Data wrangling

Data gathering :

The data were in three different sources, the first was tweets archive I manually download it as csv file,

The second one was image prediction data set I download it using code from this URL : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv,

the third one I hade to create twitter account from here https://developer.twitter.com/content/developer-twitter/en.html

after I create the developer account I used it to pulled the favorites and retweets count for each tweet I had using the tweet ID,

I had problem that tweets ID in tweets archive are wrong so I extract the tweets id using the URL path, also twitter delayed the collecting of my developer account and that affected my project summation.

data assessing

after gathering the data I started showing the information of every dataframes I have to see the issues in the data.

Data quality :

1- in tweet_archeve, the column tweet_id has wrong values.

2-tweet_archeve has retweeted tweets which we don't want them

3-in tweet_archeve, columns 'doggo','floofer', 'pupper' and 'puppo' should be in one column.

4-in image_prediction there is no column for most confedence breed of dogs

5-in tweet_archeve columns in_reply_to_status_id and in_reply_to_user_id have lots of missing data.

6-we only want the tweet with images

7-there are missing tweets since the tweets in tweet_archeve are 2356 and in tweets are 2190 and in image_prediction are 2075

8-wrong datatype in column timestamp in tweet_archeve.

Data Tidiness:

1- tweet_archeve,image_prediction and tweets_interaction should be in one dataframe

2-we should delet columns 'doggo','floofer', 'pupper' and 'puppo' after make it in one column

3-we should delet columns in_reply_to_status_id and in_reply_to_user_id after deleting replay tweets

4-we should delet columns 'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp' after deleting retweets tweets

Data cleaning :

First I made a copy of every dataframes I have , then I deleted the retweets tweets and delete the columns that has related to it, then I deleted the replay tweets and deleted the columns related to it , after that I merge the dogs stages in one column and deleted the old one, then I put

the most confecence dog breed and the confidence percentage in two columns, after that I join all the three dataframes together using the tweet id , then I deleted the tweets that has no image, then I change the data type of timestamp column