Data Analysis and Visualization for WeRateDogs

Author: Xueyun Zhou

This report illustrates the visualization for the clean dataset of WeRateDogs. This report tries to explain the factors that make a WeRateDogs tweet more favoured by people.

1. Descriptive Statistics

The descriptive statistics of numeric variables are as follows. Please note there is no tweet for the source vine.

Table 1-1 Descriptive Statistics

	rating-numerator	rating-denominator	favorite-count	retweet-count	grade	text_len
count	1993	1993	1993	1993	1993	1993
mean	12.2199	10.5118	8924.6483	2770.5845	1.1646	122.2604
std	41.4712	7.2629	12403.1664	4717.0778	4.0649	26.0331
min	0.0000	10.0000	81.0000	15.0000	0.0000	36.0000
25%	10.0000	10.0000	1969.0000	621.0000	1.0000	105.0000
50%	11.0000	10.0000	4114.0000	1348.0000	1.1000	132.0000
75%	12.0000	10.0000	11278.0000	3203.0000	1.2000	139.0000
max	1776.0000	170.0000	132318.0000	79116.0000	177.6000	167.0000

Table 1-2 Descriptive Statistics for Dog Stages

	doggo	floofer	pupper	puppo
count	1993	1993	1993	1993
mean	0.0401	0.0040	0.1189	0.0146
std	0.1963	0.0632	0.3238	0.1198
min	0.0000	0.0000	0.0000	0.0000
25%	0.0000	0.0000	0.0000	0.0000
50%	0.0000	0.0000	0.0000	0.0000
75%	0.0000	0.0000	0.0000	0.0000
max	1.0000	1.0000	1.0000	1.0000

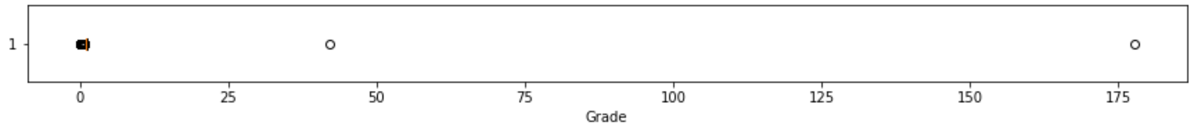
Table 1-3 Descriptive Statistics for Tweet Sources

	web	iphone	vine	tweetdeck
count	1993	1993	1993	1993
mean	0.0140	0.9804	0.0000	0.0055
std	0.1177	0.1385	0.0000	0.0741
min	0.0000	0.0000	0.0000	0.0000
25%	0.0000	1.0000	0.0000	0.0000
50%	0.0000	1.0000	0.0000	0.0000
75%	0.0000	1.0000	0.0000	0.0000
max	1.0000	1.0000	0.0000	1.0000

2. Univariate Analysis

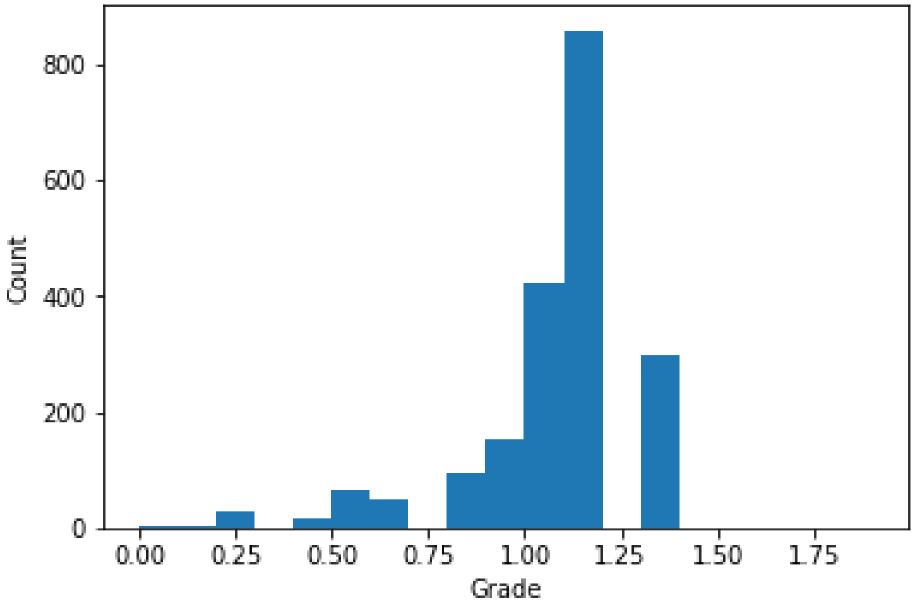
From Chart 1, the distribution of grade centralizes between 0 and 2.

Chart 1 Bloxplot for the grade



From Chart 2, zoom in the chart, it can be seen people tend to give a grade between 1.0 and 1.2.

Chart 2 Histogram for the grades smaller than 2



From Chart 3 and 4, the distributions of favorite_count and retweet_count are right-skewed.

Chart 3 Histogram for favorite_count

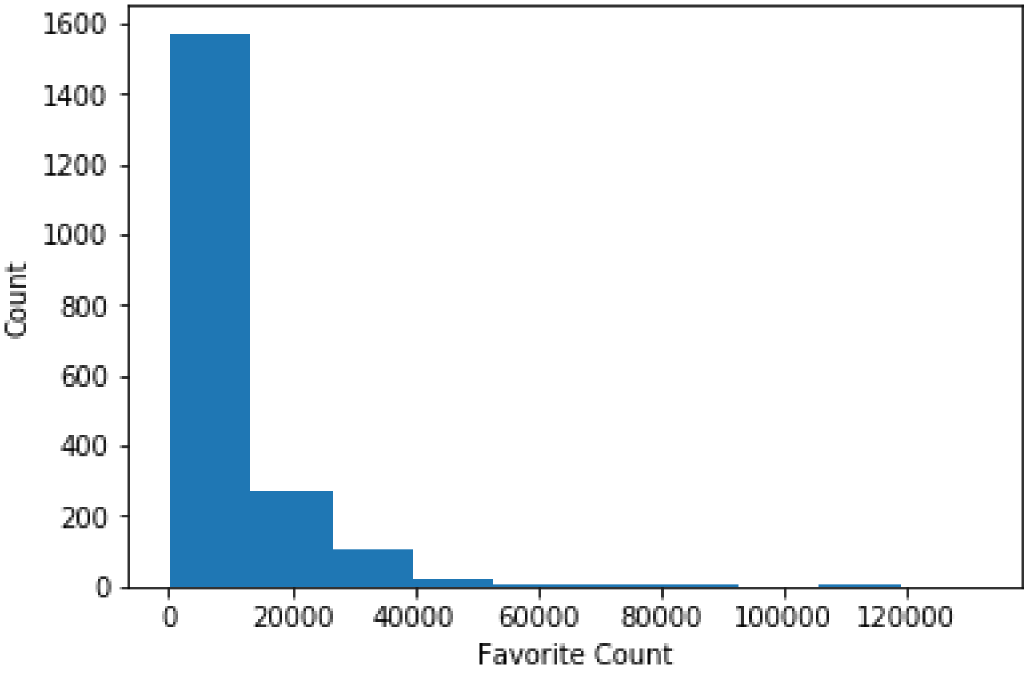
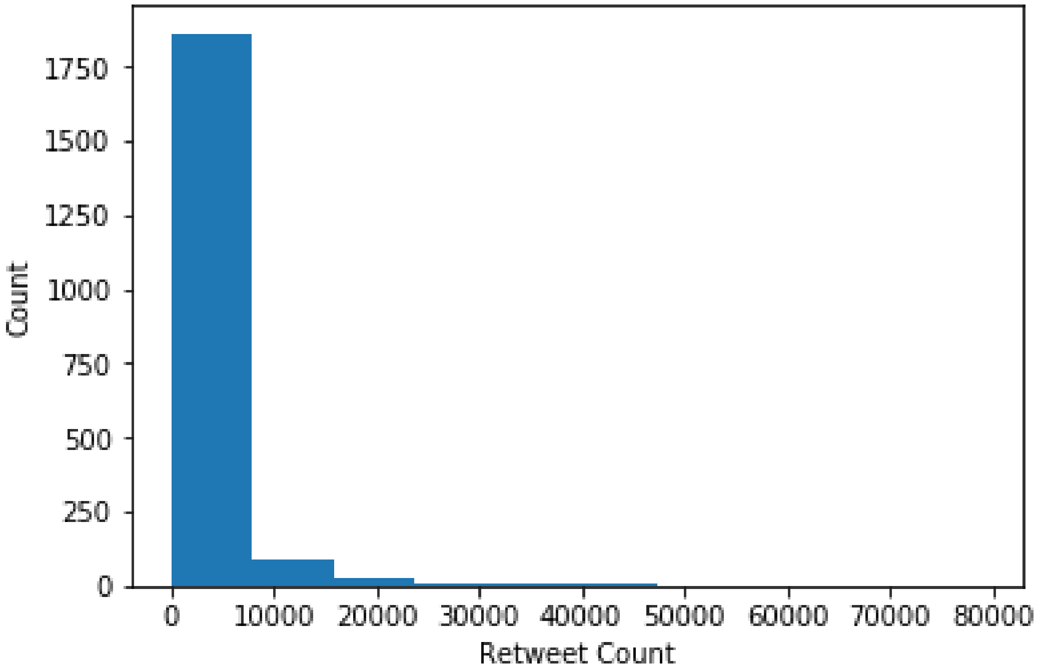
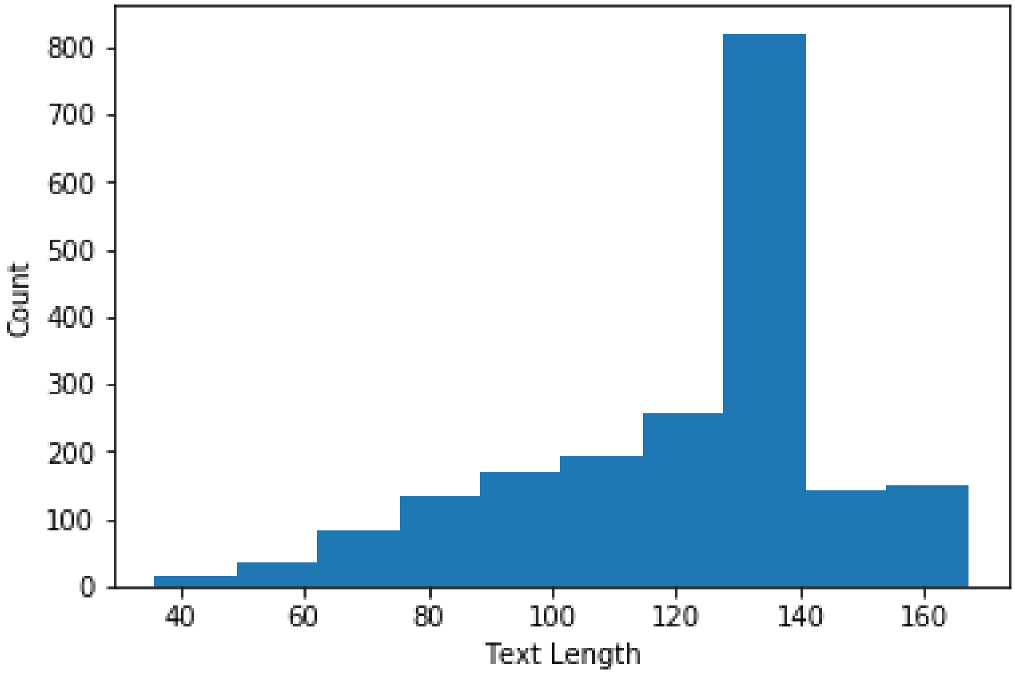


Chart 4 Histogram for retweet_count



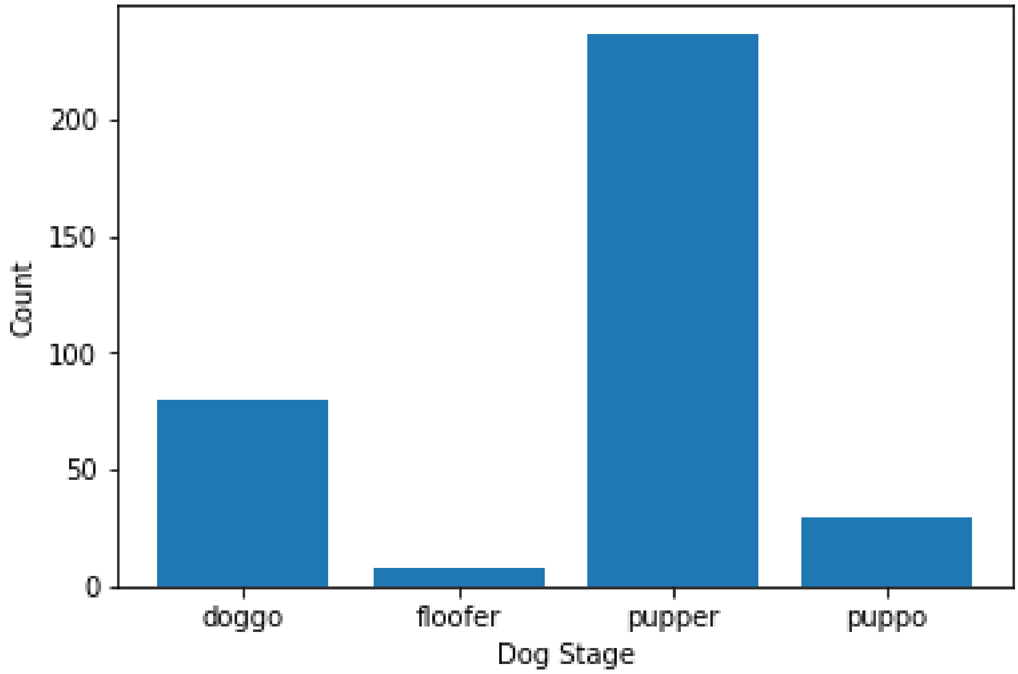
From Chart 5, the distribution of the text length is left-skewed. People tent to write a tweet with the length with 130 to 140.

Chart 5 Histogram for text lengths



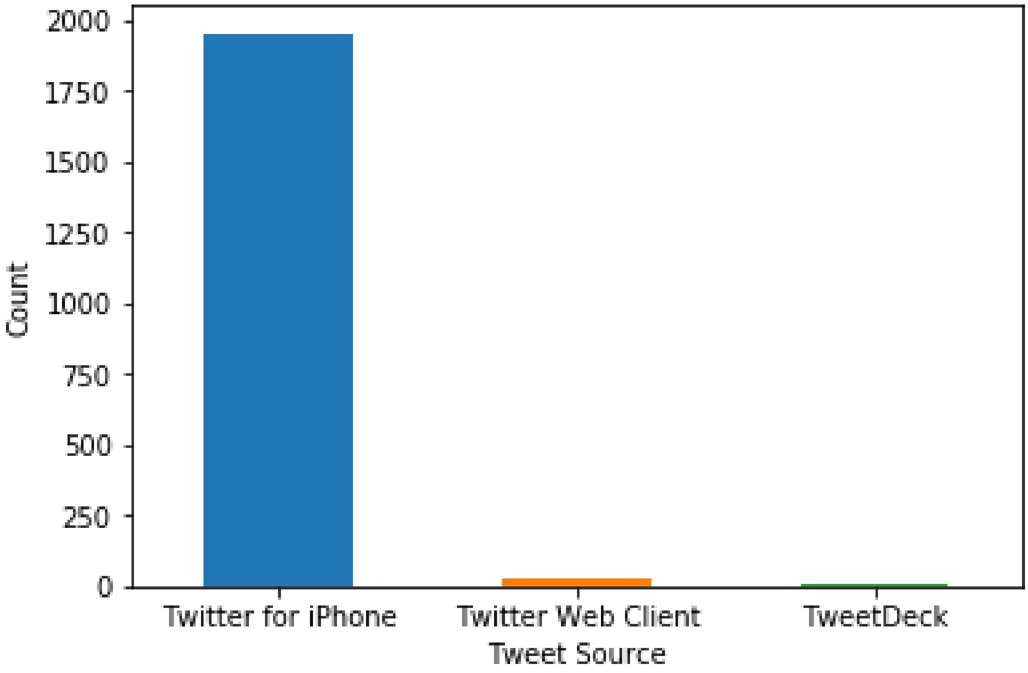
From Chart 6, the pupper is a stage that is mentioned the most.

Chart 6 Bar chart for the dog stage



From Chart 7, people mostly use iPhone to tweet.

Chart 7 Bar chart for tweet sources



3. Bivariate Analysis

From Table 2, the correlations between independent variables (text_len and grade) and dependent variables (favorite_count and retweet_count) are weak with correlation coefficients lower than 0.3.

Table 2 Correlation Coefficients

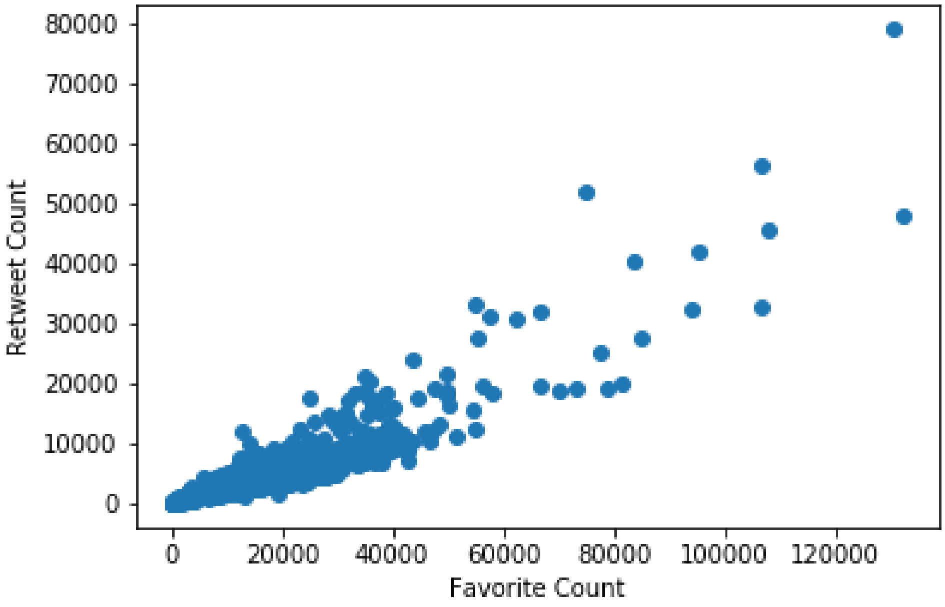
	retweet_count	favorite_count
text_len	0.04818415	0.14616841
grade	0.02342507	0.02247286

4. Conclusions

4.1 Conclusion 1

From Chart 8 and by calculation, the favorite_count and retweet_count have a strong correlation. The correlation coefficient is 0.91501956 and the scatter points concentrate in a straight line. It means the tweets which are given the 'favorate' are likely to be retweeted.

Chart 8 Scatter plot for favorite_count and retweet_count



4.2 Conclusion 2

From Table 3 and 4, the 6 favorite kinds of dogs by both favorite_count and retweet_count are golden_retriever, Labrador_retriever, Pembroke, Chihuahua, Samoyed, and French_bulldog.

Table 3 Top 10 kinds of dogs by favorite_count

Order	dog_type	favorite_count
1	golden_retriever	1950826
2	Labrador_retriever	1269771
3	Pembroke	1036321
4	Chihuahua	756715
5	Samoyed	582082
6	French_bulldog	568978
7	chow	456699
8	cocker_spaniel	413968
9	pug	382463
10	malamute	350710

Table 4 Top 10 kinds of dogs by retweet_count

Order	dog_type	retweet_count
1	golden_retriever	588494
2	Labrador_retriever	405312
3	Pembroke	290602
4	Chihuahua	253916
5	Samoyed	202313
6	French_bulldog	155290
7	cocker_spaniel	147681
8	chow	133512
9	pug	118051
10	toy_poodle	115125

4.3 Conclusion 3

From Table 5, the tweets with longer text lengths and mentioning "doggo" are more likely to be retweeted.

Table 5 OLS Regression Results for retweet_count

Dep. Variable:	retweet_count	**R-squared:**	0.028
Model:	OLS	**Adj. R-squared:**	0.027
Method:	Least Squares	**F-statistic:**	28.82
Date:	Sun, 15 Sep 2019	**Prob (F-statistic):**	4.60e-13
Time:	21:48:28	**Log-Likelihood:**	-19658.
No. Observations:	1993	**AIC:**	3.932e+04
Df Residuals:	1990	**BIC:**	3.934e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	1709.0835	500.533	3.415	0.001	727.461	2690.707
text_len	7.4134	4.008	1.849	0.065	-0.448	15.274
doggo	3864.7821	531.480	7.272	0.000	2822.467	4907.098

Omnibus:	2397.006	**Durbin-Watson:**	1.707
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	324221.524

Skew:	6.240	**Prob(JB):**	0.00
Kurtosis:	64.225	**Cond. No.**	637.

4.4 Conclusion 4

From Table 6, the tweets with longer text lengths and mentioning "doggo" are more likely to be given a "favorite". It seems people do not like a tweet mentioning "pupper". The grade in the tweet does not contribute significantly for the "favorite".

Table 6 OLS Regression Results for favorite_count

Dep. Variable:	favorite_count	**R-squared:**	0.044
Model:	OLS	**Adj. R-squared:**	0.042
Method:	Least Squares	**F-statistic:**	23.07
Date:	Sun, 15 Sep 2019	**Prob (F-statistic):**	1.17e-18
Time:	21:50:21	**Log-Likelihood:**	-21568.
No. Observations:	1993	**AIC:**	4.315e+04
Df Residuals:	1988	**BIC:**	4.317e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	563.7838	1324.119	0.426	0.670	-2033.022	3160.590

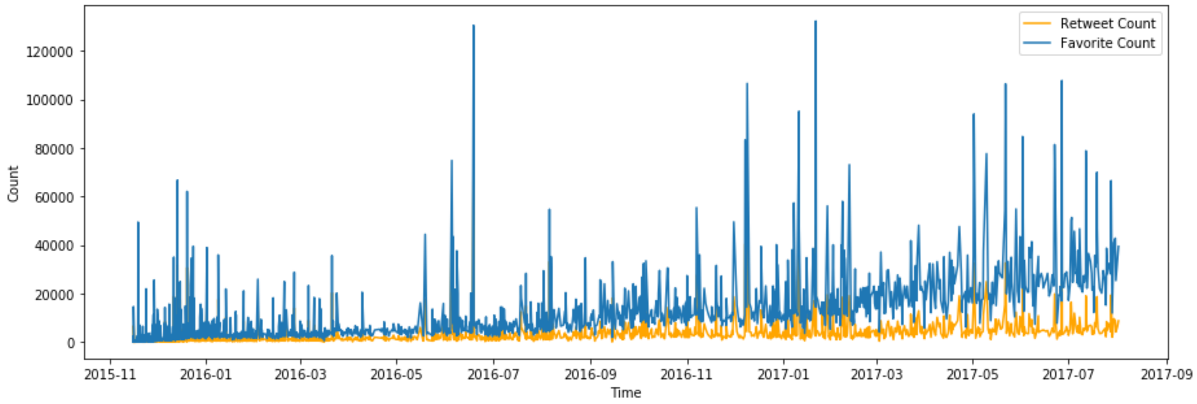
grade	87.8433	66.984	1.311	0.190	-43.522	219.209
text_len	66.1722	10.484	6.312	0.000	45.612	86.733
doggo	9008.7419	1386.485	6.498	0.000	6289.626	1.17e+04
pupper	-1625.4863	841.146	-1.932	0.053	-3275.106	24.133

Omnibus:	1643.454	**Durbin-Watson:**	1.234
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	47023.880
Skew:	3.754	**Prob(JB):**	0.00
Kurtosis:	25.581	**Cond. No.**	638.

5. Visualization

Chart 9 is a line chart for the retweet count and the favorite count over time.

Chart 9 Tendency of retweet-count and favorite-count over time



From the chart it can be seen both counts go up over time. The retweet count increases steadily, and the range of increase is lower than the favorite count. The favorite count rises greatly with wild fluctuations.

It is supposed the large fluctuations appear when there are tweets that appeal to people. By contrast, people prefer giving a 'favorate' to retweeting.