Wrangle Report

In this project we wanted to Wrangle and Analyze data from the twitter's account WeRateDogs (@dog_rates). WeRateDogs account received international media attention for its singular way to rate other people dog's photos and because it was suspended for Twitter for breaking social media copyright law. The first step was to gather data from Udacity and Twitter API. Then I cleaned and analyzed it using Python libraries.

Gathering data

Data for this project was obtained from three ways:

- The main data was provided by Udacity and downloaded manually.
- Other data as Retweets and favorite counts was obtained from Twitter's API using Tweepy.
- Twitter's image prediction file was downloaded programmatically using Request from Udacity's server.

Assessing Data

Assessing data was performed using the following methods:

.info()
.head()
.unique()
.value_counts

The following data quality and tidiness were reported:

- twitter_id is an int;
- missing a lot of data from columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp';
- name_values has 745 None as name and some wrong names like 'his', 'my', 'an', 'a';
- there are retweets that we would not like to count;
- timestamp is an object;
- it is hard to do an overview because there are too many and unnecessary columns;
- some dogs seems to have a wrong rating_numerator because in the text it was a float.
- p1, p2 and p3 have underscore where should be a space;

- missing some data.
- the columns doggo, floofer, pupper, puppo should be variables in the same column;
- the three datasets should be just one.

Cleaning and testing

Issues found in assessing phase were in new tables that was a copy from original and then fixed as follows. First I merge the three data sets together. Then a column stage was created for store the values doggo, floofer, pupper and puppo and these columns were dropped.

The type of tweet_id was fixed with function .astype() and timestamp need to be lost it times zone and then I could use pd.to_datetime(). Retweets were ignored unising np.isnan() in retweeted_status_id and unnecessary columns were dropped with .drop().

Column with dog’s names has two problems: first there was a name “None” instead of NaN and some wrong names as ‘a’, ‘such’ and so on. I fixed these two problems with .map() and .replace().

Some values in rating columns were float, so it was wrong in column. To fix that, first I changed the two rating columns with .astype() and then search in the dataframe for floats values that were wrong and fixed it.

At least I change the underscores in p1, p2, and p3 columns for space.