WeRateDogs – data wrangling

0. Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10, while numerator is almost always greater than 10 (11/10, 12/10, 13/10, etc.) Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

In this project gathering, assessing, cleaning and analysing of data was required.

1. Gathering

Data was gathered from different sources in a different ways:

- Downloading file manually from the following link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv
- Downloading file programmatically using the 'Requests' library from the following link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Each tweet's retweet count and favourite ("like") count was extracted from the Twitter API for each tweet's JSON data using Python's Tweepy library and saved as JSON data file in tweet_json.txt file.

2. Assessing

After assessing (visually and programmatically) the following issues with data was detected:

- Quality issues:
  Archive_df
  - 'timestamp' shold be datetime format;
  - 'in_reply_to_status_id' and 'in_reply_to_user_id' should be int-type;
  - 'text' is not fully displayed;
  - dataframe includes retweets;
  - 'rating_numerator' has huge numbers, e.g. 1776, 960 as well as 0s;
  - 'rating_denominator' has huge numbers, e.g. 170, 150 as well as values less than 10;
  - not all tweets has dog name;
  - dog names has strange names, such as 'a', 'an';
  - many missing dog types in columns 'doggo', 'floffer', 'pupper', 'puppo';
  - some tweets has multiple dog type.

  Image_df
  - strange dog types in col 'p1-3', such as minibus, boathouse, pickup, can opener
- Tidiness issues:
  - retweets should be deleted and then columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' can be deleted as well;
  - 'text' column in archive_df can be split into gender, dog_stage, link and score;
  - 'doggo', 'floofer', 'pupper' and 'puppo' columns in one column 'dog_stage';
  - add columns 'retweeted_count', 'favorite_count' from json_df to archive_df;
  - all dataframes have different number of rows;
  - add 'breed' column from image_df to archive_df;

There are more issues with the data, but for current project and analysis this list of issues is sufficient.

3. Cleaning

The following cleaning was performed:

- Quality issues:
  - 'timestamp' was changed be datetime format;
  - 'in_reply_to_status_id' and 'in_reply_to_user_id' was changed be int-type;
  - 'text' was fully displayed;
  - retweets were removed;
  - strange dog names, such as 'very', 'unacceptable', were removed;

    - based on dog breed data (p1-p3, p1_conf-p3_conf, p1_dog-p3_dog) non dog tweets were removed and new column containing only the most probable dog breed was generated;
    - ‘rating_numerator’ and 'rating_denominator' was fixed by extracting this values from ‘text’ column;

- Tidiness issues:
    - columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' was removed;
    - dog gender, dog stage, link and score was extracted from ‘text’ column;
    - new column column with dog stage was generated from the four columns 'doggo', 'floofer', 'pupper' and 'puppo’;
    - three dataframes was merged in one master dataframe;

Master dataframe was saved as twitter_archive_master.csv file.