“WeRateDogs” Wrangling Report

Introduction

The main data set that we are going to wrangle it contains an archived tweeter data for 2356 tweet in the period from November, 2015 to August, 2017. The twitter account “WeRateDogs” rates people's dogs. On the other hand, additional data set represents 3 different algorithms for dog image predication.

Gathering process

1. Gather process started with provided dataframe of ‘twitter-archive-enhanced.csv’ which is imported as (twitter_archive_df)
2. The second piece of data is collected by requesting tap separated file of image predication through link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archiveenhanced/twitter-archive-enhanced.csv. Then it saved as ‘image_predic_df’
3. The final chunk of data is gathered from twitter API and stored as ‘tweet_json.txt’ which has tweet_id, retweet_count and favorite_count.

Assessing process

In this stage of data wrangling, I explored the 3 aforementioned dataframes visually and programmatically. The meaning of each column in each data frame is defined. Of course df.info(), and df.series.value_count() are commonly used to identify missing values, erroneous data type, .. etc. Accordingly, I able to get *8 quality issues and 2* tidiness issues which as stated below:

Quality issues

a) `source` has html tages which is not neccessary and convert its data type to categoty.
b) `timestamp`, and `retweeted_status_timestamp` should be as datetime not object (data type error)
c) `in_reply_to_status_id` and `in_reply_to_user_id` should be integer not float (data type error)
d) keep original tweets and remove reweeted ones.
e) based on point above, remove columns of `retweeted_status_id, retweeted_status_user_id` and `retweeted_status_timestamp`
f) satisfy consistancy between `twitter_archive_df` and `image_predic_df`
g) dog `names` with lowercase and single character such as *'a'*
h) very high `rating_numerator` e.g. 1776 with respsect to `rating_denominator`==> scaling both

Tidness issues

a) `doggo, floofer, pupper`, and `puppo` should be mergerd in one column
b) table of `tweets_df` should be added to `twitter_archive_df`

cleaning stage:

The quality and tidiness issues have been encountered in ‘twitter_archive_clean’ which is just a copy of `twitter_archive_df`. While executing evert issue, 3 steps of define, code and test have been followed for organization and reproducibility.