Reporting: wragle_report

Data wrangling of the various datasets

Celaning of the Enhanced archive dataset

The ***twitter-archive-enhanced.csv*** dataset had 2356 tweets with 181 of these being retweets. Since the main objective of this project was to analyse data on original tweets, all such retweets were deleted. First, the rows with retweets and replies were delete. Then, subsequently, the ***retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp*,*in_reply_to_status_id*** and ***in_reply_to_user_id*** columns were also deleted from the dataframe.

Similarly, there are 59 tweets with missing expanded_urls and possibly no dog images. Of these tweets,56 are replies and/or retweets. These tweets were deleted since they are not originial tweets, and also since there are possibly no dog images.

The ***timestamp*** column is a string datatype. This was formated into a datetime datatype as it is just right that a date object should be a datetime datatype. This will aslo help us do a better analysis of the tweet dates if we so wish.

A visual inspection of the *source* column shows that there are 4 distinct display types in each of the texts. These display types were extracted from the text to get a categorical set of display types.

The ***rating_numerator*** has some very large values which is unusal for a rating for a single dog. Further inspections showed that those values were ratings for multiple dogs. About 25 rating_numerator values are greater than or equal to 20. A rating of 20 was chosen arbitrarily. These very large rating values were dropped since they were tweets/ratings of multiple dogs.

The ***rating_denominator*** also has some values greater than 10, which is the offical value. 23 rating_denominator values are greater than or less than 10 (20 are greater than 10 and 3 are less than 10). It's likely tweets with rating_denominator not 10 have multiple dogs in them so these tweets were dropped

The ***name column*** had invalid data for dog names (a, the, an). Futher inspection of the column showed that all valid dog names start with uppercase. To clean the data, all invalid dog names were replaced with *None*

Cleaning of the Image predictions dataset

The ***p1_dog***, ***p2_dog*** and ***p3_dog*** columns had some wrong predictions for dogs. Since this project is about analysing data about dogs, all non-dog predictions were dropped. In all, 324 instances of such predictions were dropped