Report

Data wrangling of
datasets from "WeRateDogs" tweets.

Data wrangling procedure has been completed on the datasets obtained from "WeRateDogs". There have been 3 main stages:

1. Data gathering;
2. Data assessment;
3. Data cleaning.

Data have been gathered from different resources such as a provided file, a downloaded file via the internet using the provided link and via tweepy API. As a result 3 datasets (tw_arc, prediction, tweets_api) have been collected for further assessment and cleaning.
Data assessment was conducted against Quality and Tidness issues. Standard python methods and functions were used for data assessment such as .head(), .value_countes(), .sample(), .describe(), .info() and etc. The following problems were identified:

Quality issues:

tw_arc table

- column source in tw_arc is too long for such source information
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp variables have a lot of missing data and, moreover, we do not need them for the analysis
- rating_denominator is 10 in 2333 cases out 2356 cases, consider rating_denominator to be 10
- rating_numerator in most cases is in between 0 and 15, the rest consider as outliers
- variable name has 745 None values and 55 "a" values
- in timestamp +0000 is redundent information
- name variable has some entries starting with low case letters(example: An). Is An a dog name? It occurs 7 time in the dataset, though.
- expanded_urls contains link which are not valid, possible because they are expired
- remove rows in retweet_count and favorite_count with missing values

prediction table

- variable img_num is not needed
- change variable types to the appropriate, where it is needed
- prediction table is missing one importent varaible which would show if the picture truly contains breed of dog or not like it is shown above. one picture has straus on it and algotithm classified it as it is not a breed of dog, but another picture has a dog on it; however, it was misclassified as not breed of dog

Tidiness issues

tw_arc table

- Variables doggo, floofer, pupper and puppo in one column

- tables wt_arc and tweets_api form one observational unit

prediction table

- jpg_url variable should be in tw_arc table to satisfy tidiness definition
- tw_arc and prediction tables form two different observations units and will be kept separately

One important issue came up during the assessment of the prediction dataset. This is a lack of the booling variable which confirm that the pictures either has or has not a breed of dog. As it was shown manually that a picture of straus was correctly classified as False meaning that it is not a breed of dog(p1_dog=p2_dog=p3_dog=False), but there is no indication that the picture is not a breed of a dog. On the other hand, we have a breed of dog, which was two times misclassified p1_dog=p3_dog=False and one time correctly classified as a breed of dog p2_dog=True. Again, without manual confirmation that this is a breed of dog or not we cannot assess the correctness of the algorithm.

We also discovered that source variable takes only for 4 values which can be changed to more readable values such as 'Twitter for iPhone', 'Vine', 'Twitter Web Client' and 'TweetDeck'. Also, one can note that expended_urls variable has links which are broken; hence, this variable was dropped. The timestamp was cleaned by removing “+0000”. In addition, to satisfy tideness definition tw_arc and tweets_api datasets were merge together.

To satisfy the tidiness definition jpg_url variable was moved from prediction table to tw_arc table and after that prediction table is ignored due to the fact which was described above.

During the cleaning stage for each step cleaning procedure was documented as “Define”, code was developed and tested.

Because data wrangling is an iterative process, some outliers were found and cleaned during the Analyzing and Visualizing Data stage, when histograms and scatter plots were drawn.

Finally, the dataset was stored as .csv file: twitter_archive_master.csv.