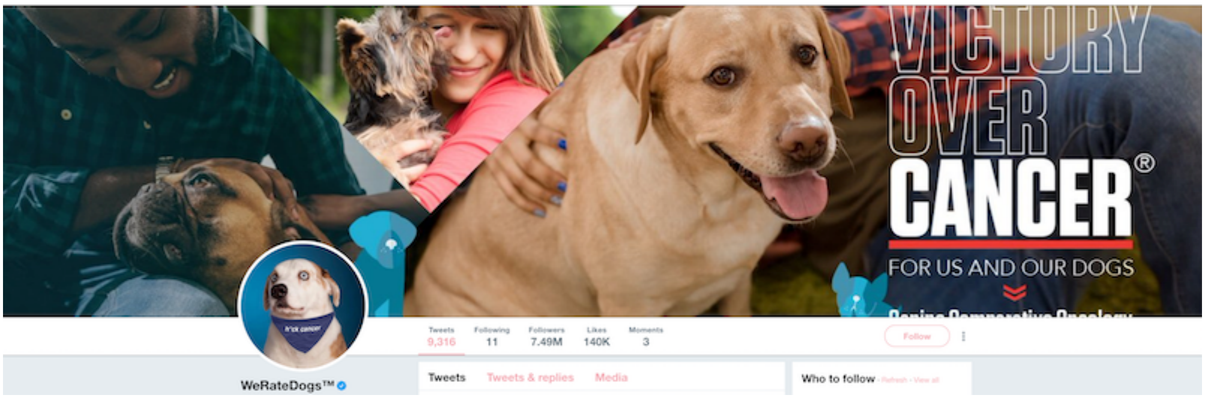
UDACITY

Data Analyst NanoDegree Program

Project #4: Wrangle and Analyse Data

by Samuel Jiménez Sanabria

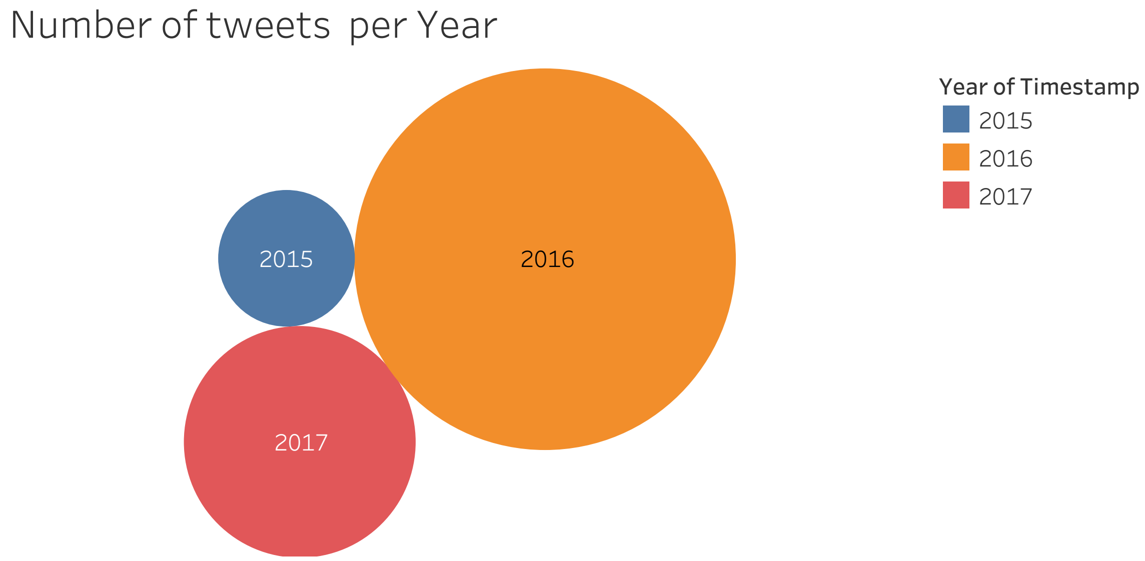


Your Only Source For Professional Dog Ratings

https://twitter.com/dog_rates

Act Report

For the visualisations in this report I've used Tableau. The data visualised comes from the **tweeter_archive_master.csv**



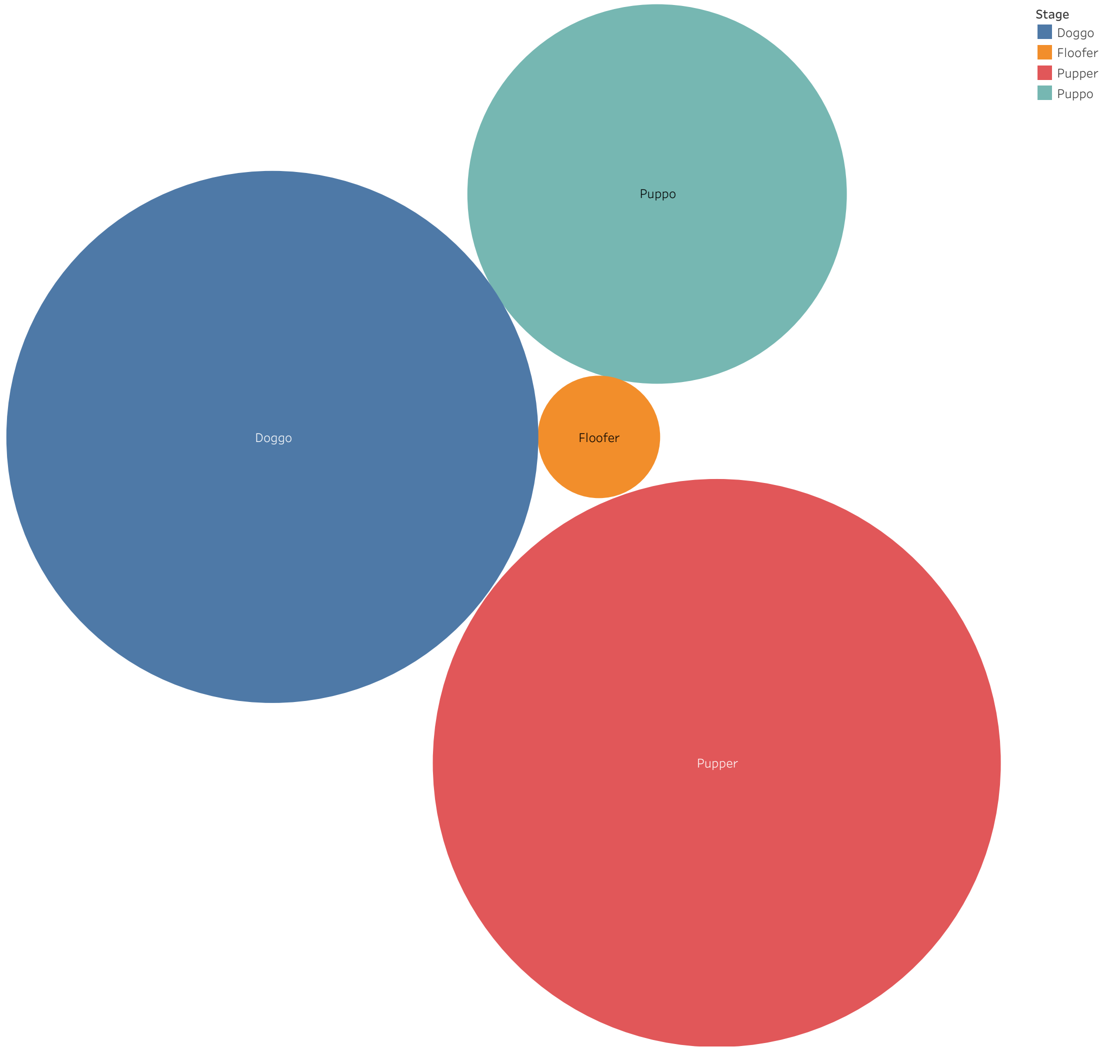
Timestamp Year. Color shows details about Timestamp Year. Size shows sum of Number of Records. The marks are labeled by Timestamp Year.

From the data gathered we can observe that 2016 was the year with the biggest amount of tweets from the WeRateDogs tweeter account.

Christy Westbrook @ChristinaWestb3 · Sep 3
Vinny, you made my day #weratedogs #goldenretriever #goldenmom

WeRateDogs™ @dog_rates
This is Vinny. He's a single father. Thinks the greatest gift he can give his children is believing in them. 13/10 you're doing amazing, Vinny

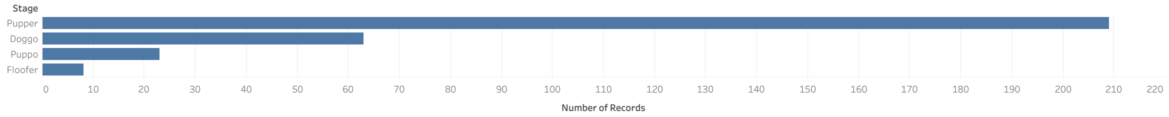
Most Favorited Stages



Stage. Color shows details about Stage. Size shows sum of Favorite Count. The marks are labeled by Stage. The view is filtered on Stage, which keeps Doggo, Floofer, Pupper and Puppo.

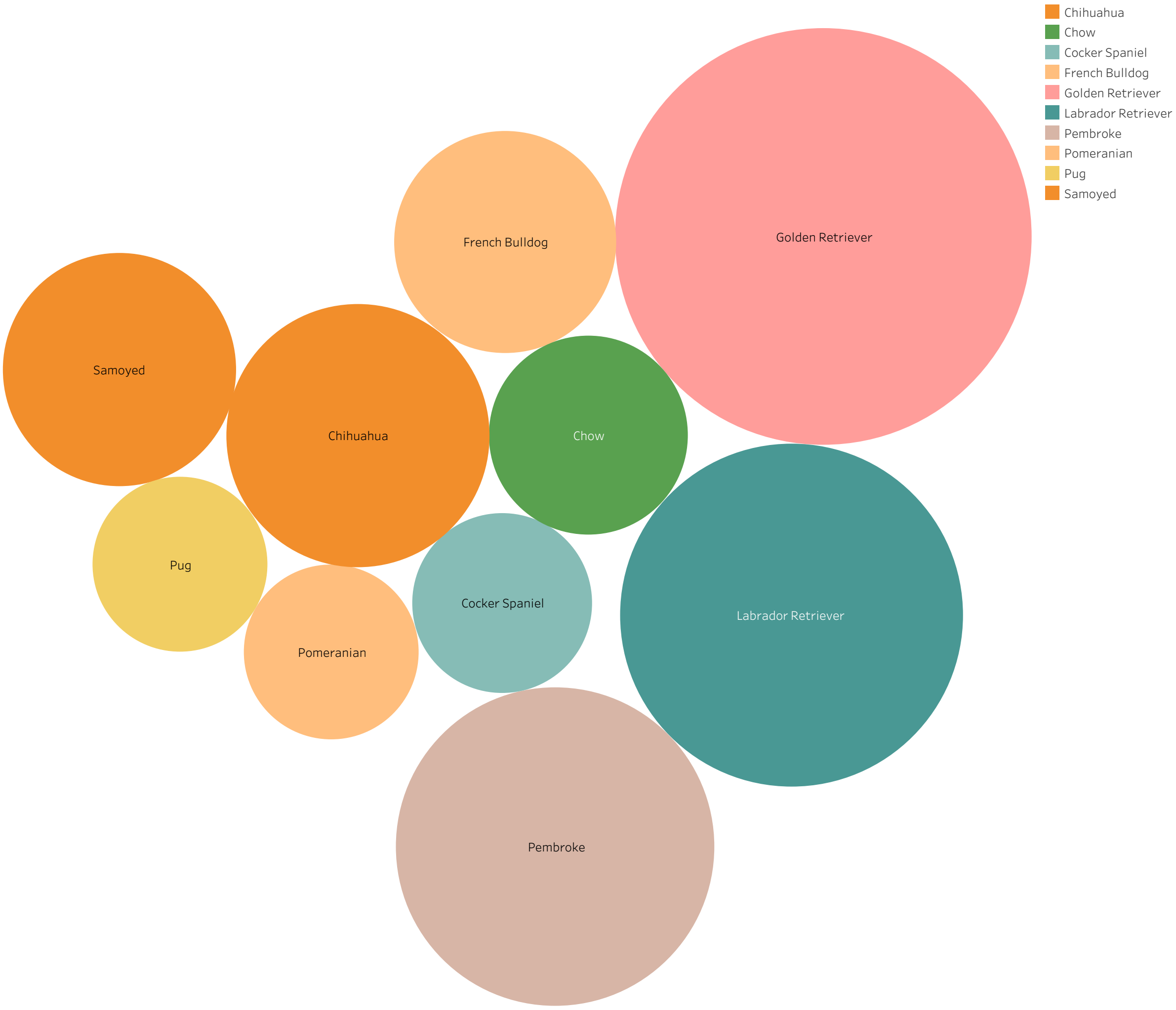
The most favorited stage was pupper, followed by doggo, puppo and lastly floofer. These are also the most common stages as we can observe below.

Most common stages



Sum of Number of Records for each Stage. The view is filtered on Stage, which excludes None and Undefined.

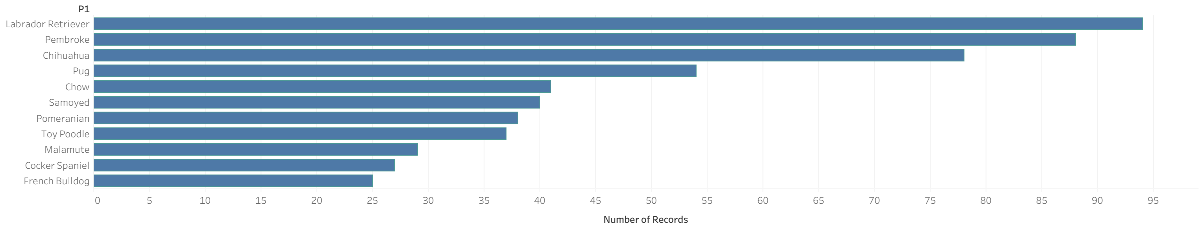
Most favorited breeds



P1. Color shows details about P1. Size shows sum of Favorite Count. The marks are labeled by P1. The view is filtered on P1, which keeps 10 of 293 members.

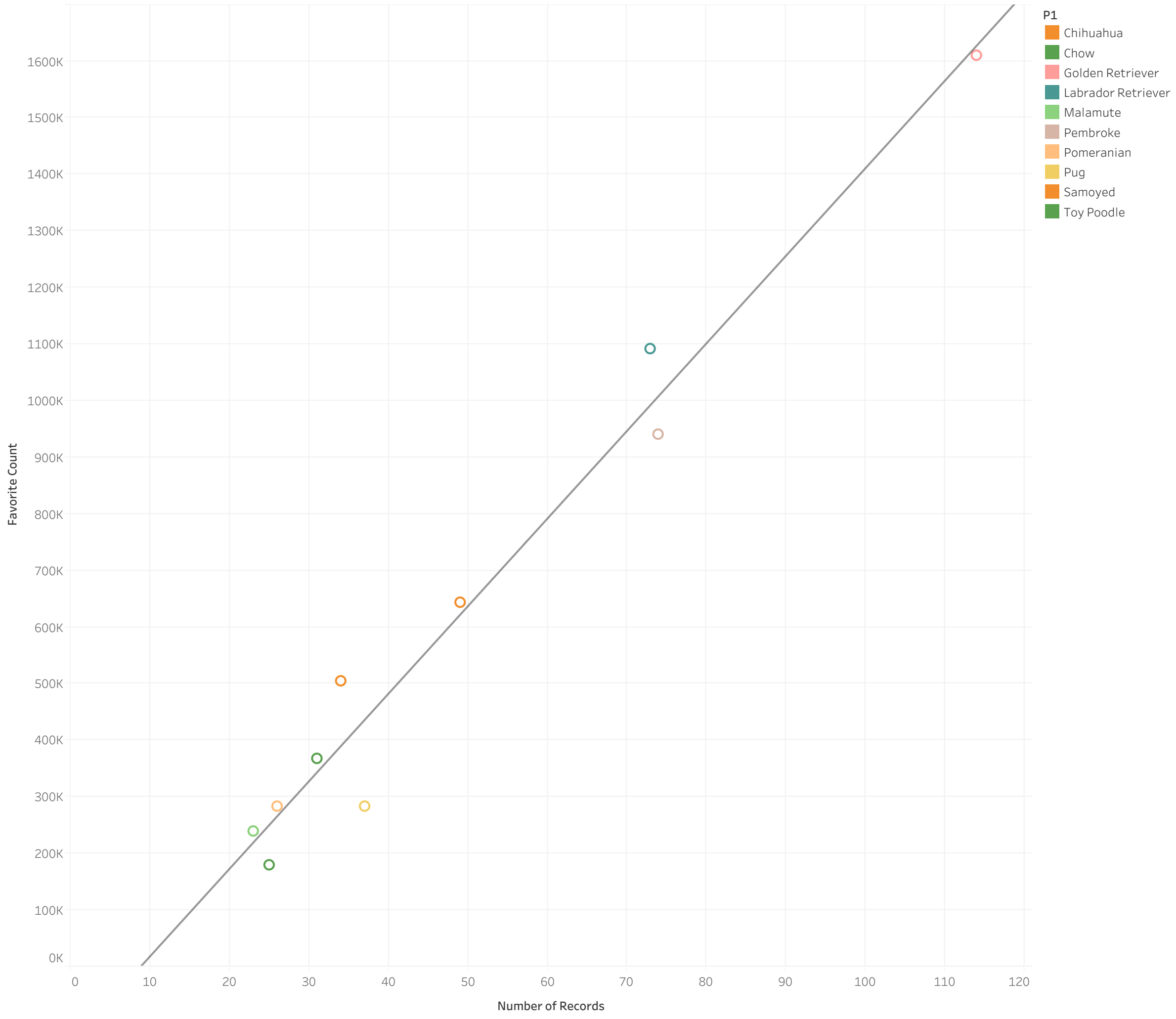
The most favorited breed was golden retriever, followed by penbroke and labrador. Looking at the chart below we can confirm that these are not only the most favorited breeds, but also the most common breeds.
We can observe this correlation (top 10 breeds) in the following graph:

Most common breeds



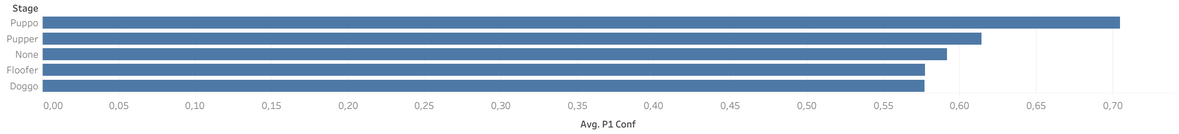
Sum of Number of Records for each P1. The view is filtered on sum of Number of Records, which ranges from 25 to 134.

Correlation - Number of Records and Favorites



Sum of Number of Records vs. sum of Favorite Count. Color shows details about P1. Details are shown for P1. The view is filtered on P1, which keeps 10 of 293 members.

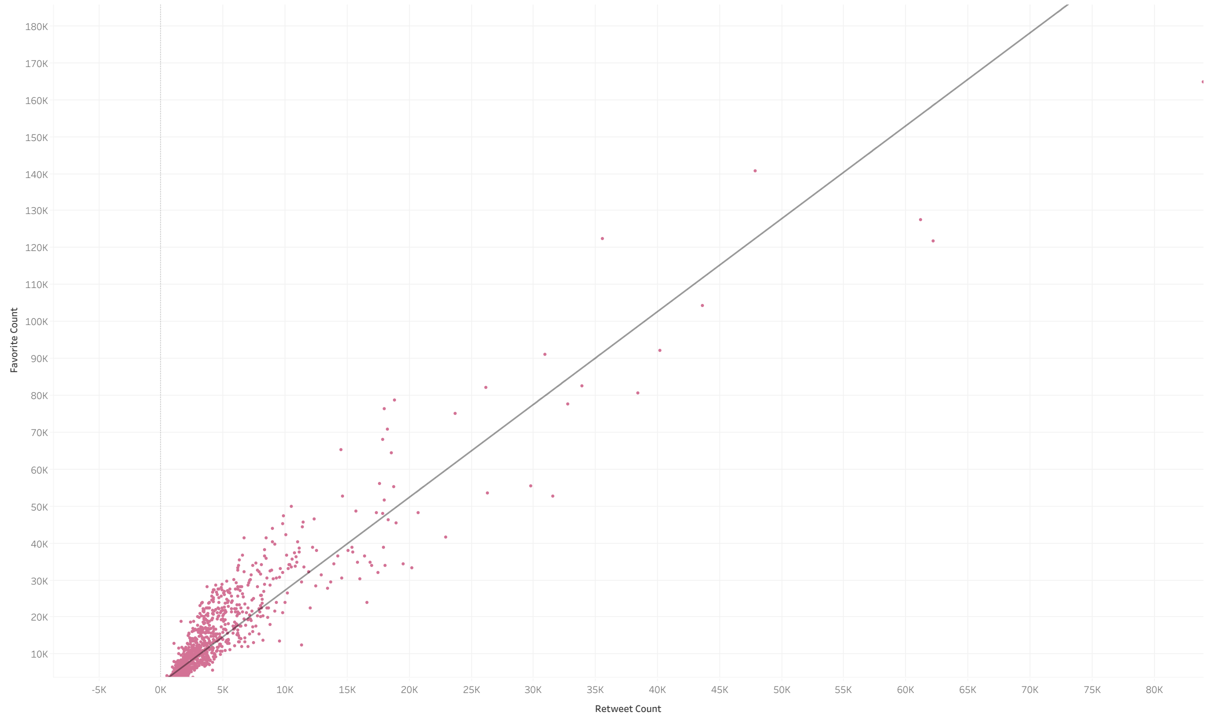
Prediction Avegare for Dog Stages



Average of P1 Conf for each Stage. The view is filtered on Stage, which excludes Undefined.

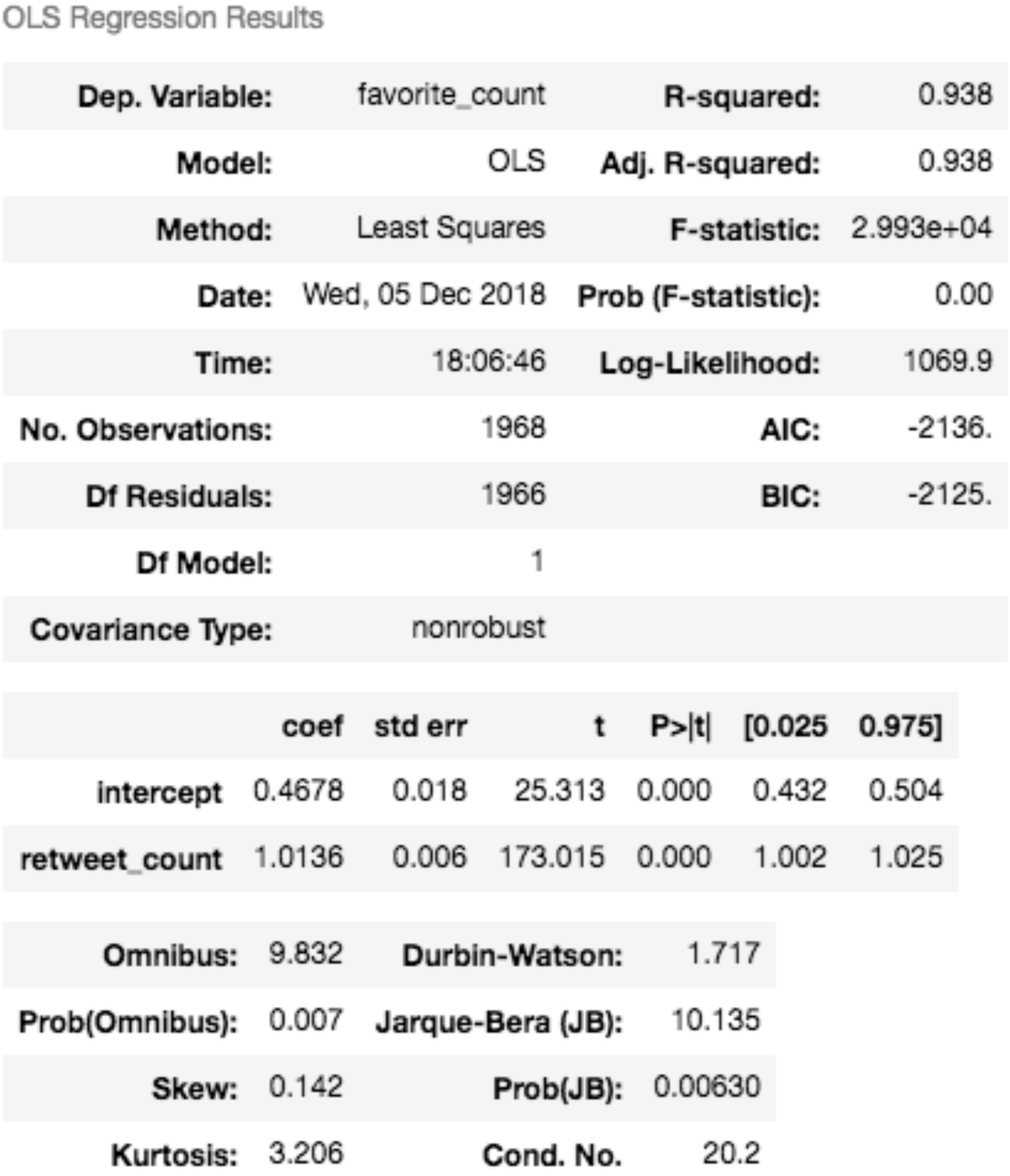
Looking at the first dog prediction (P1) from the neural network we can observe that when the image is predicted as a dog (omitted from the plot is when images are predicted as “no dogs”) the confidence of the prediction very high when the stage is doggo.

Correlation Retweet - Favorite



Retweet Count vs. Favorite Count.

A strong correlation between Retweet Count and Favourite Count. To look a bit more closer into this correlation I created a simple linear regression model using the Statsmodels library. The summary of results is shown in the next page:



OLS Regression Results

Dep. Variable:	favorite_count	**R-squared:**	0.938
Model:	OLS	**Adj. R-squared:**	0.938
Method:	Least Squares	**F-statistic:**	2.993e+04
Date:	Wed, 05 Dec 2018	**Prob (F-statistic):**	0.00
Time:	18:06:46	**Log-Likelihood:**	1069.9
No. Observations:	1968	**AIC:**	-2136.
Df Residuals:	1966	**BIC:**	-2125.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
intercept	0.4678	0.018	25.313	0.000	0.432	0.504
retweet_count	1.0136	0.006	173.015	0.000	1.002	1.025

Omnibus:	9.832	**Durbin-Watson:**	1.717
Prob(Omnibus):	0.007	**Jarque-Bera (JB):**	10.135
Skew:	0.142	**Prob(JB):**	0.00630
Kurtosis:	3.206	**Cond. No.**	20.2

The coefficients obtained allow us to predict the favourite count based on the number of tweets. The good fit of the model is confirmed by the high squared correlation value R-squared. Also, the statistical significance of the dependent variable favorite_count is shown in the low p-values. As a summary we could conclude that the observations provided by the plot are confirmed by the linear model.