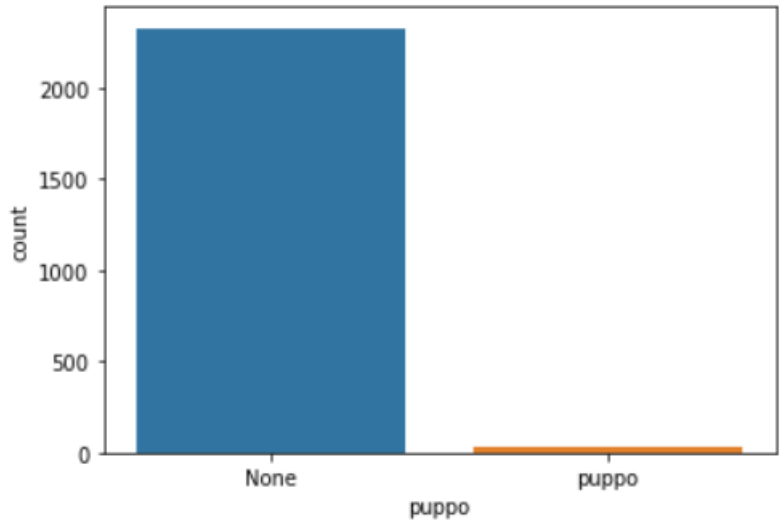
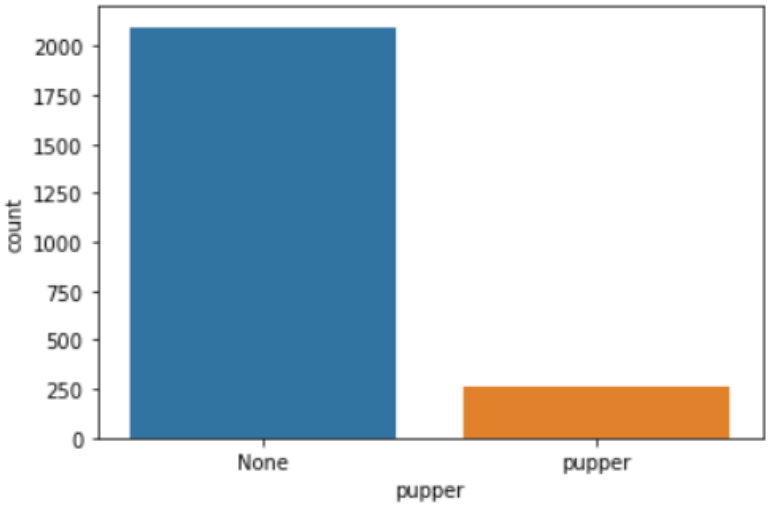
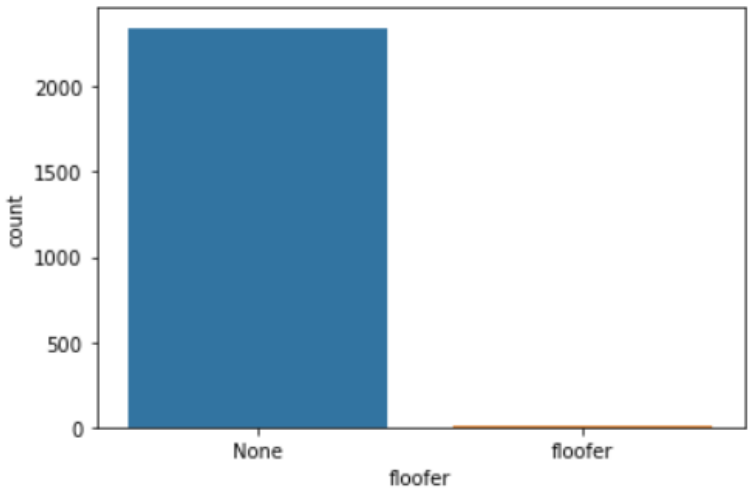
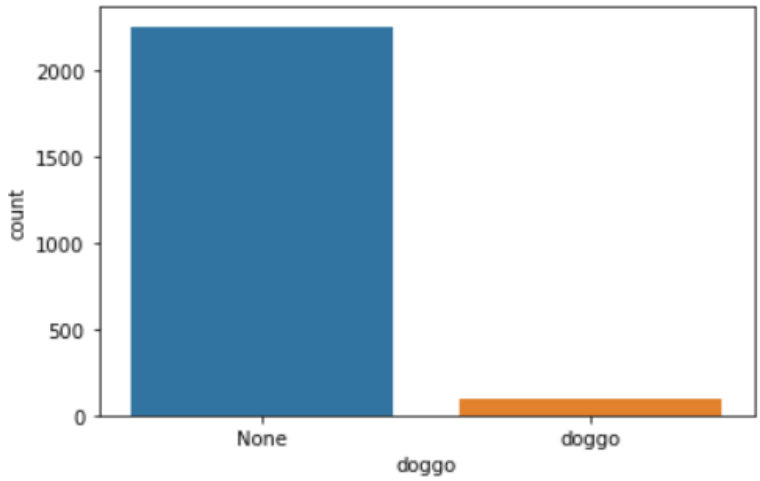
Insights, Analysis and Visualization(s) produced from the Wrangled Data - Report

- Firstly, we can see that there are many rows with 'None' listed for the stage of dog, and so when using this column we should be very cautious of any analysis (such as trying to find which stage of life dogs are usually posted there on) as the majority of values are not useful data, and so making conclusions based on those filled in









- We can see below from the frequency distribution that the favourite count is heavily left-skewed, with most results being less than 5,000, and the median at 3603. However, there are a large number of outliers in the top quartile as we can see from the box plot which cause the mean to be such much higher at 8,081.

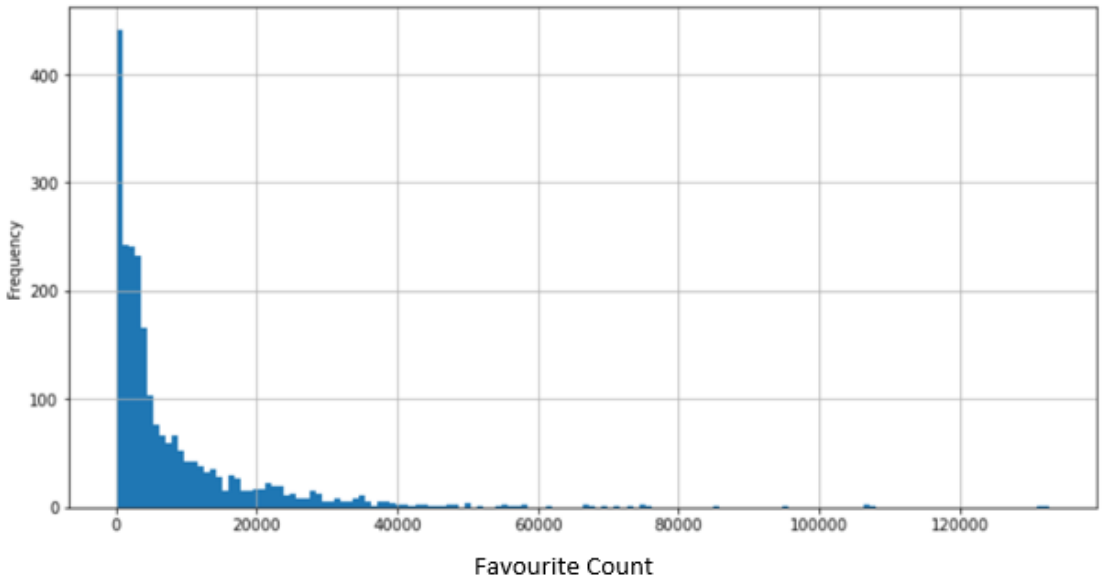
```
tweets_users_df['favorite_count'].median()
```

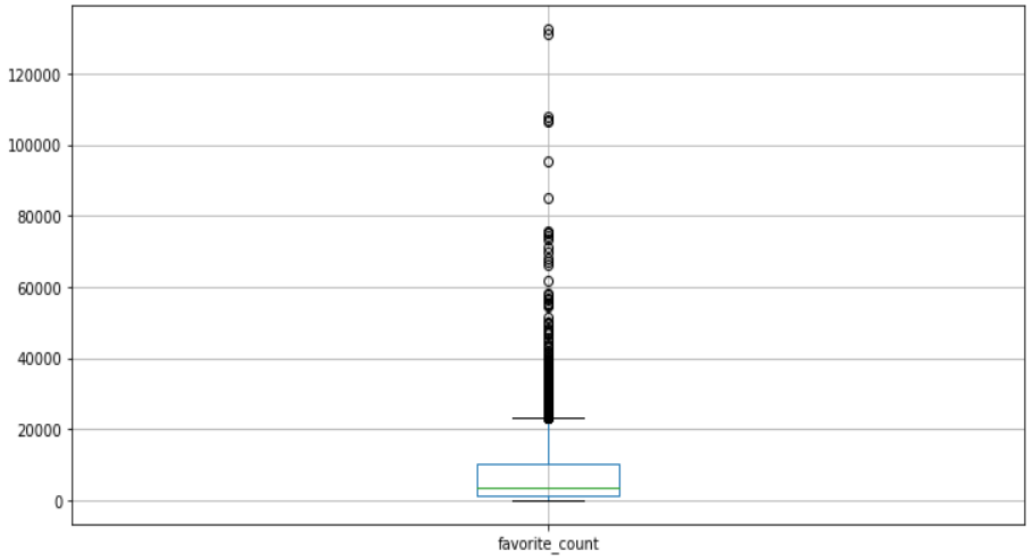
```
3603.5
```

```
tweets_users_df['favorite_count'].mean()
```

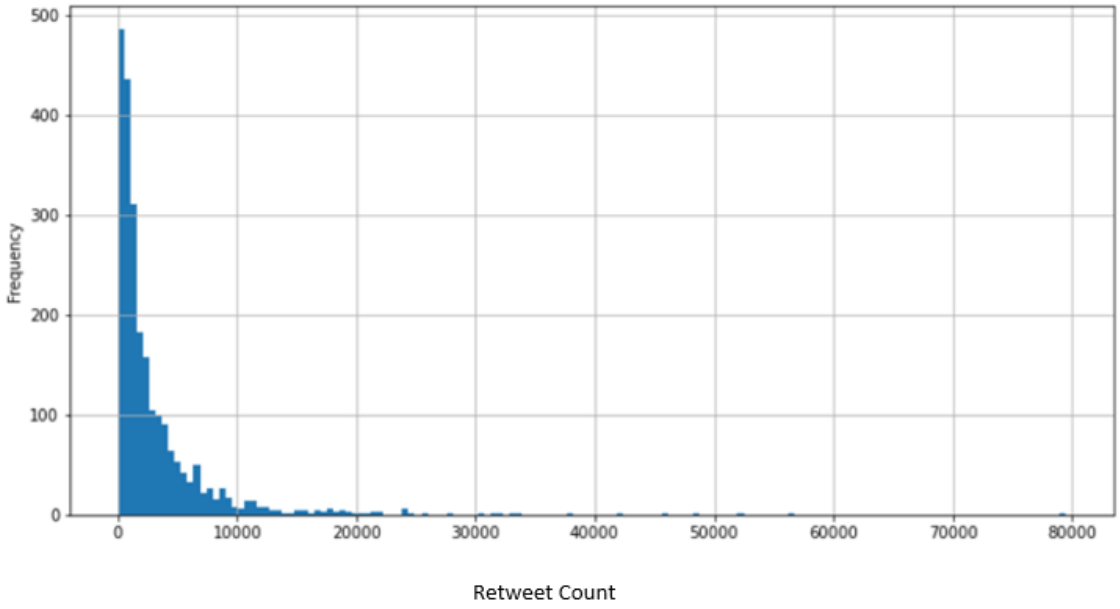
```
8080.968564146135
```

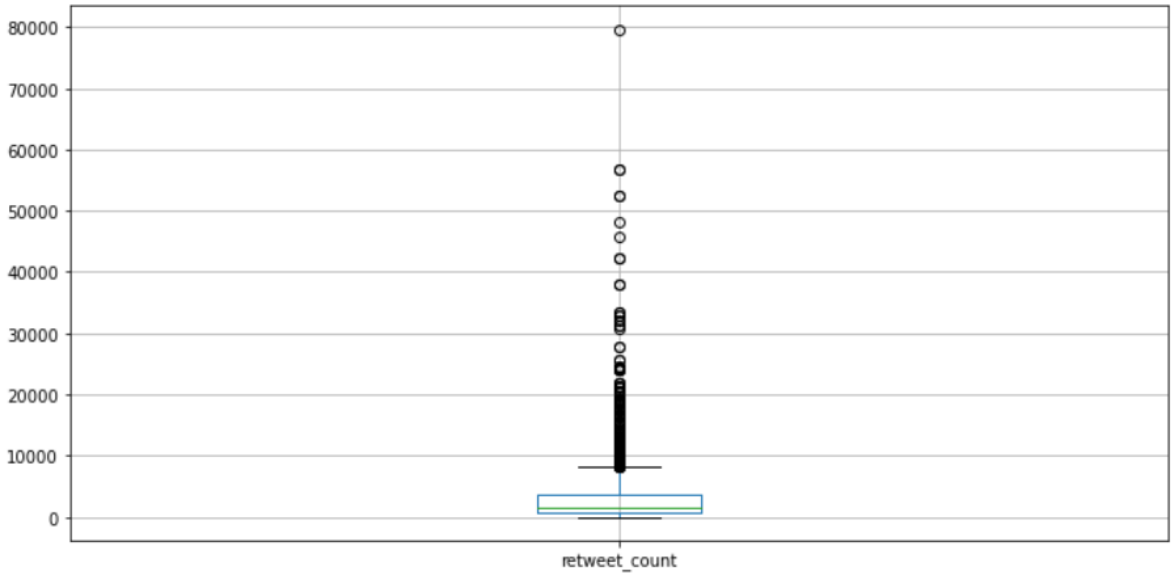
Favourite Count



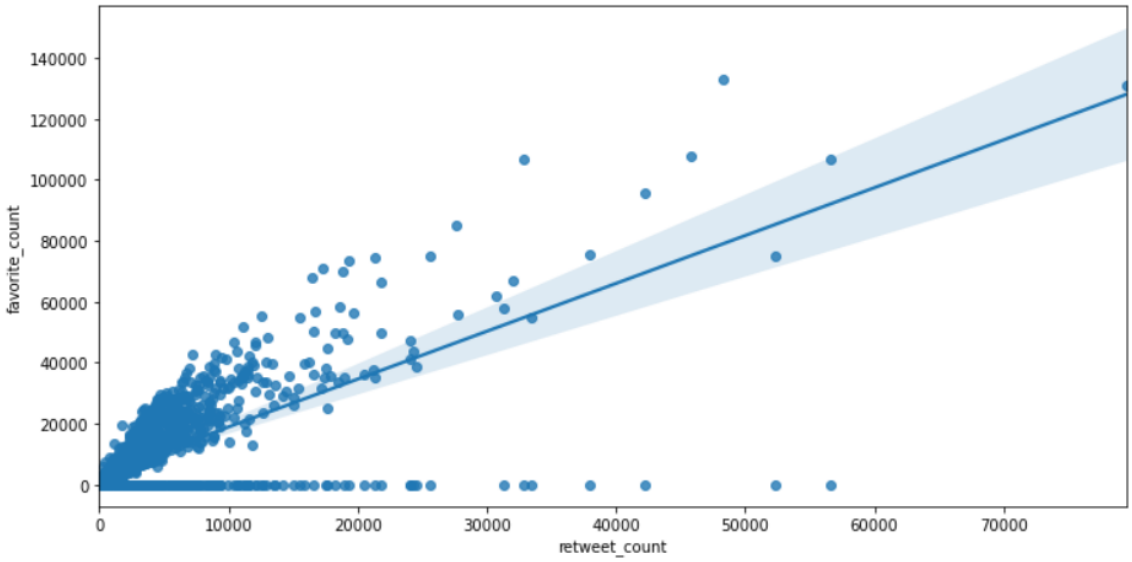


- We can see a very similar pattern with Retweets from the above frequency distribution and the boxplot that the favourite count is heavily left-skewed, with most results being less than 5,000, and a large number of outliers in the top quartile as we can see from the box plot





- We can see from the graph below that there is a linear relationship between the number of times a tweet has been favourited, and the number of retweets it has**.**



- We can see from the OLS/Linear Regression analysis below, from the Adj. R-squared number of 0.494 that this is likely a significant factor that more retweets will cause more favourites. The low P number indicates it is statistically likely to be a cause of movement/ the null hypothesis is false (I.e the coefficient is 0 and therefore not causing a change by increasing or decreasing the independent variable) - however as the condition number is so large, there may also be problems with high correlation between the factors essentially confusing the results.

OLS Regression Results

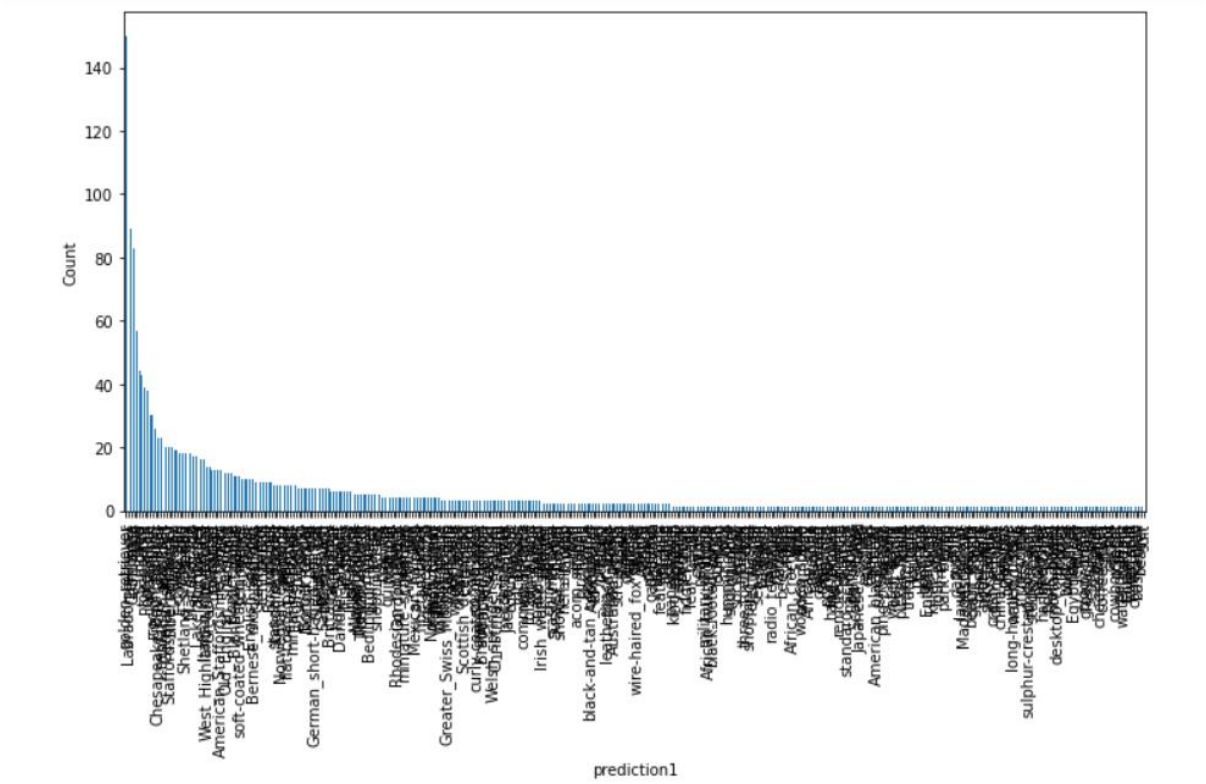
Dep. Variable:	favorite_count	**R-squared:**	0.494
Model:	OLS	**Adj. R-squared:**	0.494
Method:	Least Squares	**F-statistic:**	2297.
Date:	Sat, 19 Jun 2021	**Prob (F-statistic):**	0.00
Time:	16:10:59	**Log-Likelihood:**	-24611.
No. Observations:	2354	**AIC:**	4.923e+04
Df Residuals:	2352	**BIC:**	4.924e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3107.8692	201.951	15.389	0.000	2711.849	3503.889
retweet_count	1.5714	0.033	47.923	0.000	1.507	1.636

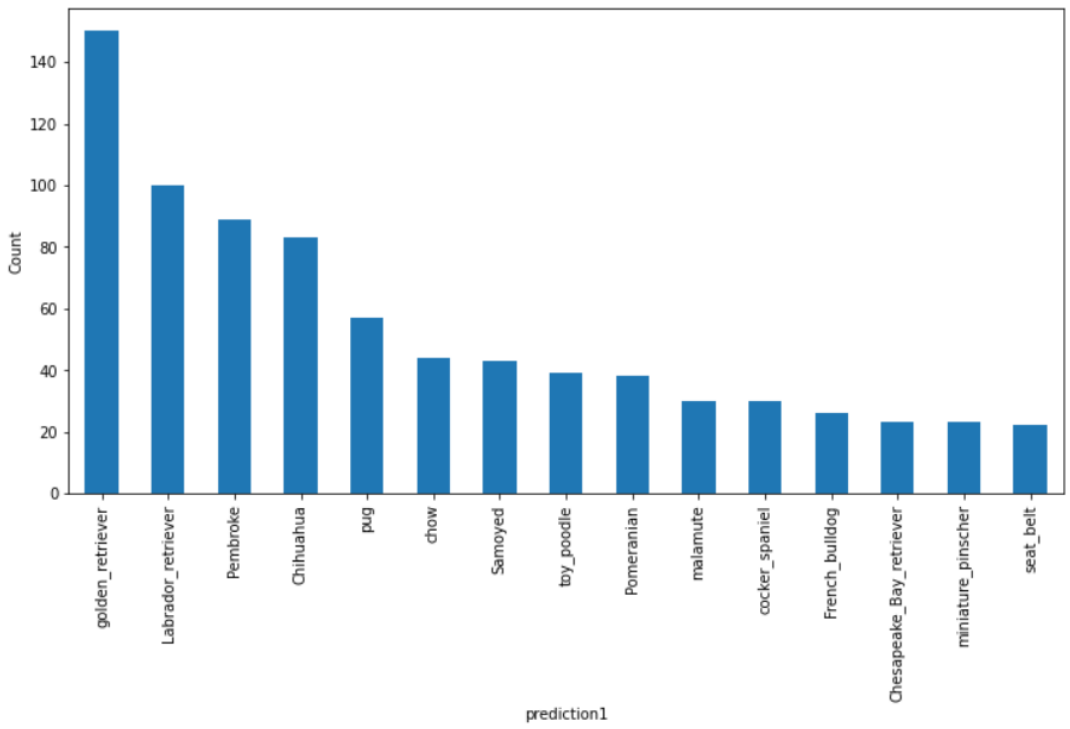
Omnibus:	1034.735	**Durbin-Watson:**	1.655
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	42336.254
Skew:	-1.368	**Prob(JB):**	0.00
Kurtosis:	23.595	**Cond. No.**	7.18e+03

- The ‘Prediction1’ column predicts what kind of dog the tweet is from the image, so I used this as a proxy to measure the number of different types of dogs on there. We can see below from the many values that the prediction algorithm is guessing things other than dogs (such as

'ice lolly' and 'seat_belt' which affects our analysis trying to compare how people reacted to different dogs - this a data quality issue.



- So I removed those that had less than 50 predictions to get the most tweeted about dogs We can see that the top 3 dogs are Golden retreivers, Labradors and Pembrokes, closely followed by Chihuanhua's. However this is assuming the algorith is correct so the results should be taken with caution.



- Below we can see a heatmap showing the correlations of different numerical columns with each other. We can see that the Tweet's Retweet and Favourite Count are highly positively correlated at 0.7. There are no other strong correlation suggestions from this heatmap.