Wrangle Report

Data gathering

df1

twitter-archive-enhanced.csv was gathered through pd.read_csv since it is the file in hand.

df2

image_prediction.tsv was gathered programmatically by using requests library and url.

df3

My request to use twitter API was rejected so I cannot gather data from twitter API. Since the code provided to gather some of the required data for the project is not working, I directly use tweet_json.txt and read it line by line to form a data frame.

Data assessing

I checked all of the three dataframe manually and programmatically.

I checked whether there is any null value, there is none.

I kept record of all the problems with each dataframe. Then I classified them into quality issues and tidiness issues.

Quality:

- [x] ~~We don't want retweets, so the rows whose retweeted_status_id is not null should be removed.~~
- [x] ~~The data type of timestamp should be datetime.~~
- [x] ~~The highest prediction which is dog should be identified in df2 since not all the prediction in df2 is dog.~~
- [x] ~~Not all ratings has image. We should remove the rows in df1 that do not have image in df2.~~
- [x] ~~None in df1 should be NAN~~
- [x] ~~Some ratings in df1 are wrong collection from the text.~~
- [x] ~~The data type of retweet_count and favorite_count in df3 should be int. The data type of tweet_id should be int.~~
- [x] ~~Some ratings in df1 is too high(>=2) or too low<=0.5. Something might be wrong.~~

Tidiness:

- [x] ~~df3 can be merged to df1.~~
- [x] ~~prediction in df2 can be merged to df1~~
- [x] ~~In df1, ratings should be one column.~~
- [x] ~~In df1,the last three columns should be one to show the stages of dogs.~~
- [x] ~~Three columns 'in_reply_to_status_id' and in_reply_to_user_id' ,'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'in df1 should be removed~~

Data cleaning

First of all, I made copies of the three uncleaned dataframes.

Issue 1

We don't want retweets.

Define

Slicing the dataframe to be the part whose retweeted_status_id is null.

Issue 2

Not all ratings has image.

Define

Get the tweet id in df2 and merge it with df1 using inner merge.

Issue 3

In df1,the last four columns should be one to show the stages of dogs.

Define

Set the None values in the last four columns in df to be ''. Then add a new columns called 'stage' whose values is the concatenation of strings of the last four columns. Then replace the '' into np.nan. Then remove the original four columns.

Issue 4

In df1, ratings should be one column.

Define

Divide the numerator with the denominator.

Issue 5

Some ratings are wrong collection from the text.

Define

Take all the text that denominator is not 10 and change the rating manually.

Issue 6

Strange ratings.

Define

Take all the ratings that is higher than 2 or smaller than 0.5. Manually check those ratings.

Issue 7

Useless columns

Define

Use drop method

Issue 8

Wrong datatype

Define

Use pd.to_datatime to change timestamp.

Use astype to change int. Id should be Int64.

Issue 9

The highest prediction

Define

Set all the prediction result to be p1. Select those p1 which are not dogs and set their prediction to be p2. Select those p2 which are not dogs and set their prediction to be p3. Select those p3 which are not dogs and set their prediction to be NAN.

Issue 10

Merge the information

Define

Use pd.merge

Store data

Store df_clean to a csv file using to_csv