Data Wrangling / Analysis

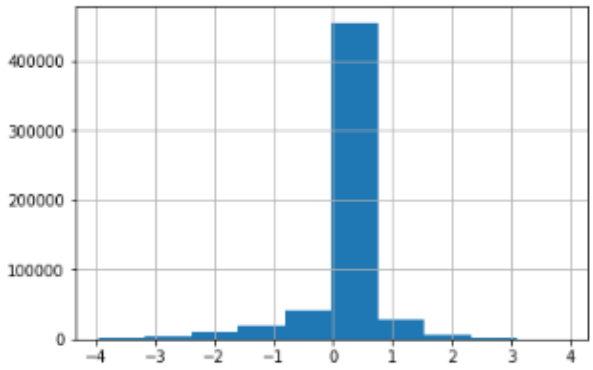


Fig 1. The y axis represents the frequency and the x axis means the normalized rating score for the amazon food review data.

Since different users might be biased towards certain ratings (users who tend to give 5 if they are satisfied vs users who give 4 for the same level of satisfaction), the data will be grouped into users and the average score per user will be calculated. The average score for each user will be used to calculate the deviation away from their average, and set as the column "normalized_score". This gives an idea of how much satisfaction/dissatisfaction a user gained from using a certain product which may be more accurate to use than the star ratings.

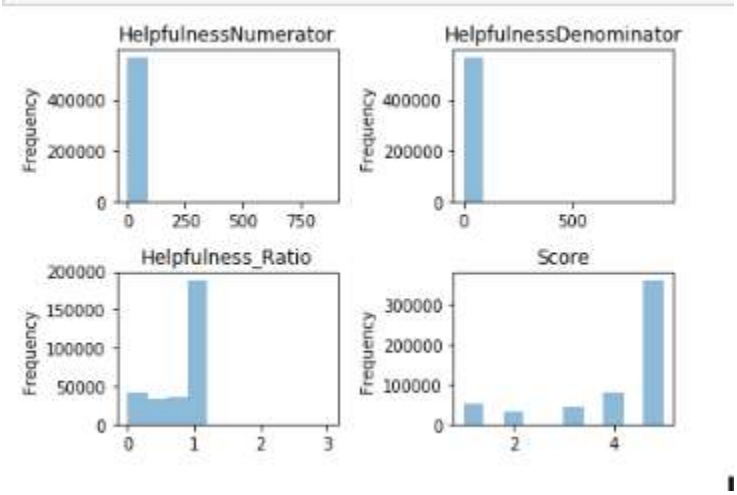


Figure 2. Distributions of the features helpfulness numerator, denominator and score were checked for anomalies. There are definite outliers in this dataset, as seen by the scaling of the histogram, and users tend to give feedback for positive reviews more than neutral or negative reviews. From boxplots, most data is within 0 and 5 for these features. Further a new feature "Helpfulness_Ratio" is created to assess the percentage of helpful tags a review received. Fortunately, there are only two rows with values larger than 1.0 and hence were removed.