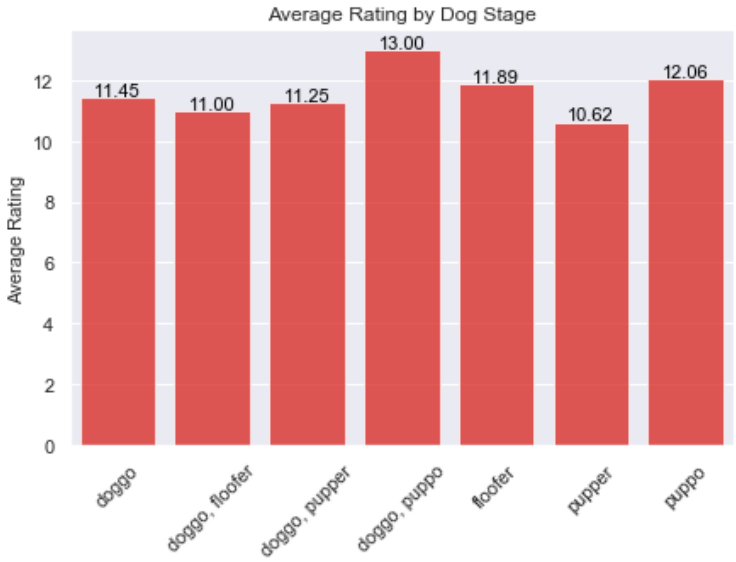
Paras Patel
Data Analytics – Udacity
Act Report

A transition into exploratory data analysis is made after the data wrangling steps are completed.

The following visualizations and insights were produced from the first DataFrame (df1):

1. Average rating by dog stages



2. Retweet and favorite count after merging with third dataframe and normalizing data.

	dog_stage	Retweet Norm	Favorite Norm	dog_count
0	doggo	5544.972973	14563.135135	74
1	puppo	5077.472222	16139.638889	36
2	doggo, pupper	4954.833333	10322.833333	12
3	floofer	3884.444444	10538.444444	9
4	pupper	2348.440476	5825.928571	252

The following visualization was produced from the second DataFrame (df2):

3. Image prediction model results - confidence distribution

```
Green = p3, Blue = p2, Red = p1
```

The following insights were produced from the third DataFrame (df3):

4. Top retweet and favorite Count by dog breed.

	dog_breed	Retweet Count	Favorite Count
1	Golden Retriever	522922.0	1582711.0
2	Labrador Retriever	398475.0	1066241.0
3	Pembroke	259245.0	944444.0
4	Chihuahua	226230.0	633571.0
5	Samoyed	189909.0	498761.0
6	French Bulldog	115851.0	460616.0
7	Cocker Spaniel	112437.0	291079.0
8	Chow	106699.0	368427.0
9	Pug	102705.0	298913.0
10	Eskimo Dog	91193.0	249328.0

5. Top retweet's per tweet average by dog breed after merging with df2

	dog_breed	Retweet Norm	Favorite Norm	dog_count
0	Standard Poodle	11915.714286	15082.571429	7.0
1	Afghan Hound	8333.000000	14197.333333	3.0
2	Eskimo Dog	6079.533333	16621.866667	15.0
3	English Springer	5255.700000	13475.300000	10.0
4	Cardigan	5012.235294	13207.764706	17.0
5	Tibetan Mastiff	4962.400000	9242.600000	5.0
6	Saluki	4805.000000	23083.000000	4.0
7	Mexican Hairless	4522.333333	15277.333333	3.0
8	Samoyed	4521.642857	11875.261905	42.0
9	Great Pyrenees	4398.642857	12271.000000	14.0

6. Top favorite's per tweet average by dog breed after merging with df2

	dog_breed	Favorite Norm	Retweet Norm	dog_count
0	Standard Poodle	15082.571429	11915.714286	7.0
1	Afghan Hound	14197.333333	8333.000000	3.0
2	Eskimo Dog	16621.866667	6079.533333	15.0
3	English Springer	13475.300000	5255.700000	10.0
4	Cardigan	13207.764706	5012.235294	17.0
5	Tibetan Mastiff	9242.600000	4962.400000	5.0
6	Saluki	23083.000000	4805.000000	4.0
7	Mexican Hairless	15277.333333	4522.333333	3.0
8	Samoyed	11875.261905	4521.642857	42.0
9	Great Pyrenees	12271.000000	4398.642857	14.0

7. Word cloud for dogs with highest ratings.

Conclusion

In the dog stage analysis, 'doggo, puppo' and 'doggo, floofer' have only one value, so these cannot be evaluated and compared to the same confidence as the other categories. 'Pupper', with the highest count size of 233, has the lowest average ranking of the categories with 10.57. 'Puppo', with a count size of 31, has the highest average ranking of 12.

Before evaluating dog_breeds, it was necessary to evaluate the image prediction model itself. Over 200 dog breeds were identified using a p2 or p3 prediction. 'p2' has an average confidence of 11%, while 'p3' has an average confidence of 5%. Because over 10%, a significant amount, of the total data comes from these predictions, the following dog breed analysis was performed for predictions with a **confidence of 25% or higher**.

The dog breed analysis revealed that the **Golden Retriever** was the most popular dog i.e. received the highest number of retweets and favorites. 9 out of 10 dogs were both on the top Favorite's list as well as the top Retweet's list. These dogs are: Golden Retriever, Labrador Retriever, Pembroke, Chihuahua, Samoyed, French Bulldog, Chow, Cocker Spaniel and the Pug.

However, it's important to note that Retweet and Favorite counts also reflect the **Number of Tweets**. Without normalizing the data, a fair comparison can't be made. After normalizing the data, it's revealed that the **Standard Poodle** receives the most Retweet's and Favorite's per tweet, 7,272 and 17,596 respectively. In this case, it's also important to consider the possibility of outliers that have counts of 1. Take for instance the scenario where WeAreDogs posts one tweet about a specific dog breed and that tweet goes viral and garners far more retweets and favorites than originally expected. To avoid making inaccurate conclusions, only dog breeds with 'dog_counts' (number of tweets for a dog breed) of greater than 2 are analyzed.

The following comparison, similar to that of the dog breed, was performed for the dog stages. After normalization, we can conclude that doggo's are the most popular dog stages, with average retweets and favorites per tweet of 5,658 and 16,084 respectively.

A word cloud was finally generated to identify any keywords that could correlate to high rating. "Great", "pup", "please", and "call" are some of the top words that are used in tweets with high ratings.