Neural Networks Performance on WeRateDogs.

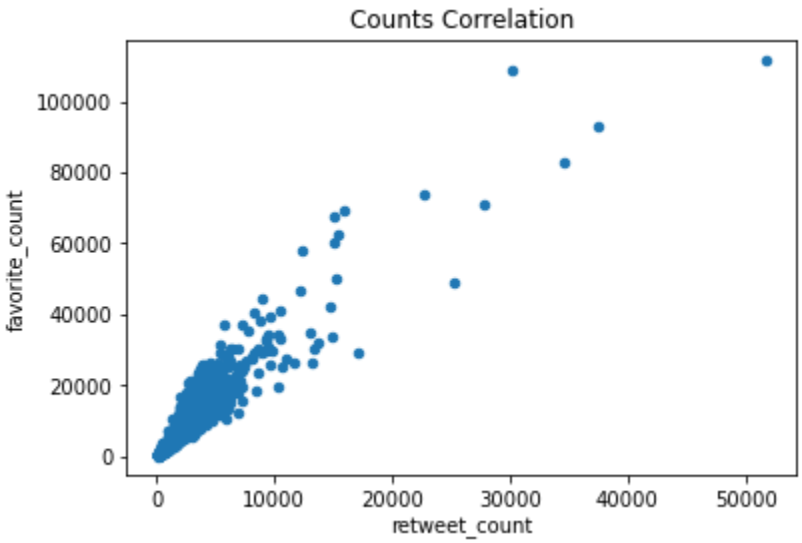
A Twitter handle 'WeRateDogs' currently has over 9 million followers. The followers' tweet about their dogs, including the dog names, a rating for the dog and images of the dog.

As part of a project for the Udacity Data Analysis Nanodegree, the twitter data was gathered. Data from the handle was gathered, up to August 2017. The data included 3 files. A CSV file archive of tweets containing tweet ids', ratings of dogs' image (usually above 10, which has helped its popularity) and dog names; An image prediction file scraped from a URL where tweets with images were analyzed in a neural network to create a table of the tweet ids, the prediction confidence and breed of dog for each image in the tweets; and a Json file of tweets from the Twitter API to include statistics of retweet count and favorite count for each tweet.

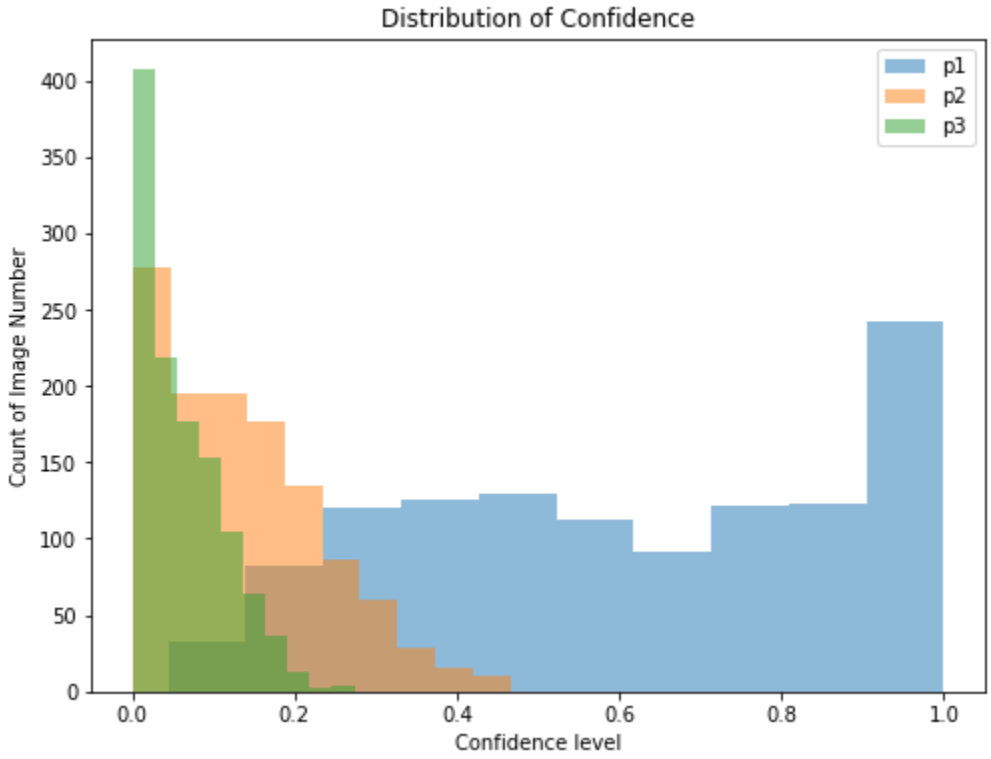
The data was loaded into 3 Pandas data frames on Jupyter Notebooks, assessed wrangled and cleaned to create a single data frame that combined elements from each data frame. Cleaning was done to improve data quality and tidiness. Some of the issues included merging the data frames, removing bad data such as tweets without images, retweets and any information not about dogs. The final data frame contained tweet ids of original tweets, retweet count, favorite counts, prediction confidence level, ratings and timestamp.

The data was analyzed by creating visuals to generate insights from the data.

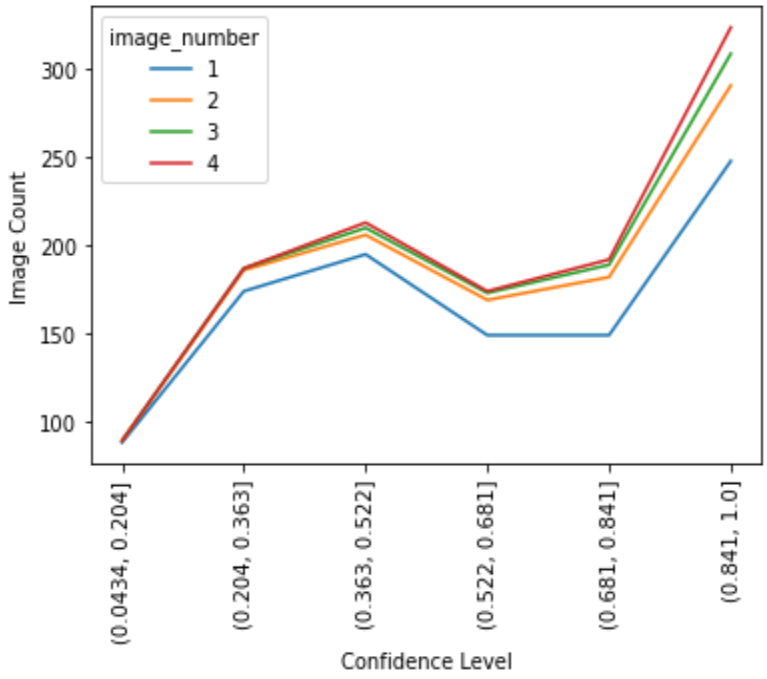
The image below is a visual of the relationship between the favorite counts and retweet counts of tweet ids. It shows a positive correlation between them indicating that most users interacting with tweets were likely to give both a favorite and a retweet to a particular tweet. There was also a positive correlation between retweet count and ratings.



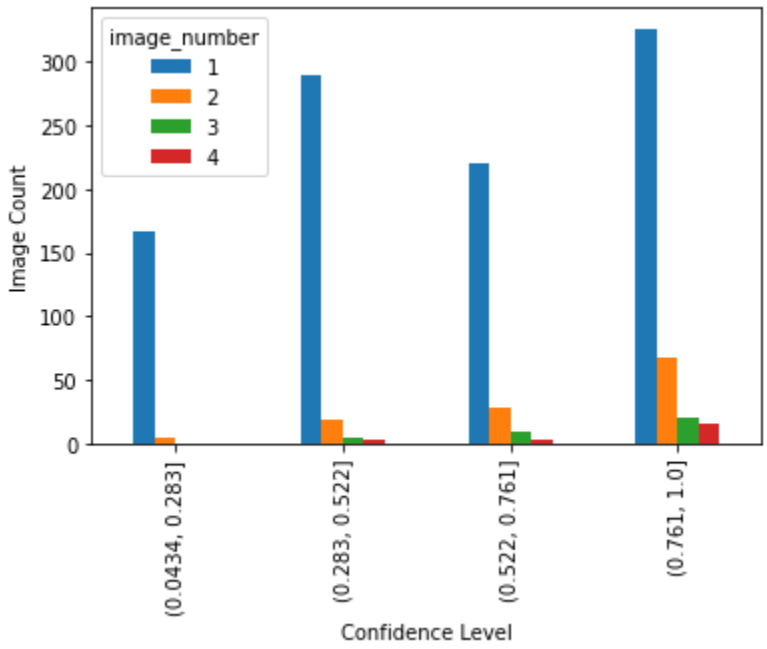
The neural network made 3 predictions p1, p2 and p3 of what an image was and whether it was true for a dog or not. The visual below shows a distribution of the confidence of each prediction. The first prediction p1 has a much higher confidence level than subsequent predictions p2 and p3.



The data was cleaned to ensure that an image was associated with each tweet. As a tweet can have up to 4 images, the neural network gave each tweet id an image number 1, 2, 3 and 4 according to its position in a tweet. The p1 confidence level associated with each image number was analyzed to create the visuals below. Image number 1 has the largest area of the line chart corresponding to most images being a number 1.



However, image number 4 has the highest minimum p1 confidence level.



A number of reasons can be adduced to this including the picture quality, the number of pictures and the various objects contained in each picture.

Analysis such as this seems useful in investigating the performance of neural networks and give insights into how systems can be designed and regulated to assure reliability.