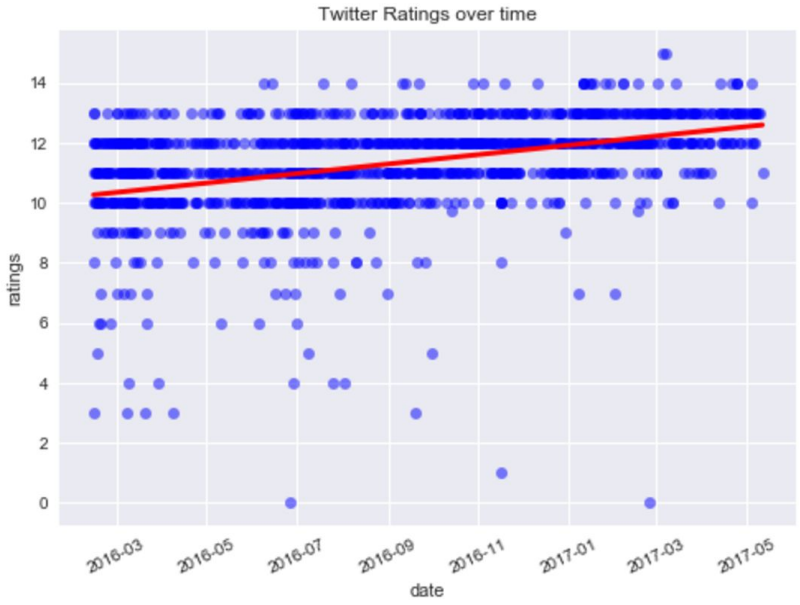
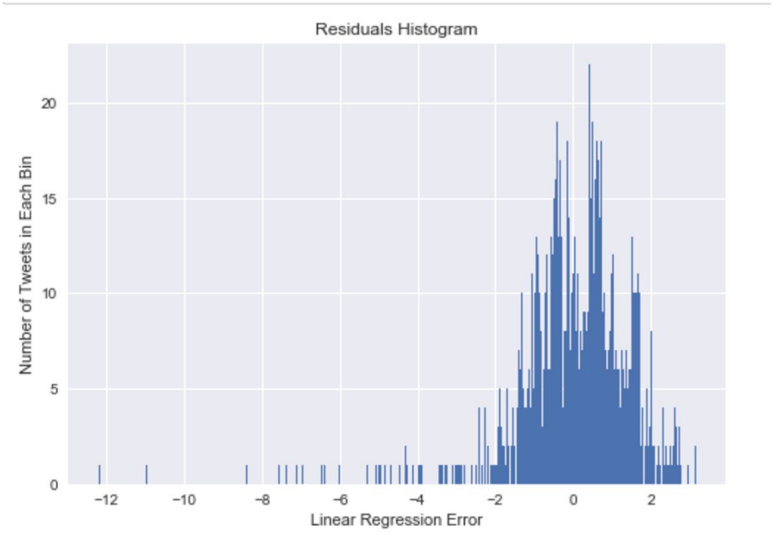
Pup Inflation

Let's rate our puppies! We have a csv of twitter data where user comments provide a n/10 rating for a puppy photo. After analyzing data, we are looking for evidence for ratings inflation over time. To do this, we want to see whether the correlation is statistically significant. Let's first look at the implementation.

Data scraped from Twitter API must first be cleaned first. We need to extract the true rating from each comment. We can use python regex to extract any n/10 out of the comments, exclude those that don't have comments. Limit ratings to at most 20/10 and any negative ones. Now we should have a decent set of ratings as well as when they are posted. After some time string conversion, you can find a way to plot this data. Use scipy linear regression to plot the line back to the data to see if the correlation exists.



From the red line in the plot above, it seems like overtime ratings do inflate. But it is hard to eyeball the positive correlation just looking at the blue dots. To check to see if this is significant, we will need to plot the difference between our prediction and the actual rating. This is called a residual and we want to plot the residuals in a histogram. Use matplotlib histogram to get something like this



We can see that most of tweet errors ranges between -2 to 2 and residual histogram appears to be normally distributed. We can conclude that positiv correlation is statistically significant.