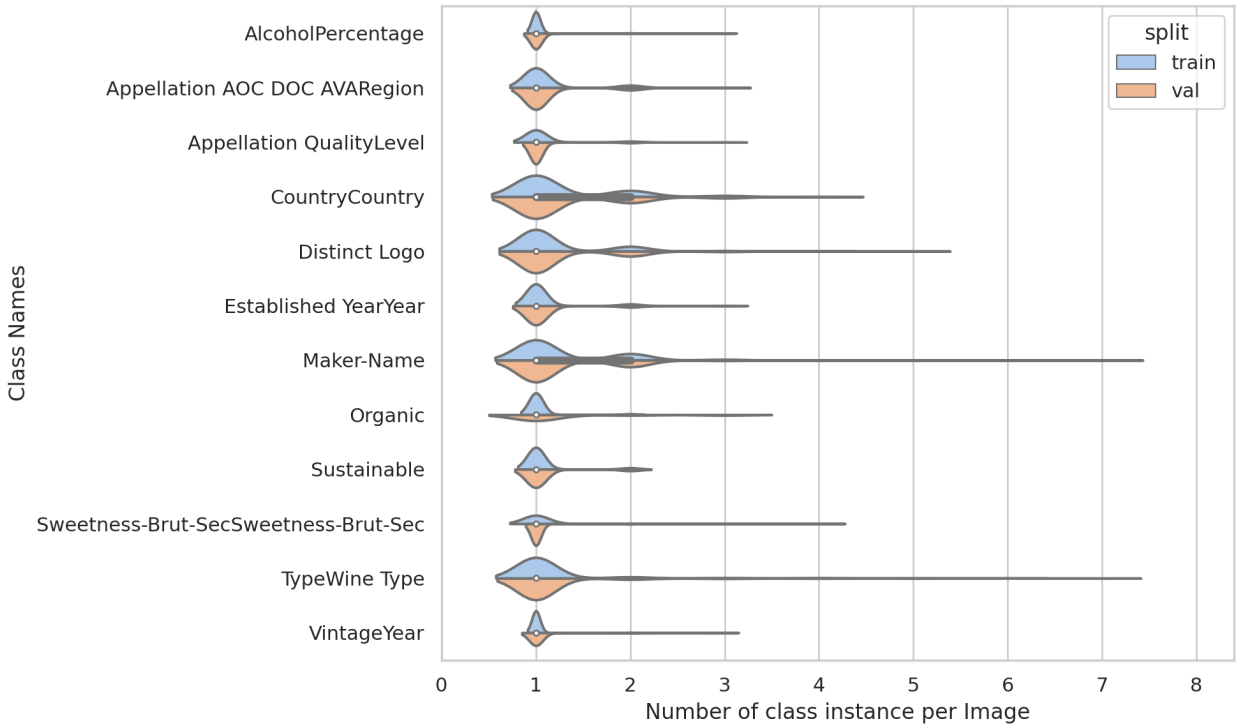
These heat maps illustrate the distribution of bounding box width and height per class.
Large variations in object size can affect the model's ability to accurately recognize objects.

1.11. Class Frequency



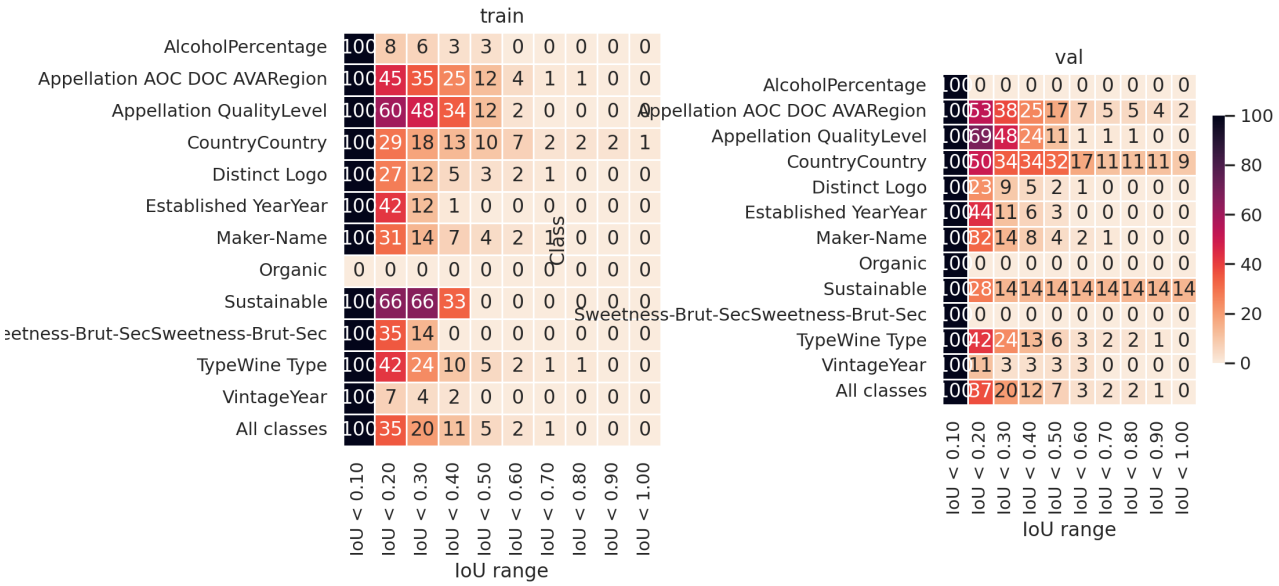
Frequency of appearance of each class. This may highlight class distribution gap between training and validation splits. For instance, if one of the class only appears in the validation set, you know in advance that your model won't be able to learn to predict that class.

1.12. Distribution of Class Frequency per Image



This graph shows how many times each class appears in an image. It highlights whether each class has a constant number of appearances per image, or whether there is variability in the number of appearances from image to image.

1.13. Intersection of Bounding Boxes



The distribution of the box Intersection over Union (IoU) with respect to other boxes in the sample. The heatmap shows the percentage of boxes that overlap with IoU in range [0..T] for each class. Intersection of all boxes is considered (Regardless of classes of corresponding bboxes).

Notice: To better understand how to tackle the data issues highlighted in this report, explore our comprehensive course on analyzing computer vision datasets.