**Module 4: Assessing dog breeds for maximum social media outreach**

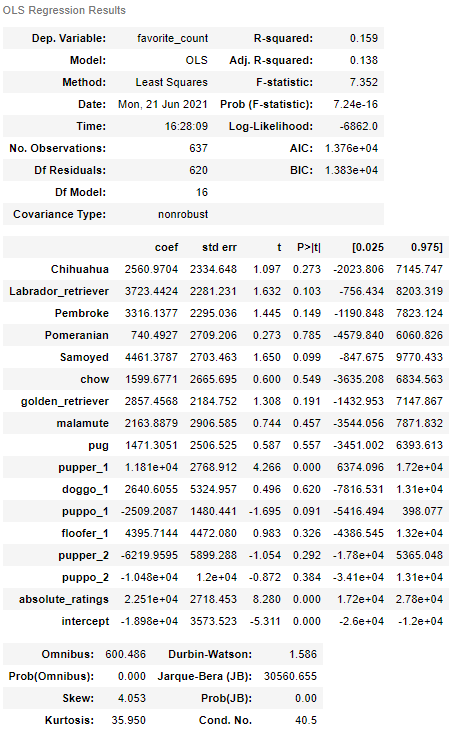
I will devise a model to predict how many favorite\_counts a dog of a given breed is likely to get based on the breed, confidence level of predictions and rating. I'll also do a similar analysis for retweet\_count.

On cleaning the data, I can conclude that any kind of model I build on top of these predictions would not be a good one since we don't have significant number of observations for a particular breeds/predictions. We can however, try analysing the top predictions appearing in the dataframe. Through this we can get an estimate of what breeds tend to get submitted to WeRateDogs. We can use this to build a flawed yet functional estimator for the metric we desire to obtain. This is the billionth time I wish we had more observations and less variability!!

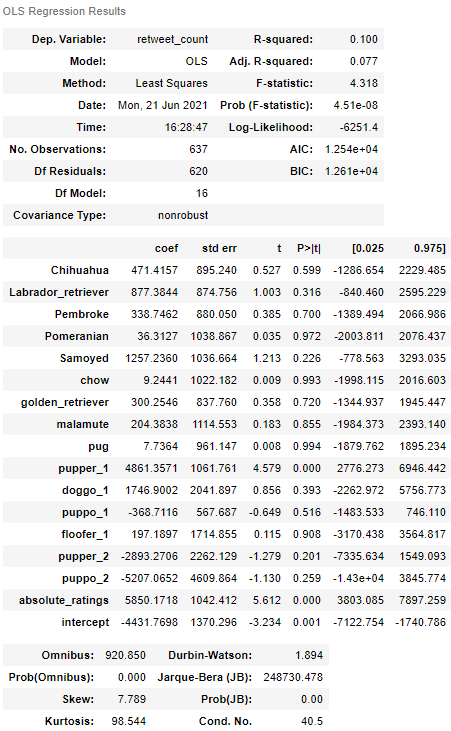
I will take toy\_poodle and Unidentified as the baseline variables. The features planned to be included in our analysis are:

1. absolute\_ratings which is rating\_numerator/rating\_denominator
2. Dog breeds with toy poodle as the baseline variable
3. Categories that the dog is presumed to belong to.
4. **Fitting Linear Models to the Data**

On fitting a linear regression to the data, we find the following coefficients:

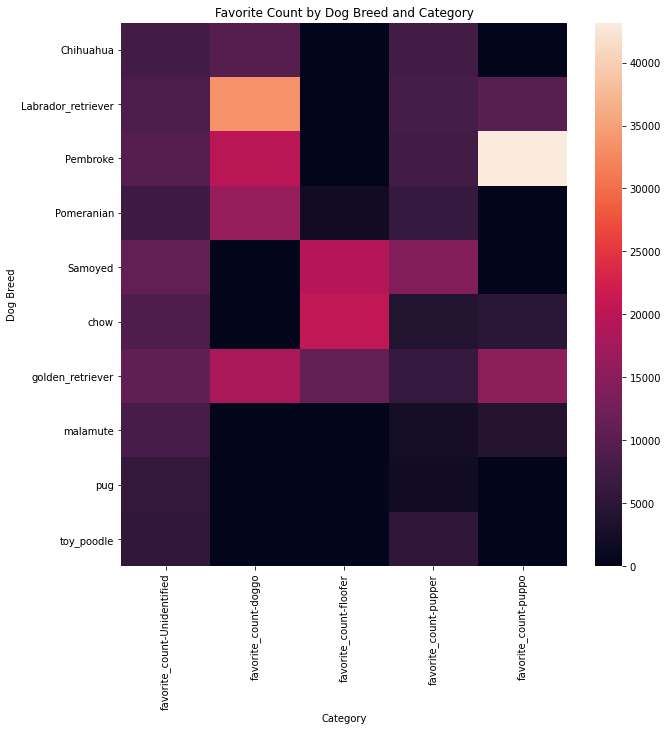


Most of the p values are above the alpha value (assumed 5% in this case). This means we cannot reject the null hypothesis that breeds other than toy\_poodle, categories other than Unidentified's result in greater favorite counts. We can however roughly gauge which features are negatively and positively correlated with the dependent variable with respect to the baseline. This inabilility to draw conclusions stems from the fact that we don't have enough data

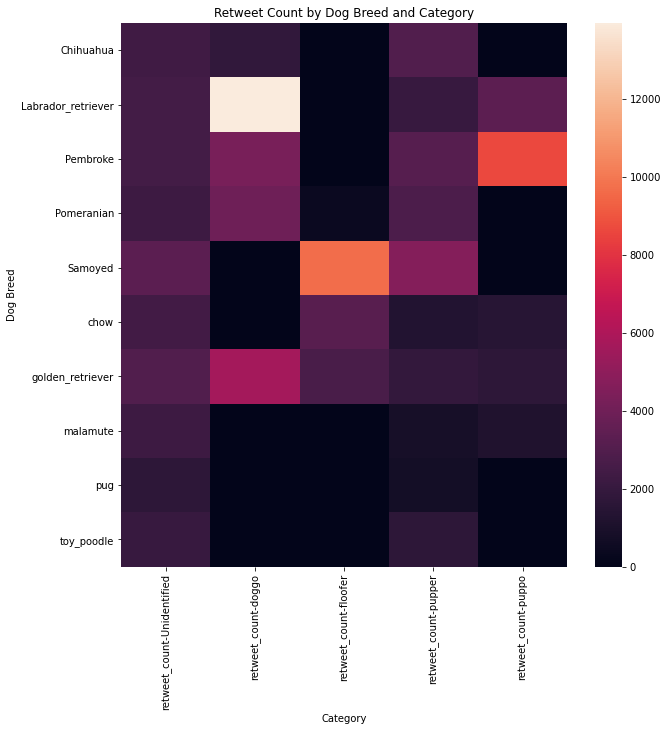


Here again, results are not statistically significant due to the same reasons

### **Visualizing Heatmap of Data Based on Dog Breed and Category**



Here we can observe that Pembroke in the puppo stage receives maximum number of likes. Labrador Retriever in doggo stage receives considerable likes. The safest option to go by is the Golden Retriever since it receives consistent favorites throughout its stages



Similar stats are observed for retweet\_count suggesting that retweet\_count and favorite\_count are positively correlated (pretty intuitive). Here as well Golden Retriever does consistently well across all its stages and Labrador retriever and Pembroke achieve high status in the pupper and puppo stages respectively.

One may attribute the drop in performance of other dog breeds in specific stages to a lack of data for those breeds in those stages. This is a perfectly valid observation.