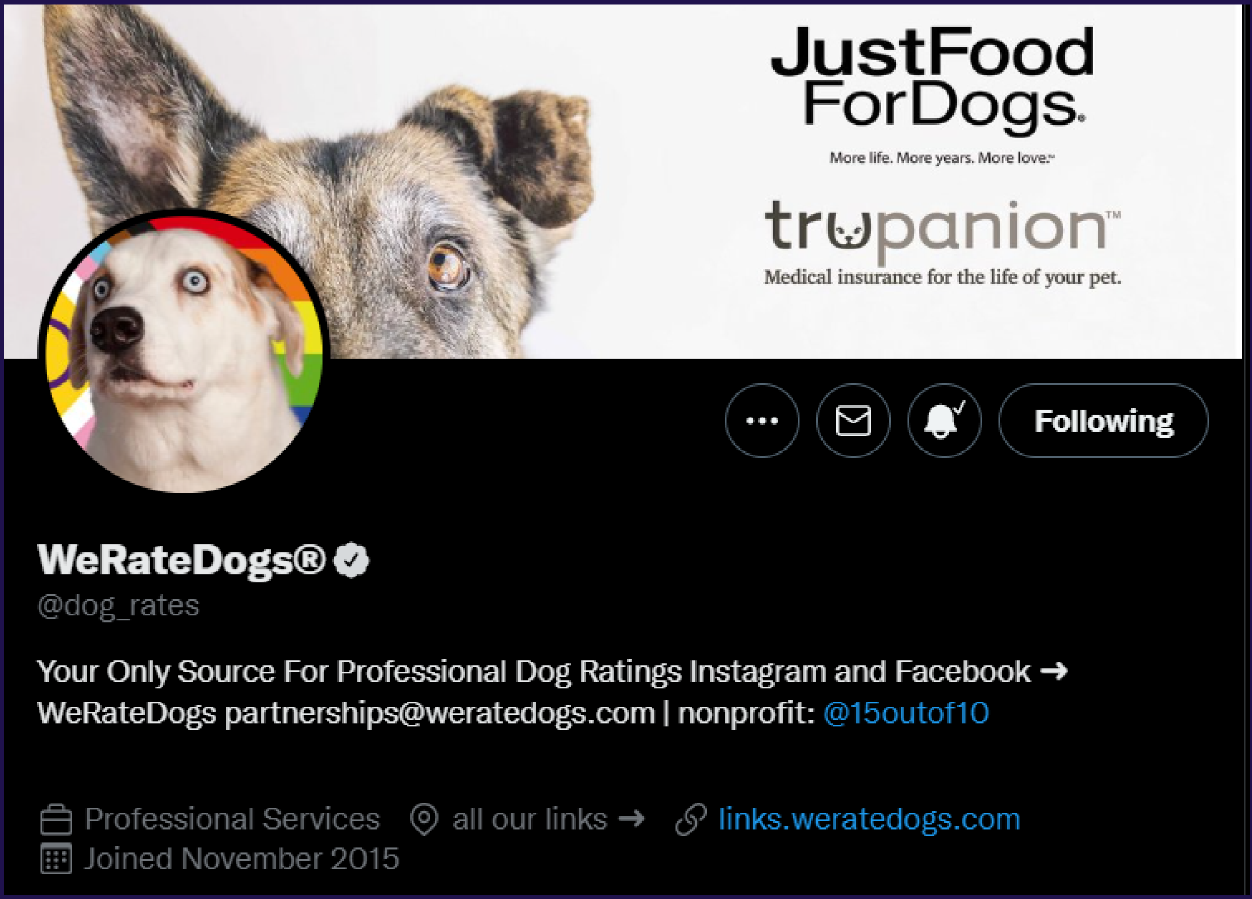
Data Analyst Nanodegree: Project Report

ANALYZING AND VISUALIZING INSIGHTS FROM WERATEDOGS' TWITTER ARCHIVE

Tevin Aduma 29/06/22


JustFood ForDogs.
More life. More years. More love.
trupanion
Medical insurance for the life of your pet.
Following

WeRateDogs®

@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook ➜ WeRateDogs partnerships@weratedogs.com | nonprofit: @15outof10

Professional Services all our links ➜ links.weratedogs.com

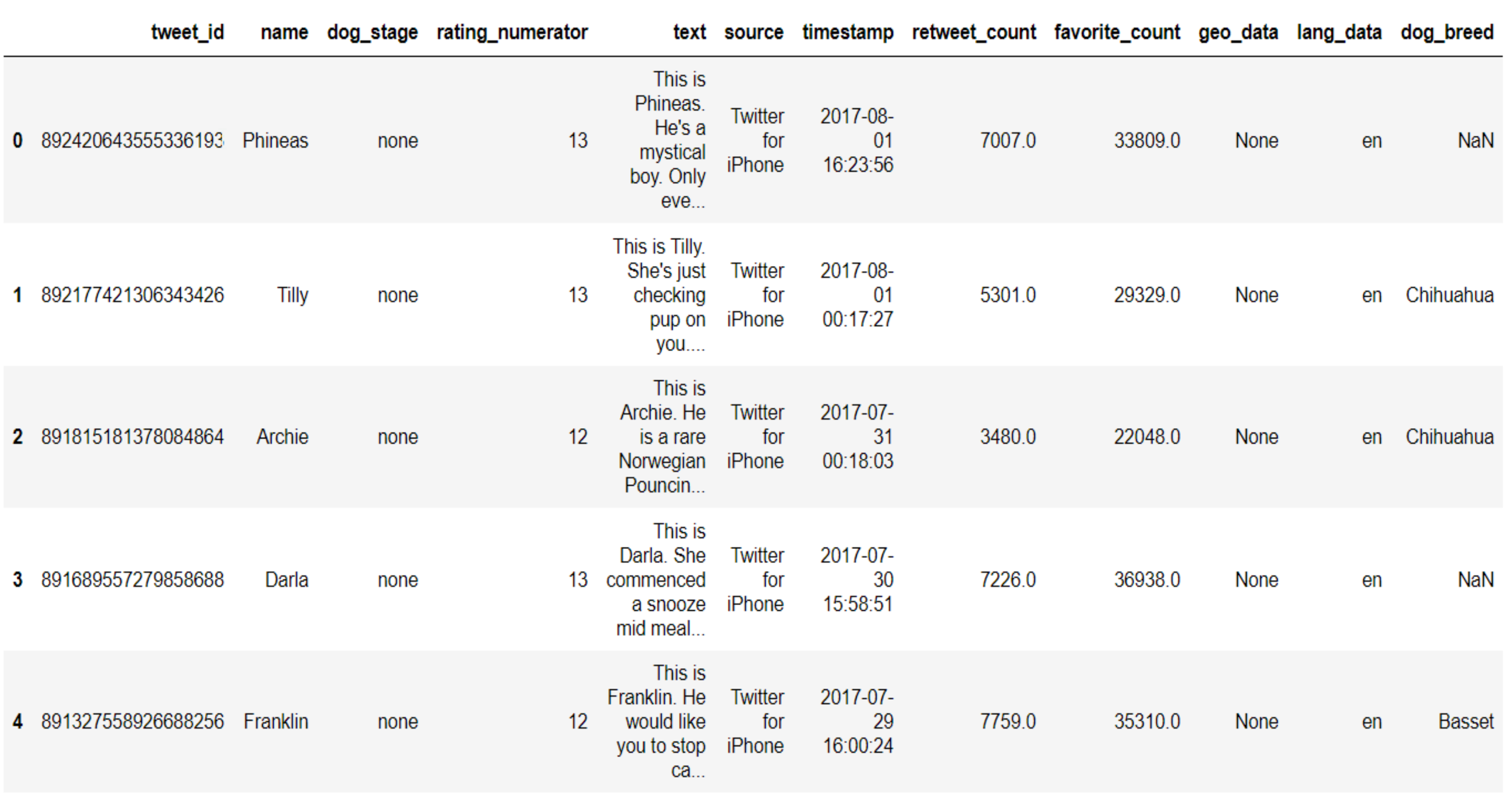
Joined November 2015

Introduction

This document entails the Exploratory Data Analysis techniques to obtain insights from hypotheses I put forward prior to wrangling WeRateDogs' (later referred to as WRD) Twitter archive dataset.

Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as 13/10, 14/10. etc.

This is the master dataset that I will be exploring for actionable insights 👇🏾



	tweet_id	name	dog_stage	rating_numerator	text	source	timestamp	retweet_count	favorite_count	geo_data	lang_data	dog_breed
0	892420643555336193	Phineas	none	13	This is Phineas. He's a mystical boy. Only eve...	Twitter for iPhone	2017-08-01 16:23:56	7007.0	33809.0	None	en	NaN
1	892177421306343426	Tilly	none	13	This is Tilly. She's just checking pup on you....	Twitter for iPhone	2017-08-01 00:17:27	5301.0	29329.0	None	en	Chihuahua
2	891815181378084864	Archie	none	12	This is Archie. He is a rare Norwegian Pouncin...	Twitter for iPhone	2017-07-31 00:18:03	3480.0	22048.0	None	en	Chihuahua
3	891689557279858688	Darla	none	13	This is Darla. She commenced a snooze mid meal...	Twitter for iPhone	2017-07-30 15:58:51	7226.0	36938.0	None	en	NaN
4	891327558926688256	Franklin	none	12	This is Franklin. He would like you to stop ca...	Twitter for iPhone	2017-07-29 16:00:24	7759.0	35310.0	None	en	Basset

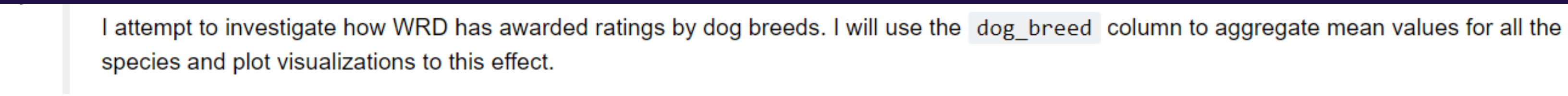
I am going to examine the following features from an efficiently enriched dataset that has been wrangled (but it is important to remember that is an iterative process):

- WRD's ratings for each dog in order to see what breeds Matt and his gang have a liking to.
- Engagement numbers for each dog to gauge what breeds the Twitter audience is quite fond of.
- Words in WRD's tweets to form a wordcloud and ascertain their sense of polarity and subjectivity.

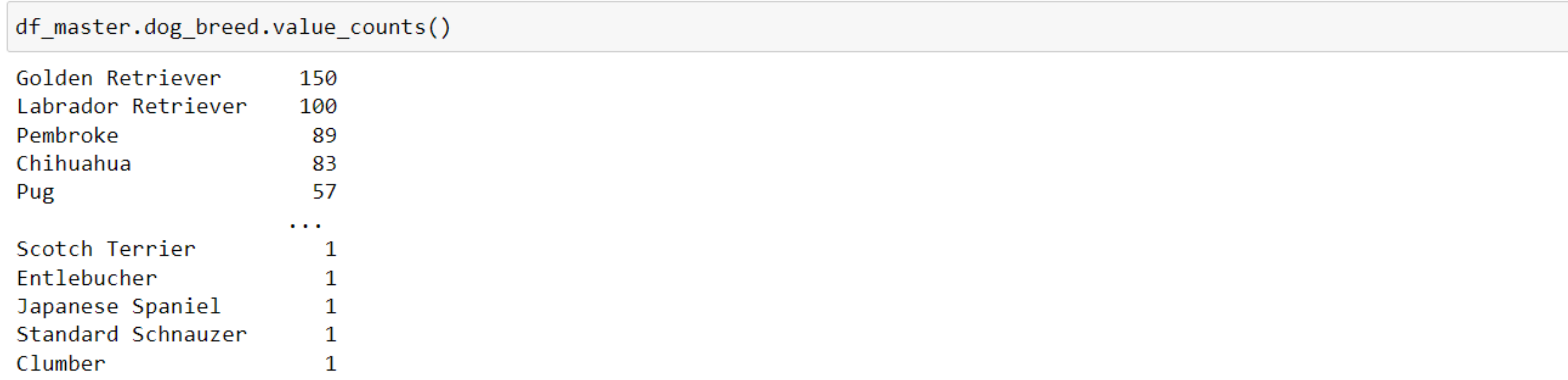
Since Krypto is getting a movie release this year, I will grant a Marvel hero moniker for the dog that will feature the most in our top 10 through all the years on an aggregate and also by the yearly analyses i.e 2016 and 2017 i.e. Tony Bark 🐶

Q1: Which dog breeds have been awarded the highest ratings? 🐾

For this analysis, I will only use records that have values in the `dog_breed` column so all records that have null values in this column will not be used.



> I attempt to investigate how WRD has awarded ratings by dog breeds. I will use the `dog_breed` column to aggregate mean values for all the species and plot visualizations to this effect.

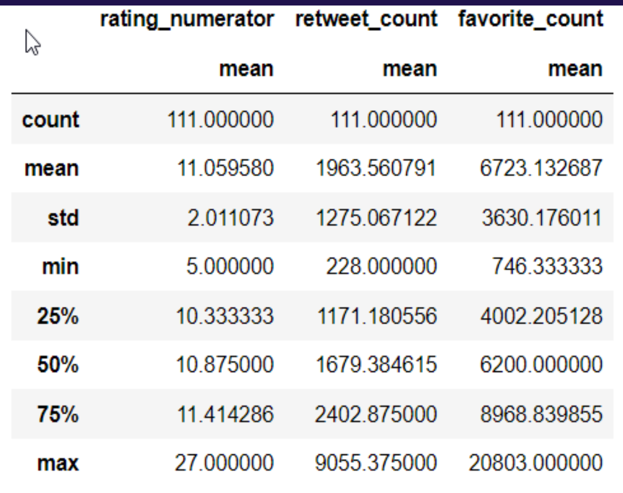


```
df_master.dog_breed.value_counts()
```

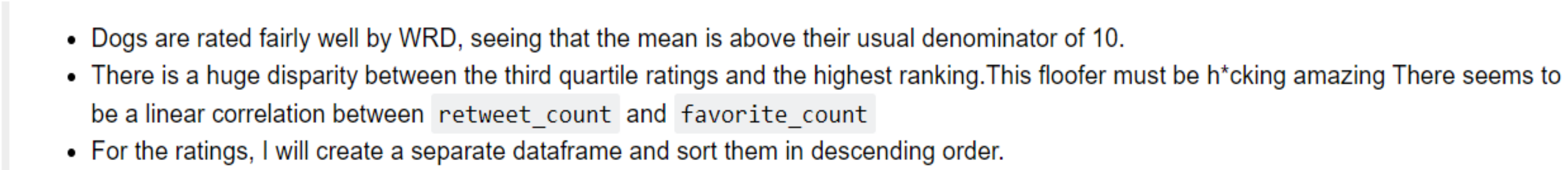
```
Golden Retriever      150
Labrador Retriever    100
Pembroke               89
Chihuahua              83
Pug                    57
                     ...
Scotch Terrier          1
Entlebucher             1
Japanese Spaniel        1
Standard Schnauzer      1
Clumber                 1
```

- Assuming that the neural network was accurate, Golden Retrievers are the most common breeds rated by WRD in our dataset.

I intend to make another dataframe of the aggregate engagements, `df_agg_stats` that will group all the records by their `dog_breed` and calculate the mean for their `rating_numerator`, `retweet_count` and `favorite_count`

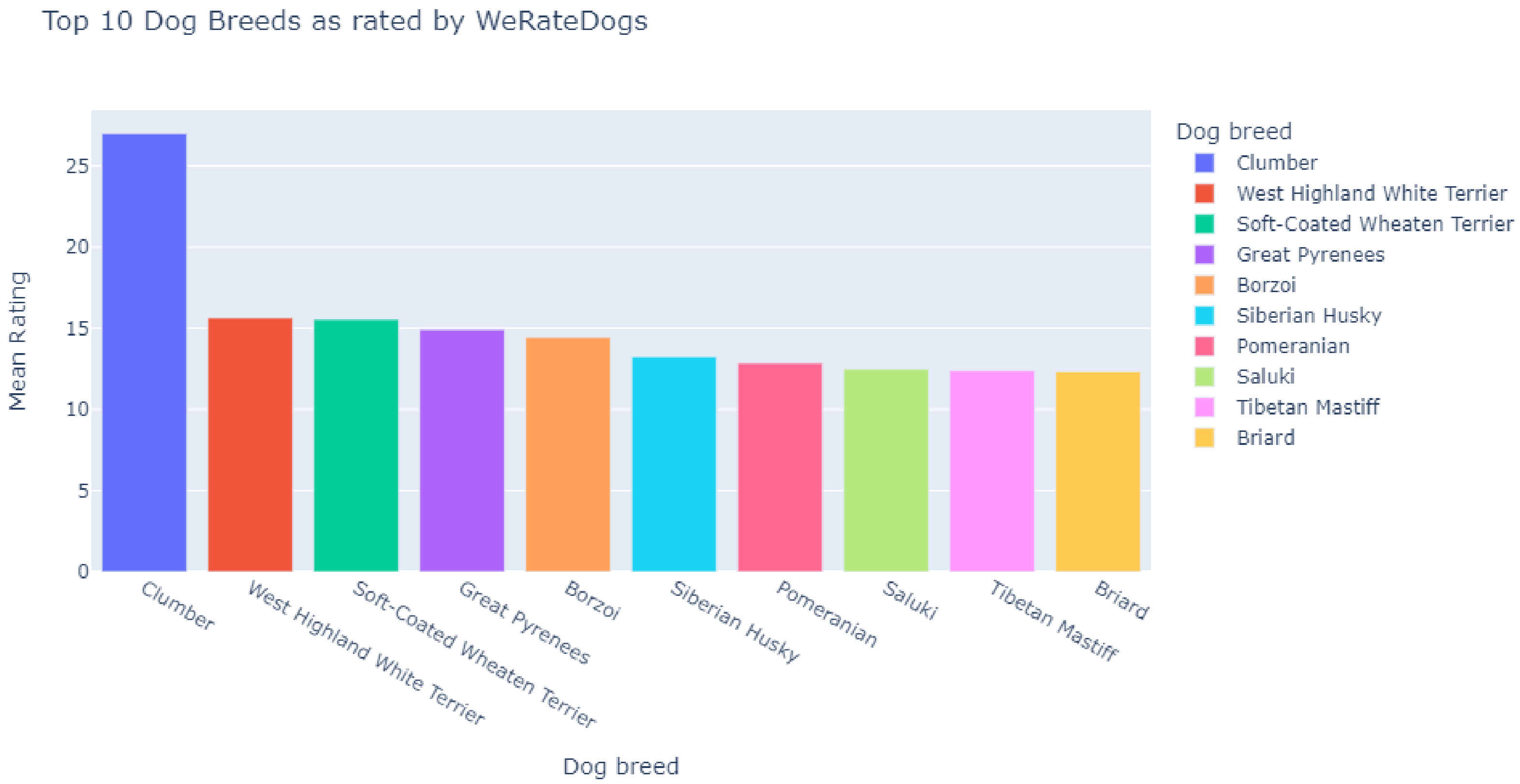


	rating_numerator	retweet_count	favorite_count
	mean	mean	mean
count	111.000000	111.000000	111.000000
mean	11.059580	1963.560791	6723.132687
std	2.011073	1275.067122	3630.176011
min	5.000000	228.000000	746.333333
25%	10.333333	1171.180556	4002.205128
50%	10.875000	1679.384615	6200.000000
75%	11.414286	2402.875000	8968.839855
max	27.000000	9055.375000	20803.000000



- Dogs are rated fairly well by WRD, seeing that the mean is above their usual denominator of 10.
- There is a huge disparity between the third quartile ratings and the highest ranking.This floofer must be h*cking amazing There seems to be a linear correlation between `retweet_count` and `favorite_count`
- For the ratings, I will create a separate dataframe and sort them in descending order.

Here are the top 10 dog breeds by rating 👇🏾 🐕



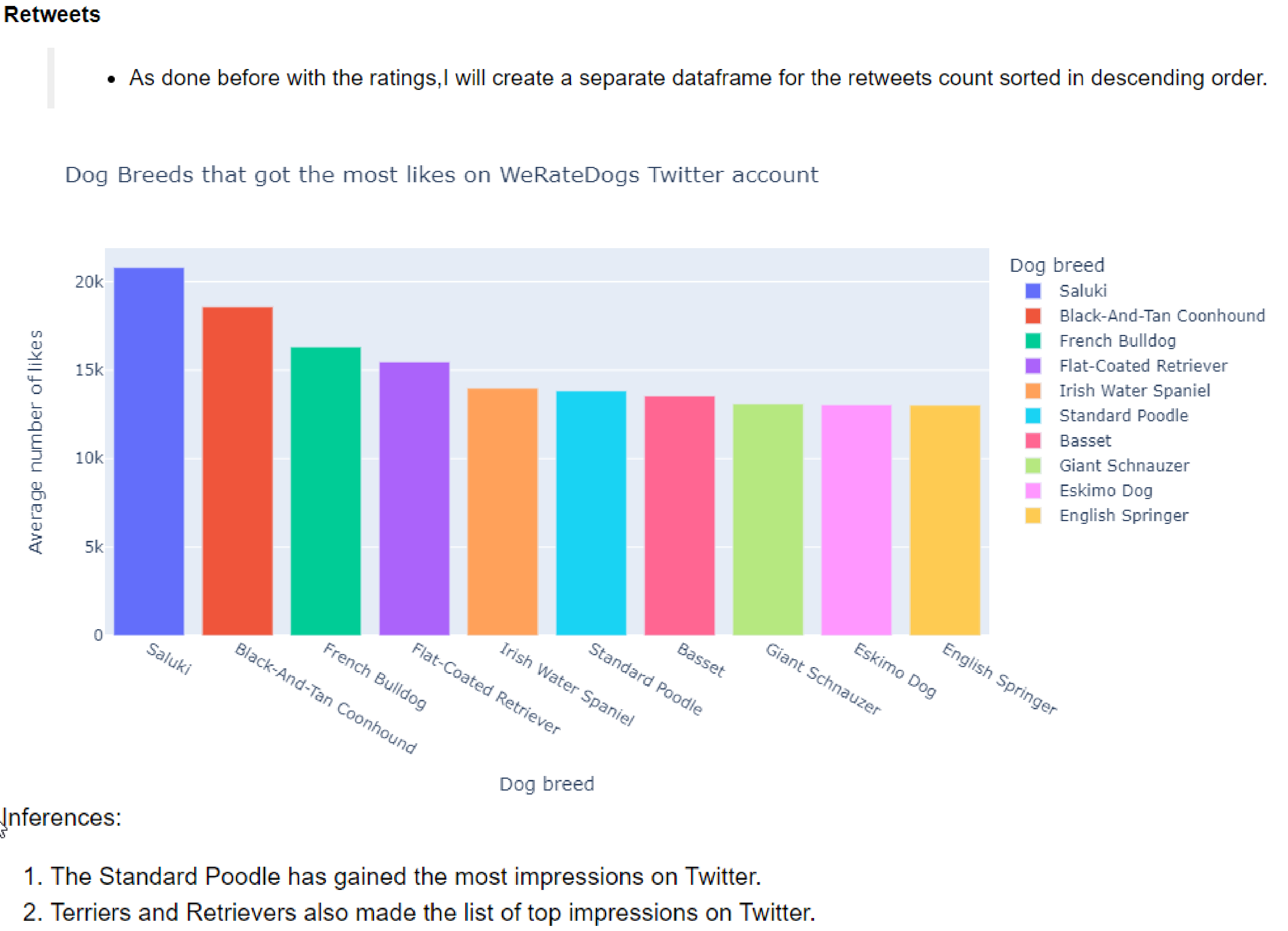
Inferences:

1. The Clumber despite being the only 1 on our dataset holds the top spot for the highest average rating.
2. Terriers, Pyrennes, Borzoi and the Husky are pretty popular dogs.
3. The Japanese Spaniel is the lowest ranked dog breed on WRD.

Q2: Which dog breeds have attracted the most engagement on WRD? 🐾

I will visualize the engagements by two spectrums:

- Retweets 🔁
- Favorites ❤️



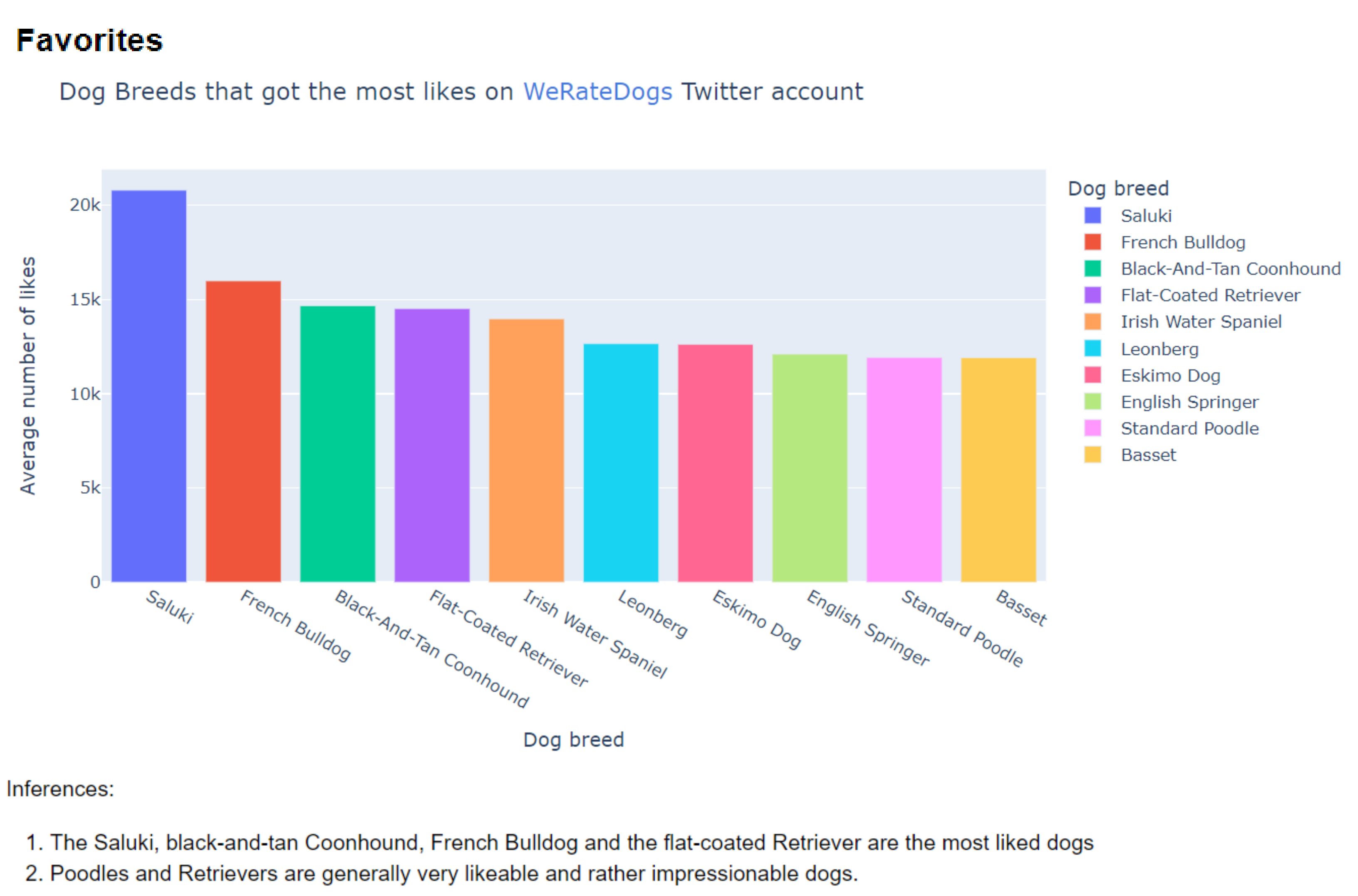
Retweets

> - As done before with the ratings,I will create a separate dataframe for the retweets count sorted in descending order.

Inferences:

1. The Standard Poodle has gained the most impressions on Twitter.
2. Terriers and Retrievers also made the list of top impressions on Twitter.

- The Standard Poodle is outrightly the most impressionable dog on WRD so far.



Favorites

Dog Breeds that got the most likes on WeRateDogs Twitter account

Inferences:

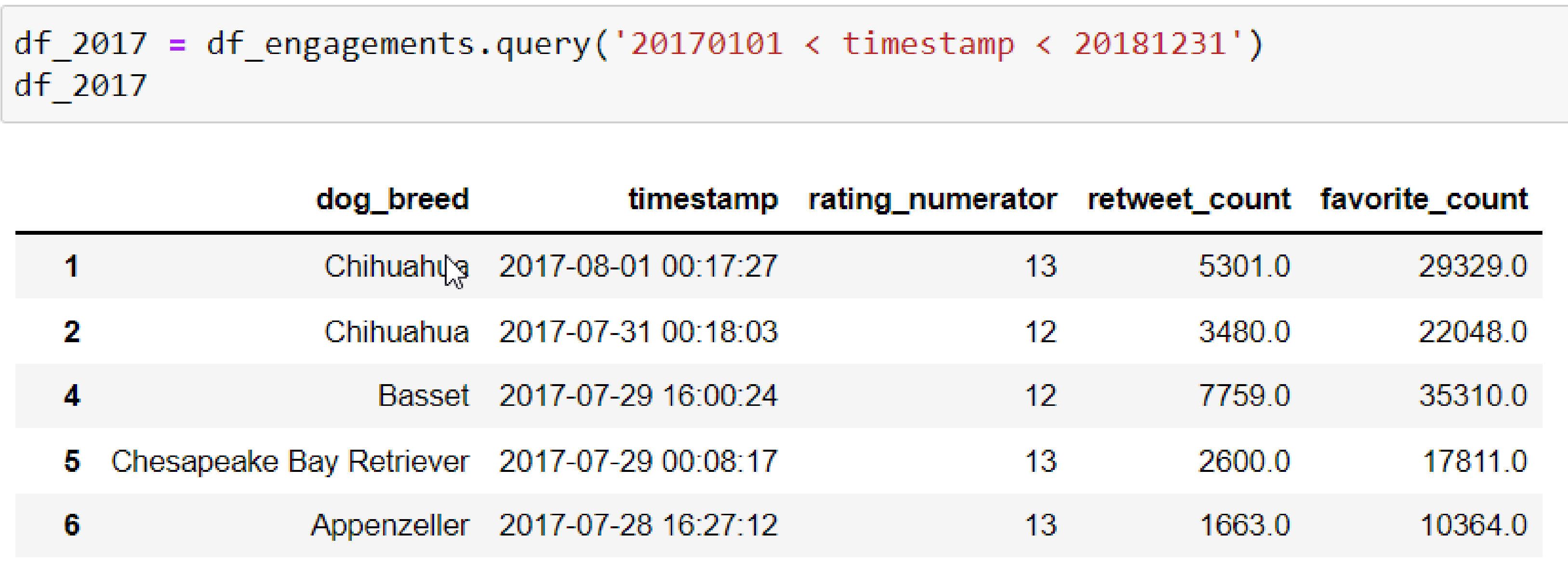
1. The Saluki, black-and-tan Coonhound, French Bulldog and the flat-coated Retriever are the most liked dogs
2. Poodles and Retrievers are generally very likeable and rather impressionable dogs.

- Based on these metrics, chances are someone would get a Saluki, Terrier or Retriever as their first dog due to their likability in nature.

Metrics aggregated in the most recent year: 2017

To get an in-depth analysis on the ratings, retweets and favorites, I classified the engagements dataframe through the years to see which dog breeds ranked highest over different periods.

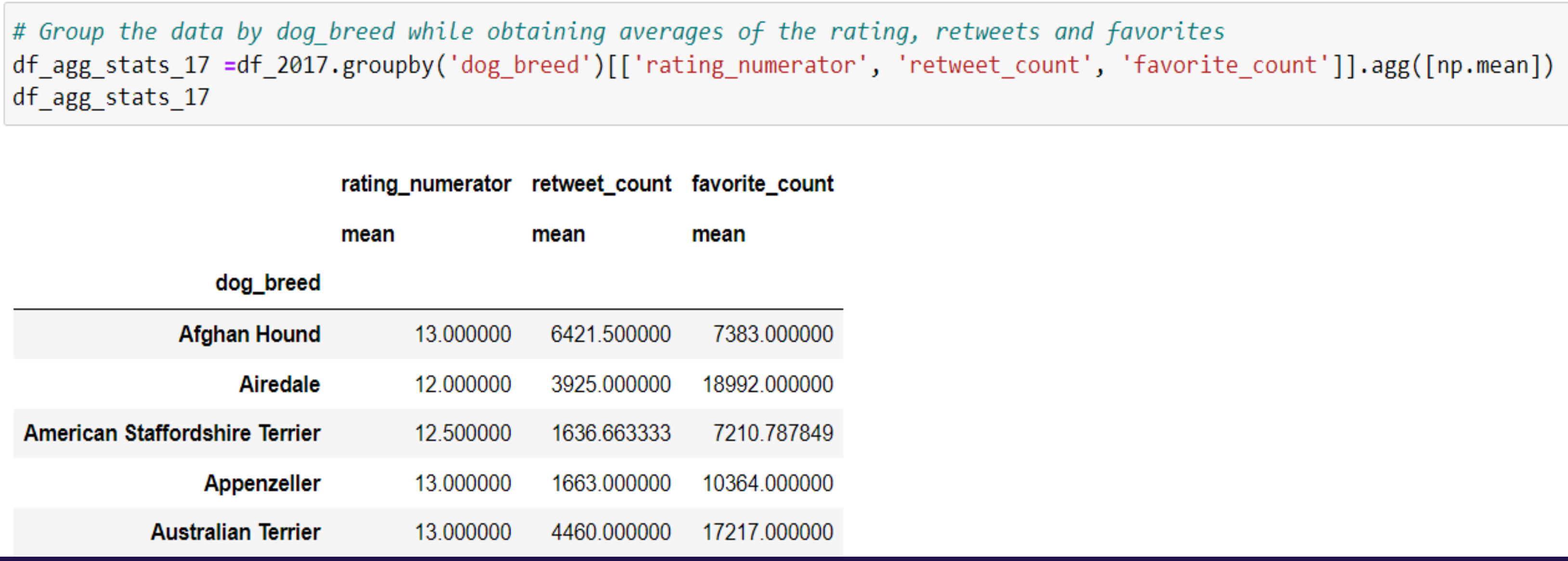
For purposes of being concise I show visualizations for 2017 only. Here's a snippet of how I query the data from the `engagements` dataframe and create separate years' aggregated.



```
df_2017 = df_engagements.query('20170101 < timestamp < 20181231')
df_2017
```

	dog_breed	timestamp	rating_numerator	retweet_count	favorite_count
1	Chihuahua	2017-08-01 00:17:27	13	5301.0	29329.0
2	Chihuahua	2017-07-31 00:18:03	12	3480.0	22048.0
4	Basset	2017-07-29 16:00:24	12	7759.0	35310.0
5	Chesapeake Bay Retriever	2017-07-29 00:08:17	13	2600.0	17811.0
6	Appenzeller	2017-07-28 16:27:12	13	1663.0	10364.0

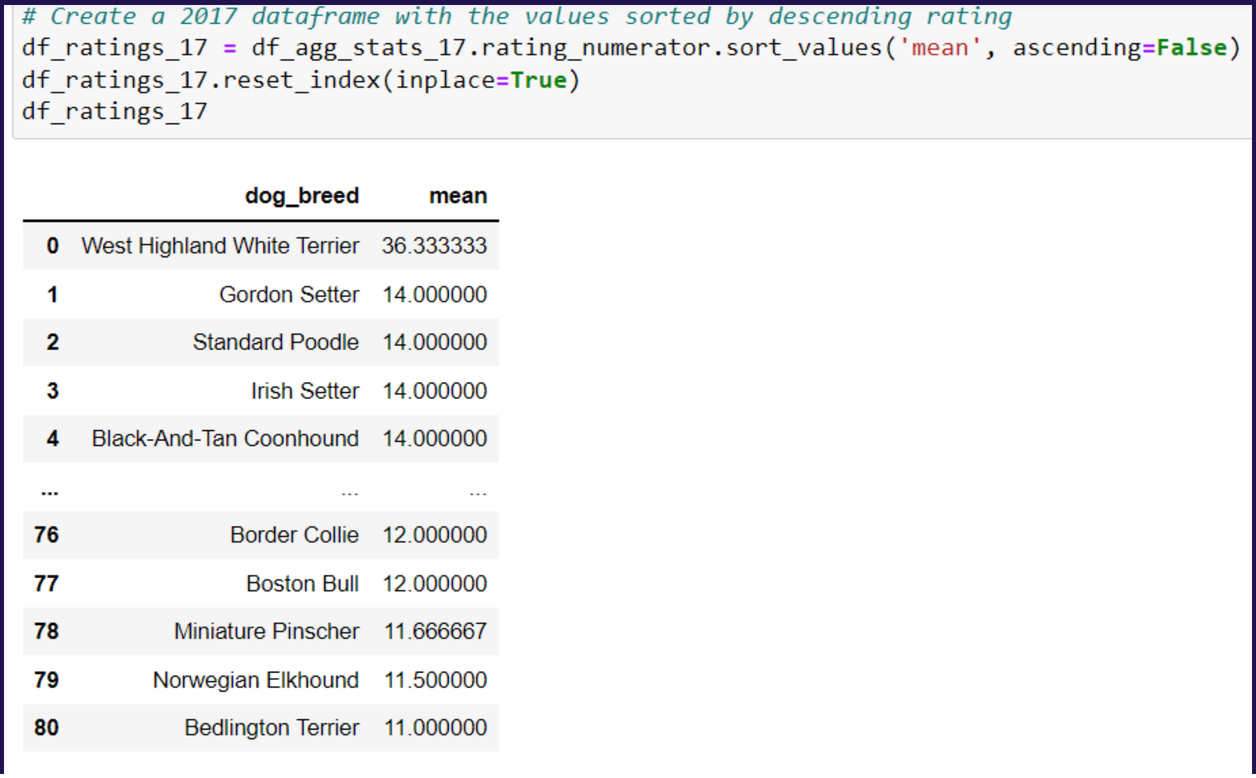
I will group the data by dog_breed and calculate aggregated statistics using Numpy's mean() function. Below is an illustration snippet code describing how I do this.



```
# Group the data by dog_breed while obtaining averages of the rating, retweets and favorites
df_agg_stats_17 =df_2017.groupby('dog_breed')[['rating_numerator', 'retweet_count', 'favorite_count']].agg([np.mean])
df_agg_stats_17
```

	rating_numerator	retweet_count	favorite_count
	mean	mean	mean
dog_breed			
Afghan Hound	13.000000	6421.500000	7383.000000
Airedale	12.000000	3925.000000	18992.000000
American Staffordshire Terrier	12.500000	1636.663333	7210.787849
Appenzeller	13.000000	1663.000000	10364.000000
Australian Terrier	13.000000	4460.000000	17217.000000

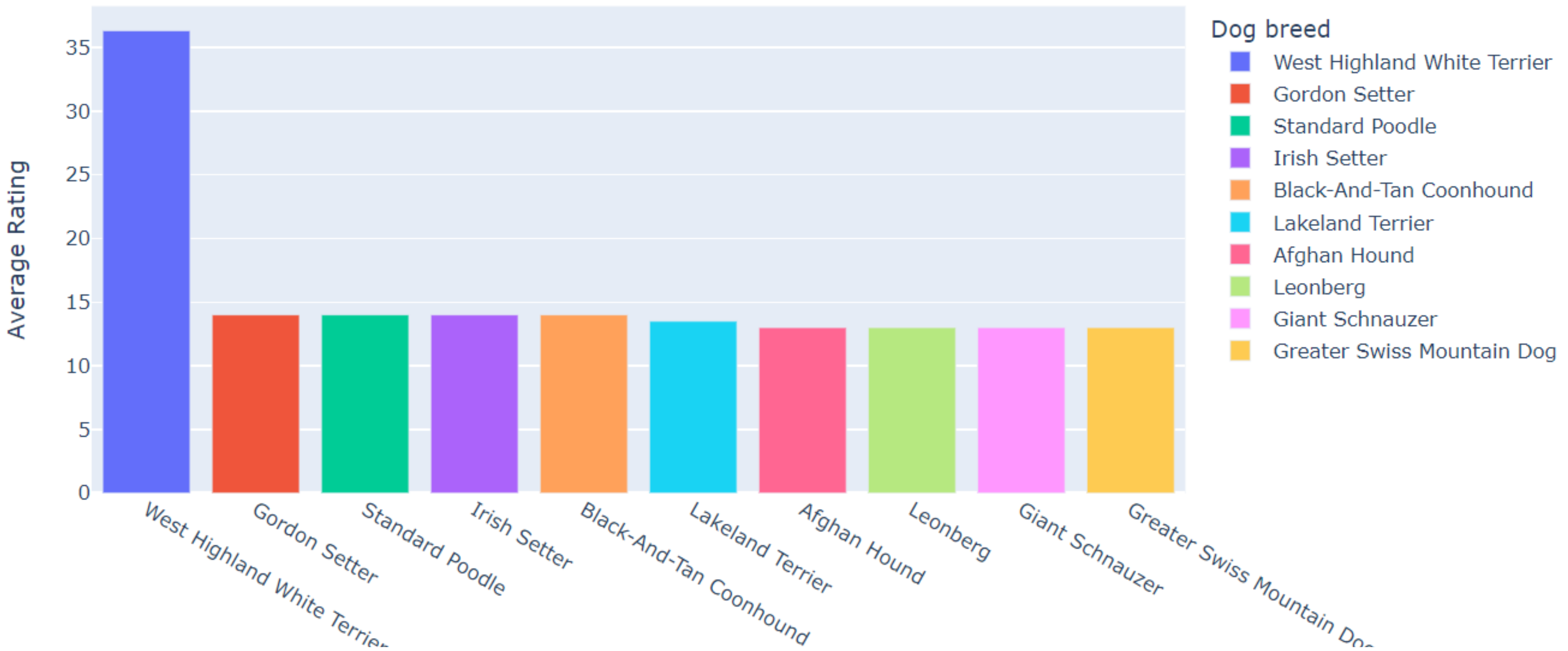
I will filter each feature out of the aggregated stats dataframes and sort their mean values by descending order.



```python
# Create a 2017 dataframe with the values sorted by descending rating
df_ratings_17 = df_agg_stats_17.rating_numerator.sort_values('mean', ascending=False)
df_ratings_17.reset_index(inplace=True)
df_ratings_17
```

	dog_breed	mean
0	West Highland White Terrier	36.333333
1	Gordon Setter	14.000000
2	Standard Poodle	14.000000
3	Irish Setter	14.000000
4	Black-And-Tan Coonhound	14.000000
...	...	...
76	Border Collie	12.000000
77	Boston Bull	12.000000
78	Miniature Pinscher	11.666667
79	Norwegian Elkhound	11.500000
80	Bedlington Terrier	11.000000

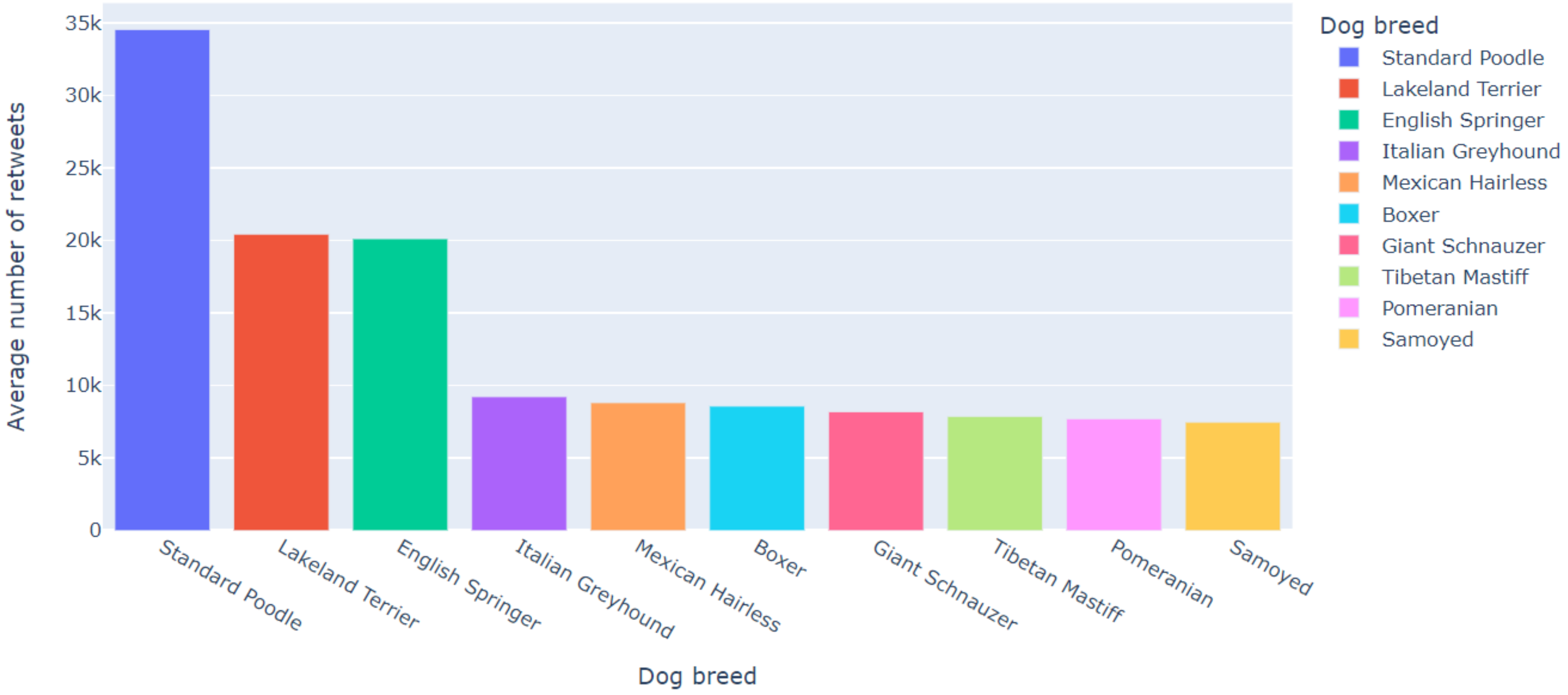
Dog Breeds that got the highest ratings on WeRateDogs in 2017



Inferences:

- The West Highland White Terrier, Standard Poodle, and Black-and Tan Coonhound are still among WRD's **most loved dogs** 🐶

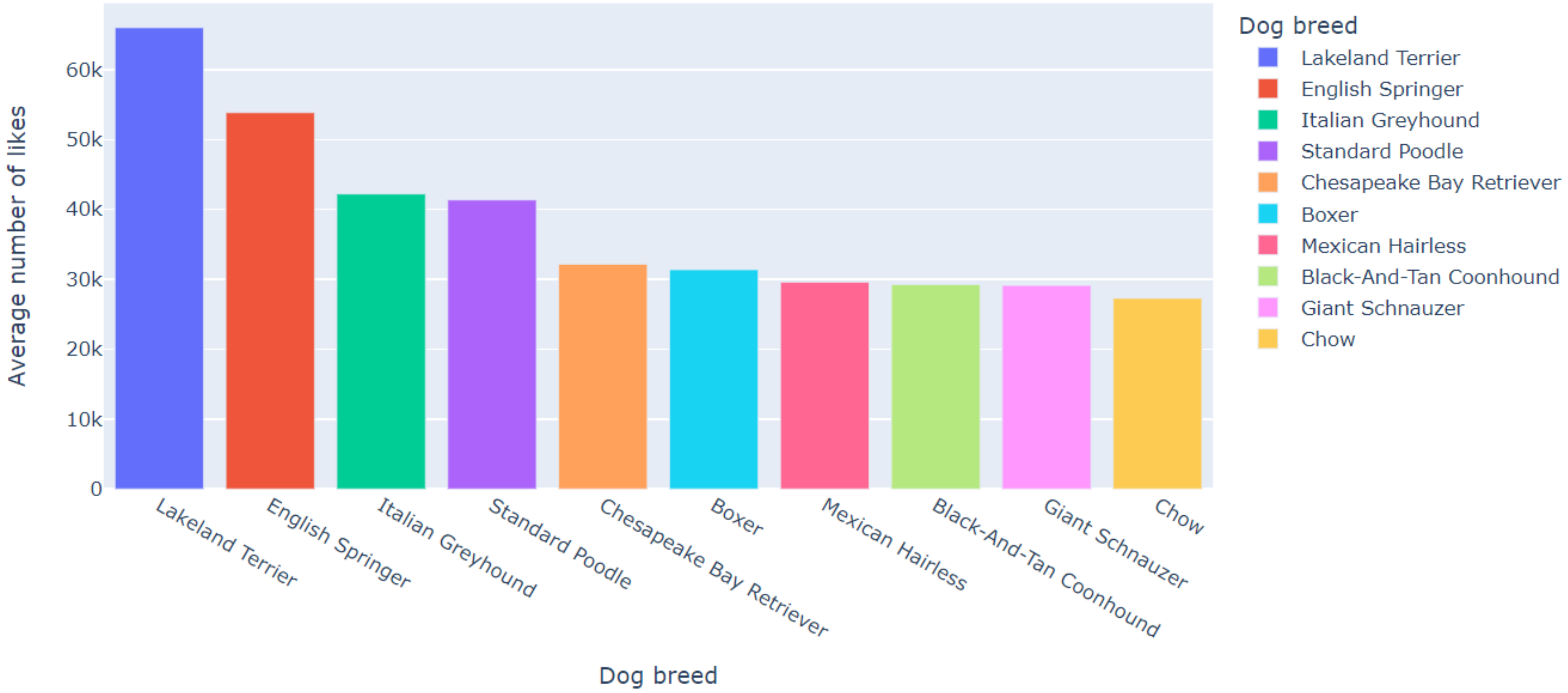
Dog Breeds that got the most retweets on WeRateDogs in 2017

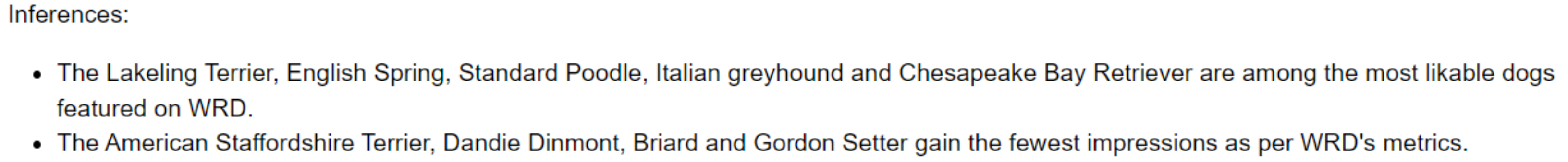


Inferences:

- The Standard Poodle was the **most impressionable dog of 2017**
- Generally, Terriers are **really impressionable and lovable dogs**

Dog Breeds that got the most likes on WeRateDogs in 2017





Inferences:

- The Lakeling Terrier, English Spring, Standard Poodle, Italian greyhound and Chesapeake Bay Retriever are among the most likable dogs featured on WRD.
- The American Staffordshire Terrier, Dandie Dinmont, Briard and Gordon Setter gain the fewest impressions as per WRD's metrics.

From these analyses, the Standard Poodle comes across as the best rated, most impressionable and most likable dog featured on WRD 🐶🐾

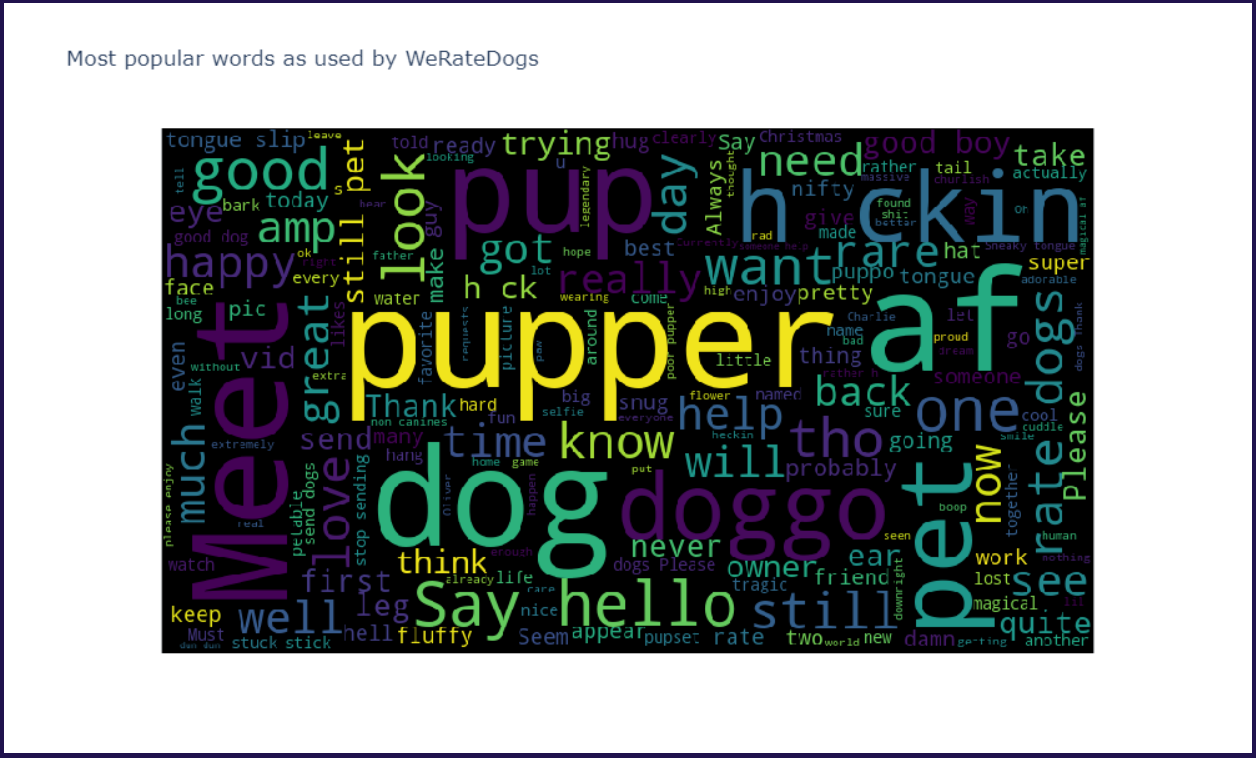
Here’s where I ‘m gonna crown it as **Poolie** 👑

Photo by Samia Liamani on Unsplash

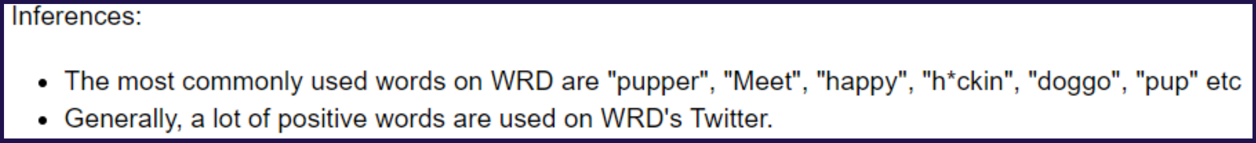
Q3: What are the most used words on WRD?

Word Clouds (also known as wordle, word collage, or tag cloud) are visual representations of words that give greater prominence to words that appear more frequently. The goal is to understand how WRD's author and audience feel about dogs and topics revolving around dogs. I will use the cleaned text to figure this out in addition to Python's WordCloud library

I intend to filter a dataset that does not contain any null values in the `text` column. The `timestamp` field will be used for a future analysis of sentiment by years.



Most popular words as used by WeRateDogs



Inferences:

- The most commonly used words on WRD are "pupper", "Meet", "happy", "h*ckin", "doggo", "pup" etc
- Generally, a lot of positive words are used on WRD's Twitter.

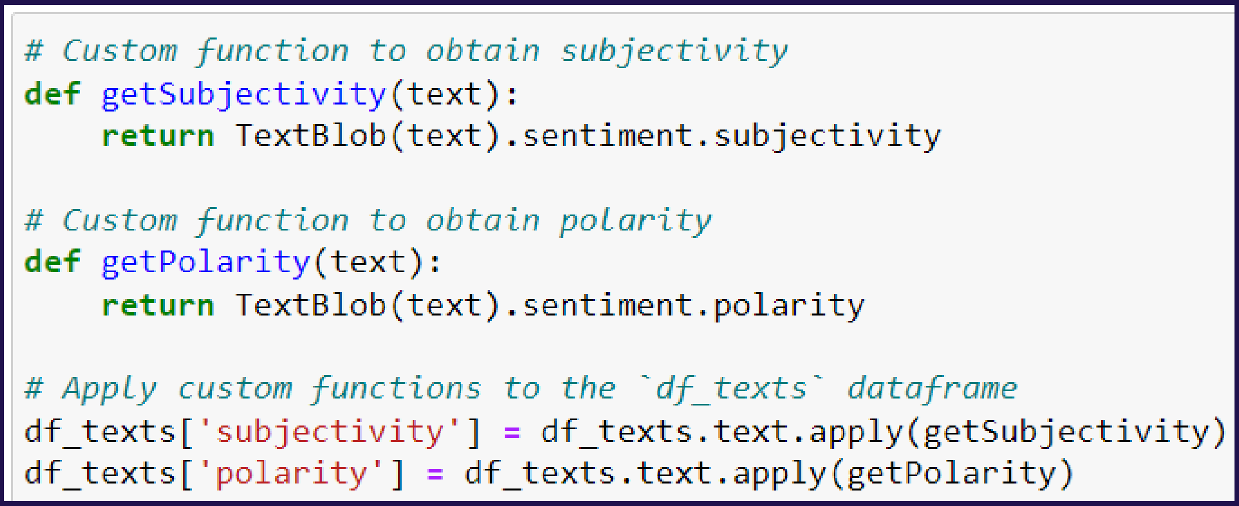
Q4: Generally, what's the sentiment given off by WRD? Is it positive, neutral or negative? Is it subjective (personal and opinionated) or objective (factual)?

I will use a binary classifier using the Twitter data to detect the sentiment of each tweet. The input data is the text and the library in use will be Python's TextBlob. I will get the score of each tweet's polarity.

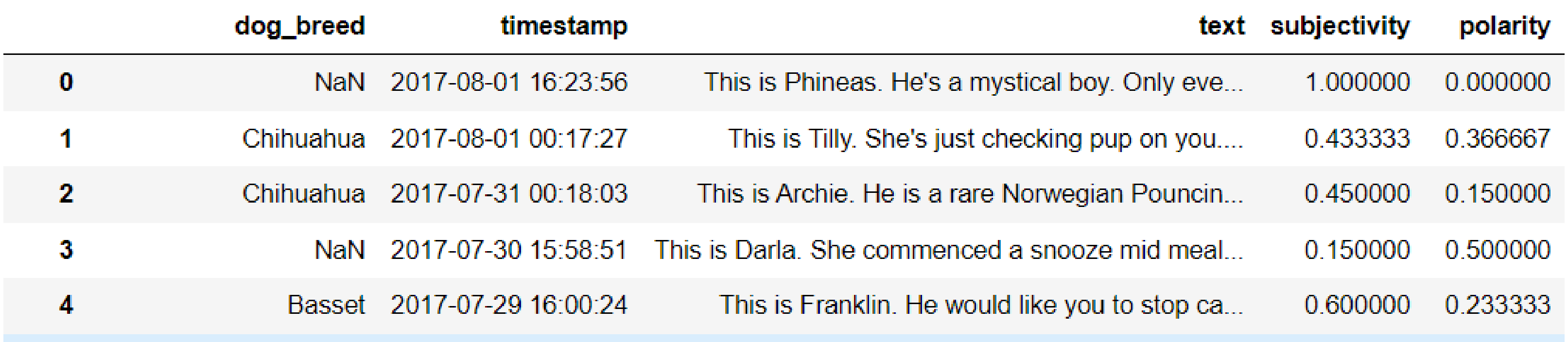
Polarity is the output that lies between [-1,1], where -1 refers to negative sentiment and +1 refers to positive sentiment.

Subjectivity quantifies the amount of personal opinion and factual information contained in the text. The higher subjectivity means that the text contains personal opinion rather than factual information. Subjectivity output that lies within [0,1] and refers to personal opinions and judgments.

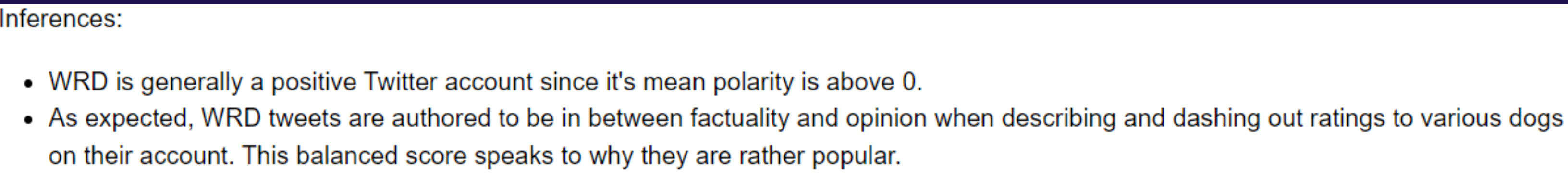
I will use two custom functions to obtain these scores into new columns named `subjectivity` and `polarity` respectively.



```python
# Custom function to obtain subjectivity
def getSubjectivity(text):
    return TextBlob(text).sentiment.subjectivity

# Custom function to obtain polarity
def getPolarity(text):
    return TextBlob(text).sentiment.polarity

# Apply custom functions to the `df_texts` dataframe
df_texts['subjectivity'] = df_texts.text.apply(getSubjectivity)
df_texts['polarity'] = df_texts.text.apply(getPolarity)
```



	dog_breed	timestamp	text	subjectivity	polarity
0	NaN	2017-08-01 16:23:56	This is Phineas. He's a mystical boy. Only eve...	1.000000	0.000000
1	Chihuahua	2017-08-01 00:17:27	This is Tilly. She's just checking pup on you....	0.433333	0.366667
2	Chihuahua	2017-07-31 00:18:03	This is Archie. He is a rare Norwegian Pouncin...	0.450000	0.150000
3	NaN	2017-07-30 15:58:51	This is Darla. She commenced a snooze mid meal...	0.150000	0.500000
4	Basset	2017-07-29 16:00:24	This is Franklin. He would like you to stop ca...	0.600000	0.233333

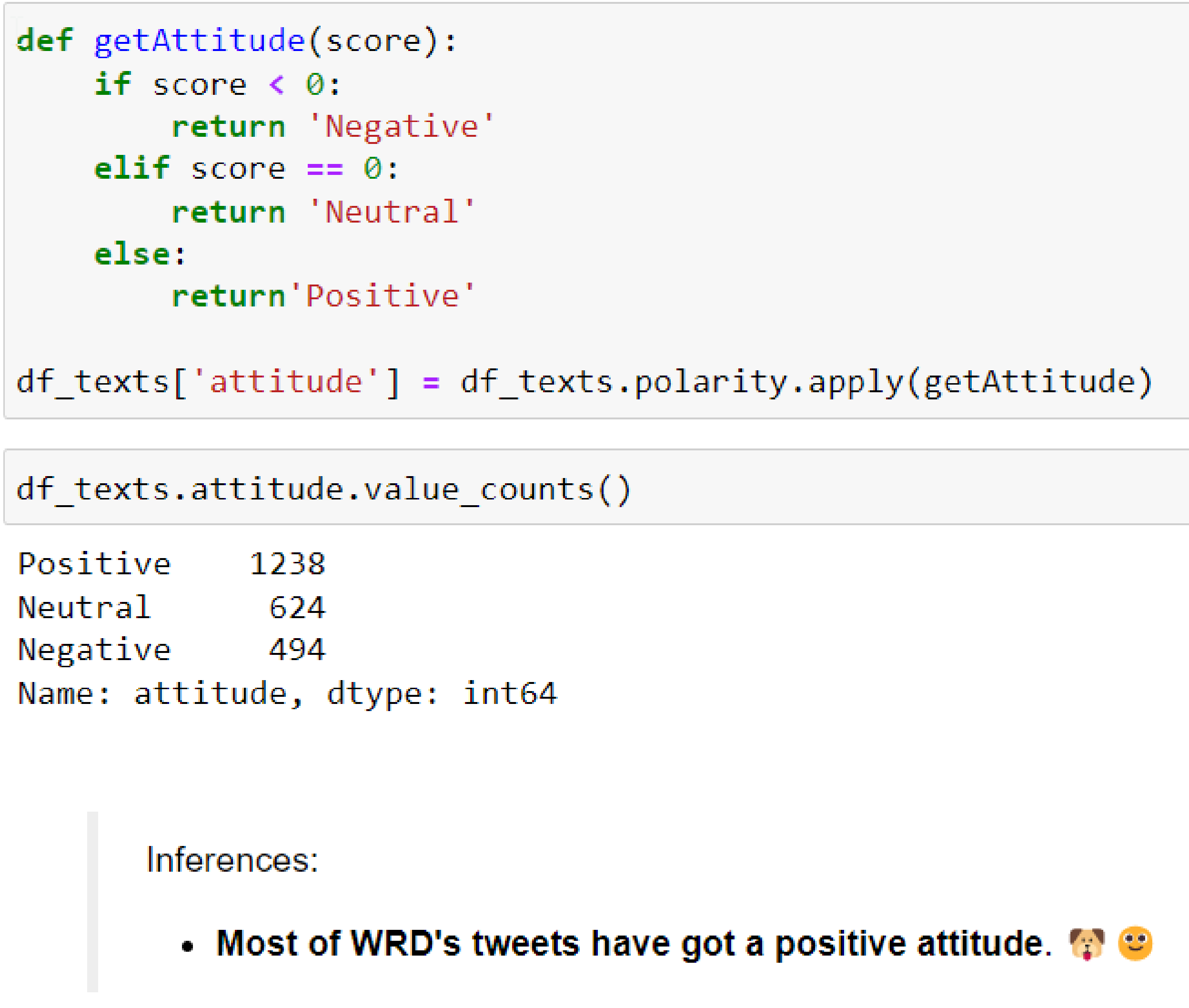


Inferences:

- WRD is generally a positive Twitter account since it's mean polarity is above 0.
- As expected, WRD tweets are authored to be in between factuality and opinion when describing and dashing out ratings to various dogs on their account. This balanced score speaks to why they are rather popular.

I will create a custom function to get a better read of each tweet's polarity in a new column named `Attitude`.

The logic will be dependent on the polarity score such that values below 0 will be awarded `Negative` attitude, values that are 0 will be awarded `Neutral` attitude while all values above 0 will be awarded `Positive` attitude.



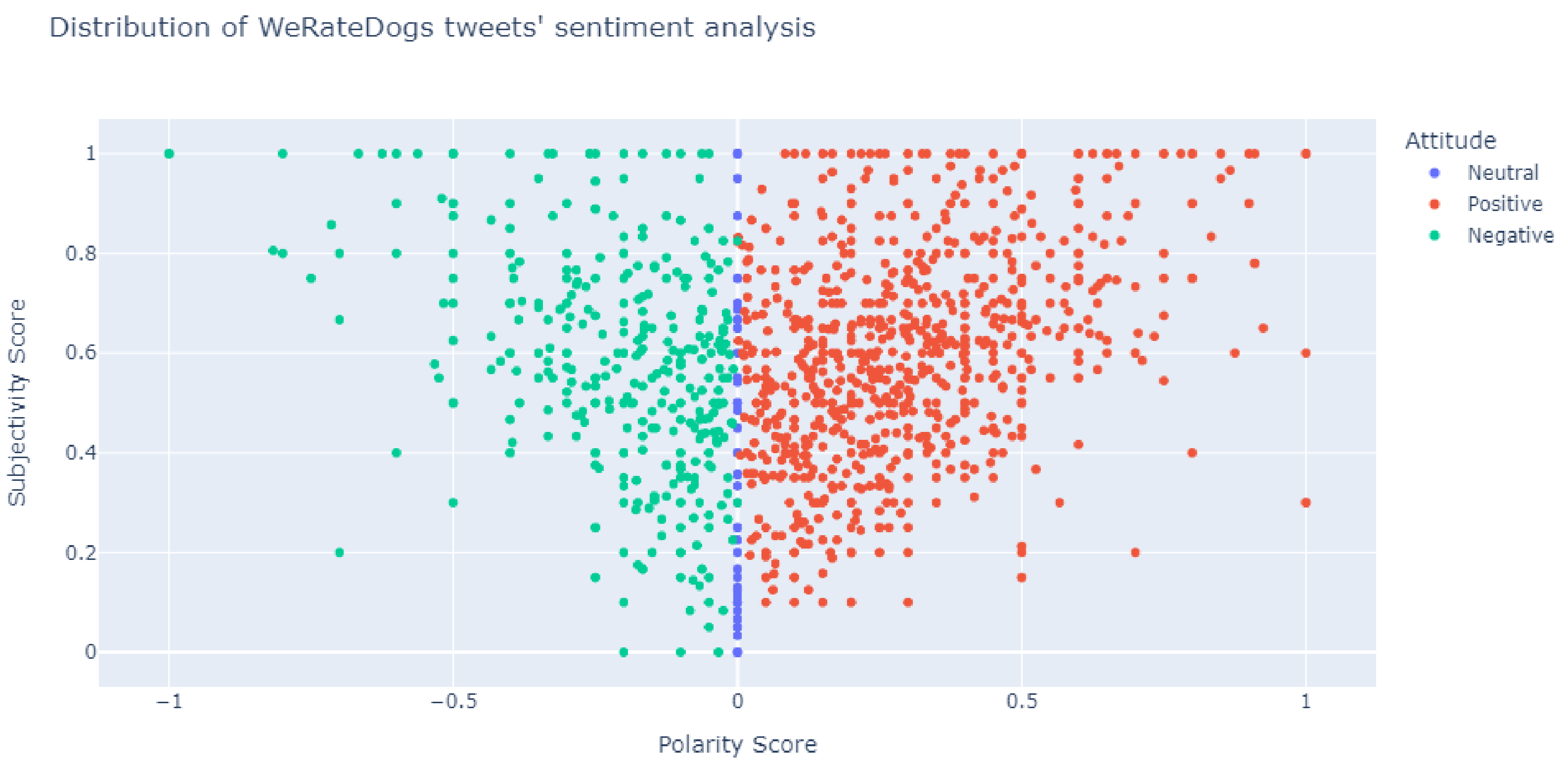
```
def getAttitude(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return'Positive'

df_texts['attitude'] = df_texts.polarity.apply(getAttitude)
```

```
df_texts.attitude.value_counts()
```

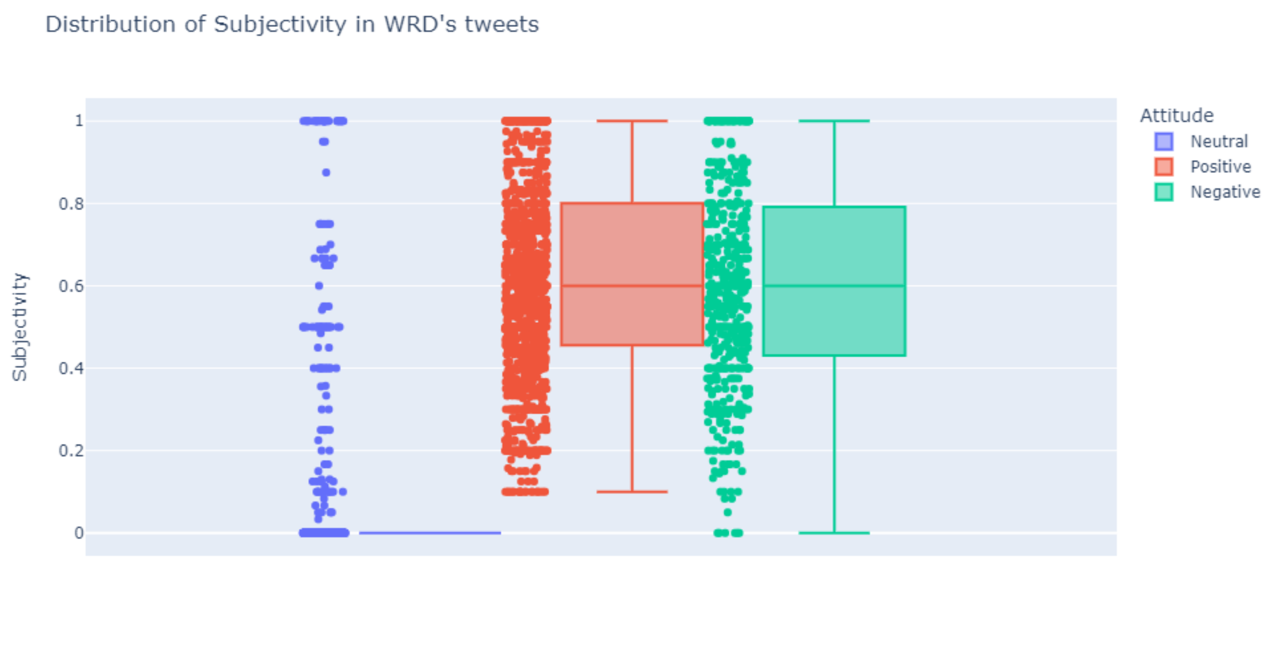
```
Positive    1238
Neutral      624
Negative     494
Name: attitude, dtype: int64
```

> Inferences:
>
> - **Most of WRD's tweets have got a positive attitude.** 🐶 🙂

Here’s a scatter plot of the same data sentiment distribution.



There’s a valid argument regarding the overall sentiment on WRD’s being skewed toward the positive spectrum with respect to attitude. From this scatter plot, there seem to be more tweets with an opinionated subjectivity compared to factual subjectivity. The plot is skewed going upward.



There seems to be a bilinear correlation between subjectivity and polarity. The more opinionated a tweet is, the more chances its polarity falls either on the higher end of the positive spectrum or the negative spectrum. Rarely does it fall on the neutral scale.

Limitations

- The major drawback in these visualizations would be the currency of the data and its relevance to the modern WRD audience and their preferences. One would argue that maybe even the WRD team might have shifted their preferences.

  I intend to carry out an analysis of WRD's most recent ratings from the start of the new decade.

Resources

- WeRateDogs
- Sentiment Analysis with Python
- Are dogs rated fairly on @dog_rates?

Additional Info

Here's a link to the Github repo should you want to replicate this work.

You can reach me via Gmail, Outlook, LinkedIn or Twitter.