We Rate Dogs Project: Data Analysis Report

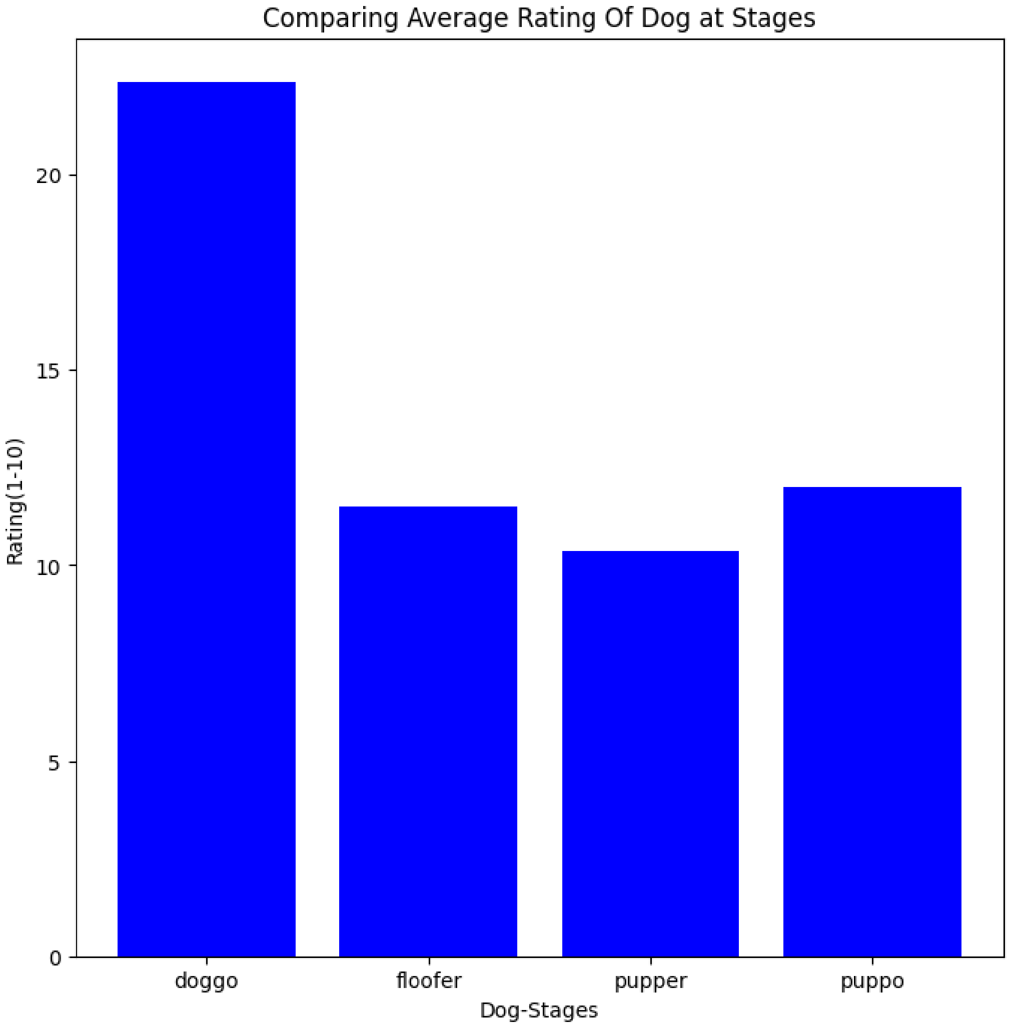
Introduction

In this document, I am going to summarize the work that I have done in this project.
The dataset I examined pertains to the tweet archive of a popular Twitter user, @dog_rates, also known as WeRateDogs. WeRateDogs is a well-known Twitter account that assigns ratings to users' dogs, accompanied by witty and humorous comments about the dogs. Notably, these ratings consistently feature a denominator of 10, adhering to a standardized scale. However, the numerators associated with the ratings tend to surpass 10 quite frequently. For instance, ratings such as 11/10, 12/10, 13/10, and so on are commonly assigned. The rationale behind these inflated ratings is humorously justified by the account's catchphrase, "they're good dogs Brent." WeRateDogs boasts a substantial following of over 4 million users and has garnered significant attention from media outlets worldwide.The analysis conducted on the Twitter archive encompassing all tweets as of August 1, 2017, has led to the following findings, which are summarized in this report.

Data Analysis

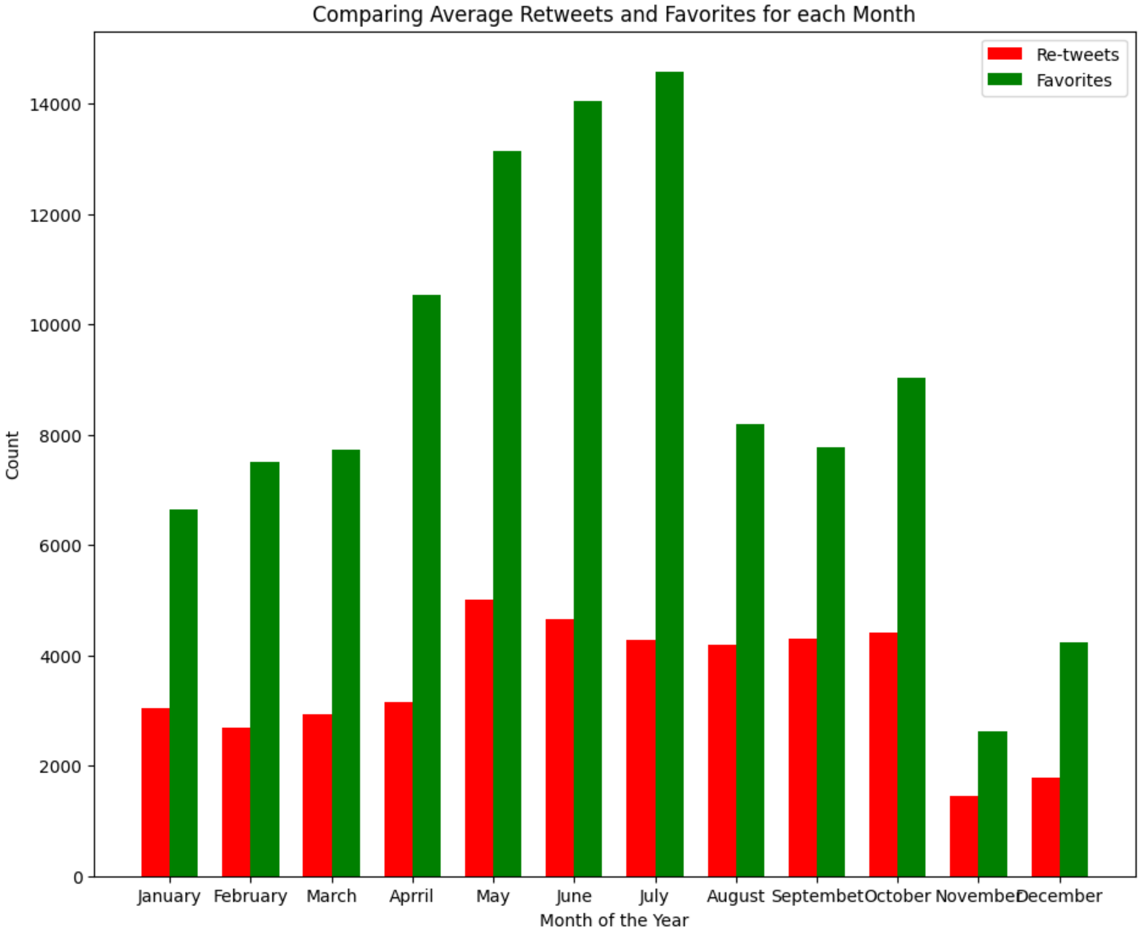
In this analysis, the dataframes tweet_data_clean and img_pred_clean were examined and subsequently saved as separate .csv files named twitter_archive_master.csv and image_predictions_master.csv, respectively. Furthermore, the analysis encompassed the consolidation and cleansing of diverse datasets collected from various origins.
I had several questions of interest following my analysis:

1. Analyzing the average ratings of different dog stages to draw comparisons



Based on the above bar graph, a notable distinction in average ratings among the different categories. Specifically, doggo demonstrates the highest average rating, while pupper exhibits the lowest.

2. Analyzing the average count of retweets and favorites over a month.



From the above plot, the following observations can be noted:

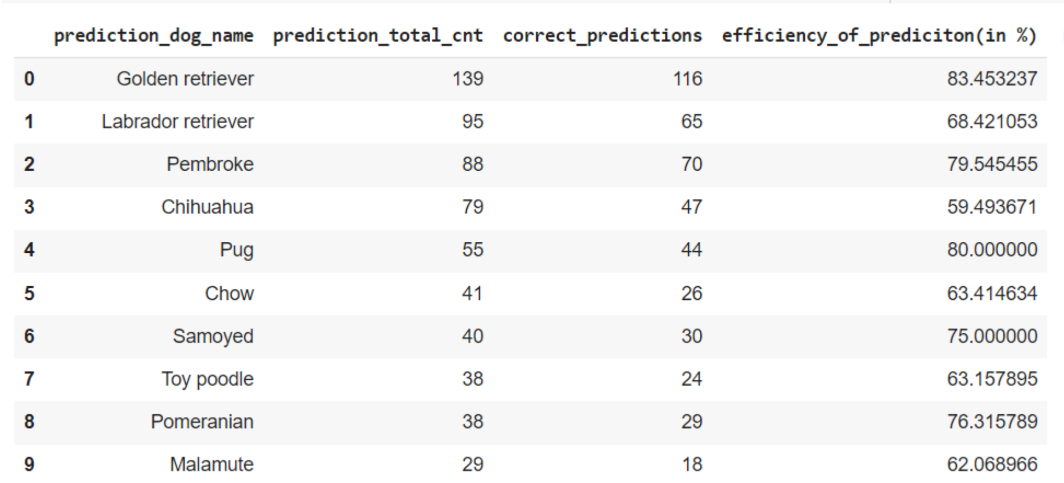
1. The average number of favorites is consistently higher than the average number of retweets. This suggests that users tend to show more preference for favoriting tweets rather than retweeting them.

2. Among the months analyzed, July stands out with the highest average number of favorites, indicating that tweets posted during this month are more likely to receive favorable engagement from users. On the other hand, November has the lowest average number of favorites.

3. When it comes to retweets, May emerges as the month with the highest average number, suggesting that tweets posted in May have a higher chance of being shared by users. Conversely, November experiences the lowest average number of retweets.

4. June exhibits the second highest average number of both favorites and retweets, indicating that tweets posted in June tend to receive a good level of engagement from users, although not as high as those in July.

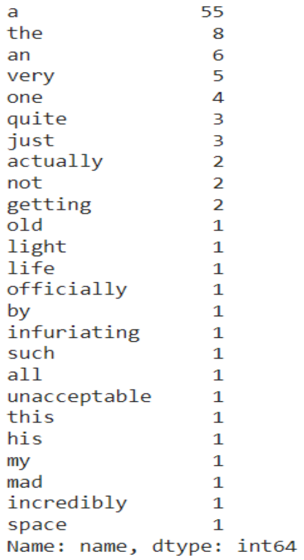
3. Efficiency of Algorithm's Initial Predictions for the Top 10 Most Frequent Outcomes



	prediction_dog_name	prediction_total_cnt	correct_predictions	efficiency_of_prediciton(in %)
0	Golden retriever	139	116	83.453237
1	Labrador retriever	95	65	68.421053
2	Pembroke	88	70	79.545455
3	Chihuahua	79	47	59.493671
4	Pug	55	44	80.000000
5	Chow	41	26	63.414634
6	Samoyed	40	30	75.000000
7	Toy poodle	38	24	63.157895
8	Pomeranian	38	29	76.315789
9	Malamute	29	18	62.068966

After analyzing the above table, noticeable patterns arise regarding the algorithm's effectiveness in predicting various dog breeds. An interesting observation is that the algorithm performs exceptionally well in accurately predicting Golden Retrievers, as indicated by the consistently correct predictions among the top 10 most frequent results. On the other hand, the algorithm demonstrates relatively lower accuracy when it comes to predicting Chihuahuas, with a higher number of incorrect classifications. It is important to note that the dataset exhibits a considerable frequency of predictions for Golden Retrievers, while Malamutes receive comparatively fewer predictions. These findings imply a potential bias towards successful predictions for Golden Retrievers within the algorithm's performance on this particular dataset.

4. Are there any invalid data entries in names of dogs?



```
a               55
the              8
an               6
very             5
one              4
quite            3
just             3
actually         2
not              2
getting          2
old              1
light            1
life             1
officially       1
by               1
infuriating      1
such             1
all              1
unacceptable     1
this             1
his              1
my               1
mad              1
incredibly       1
space            1
Name: name, dtype: int64
```

Yes, there are several invalid data entries in the name feature, as there are some invalid names found like a, the, an, very etc so these were handled during the process too.

In addition to the previously mentioned findings, it was observed that there is a notable positive association between the quantity of tweets and the number of likes received. This suggests that as the number of tweets increases, there is a tendency for a higher level of appreciation or approval towards the dog. Also, a good positive correlation was observed between number of likes and number of retweets . This indicates that as the count of retweets increases, there is a corresponding tendency for an increased level of liking or positive response towards the dog.