Act Report

Project: Wrangle and Analyze Data

Jungun Goo

September 2018

The clean dataframe contains 2283 observations. The key information contains rating, doggy type, the number of favourites and retweets.

Basic information

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2283 entries, 0 to 2282
Data columns (total 23 columns):
tweet_id                 2283 non-null object
in_reply_to_status_id    23 non-null object
in_reply_to_user_id      23 non-null object
timestamp                2283 non-null datetime64[ns]
source                   2283 non-null object
text                     2283 non-null object
expanded_urls            2283 non-null object
name                     1604 non-null object
type                     371 non-null object
favorites                2283 non-null int64
retweets                 2283 non-null int64
rating                   2283 non-null float64
jpg_url                  2067 non-null object
img_num                  2067 non-null float64
p1                       2067 non-null object
p1_conf                  2067 non-null float64
p1_dog                   2067 non-null object
p2                       2067 non-null object
p2_conf                  2067 non-null float64
p2_dog                   2067 non-null object
p3                       2067 non-null object
p3_conf                  2067 non-null float64
p3_dog                   2067 non-null object
dtypes: datetime64[ns](1), float64(5), int64(2), object(15)
```

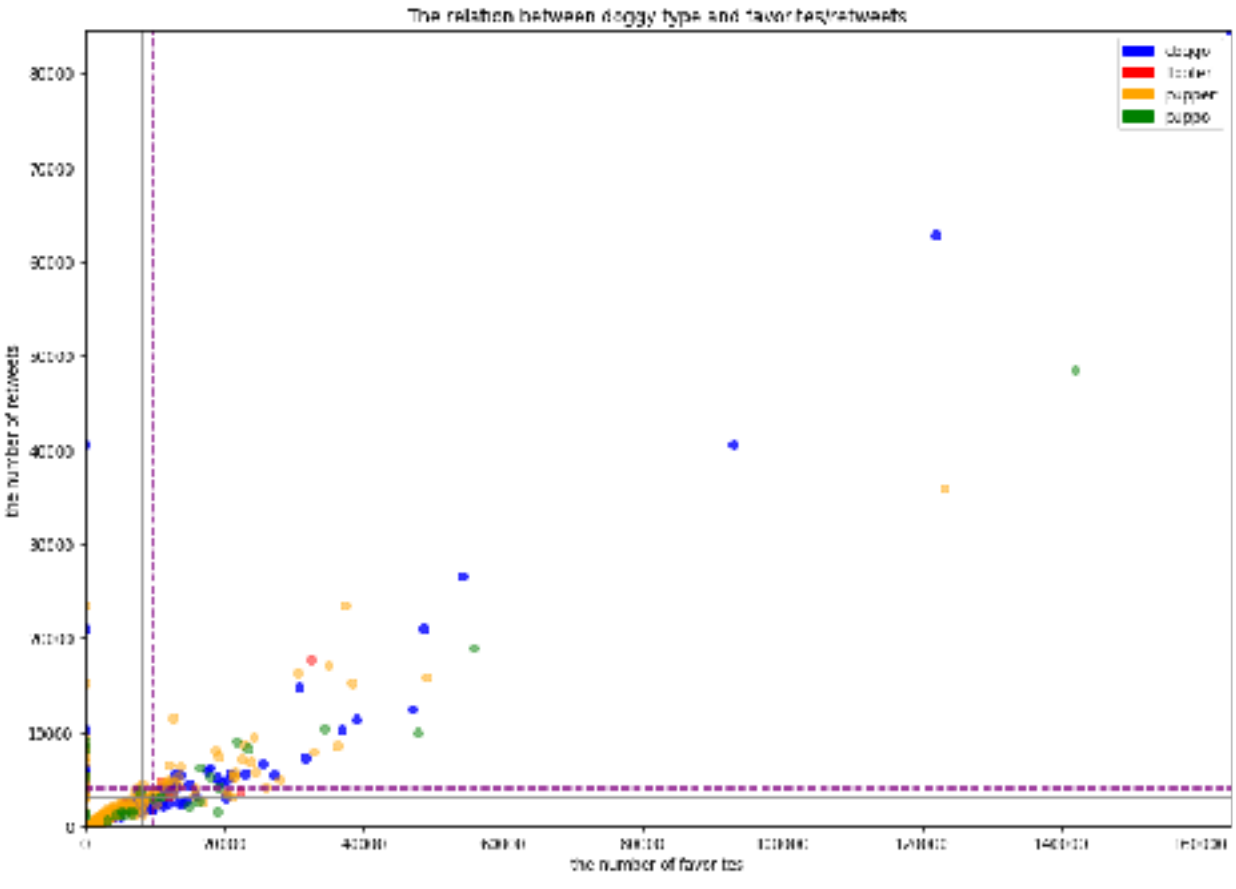
Basic statistics

	favorites	retweets	rating	img_num	p1_conf	p2_conf	p3_conf
count	2283.000000	2283.000000	2283.000000	2067.000000	2067.000000	2.067000e+03	2.067000e+03
mean	8199.341858	3046.429260	1.171521	1.203677	0.595032	1.345858e-01	6.029258e-02
std	12477.996182	5090.752634	3.603531	0.562309	0.271238	1.007836e-01	5.095999e-02
min	0.000000	12.000000	0.000000	1.000000	0.044333	1.011300e-08	1.740170e-10
25%	1461.500000	633.000000	1.000000	1.000000	0.364412	5.370120e-02	1.514795e-02
50%	3602.000000	1463.000000	1.100000	1.000000	0.589011	1.181810e-01	4.934910e-02
75%	10179.000000	3641.500000	1.200000	1.000000	0.846942	1.955655e-01	9.203545e-02
max	154340.000000	84546.000000	177.600000	4.000000	1.000000	4.880140e-01	2.734190e-01

Visualizations and Insights

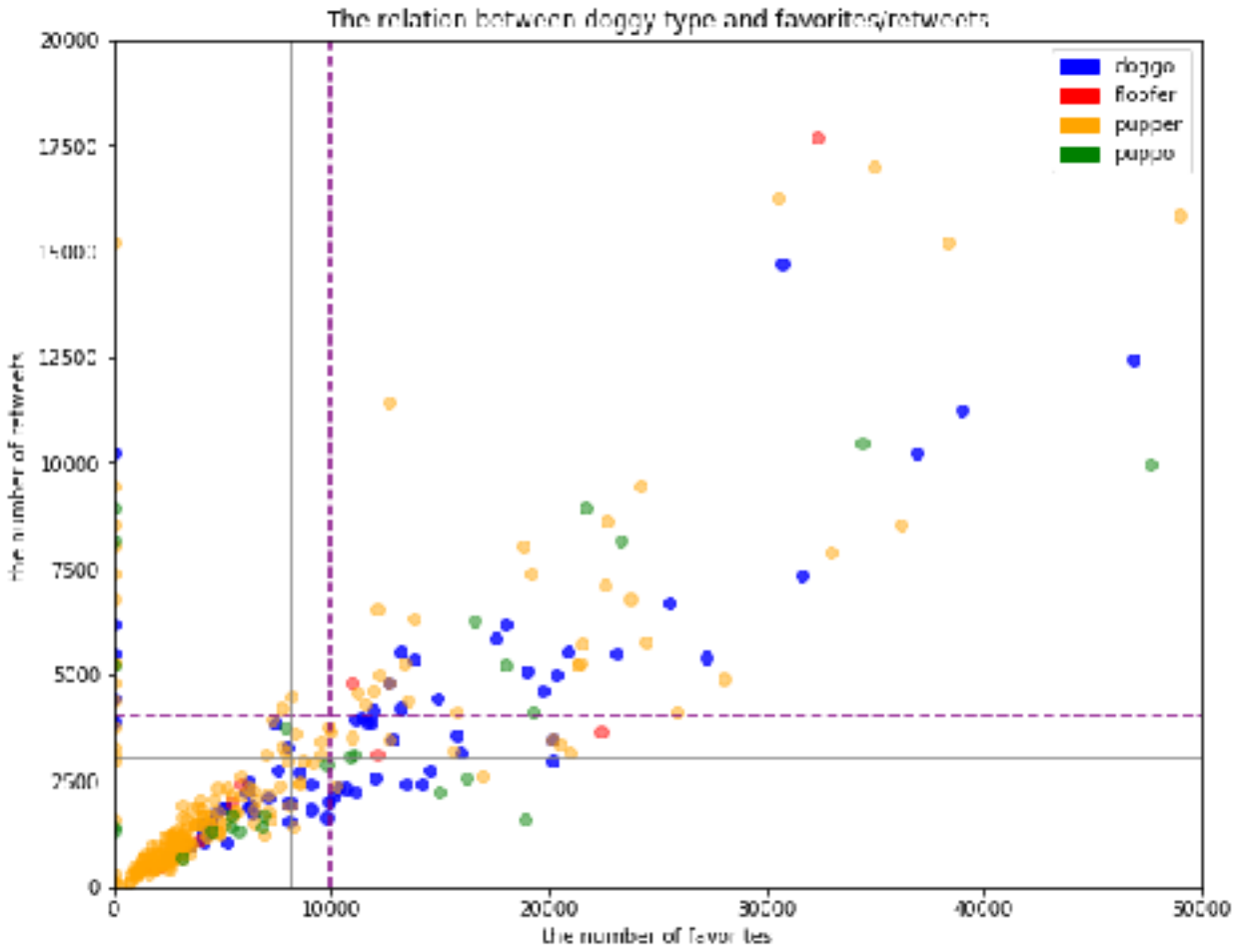
Chart 1: Doggy type in relation to favourites/retweets

I want to research how favorites and retweets are related to each other. In addition, I'd like to investigate how it differs depending on the type of doggy(doggo, pupper, puppo, floofer)



The below is an enlarged graph of the favourites and retweets below than 50000 and 20000 respectively.

- Purple dotted line: the averages of data which has doggy type information.
- Grey solid line: the averages of all data

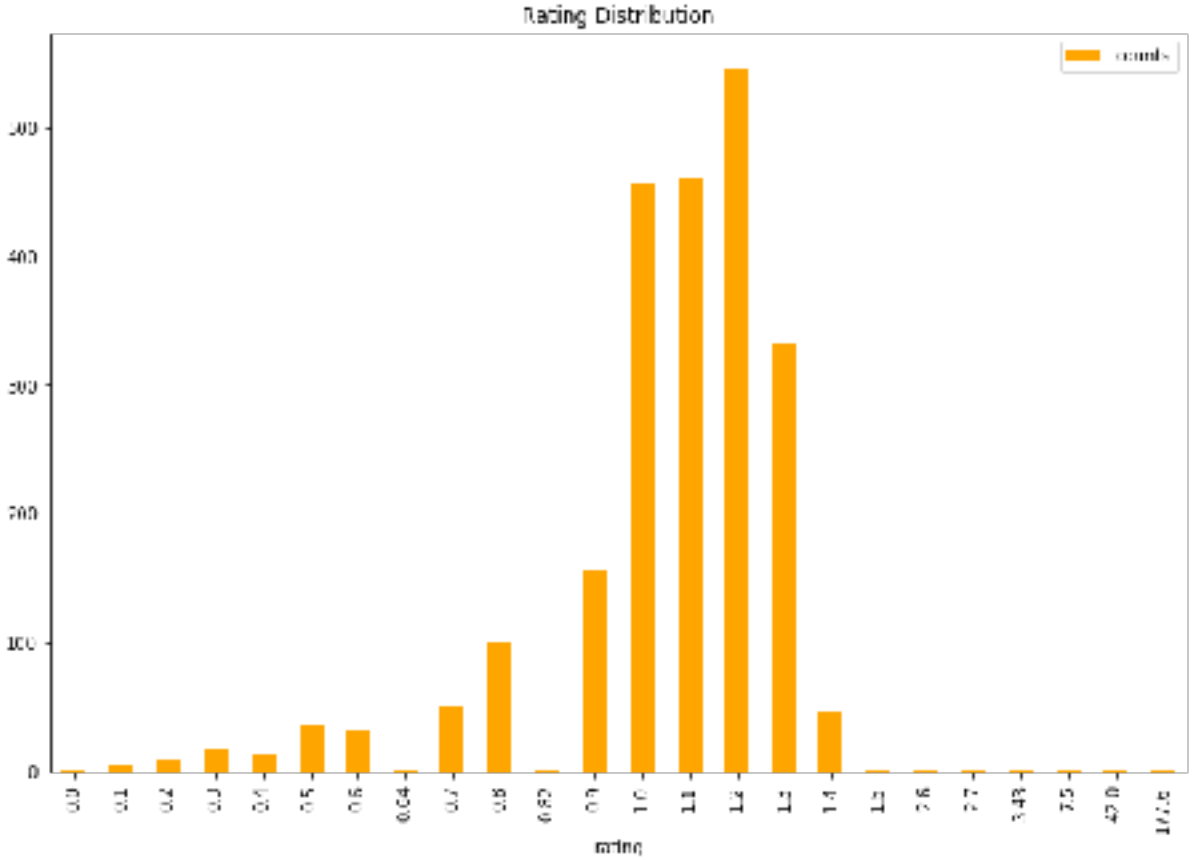


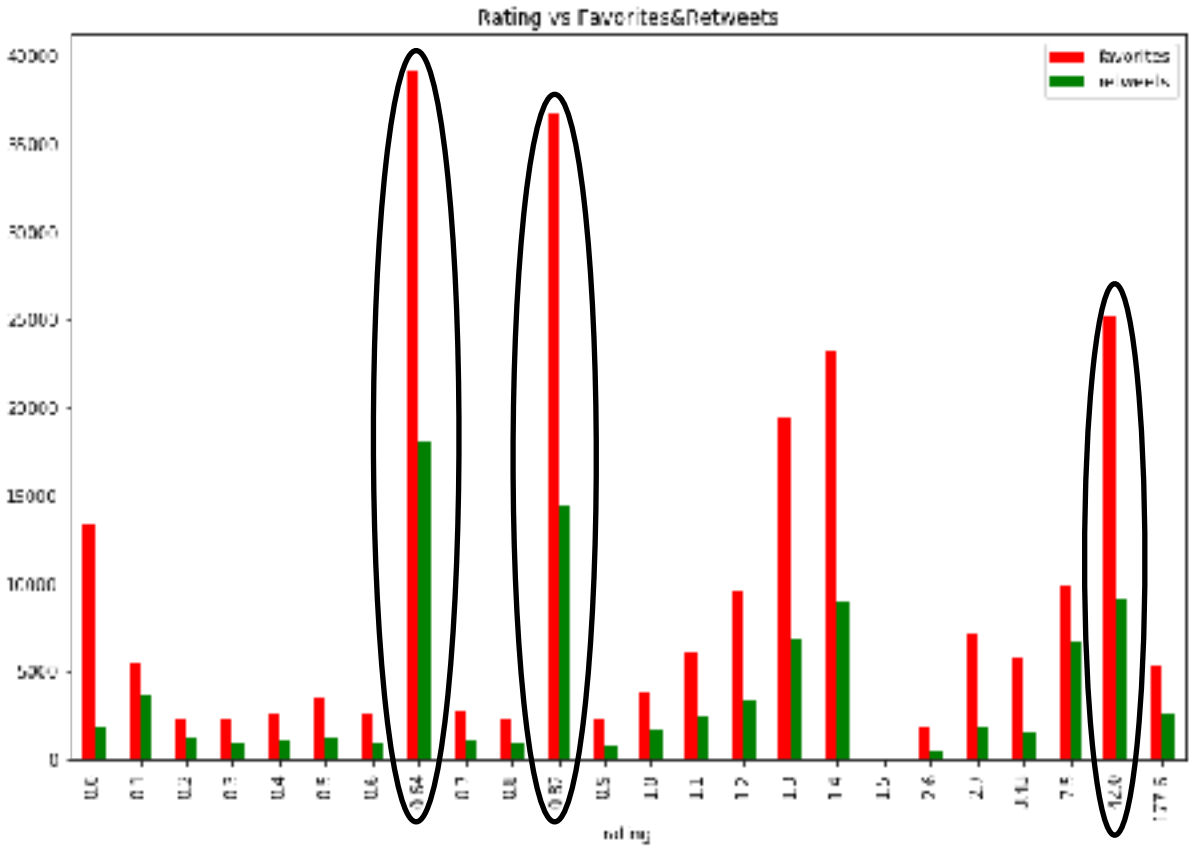
Insight from Chart 1:

- The average number of favorites and retweets is higher if the entry has the information of doggy type than the others which do not have.
- There is a linear relationship between favorites and retweets.
- Overall, however, there tends to be 2.5 times as many favorites as retweets.
- In addition, there are some data retweeted with no favorites received, even though there is no case the other way around.
- *pupper* tends to receive less favorites and retweets as you can see orange dots are highly dense under the average bars.
- *doggo* is the most popular among other doggy types many of blue dots are appeared in the upper part of the averages.

Chart 2. The relation between rating and favorites/retweets

Can we anticipate the ratings and is there any correlation with favorites and retweets?



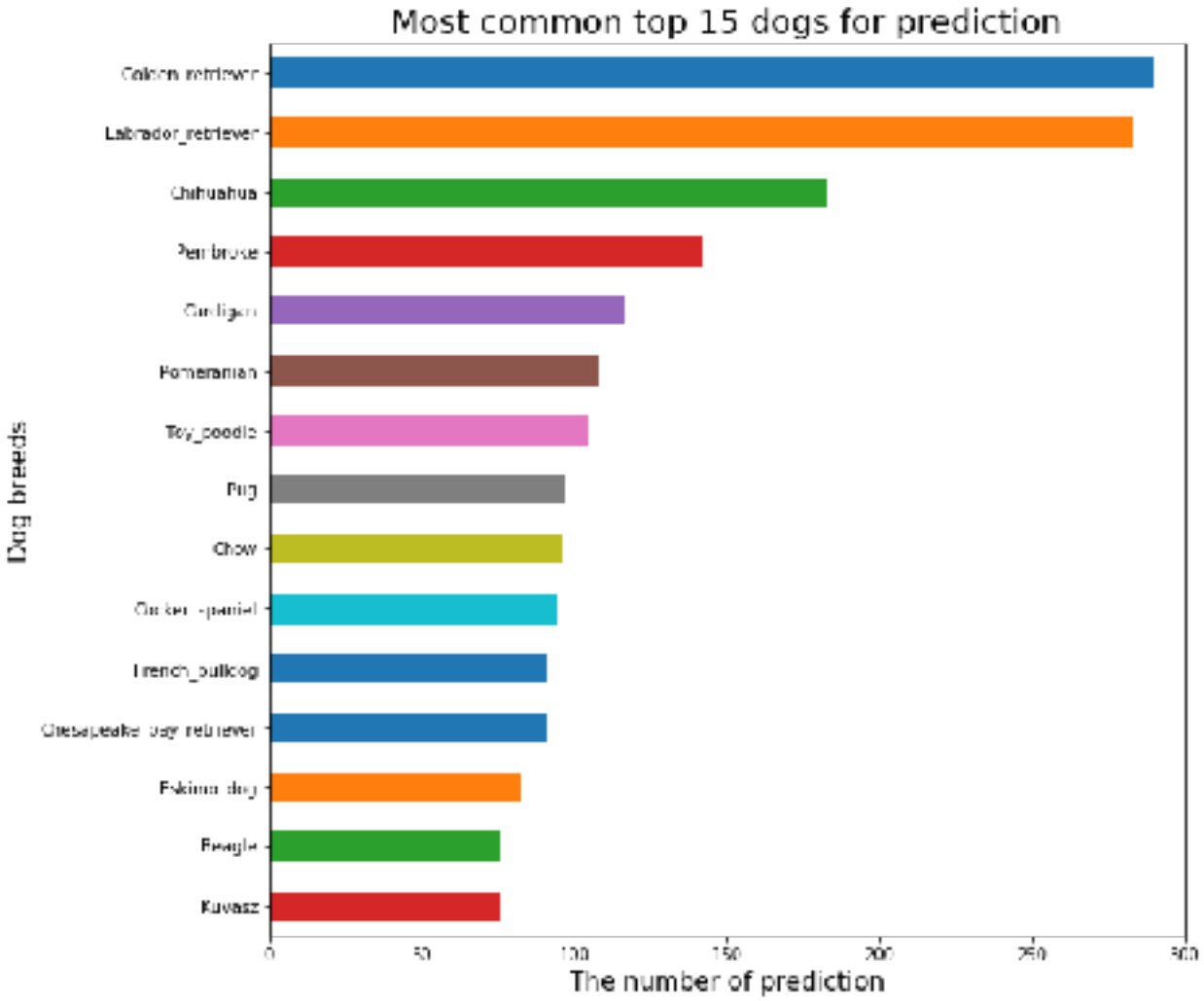


	rating	counts	favorites	retweets
0	0.00	2	13466.500000	1810.500000
1	0.10	5	5509.000000	3725.600000
2	0.20	10	2390.900000	1326.800000
3	0.30	19	2312.842105	909.421053
4	0.40	15	2646.733333	1065.800000
5	0.50	36	3605.166667	1200.694444
6	0.60	32	2627.312500	999.093750
7	0.64	1	39280.000000	18168.000000
8	0.70	52	2867.519231	1071.903846
9	0.80	102	2265.774510	977.196078
10	0.82	1	36799.000000	14432.000000
11	0.90	155	2354.658065	827.929032
12	1.00	457	3851.308534	1660.356674
13	1.10	461	6135.555315	2491.390456
14	1.20	546	9674.287545	3312.694139
15	1.30	393	19415.765765	6923.684685
16	1.40	48	23206.875000	8070.958333
17	1.50	1	0.000000	36.000000
18	2.60	1	1813.000000	523.000000
19	2.70	1	7072.000000	1800.000000
20	3.43	1	6766.000000	1661.000000
21	7.50	2	9846.000000	6760.000000
22	42.00	1	25242.000000	9150.000000
23	177.60	1	5452.000000	2676.000000

Insight from Chart 2:

• Most of dogs have rated between 0.9 and 1.3 as you can find it in the first graph.
• In the second graph, the number of favorites and retweets has a significant linear correlation in the section between 0.9 and 1.4 rating.
• However, some ratings(0.64, 0.82 and 42) have unusual high value as you find it in the second graph. And this is unreliable data, because there was just only one person to mark this rate according to the table left.
• Therefore, we can consider this is not a representative figure of the user's tendency.

Chart 3. Does the accuracy rate for the breed prediction differ depending on the dog breeds?



The below chart shows the average accuracy rate of divided into 5 groups (depending on the popularity)	
1 ~ 50	94.0%
50 ~ 100	78.0%
100 ~ 200	25.0%
200 ~ 400	3.5%
400~	0.0%

breed	accuracy
Golden_retriever	100.000000
Labrador_retriever	100.000000
Chihuahua	100.000000
Pembroke	100.000000
Cardigan	99.137931
Pomeranian	100.000000
Toy_poodle	100.000000
Pug	100.000000
Chow	100.000000
Cocker_spaniel	100.000000
French_bulldog	100.000000
Chesapeake_bay_retriever	100.000000
Eskimo_dog	100.000000
Beagle	100.000000
Kuvasz	100.000000
Siberian_husky	100.000000
Samoyed	100.000000
Staffordshire_bullterrier	100.000000
Malamute	100.000000
Pekinese	100.000000
Kelpie	100.000000
American_staffordshire_terrier	100.000000
Miniature_pinscher	100.000000
Great_pyrenees	100.000000

Insight from Chart 2:

- Overall, the accuracy rate for predicting dog breeds is about *74%* which seems pretty reliable. (It can guess the type of dogs correctly 3 cases out of 4)
- There is no strong evidence for explaining whether there's the differentiation of accuracy depending on dog breeds.
- If a learning system has more data, then becomes more accurate to predict. It doesn't necessarily mean that a system recognize some specific breeds more precisely. Because the accuracy is 100% for the most of dog breeds of the top 50 common ones. (You can find it the table left)
- We can assume that the prediction system might be trained with so many data to be correct not because some dog breeds have distinctive features to be recognized better than other breeds.
- We plotted the 15 most common dog breeds in a bar chart. We see that according to this chart, the most common dog breed is the golden retriever with almost 300 times.