Simon Fraser University

CMPT 353 D100: Computational Data Science

Exercise 12

Pup Inflation Visualization Report

Instructor: Dr. Greg Baker

August 6, 2021

Contents

I. Background

This visualization task is followed by Exercise 7 of this term, and I took the dataset of dog rates tweets from this exercise. The dataset is collected from the @dog_rates Twitter, which rates the cuteness of users' dog pictures [1]. And the question raised from the dataset is whether there is pup inflation. Or in other words, is it true that good dogs getting better [2]?

II. Data Preprocessing

First of all, I do the data preprocessing. The steps are as follows [3]. Load the data from the CSV into a DataFrame. Find tweets that contain an “n / 10” rating. Extract the numeric rating. Exclude tweets that don't contain a rate. Make sure the 'created_at' column is a datetime value, not a string. And now, we can do the first visualization which is a scatter plot and box plot to view the distribution of the raw data. The visualization is as follows.

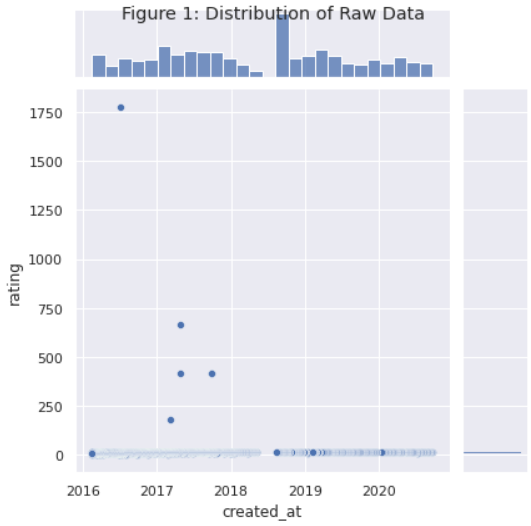


Figure 1: Distribution of raw data.

Some outliers significantly affect the distribution. By looking back at the dataset, we believed that any rating which is larger than 25 is an outlier [3]. After removing the outliers, we have the second visualization of the processed data, which is a scatter plot and box plot to view the distribution of the processed data. The visualization is as follows.

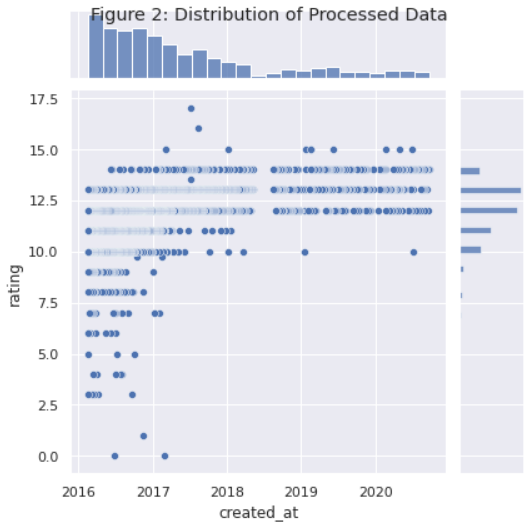


Figure 2: Distribution of processed data.

The distribution of processed data seems acceptable, even though it is right-skewed instead of normal.

III. Data Analysis

Based on the processed data, I generated the first visualization of the dataset, which is a scatter plot of date vs rating as well as a fit line across the data. The visualization is as follows.

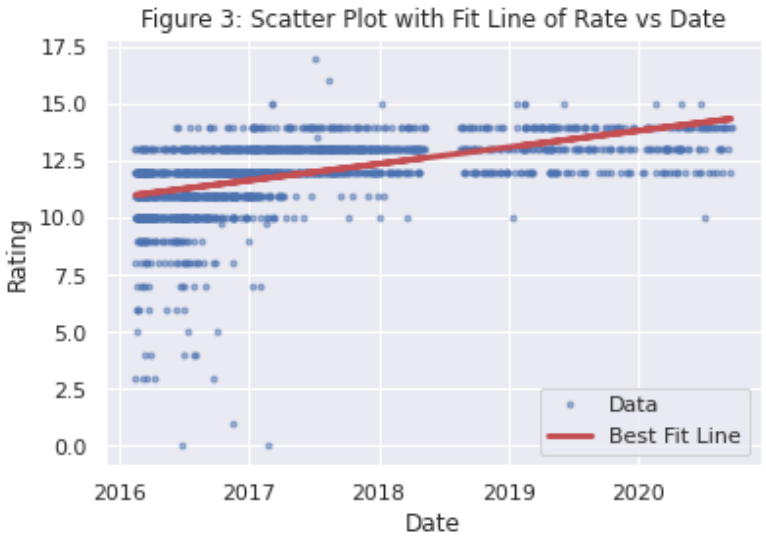


Figure 3: Scatter plot with a best-fit line of dog rates vs date.

From Figure 3, we can have a basic understanding of the data. After preprocessing, especially removing all outliers, it seems that there is a positive linear relationship between rating and date. So, we did a linear regression. But we still need some other analysis to justify our hypothesis. I calculate the p-value, which is much less than the significance level. Hence, the slope is different from 0. In other words, there is a positive linear correlation between rate and date. But this might not be true.

The linear regression model we used here is called ordinary least square, or OLS. Apart from the slope and the tendency of rating, we also analyze the residuals. Only if residuals are normally distributed, the p-value is meaningful and can be used to interpret. The fourth visualization is about the distribution of residuals using OLS. To view it, we used a histogram as follows.

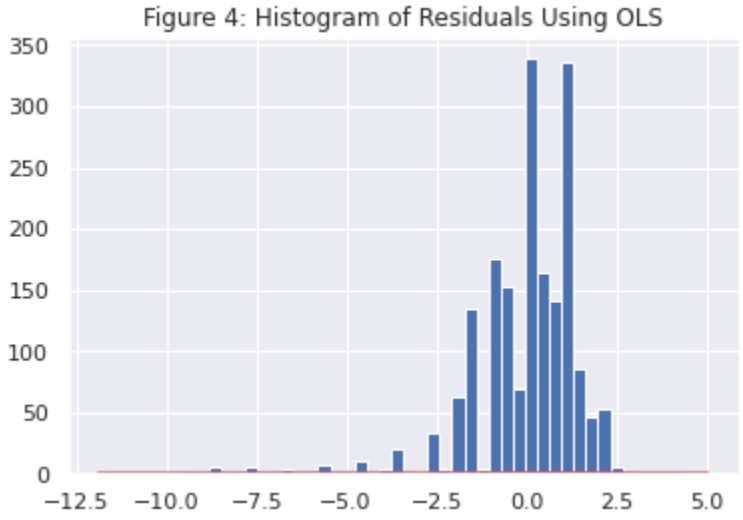
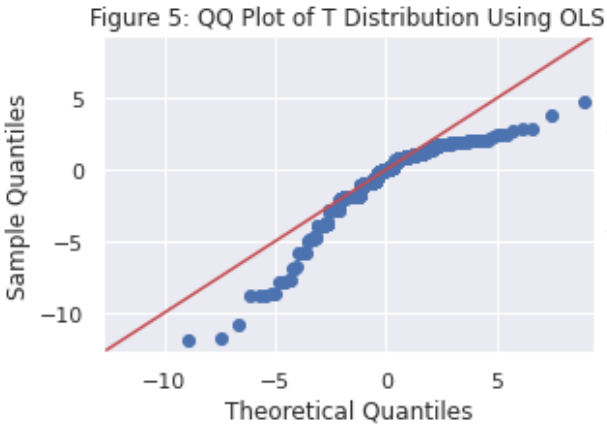


Figure 4: Histogram of residuals using OLS.

It seems that the distribution of residuals is left-skew instead of normal. But we have to make two QQ plots and a normality test to justify our guess, which is the fifth and sixth visualization. Both QQ plots are presented as follows.



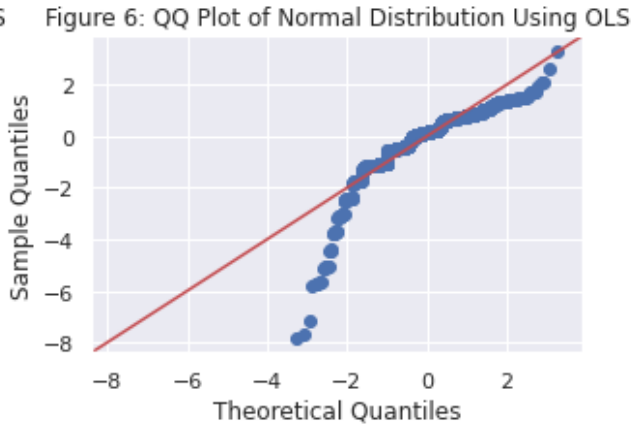


Figure 5: QQ Plot of T Distribution

Figure 6: QQ Plot of Normal Distribution

After conducting a QQ plot of residuals and normality test, residuals are NOT normally distributed. Given the OLS is based on the normal-distributed residuals, the requirement of OLS does NOT satisfied. So, we cannot look at the OLS p-value.

Luckily, the 95% Confidence Interval of the slope is 0.0000000230 +/- 0.0000000018. So, we are 95% confident that the slope lies in an interval whose infimum (i.e., greatest lower bound) is larger than 0. And therefore, we can still conclude that the ratings are increasing. Now, we are very close to concluding that the hypothesis of pup inflation holds.

We want to further view how average ratings are raising with the increase of date. So, inspired by David H. Montgomery's article [2], I made a line chart of monthly average ratings, which is the seventh visualization. The chart is as follows.

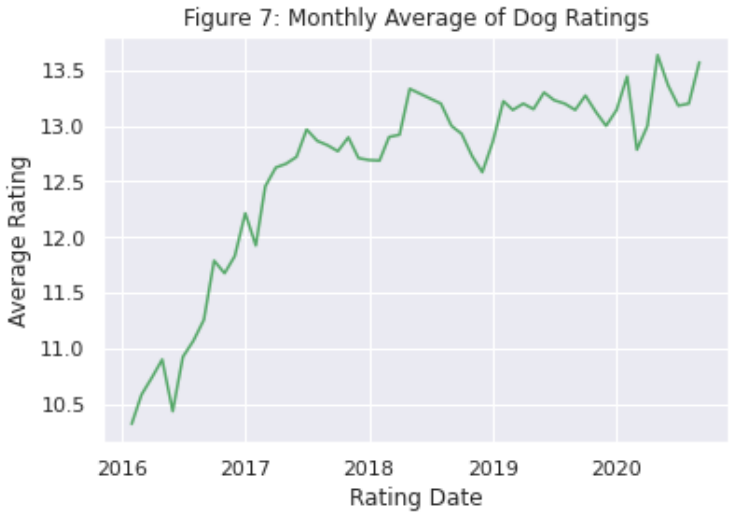


Figure 7: Monthly average of dog ratings.

Although it fluctuates, the average rating is generally increasing. With the increase of average ratings, the number of bad ratings is decreasing. I made a line chart showing the number of bad ratings (less than 10), which is the last visualization of this dataset. This visualization is also inspired by David H. Montgomery's article [2].

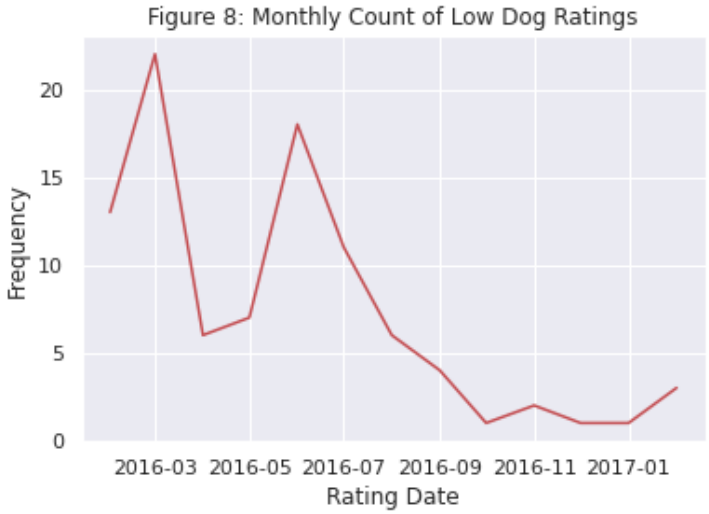


Figure 8: Monthly count of low dog ratings.

The number of low ratings does decrease, even though some fluctuations exist. After February of 2017, there are no low ratings. In other words, all the ratings made after February of 2017 are greater than 10.

IV. Conclusion

We can now draw a conclusion. Given the slope of the linear fit is larger than 0, we can say that there is a positive correlation between ratings and dates. From the monthly average ratings, we can find the ratings are gradually increasing with the increase of date. According to the monthly count of low ratings, we realize that no one rates a dog less than 10/10 after February of 2017.

All the above phenomena lead to a conclusion that pup inflation EXISTS.

V. References

[1] G. Baker, "CMPT 353: Exercise 7," Simon Fraser University, 28 June 2021. [Online]. Available: https://coursys.sfu.ca/2021su-cmpt-353-d1/pages/Exercise7#h-dog-rates-significance. [Accessed 6 August 2021].

[2] D. H. Montgomery, "Pup inflation: Good dogs getting better," Jekyll, 28 Mar 2017. [Online]. Available: http://dhmontgomery.com/2017/03/dogrates/. [Accessed 6 August 2021].

[3] G. Baker, "CMPT 353: Exercise 2," Simon Fraser University, 7 April 2021. [Online]. Available: https://coursys.sfu.ca/2021su-cmpt-353-d1/pages/Exercise2#h-pup-inflation-analysing-tweets. [Accessed 6 August 2021].