Phase 4 Part C: Summary

We used the Local Outlier Factor (LOF) algorithm to classify outliers in our dataset. This algorithm assigns a score to each data point in the dataset based on how different it is compared to its cluster of neighbours. Outliers are data points that have a much lower score than their neighbours.

For the outlier detection, we considered every column in our dataset except for the review_text. We set two of LOF's hyperparameters: n_neighbors (the number of neighbours to consider when computing the LOF score for each data point) and contamination (the proportion of the dataset that is assumed to be outliers). We chose a large value of 800 for n_neighbors, so that the algorithm was more strict in identifying outliers (it would require a larger number of nearby data points to form a cluster before considering a data point as an outlier). We examined the distribution of the review ratings and noticed it is highly skewed towards 5 star reviews; there is an extremely low amount of 1 and 2 star reviews. From this distribution analysis, if we only consider 1 and 2 star reviews as outliers, then we can expect 3.59% + 4.26% (7.85%) of our data to be outliers (7085 of our 90355 reviews). This is how we determined our contamination value of 0.0785.

The algorithm detected 7091 outliers. Not all the outliers are 1 or 2 star-rated reviews; about 70% of them are. We believe this is because, not only is the algorithm imperfect, but there are also other factors at play to identify an outlier (all the columns other than 'rating'). For example, there were only 284 reviews written in 2021, therefore they may be regarded as outliers (regardless of their rating). We checked, and the Local Outlier Factor categorized all 284 as outliers.