REPORT ON ANALYSIS AND VISUALIZATION

By Priyanka Krishnan

Introduction:

For the Wrangle and Analyze Data project, I have wrangled (and analyzed and visualized) the tweet archive dataset of Twitter user dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

As part of this project, I have had to do the below steps:

- Data wrangling
    - Gathering data
    - Assessing data - quality & tidiness
    - Cleaning data
- Storing, analyzing, and visualizing wrangled data
- Reporting on
    - data wrangling efforts and
    - data analysis and visualizations

This report pertains to my data analysis and visualization efforts.

Storing, analyzing, and visualizing wrangled data:

a. Storing:

As part of this project, the cleaned data were stored into multiple csv files:

- twitter_merged_master.csv → This is the merged data created at the end of cleaning process.
- twitter_archive_master.csv → This is twitter archive dataset after cleaning process.
- twitter_image_prediction_master.csv → This is tweet image predictions dataset after cleaning process.
- twitter_additional_details_master.csv → This is tweet additional information after cleaning process.

The merged data was moved into another dataframe after some columns (namely 'source', 'text', 'rating_numerator', 'rating_denominator', 'jpg_url' and 'img_num') were deleted that was not needed for my analysis and visualization.

b. Analyzing and visualizing data:

I have analyzed the following:

- **Descriptive Statistics:**
  This gives some statistics like:
  - The max rating is 1.4 and min is 0.1; the average rating is 1.05
  - The average number of retweet is 2339; min is 11 and max is 73560
  - The mean of favourite count is 7991; min 69 and max 149439
  - The p1_conf, p2_conf and p3_conf are within 0 and 1. This is correct as these values are supposed to be confidence interval which shows whether the p1, p2 and p3 are dogs or not.

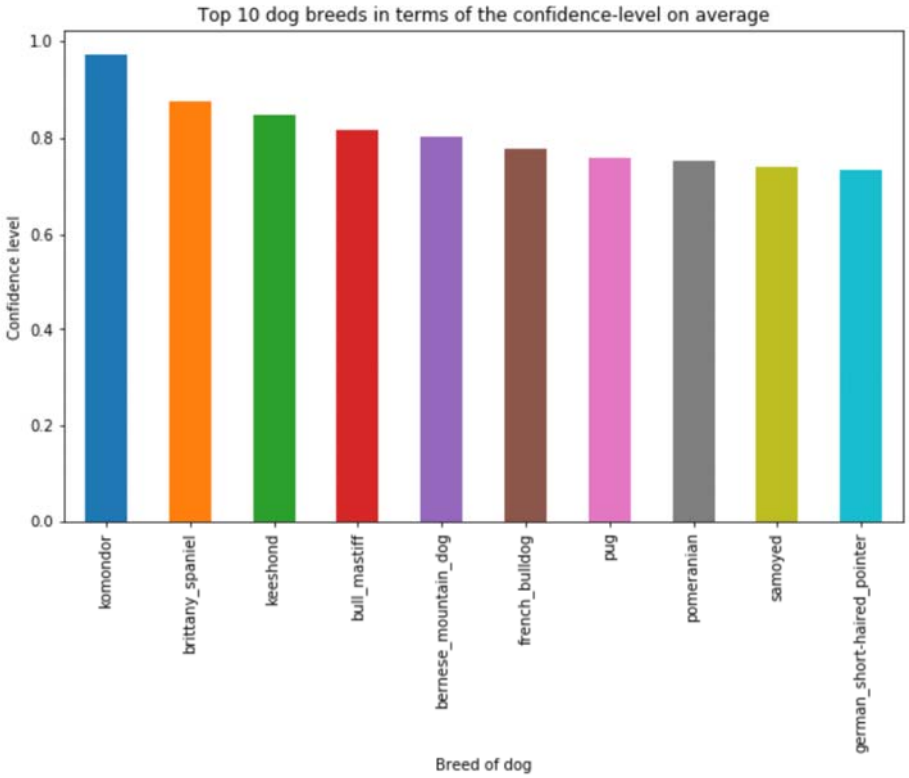
	rating	retweet_count	favorite_count	p1_conf	p2_conf	p3_conf
count	1941.0000	1941.0000	1941.0000	1941.0000	1941.0000	1941.0000
mean	1.0537	2339.0263	7991.7161	0.5935	0.1351	0.0603
std	0.2163	4193.2247	11787.2881	0.2722	0.1011	0.0509
min	0.1000	11.0000	69.0000	0.0443	0.0000	0.0000
25%	1.0000	520.0000	1684.0000	0.3604	0.0539	0.0161
50%	1.1000	1128.0000	3599.0000	0.5873	0.1182	0.0495
75%	1.2000	2674.0000	9968.0000	0.8453	0.1964	0.0917
max	1.4000	73560.0000	149439.0000	1.0000	0.4880	0.2710

- **Top 10 dog (names) that have highest mean rating:**
  All the below dog names have an average rating of 1.4

```
name
Kuyu        1.4
Smiley      1.4
Iggy        1.4
Sundance    1.4
General     1.4
Doobert     1.4
Clifford    1.4
Cermet      1.4
Laika       1.4
Emmy        1.4
Name: rating, dtype: float64
```

- **Top 10 dog breeds in terms of the confidence-level on average**
  The dog breed with highest average confidence level is komondor at 97% confidence, followed by brittany_spaniel with confidence level at 87%.

  Please note: This is populated using only p1 dog breeds with p1_dog as True which indicates p1 is a dog.



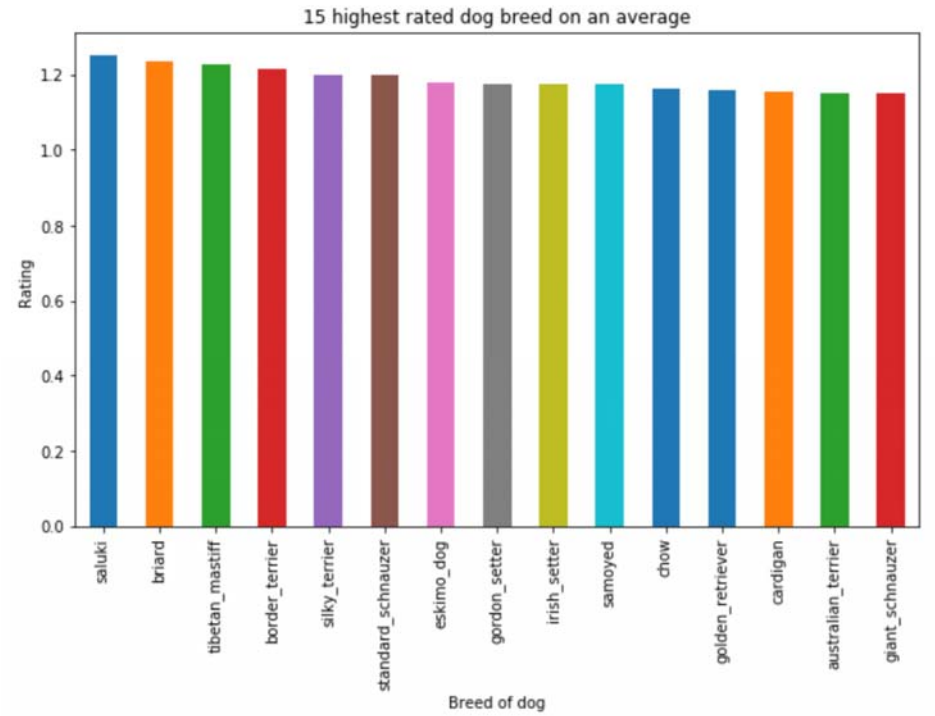

```
p1
komondor                    0.972531
brittany_spaniel            0.874545
keeshond                    0.844431
bull_mastiff                0.815618
bernese_mountain_dog        0.801816
french_bulldog              0.777413
pug                         0.759223
pomeranian                  0.751073
samoyed                     0.740719
german_short-haired_pointer 0.732425
Name: p1_conf, dtype: float64
```

- **15 highest rated dog breed on an average.**

  The highest rated dog breed on an average is saluki with a rating of 1.25 followed by briard with rating of 1.23

  Please note: This is populated using only p1 dog breeds with p1_dog as True which indicates p1 is a dog.

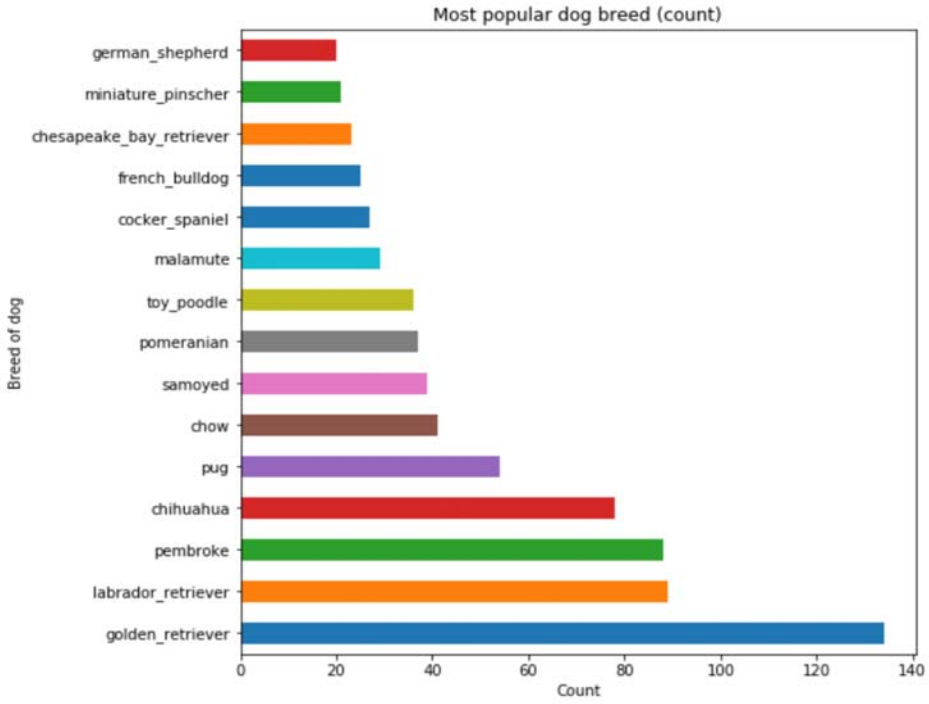
```
p1
saluki                1.250000
briard                1.233333
tibetan_mastiff       1.225000
border_terrier        1.214286
silky_terrier         1.200000
standard_schnauzer    1.200000
eskimo_dog            1.177778
gordon_setter         1.175000
irish_setter          1.175000
samoyed               1.174359
chow                  1.160976
golden_retriever      1.156716
cardigan              1.152941
australian_terrier    1.150000
giant_schnauzer       1.150000
Name: rating, dtype: float64
```



- **Most popular dog breed (count)**

  The most popular dog breed on the basis of count is golden_retriever at 134 followed by labrador_retriever at 89

  Please note: This is populated using only p1 dog breeds with p1_dog as True which indicates p1 is a dog.



```
golden_retriever           134
labrador_retriever          89
pembroke                    88
chihuahua                   78
pug                         54
chow                        41
samoyed                     39
pomeranian                  37
toy_poodle                  36
malamute                    29
cocker_spaniel              27
french_bulldog              25
chesapeake_bay_retriever    23
miniature_pinscher          21
german_shepherd             20
Name: p1, dtype: int64
```

- **Most common dog names**

  The most common dog names seem to be Cooper, Charlie and Oliver. The names 'None' and 'a' were fetched incorrectly while gathering data. This was not cleaned as there were too many of them.

```
None       510
a           54
Charlie     10
Oliver      10
Cooper      10
Name: name, dtype: int64
```

- **Which dog breeds have the highest average retweet and favorite-count**

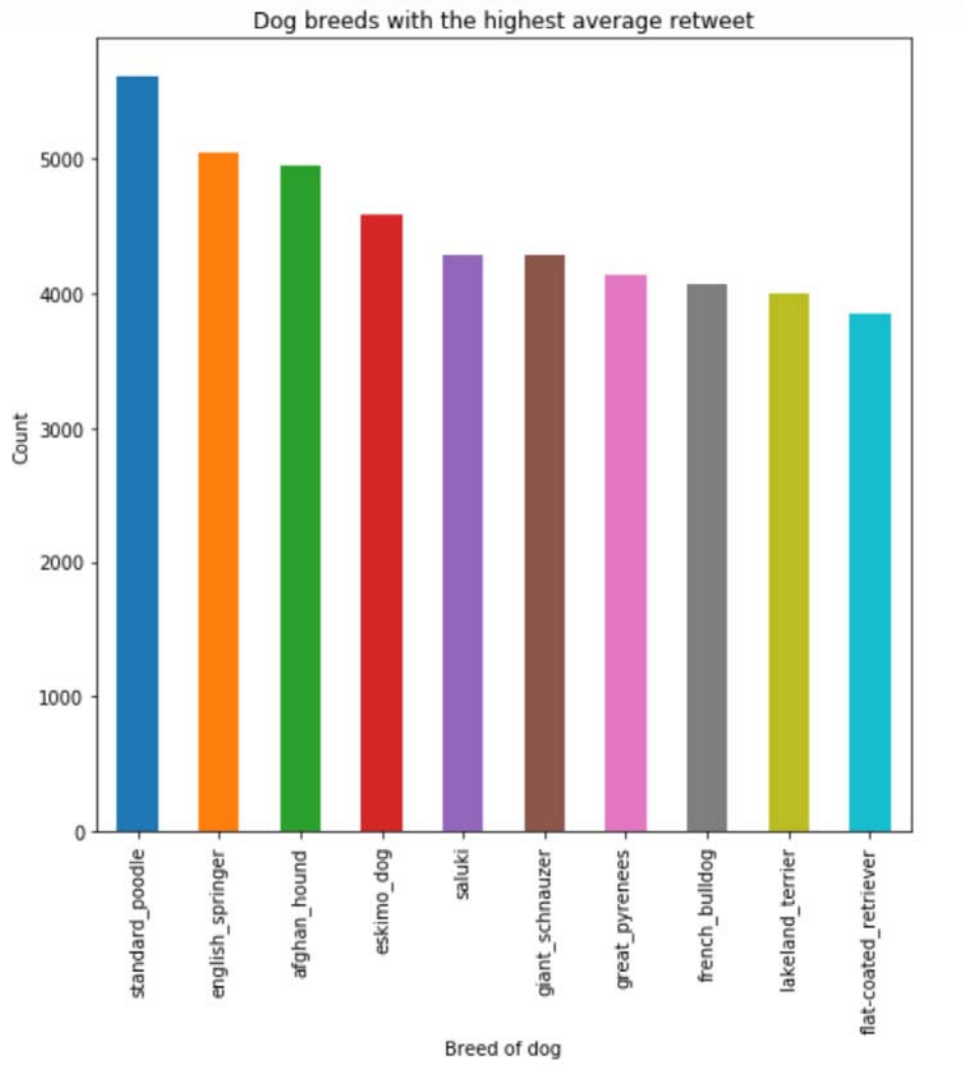
  On average the breed standard_poodle appears to be retweeted most followed by english_springer.

  On average the breed saluki appears to be most favorite followed by french_bulldog

  Please note: This is populated using only p1 dog breeds with p1_dog as True which indicates p1 is a dog.

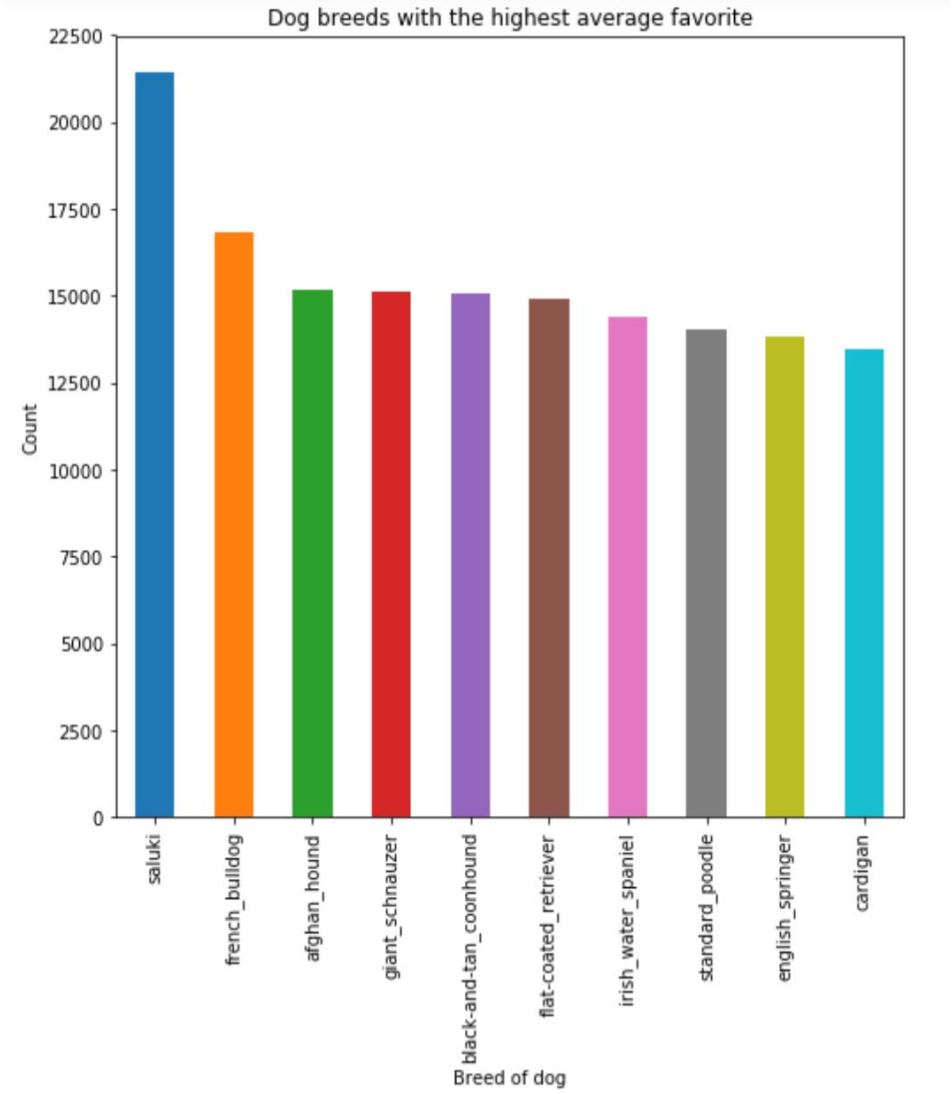
  Retweet:

```
p1
standard_poodle          5610.000000
english_springer         5043.111111
afghan_hound             4954.666667
eskimo_dog               4578.833333
saluki                   4285.500000
giant_schnauzer          4281.500000
great_pyrenees           4135.153846
french_bulldog           4070.760000
lakeland_terrier         4001.533333
flat-coated_retriever    3847.500000
Name: retweet_count, dtype: float64
```

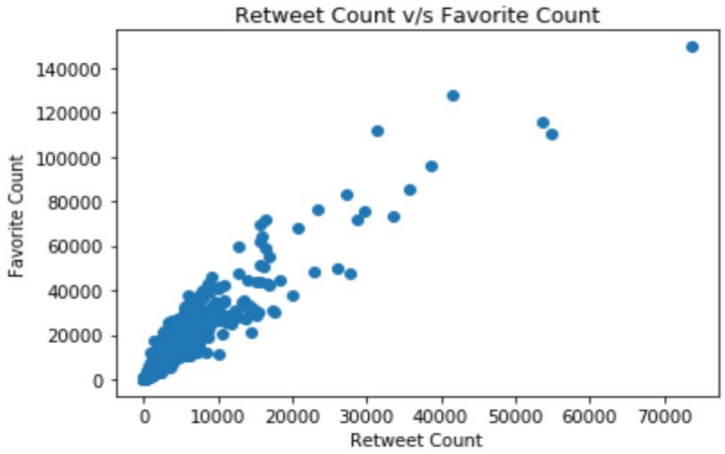


Favorite:

```
p1
saluki                      21422.000000
french_bulldog              16829.480000
afghan_hound                15163.333333
giant_schnauzer             15132.000000
black-and-tan_coonhound     15087.000000
flat-coated_retriever       14942.125000
irish_water_spaniel         14393.666667
standard_poodle             14039.571429
english_springer            13852.666667
cardigan                    13450.647059
Name: favorite_count, dtype: float64
```

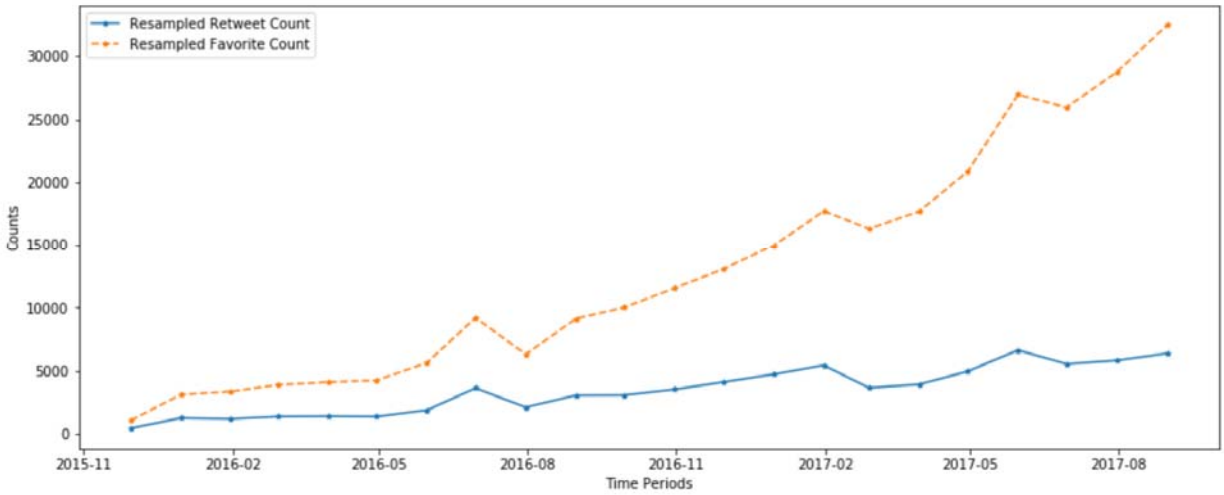


- **The retweet counts and favorite counts comparison using scatter plot**
  There is a positive correlation between retweet count and favorite count.



- **Plot time series - retweet count and favorite count**

  The time series graph shows a positive correlation between favorite_count and retweet_count. Looks like people retweet posts/tweets they like so that others can also view them in their feed.



resampled retweet data (by month)

```
timestamp
2015-11-30     437.774744
2015-12-31    1258.612360
2016-01-31    1195.280488
2016-02-29    1372.358491
2016-03-31    1399.339130
2016-04-30    1370.192308
2016-05-31    1878.727273
2016-06-30    3626.602564
2016-07-31    2104.541176
2016-08-31    3060.067797
2016-09-30    3073.983607
2016-10-31    3522.421875
2016-11-30    4102.673077
2016-12-31    4718.528302
2017-01-31    5431.707692
2017-02-28    3646.096774
2017-03-31    3924.155556
2017-04-30    4957.525000
2017-05-31    6630.000000
2017-06-30    5535.767442
2017-07-31    5812.224490
2017-08-31    6369.000000
Freq: M, Name: retweet_count, dtype: float64
```

resampled favorite data (by month)

```
timestamp
2015-11-30     1056.883959
2015-12-31     3115.014045
2016-01-31     3358.689024
2016-02-29     3886.688679
2016-03-31     4097.400000
2016-04-30     4243.365385
2016-05-31     5633.600000
2016-06-30     9201.012821
2016-07-31     6304.564706
2016-08-31     9122.847458
2016-09-30    10019.409836
2016-10-31    11546.468750
2016-11-30    13049.115385
2016-12-31    14946.528302
2017-01-31    17716.261538
2017-02-28    16316.516129
2017-03-31    17691.133333
2017-04-30    20880.975000
2017-05-31    26956.476190
2017-06-30    25946.906977
2017-07-31    28750.102041
2017-08-31    32466.500000
Freq: M, Name: favorite_count, dtype: float64
```

Conclusion:

There are many more insights that you can infer from the dataset. I have visualized a few insights from the cleaned dataset as part of this project. The web page admin could use these insights in numerous ways like to increase user traffic or to use this for targeted marketing or to enhance the website to make it more user friendly and so on.