Wrangle Reports

Our goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. We need to gather, then assessing and cleaning data from various sources.

GATHER

First I open a twitter archives .csv file and convert to a dataframe.

We need some information to complement the twitter archives, like the favorite and retweets of each tweet id. So, we connect with the API from twitter throw the package tweepy. Then we can extract the data and convert to a JSON file and then make it a dataframe.

Al last, we have a predictive model that has the breed of the dogs throw the images that people post. We need to open it in a URL and extract a image_predictions.tsv file.

ASSESS

With visual and programming assessment I found some quality a tidiness issues than need to be corrected to make analysis correct.

Quality:

1) tweet_id is a integer, it should be a object, we don´t want to do calculations with them

2) Variables with a lot of NaN, in_reply_to_status_id,in_reply_to_user_id

3) Some tweets (59) don´t have images attach

4) Rating denominator equal to 0

5) Scrapping errors in Rating numerators

6) Variable names has NaN values represented by the word 'None'

7) In df_img, tweet_id is a integer and it should be a object to use it as a principal key and we don´t want to do calculations with them

8) Some predictions are not dog breeds

9) In Tweet_counts, tweet_id is a integer and it should be a object to use it as a principal key and we don´t want to do calculations with them.

10) In the retweeted_status_id, we don't need when it has data. Only use the NaN

Tidiness:

1) Doggo, floofer, pupper, puppo should be one variable, not headers

2) Retweet and favorite should be appended to twitter-archive-enhanced Dataframe(df) and breed image prediction to.

CLEANING

With this issues addressed, we now clean each one at the time, describing the action, doing the code and then testing that the job it's done.