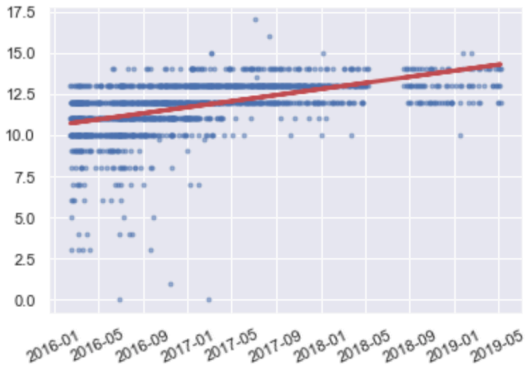
Dog Rates Summary

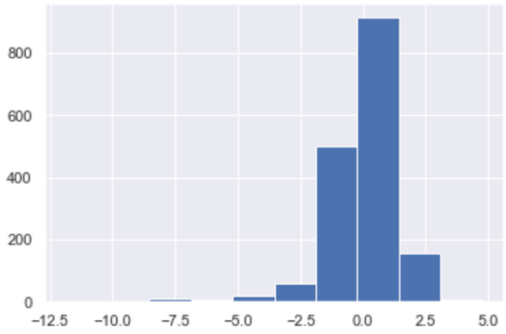
As we have a database of dog rates from @dog_rates Twitter Account, I'm going to work on determine whether there is a rating inflation occur during the period.

Raw data is saved in dog_rates_tweets.csv, which can be read into my python program. After load the dataset into code, I decide to use regular expression to draw numeric rating from the whole tweet. Then, I decided to do data filtering since there should be some outliers or unreliable ratings could twist my following analysis. I choose to set the upper bound of reliable rating score to be 25, all data above 25 are going to be dropped. For the x-axis value for regression, I choose to convert time data into a time stamp for numerical calculation.

With linear regression function, I get a slope and intercept of the regression line. After plotting it out, I realize the regressed rate increase from 10.5 to 14.5 during the period from 2016-1 to 2019-5. Moreover, with taking look at p-value of the regression model which is 1.5139606492959894e-106, there is a significant linear relation between time and ratings.



After the linear regression model has been built, I calculate the residual of each data point to see if residuals are normal distributed along the regression line. By doing normal test on the residuals, the p-value result shows as 2.07953030594431e-192, which very close to 0 and indicates it's not normally distributed. But, when I review the dataset, I notice there are more than 8000 ratings included. Based on central limit theory theorem, whenever a dataset has more than 40 individuals, we can still assume it normal distributed.



Therefore, the linear regression model is valid as its p-value < 0.05 and residual has a normal distribution. Since the slope of the regression model is positive, we can conclude there is an inflation on the rating of dog.