WeRateDogs DataSet Wrangling

Data Gathering

This is a data wrangling project aiming to gathering data from three separated datasets : twitter-archive-enhanced dataSet that contains 2356 row and 17attributes (tweet_id , in_reply_to_status_id , in_reply_to_user_id , timestamp , source , text , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp , expanded_urls , rating_numerator , rating_denominator , name , doggo , floofer , pupper and puppo) , image-predictions dataSet thst contain 2074 row and 12 column (tweet_id , jpg_url , img_num , p1 , p1_conf , p1_dog , p2 , p2_conf , p2_dog , p3 , p3_conf and p3_dog) and tweet_json dataSet that is supposed to be by Twitter's API which contains 2354 row and 31 column .

Data Assessment

Assessing is the precursor to cleaning. I need to identify and categorize common data quality and tidiness issues .

Quality issues

twitter_archive table

1. Columns (in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp) are missing many values, therefore we can analyze this table without these columns .

2. Missing records in the expanded_urls column (2297 non-null out of 2356) .
3. Erroneous data type (tweet_id , timestamp) .
4. There are extreme values in rating_numerator(larger than 14) .
5. There are Erroneous values in rating_denominator(larger than 10) .
6. NaN values in name column represented by the word 'None' .
7. Drop unneeded columns .
8. Lowercase of some column "name" .

image_predictions table

1. 1532 True values and 543 False values of p1_dog column , 1553 True values and 522 False values of p2_dog column and 1499 True values and 576 False values of p3_dog column , the False values are probably not dogs.
2. Erroneous data type (tweet_id) .
3. There are strange values like car_wheel and can_opener that need to be checked in p3 column .
4. Drop unneeded columns .

tweet_json table

1. There is no need to use columns like(coordinates , contributors , in_reply_to_screen_name , place , geo) to analyze.
2. Erroneous data type (id) .

Tidiness issues

twitter_archive table

1. The columns (doggo, floofer, pupper and puppo) must be values and not separated columns .
2. timestamp column needs to be separated to two columns (day column and time column) .

image_predictions table

3. The names of columns (p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog and p3_dog) need to have appropriate descriptive names .

4. Mereg all tables usig tweet_id .

Data Cleaning

After copying the three dataSets I started to clean what i assessed on the three copies .

1. The columns (doggo, floofer, pupper and puppo) got changed to values in column("dogs") and not separated columns .
2. Change erroneous data type (tweet_id , timestamp) .
3. Change timestamp column to be separated in to two columns (day column and time column) .
4. Change the names of columns (p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog and p3_dog) need to appropriate descriptive names (prediction_1 , prediction_1_confidence , prediction_1_dog , prediction_2 , prediction_2_confidence , prediction_2_dog , prediction_3" , prediction_3_confidence" , prediction_3_dog) .
5. Drop Columns (in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp , expanded_urls , source , timestamp ,) as

some of them are missing many values, and others we can analyze this table without them .

6. Change the extreme values in rating_numerator(larger than 14) to 10 .
7. Change Erroneous values in rating_denominator(any number but 10) to 10 .
8. Change None values in name column to np.NaN .
9. Filter 1532 True values of prediction_1_dog column , 1553 True values of prediction_2_dog column and 1499 True values and of prediction_3_dog column .
10. Check strange values like car_wheel and can_opener if there are still exist in p3 (prediction_3) column .
11. Change datatype of (tweet_id) to str in Cleaned_image_predictions dataset .
12. Drop unneeded columns in Cleaned_image_predictions dataset .
13. Drop columns (coordinates , contributors , in_reply_to_screen_name , in_reply_to_status_id , in_reply_to_status_id_str,quoted_status_id_str,truncated ,retweeted,quoted_status, is_quote_status,in_reply_to_user_id_str ,created_at, in_reply_to_user_id , place , geo , favorited , id_str) .
14. Change datatype of (id) column in Cleaned_tweet_json dataset .
15. Mereg all tables usig tweet_id .