Data Collection and Preprocessing Phase

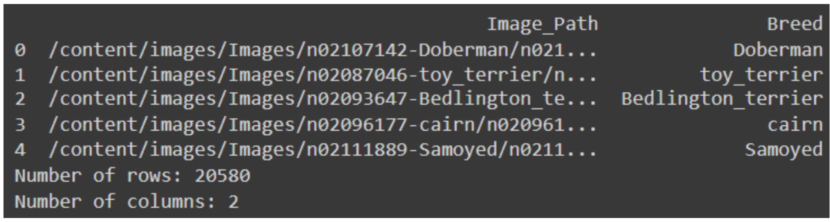
Date	27th June 2024
Team ID	SWTID1720073336
Project Title	Dog breed identification using transfer learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report:

In the dog breed identification project, dataset variables will be statistically analyzed to identify patterns and outliers. Python will be employed for preprocessing tasks such as normalization and feature engineering. This will ensure that the data is standardized and suitable for machine learning algorithms. Through data cleaning processes, missing values and outliers will be addressed, maintaining high data quality for subsequent analysis and modeling. These steps form a strong foundation, enabling the extraction of meaningful insights and the development of accurate predictive models for dog breed identification.

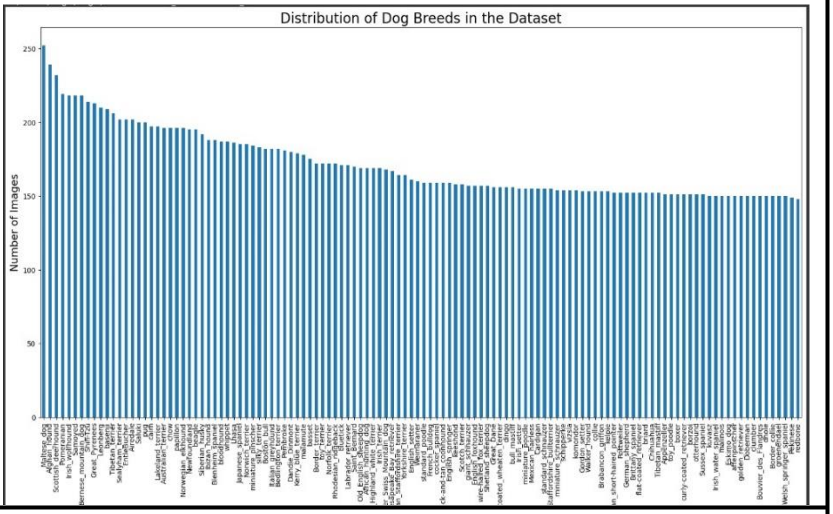
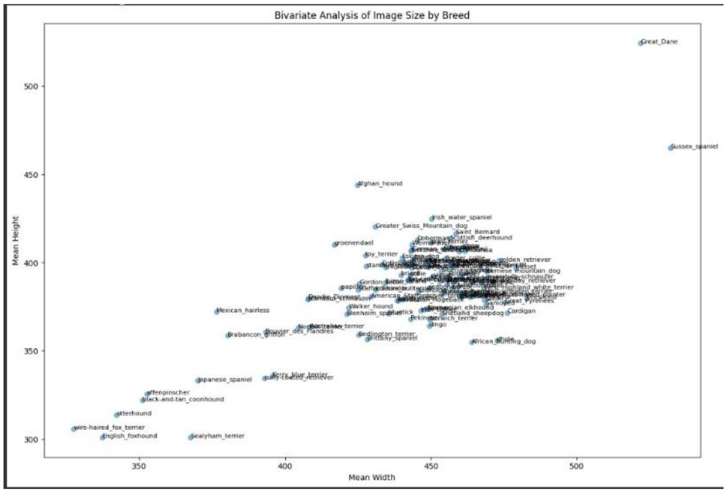
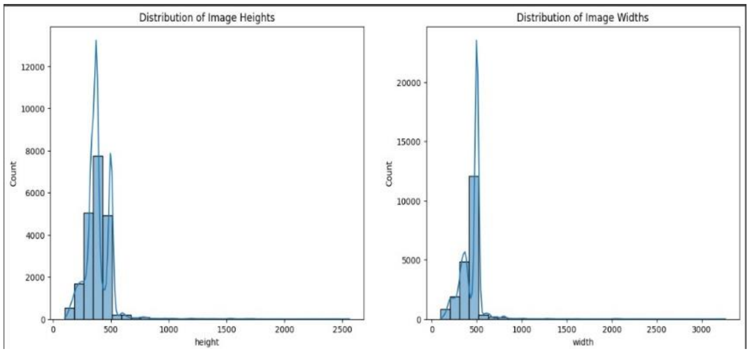
<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>

<u>Dimension:</u>
Rows: 20580 x
columns: 2
<u>Descriptive statistics:</u>



```
                                          Image_Path               Breed
0  /content/images/Images/n02107142-Doberman/n021...            Doberman
1  /content/images/Images/n02087046-toy_terrier/n...         toy_terrier
2  /content/images/Images/n02093647-Bedlington_te...  Bedlington_terrier
3  /content/images/Images/n02096177-cairn/n020961...               cairn
4  /content/images/Images/n02111889-Samoyed/n0211...             Samoyed
Number of rows: 20580
Number of columns: 2
```

</td></tr>
<tr><td>Univariate Analysis</td><td></td></tr>
</table>

	
Bivariate Analysis	
Multivariate Analysis	

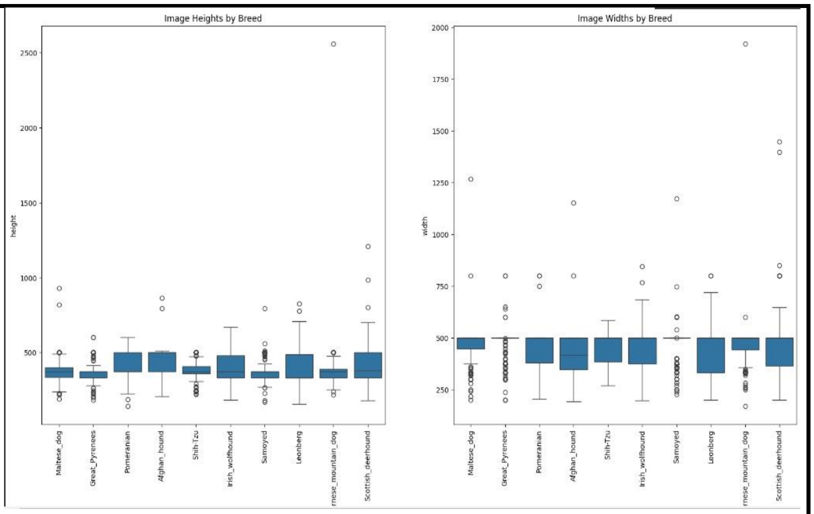
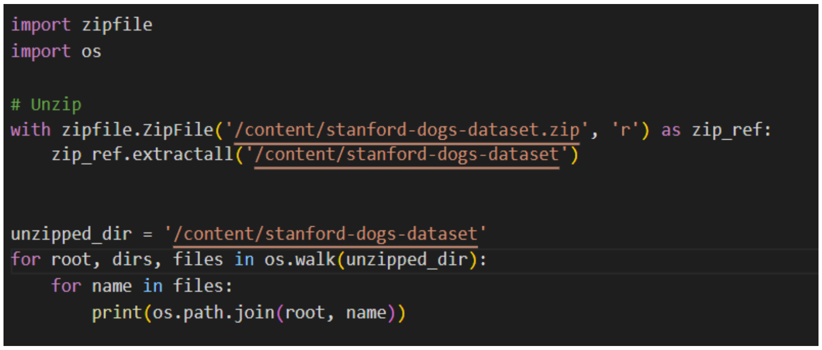
	

Image Heights by Breed
height
2500
2000
1500
1000
500
Maltese_dog
Great_Pyrenees
Pomeranian
Afghan_hound
Shih-Tzu
Irish_wolfhound
Samoyed
Leonberg
rnese_mountain_dog
Scottish_deerhound
Image Widths by Breed
width
2000
1750
1500
1250
1000
750
500
250
Maltese_dog
Great_Pyrenees
Pomeranian
Afghan_hound
Shih-Tzu
Irish_wolfhound
Samoyed
Leonberg
rnese_mountain_dog
Scottish_deerhound

<table>
<tr><td>Outliers and Anomalies</td><td>-</td></tr>
<tr><td colspan="2"><b>Data Preprocessing Code Screenshots</b></td></tr>
<tr><td>Unzipping Data</td><td>

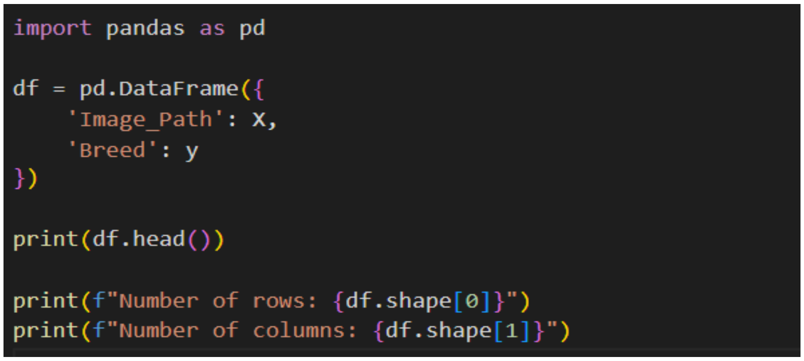


```
import zipfile
import os

# Unzip
with zipfile.ZipFile('/content/stanford-dogs-dataset.zip', 'r') as zip_ref:
    zip_ref.extractall('/content/stanford-dogs-dataset')


unzipped_dir = '/content/stanford-dogs-dataset'
for root, dirs, files in os.walk(unzipped_dir):
    for name in files:
        print(os.path.join(root, name))
```

</td></tr>
<tr><td>Data Transformation</td><td>



```
import pandas as pd

df = pd.DataFrame({
    'Image_Path': X,
    'Breed': y
})

print(df.head())

print(f"Number of rows: {df.shape[0]}")
print(f"Number of columns: {df.shape[1]}")
```

</td></tr>
</table>