We Rate Dogs Twitter Data

Wrangling Report

SOURCES:

wrd_df: Data is collected by the csv file given by Udacity **predict_df**: Data

was obtained by the Url of the website provided by udacity

tweets_df: For this data first, I did was to request for developer keys and then extracted the twitter data against the ids available in csv file, which was then loaded in to pandas data frame

Process:

Analysed all three of the datasets for the quality and tidiness issues:

I audited the data by check datatype, info and describe function ,some of the tweets are resulted I error while retrieval .

I have tackled with two kind of issues in the dataset:

1) Tidiness issue
2) Quality issue

Tidiness issue:

1) 7) columns like,'in_reply_to_user_id','in_reply_to_status_id','retweeted_status_id','retweeted_status_user_id','expanded_urls','doggo','floofer','pupper','puppo' should be dropped

2) none values have to be converted to nan values

3) doogo, floofer,pupper,puppo can be combined in to a single column

Quality issue:

1) if we look at the **rating_numerator** and **rating_denominator** columns, the maximum and minimum are out of range

2) drop the rows with denominator less than 10 rating

3) drop the rows with 0 denominator and numerator

4) drop the rows with numerator and denominator greater than 20

5) tweet_id is integer type

6) names like 'a', 'an', 'the', 'just', 'one', 'very', 'quite', 'not', 'actually', 'mad', 'space', 'infuriating', 'all', 'officially', '0', 'old', 'life', 'unacceptable', 'my', 'incredibly', 'by', 'his', 'such ' should be manually changed

8) combining the dog's class in to one column

9) Egyptian Cat is not a dog and should be removed

STORAGE:

I have stored the data in to csv one for tweets data as tweets_df and another for the image data as predict_df