Wrangle Report

Project Overview

This report documents the data wrangling process for the WeRateDogs Twitter dataset. The goal was to gather, assess, clean, and combine multiple data sources into a high-quality, tidy master dataset suitable for analysis. The process followed the standard data wrangling steps: **gathering**, **assessing**, **cleaning**, and **storing**.

Data Gathering

Three main datasets were collected:

1. **Twitter Archive**: Downloaded directly as `twitter_archive_enhanced.csv`, containing tweet-level information and ratings.

2. **Image Predictions**: Downloaded as `image_predictions.tsv`, containing the results of a neural network's predictions on tweet images.

3. **Tweet Metadata**: Queried from the Twitter API and loaded from `tweet_json.txt`, providing additional tweet-level metrics such as retweet and favorite counts.

Data Assessment

Both visual and programmatic assessments were performed to identify quality and tidiness issues:

- **Quality Issues:** included missing values, incorrect data types, invalid dog names, duplicate entries, and inconsistent breed/stage information.

- **Tidiness Issues:** included the need to combine multiple columns representing dog stages into a single column and the need to merge datasets for a unified analysis.

Data Cleaning

Key cleaning steps included:

- **Removing Retweets**: Only original tweets were retained by filtering out rows with non-null retweet identifiers.

- **Handling Duplicates**: Duplicate tweet entries were removed based on unique tweet IDs.

- **Fixing Data Types**: Columns such as timestamps were converted to appropriate datetime formats.

- **Cleaning Dog Names**: Non-name placeholders (e.g., 'a', 'the', 'None') were replaced with null values.

- **Combining Dog Stages**: The four dog stage columns (`doggo`, `floofer`, `pupper`, `puppo`) were consolidated into a single `dog_stage` column.

- **Standardizing Breed Names**: Breed predictions were converted to lowercase and underscores replaced with spaces for consistency.

- **Filtering Predictions**: Only image predictions where at least one prediction was a dog were retained.

- **Selecting Relevant Columns**: Only necessary columns for analysis were kept from each dataset.

Data Merging and Storage

The cleaned datasets were merged into a single master DataFrame using tweet IDs as keys. The final master dataset was saved as `twitter_archive_master.csv` for subsequent analysis.

Summary

The wrangling process resulted in a tidy, high-quality dataset ready for analysis. Key challenges included handling missing and inconsistent data, merging datasets with different structures, and ensuring all transformations preserved the integrity of the original information. The resulting dataset enables robust analysis of dog ratings, breed predictions, and tweet engagement metrics.