After gathering, analyzing, and cleaning the three datasets, I consolidated them into one and gave it the name "twitter archive.csv". I opened pandas and read the file to get started. I updated the master csv file timestamp's datatype from object to datetime. After gathering, assessing and cleaning the data, I stored the data into csv file. The master csv file was then read into pandas DataFrame. I did find some three insights to work on then and then did some visualizations.

```
In [248]: master = Twitter_merge_df.join(image_df_clean.set_index('tweet_id'), on='tweet_id')
```

Test

```
In [249]: master.sample(8)
```

Out[249]:

tweet_id	timestamp	source	text	expanded_urls	rating_numerator	rating_denc
833722901757046785	2017-02-20 17:00:04	<a href="http://twitter.com/download/iphone" r...	This is Bronte. She's fairly h*ckin aerodynami...	https://twitter.com/dog_rates/status/833722901...	13	
670833812859932673	2015-11-29 05:16:59	<a href="http://twitter.com/download/iphone" r...	This is Jett. He is unimpressed by flower. 7/10	https://twitter.com/dog_rates/status/670833812...	7	
666454714377183233	2015-11-17 03:16:00	<a href="http://twitter.com/download/iphone" r...	I'll name the dogs from now on. This is Kreggo...	https://twitter.com/dog_rates/status/666454714...	10	
	2015-12-	<a	This is Hector. He			

```
In [250]: master.info()
```

INSIGHTS

1. What is the most popular dog stage?

As seen below we found the popper stage to be most popular stage of the four dog stages. I

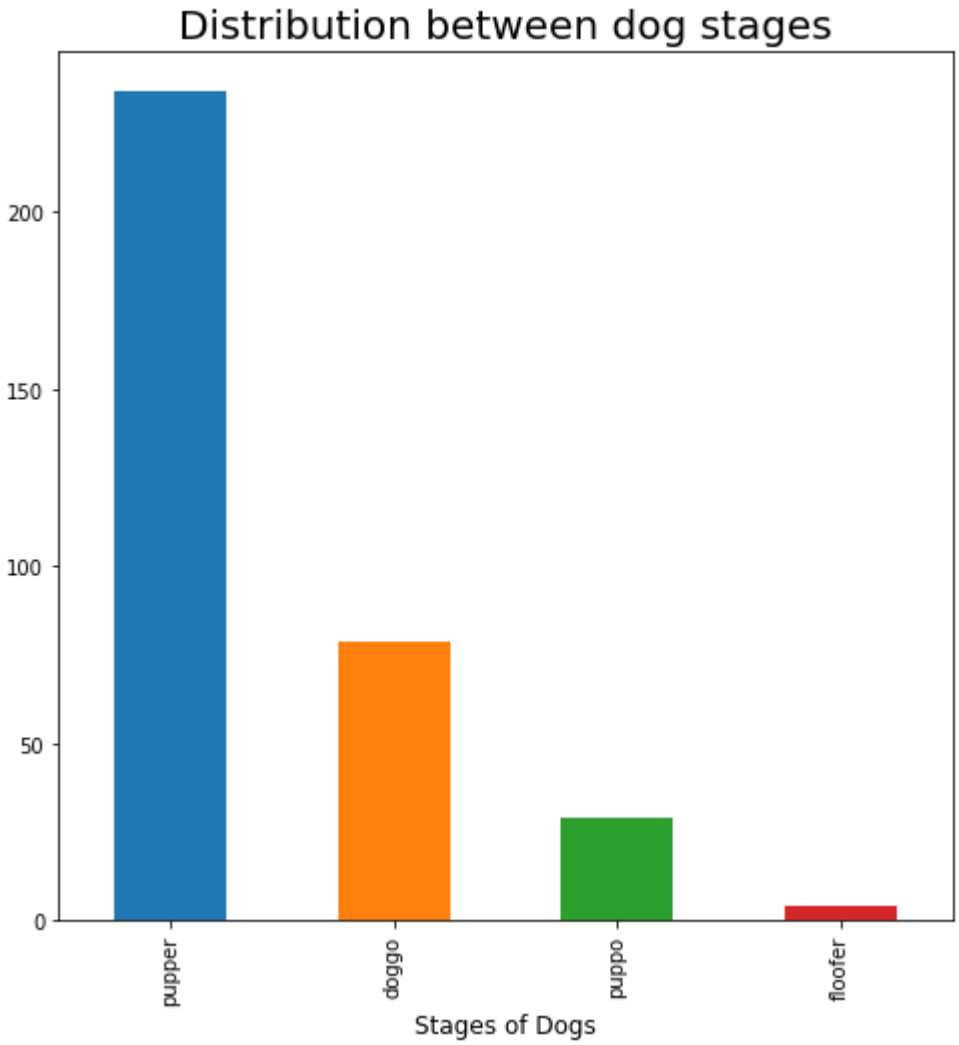
```
twit_arch_clean.dog_stage.value_counts()
```

```
pupper     234
doggo       79
puppo       29
floofer      4
Name: dog_stage, dtype: int64
```

```
plt.figure(figsize=[8,10])
master.dog_stage.value_counts().sort_values(ascending=False).plot(kind ='bar')
plt.xticks(rotation=90)
plt.xlabel('Stages of Dogs', fontsize=12)
plt.ylabel('counts', fontsize=12)
plt.title('Distribution between dog stages', fontsize=20)
```

```
Text(0.5,1,'Distribution between dog stages')
```

I used .value counts to calculate the number of dogs in each stage. The popper dog was first with the count of 234 followed by doggo which had the count of 79. The least popular stage was the floofer stage which only had the count of 4.



2. Which dog tweets got the least retweets
   I used. head () and. tail () to get the least retweets by sorting the retweet count and also the top 5 dog tweets that got the highest favorite count.

```
# check 10 least retweets
master.sort_values(by = 'retweet_count', ascending = True).tail(10)
```

	timestamp	source	text	expanded_urls	rating_n
1572	NaT	<a href="http://twitter.com/download/iphone" r...	This is Kenneth. He's stuck in a bubble. 10/10...	https://twitter.com/dog_rates/status/676219687...	
115	NaT	<a href="http://twitter.com/download/iphone" r...	This is Jamesy. He gives a kiss to every other...	https://twitter.com/dog_rates/status/866450705...	
668	NaT	<a href="http://twitter.com/download/iphone" r...	Ohboyohboyohboyohboyohboyohboyohboyohboyo...	https://twitter.com/dog_rates/status/761672994...	
418	NaT	<a href="http://twitter.com/download/iphone" r...	"Good afternoon class today we're going to lea...	https://twitter.com/dog_rates/status/806629075...	
348	NaT	<a href="http://twitter.com/download/iphone" r...	This is Bo. He was a very good First Doggo. 14...	https://twitter.com/dog_rates/status/819004803...	
60	NaT	<a href="http://twitter.com/download/iphone" r...	This is Duddles. He did an attempt. 13/10 some...	https://twitter.com/dog_rates/status/879415818...	
324	NaT	<a href="http://twitter.com/download/iphone"	Here's a super supportive puppo participating ...	https://twitter.com/dog_rates/status/822872901...	

3. Which dog tweets got the highest favorite count

```
]: master.sort_values(by = 'retweet_count', axis=0, ascending = True, inplace = True, na_position ='last')
master.head()
```

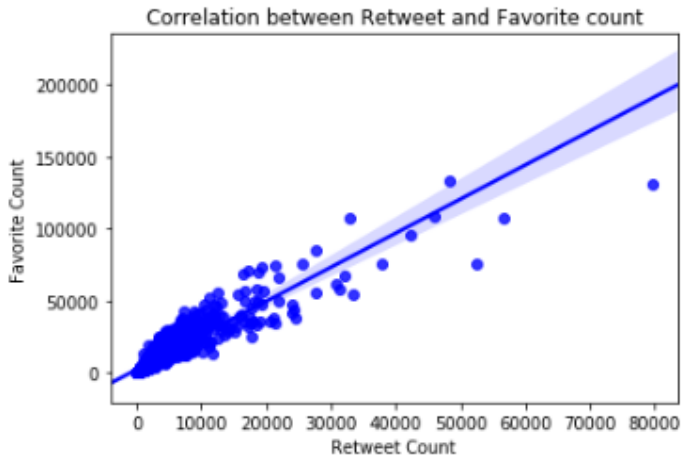
]:

	timestamp	source	text	expanded_urls	rating_numerator	rating_denominator	name	dog_
2060	NaT	href="http://twitter.com/download/iphone" r...	Oh my. Here you are seeing an Adobe Setter giv...	https://twitter.com/dog_rates/status/666102155...	11	10	None	
2041	NaT	href="http://twitter.com/download/iphone" r...	This is Scout. She is a black Downton Abbey. I...	https://twitter.com/dog_rates/status/666447344...	9	10	Scout	
1010	NaT	href="http://twitter.com/download/iphone" r...	Reminder that we made our first set of sticker...	https://twitter.com/stickergrub/status/7099191...	12	10	None	
1949	NaT	href="http://twitter.com/download/iphone" r...	I can't do better than he did. 10/10	https://twitter.com/dog_rates/status/668291999...	10	10	None	
2058	NaT	href="http://twitter.com/download/iphone" r...	Very concerned about fellow dog trapped in com...	https://twitter.com/dog_rates/status/666268910...	10	10	None	

5 rows × 21 columns

Retweet and favorite counts for a tweet are strongly correlated; in other words, if a tweet has a high favorite count, it will also have a high retweet count. The figure below, which was created by comparing these two data, clearly demonstrates a high number of connection by the tightly packed collection of dots as from 0 to 2500.

```
]: sns.regplot(x= Ret, y= Fav, marker= 'o', color= 'blue')
plt.xlabel('Retweet Count')
plt.ylabel('Favorite Count')
plt.title('Correlation between Retweet and Favorite count')
plt.legend
plt.show(),
```



The correlation between favorite count and Retweet count was found to be 9. Which is a strong positive correlation.