Act Report

This document is about key findings regarding the WeRateDogs tweets data. WeRateDogs twitter account is about rating the charm of every dog in the world of twitter. Sure, it has one uniqueness. That is the rating system. The rating system usually out of the range, greater than the range. WeRateDogs scale using 10-valued scale, 1 to 10 with 1 is the least and 10 is the most favorable. The ratings are always above the scale, 11 or 12. The dogs vary in phase – from young to old: pupper puppo doggo flooper. That dog phase distribution is interesting because we can know what is the most mentioned dog_phase. In addition, I interested in the twitter favorite and retweet count associated with every tweet and their correlation.

To recap, these research questions is asked:

1. What is the distribution for dogs rating?

2. What is the highest mentioned dog phase?

3. What is the distribution for retweet count and favorite count dog tweets?

4. How is the correlation between retweet count and favorite count?

For the first question, the distribution of dogs rating is in figure below. The mean for the dog rating is 1.174, which is not surprising because of the rating system. This observation implied that every dog is truly a good dog that deserve out of the scale rating. Noteable abservation is the minimum and the maximum scale. The minimum scale is 0.9 and maximum is 42. This is a huge gap between the rating.

```
count    1678.000000
mean        1.174553
std         1.016830
min         0.900000
25%         1.000000
50%         1.100000
75%         1.200000
max        42.000000
Name: rating, dtype: float64
```

Figure 1 Dogs rating statistics

For the second question, the proportion associated with each dog phase is in figure 2. The highest proportion is pupper. That means pupper dog phase is usually mentioned in the twitter account. The second highest mentioned is the doggo dog phase. This hinted us that dog in early phase, pupper will be mentioned more than other phases. More data is required because the data count to compute this proportion is 292.

pupper	0.623288
doggo	0.236301
puppo	0.075342
floofer	0.030822
doggopupper	0.027397
flooferdoggo	0.003425
doggopuppo	0.003425

Figure 2 Dogs phase proportion

For the thirth question, statistics regarding retweet and favorite count is in figure 3 and 4. The statistic hinted that the mean of favorite count is higher than retweet count. That implied that on average, twitter user is more likely to use favorite button to share the dogs. But, the standard deviation in favorite count is higher than in retweet count. That means favorite count is less predictable than retweeet count.

count	1678.000000
mean	3296.330751
std	5124.572717
min	16.000000
25%	790.500000
50%	1799.000000
75%	3823.500000
max	79515.000000

Figure 3 Retweet statistics

count	1678.000000
mean	10520.373063
std	13059.629762
min	81.000000
25%	2610.000000
50%	5727.500000
75%	13817.250000
max	132810.000000

Figure 4 Favorite statistics

For the fourth question, the correlated data is visualized as below.

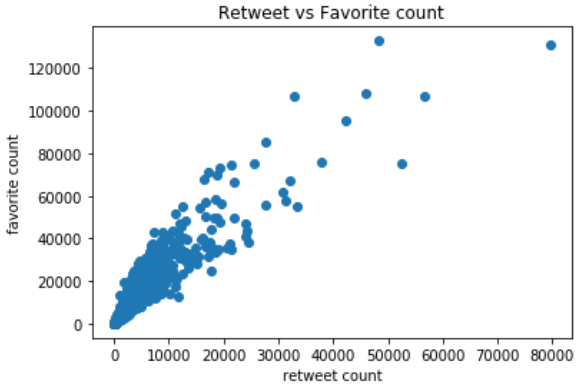

Figure 5 Retweet vs favorite count scatter

The diagram hinted that there is a correlation between these variables. Linear regression model will be used to modelled the linear relationship between theses variables. The linear regression model si shown below.

Dep. Variable:	retweet_count	**R-squared:**	0.826
Model:	OLS	**Adj. R-squared:**	0.826
Method:	Least Squares	**F-statistic:**	7945.
Date:	Sat, 18 Apr 2020	**Prob (F-statistic):**	0.00
Time:	01:14:49	**Log-Likelihood:**	-15247.
No. Observations:	1678	**AIC:**	3.050e+04
Df Residuals:	1676	**BIC:**	3.051e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
favorite_count	0.3566	0.004	89.137	0.000	0.349	0.364
intercept	-455.0969	67.075	-6.785	0.000	-586.657	-323.537

Omnibus:	1635.327	**Durbin-Watson:**	1.256
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	217467.038
Skew:	4.210	**Prob(JB):**	0.00
Kurtosis:	58.131	**Cond. No.**	2.15e+04

Figure 6 Retweet, favorite regression model

The model implied that favorite count is correlated with retweet count. The p-value indicated that favorite count significantly correlated with retweet count. There is a strong and positive relationship between these variables (R = 0.9). Furthermore, 82.6% of retweet count variability is explained by favorite count variability. Which means linear regression model will be adequate to model the variables relationship.