REPORT

ON

Wrangle 'WeRateDogs' Twitter Data Project

by Swati Chanchal

Steps Involved in this Project :

1. Gathering Data
2. Accessing Data
3. Cleaning Data
4. Storing Cleaned Data
5. Analyzing, and Visualizing Data

Analyzing, and Visualizing Data for this Project

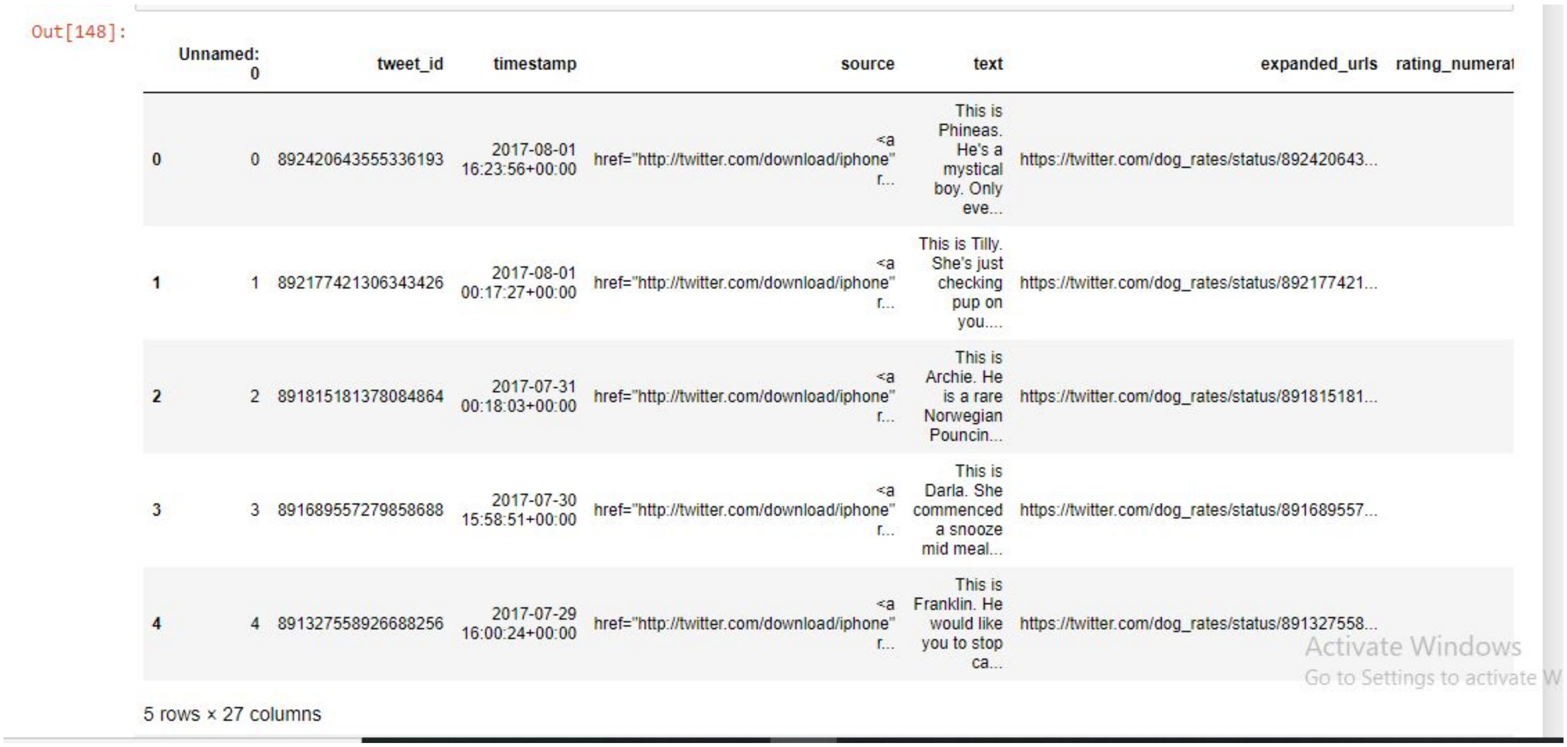
Stored the clean DataFrame in a CSV file with the main one named `twitter_archive_master.csv`.Imported the cleaned dataset .

In [147]:

```
df = pd.read_csv('twitter_archive_master.csv')
```

In [148]:

```
df.head()
```



Out[148]:

	Unnamed: 0	tweet_id	timestamp	source	text	expanded_urls	rating_numerat
0	0	892420643555336193	2017-08-01 16:23:56+00:00	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	https://twitter.com/dog_rates/status/892420643...	
1	1	892177421306343426	2017-08-01 00:17:27+00:00	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	https://twitter.com/dog_rates/status/892177421...	
2	2	891815181378084864	2017-07-31 00:18:03+00:00	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	https://twitter.com/dog_rates/status/891815181...	
3	3	891689557279858688	2017-07-30 15:58:51+00:00	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	https://twitter.com/dog_rates/status/891689557...	
4	4	891327558926688256	2017-07-29 16:00:24+00:00	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	https://twitter.com/dog_rates/status/891327558...	

5 rows × 27 columns

The shape of the new dataset is (2175, 26) . i.e Rows = 2175 and Columns = 26 .

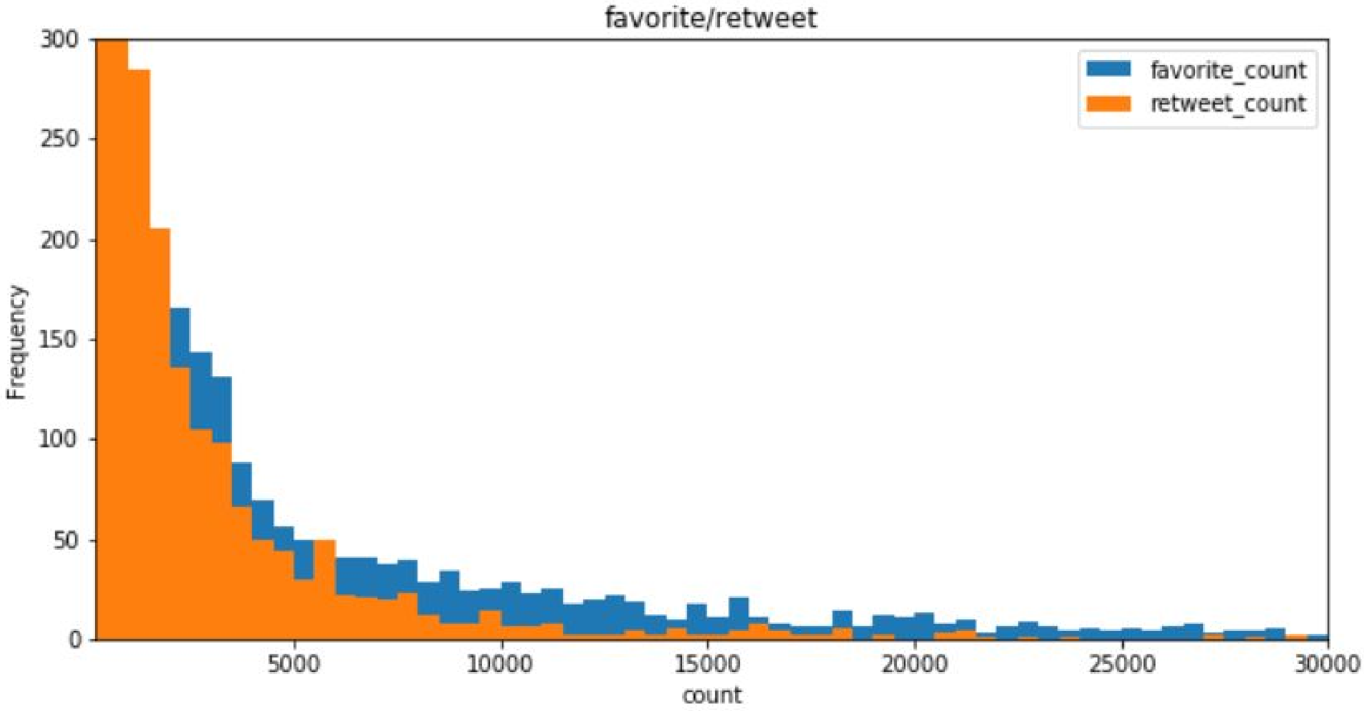
Insight 1

Plot between Count of Retweet and Count of Favourites :

In [151]:

```
plt.figure(figsize = [10,5])
bins = np.arange(df['favorite_count'].min() ,
df['favorite_count'].max() + 500, 500)
```

```
df.favorite_count.plot(kind='hist',  bins=bins )
bins = np.arange(df['retweet_count'].min() , df['retweet_count'].max()
+ 500, 500)
df.retweet_count.plot(kind='hist', bins=bins )
```



Observation :

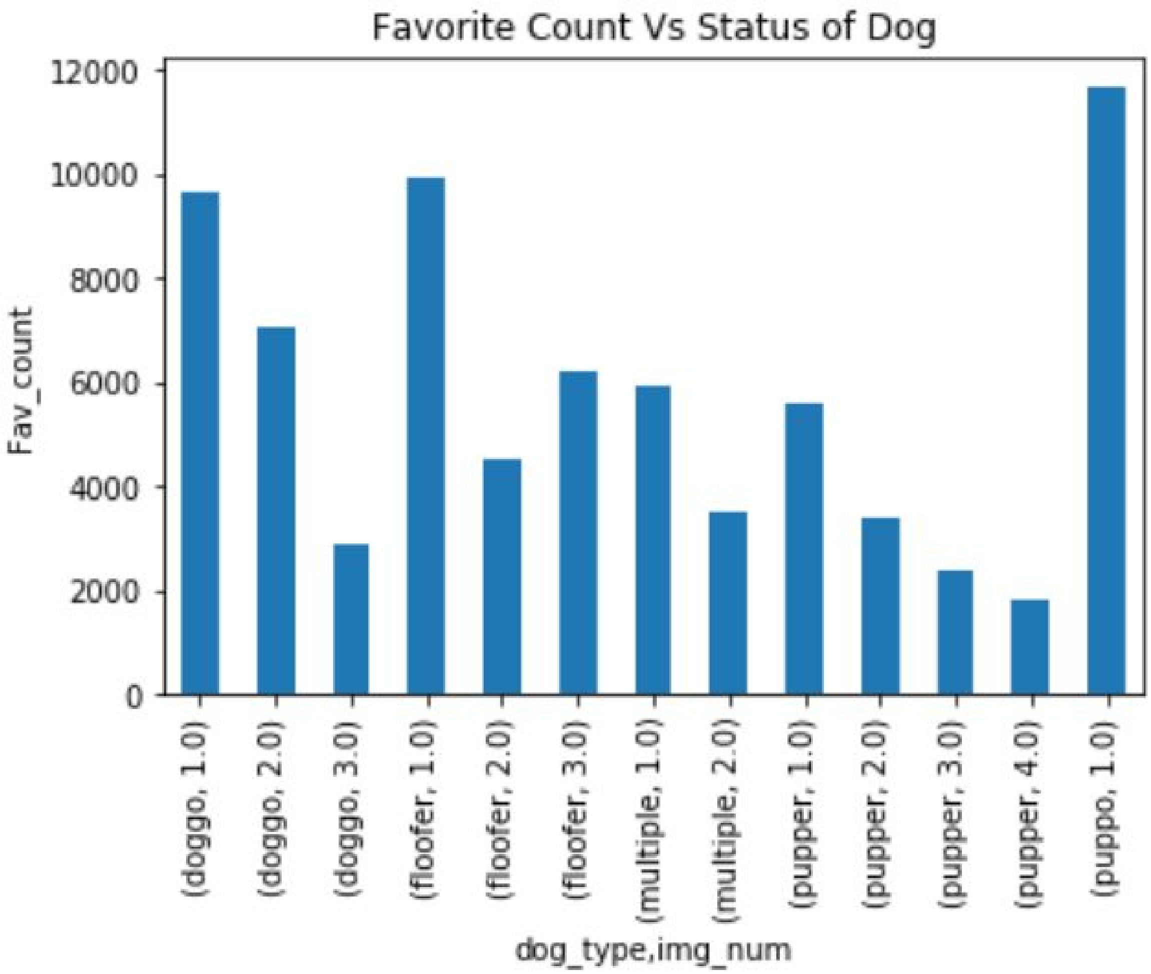
- it is clearly shown that the counts of favourites are more than the counts of retweet .

Insight 2

Plot between Dog type and Image Num vs Favourite Count

In [152]:

```
count = df.groupby(['dog_type','img_num']).favorite_count.mean(
count.plot(kind='bar')
```



Observation :

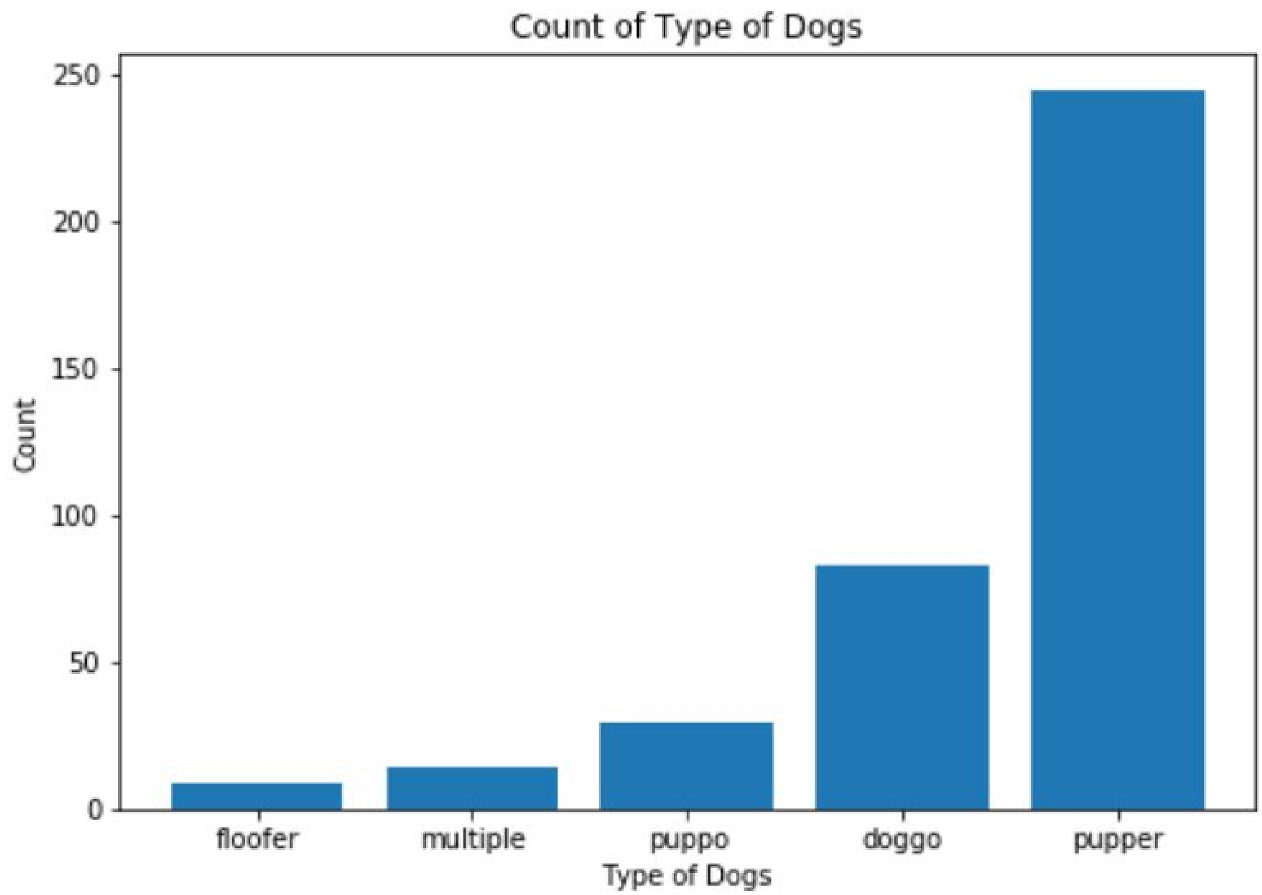
- Image 1 have highest sample size among all the images , also Pupper type dog having more favourite counts .

Insight 3

Plot on Count of Dog Types.

In [158]:

```
count=list(df['dog_type'].value_counts().sort_values())
label=list(df['dog_type'].value_counts().sort_values().index)
```



Observation :

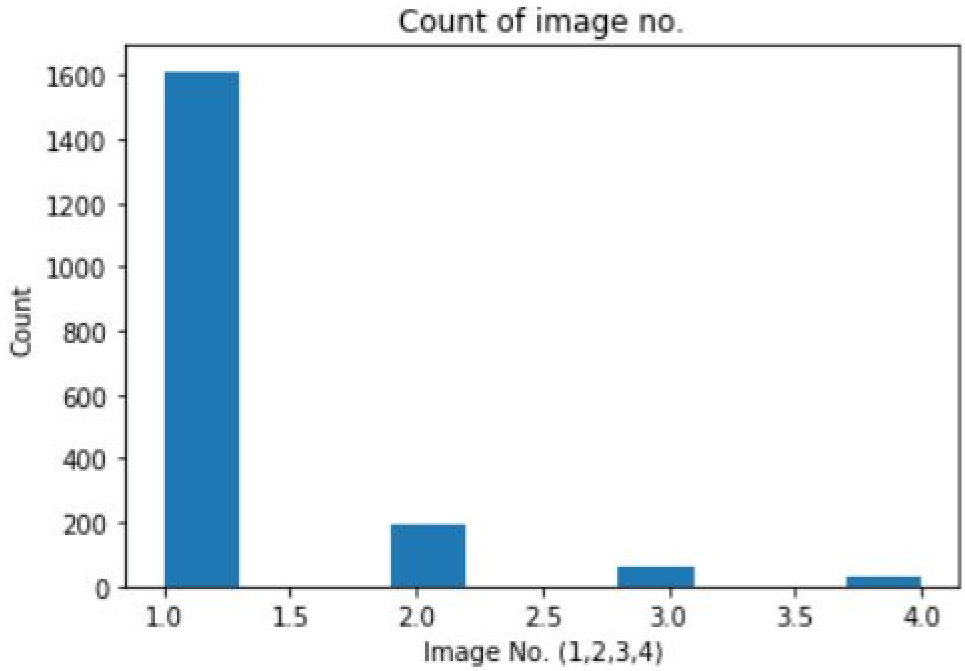
- we can clearly see that the most famous dog type is PUPPER followed by Doggo .

Insight 4

Plot of count of Image no.

In [163]:

```
plt.hist(data =df , x='img_num')
```

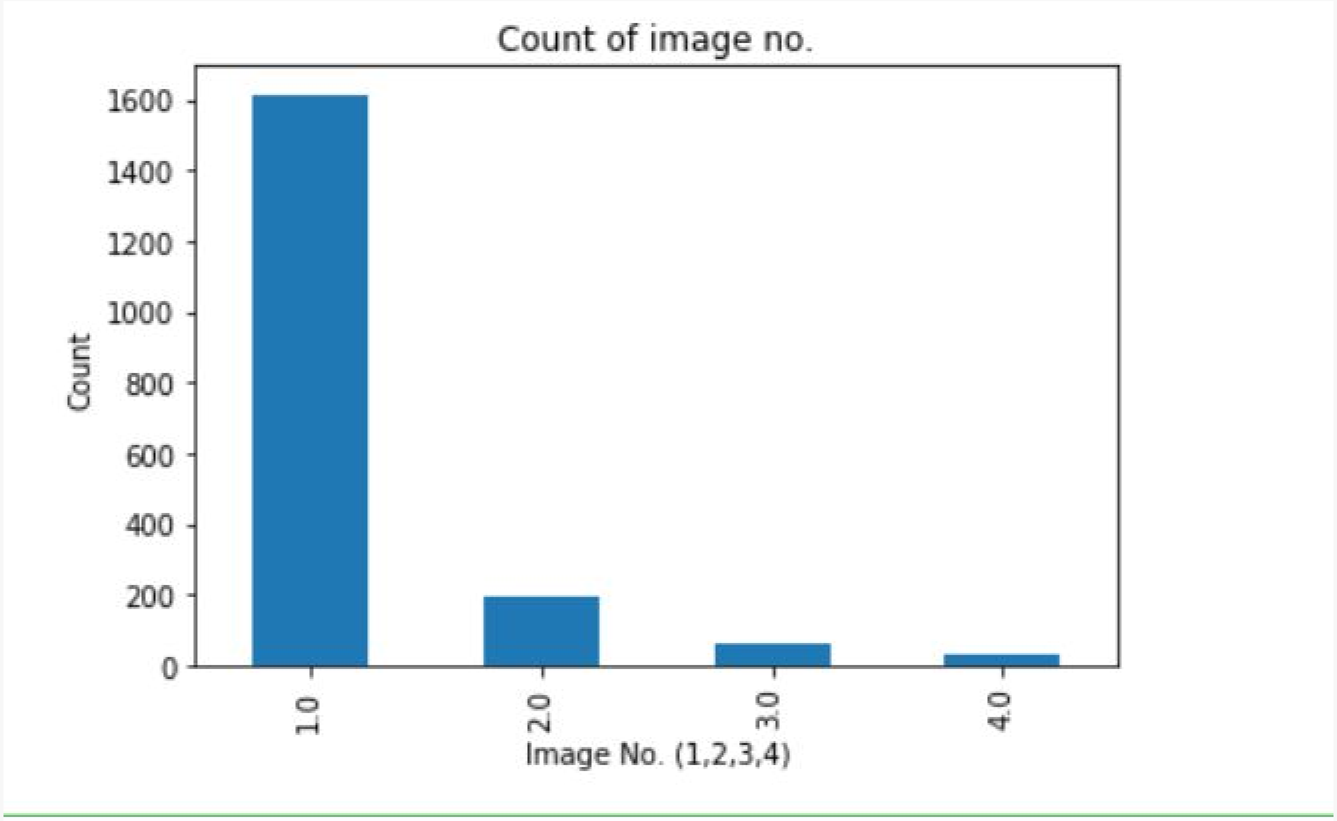


Reviewd Solution :

since img_num is not a continuous variable, it is more appropriate to use a bar graph than a histogram.

In [7]:

```
g = df['img_num'].value_counts()
g.plot(kind='bar')
```



Observation :

- clearly the Image No. 1 is the most frequent image .