General thoughts before starting analysis:
The analysis center around the dogs. There is no way to unambiguously identify a dog by name or a dog ID, so I am assuming there is only one (original) tweet per dog. Owners don't tweet twice about their dog, so we would use the tweet_id to recognize the dog.
Quality issue 1: Dog names column contain errors. I have decided to not fix this issue as there is no meaningful replacement for a dog name
Quality issue 2: The dogtionary mentioned the following type of dogs: doggo, pupper, puppo, floof The provided file contains the column names: doggo, floofer, pupper, and puppo. Column name floofer was replaced with floof.
Quality issue 3 and 4: Not all images uploaded are for dogs. This has three different reasons, quality issue 3 is about the first two

- User uploaded drawings or screenshots of dogs
- Users uploaded animals that are not dogs (https://pbs.twimg.com/media/CT-jNYqW4AAPi2M.jpg)
- Image recognition failed

I am making the assumption, that we are only interested in dogs. As we are interested in dogs and dogs only: Users violated the schema when uploading non-dogs. Which makes this a validity issue.
The third reason for non-dogs is that the image recognition failed, a dog is present but the image recognition identifies a more prominent object. Example: A dog was identified as seatbelt https://pbs.twimg.com/media/CUDeDoWUYAAD-EM.jpg This is an accuracy problem, a dog is pictured, but not recognized correctly.
This can be solved by removing all non-dogs from the data frame. I have decided on the most strict approach, all three recognitions need to recognize a dog.
Quality issue 5 Image_pred has a column for img_num. Seems not to be used and will be removed.
Quality issue 6 I am making the assumption, that 'in reply to' means that this tweet is a retweet. As it was stated that we are only interested in original tweets. Need to remove retweets
Quality issue 7: According to twitter_arch there 181 retweeted tweets, according to tweet_json_df_sv there are 2341 retweeted tweets. I trust the data from the Twitter API more, will not use "retweeted_tweets" from the Twitter_archive of weratedogs.
Quality issue 8: Several tweet IDs threw an error when querying from the API. Only use tweet_ids that were returned from API
Tidiness issue 1 In twitter_arch the dog categories (doggo, floofer, pupper and puppo) should be in one column, e.g named "dog category".
Tidiness issue 2 In twitter_arch, some Twitter URLs are doubled.
Tidiness issue 3: twitter_arch and tweet_json should be one dataframe.