wrangle_report

June 4, 2025

1 Data Wrangling Summary Report

produced by: Abdalrahman Samir **Project:** Wrangling and Analyze Data

1.1 Overview

This report summarizes the data wrangling activities I have done in this project to prepare and clean the following datasets: `twitter-archive-enhanced.csv`, `image-predictions.tsv` and `tweet_json.txt` for use in wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The project involved gathering, assessing, cleaning, and analyzing Twitter data to create a high-quality dataset for analysis. The primary objective was to transform raw, inconsistent, and incomplete data into a structured, clean, and ready for analysis format.

1.2 Data Gathering

Three primary data sources were collected:

1. **Twitter Archive (Direct Download)**

- File: `twitter_archive_enhanced.csv`
- Contains basic tweet information and dog ratings

2. **Image Predictions (Downloaded via Requests)**

- TSV file hosted on Udacity servers
- Contains machine learning predictions about dog breeds in images

3. **Additional Tweet Data (via Twitter API)**

- JSON data containing engagement metrics (retweets, favorites)
- Stored in `tweet_json.txt` and processed into `tweet_data.csv`

1.3 Data Assessment

1.3.1 Quality Issues Identified:

1. Missing values in reply/retweet columns (irrelevant for original content)
2. Tweets without images (2297/2356 had images)
3. Presence of retweets (181 instances)

4. Inaccurate dog names (placeholders like "a", "an", "the", "None")
5. Incorrect data type (tweet_id as integer)
6. Incorrect data type (timestamp as string)
7. Useless retweet_status_id column after filtering
8. Non-dog images in predictions (543/2075 images)

1.3.2 Tidiness Issues Identified:

1. Dog stage information spread across four columns (doggo, floofer, pupper, puppo)
2. Data spread across three separate tables: `twitter_archive_clean`, `image_preds_clean` and `tweets_data_clean`

1.4 Data Cleaning

1.4.1 Quality Issues Addressed:

1. **Dropped irrelevant columns** (reply/retweet metadata)
2. **Removed tweets without images** (kept 2297 entries)
3. **Filtered out retweets** (final 2117 original tweets)
4. **Standardized dog names** (replaced placeholders with "Unknown")
5. **Corrected data type** (Converted tweet_id to string)
6. **Corrected data type** (Converted timestamp to datetime)
7. **Dropped retweet_status_id column**
8. **Filtered non-dog images** (kept 1532 dog predictions)

1.4.2 Tidiness Issues Addressed:

1. **Consolidated dog stages** into single "dog_stage" column
2. **Merged the three datasets** into one master dataset: `df_combined`

1.5 Storage

The final cleaned dataset was saved as `twitter_archive_master.csv` containing: - Original tweet data - Image prediction data - Engagement metrics (from: the `tweet_data.csv`) - Cleaned/standardized columns

Attachments:

- `twitter_archive_master.csv` (cleaned dataset) - Jupyter Notebook with full wrangling process