Wrangling the WeRateDogs twitter account

In this project I have gathered, assessed, cleaned, stored, analyzed and visualized data from WeRateDogs (https://twitter.com/dog_rates), a well known twitter account where images of dogs and sometimes of other animals (or even of objects) are posted usually with humorous comments and a score given to the dogs.
The scoring system on WeRateDogs is ideally in tenths, but scores exceeding 10/10 are very often given to dogs and considered valid.

In the **Gathering** section of the project I have gathered data from three sources:

1) A csv file provided by Udacity with the archive of tweets from WeRateDogs, containing data such as the id associated with every tweet, the timestamp, the score given to the dog and sometimes the names of the dogs and a stage classification: doggo for adult dogs, puppo for very young dogs, pupper for young not yet adult dogs and floofer for furry dogs.
   From this file I created a twitter_archive dataframe.
2) A tsv file gathered from an url, containing the prediction of a neural network about the content of an image associated with a tweet, in particular the neural network tried to guess if the image was really of a dog or not.
   From this file I created an image_predictions dataframe.
3) Complementary data about the number of retweets and favorites for every tweet has been gathered with the Tweepy API from Twitter, using the tweet ids present in the archive of tweets.
   From this file I created a tweet_extra dataframe.

In the **Assessing** phase of the project I searched for possible quality problems in the dataframes created from the sources of data, like mismatches in data types for some columns such as dates represented as strings instead of datetime objects, or null values represented as strings. I maintained the possibility for the score to be greater than ten, but I also normalized the scores that have been expressed in other formats (like 44/40).

I also tried to identify tidiness problems: like a four columns for describing one variable, the dog stage, instead of one column. After normalizing the score I dropped the column for the denominator, in order to have again one column for one variable instead of two.

In the **Cleaning** phase of the project I created copies of the original dataframes and cleaned one by one every quality and tidiness problems, proceeding first defining the

procedure (Define), then coding (Code) and finally testing that every change has been implemented correctly (Test).

After the cleaning phase I merged the useful data from the three dataframe in one main dataframe and **saved it** to a csv file with the filename twitter_archive_master.csv.

I then proceeded to **analyze and visualize** data from cleaned dataframe.

The complete data wrangling phases can be found in the wrangle_act.ipynb file, while a brief summary of the analyzing and visualizing phases can be found in the act_report.pdf file.