wrangle_report

August 16, 2020

1 wrangle_report

1.1 introduction:

In this Udacity project we wrangling data that provided by three steps. First, gathering data then assessing data and the last clean it. after that we visually and analysed the data.

1.2 Gather data:

Gather data from three files described below in a Jupiter Notebook title wrangle_act.ipynb: The WeRateDogs Twitter archive (manual upload of 'twitter_archive_enhanced.csv' to work space and read it by read_csv()) The tweet image predictions (manual upload of 'image-predictions.tsv' and read it by read_csv().Also, I add (sep= '^') for file.tsv). Each tweet's retweet count , favourite count and tweet ID(upload zipfile to work space and then extract all contents from zip file by ZipFile() method. In last saved the file with DataFrame named tweet_json)

1.3 Assessing Data:

In assessment phase we do two type of Assessment which is : Visual assessment and Programmatic assessment.

- Frist, lets took about Visual assessment: I scrolling through the data frames that listed above and have a quick look at the structure of data.
- Second, programmatic assessment: I using code to view specific portions and summaries of the data (head , info, describe and sample methods, for example).

after assessing data I note some issue in data it is also called unclean data.

1.3.1 Quality issues:

twitter-archive-enhanced table:

- Completeness issue: in_reply_to_status_id & in_reply_to_user_id is 78 instead of 2356 retweeted_status_id ,retweeted_status_user_id & retweeted_status_timestampis is 181 instead of 2356
- data type issue: id shoud be string instead of int timestamp & retweeted_status_timestamp shoud be timestamp instead of object change rating numerators to float instead of int
- Validity issue: only want original ratings (no retweets) only want original ratings that have images there are names of dogs do not quite names source column containing <a html tags

imge_predict table:

- data type issue: id shoud be string instead of int
- Consistenc: rename colums (p1 &p2& p3)

tweet_json table:

- data type issue: id shoud be string instead of int
- Consistenc: rename id colum to tweet_id

1.3.2 Tidiness issue:

- The columns (doggo,floofer,pupper,puppo) should be in one column dog_type with a category datatype.
- Merging the three table, using the tweet_id column

1.4 Clean Data:

For cleaning data we did the following: There are some column having null value and we will drop them by drop() method, change tweet_id data type to object & rename id column to tweet_id in tweet_json table ,change datatype of rating_numerator to float by astype('float'). only we want original ratings (no retweets) so I select all text with no RT then save it on dataframe maste Also, we only want original ratings that have images. So we select imge url not null and save it in master dataframe.There are names of dogs do not quite names like a ,an ,the, by. we observed that all the lowercase are not names so we will remove it. The source column containing < a html tags> so we clean it by using method to remove tags and apply it with apply method.

In [ ]: