Wrangle Report

Introduction

This report outlines the data wrangling process undertaken to prepare the WeRateDogs Twitter d
for analysis. The primary objective was to gather data from multiple sources, assess its quality ar
structure, clean it to ensure consistency and reliability, and merge it into a single dataset suitable
analysis.

Data Gathering

Three datasets were utilized:

1. **Twitter Archive Enhanced**: A CSV file containing basic tweet data such as tweet ID, timesta
   text, and extracted information like dog names and ratings.
2. **Image Predictions**: A TSV file with image prediction data generated by a neural network, in
   the top three predictions for each image and their respective confidence levels.
3. **Tweet JSON Data**: A JSON file containing additional tweet information, notably retweet and
   counts, obtained via the Twitter API.

Data Assessment

Quality Issues

- **Missing Data**: Several columns had missing values, particularly in the `in_reply_to_status_id`, `in_reply_to_user_id`, and `retweeted_status_id`
- **Incorrect Data Types**: Columns such as `timestamp` were not in datetime format, and nu
  fields like `rating_numerator` and `rating_denominator` were stored as integers wit
  considering decimal values.
- **Inaccurate Ratings**: Some tweets had incorrect ratings due to extraction errors, e.g., rating
  1776/10.
- **Invalid Dog Names**: The `name` column contained entries like "a", "an", or "the", which are
  dog names.

Tidiness Issues

- **Multiple Variables in One Column**: The `doggo`, `floofer`, `pupper`, and `puppo` colu
  represented different stages of a dog but were spread across multiple columns instead of a s
  categorical column.