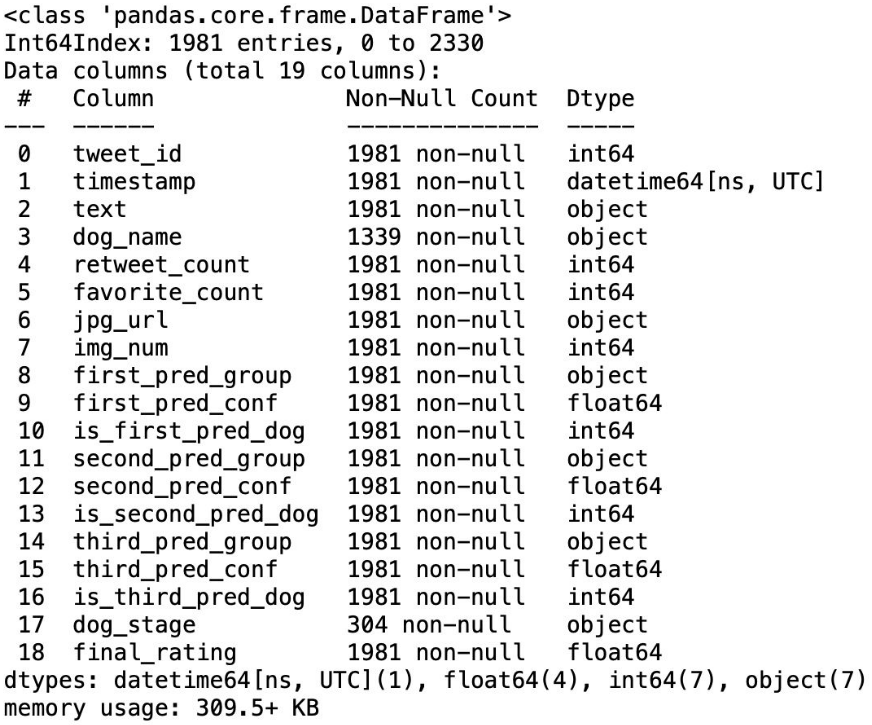
Twitter Data Analysis

Based on the previous step which is Data Wrangling, we now have the clean data as below

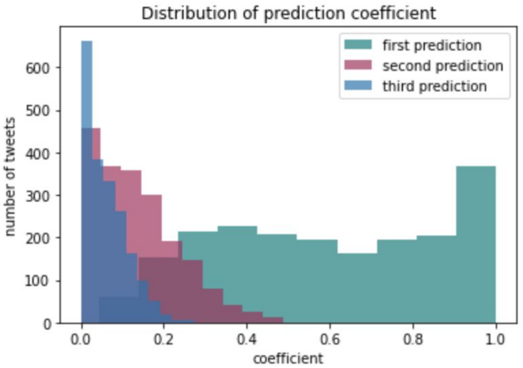


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1981 entries, 0 to 2330
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1981 non-null   int64
 1   timestamp           1981 non-null   datetime64[ns, UTC]
 2   text                1981 non-null   object
 3   dog_name            1339 non-null   object
 4   retweet_count       1981 non-null   int64
 5   favorite_count      1981 non-null   int64
 6   jpg_url             1981 non-null   object
 7   img_num             1981 non-null   int64
 8   first_pred_group    1981 non-null   object
 9   first_pred_conf     1981 non-null   float64
 10  is_first_pred_dog   1981 non-null   int64
 11  second_pred_group   1981 non-null   object
 12  second_pred_conf    1981 non-null   float64
 13  is_second_pred_dog  1981 non-null   int64
 14  third_pred_group    1981 non-null   object
 15  third_pred_conf     1981 non-null   float64
 16  is_third_pred_dog   1981 non-null   int64
 17  dog_stage           304 non-null    object
 18  final_rating        1981 non-null   float64
dtypes: datetime64[ns, UTC](1), float64(4), int64(7), object(7)
memory usage: 309.5+ KB
```

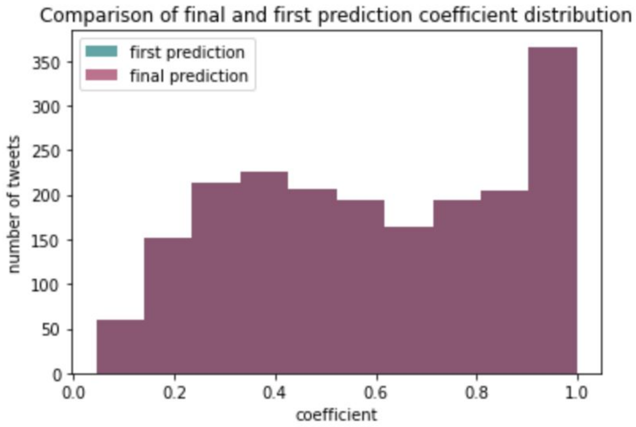
In this step, we will try to work on the analysis so we can get some insights from the data that we cleaned with effort. To analyze the data, we can come up with some questions

How is the proportion of dog's breed?

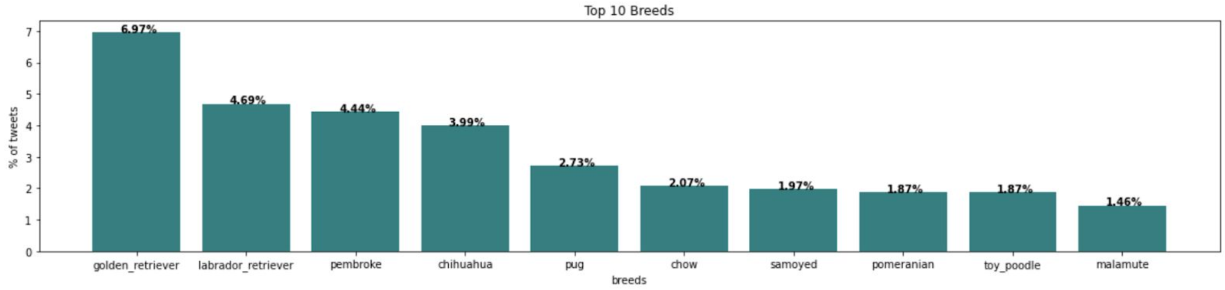
If we see our data structure, there are 3 different algorithms to identify or predict the breed of dog(s) in the image. To get to know which is the most frequent breed in our data, we need to choose the highest coefficient of prediction amongst those 3 algorithms. firstly I, we can see the distribution of coefficient from each algorithm



We can see that the **first algorithm produces a higher coefficient** compared to the second and third algorithm. The second and third algorithms tend to have positive skewed distribution. This got us wondering whether we could use the first algorithm alone to predict the breed. But, after this we will still choose the higher coefficient for each algorithm



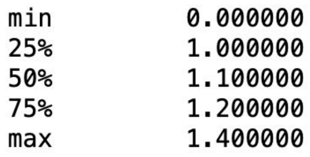
Seems like **we can use the first algorithm as our prediction** since the distribution of the first and final algorithm are the same. After this we will see the top 10 breeds in our data. Since the total of breeds is too many, we will choose only 10 breeds that have higher proportion. The result is shown below



From the picture above, 6.97% of tweets are showing golden retriever dogs, 4.69% are showing labrador retriever dogs, 4.44% are talking about pembroke dogs, 3.99% are talking about chihuahua, and the rest can be seen in the chart.

What is the breed of dogs that have higher ratings and low ratings?

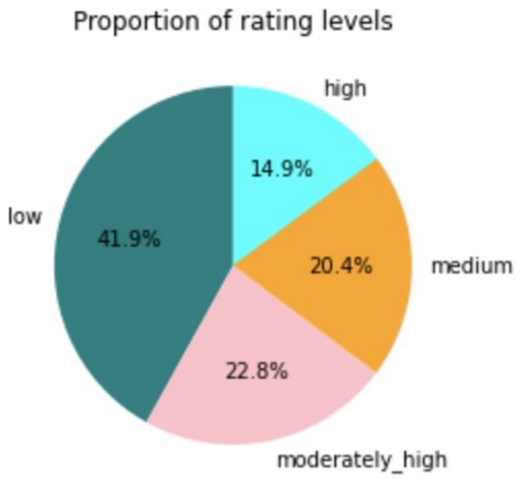
Next, we want to see what is the breed of dog that got a high or low rating. The problem is we can't define how low and high a rating is. So, we set values for each level using five number summary.



```
min        0.000000
25%        1.000000
50%        1.100000
75%        1.200000
max        1.400000
```

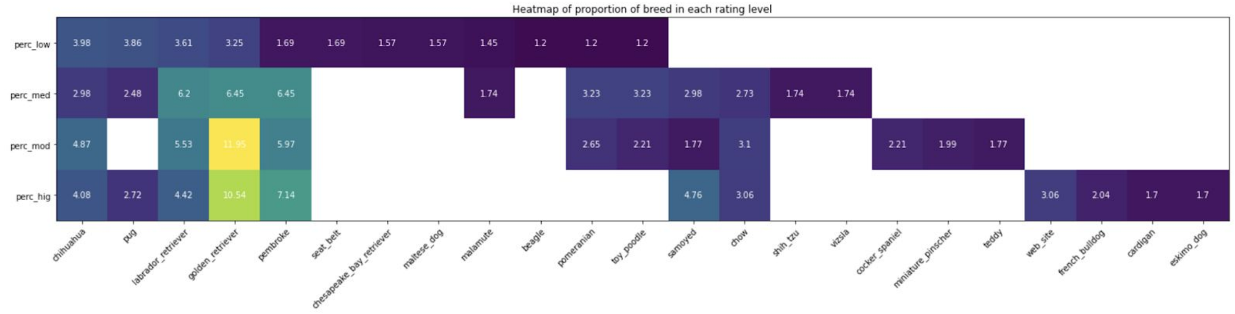
Using those numbers, we define the group as below

- Low rating : 0 - 1
- Medium rating : 1 - 1.1
- Moderately high rating : 1.1 - 1.2
- High rating : 1.2 - 1.4



- 41.9% of tweets are low rated dogs, 20.4% are medium rated dogs, 22.8% are moderately high rated dogs, and 14.9% are high rated dogs

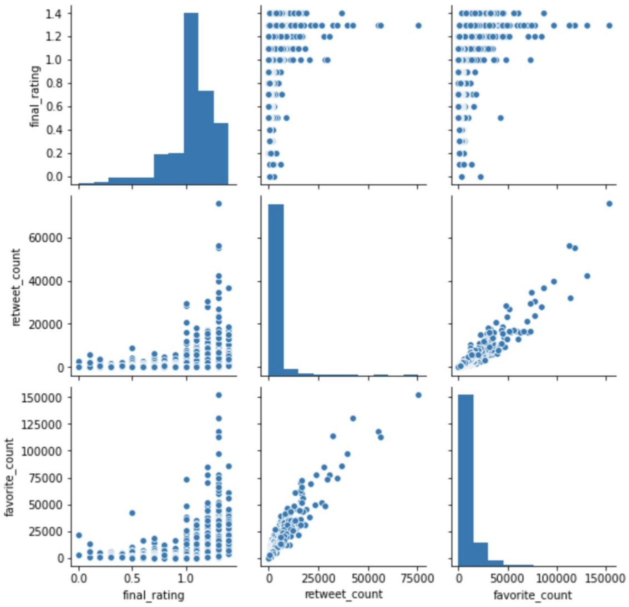
After that, we split the data into each rating group and calculate the proportion of dog's breed within that group

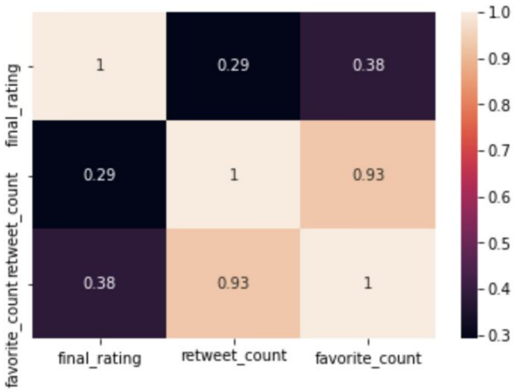


- Generally, golden retriever, labrador retriever, and pembroke as the top 3 of dog's breed are also the top 3 dog's breed for medium, moderately high, and high rating level (except the third rank in high rating level is samoyed)
- Chihuahua and pug are top 2 dog's breed in low rating level
- Since the proportion of low rating level is quite huge, the gap of proportion for each breed as not that far (for example the first rank is chihuahua at 3.98%, the second rank is pug at 3.86%, and so on)
- There are several dog's breeds in each level that are not shown in other levels
    - Seat_belt (not a dog), chesapeake bay retriever, maltese dog, and beagle only appear in low rating level
    - Shih tzu and vizsla only appear in medium rating level

- Cocker spaniel, miniature pinscher, and teddy (not a dog) only appear in moderately high rating level
- Web site (not a dog), french bulldog, cardigan, and eskimo dog only appear in high rating level

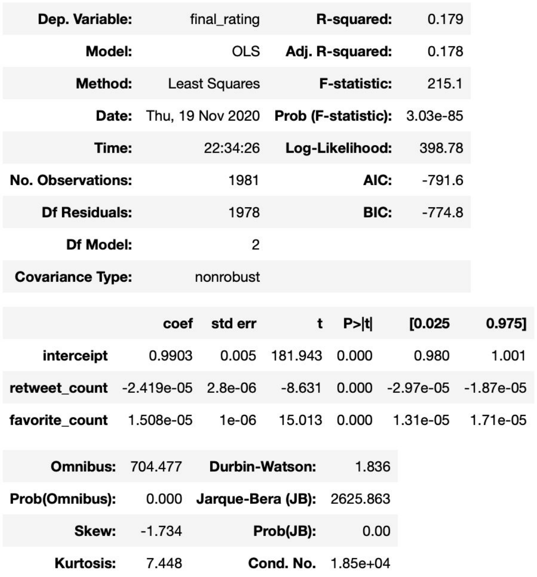
How is the correlation between retweet count, favorite count and final rating?





- All final rating, retweet count, and favorite count are skew.
- Retweet count and favorite count are positively skewed and final rating is negatively skewed.
- The correlation between retweet count and favorite count is strong positive (0.93)

We also tried to build a multilinear regression model using those attributes



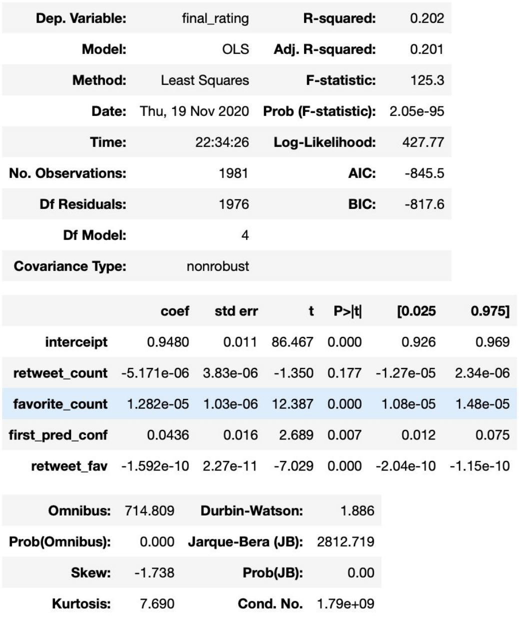
Dep. Variable:	final_rating	**R-squared:**	0.179
Model:	OLS	**Adj. R-squared:**	0.178
Method:	Least Squares	**F-statistic:**	215.1
Date:	Thu, 19 Nov 2020	**Prob (F-statistic):**	3.03e-85
Time:	22:34:26	**Log-Likelihood:**	398.78
No. Observations:	1981	**AIC:**	-791.6
Df Residuals:	1978	**BIC:**	-774.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
interceipt	0.9903	0.005	181.943	0.000	0.980	1.001
retweet_count	-2.419e-05	2.8e-06	-8.631	0.000	-2.97e-05	-1.87e-05
favorite_count	1.508e-05	1e-06	15.013	0.000	1.31e-05	1.71e-05

Omnibus:	704.477	**Durbin-Watson:**	1.836
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2625.863
Skew:	-1.734	**Prob(JB):**	0.00
Kurtosis:	7.448	**Cond. No.**	1.85e+04

- The p-values of retweet count and favorite count are 0. Meaning that those variables are useful in predicting the final rating
- The CIs of retweet count and favorite count are not overlap. There is statistically significant evidence that retweet count differs from favorite count
- The r-square is 0.179 means that 17.9% of the variability in final rating is explained by retweet and favorite count

To make r-squared higher, we tried to include on more variables which are first_pred_conf and retweet_fav (multiplication of retweet count and favorite count)



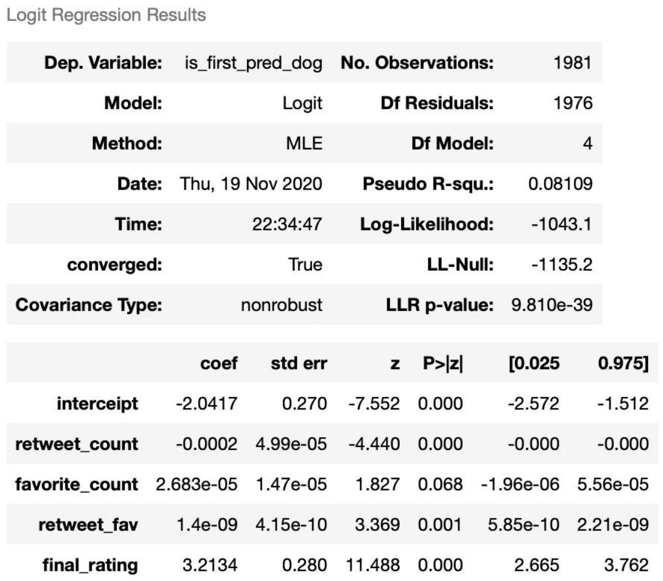
Dep. Variable:	final_rating	**R-squared:**	0.202
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	125.3
Date:	Thu, 19 Nov 2020	**Prob (F-statistic):**	2.05e-95
Time:	22:34:26	**Log-Likelihood:**	427.77
No. Observations:	1981	**AIC:**	-845.5
Df Residuals:	1976	**BIC:**	-817.6
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
interceipt	0.9480	0.011	86.467	0.000	0.926	0.969
retweet_count	-5.171e-06	3.83e-06	-1.350	0.177	-1.27e-05	2.34e-06
favorite_count	1.282e-05	1.03e-06	12.387	0.000	1.08e-05	1.48e-05
first_pred_conf	0.0436	0.016	2.689	0.007	0.012	0.075
retweet_fav	-1.592e-10	2.27e-11	-7.029	0.000	-2.04e-10	-1.15e-10

Omnibus:	714.809	**Durbin-Watson:**	1.886
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2812.719
Skew:	-1.738	**Prob(JB):**	0.00
Kurtosis:	7.690	**Cond. No.**	1.79e+09

- After adding first_pred_conf the r-squared becomes 0.202. 20.2% of the variability in final rating is explained by retweet count, favorite count, retweet fav, and first prediction coefficient
- The p-value of first_pred_conf is 0.007 and retweet_fav is 0, this variables are also helpful in predicting final rating

How is the correlation between retweet count, favorite count, final rating and whether it's dog or not?



Logit Regression Results

Dep. Variable:	is_first_pred_dog	**No. Observations:**	1981
Model:	Logit	**Df Residuals:**	1976
Method:	MLE	**Df Model:**	4
Date:	Thu, 19 Nov 2020	**Pseudo R-squ.:**	0.08109
Time:	22:34:47	**Log-Likelihood:**	-1043.1
converged:	True	**LL-Null:**	-1135.2
Covariance Type:	nonrobust	**LLR p-value:**	9.810e-39

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
interceipt	-2.0417	0.270	-7.552	0.000	-2.572	-1.512
retweet_count	-0.0002	4.99e-05	-4.440	0.000	-0.000	-0.000
favorite_count	2.683e-05	1.47e-05	1.827	0.068	-1.96e-06	5.56e-05
retweet_fav	1.4e-09	4.15e-10	3.369	0.001	5.85e-10	2.21e-09
final_rating	3.2134	0.280	11.488	0.000	2.665	3.762

- Retweet count, retweet fav and final rating are statistically significant in helping to predict whether the image is dog's image or not since the p-values are less than 0.05
- However, favorite count is not statistically significant since the p-value is 0.068

Conclusion

1. How is the proportion of dog's breed?
    a. 6.97% of tweets are showing golden retriever dogs, 4.69% are showing labrador retriever dogs, 4.44% are talking about pembroke dogs, 3.99% are talking about chihuahua
2. What is the breed of dogs that have higher ratings and low ratings?
    a. 41.9% of tweets are low rated dogs, 20.4% are medium rated dogs, 22.8% are moderately high rated dogs, and 14.9% are high rated dogs

b. Generally, golden retriever, labrador retriever, and pembroke as the top 3 of dog’s breed are also the top 3 dog’s breed for medium, moderately high, and high rating level (except the third rank in high rating level is samoyed)

c. Chihuahua and pug are top 2 dog’s breed in low rating level

d. There are several dog’s breeds in each level that are not shown in other levels

3. How is the correlation between retweet count, favorite count and final rating?

   a. The correlation between retweet count and favorite count is strong positive (0.93)

   b. The p-values of retweet count and favorite count are 0. Meaning that those variables are useful in predicting the final rating

   c. The CIs of retweet count and favorite count are not overlap. There is statistically

   d. The p-value of first_pred_conf is 0.007 and retweet_fav is 0, this variables are also helpful in predicting final rating

4. How is the correlation between retweet count, favorite count, final rating and whether it's dog or not?

   a. Retweet count, retweet fav and final rating are statistically significant in helping to predict whether the image is dog’s image or not since the p-values are less than 0.05

   b. However, favorite count is not statistically significant since the p-value is 0.068

Resources

- https://matplotlib.org/3.1.1/gallery/images_contours_and_fields/image_annotated_heatmap.html
- https://www.dataforeverybody.com/matplotlib-seaborn-pie-charts/
- https://datatofish.com/if-condition-in-pandas-dataframe/