1. Introduction

This document serves as the Data Appendix for the image data project "Pet-Breed Classification". It provides a structured overview of the datasets and transformations applied, from raw input data to analysis data.

2. Data Pipeline Workflow

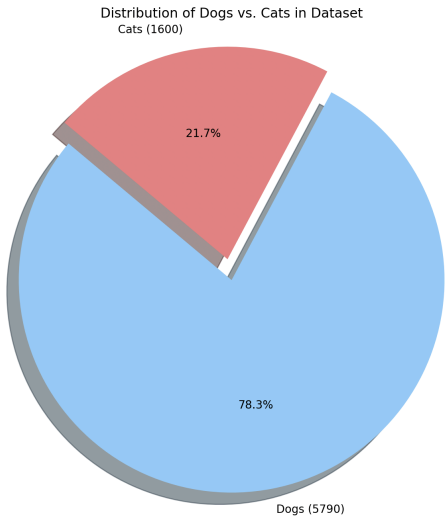
Step 1: Raw Image Archive (InputData/images)

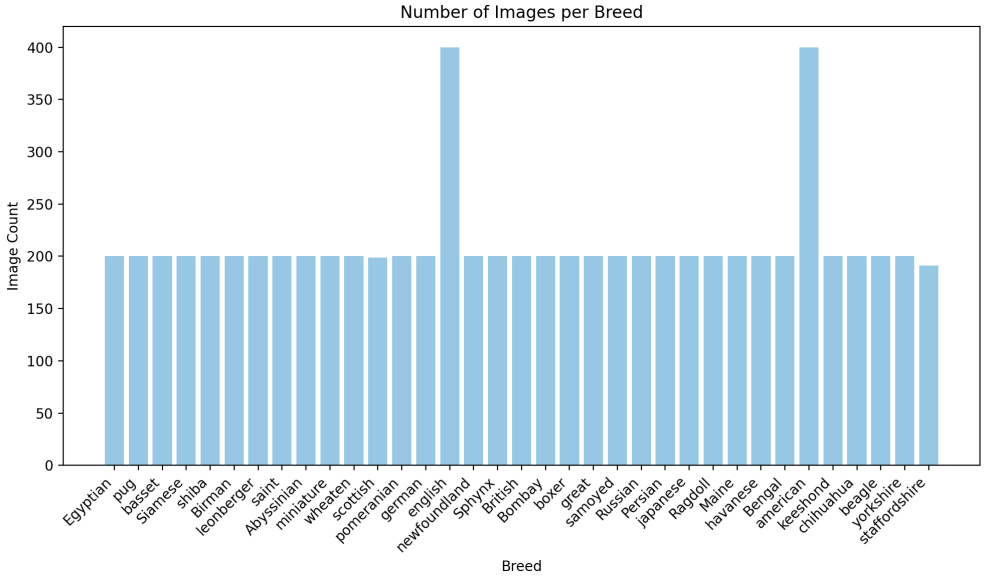
- **Unit of Observation**: One JPEG file, each a single photo of a dog or cat.
- **Key Variables** (extracted from filename or EXIF):
    - file_name – Original image filename (e.g., Abyssinian_1.jpg).
    - breed_label – Breed extracted from filename (text before first "_").
- **Purpose**: This raw archive is the primary source of images used to train and evaluate the breed-classification CNN.
- **Processing Steps**:
    - Downloaded the Images archive from the Oxford-IIIT Pet Dataset webpage.
    - Extracted all 9,687 JPEGs into the project's images/ folder.

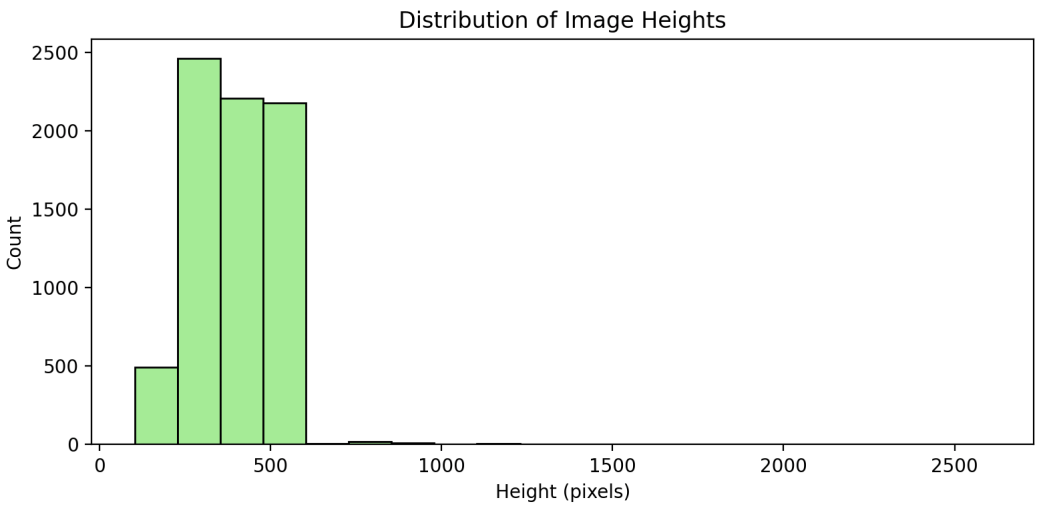
Step 2: Organized Image Folders (AnalysisData/organized_images)

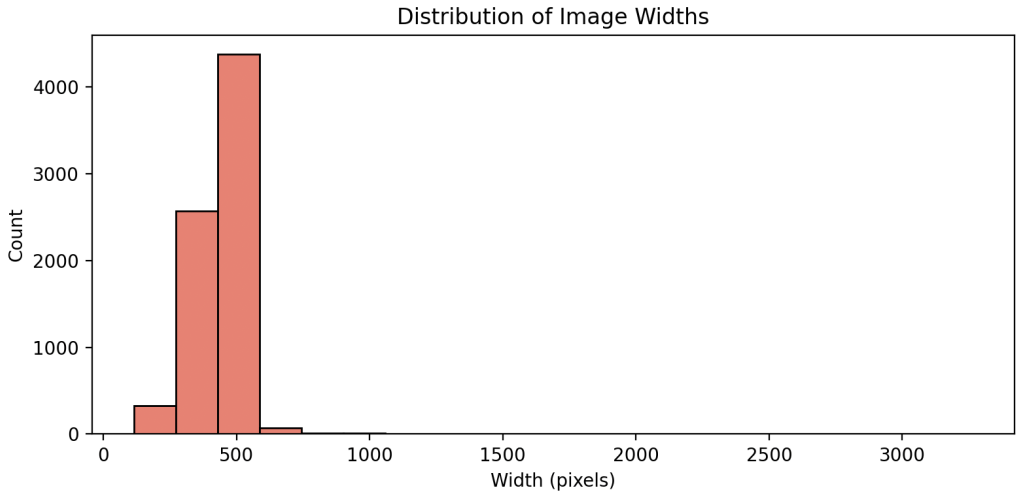
- **Unit of Observation:** Each row (file) represents one JPEG stored inside a breed-specific subfolder (organized_images/<breed>/).
- **Key Variables:**
    - file_name – Original image filename (e.g., Abyssinian_1.jpg).
    - breed_label – Breed extracted from filename (text before first "_").
    - breed_folder – Name of subfolder; becomes the class label for the CNN.
- **Purpose:** Provides a directory structure where each subfolder is automatically recognized as a separate class.
- **Processing Steps:**
    - Ran data_organization.py, which:
        1. Parsed each filename to obtain breed_label.
        2. Created a sub-directory for each breed (35 total).
        3. Moved images into their respective breed folders.

3. Summary Statistics and Visualizations






Distribution of Image Heights
Count
Height (pixels)
0
500
1000
1500
2000
2500


Distribution of Image Widths
Count
Width (pixels)
0
1000
2000
3000
4000
500
1500
2500