Internal Data Wrangling Report: WeRateDogs X Data

Overview

This report summarizes the data wrangling process for the WeRateDogs X dataset, which involved gathering, assessing, cleaning, and merging multiple data sources to create a master dataset.

Data Gathering

Three datasets were collected:

1. Twitter Archive Enhanced: Downloaded directly as a CSV file, containing tweet metadata and dog ratings.
2. Image Predictions: Acquired via a provided URL as a TSV file, containing algorithmic predictions of dog breeds from tweet images.
3. Tweet JSON Data: Due to software and hardware limitations could not be downloaded from the tweepy API so it was Loaded from a local file, containing additional tweet metadata such as favorite and retweet counts.

Data Assessment

Both visual and programmatic assessments were done to identify quality and tidiness issues. Some issues included:

1. Presence of retweets and replies, which are not relevant for original dog ratings.
2. Missing or placeholder values in columns (e.g., dog names, dog stages).
3. Inconsistent data types (e.g., tweet IDs as integers instead of strings).
4. Duplicate rows and duplicate images.
5. Multiple columns representing dog stages instead of a single categorical column.
6. Columns with excessive missing data or irrelevant to the analysis.

Data Cleaning

The following cleaning steps were performed:

1. Removed retweets and replies by filtering out rows with non-null retweet or reply identifiers.
2. Dropped columns related to retweets, replies, and other irrelevant or sparsely populated fields.
3. Removed duplicate rows and duplicate images based on the image URL.
4. Replaced invalid or placeholder dog names (e.g., "None", "a", "an") with a generic value ("dog").
5. Filled missing values in favorite_count and retweet_count with zeros.
6. Merged the doggo, floofer, pupper, and puppo columns into a single dog_type column.
7. Corrected inaccurate data types for columns.

Data Storage

The cleaned datasets were merged into a single master DataFrame using the tweet ID as the key. Redundant columns were dropped after merging. The final master dataset was saved as X_archive_master.csv.

Conclusion

Through wrangling, the WeRateDogs X data was transformed from a collection of messy sources into a unified and tidy dataset. This process addressed key quality and tidiness issues, making insights and visualizations about dog ratings, breeds, and popularity trends on X.