Data Wrangling Report

We wrangled data from WeRateDogs, a Twitter account with the handle, @dogrates. This Twitter account rates people's dogs with humorous comments about the dogs with most dogs scoring 10/10 or higher because '*they are good dogs*…'

This report documents the steps taken in gathering, assessing, cleaning and storing the data.

GATHERING

In this step, three pieces of data were gathered and represented as pandas DataFrames.

- The first piece of data, provided by Udacity, was a Twitter archive of WeRateDogs in CSV format, and this file was manually downloaded as '*twitter-archive-enhanced.csv*'.
- The second piece of data was a TSV file containing tweet image predictions which was programmatically downloaded from a provided URL as '*image-predictions.tsv'*.
- The last piece of data was a JSON data of the archived tweets which was pulled from Twitter API using tweepy. Each tweet's data was written to its own line and stored as '*tweet_json.txt*'.

ASSESSING & CLEANING

Visual and programmatic assessments were done and copies of data were made. The table below outlines the content and structural issues identified and the cleaning done using the define, code and test process:

Quality Issues

DataFrame	Issue	Cleaning Step
df_counts	Incomplete data	Dropped entries from **df_tweets** that were missing in **df_counts**
df_tweets	Erroneous datatype for '*timestamp*'	Converted datatype to datetime
	Name misspellings	Corrected name misspellings
	Invalid dog names	Changed all invalid names to 'NaN'
	Retweets present in some columns	Remove entries with null values in retweet related columns
	Irrelevant columns	Dropped columns not relevant for our analysis
	Text in '*source'* values	Extracted text from '*source*' URL and converted datatype to category
	Hyperlink in '*text'* values	Removed hyperlinks in '*text'*
	Multiple dog stages	Manually cleaned *'dog_stage'*
df_images	Incomplete data	Dropped entries from **df_tweets** and **df_counts** that were missing in **df_images**
	Invalid tweets with images of other animals	Drop all entries with image(s) of animals that are not dogs
	Percentage values in hundredths	Multiplied all values in '**_conf*' columns by 100
	Lowercase names	Changed first character of all names to uppercase

Tidiness Issues

DataFrame	Issue	Cleaning Step
df_tweets	Dog stages, *'doggo', 'floofer', 'pupper'*, and *'puppo'* in four separate columns	Created a single column '*dog_stage*' containing the different dogstages and converted to category datatype
	Invalid data for '*rating_numerator*' and '*rating_denominator*'	Correct invalid *rating_denominator* values for entries with correct denominator in '*text*' and dropped the rest. Correct invalid '*rating_numerator*' values for only entries greater than 14 within valid score range in '*text*' and dropped the rest. '*rating_numerator*' was renamed to '*rating*'
df_images	*'p1', 'p1_dog', 'p2', 'p2_dog', 'p3', 'p3_dog* in six different columns *'p1_conf', 'p2_conf', 'p3_conf'* in three different columns	Create '*dog_breed*' column for breeds of dog and '*percentage_conf*' for prediction confidence values. Although both issues were addressed under tidiness, multiple prediction columns (*'*_conf'*) was identified as a quality issue.
df_images **df_counts** **df_tweets**	'*Tweet_id*' duplicated Entire dataset should be one table	Tidied both issues by concatenating all tables into **twitter_clean**

Our tidy dataset, **twitter_clean**

```
twitter_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1663 entries, 0 to 1662
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1663 non-null   int64
 1   timestamp           1663 non-null   datetime64[ns]
 2   source              1663 non-null   category
 3   text                1663 non-null   object
 4   rating              1663 non-null   int64
 5   rating_denominator  1663 non-null   int64
 6   name                1585 non-null   object
 7   dog_stage           259 non-null    category
 8   retweet_count       1663 non-null   int64
 9   like_count          1663 non-null   int64
 10  jpg_url             1663 non-null   object
 11  dog_breed           1663 non-null   object
 12  percentage_conf     1663 non-null   float64
dtypes: category(2), datetime64[ns](1), float64(1), int64(5), object(4)
memory usage: 146.8+ KB
```

STORAGE

The cleaned master DataFrame was saved in a CSV file as '*twitter_archive_master.csv*'.