Wrangling Report

This report briefly summarizes the identified quality and tidiness issued identified in the three files and the actions taken in the data cleaning process. The Jupyter notebook wrangle_act.ipynb provides more details.

Quality Issues

The following quality issues were identified and fixed in the wrangling process

Twitter Archive

- Column **tweet_id** is not of type String → changed to String
- Column **timestamp** is not of type DateTime object → changed to DateTime object
- Some values in column **rating_denominator** are not 10 (23 rows) → rows deleted
- Ratings higher or equal to 26 are either due to wrong formatting or are outlier ratings in column **rating_numerator** (9 rows) → rows deleted
- Erroneous entries in **name** column, such as ‘a’, ‘an’, ‘such’ and ‘quiet’. Since all visibly identified erroneous entries start with a lower case letter, this was used to identify erroneous entries and store them in an array (archive_name_none) → Erroneous entried changed to None
- Not all columns are relevant for later analysis → columns dropped
- Note: Since out of the 2,356 available data records only 394 records have information on the type of dog this information will not be considered in the further analysis. If it were to be considered, one possible cleaning action would be to combine columns **doggo, floofer, pupper, and puppo** into one 'dogtype' column.

Tweet Image Prediction

- Column **tweet_id** is not of type String → changed to String
- Some dog breeds might be stored with a lower case first letter → all entries capitalized
- Column names are not informative → use better column names
- Not all columns are relevant for later analysis → columns dropped

Tweet Retweet Count

- Column **tweet_id** is not of type String → changed to String

Tidiness Issues

The following tidiness issues were identified and fixed in the wrangling process

- Most confident prediction for dog (and breed) is stored in multiple columns (**Tweet Image Prediction** columns **p1, p2, p3, p4**); for img_num is 4, p3 should be used instead. → Rename existing columns **px_dog** to 'isdog', **px** to'dogbreed', and **px_conf** to 'p_value' to store most confident prediction for dog and dog breed
- One single dataframe is created by merging the three dataframes.