wrangle_report

March 1, 2018

1 Wrangle_report

1.0.1 Quality

Twitter_Archive Table

1. in_reply_to_status_id,in_reply_to_user_id all only have 78 non-null value, 2278 missing value, 97% field in these columns are missing. retweeted_status_id,retweeted_status_user_id and retweeted_status_timestamp have 2175 missing value out of 2356, about 92% fields are missing. Since we don't use reweet data, I will delete the rows populated with values under these columns and remove these columns which have a large amount of missing value.

2. expanded_urls have 59 url with missing value. I will fill these missing expanded_urls up by combining 'https://twitter.com/dog_rates/status/' with their tweet id + '/photo/1'

3. Tweet_ID should be string rather than numeric value in three tables.

4. 'timestamp' variable has unnessary suffix '+0000',which needed to be removed. This columns should be time type, not string type. Year varibale can extracted and used for further analysis.

5. There is one denominator equal to 0, which is invalid rating. The rating(Both rating_numerator and rating_denominator) should be changed to 13/10. The rating_numerator and rating_denominator of tweet with denomintor = 2 should be 9/10, based on the original tweets. The rating_numerator and rating_denominator of 3 tweet' with denominator =11 are all wrong, the correct rating should be 14/10(tweet_id 775096608509886464) ,14/10(tweet_id 740373189193256964) and 10/10 (tweet_id 682962037429899265)

6. A New Rating column should be created using rating_numerator/rating_denominator for futher data analysis.

7. 'name' variable has 745 'None', 55 'a', 7 'an' name and 8 'the',these value are not names. Name is extracted from the word after 'This is', a lot of text does not follow this rule,some texts just don't include name. Hence, I will replace 'a', 'an', 'the', 'Non' with nan value.

8. I noticed that the value of source column has the same values in some rows in the visual assessment, after checking in programmatical assessment, I found out that there are 4 different

kinds of source – 'Twitter for iPhone','Vine - Make a Scene','Twitter Web Client',and 'TweetDeck'. Therefore, the category string should be extracted from source column and create a new column -Source_Category to label them.

9. We can see from tweet_id = 883482846933004288, the text states the rating should be 13.5/10, but the rating numerator is 5. Several tweet have the same issue and I will fix by extracting data from text.

imagePrediction Table

8. In visual assessment and programatical accessment, I found that for p1,p2,p3 columns, while some value have 1st letter capitalized, and some are not, which may cause issue if we want to label data. To keep the value format consistent, I will convert them all into all lower case character.

9. Also, p1_conf, p2_conf and p3_conf display as 6 digit decimal value, which is not easy to observe, we should keep only 4 decimal value.

10. 'newfoundland', 'laptop','sea_lion','lhasa','mitten','feather_boa' in p2 and p1 are not a breed of dog, their indicator should be equal to TRUE, not False.

1.0.2 Tidyness

1. In Twitter_Archive table, the doggo, floofer, pupper, and puppo column should be merged into one column called dog_stage to reduce redundency.

2. The column names in imagePrediction are very random and don't make sense to end users if a person did not know the background information, hence, i will rename with more meaninful names.

3. Lastly, in order to include reweets_counts and favorite_count into tweet data analysis, I will merge tweet_json table with Twitter_Archive Table by tweet_id , so we will have one Twitter_Achive table and imagePrediction in the end.