Data Appendix

I. Dataset: TrashNet

Derived from the Hugging Face website, the publicly available dataset titled TrashNet was downloaded as JPEG folders. There were six folders for each of the classifications of recycling (glass, plastic, metal, plastic, cardboard paper, and trash). Each image in all of the folders demonstrates an item that can be recycled and sorted into the aforementioned groups. The unit of observation is "waste." The project is focusing on analyzing images in order to train a model to classify the different types of waste to improve recycling practices.

Dataset (Uncleaned and Cleaned):

The images were generated in a controlled environment where objects were placed on a white poster board under natural sunlight or room lighting and all images were resized to 512 x 384 pixels. Therefore, no data cleaning was required.

Variable Name	Variable Type	Description
Cardboard	Image	512 x 384 pixels
Glass	Image	512 x 384 pixels
Plastic	Image	512 x 384 pixels
Metal	Image	512 x 384 pixels
Paper	Image	512 x 384 pixels
Trash	Image	512 x 384 pixels

II. Variables in Dataset

The TrashNet dataset contains the following variables:

- Cardboard - labeled as "0" in the dataset
- Glass - labeled as "1" in the dataset
- Metal - labeled as "2" in the dataset
- Paper - labeled as "3" in the dataset
- Plastic - labeled as "4" in the dataset
- Trash - labeled as "5" in the dataset

III. Summary Statistics

```
Count of Each Category:
cardboard: 806
glass: 1002
metal: 820
paper: 1188
plastic: 964
trash: 274
```

Figure 1. Summary Statistics of Category Distribution

```
Pixel Intensity (RGB) Statistics:
Red - Mean: 171.59, Std: 19.50
Green - Mean: 163.19, Std: 20.96
Blue - Mean: 154.20, Std: 27.52
```

Figure 2. Summary Statistics of RGB Pixel Intensities

IV. Frequency Tables on Categorical Variables

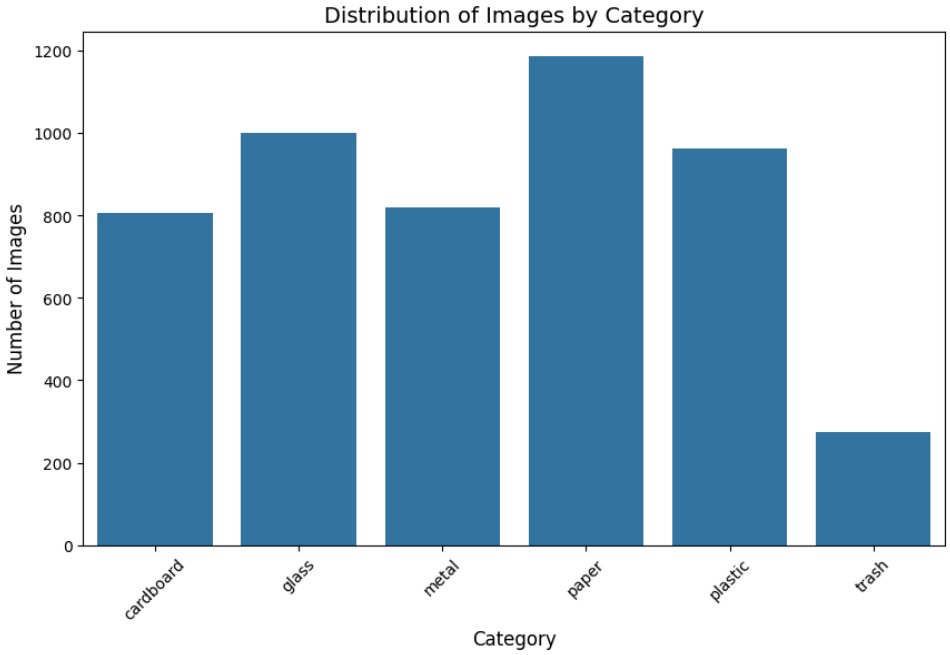

Figure 3. Bar Graph of Images by Category (including training data)

V. Code Scheme

The dataset, sourced from the Hugging Face repository, required no significant cleaning due to its high quality and controlled environment. Images were

uniformly captured on a white poster board under consistent lighting and resized to 512 x 384 pixels, ensuring a standardized format for immediate analysis. No anomalies or missing data allowed us to skip extensive cleaning and focus on preparing the data for modeling.

To begin, preliminary EDA was conducted to understand the dataset's structure and distribution across the six waste categories: cardboard, glass, metal, paper, plastic, and trash. Visualizations of class and RBG distributions provided insights into potential imbalances, helping guide augmentation strategies during preprocessing. Data augmentation techniques, such as random flipping and rotation, were used to enhance generalization and address minor class imbalances. The analysis used a Convolutional Neural Network (CNN), a deep learning model specifically designed for image data. Pre-trained models, EfficientNet and ResNet50, were used to classify images into the six target categories. The CNN extracts hierarchical spatial features (e.g., texture, shape, color) through convolutional, pooling, and fully connected layers, making it ideal for distinguishing between classes with overlapping visual traits. By iteratively training and validating the model over four epochs, performance metrics like accuracy, precision, recall, and F1-score were recorded to refine the model and ensure robust generalization.

VI. Exploratory Data Analysis

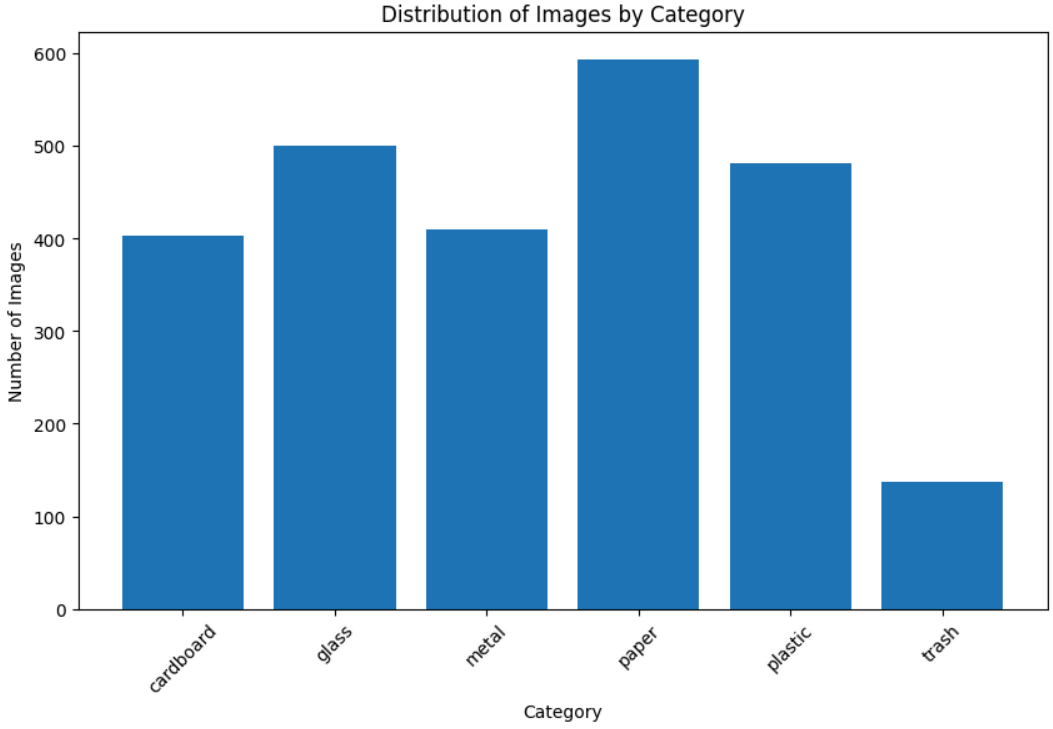

Figure 4. Distribution of Images by Category (excluding training data)

The distribution of images across the different categories (figure 4) shows that paper has the highest number of images, close to 600. Plastic and glass also have a relatively high count, slightly below 600. Metal and cardboard have moderate counts, about 400 images. Trash has the lowest count with fewer than 200 images. This shows an imbalance between categories with significantly fewer images in the trash category which could potentially affect model performance as the model might have less data to learn from for the trash category.

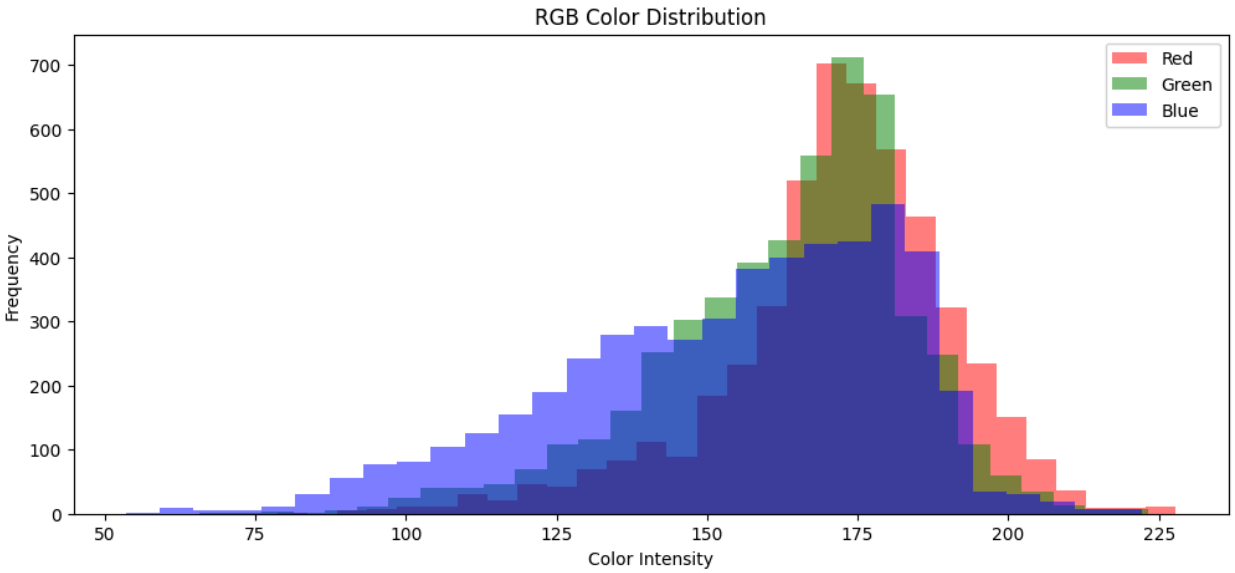

Figure 5. Red, Green, Blue Color Distribution

The RGB color distribution graph (figure 5) shows that the blue generally has higher intensity than red and green, meaning that the images may have cooler tones overall. The peak color intensities for all channels are around 150-175, suggesting the images are relatively balanced in brightness but with a slight blue bias. This tells us that the dataset is pretty consistent for modeling.