Wrangle Report WeRateDogs

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The goal of this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Data

The data for this project consisted of three pieces:

1. The WeRateDogs Twitter archive with the contents and metadata for the WeRateDogs tweets (called “archive”)
2. Tweet image predictions: what breed of dog is present in each tweet according to a neural network (called “images”)
3. The tweet's retweet count and favorite count, acquired by querying the Twitter API (called “rtfav”)

Wrangling: issues covered

Please note that the list below is not exhaustive.

<u>Quality</u>

- Archive: around 180 of the tweets in the archive table were retweets. Requirement was to not include retweets, so these tweets were removed.
- Archive: around 125 of the tweets in the archive table did not contain an image. Requirement was to not include tweets without images, so these tweets were removed.
- Archive: the rating denominator was not always 10, due to the fact that some strings like "24/7" and "9/11" were used as a rating, and tweets about more than one dog had a higher denominator. To be able to use rating in analysis, all denominators were set to 10.
- Archive: there were missing values and incorrect values for dog names. New code to extract dog names was run, resulting in slightly less missing values. The incorrect names were replaced by either the right name or by “None”.
- Archive: there were many missing values for dog stages. New code to extract dog stages was run, resulting in a few more values for dog states.
- Archive: timestamp was a string, it was converted to a datetime object and can now be used for analysis.

- Archive: columns that will not be used for analysis were dropped ('in_reply_to_status_id', 'in_reply_to_status_id', 'retweeted_status_id', 'retweeted_status_user_id' and'retweeted_status_timestamp']
- Images: the mean confidence levels for p2 and p3 were very low (13% and 6% respectively), which means the algorithm is not very confident in these predictions. These columns (including the names and confidence levels) were removed.
- Images: images cannot be used if p1_dog is False. All rows where p1_dog was False were removed.
- Images: dog breed names in column p1 were sometimes capitalized, sometimes lower case only. All breed names were set to lowercase.
- Images: there are tweet id's missing as compared to the archive table (the images table has 2075 rows, the archive table 2359). This is probably because no prediction was possible for these tweet id's. Therefore, this issue was not handled during cleaning.
- Images: columns that will not be used for analysis were dropped ("img_num", "p2", "p2_conf", "p2_dog", "p3", "p3_conf" and "p3_dog")
- The table with retweet_count and favorites count had 14 missing rows compared to the archive table, which means that these tweets were deleted in the meantime.

Tidiness

- Archive: each dog stage had its own column (rule: every variable -in this case: dog stage- forms a column). One column "stage" was created.
- Archive: this table holds data on tweets as well as on dogs (rule: each type of observational unit forms a table). This table was split in a tweets table and a dog table. Most of the columns of the images table were added to the dog table.
- The retweets/favorites table was added to the tweets table.