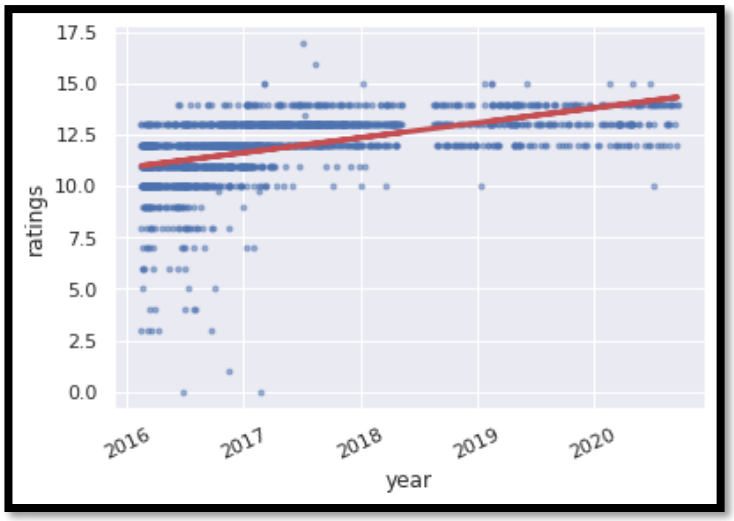
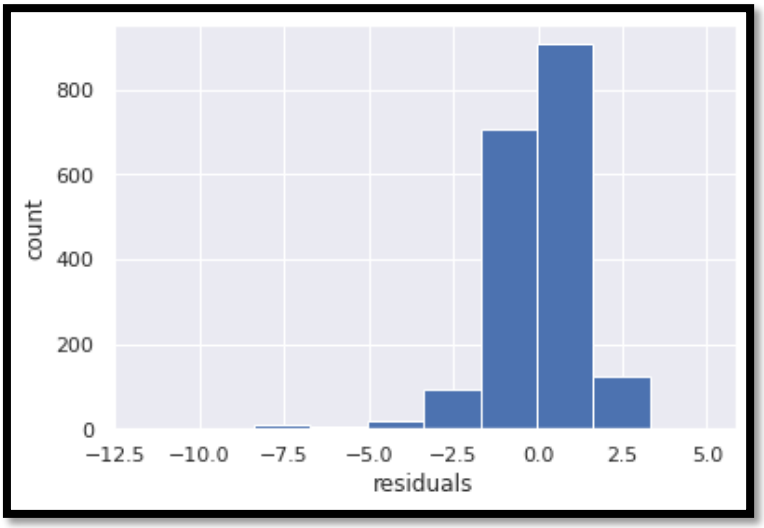
Explaining Pup Inflation





In Exercise 2 and 7, the statistically significant trend in the dog ratings from the @dog_rates Twitter account has been calculated. Data is cleaned by extracting the numeric ratings and removing the outliers (very large ratings) that are too large to make sense. After that linear fitting is applied on the timestamp and rating data to get the slope and intercept for a best-fit line (red line) which in turn is used to calculate the predictions of the ratings. The data is plotted as a scatter plot (top left graph) where the x-axis represents the date (year) of the tweet and the y-axis represents the number of ratings. In the plot, the blue dots represent the actual data and date-time plotting, and the linear red line is the linear best fit line representing a positive linear relationship between date-time and the number of ratings. The left plot represents that ratings increase over time.

After that, the statistics are performed where the p-value is being calculated. P-value represents the probability of occurrence of any given event, and here the question that concerns for p-value is that “Is the slope different from zero?” where a low p-value will determine high chances of this statement to be true. Since the p-value is found to be less than 0.05, it tells us that the slope is indeed different from zero. In order to show what is actually happening in the data, the histogram is plotted (top right graph), where the x-axis is the residuals i.e, difference of observed value and predictions, that are (actual ratings – predictions) plotted against the count on the y-axis. The significance of this graph is to make sure that the outcome we get from the fit-line graph is actually correct. The histogram suggests that residuals are somewhat normally distributed with a positive slope. This tells us that the slope is different than zero which answers our questions. Hence both the plots show that the ratings increase with time and have a positive slope.