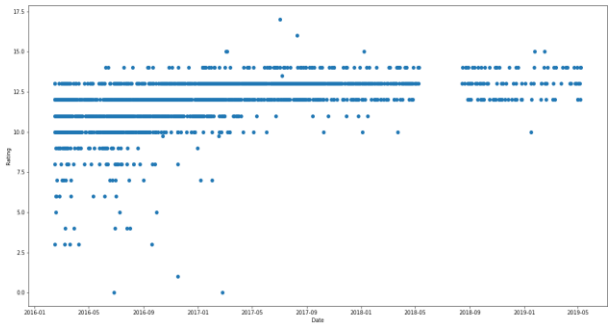
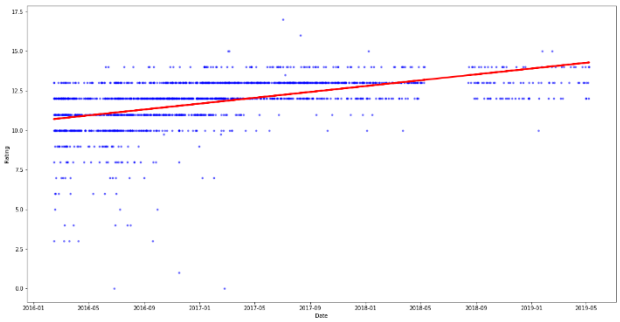
Data Science is a mixture of different tools, algorithms, and machine learning standards with the objective to find concealed information from crude data. A data scientist can discover patterns by examining data from various perspectives and predict future events using different advanced machine learning algorithms. Data analysis comes with a predefined process. The steps include defining the problem, collecting raw data, processing data for analysis, performing in-depth analysis and communicating results of the analysis.

In this blog, I'll share the result of data analysis performed on @dog_rates Twitter feed, which rates the cuteness of the user's dog. The purpose is to find out if there has been rate inflation over the years. I started with the data, tweets related to dogs, scrapped from @dog_rates Twitter feed. I cleaned it up by excluding tweets that didn't have a rating and removed outliers to exclude ratings that were too large, i.e. ratings larger than 25/10.

Now that the data was in the expected state, it was time to perform analysis. I began by plotting a scatterplot of date vs rating to visualize any evident trend before applying further statistics. The scatterplot showed an upwards trend in ratings. The ratings prior to 2016 ranged between 2.5 and 13, however, from 2017 onwards, the ratings were on the higher end of the scale.



To test the claim, I applied linear regression model to create the best fit line. The model provided some useful information such as Y-intercept, gradient, and p-value. In statistics, a p-value helps determine the significance of the results. A small p-value (< 0.05) indicates strong evidence against the null hypothesis which means the dog ratings indeed inflated over the years. The p-value for the given data set was 1.5139606492959894e-106.



To further solidify the claim, I calculated the R-squared value to find out what percentage of dog ratings varies with time. The 'r' in R-squared is called residuals. Residuals is the difference between the observed value of the dependent variable and the predicted value given by the best fit line. The histogram shows the distribution of residual values. The R-squared value was 0.25, meaning 25% of dog ratings variation is explained by the model. A value strong enough to conclude that the model is significant and dog ratings has indeed inflated over the years.