Analyzing Pup Grade Inflation on Twitter

Introduction

The user @dog_rates on Twitter takes other user-submitted pictures and videos of dogs and rates them with a numerical score, out of 10, but can be higher than that. In this analysis, we will determine whether there has been grade inflation from @dog_rates' ratings over the years.

Methodology

With any statistical analysis involving large amounts of data, we need to filter out any invalid values, and determine a step for major outliers.

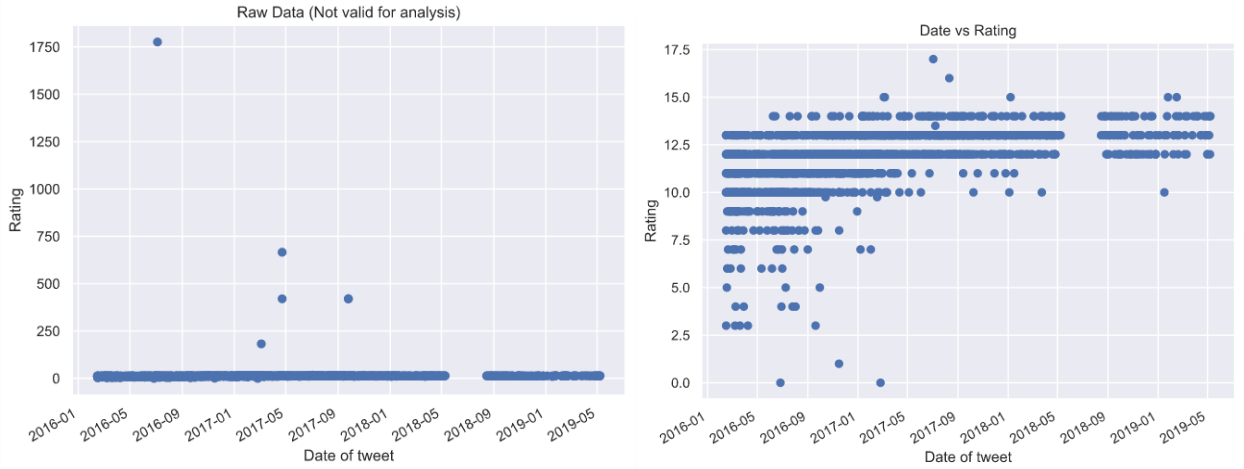

Figure 1. Comparing unfiltered (left) and filtered (right) data points.

A first glance at the unfiltered data shows a near consistent line representing the values close to the original "out of 10" rating system. There are a few outliers that need to be removed to get a better representation of the data. Upon removing these outliers with rating > 25, we find that all the data points are closer to each other in rating without adversely affecting our ability to make conclusions from the data.

Now that we have our filtered data, analytical techniques can now be applied. Linear regression generates a best fit line prediction and a p-value of 1.51e-106 based on the data.

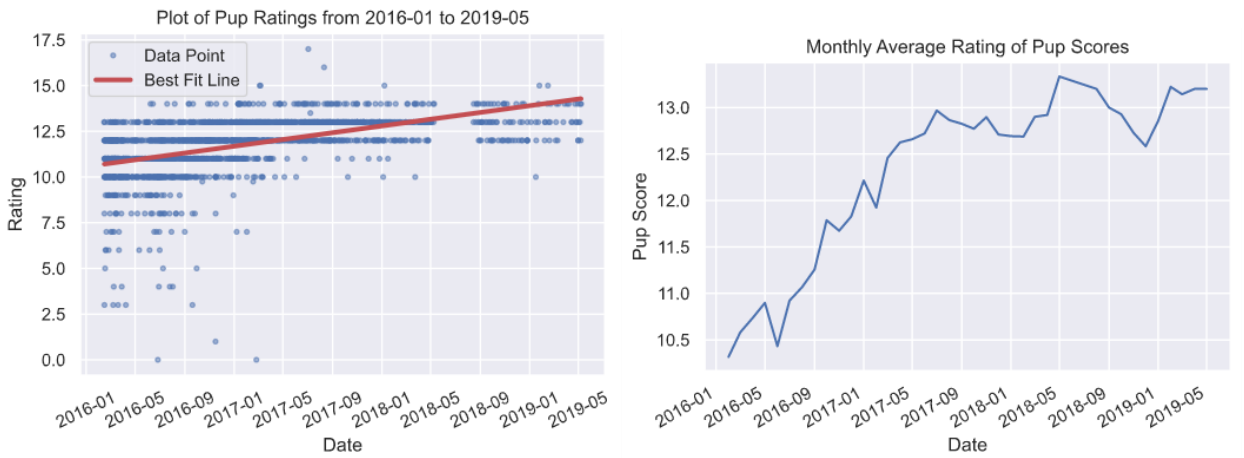

Figure 2. Visualizing the data points and the OLS best fit line, and the monthly average values.

Applying Ordinary Least Squares linear regression (OLS) generates a best fit line that serves as a predictive model for the dog ratings. The best fit line clearly trends upwards from 2016 to 2019. In order to make conclusions from OLS there were some made assumptions about the data:

1. The sample is representative of the population
2. The relationship between X and Y is linearly related
3. The residuals are normally distributed

From the data gathering and visualization, we can conclude that the sample is representative of the population, and the ratings are linearly related by rejecting the null hypothesis that there is no correlation between date and rating based on our p-value of 1.514e-106 which is less than 0.05.

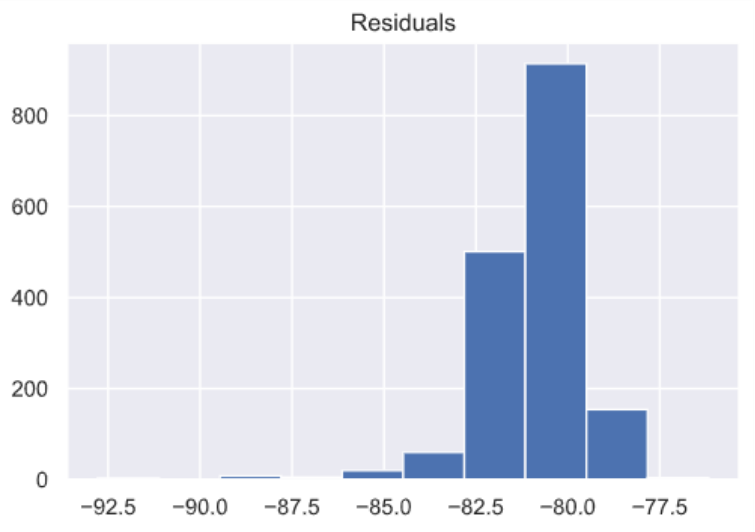

Figure 3. Plotting the residuals from the data and the OLS model.

By visually inspecting the residuals plot shows that the residuals are normally distributed enough. With the three requirements satisfied, we can use the linear regression model to make meaningful conclusions.

Conclusion

Based on the upwards trend of the OLS best fit line and supplemented by the increasing month to month average seen in Figure 2, we can conclude that there is indeed grade inflation in the dog ratings given by Twitter user @dog_rates. This however, does not investigate the reasons behind the grade inflation and a deeper dive at the data itself is needed to investigate that.