Data Wrangling and Analysis of @WeRateDogs Archived Tweets

WeRateDogs is a Twitter account that requires users to send in dog pictures in other to get a rating based on a defined rating system. The rating comprises an integer numerator from 0 and a constant denominator of 10. In cases where there is more than one dog in a picture, the denominator is 10 multiplied by the number of dogs.

To perform data wrangling and analysis, archived tweets of the Twitter account were provided, and more tweets metadata was obtained using Tweepy and the Twitter API. After successfully gathering, assessing and cleaning the datasets, some insights and visualisations were produced from the cleaned version of the data

The following are the insights discovered:

1. Average retweets and favourites counts are approximately 2306 and 7785, respectively; this implies that, on average, dogs rated by the account are likely to get 2306 retweets and 7785 favourites. Outliers were not considered, so this is a mere assumption

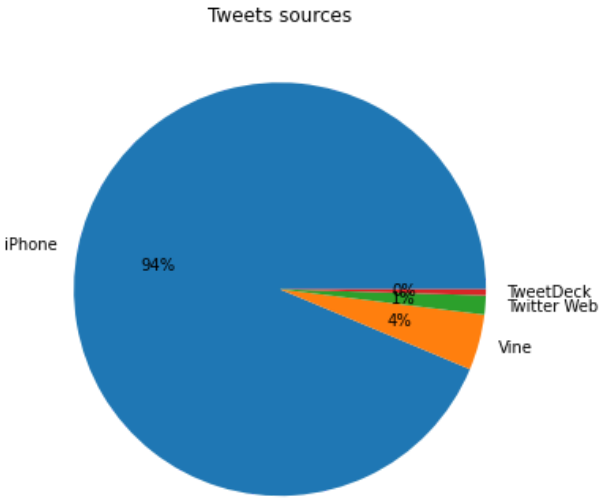
	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count
count	2.088000e+03	2088.000000	2088.000000	2088.000000	2088.000000
mean	7.363013e+17	12.192050	10.454981	2306.331418	7785.028736
std	6.703071e+16	40.450186	6.656316	4042.436246	11338.465996
min	6.660209e+17	0.000000	10.000000	11.000000	66.000000
25%	6.767367e+17	10.000000	10.000000	511.000000	1716.750000
50%	7.094844e+17	11.000000	10.000000	1114.000000	3536.000000
75%	7.870529e+17	12.000000	10.000000	2623.500000	9699.250000
max	8.924206e+17	1776.000000	170.000000	70784.000000	144952.000000

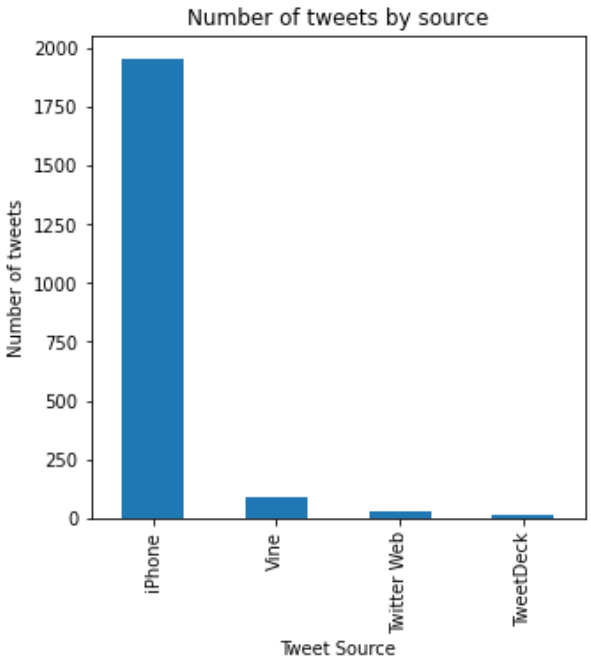
This was achieved by calling the DataFrame.describe() function on the cleaned DataFrame.

The max and minimum retweet and favourite counts can also be spotted from the summary table.

2. The iPhone is the primary device used for the tweets to rate the dogs in the cleaned dataset. Calling Series.value_counts() on the source column displays the counts of tweets for each tweet source. `tweet_archive_merged.source.value_counts()`

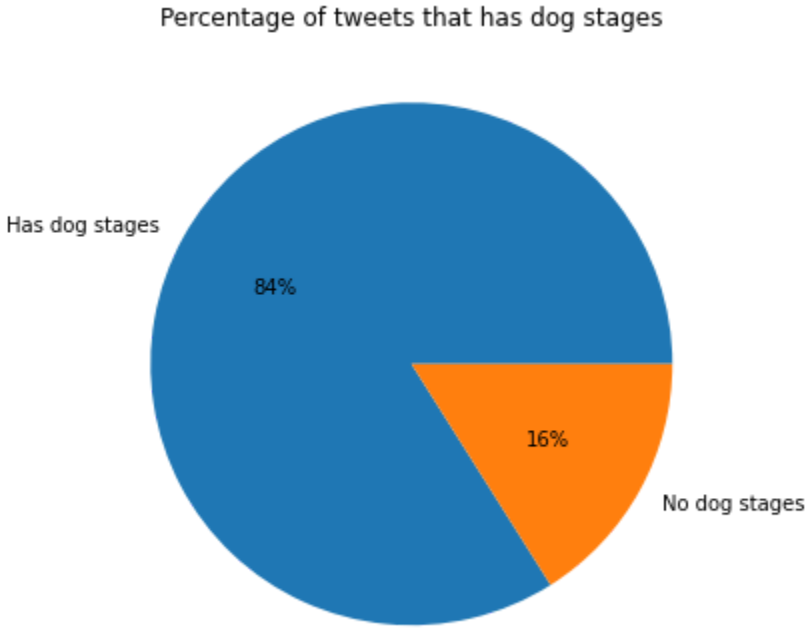
```
iPhone          1956
Vine              91
Twitter Web       31
TweetDeck         10
```





3. 84% of the tweet with ratings does not indicate the dog stage this is equivalent to 1753 out of 2088 of tweet has no dog stage. This can be seen by running:

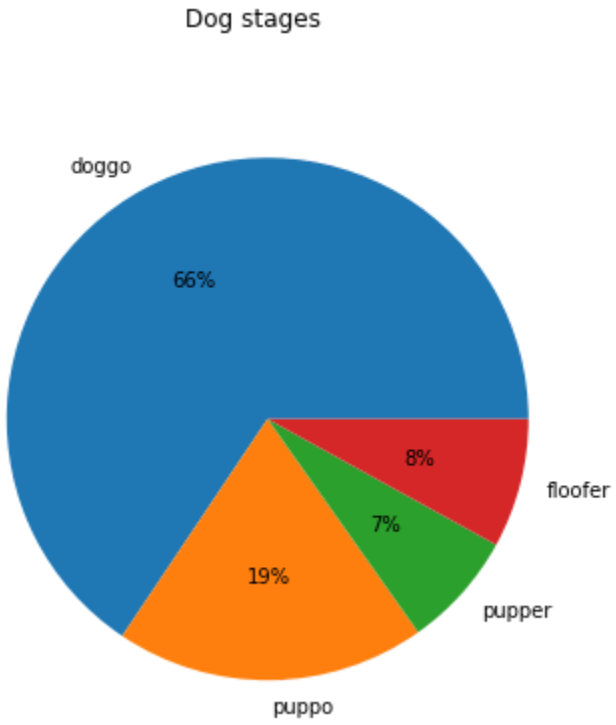
   tweet_archive_merged.dog_stage.isna().value_counts()



4. To determine the percentage of dog stages, we select all records with a dog stage present and run the below code to calculate the total count for each dog stage.

```
# Insight 4
dog_stages_count  = tweet_archive_merged.dog_stage.value_counts()
dog_stages_count_dict = {}
for stages in dog_stages_count.index[1:]:
    for stage in stages.split(','):
        dog_stages_count_dict[stage] = dog_stages_count_dict.get(stage, 0) +
dog_stages_count[stages]

pd.DataFrame.from_dict(dog_stages_count_dict, orient='index').plot.pie(subplots=True,
figsize=(8, 6), title="Dog stages", ylabel='', autopct='%.0f%%', legend=False)
```



The above explanation shows us sample of the insights we can get from the dataset, more insight can be extracted by looking at other features and presenting our findings in suitable visualizations.