Data is the new “gold” of this digital era and student from computer science and statistics background are using the buzz words “data science”, “big data”, “machine learning” and etc. to sound cool during any conversation. In this blog post, I would talk about a small data science task and show you how information that most people would think is unnecessary will give us some cool results. I am not just writing it to sound trendy, it is worth 3.5% of my a grade.

Analyzing tweets is a fun exercise and can tell a lot about what people think about different things happening around world. Here, we will analyze that how people rating(cuteness) dogs have change over the years. We start with a data dump which has all the tweets related to dogs. Then we filter out the data to keep what we only need i.e the tweets that has some ratings for dogs for e.g 10/10. After that, we removed some “outliers” from those ratings for e.g 1776/10 is not a realistic rating and might affect our results. After filtering some unwanted data, plotting a scatterplot is the first thing one should do to visualize any data to have some initial idea before applying any statistics or data science. Figure1. Shows that initial scatterplot. In 2016, there was a mix of high and low ratings and, from 2017, the ratings are on the higher side. We can apply most common stat model i.e linear regression to see if there is any relation between the ratings and time. We get a best fit line shown in Figure2, and from the figure we see that ratings are increasing over time. But can we prove it using any other tools? Yes, statistics gives us some useful tools to see the significance of our results. So linear regression models gives us some other good information, the intercept, slope and p-value. Slope and intercept are the terms from grade level mathematics but how are they significant to this result. For that, we have p-values that ranges from 0 to 1, think of it as some kind of probability. If it is less 0.05, then our slope is meaningful otherwise it is not. P-values using this python library gives us the number $1.5139606492959894 \times 10^{-106}$. A very small number meaning that this model is significant. We will need another statistic to see if our variation is also justified meaning what percentage of dog ratings varies with time. The statistics is called “r-squared” where r is residuals meaning the difference between the actual value and the value given by the best fit line. Figure3 shows the distribution of residual values. The r-squared value from the same python library, it gives 0.25. This means that 25% of dog ratings variation is explained by our model. Concluding our model is significant but it does not tell a lot of variation in our data.

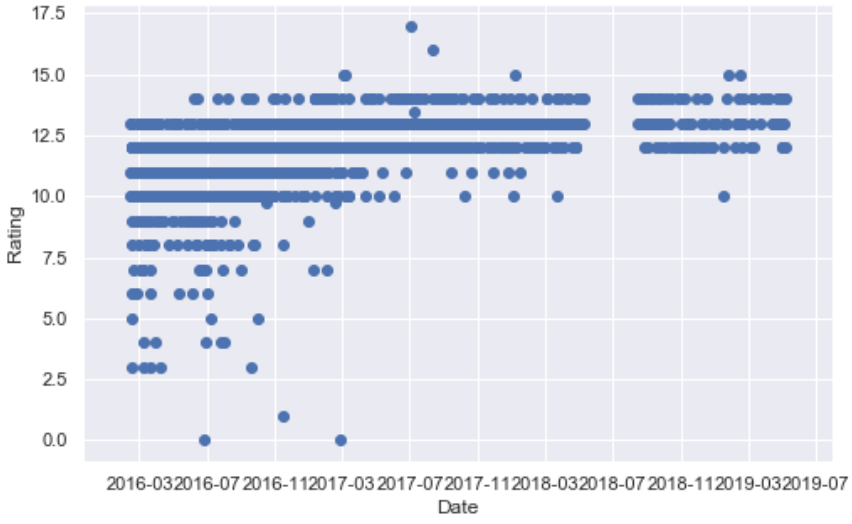

Figure1

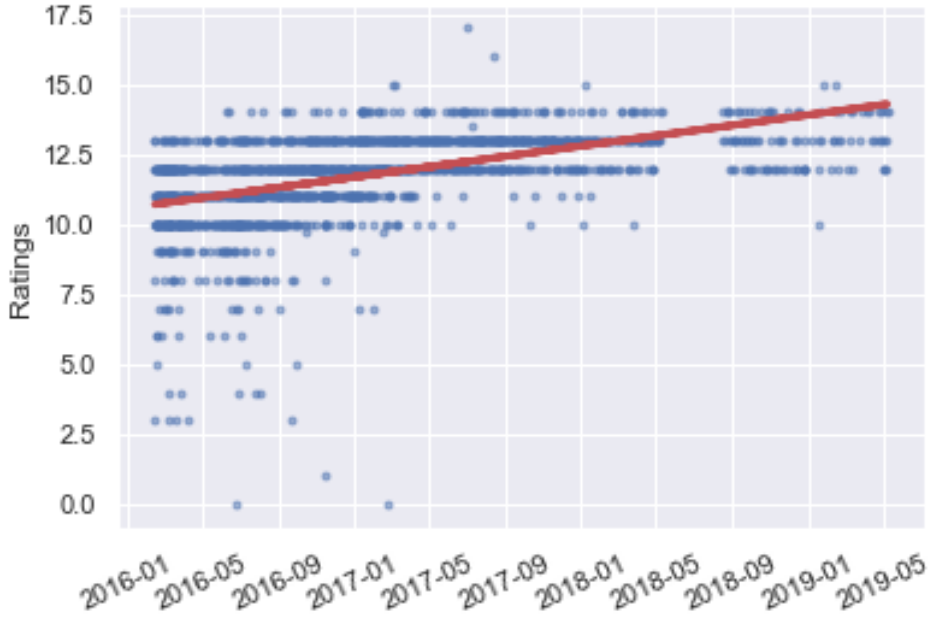

Figure2

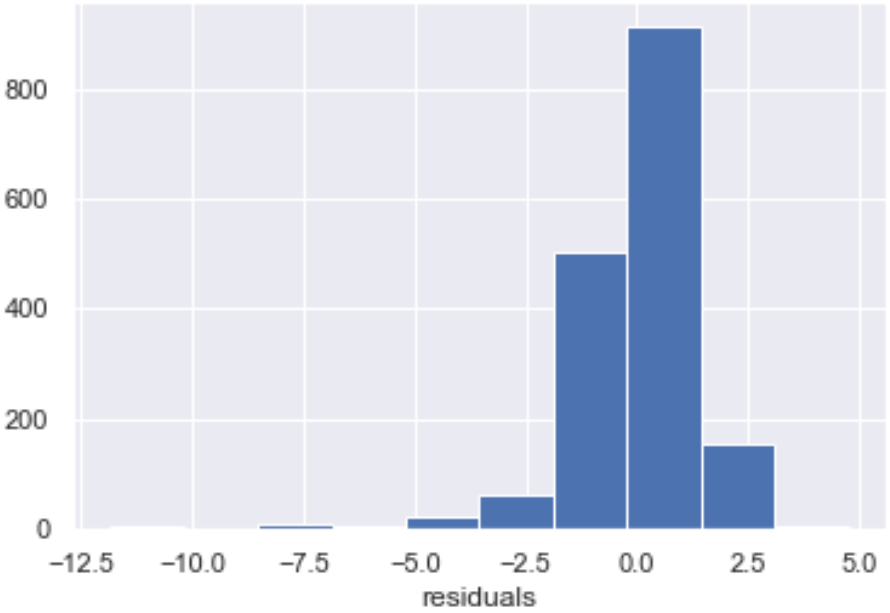

Figure3