UDACITY DATA ANALYSIS COURSE

PROJECT 2

DATA WRANGLING

-

INSIGHT AND VISUALISATIONS REPORT

By

TIJANI MOHAMMED

JUNE 2022

INTRODUCTION

As part of the course requirement, I am supposed to undertake several projects as part of my learning process. In this second project, my task was to wrangle data from 3 different sources all centered on data from @weratedogs twitter page.

MOTIVATION

Objective is to wrangle @WeRateDogs Twitter data to create interesting and "*Wow!*"- worthy analyses and visualizations.

DATA SOURCES

Three data sources were used":

a. **Enhanced Twitter Archive.** This was provided to be manually downloaded.

b. **Tweets Data.** Additional Data via the Twitter API.

c. **Image Predictions File.** Data consisting of probability of dog predictions using 3 different neural networks.

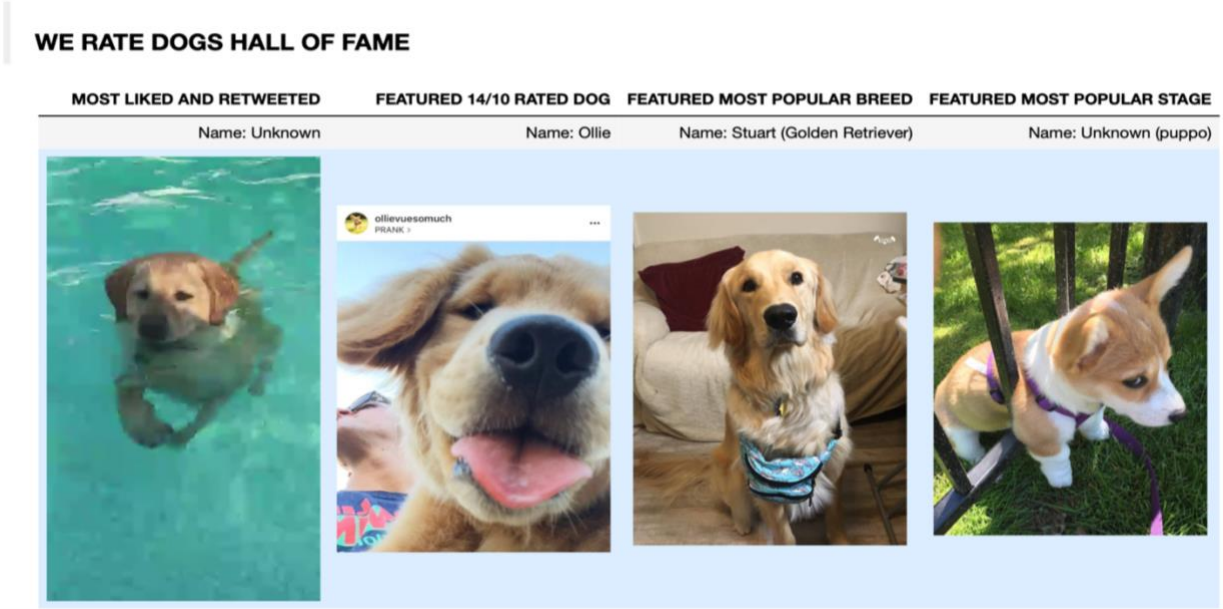
CLEANED DATA

Below is info and sample overview of my cleaned data:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1959 entries, 0 to 1958
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   tweet_id        1959 non-null   int64
 1   date_time       1959 non-null   datetime64[ns]
 2   source          1959 non-null   category
 3   name            1861 non-null   object
 4   breed           1657 non-null   object
 5   stage           295 non-null    category
 6   rating_num      1959 non-null   float64
 7   retweet_count   1959 non-null   int64
 8   favorite_count  1959 non-null   int64
 9   img_url         1959 non-null   object
 10  dog_count       1959 non-null   int64
 11  text            1959 non-null   object
dtypes: category(2), datetime64[ns](1), float64(1), int64(4), object(4)
memory usage: 157.3+ KB
```

	tweet_id	date_time	source	name	breed	stage	rating_num	retweet_count	favorite_count	img_url	dog_count	text
1480	675501075957489664	2015-12-12 02:23:01	Twitter for iPhone	None	None	NaN	13.0	5228	15575	https://pbs.twim...	1	I shall call him...
791	739606147276148736	2016-06-05 23:53:41	Twitter for iPhone	Benji	Bienheim spaniel	pupper	9.0	1496	5003	https://pbs.twim...	1	Meet Benji. He j...
1752	669972011175813120	2015-11-26 20:12:29	Twitter for iPhone	None	None	NaN	10.0	134	396	https://pbs.twim...	1	Here we see real...

INSIGHTS



A summary of the insights are highlighted below:

a. The data used was between 15 November 2015 and 01 August 2107 and cleaned to a total of 1959 records.

b. Most dogs names were missing, however Cooper, Charlie, Oliver were the most popular dog names all with counts of 10.

c. Most popular dog breed was Golden retriever & Labrador retriever followed by Pembroke & Chihuahua. It indicates Retrievers are the most popular dogs.

d. The highest rating was 14/10 and went for 34 dogs. However the most rated number is 12/10.

e. Highest retweet was 70780 and likes was 144938 by the same dog with unkown name, a puppo and Labrador retriever and and rating of 13/10.

f. There were about 10 multiple dogs pictures whose rating where a multiple of the number of dogs present.

g. Most tweets were made by mobile specifically iphones - Twitter for iphone.

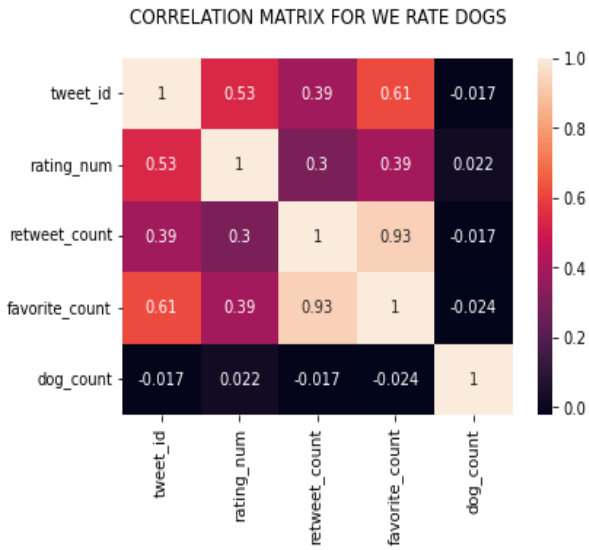
h. Puppers are the most adorable dogs with highest counts, retweets and likes.

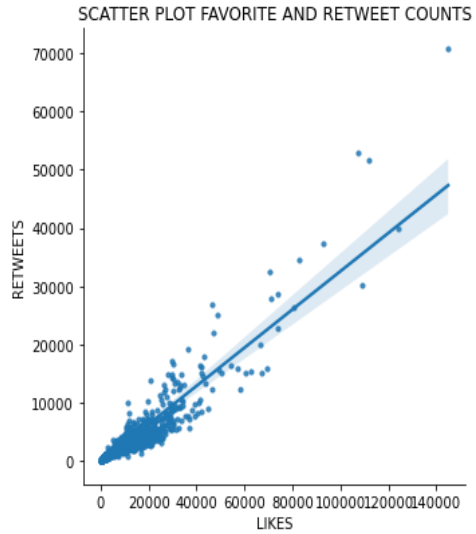
i. Golden retriever at the puppo stage is the most highly rate breed of dogs.

j. Labrador Retriever at the puppo stage are the most liked dogs.

CHARTS AND GRAPHS

Correlation Matrix

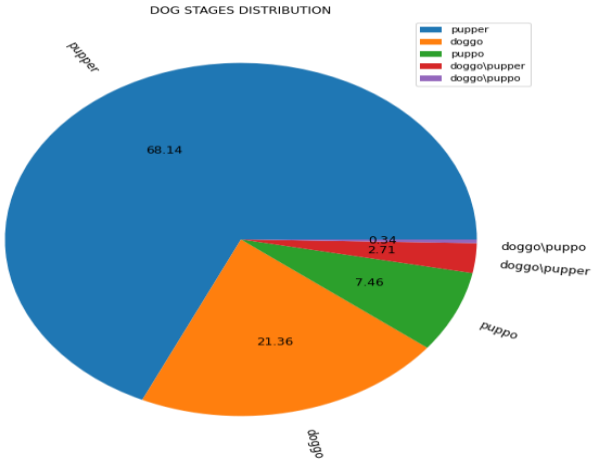




Observation. The highest correlation was 0.93 and was between retweet count and favorite count showing a positive correlation seen in the scatter plot.

Distribution of Stages of Dogs

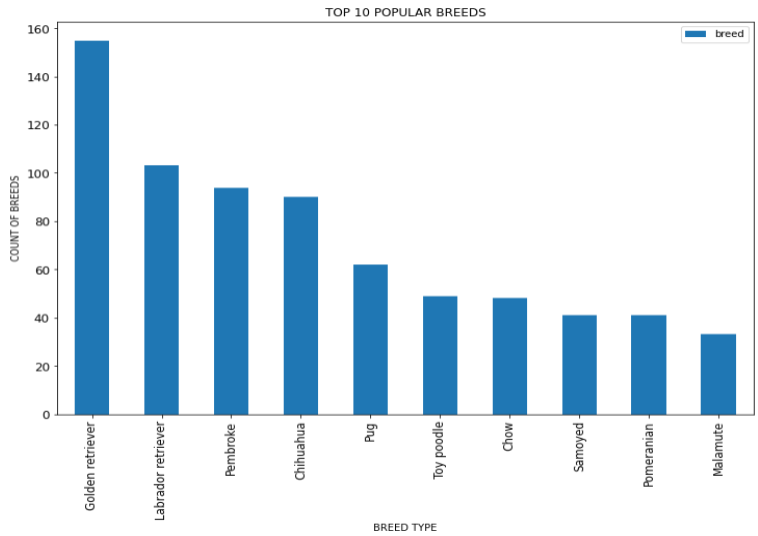
	index	stage
0	pupper	201
1	doggo	63
2	puppo	22
3	doggo\pupper	8
4	doggo\puppo	1



Observation. Puppers are the most dogs amongst the ones identified by the prediction algorithm.

Distribution of breeds of Dogs

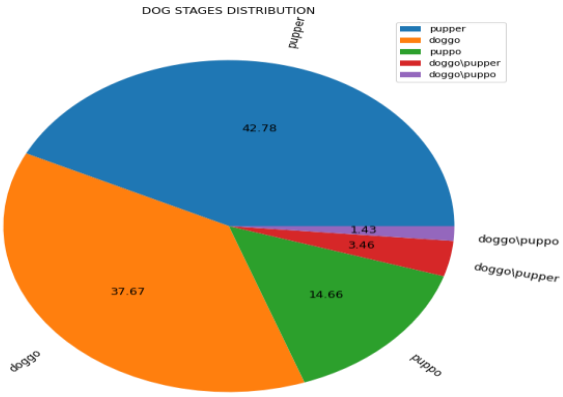
	breed
Golden retriever	155
Labrador retriever	103
Pembroke	94
Chihuahua	90
Pug	62
Toy poodle	49
Chow	48
Samoyed	41
Pomeranian	41
Malamute	33



Observation. Retrievers are the most dogs amongst the ones identified by the prediction algorithm.

Stages and Favorite Counts

	stage	tweet_id	rating_num	retweet_count	favorite_count	dog_count
0	pupper	1.444591e+20	2138.27	387129	1256979	201
1	doggo	5.027383e+19	747.00	373335	1106638	63
2	puppo	1.779792e+19	264.00	117231	430763	22
3	doggo\pupper	6.198193e+18	88.00	29821	101762	8
4	doggo\puppo	8.558515e+17	13.00	16143	41927	1

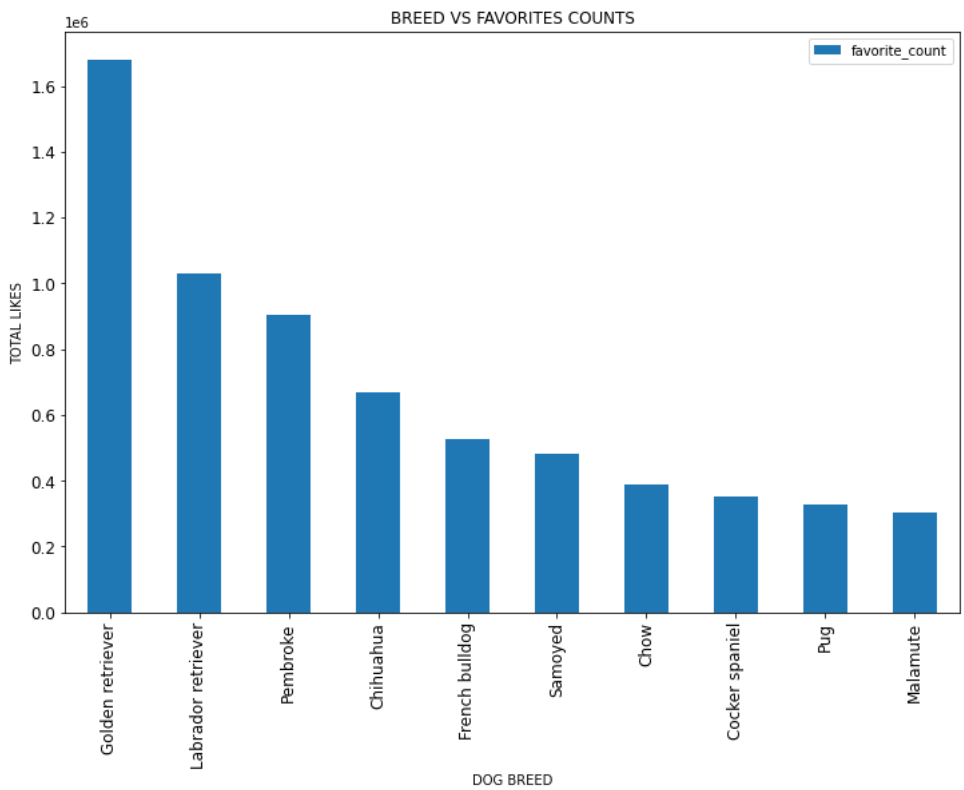


Observation. Puppers dogs amongst the ones identified by the prediction algorithm. From the data, the average counts will follow same pattern. Retweets will follow similar trend as favorite counts since they are highly correlated.

Stages and Favorite Counts

Top 10 breeds:

	breed	tweet_id	rating_num	retweet_count	favorite_count	dog_count
0	Golden retriever	1.173593e+20	1801.5	475517	1681304	163
1	Labrador retriever	7.706275e+19	1151.0	311924	1031647	127
2	Pembroke	7.103179e+19	1074.0	236772	905630	94
3	Chihuahua	6.487619e+19	948.0	210106	667233	90
4	French bulldog	2.355418e+19	335.0	132258	526368	30
5	Samoyed	3.086788e+19	481.0	156168	482328	41
6	Chow	3.587444e+19	548.0	107290	389759	59
7	Cocker spaniel	2.260826e+19	340.0	118815	352347	30
8	Pug	4.449408e+19	635.0	94463	325571	62
9	Malamute	2.466406e+19	359.0	88573	304856	33



REFERENCES

- Udacity Notes
- Stackoverflow.com
- Geegforgeeks.com
- Pandas Official Documentation