Wrangle Report

Hereby report that the following actions were made.

1. Gathering data

Following three files gathered;

- ‘twitter_archive_enhance.csv’ by download
- ‘image prediction.tsv’ from Udacity by Requests
- ‘tweet_json.text’ from Twitter API by Tweety

Those are read as Pandas DataFrame as follows;

- df_main(‘twitter_archive_enhance.csv’)
- df_class(‘image prediction.tsv’)
- df_tweet(‘tweet_json.text’)

2. Assessing data

Based on the two categories, Quality related and Tidiness related, assessment was made. The Following are the summaries of those findings.

Quality(11 issues found)

df_main(twitter_archive_enhanced.csv)

- datatype of ‘timestamp’ is object
- ‘source’ column,’ and ‘’ are not necessary.
- ‘name’ column, ‘None’ or ‘a’ etc. are in it.
- ‘rating_numerator’ column, the distribution of the values looks abnormal(max=1776 , mean=13.1, min=0)
- There are unnecessary columns(majority of data in those columns is Nan, ‘in_reply_to_status_id’, ‘in_reply_to_user_id, ‘retweeted_status_id’, ‘retweeted_status_user_id’, ‘retweeted_status_timestamp’)

df_class(image predictions.tsv)

- There are rows except dogs
- The column ‘img_num’ is not necessary
- ‘p1’, ‘p2’, ‘p3’, lowercase and uppercase are mixed

- Those columns such as 'p1' and 'p1_conf' are not easy to understand(should be renamed as 'breed' and 'p1_conf' as 'confidence_level')
- p3' is not necessary because 'p1' + 'p2' covers majority

df_tweet(tweet_json.text)

- columns name 'favorites' and 'retweet' should be renamed as 'favorites_counts' and 'retweet_counts'

df_weratedogs

- All retweets were removed.

Tidiness(2 issues found)

df_main(twitter_archive_enhanced.csv)

- dog stages such columns as 'doggo', 'floofer', 'pupper' and 'puppo' are untidy. tidying

3 dfs(df_main, df_class, df_tweet) merge

- those dfs are merged as 'df_weratedogs'

3. Cleaning data

Those issues found in the above process were rectified by this cleaning process as follows;

Quality(10 issues found)

df_main(twitter_archive_enhanced.csv)

- datatype of 'timestamp' is object.: datatype was changed to 'datetime'
- 'source' column,' and '' are not necessary.: tabs were removed
- 'name' column, 'None' or 'a' are in it.: 'None' and 'a' etc. are replaced by Nan.
- 'rating_numerator' column, the distribution of the values looks abnormal(max=1776 , mean=13.1, min=0): Try to pick up values of rating from text and compare if there is any difference for replacement. But both are similar, therefore leave them as they are by understanding that those outlier figures are normal ones.

df_class(image predictions.tsv)

- There are rows except dogs.: Those rows were dropped(excluded from pickups)

- The column ‘img_num’ is not necessary.:This column dropped
- ‘p1’, ‘p2’, ‘p3’, lowercase and uppercase are mixed. All were lowercase.
- Those columns such as ’p1’ and ‘p1_conf’ are not easy to understand(should be renamed as ‘breed’ and ‘predictions).: Columns renamed.
- ‘p3’ is not necessary because ‘p1’ + ‘p2’ covers majority.:’p3’ and ‘p3_conf’ are dropped
-

df_tweet(tweet_json.text)

- columns name ‘favorites’ and ‘retweet’ should be renamed as ‘favorites_counts’ and ‘retweet_counts’ .: columns renamed
- There are unnecessary columns(majority of data in those columns is Nan, ‘in_reply_to_status_id’, ‘in_reply_to_user_id, ‘retweeted_status_id’, ‘retweeted_status_user_id’, ‘retweeted_status_timestamp’): Those columns are dropped(excluded from pickups)

Tidiness(2 issues found)

df_main(twitter_archive_enhanced.csv)

- dog stages such columns as ‘doggo’, ‘floofer’, ‘pupper’ and ‘puppo’ are untidy. tidying: Those four columns are melted to one column as ‘dog_stage’

3 dfs(df_main, df_class, df_tweet) merge

- those dfs are merged.

4. Storing, Analyzing and Visualizing Data

Cleaned three data frames are merged as ‘df_weratedogs’ and saved as a csv file as ‘weratedogs.csv’.