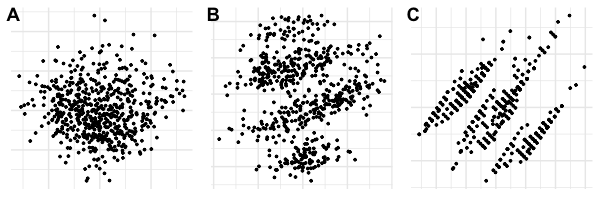
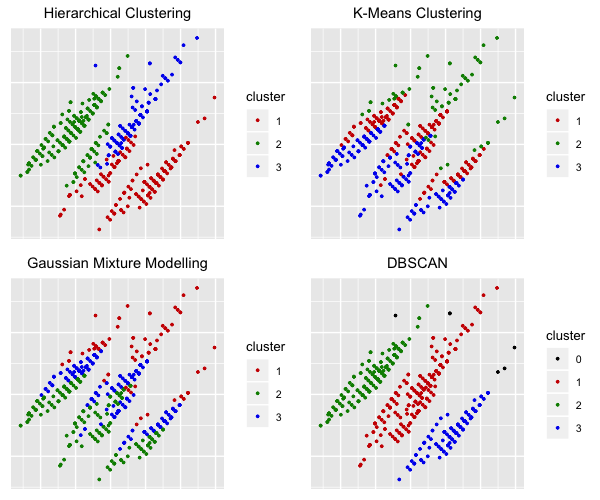
**Clustering**

In the first step three columns were added to the dataset, each containing the mean of the brand awareness, knowledge and involvement values. This will be useful for the evaluation and interpretation of the clusters.

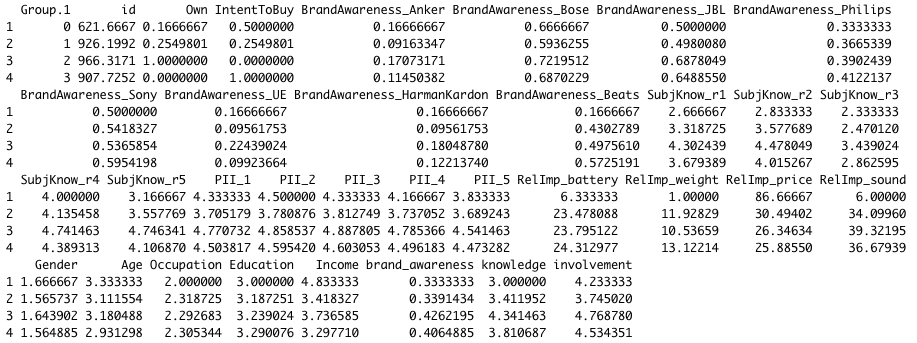
For the clustering it was first necessary to find a subset of the data that contained a sufficient amount of meaningful variables to yield good clusters but didn’t contain too many variables, so the clusters would be sufficiently spaced for the algorithm to correctly indentify them. In the graph below one can see the MDS of three subsets.



A is the MDS of the entire dataset. No clusters are visible. C is the MDS of a subset containing the variables Own, IntentToBuy, RelImp\_Price and Income. Three distinct and well spaced clusters are visible. The subset used for B contains the same variables as C but additionally the above mentioned averages of brand awareness, knowledge and involvement. One can see the outlines of four different clusters, which is promising. Unfortunately the clusters are not sufficiently distinct for any algorithm to yield a good result. Therefore we settled on the subset whose MDS is displayed in graph C.



In the next step we tried out several different clustering algorithms to find the one which achieves the best result. The graphical results of the cluster analysis can be seen in the graph above. While Hierarchical Clustering, K-Means Clustering and Gaussian Mixture Modelling (GMM) have achieved decent results, only Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was able to correctly indentify the three clusters. This makes sense as the clusters are well spaced and have roughly the same density. A great thing about DBSCAN is also that it marks as outliers points that lie alone in low-density regions (i.e. whose nearest neighbors are outside of the set epsilon neighborhood) and adds them to another cluster. As outliers might have something in common too, this is worth taking a look into and opens additional opportunities in the analysis. Below one can see the mean values for each cluster and each variable, including the averages of brand awareness, knowledge and involvement at the end.



Cluster 0 are the outliers, or „Ignorant Rich People“ and contains 6 elements. The have the highest income, very little knowledge, mostly don’t own a speaker and put high importance on price. Cluster 1 are the „Scrimpers“ with 251 elements. They mostly don’t own and don’t want to buy a bluetooth speaker. Overall they are quite average, but have little knowledge and generally care about the price. Cluster 2 are the „Owners“ with 205 elements. They all own and don’t want to buy a speaker, have high knowledge, high involvement and care about sound. Furthermore they have the highest age and the highest income out of the three main clusters. Cluster 3 are the „Poor Graduates“ with 131 elements. They all don’t own a bluetooth speaker, but want to buy one and show rather high involvement. They are the youngest and have the lowest income, but also the highest education level.