Wrangle report

Wrangling data is an important skill for every data analyst, as part of Data Analyst Nanodegree program we are tasked in this project to work on WeRateDogs tweets archive contains basic tweet data for all 5000+ of their tweets, but to complete the other pieces of data we are tasked to gather extra data using different methods. This is a short report describing wrangling steps carried out in this project.

Gathering data

Our task in this step was to obtain data from three different ways:

1. Downloading manually ***twitter_archive_enhanced.csv*** and upload it in jupyter Notebook
2. Using ***Requests*** library to download programmatically ***image_predictions.tsv*** , this file is hosted on Udacity's servers and it's a dogs breeds predictions using tweets images and neural network.
3. Using ***Tweepy*** library we are tasked to query Twitter's API to obtain a data in json format for each tweet ID, then using lists and dictionary to extract favorites and retweets counts, for finely append it in new data frame.

Assessing data

As part of wrangling process the assessing data is carried out to detect tidiness and quality issues, to achieve this task we used two different techniques:

1. Visual assessment: detecting issues using visual techniques like sampling…
2. programmatic assessment: detecting advanced quality and tidiness issues

Detected quality and tidiness issues in 3 datasets:

Dateframe	Issue	issue type
df_arch	Columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id are floats insted of int	Quality
df_arch	Data in source column are not clear and unreadable	Quality
df_arch	timestamp is object insted of datetime	Quality
df_arch	missing data in in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id columns	Quality
df_arch	invalid rating in rating_numerator and rating_denominator columns	Quality
df_arch	'None' used to indicate a missing value instead of 'NaN' in name column	Quality
df_arch	invalid 'names' in name columns e.g. (55 names = 'a')	Quality

df_arch	source column are 'object' instead of 'category' data type (there is only 3 repeated sources)	Quality
df_img	- p1, p2 and p3 columns must be 'Category' instead of 'object' data type	Quality
df_img	- some of images are duplicated in jpg_url column	Quality
df_arch	- Columns doggo, floofer, pupper and puppo represente the dog stage it must be grouped in one column (dog_stage)	Tideness
df_img	- All columns from p1_dog to p3_conf are used to predict the dog breeds instead of one column	Tideness
tweet_df	- Rename status_id column to tweet_id to be able to join data frames	Tideness
All	- All tables must be merged to one table	Tideness

Cleaning data

Cleaning data is the last part of wrangling process, and for each defined cleaning action we have to:

1) Define: specify the action in words
2) Code: wright the code executing defined cleaning action
3) Test: testing the efficiency of executed code

Hereafter the most important achieved cleaning actions:

- Group columns `doggo`, `floofer`, `pupper` and `puppo` to one column called `stage`, then drop variable column and duplicates generated from this process.
- Use all columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_confand p3_dog to predict the dog breeds in one column called breeds.
- Group all data in one unique data frame
- Extract a meaningful word from `source` column and make it readable
- Reduce the number of invalid dog names using regular expression with pattern
- Detect and fix rating numerator and denominator using regex to extract rating from text