Gathering data

The first stage of data wrangling process is gather the data. Gather the data can be simple, like click in a csv file to download, or complex, when querying an API to extract data. Both types of tasks were used during the construction of wrangle_act.ipynb.

Initially, we downloaded manually the twitter-archive-enhanced.csv to access the twitter archive, saving it in the local folder. We were able to open the file easily and renaming the file as df_archive.

The second file, the tweet image predictions was made available in a specific url**.** Through the command requests.get (url), we reach the file image-predictions.tsv and downloaded programmatically, saving the file in a local folder, renaming it as df_image.

The third and last file was extracted using tweepy to query the Twitter API, using the tweet_ids stored in df_archive in the search. That way, we get the data for each tweet in JSON format, saving the file as tweet_json.txt. This file contains the count of favorites and retweets for each tweet that we will analyze from the WeRateDogs profile. The JSON file was renamed to df_tweet.

Assessing data

df_archive, df_image and df_tweet were opened and the search for quality problems and tidiness problems started. Using pandas commands info, duplicated, isnull, head, tail we arrive in the following list that will guide us in correcting in data cleaning process.

Quality problems:

- Missing data. Some columns refer to the reply are incomplete and are not needed;
- Missing data. Some columns refer to the are incomplete and are not needed;
- Missing data. The expanded_urls column is incomplete and is not needed;
- Data type incorrect in timestamp. Should be datetime;
- The numerator and denominator values must be refined;
- Missing data. 2356 tweets in df_archive, but only 2075 tweets in df_image => 281 missing;
- The breeds of dogs names in p1, p2 and p3 columns are not standardized. All names should be lowercase;
- Duplicated data. Some tweets have the jpg_url duplicated;
- Missing data. 2356 tweets in df_archive, but only 2331 tweets in df_tweet => 25 missing.

Tidiness problems:

- Structure issues. Doggo, floofer, pupper, puppo columns should be one column (one variable in four columns);
- Structure issues. Join the p1, p2, p3 columns into a new one and p1_conf, p2_conf, p3_conf columns into another. These structural changes will facilitate the analysis at the end;
- Structure issues. df_archive, df_image and df_tweet should be merged. The three datasets contain information from the same tweets list, and it makes no sense for them to be separated.

Cleaning data

We made a copy of the datasets in order to clean and change the structure. Some columns in df_archive would not be used in the analysis and could be dropped: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls, source.

The timestamp column has the data type changed from string to datetime.

Using regular expression, we identified the rating_numerator that are decimal numbers. We extracted the rating and splitted it into rating_numerator and rating_denominator.

We replaced the None value found in doggo, floofer, pupper, puppo columns to blank, concatenated all 4 columns to 1 column dog_stage, updated multiple dog_stages and, finally, dropped doggo, floofer, pupper, puppo columns.

In df_image the columns with breeds dogs names were standardized when placed in lower case. Some tweets have the jpg_url duplicated, so we dropped them, keeping the last one. We joined the p1, p2, p3 columns into a new one and p1_conf, p2_conf, p3_conf columns into another.

Following what was instructed in Project Motivation, the numerator and the denominator are not 100% correct and need to be revised, in order to be used in analysis. Using pandas command with pd.option_context("max_colwidth", 200), we increased the width of the text column displayed to be able to read the entire tweet in Jupyter Notebook. Thus, we found that some ratings in text column were decimal and didn’t match with the rating stored in rating_numerator column, some rating_denominator different than 10 were extracted wrongly, some tweets were indeed retweets and we didn't want analyze them. Using regular expression, we identified the rating_numerator that are decimal numbers. We extracted the rating and splitted it into rating_numerator and rating_denominator. We corrected the rating_denominator, deleted the tweets without a rate, deleted too the RT tweets.

Finally, we merged the three tables in one - df_twitter, ending the clean data process.

Storing data

We stored the final dataframe df_twitter as a comma-separated values (csv) file named twitter_archive_master.csv.