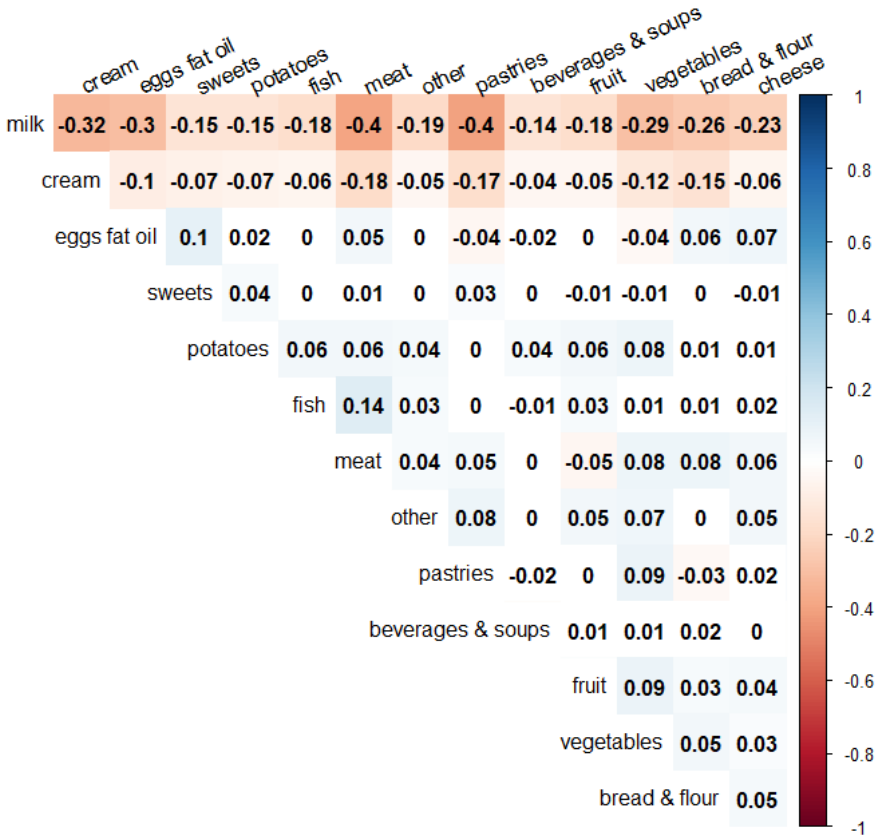
Classifying Households by Diet

In order to understand whether applying clustering to the data would yield meaningful results, I need to analyze whether groups of households have a preference for a certain bundle of foods.

This plot shows the correlation between percent of purchases made in each food category on a household level. Note these are coarse categories used for EDA but the underlying data contains hundreds of categories. Because buying an apple does not have the same meaning as buying a filet-mignon, I weighted all of the purchases by the nutritional value provided by the British government. The purpose of this table is to show whether there are households who prefer similar types of foods. If this turns out to be the case, then a machine learning classification algorithm would be successful at identifying distinct types of purchasers.

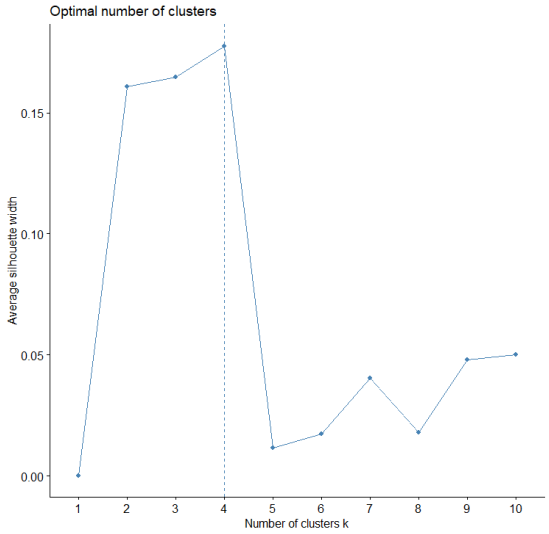
Correlation Between Types of Food Purchases (weighted by calories)



Some insights from the correlation table:

- Purchases in most categories are negatively correlated with one another, with milk as exception.
- The relative strength of correlation between some of the category pairs indicate that there are potential clusters of customers that might be detected in the data.

I took use a k-means clustering algorithm to segment the households into groups based on the percentage of all food spending on a particular food item. One of the choices the machine learning scientist has to make is the number of clusters to use. I consider the optimal cluster number by plotting the silhouette measure resulting from using k=1, 2,…10 clusters:



The plot indicates that running k-means over 4 clusters produces the best balance between the closeness of within-cluster points as well as separation amongst clusters. Note that using the elbow method also produced similar conclusions.

Applying the k-means clustering routine with k=4, I obtain the following characteristics about groups:

Characteristics of Four Clusters of Customers

Name (% of data)	**Top 5 Foods Purchased (% of spending)**				
Cluster 1 (1%)	Semi & Other Skimmed Milk (76%)	Standard white sliced bread (6%)	Fully Skimmed Milk (1%)	Premium sliced bread (1%)	Brown sliced bread (1%)
Cluster 2 (7%)	Wine (27%)	Broiler chicken (2%)	Semi and other skimmed milks (2%)	Fruit juices – (1%)	Lagers & Continental Beers (1%)
Cluster 3 (89%)	Semi and other skimmed milk (4%)	Broiler chicken (3%)	Soft drinks (2%)	Cakes Pastries (2%)	Cheese Natural hard cheddar & cheddar type (2%)
Cluster 4 (3%)	Spirits (27%)	Wine (3%)	Beers (2%(	Soft drinks (2%)	Lager & Continental Beer (2%)

The “typical” household is cluster 3, making up almost 90% of the data, uses a small percentage of their overall spending on one specific food item. Amongst the many items they buy, milk and chicken makeup the largest portion of their spending. Cluster 1 is very small, capturing the small numbers of households who spend almost all their budget on milk and bread. These are likely the households living in poverty. Clusters 2 and 4 collectively make up 10% of the households. I interpret these households as “luxury” consumers who spend far more on wines, beers and other alcoholic beverages than on non-beverage food items.