Data Appendix

The purpose of this appendix is to provide a detailed description of the datasets used in the Exploratory Data Analysis (EDA) and our Image Data Analysis. This document details information regarding the variables of the datasets, descriptive statistics, and general observations and descriptions of all datasets used in this project.

Appendix A. Image Dataset ("102flowers.tgz")

Our dataset comes from the Oxford 102 Category Flower Dataset, which includes 8,189 images across 102 distinct flower categories. The dataset was developed by Maria-Elena Nilsback and Andrew Zisserman at the University of Oxford. It is provided as a compressed .tgz file titled **"102flowers.tgz"** and can be found via the link in the **DATA** folder of our GitHub repository.

Each image is labeled with a numeric class identifier, which corresponds to a specific flower category. These labels are stored in a file titled **"flowerimagelabels.csv"**, also available in the GitHub repository and on the dataset website under the name **"The image labels"**.

To facilitate easier flower classification, we matched each numeric label with its corresponding flower name using the category statistics provided on the dataset website. The resulting merged file, which contains both the numeric class and the flower name, is titled **"label_to_name.csv"** and can also be found in the **DATA** folder of our GitHub repository.

Descriptive Statistics:

- Total Images: 8,189 flower images, organized by 102 different flower classes
- Top 3 flowers with highest counts:
    - Petunia, label 51, 258 images

      
    - Passion Flower, label 77, 251 images

      
    - Wallflower, label 46, 196 images

      

Cleaned Data Dictionary:

The following data dictionary contains information about variables deemed relevant for our analysis and providing proper contextualization.

- **Label Variable:** This is a numeric class identifier ranging from 1 to 102, where each number corresponds to a specific flower category. These were assigned by the original dataset authors.
- **Flower Name Variable:** The name of the flower corresponding to the numeric label. This was mapped using the dataset's category statistics and merged into the main dataset.

Label	Flower Name
1	pink primrose
2	hard-leaved pocket orchid
3	canterbury bells
4	sweet pea
5	english marigold
6	tiger lily
7	moon orchid
8	bird of paradise
9	monkshood
10	globe thistle
11	snapdragon
12	colt's foot
13	king protea
14	spear thistle
15	yellow iris
16	globe-flower
17	purple coneflower
18	peruvian lily
19	balloon flower
20	giant white arum lily
21	fire lily
22	pincushion flower
23	fritillary
24	red ginger
25	grape hyacinth
26	corn poppy
27	prince of wales feathers
28	stemless gentian
29	artichoke

30	sweet william
31	carnation
32	garden phlox
33	love in the mist
34	mexican aster
35	alpine sea holly
36	ruby-lipped cattleya
37	cape flower
38	great masterwort
39	siam tulip
40	lenten rose
41	barberton daisy
42	daffodil
43	sword lily
44	poinsettia
45	bolero deep blue
46	wallflower
47	marigold
48	buttercup
49	oxeye daisy
50	common dandelion
51	petunia
52	wild pansy
53	primula
54	sunflower
55	pelargonium
56	bishop of llandaff
57	gaura
58	geranium
59	orange dahlia
60	pink-yellow dahlia?
61	cautleya spicata
62	japanese anemone
63	black-eyed susan
64	silverbush
65	californian poppy
66	osteospermum

67	**spring crocus**
68	**bearded iris**
69	**windflower**
70	**tree poppy**
71	**gazania**
72	**azalea**
73	**water lily**
74	**rose**
75	**thorn apple**
76	**morning glory**
77	**passion flower**
78	**lotus**
79	**toad lily**
80	**anthurium**
81	**frangipani**
82	**clematis**
83	**hibiscus**
84	**columbine**
85	**desert-rose**
86	**tree mallow**
87	**magnolia**
88	**cyclamen**
89	**watercress**
90	**canna lily**
91	**hippeastrum**
92	**bee balm**
93	**ball moss**
94	**foxglove**
95	**bougainvillea**
96	**camellia**
97	**mallow**
98	**mexican petunia**
99	**bromelia**
100	**blanket flower**
101	**trumpet creeper**
102	**blackberry lily**

Table 1: Data Dictionary

Visualizations:

For our visualizations, we explored image properties by plotting the distributions of image widths and heights using a log scale, which revealed consistent sizing with some variability (Figure 1). A log-transformed plot of image areas showed that most images clustered around a typical size (Figure 2), while aspect ratio analysis indicated that the majority were nearly square, with slight variation (Figure 3). We also examined class balance using a count plot of flower labels, which revealed some overrepresented classes—most notably “petunia” (label 51) and “passion flower” (label 77), as shown in Figure 4. Finally, we displayed five random images with their class labels to visually confirm dataset integrity and assess image quality and diversity (Figure 5).

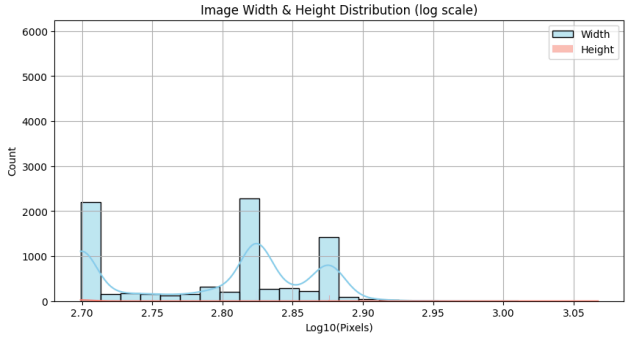

Figure 1: Distribution of Image Width & Height

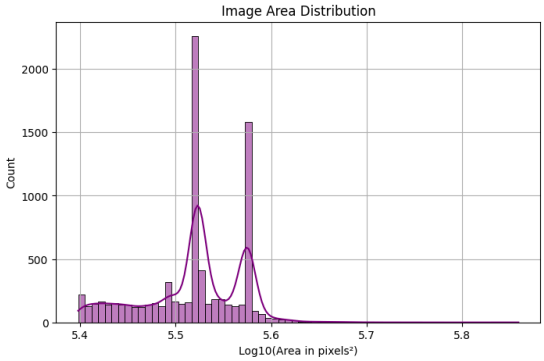

Figure 2: Distribution of Image Pixel Areas

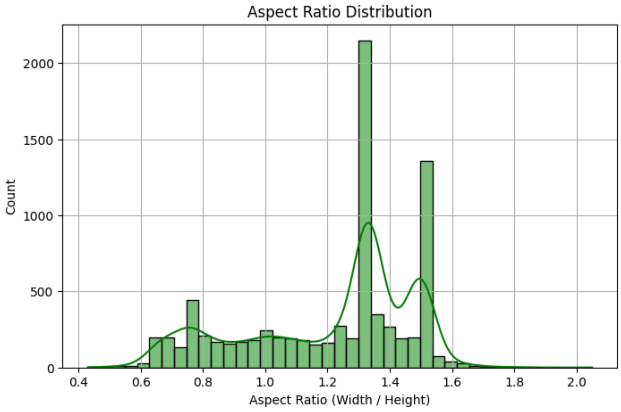

Figure 3: Distribution of Image Aspect Ratios

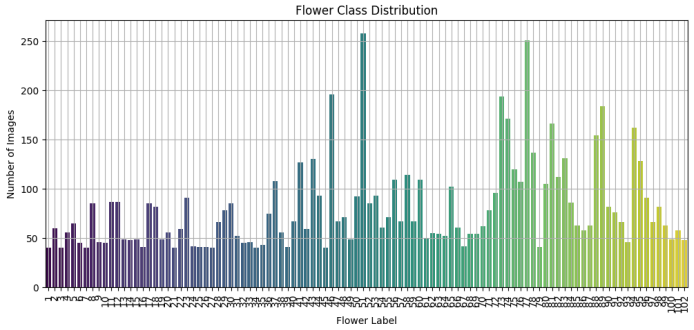

Figure 4: Distribution of Flower Class Labels

Figure 5: Sample of Five Random Images From Dataset