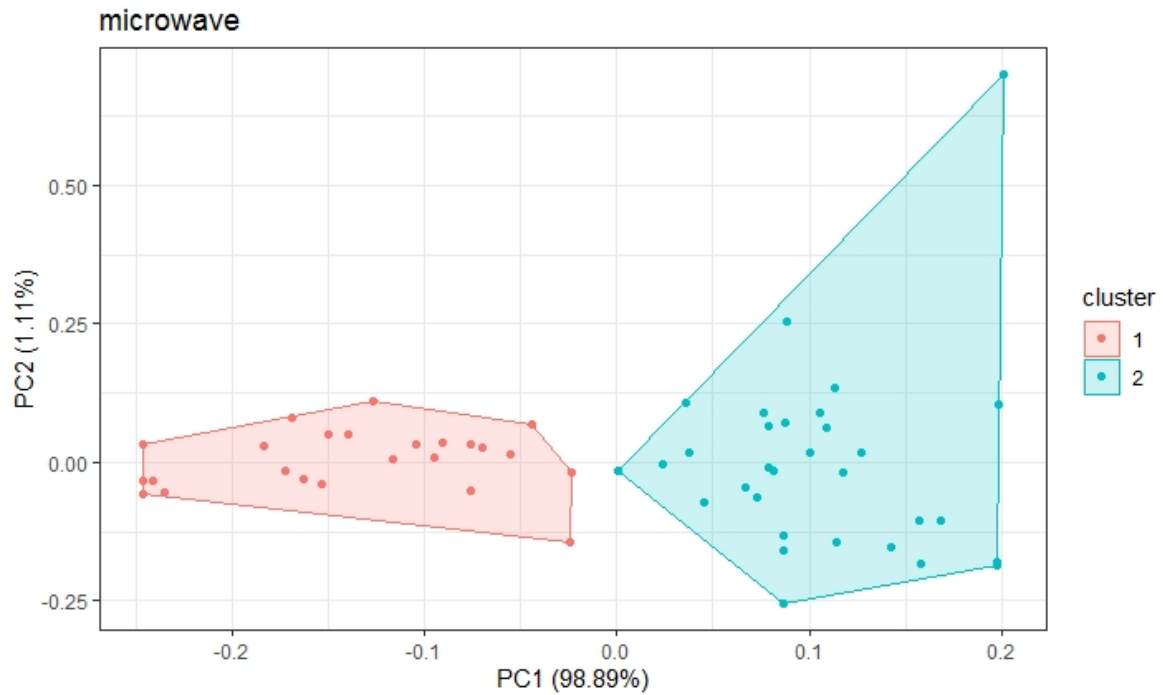# K - means clustering

The comments were clustered according to 'wscore' (weighted sentiment score) and 'star_rating'. Considering the comment homogeneity of products with the same 'product_parent', we average the values of these two variables, then do clustering. As a result, we cluster 55 kind of microwave product.

The variance contribution rate of the first principal component was 98.89%. The variability of the data is basically explained. Since PC1 is actually calculated by 'star_rating' and 'wscore', which is weighted by the sentiment score of the review text, and 'star_rating' is also related to the tendency of emotion, it is reasonable to guess that these two categories may be the division of favorable and unfavorable comments. To test this idea, the next step is to perform a text analysis of the comment text review_body owned by each of the two categories of products, which is explained in Model 3.