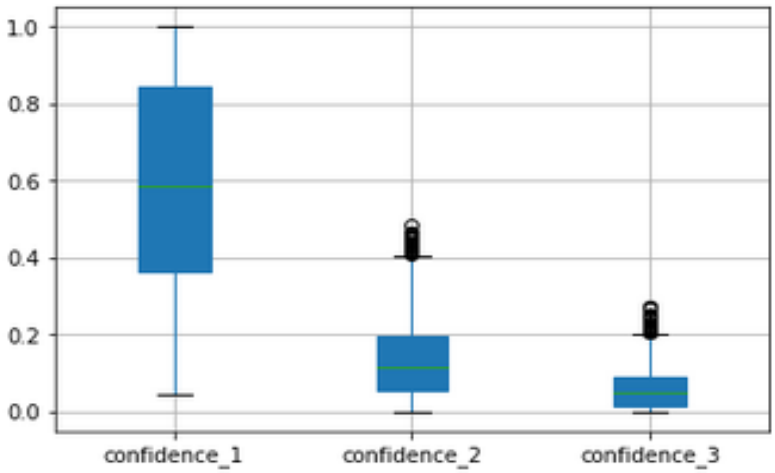
Final Report: Insights and Visualizations
"They're Good Datapoints, Brent."

Insight 1

I examined the section of the dataset that concerns image predictions, and analyzed the range of confidences in each prediction. My alternate hypothesis was that the strength of the confidence would go down between each prediction; so that the mean confidence for the second prediction would be lower than the first, and the third lower yet. My null hypothesis was that the numbers would be roughly the same across all predictions.

Using the describe function, my alternate hypothesis was shown to be correct – there is a distinct fall-off between predictions, with almost no overlap in the bulk of the values. I created a boxplot to better illustrate these differences, as shown below:




Insight 2:

For my second insight, I looked at how dogs who are named in the tweet compare to dogs who are not named across the following metrics: rating, number of retweets, number of favorites, and confidence. My null hypothesis for each metric is that there would be no significant difference between the two numbers. My alternative hypothesis was that there would be a noticeable difference, based on the theory that dogs with names would be likely to be more clearly pictured and more highly rated; they would be more personal, and thus invite higher levels of engagement and affection.

My analysis showed that the null hypothesis was correct for all but number of favorites, where dogs with names were retweeted noticeably more often than those without names.

Insight 3:

I wanted to see if specific breeds were predicted most for each stage (pupper, doggo, puppo and floofer), perhaps indicating a specific type of dog that was most likely to get each stage assigned. (For example, long-haired dogs might be more likely to be floofers, and small breeds more likely to be puppos.)

For brevity, I have summarized my results in the table below:

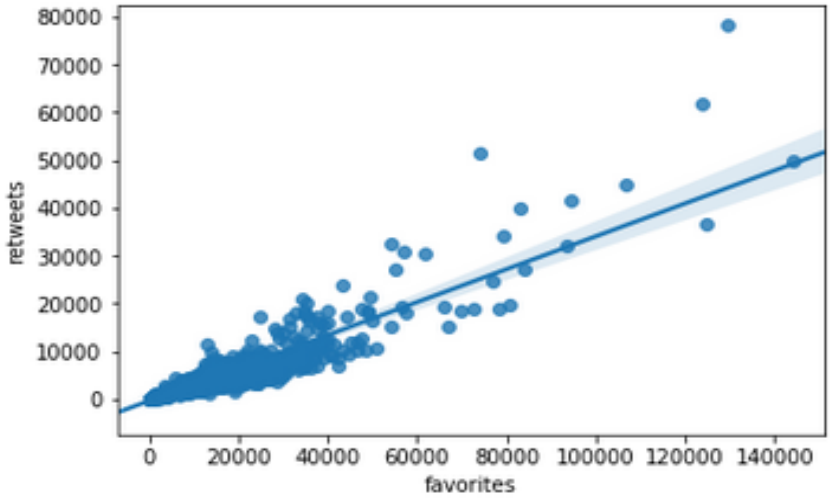
| | Pupper | Doggo | Puppo | Floofer |
|---|---|---|---|---|
| Prediction 1 | Golden Retriever | Golden Retriever | Eskimo Dog | Samoyed |
| Prediction 2 | Labrador Retriever | Labrador Retriever | Siberian Husky | Pomeranian |
| Prediction 3 | Labrador Retriever | Labrador Retriever | Malamute | Keeshond |

Golden Retrievers and Labrador Retrievers hold most of the top spots (there are far more puppers than any other dog stage); as they are numbers 3 and 1 in the American Kennel Club's most popular dogs, this is not terribly surprising[i]. (Although dog_rates of course receives submissions from all over the world, I chose this as a metric that likely represents the majority of submissions.)

(Having first-hand experience with Golden Retrievers, they are almost universally *extremely* photogenic and personable.)

It's notable, too, that all three top floofer predictions are indeed long-haired dogs.

Visualization:




For this analysis, my null hypothesis was that there was no linear relationship between number of favorites and number of retweets. My alternative hypothesis was that there was a relationship – that as number of favorites changed, so would number of retweets. Although there is slight fanning towards the higher end, this scatterplot shows a distinct linear regression and suggests a positive relationship between the two metrics, supporting my alternative hypothesis.

i http://www.akc.org/content/news/articles/the-labrador-retriever-wins-top-breed-for-the-26th-year-in-a-row/, accessed 26 February, 2018.