WIDS 5.0 – WEEK 1

EDA

PLANT VILLAGE DATASET

ABOUT THE DATASET

The dataset contains images of plant leaf samples collected under two circumstances.

1. When the plant is healthy
2. When the plant is suffering from a disease

The dataset is highly diverse encompassing various species and their condition during different types of diseases.

The dataset includes a total of 14 species of plants and 38 different classes of images.

INITIAL ANALYSIS

The dataset consists of 14 different species of plants.

```
['Apple', 'Blueberry', 'Cherry', 'Corn', 'Grape',
'Orange', 'Peach', 'Pepper,', 'Potato', 'Raspberry',
'Soybean', 'Squash', 'Strawberry', 'Tomato']
```

ASPECT RATIO AND RESOLUTION

The images in dataset are uniform in size and resolution. The images share a common resolution and aspect ratio.

ASPECT RATIO : 1.0
RESOLUTION : 256 x 256 = 65536 pixels

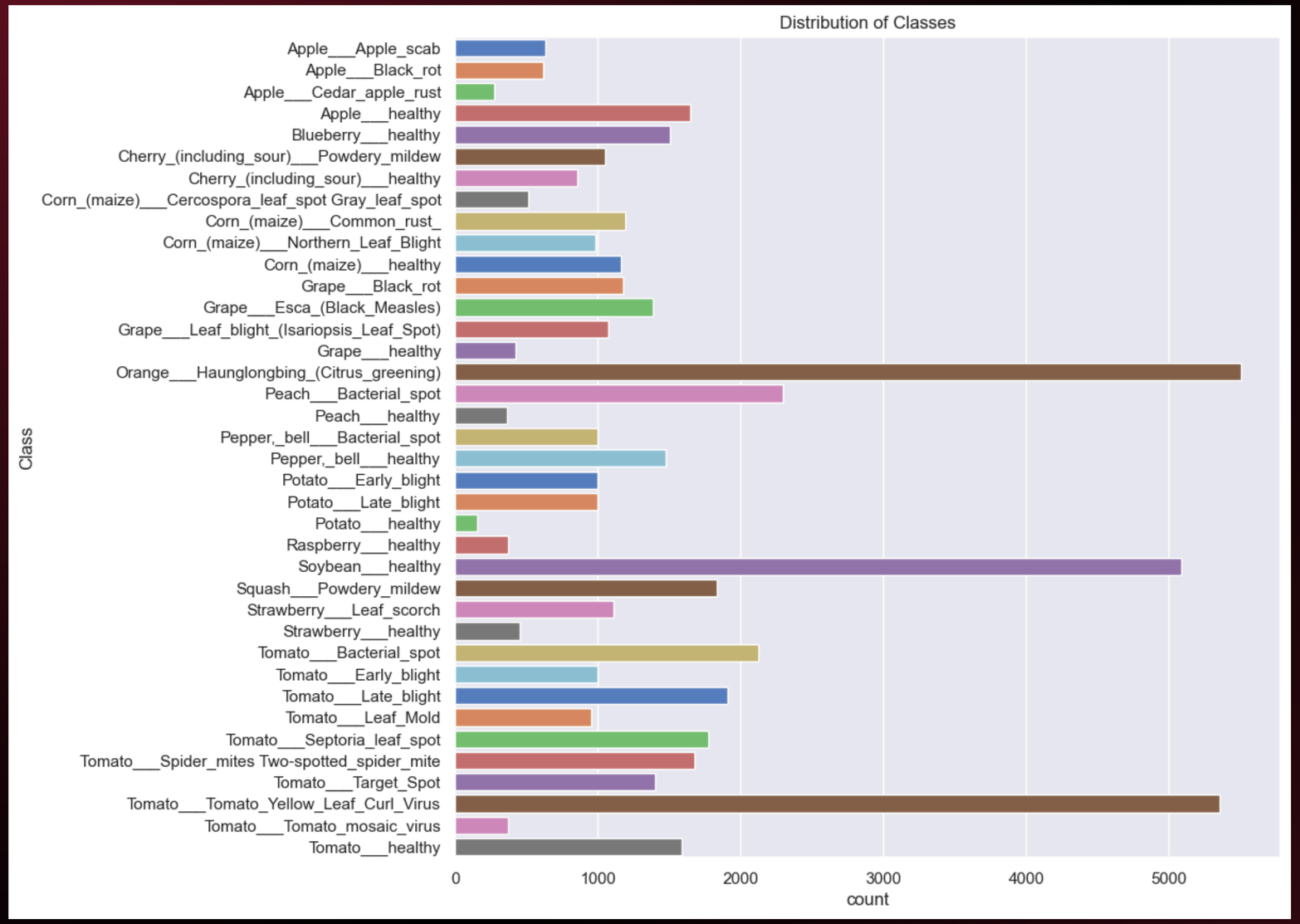
The dataset takes into account multiple disease conditions for each plant which amount to around 38 different classes in which the dataset is divided.

```
['Apple___Apple_scab' 'Apple___Black_rot' 'Apple___Cedar_apple_rust'
 'Apple___healthy' 'Blueberry___healthy'
 'Cherry_(including_sour)___Powdery_mildew'
 'Cherry_(including_sour)___healthy'
 'Corn_(maize)___Cercospora_leaf_spot Gray_leaf_spot'
 'Corn_(maize)___Common_rust_' 'Corn_(maize)___Northern_Leaf_Blight'
```

```
'Corn_(maize)___healthy' 'Grape___Black_rot'
'Grape___Esca_(Black_Measles)'
'Grape___Leaf_blight_(Isariopsis_Leaf_Spot)' 'Grape___healthy'
'Orange___Haunglongbing_(Citrus_greening)' 'Peach___Bacterial_spot'
'Peach___healthy' 'Pepper,_bell___Bacterial_spot'
'Pepper,_bell___healthy' 'Potato___Early_blight' 'Potato___Late_blight'
'Potato___healthy' 'Raspberry___healthy' 'Soybean___healthy'
```

```
'Squash___Powdery_mildew' 'Strawberry___Leaf_scorch'
'Strawberry___healthy' 'Tomato___Bacterial_spot' 'Tomato___Early_blight'
'Tomato___Late_blight' 'Tomato___Leaf_Mold' 'Tomato___Septoria_leaf_spot'
'Tomato___Spider_mites Two-spotted_spider_mite' 'Tomato___Target_Spot'
'Tomato___Tomato_Yellow_Leaf_Curl_Virus' 'Tomato___Tomato_mosaic_virus'
'Tomato___healthy']
```

CLASS DISTRIBUTION



INFERENCE :

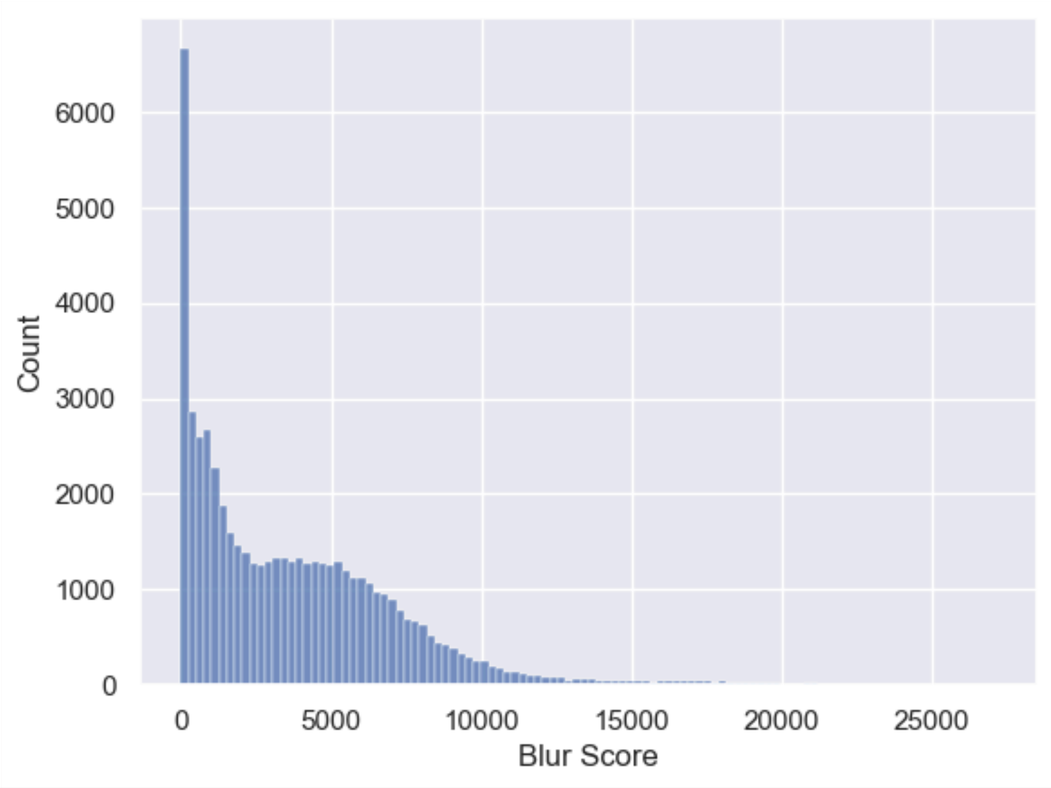
The dataset presents a classic case of class imbalance.

Certain classes are very highly represented having more than 5000 samples.

On the other hand certain classes are very poorly represented having less than 500 samples.

This class imbalance is very unfavorable as this makes the model biased towards the dominant classes.

BLURRY IMAGES



The histogram suggests that many images have very low blur score indicating that a substantial portion of dataset comprises of blurry photos. Such photos if used will yield poor results.

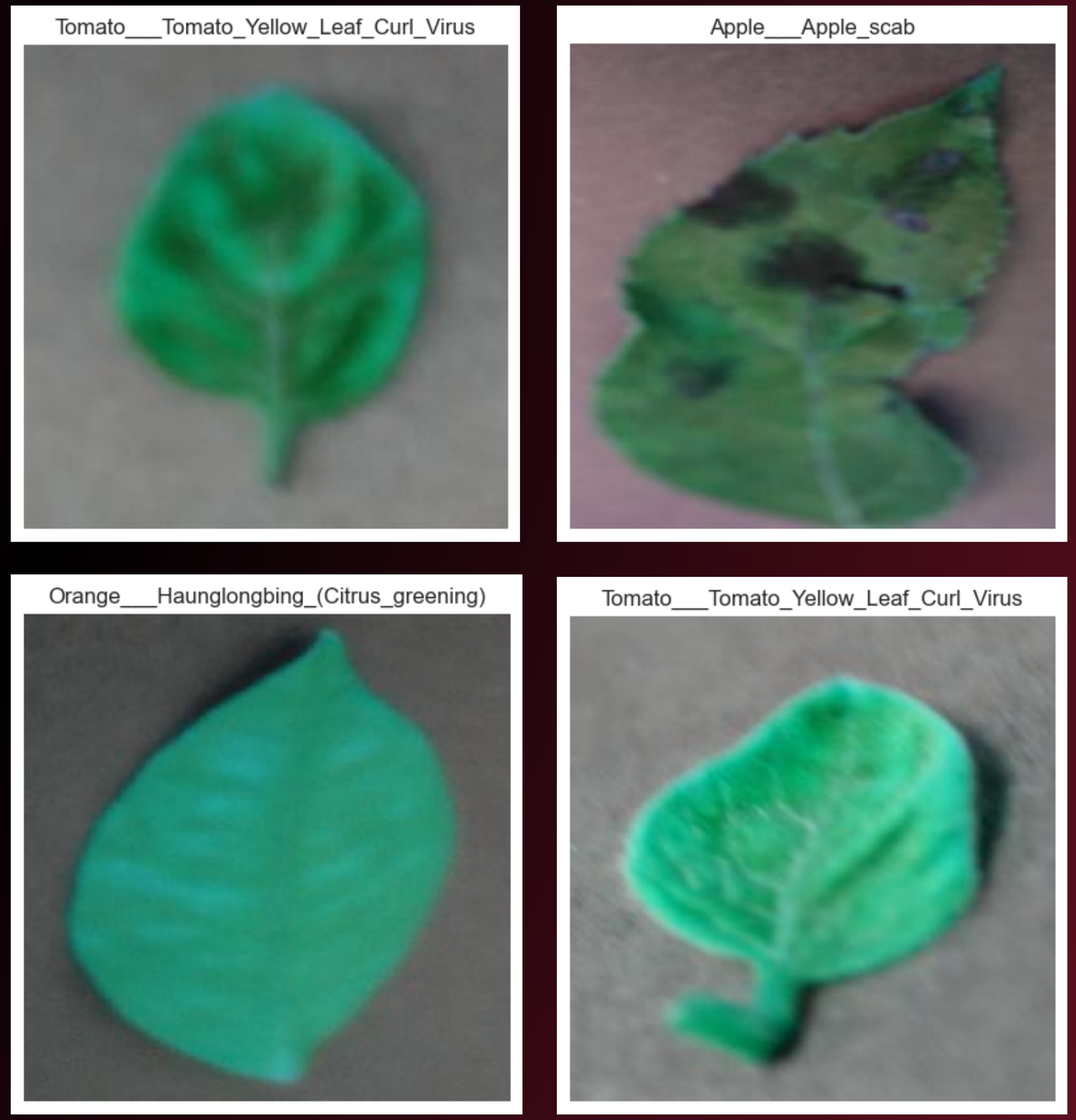


Fig. Examples of blurry images

VARIATION IN QUALITY

Fig. Sharpest Images

Fig. Most blurry images

Fig. Average Sharpness

VARIATION IN BRIGHTNESS

Fig. Brightest Images

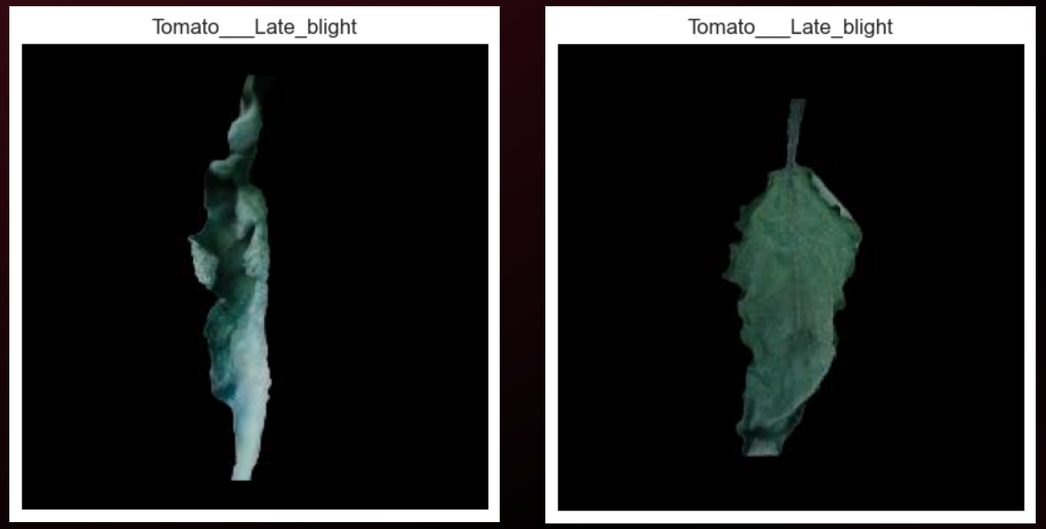
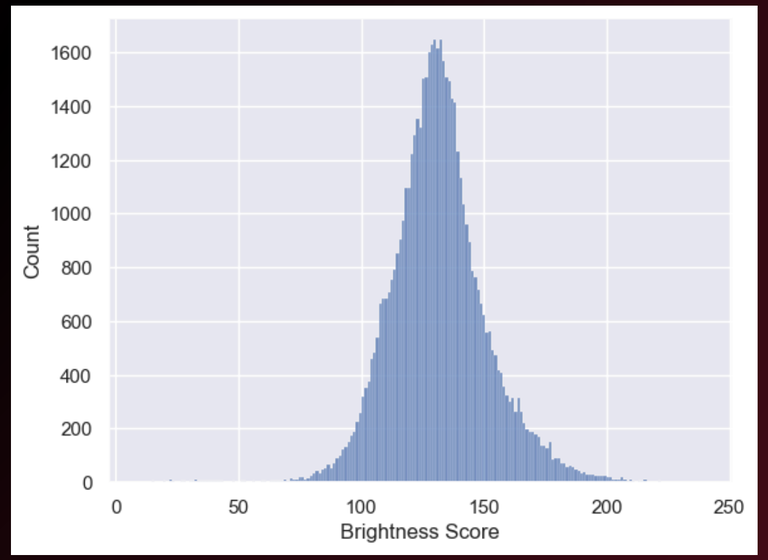


Fig. Dullest Images

Fig. Average Brightness



The plot suggests that most of the images have brightness around the mean. Short left tail suggests that darkness is rare in the dataset and the images are more or less well lit.

VARIATION IN BACKGROUND

Background in an image is an important aspect that must not be overlooked. If the images have varying backgrounds, it becomes a proxy variable for the model.

This means that, instead of learning what the leaf looks like (texture, veins, disease spots), the model might learn to cheat by just looking at the background color.

Our dataset consists of majority of images with light background.
About 2.5 % of images have black background.
A certain portion of images contain leaf in the entire region.

THANK YOU