WeRateDogs Data Wrangling Report

Gather

Three sources of data were gathered to complete this project:

1. **twitter enhanced archive file** - csv downloaded from udacity course site, which includes various information about tweets of WeRateDogs account
2. **tweet image predictions file** - tsv requested from udacity site, which includes predictions of objects on images included in WeRateDogs tweets
3. **tweet details file** - json file downloaded from udacity, which includes information missing from the enhanced archive file, namely retweets and likes counts

Data was gathered using different methods:

1. csv file was read from a file and stored in df_archive
2. tsv file was downloaded using the requests library, written to a file and stored in df_img_preds
3. The json file available in udacity was downloaded and stored as a json file name tweet_json and was loaded using json library and finally saved as df_json

Assess

All three files were assessed separately to understand their structure and to find quality and tidiness issues and a few are listed below.

We used both visual (namely looking at the data using sample and head methods and also MS Excel) and programmatic methods (for example with describe, info, value_counts, nunique, mean methods) to assess the data.

Most of the identified issues were from df_archive as it was the main file with the most information which was extracted incorrectly.

Quality Issues

Completeness of data: No issues found

Validity:

- df_archive: Exclude REtweets
- df_archive: Exclude tweets without images
- df_preds: we require ratings of dogs, i.e. where **px_dog** is True

Accuracy:

- df_archive: ratings of more than 10/10 are acceptable, but there were some extreme or decimal values of **rating_numerator**, which were wrongly extracted from the text
- df_archive: ratings can have various denominator, but some **rating_denominator** values are inaccurate based on the text
- df_archive: some names and stages had incorrect values

Consistency:

- df_archive: incorrect data types - timestamp should be datetime, rating columns should be floats to allow for decimal values to use in required calculations.
- df_preds: some dog names start with an uppercase, some with a lowercase letter, some had special characters in it

Tidiness issues

- df_archive: columns doggo, floofer, pupper, puppo are actually values of a single column **dog_stage**
- df_archive: **rating** would make more sense in its own column by combining the numerator and denominator columns
- df_preds: the columns relevant for analysis can be merged with the main df_archive file.
- df_preds: unwanted columns can be deleted.
- df_likes: information should be included in the main archive table, there is no reason to have it in a separate data frame

Clean

We cleaned a total of 12 issues and are listed below.

- **tidiness**: merging all files together, creating a dog stage column, creating a rating column
- **quality**: deleting retweets, deleting tweets without an image, deleting tweets with an image not of a dog, cleaning dog stage column, cleaning rating columns, changing datatypes, deleting columns, cleaning name column, making all predicted breed names lowercase

Each issue was handled separately using the define-code-test workflow and the cleaned dataframe was exported as twitter_archive_master.csv.