ANALYSIS AND VISUALIZATION Udacity Project4 :

Wrangle and Analyze Data Project

By: Nikita Jain

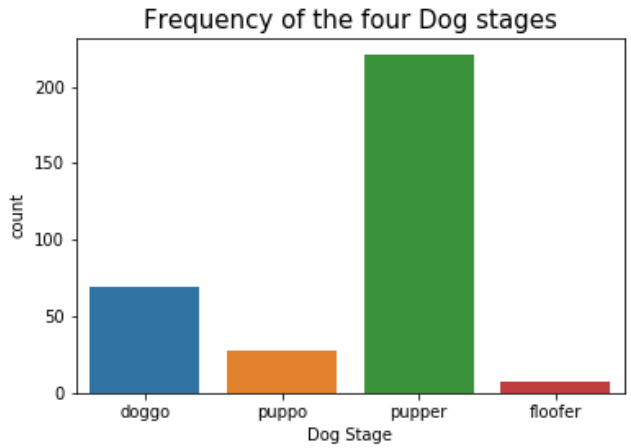
1. INTRODUCTION

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree Term 2. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, one from the twitter archive file and one from the tweet image predictions file. After scraping together the data, quality and tidiness issues were assessed and then cleaned. Finally, the data was analysed and visualised using the python library and analysis are explained below.

The final dataset that was created that had the following columns.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1987 entries, 0 to 2058
Data columns (total 14 columns):
tweet_id              1987 non-null object
timestamp             1987 non-null datetime64[ns, UTC]
text                  1987 non-null object
expanded_urls         1987 non-null object
rating_numerator      1987 non-null float64
rating_denominator    1987 non-null int64
name                  1343 non-null object
favorite_count        1987 non-null int64
retweet_count         1987 non-null int64
jpg_url               1987 non-null object
dog_stage             325 non-null object
breed                 1679 non-null object
breed_confidence      1679 non-null float64
rating                1987 non-null float64
dtypes: datetime64[ns, UTC](1), float64(3), int64(3), object(7)
memory usage: 312.9+ KB
```

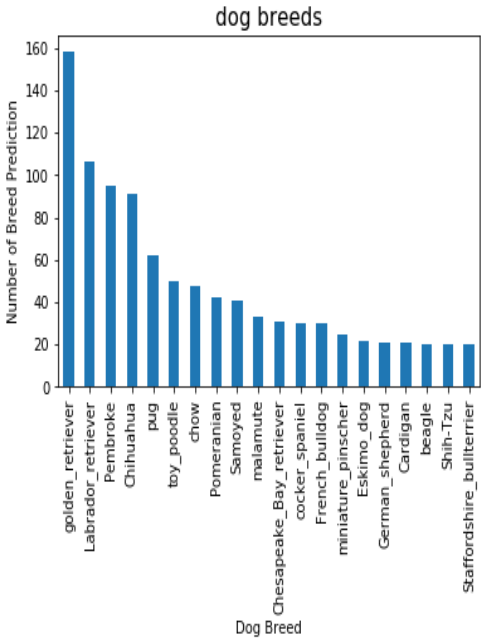
1. What is the most frequent dog stage?



As we can see from visualization drawn above pupper is most common dog stage followed by doggo then puppo and then floofer.The counts of each dog _stage are mentioned below:

```
pupper      221
doggo        69
puppo        28
floofer       7
Name: dog_stage, dtype: int64
```

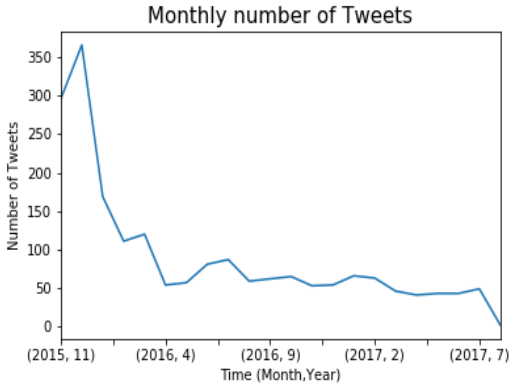
2. What are top 20 dog breeds?




```
golden_retriever              158
Labrador_retriever            106
Pembroke                       95
Chihuahua                      91
pug                            62
toy_poodle                     50
chow                           48
Pomeranian                     42
Samoyed                        41
malamute                       33
Chesapeake_Bay_retriever       31
cocker_spaniel                 30
French_bulldog                 30
miniature_pinscher             25
Eskimo_dog                     22
German_shepherd                21
Cardigan                       21
beagle                         20
Shih-Tzu                       20
Staffordshire_bullterrier      20
Name: breed, dtype: int64
Unique breeds in this dataset are   113
```

The breeds shown in the graph on left hand side are top 20 dog breeds out of total 113 breeds. The count of each breed is mentioned below:

3. What is the monthly number of tweets?

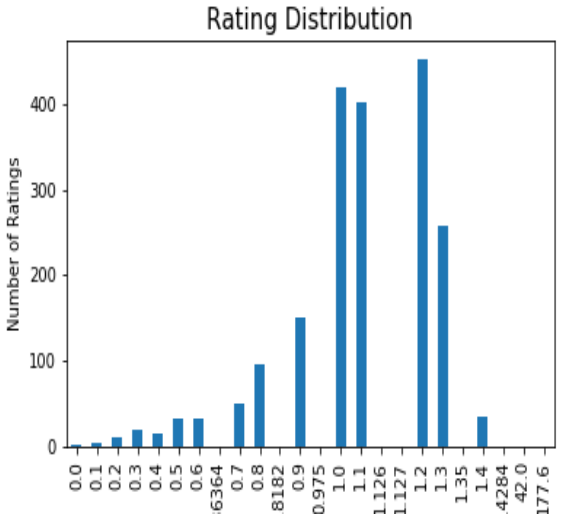


timestamp	timestamp	
2015	11	296
	12	366
2016	1	169
	2	111
	3	120
	4	54
	5	57
	6	81
	7	87
	8	59
	9	62
	10	65
	11	53
	12	54
2017	1	66
	2	63
	3	46
	4	41
	5	43
	6	43
	7	49
	8	2

Name: tweet_id, dtype: int64

Most tweets were posted in December 2015 (366 tweets). Afterwards the number of tweets decreased rapidly April 2016 and remained fairly constant since then until July 2017.

4 .Distribution of ratings

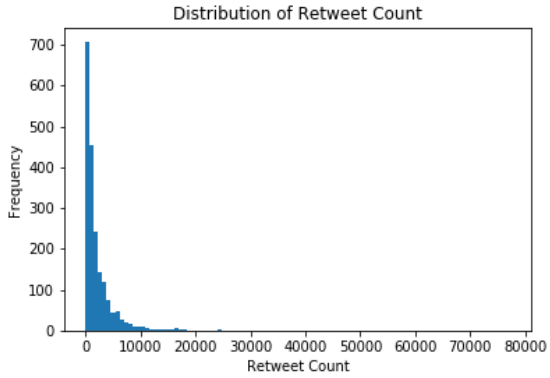


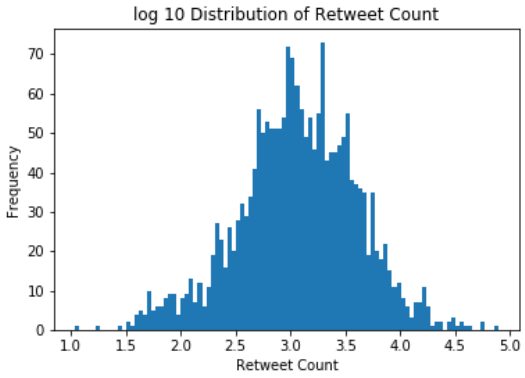
1.200000	452
1.000000	420
1.100000	402
1.300000	257
0.900000	151
0.800000	95
0.700000	51
1.400000	35
0.500000	33
0.600000	32
0.300000	19
0.400000	15
0.200000	10
0.100000	4
0.000000	2
42.000000	1
1.127000	1
1.350000	1
3.428571	1
0.636364	1
0.818182	1
1.126000	1
0.975000	1
177.600000	1

Name: rating, dtype: int64

These were the ratings obtained when rating_numerator was divided by rating_denominator and new column of rating was created.
1.2 is most assigned rating for 452 dogs followed by 1.0 for 420 dogs.

5 .Distribution of retweets





The above two graphs show the distribution of retweet counts.

First graph shows the actual retweet counts whereas the second one shows the retweets when they are normalized to log10.

We can clearly see that retweet_ counts follow a trend of normal distribution.