WeRateDogs in Twitter

Data wrangling report

1. Data gathering:

In this project, the data was collected through different sources. The twitter_archive_enhanced file downloaded directly contains basic tweet data (tweet_id, rating, name and dog stages) from twitter users. Additional data including retweet count and favorite count ,was obtained via the Twitter API. Finally, the image-prediction file was downloaded programmatically using the Requests library and provided URL. This file provided the dog images links and the prediction results based on neural network algorithms.

2. Data assessment

After import to jupyter notebook as pandas' dataframe. All files are assessed visually and programmatically. The main problems are listed:

Quality issues

- Erroneous data types across all three tables (eg: twitter_id, id, id_str should be str but not float)
- Missing or inaccurate values (eg, name in table twitter_archive_enhanced)
- Some denominators are equal 0
- Columns with no or very few values (eg, 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id')

- Reweeted records should be removed and keep original only
- Inconsistence: Mixed lowercase and uppercase characters in dog breed name (p1,p2,p3)

Tidiness issues

- For table ‘twitter_ archive_enhanced ’, four variables (doggo, floofer, pupper, puppo) should be in one column as Dog_stage
- All three tables should be combined together before analysis

3. Data Cleaning

Quality issues

- Erroneous in data types, astype () method was employed to correct the data types
- Missing names may not be able to re-extracted or found .
- For the inaccurate rating number, if the accurate values could not be found, drop all those rows.
- Using upper or lower () method to change the dog breed name
- Drop all those columns without values

Tidiness issues

- Reorganized all four dog stages columns as one
- Using merge method to combine all three tables based on tweet_id