Wrangle Report

# Gathering data

We were provided with WeRateDogs Twitter archive with 2356 rows and 17 columns. I programmatically downloaded data containing 2075 of original archived tweets’ image predictions. I also used additional data, received using Twitter API, for 1177 of original archived tweets. All three datasets I loaded into Pandas DataFrames **tweets**, **img\_predictions** and **tweets\_info** correspondently.

# Assessing data

First, I performed *visual assessment* of this data, which helped me to find a few quality and tidiness issues.

In original archive of tweets some dog names got extracted as "a", "an" and "the".

Also, there are 4 columns, **doggo, floofer, pupper** and **puppo**, representing one variable, dog "stage.

Nulls are represented as string "None" for columns **name, doggo, floofer, pupper** and **puppo**.

In image predictions dataset I noticed inconsistency in dog bread names. Some of them start with capital letter and some contain "\_".

I didn’t find any issues in additional tweeter data after visual assessment.

After seeing all three datasets I noticed that they all represent tweet’s properties, so they should not be in separate datasets.

Also, I noticed that there are missing records (1177 in image predictions dataset vs 2356 in original archive of tweets). Maybe some tweets got deleted from Tweeter server. We can’t clean that data.

Then, I performed *programmatical assessment*. I found more data quality issues.

There are 181 retweet related records. We are not interested in retweets, so we don't need records with not empty **retweeted\_status\_id**. Also, we don't need columns with information about retweet: **retweeted\_status\_id, retweeted\_status\_user\_id, retweeted\_status\_timestamp**.

Column **timestamp** has type "string" instead of "datetime".

Min value for **rating\_denominator** equals 0. That doesn't seem right.

During visual assessment I noticed, that tweet with **tweet\_id** = 666287406224695296 has rating 1/2. That doesn't look right. I checked the tweet text and noticed, that it contains 2 pieces matching pattern “<number>/<number>” and first was picked as rating. Although, it seems, that the last piece more likely represents rating.

# Cleaning data

First, I created copies of DataFrames.

I started with fixing tidiness issues.

I combine values from columns **doggo**, **floofer**, **pupper** and **puppo** in one new column **stage** and then droped the old ones.

Then I merge all 3 tables into one.

I removed all information related to retweets.

I converted column **timestamp** to datetime type.

All strings "None" were replaced with np.nan.

I replaced strings "a", "an" and "the" in dog’s name column with np.nan.

I reextracted rating from **text** column as last piece, matching pattern “<number>/<number>”, and refilled **rating\_numerator** and **rating\_denominator** columns with new values.

I transformed dog breads to lower case and replaced "\_" with space.