Wrangle Report

This document briefly describes the efforts made to wrangle the information gathered from twitter posts made by the account “WeRateDogs”.

Gathering

First, data was gathered from 3 different sources:

- Post record from the account in a .csv file (*df_twitter_archive*)
- Retweets and favs record from a json file. (*df_retweets_favs*)
- Dog breed prediction as a result from a machine learning algorithm as a .tsv file. (*df_dog_breeds*)

After gathering, an assessment of the data was made, to find any possible corrections. The problems found are all recorded in the assessment section.

Assessment

Tidiness Issues

- There are three different datasources, but only two objects of analysis: Dogs and Tweets.
- *created_at* and *timestamp* columns refer to the same event: posting date and time
- *text* and *full_text* columns contain the same information.

Quality issues

df_twitter_archive

- Column name has Names such as "a,an,the" which are not dog names.
- Some tweet had no name included or the algorithm failed to find the name, "None" is not valid.
- timestamp column has object type instead of date or datetime.

df_retweets_favs

- Some columns have no info (*contributors, coordinates,geo*)
- Some columns have not enough info to get any conclusions from them (*in_reply_to_user_id, in_reply_to_status_id*, etc.)

df_dog_breeds

- Some classifications are not dog breeds
- Columns with classifications are not consistent in the usage of upper and lower caps
- *created_at* column has object type instead of date or datetime.

Finally, most of the problems were addressed, and a few remained untouched, due to the high amount of effort needed to solve it, or the possible loss of information if the rows were deleted.

Cleaning

Tidiness issues:

-The two tables containing dog's information were merged using the *tweet_id* as key. A new table named *df_dogs* was created.

-*timestamp* column was deleted

-Post text was kept only in the table with tweet information.

Quality Issues:

-In *df_dogs*, rows containing wrong names were deleted, although rows that failed to recognize the dog's name were not. Deleting rows with "None" as a name implied deleting one third of the database.

-*timestamp* column was corrected using the function strptime to modify it from string to datetime

-A new subset of the *df_retweets_favs* dataframe was created containing only the columns with sufficient information for an analysis.

-Tweets containing photos no categorized as dogs were not deleted. This was not possible due to the amount of posts containing photos not recognized as dog breeds.

-Categories using lower cases were fixed using the title function.

-Column *created_at* was deleted (duplicated).

The original and clean tables were kept, preserving all sources of information. All of them are present in the same file of this document.

Originals: *df_twitter_archive_original.csv, df_retweets_favs.csv, df_dog_breeds_original.csv*

Clean files: *df_retweets_favs_mod.csv, df_dogs.csv*