@dog_rates (aka **WeRateDogs) Twitter User Data Analytics Activity**

Data analytics is an interesting activity that begins with having goals in the form of questions. Then data is collected, accessed and cleaned before analysis is performed. Insights are then obtained and interpreted in stories and visualizations.

This publication comes from a data analytics activity performed with twitter user data for @dog_rates (aka WeRateDogs). *@dog_rates* is a twitter user that rate dogs using a unique rating system where the rating numerator is greater than the denominator. The reason for this rating system is that WeRateDogs believe all dogs are good. Dogs are rated with humorous comments and some tweets contain dog stages know as doggo, puppo, floofer and pupper.

Data Gathering

Data was gathered from three file source of different formats. These files were assessed visually and programmatically to identify quality and tidiness issues that required cleaning. In the cleaning process, a master dataset that comprise the merger of data from all three files emerged. This was then analyzed for insights and visualization. This document, conveys the insights with visualizations.

Computing Resource

Data analysis were performed on a windows computer with Anaconda environment setup. Notepad and an online JSON viewer was used to visually assess the tweet-json.txt file downloaded from Udacity server. Microsoft Excel 2019 and a web based tabular file viewer was used to visually assessed a twitter archive file, containing @dog_rates tweet up to the year 2017 and image_predictions.tsv file containing dog breed predictions using dog image file extracted from tweets.

Method

The dataset used for this project was gathered, assessed and cleaned, from three files, namely; tweet-json.txt, image_predictions.tsv, and twitter-archive-enhanced.csv, all available on Udacity. The project required that twitter API be used to extract twitter data for @dog_rates, up to the

year 2017. This requires a twitter developer account, which I was not able to set up due twitter's web system verification process. Alternatively, a similar data that can be used for the project, was provided in the file tweet-json.txt. This file was then downloaded directly from Udacity server and used for this project.

The twitter-archive-enhanced.csv contains tweet archive of Twitter user @dog_rates up to the year 2017, in a state that required cleaning before meaningful use. While the image-predictions.tsv file contained records of dog breed predictions for dog images in each original tweet, obtained with a neural network algorithm.

Let me define a useful activity, known as ***Data wrangling***. According to Wikipedia, *"it is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics."*

After subjecting the three datasets - from the files earlier mentioned, through a data wrangling process, a master dataset was produced and analyzed for insights. The master dataset contained 1,914 tweet records and 21 columns/variables. The columns are identified below;

1. tweet_id – the message id
2. timestamp – the time of tweet
3. source – device twitter application
4. text – the tweet content
5. rating_numerator – the dog rating
6. rating_denominator – the rating denominator
7. dog_stages – doggo, puppo, floofer, pupper
8. retweet_count – number of retweets
9. likes – favorite count
10. language – the language (e.g., "en" for English)
11. image_url – URL of image used in breed prediction algorithm
12. img_num – the number of the image used
13. p1_dog_breed – first prediction dog breed
14. p1_conf' - first prediction confidence level
15. p1_dog – if first prediction is a dog or not
16. p2_dog_breed – second prediction dog breed
17. p2_conf – second prediction confidence level

18. p2_dog – if second prediction is a dog or not
19. p3_dog_breed – third prediction dog breed
20. p3_conf – third prediction confidence level
21. p3_dog – if third prediction is a dog or not

Note: Assessing the datasets, it was noticed that some of the tweets contained double dog stage such as doggo with pupper and doggo with puppo. Records in this category were classified in dog stage as "doggo pupper" and "doggo puppo". These two dog stages will be observed in the insights below.

Insights and Visualizations

A correlation matrix obtained while analyzing the dataset, showed a strong positive correlation between retweet_counts and likes (figure 1). This was made clearer with a scatter plot of these two variables as shown in figure 2.

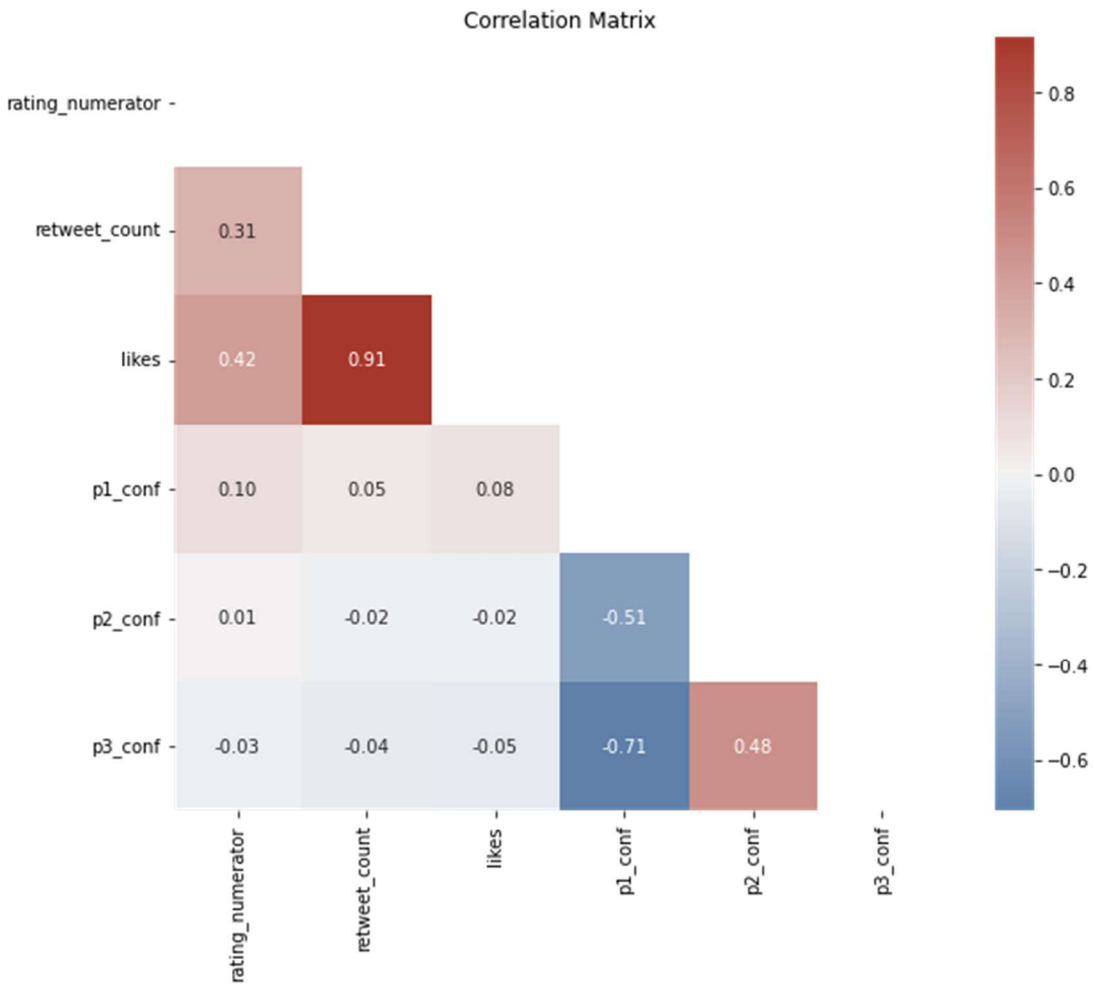


Figure 1

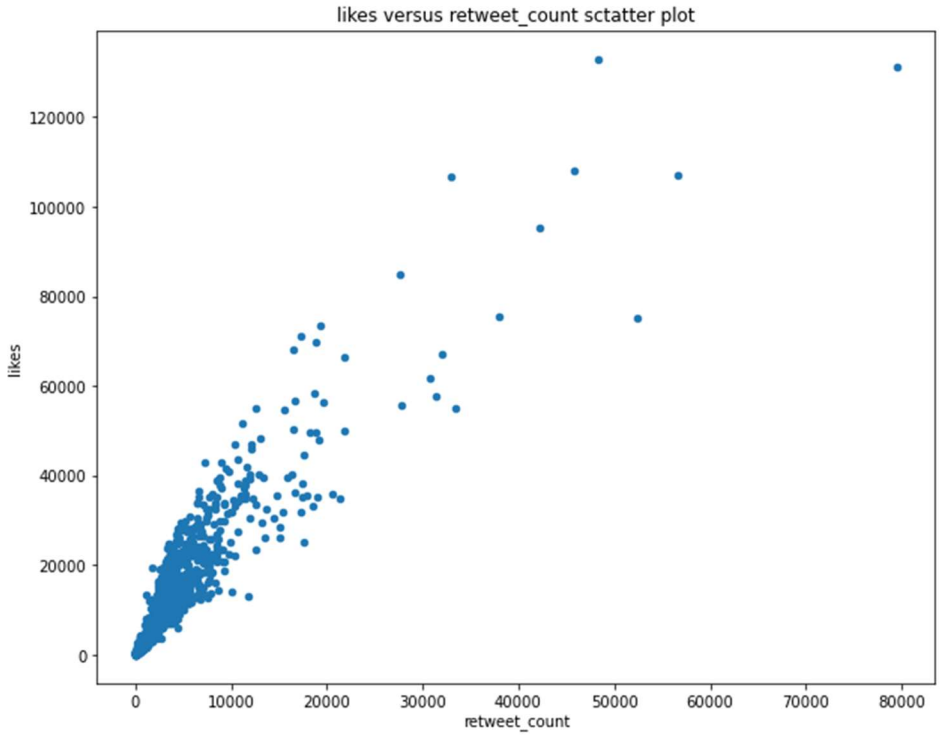
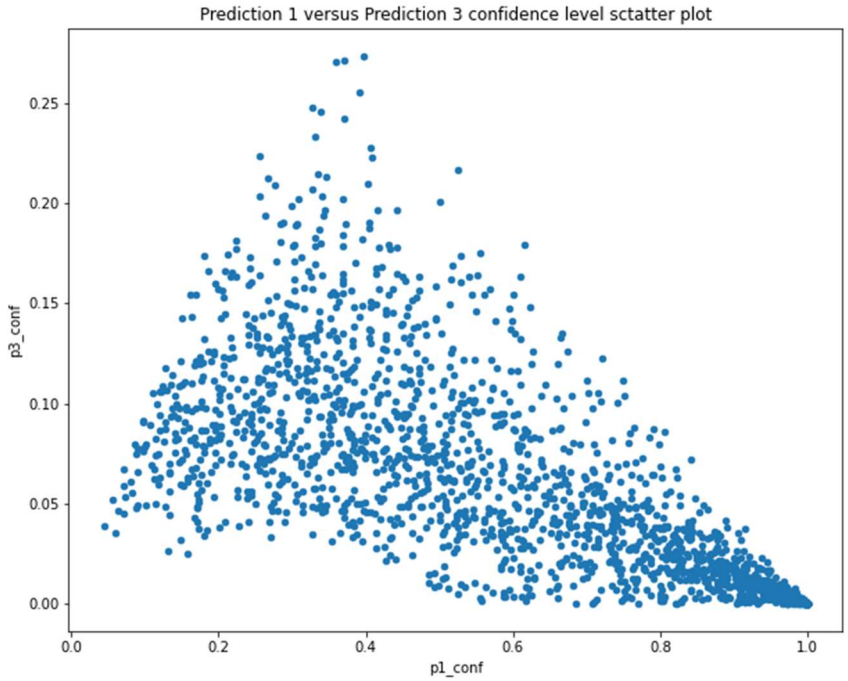


Figure 2

Prediction 1 confidence level (p1_conf) has a negative correlation with Prediction 3 confidence level (p3-conf) but prediction 2 and 3 confidence levels show a little positive correlation. This indicates the insight based on prediction 2 and 3 are going to be quite similar, with both being different from what will be observed from insights based on prediction 1. A scatter plot of prediction 1 versus prediction 3 confidence levels is shown below.



The dataset was further grouped by dog stages (i.e., doggo, puppo, floofer, pupper, doggo pupper and doggo puppo) to see what insight lies withing them.

1. For all the dog stages, pupper had more retweet counts followed by doggo. The rest come in the order puppo, floofer, doggo pupper, doggo puppo. This also reflets their order of number of records except doggo pupper had more records than floofer but floofer had more retweets.

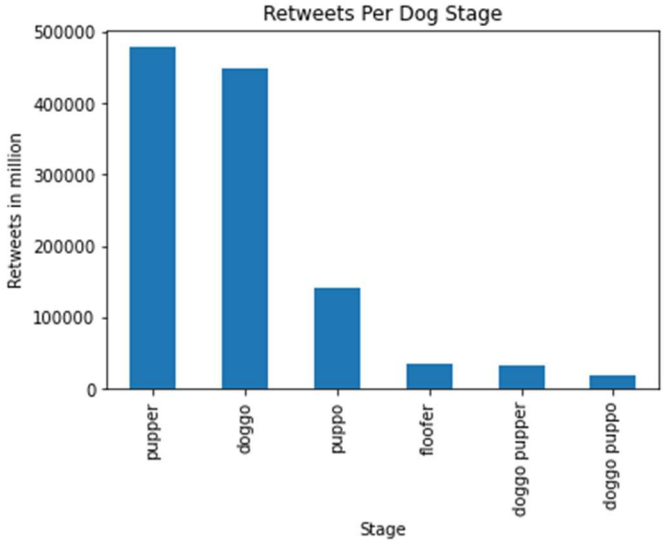


Figure 3

2. For likes, tweet relating to pupper had the highest, followed by doggo, then puppo before doggo pupper, and next is floofer before doggo puppo

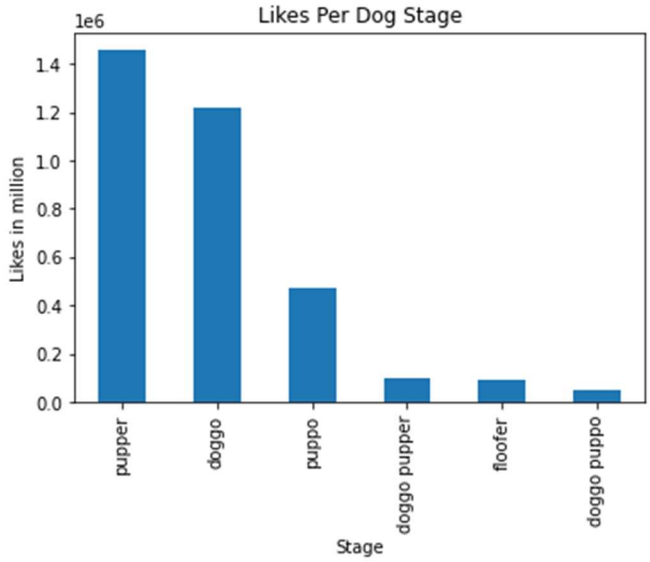


Figure 4

3. Records of dogs in pupper stage where more than others but its mean is 4th in first algorithm prediction confidence level. On a reverse case doggo pupper was 4th in number of records distribution but had the highest prediction confidence level mean with the first algorithm.

   In the 2nd and 3rd predictions, pupper had a higher confidence level mean. Quite similar pattern is observed with the median of each dog stage for each prediction, where pupper stage is better in first algorithm and not so better in second and third algorithm's confidence level.

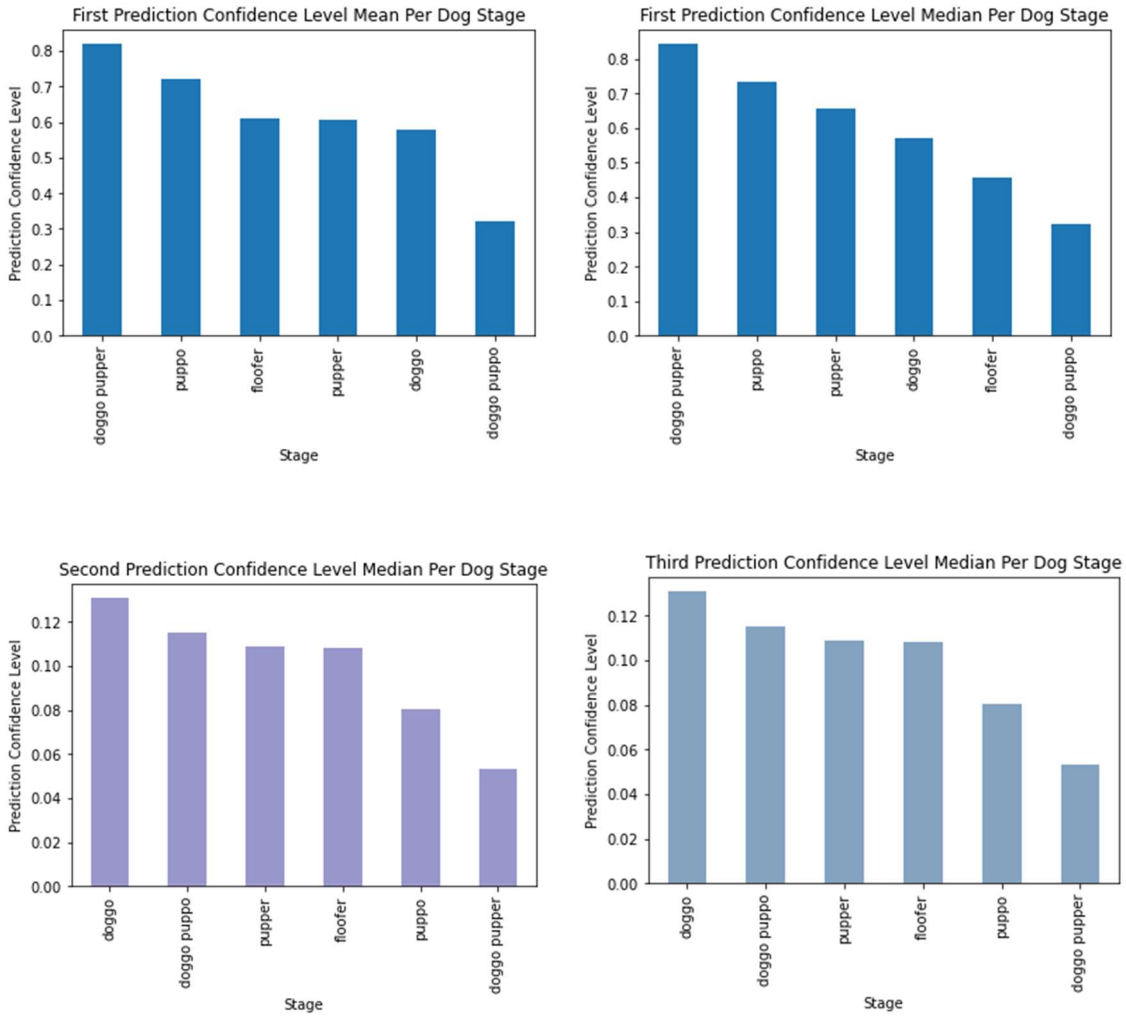


Figure 5

4. Considering dog breeds by first prediction algorithm, breeds with the highest likes for the dog stages are as follows;

Dog Class	Breed	Likes
pupper	French_bulldog	131,626
doggo	Golden_retriever	184,533
puppo	Lakeland_terrier	132,810
floofer	Samoyed	44,647
doggo pupper	golden_retriever	71,282
doggo puppo	flat-coated_retriever	47,844

5. Of all dog breeds predicted from the tweets in the dataset, Golden_retriever, had the most likes in prediction 1 while it was Labrador_retriever for prediction 2 and 3. One is likely to go with prediction 1 since the median and mean confidence level is better than the other two predictions.

Conclusion

The insights above are just a few of many insights that can be obtained from the dataset prepared from this project. Take note, the insights are not inferential since no statistical tests were conducted. Also, further data wrangling - specifically gathering activity, can be performed on the dataset used, to make it much better quality due to issues such as no dog stage in most tweets.