Act report

Three files downloaded and made following cleaning were made, and all three data frames were merged.

cleaning operation made;

- 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’ 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.ss(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.
- columns name ‘favorites’ and ‘retweet’ should be renamed as ‘favorites_counts’ and ‘retweet_counts’ .: columns renamed
- dog stages such columns as ‘doggo’, ‘floofer’, ‘pupper’ and ‘puppo’ are untidy. tidying: Those four columns are melted to one column as ‘dog_stage’
- 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)
- ‘p3’ is not necessary because ‘p1’ + ‘p2’ covers majority.:’p3’ and ‘p3_conf’ are dropped

There are still many missing values on the columns such as names of dogs, stage of dogs, and breed of dogs. The breed may be possible to fill in by deep learning prediction that is, however, beyond my capability at this stage. Besides breed, there is no way to fill (guess) missing data such as name or dog stage. That means that this data is still far away from ‘clean’.

Regarding the analysis of the data such as ranking of dogs in this dataframe, it is very difficult to carry on because of the availability and inconsistency of numerical data. Potential data to be used for the ranking of the dogs are rating, counts of favorites and retweet. But, the distribution of values of ranking, they range between o to 1766(mean=13.1, std=45.7)against their denominators are 10. It is hard to distinguish outliers. It was shown in Figure1(a box plot of the distribution of counts of dog rating).

The similar situation is for the data on both favorites counts and retweet counts as were shown below Figure2.

Having known that it is not reasonable to judge what is the best dog due to the above mentioned reason but , the following is a photo of the 2nd best dog judging from the favorite and retweet counts.

Figure.1

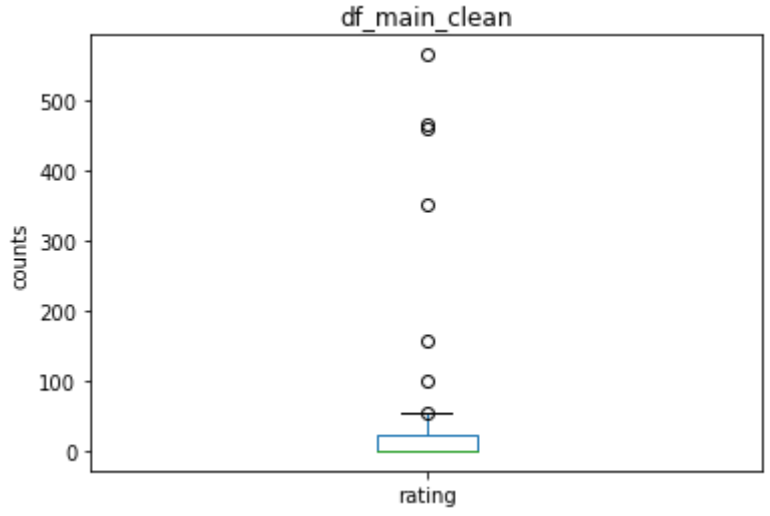
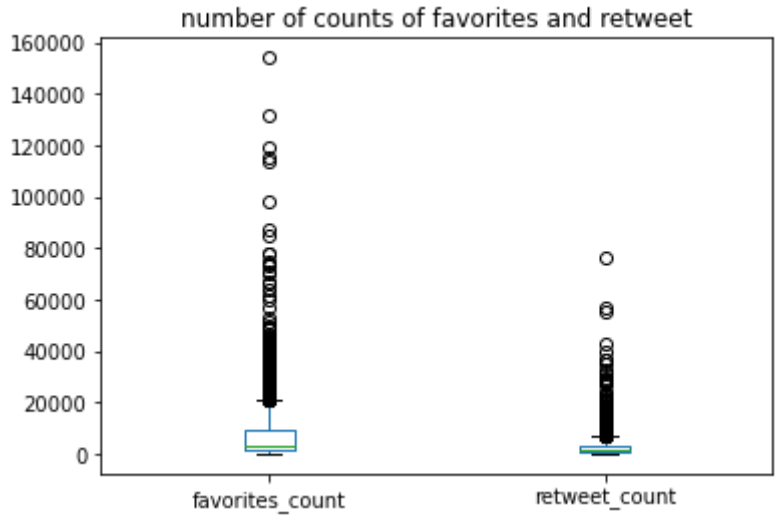


Figure 2.


number of counts of favorites and retweet
160000
140000
120000
100000
80000
60000
40000
20000
0
favorites_count
retweet_count