Wrangle report

July 4, 2019

The WeRateDogs Twitter data project consists of the following parts:

1 Data gathering

The following data have been gathered:

- The WeRateDogs Twitter archive was read from *twitter_archive_enhanced.csv* file and stored in **tweet_data**
- The tweet image predictions were downloaded programmatically from the Internet using Requests library and URL and stored in **predictions**.
- Twitter API data were gathered using Python's tweepy library and stored in **tweets_api**

2 Data assessment

During data assessment the following quality and tidiness issues have been found:

Data quality issues:

For tweet_data:

- Remove retweets
- Data completeness: columns 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp' should be dropped because they have a lot of missing data and we need only original data for this project.
- Data correctness: 23 tweets have rating denominator not equal to 10. These tweets will be removed.
- Data correctness: 24 tweets have rating numerator more than 20. These tweets will be removed
- Keep only columns where first prediction p1_dog is True
- Columns shoud be renamed: 'timestamp': 'tweet_timestamp', 'text': 'tweet_text','name': 'dog_name', 'source': 'tweet_source'
- Timestamp should be datetime format
- For source column url should be extracted from html tags.

For predictions:

- columns should be renamed: p1 - prediction1, p1_conf- prediction1_CI, p1_dog - prediction1_dog, p2 - prediction2, p2_conf- prediction2_CI, p2_dog - prediction2_dog, p3 - prediction3, p3_conf- prediction3_CI, p3_dog - prediction3_dog.

Tidiness issues

- For dog stage we have 4 columns: Doggo, floofer, puppo, pupper. We can create only one column 'dog_stage', where doggo, puppo, pupper are the values.
- Column 'rating numerator' and 'rating denominator' can be merge in one column called 'rating'. Column 'rating_denominator' will be removed because all values are 10.

3 Data cleaning

tweet_data, *predictions* and *tweets_api* table were copied to *tweet_data_clean*, *predictions_clean* and *tweets_api_clean* tables. After all quality and tideness issues were solved , all three tables were merged on tweet_id in a single dataframe **twitter_archive_master**.

4 Data storing

The merged dataframe was stored in ***twitter_archive_master.csv*** file

```
In [ ]:
```