OLS Regression Results

Dep. Variable:	rating	**R-squared:**	0.215
Model:	OLS	**Adj. R-squared:**	0.211
Method:	Least Squares	**F-statistic:**	67.17
Date:	Wed, 15 May 2019	**Prob (F-statistic):**	1.10e-97
Time:	20:18:12	**Log-Likelihood:**	429.13
No. Observations:	1975	**AIC:**	-840.3
Df Residuals:	1966	**BIC:**	-790.0
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	0.9060	0.013	68.894	0.000	0.880	0.932
doggo	0.0679	0.023	2.891	0.004	0.022	0.114
floofer	0.1100	0.069	1.589	0.112	-0.026	0.246
pupper	0.0253	0.014	1.776	0.076	-0.003	0.053
puppo	0.0364	0.041	0.880	0.379	-0.045	0.117
img_num	0.0480	0.008	5.965	0.000	0.032	0.064
p1_conf	0.0298	0.017	1.798	0.072	-0.003	0.062
retweet_count	-1.885e-05	2.3e-06	-8.182	0.000	-2.34e-05	-1.43e-05
favorite_count	1.327e-05	8.87e-07	14.963	0.000	1.15e-05	1.5e-05

Omnibus:	711.792	**Durbin-Watson:**	1.908
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2818.321
Skew:	-1.733	**Prob(JB):**	0.00
Kurtosis:	7.716	**Cond. No.**	2.52e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.52e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Importing statsmodels as library in python, calculated OLS model for the data.

No stage is used as baseline for dummy variables

By holding other stage as 0 doggo, floofer, pupper, puppo stage has positive effect on rating. However, puppo and floofer has too high p-value to pass the 0.1 confidence level.

Img_num and p1_conf has positive effect on rating.

Retweet_count has negative effect and favorite_count has positive effect on rating. However, these two have strong multicollinearity problem.

OLS Regression Results

Dep. Variable:	rating	**R-squared:**	0.058
Model:	OLS	**Adj. R-squared:**	0.055
Method:	Least Squares	**F-statistic:**	20.13
Date:	Wed, 15 May 2019	**Prob (F-statistic):**	5.83e-23
Time:	20:36:19	**Log-Likelihood:**	249.33
No. Observations:	1975	**AIC:**	-484.7
Df Residuals:	1968	**BIC:**	-445.5
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	0.9349	0.014	65.373	0.000	0.907	0.963
doggo	0.1235	0.025	4.877	0.000	0.074	0.173
floofer	0.1321	0.076	1.744	0.081	-0.016	0.281
pupper	0.0097	0.016	0.624	0.533	-0.021	0.040
puppo	0.1317	0.045	2.931	0.003	0.044	0.220
img_num	0.0678	0.009	7.750	0.000	0.051	0.085
p1_conf	0.0499	0.018	2.759	0.006	0.014	0.085

Omnibus:	558.375	**Durbin-Watson:**	1.592
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1553.457
Skew:	-1.471	**Prob(JB):**	0.00
Kurtosis:	6.198	**Cond. No.**	27.4

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

retweet_count and favorite_count variables were dropped to eliminate multicollinearity problems.

Using no stage as a baseline for stage dummy variables, doggo, floofer, pupper and puppo has positive effects on rating. However, pupper stage had too high p-value to pass 0.1 confidence level.

Img_num and p1_conf also has positive effect on rating.

OLS Regression Results

Dep. Variable:	rating	**R-squared:**	0.058
Model:	OLS	**Adj. R-squared:**	0.055
Method:	Least Squares	**F-statistic:**	24.09
Date:	Wed, 15 May 2019	**Prob (F-statistic):**	1.35e-23
Time:	20:40:16	**Log-Likelihood:**	249.14
No. Observations:	1975	**AIC:**	-486.3
Df Residuals:	1969	**BIC:**	-452.7
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	0.9357	0.014	65.678	0.000	0.908	0.964
doggo	0.1237	0.025	4.884	0.000	0.074	0.173
floofer	0.1311	0.076	1.731	0.084	-0.017	0.280
puppo	0.1306	0.045	2.909	0.004	0.043	0.219
img_num	0.0679	0.009	7.767	0.000	0.051	0.085
p1_conf	0.0501	0.018	2.772	0.006	0.015	0.086

Omnibus:	560.889	**Durbin-Watson:**	1.593
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1567.437
Skew:	-1.476	**Prob(JB):**	0.00
Kurtosis:	6.215	**Cond. No.**	27.3

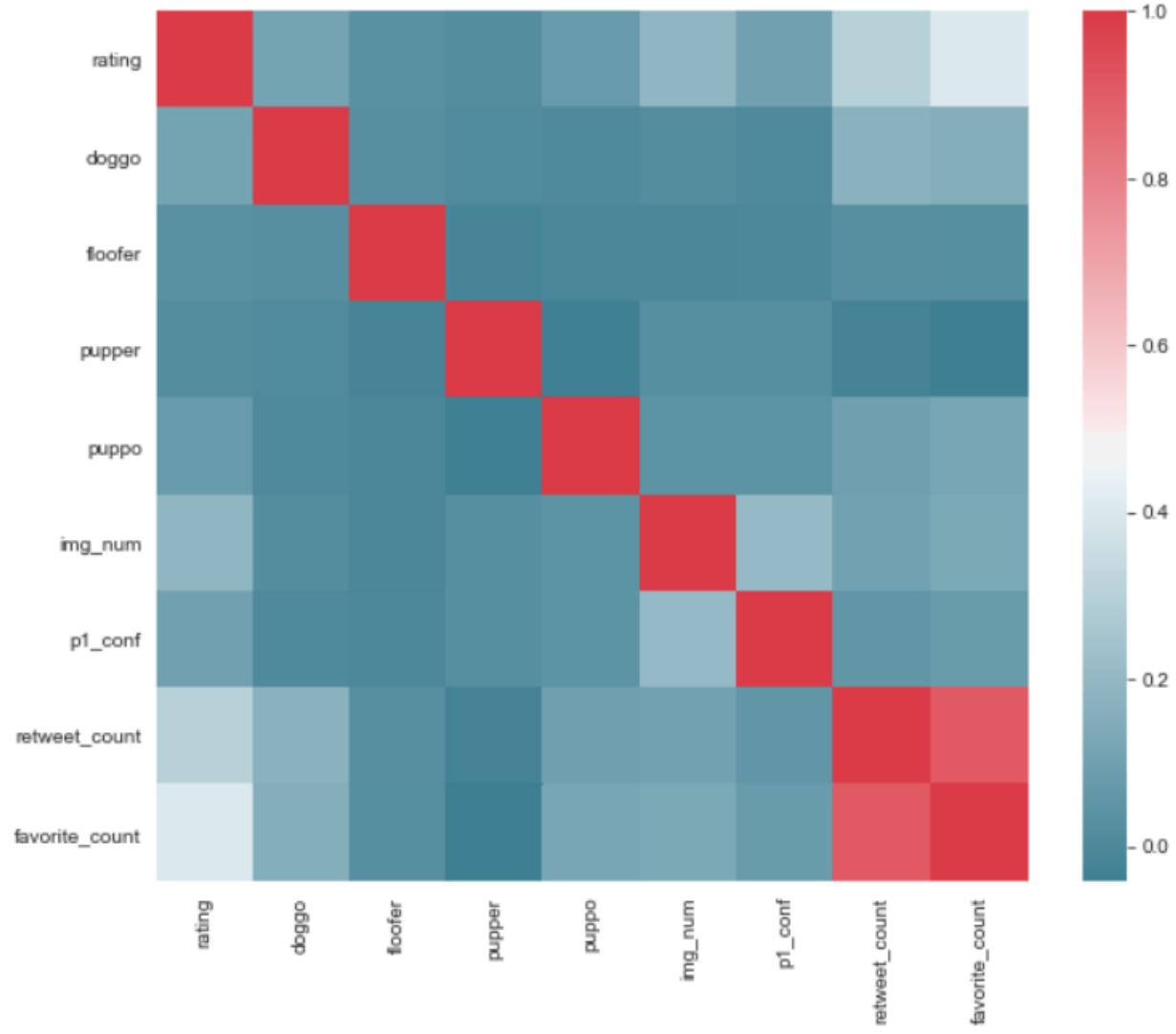
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

By using backward elimination on the model, pupper stage were removed.

Floofer stage had highest positive effect on rating than other stage puppo were second and doggo were last.

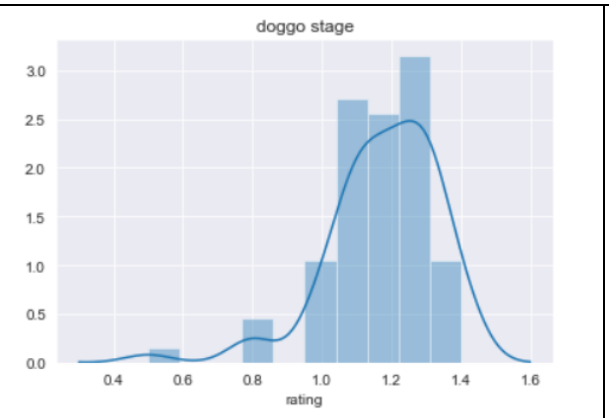
As image number increase there were higher rating and also best prediction confidence.

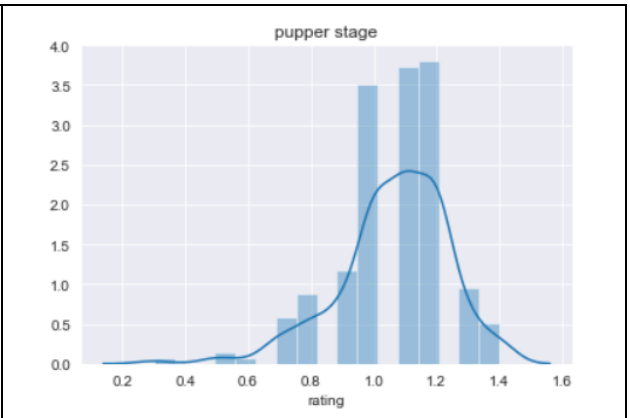
Due to high multicollinearity and p-value, retweet_count, favorite_count and pupper variables were removed.

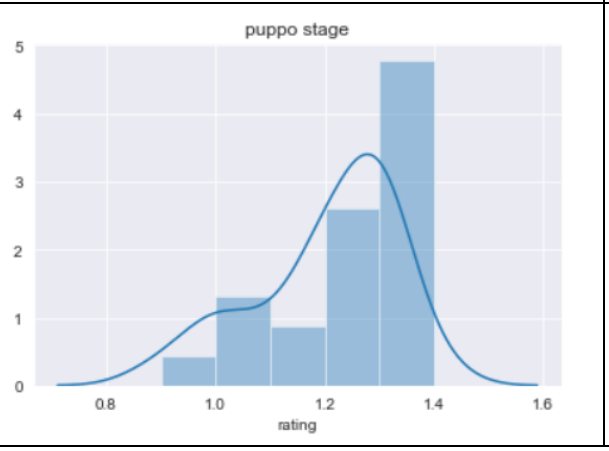


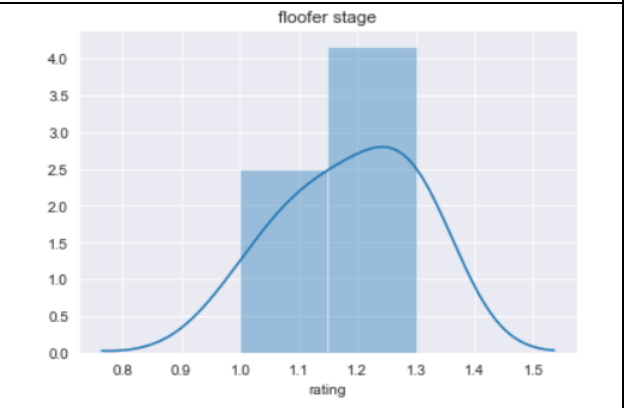
Correlation matrix were plotted on the project.

As discussed in OLS model, there is high multicollinearity problems on retweet_count and favorite_count varibles.









Floofer stage had highest positive effect on rating on our OLS model. However, floofer stage had only 8 rows. It is not accurate to predict with such a small variance.

Puppo stage were second highest rating on the OLS model. However, puppo stage also had low observations 23 rows.

Doggo stage recorded lowest rating on the OLS model. Doggo stage had 74 observation which is still low.

Pupper stage had highest observation 211 row. However, it was dropped out of our OLS model because of high p-value.

In conclusion, it is not accurate to predict rating with stage correlation.
There has been bias variance trade of in stage.