Report: act_report

```
My initial or prelimary work was to gather data from different sources which includes flat file containing WeRateDogs archived
tweets.
Download file programmatically from Udacity server and using twitter API to gather tweets from twitter real time.
Other resources from Udacity and from internet for supports.
```

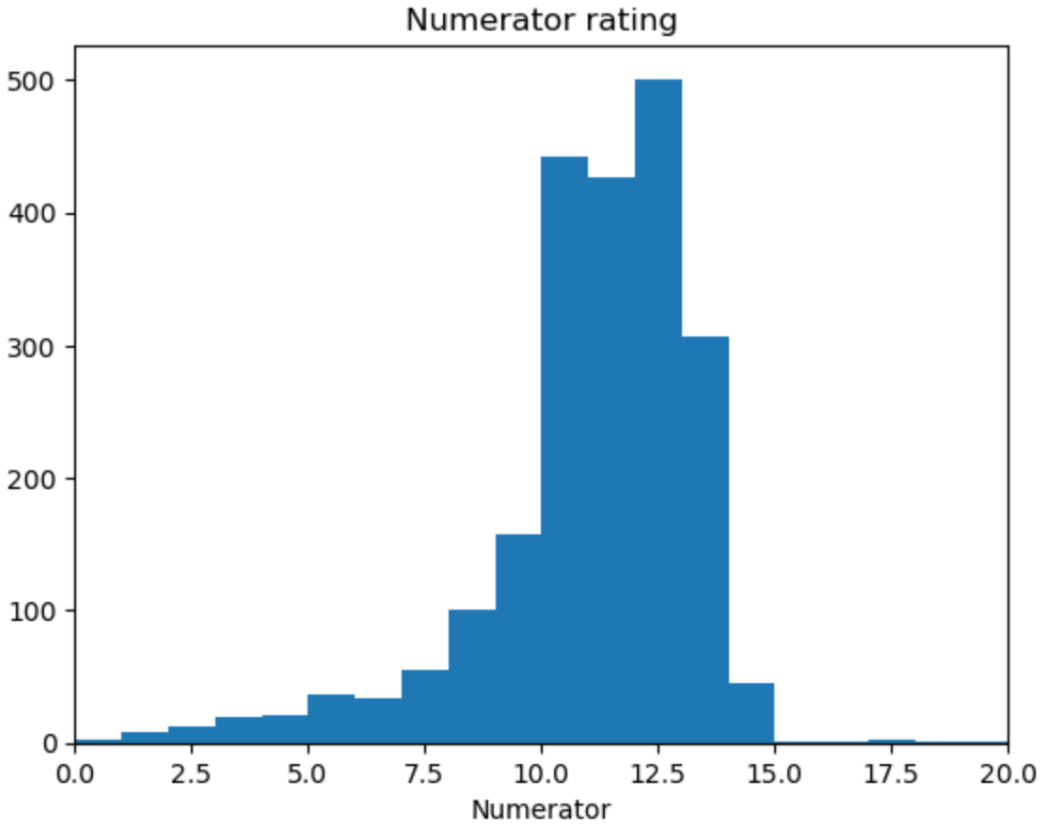
TIDINESS ASSESSMENT: For predictions_df: columns [prediction 1, prediction 2, and prediction 3] are untidy twitter_archive_df: stage columns [doggo, floofer, pupper, and puppo] are untidy IN predictions_df table jpg_url variable should be in twitter_archive_df table to satisfy tidiness definition twitter_archive_df and predictions_df tables form two different obseravations units and will be kept seperately IN twitter_archive_df table: doggo, floofer, pupper and puppo will be merged into one column tables twitter_archive_df and tweet_in_json_df will be under one observational unit There should have been a dataframe for each of these observational units [tweet data, dog data, and image predictions]

QUALITY ASSESSMENT: IN predictions_df: the prediction number needs to have data type integer One row has total confidence greater than 1 which is not normal The values in p1, p2, and p3 should be properly named Change 'None' to empty cell in doggo,floofer, pupper, puppo then delete There are duplicated jpg_url IN twitter_archive_df: some tweets in archive have missing data for retweet_count or favorite_count alot of 0 in the favorite_count IN twitter_archive_df: the dog_stage columns (doggo, floofer, pupper, and puppo) should have a categorical data type using data type float for rating so that the column can be uniform All dogs should have a name with proper letter casing dropping timestamp column not needed

```
Cleaning was performed to take care of the following:
removing retweets and duplicate values in jpg_url column.
removing duplicate values in jpg_url column
Removing pictures that are not of dogs when at least one value of p1_dog, p2_dog, and p3_dog is true with confidence greater
than 1.
Remove any invalid names and text related to lowercase names that do not have the phrase "We only rate dogs"
timestamp column was deleted

IN predictions_df: the prediction number was changed integer
One row has total confidence greater than 1 which is not normal
The values in p1, p2, and p3 was properly named
Change 'None' to empty cell in doggo,floofer, pupper, puppo then delete

IN twitter_archive_df: some tweets in archive have missing data for retweet_count or favorite_count
alot of 0 in the favorite_count so retweet_count and favorite_count was removed from the datafame.

rating number column was converted to float so that the column can be uniform
All dog names were converted to proper letter casing
dropping timestamp column because it was not needed
```

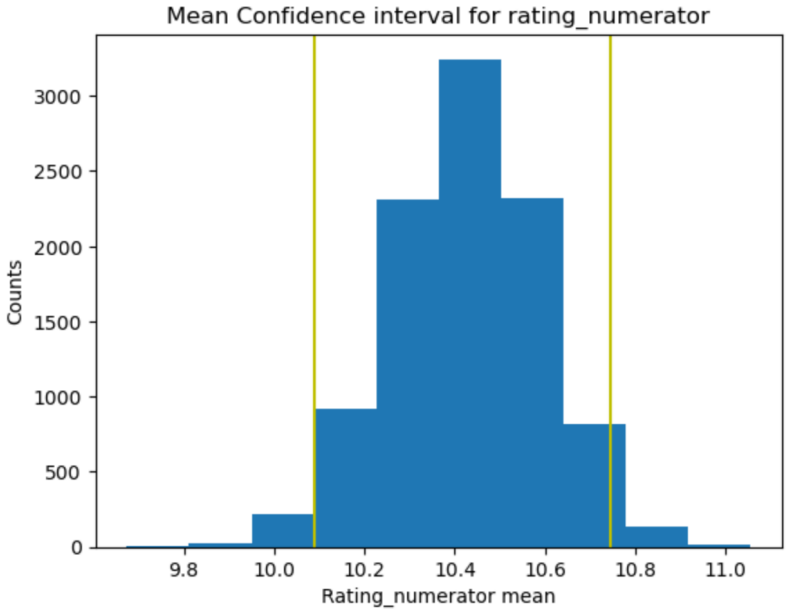
It was obvious a lot of names are in between 5 and 15 in their Numerator rating as shown below Because of the able we can see that the rating numerator histogram is skewed to the left.

In [73]:
```
bins = np.arange(0, twitter_archive_df_clean['rating_numerator'].max()+1, 1)
plt.hist(data=twitter_archive_df_clean, x='rating_numerator', bins=bins)
plt.title('Numerator rating')
plt.xlabel('Numerator')
plt.xlim(0, 20);
```



By defining confidence interval mean of 95% accuracy. I can confidently say the mean value of the rating_numerator is inbetween 10.09, 10.75.

In [77]:
```
fig, ax = plt.subplots()
left, right = np.percentile(sample_numerator_rating_means, 2.5), np.percentile(sample_numerator_rating_means, 97.5),
plt.hist(sample_numerator_rating_means);
plt.title('Mean Confidence interval for rating_numerator')
ax.set_xlabel('Rating_numerator mean');
ax.set_ylabel('Counts');
plt.axvline(left, color='y');
plt.axvline(right, color='y');
```



And finally, the clean data set was saved as twitter_archive_clean_master.csv