Lab 3

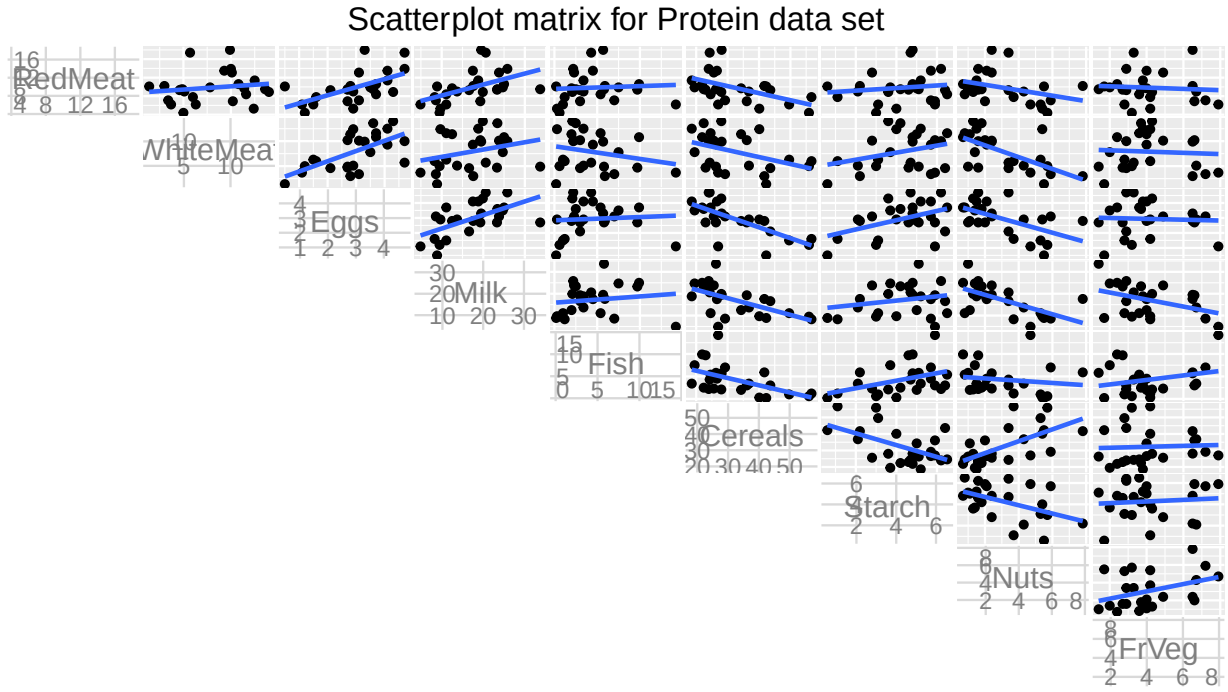
Gustav Sternelöv

September 22, 2016

Assignment 1

1.

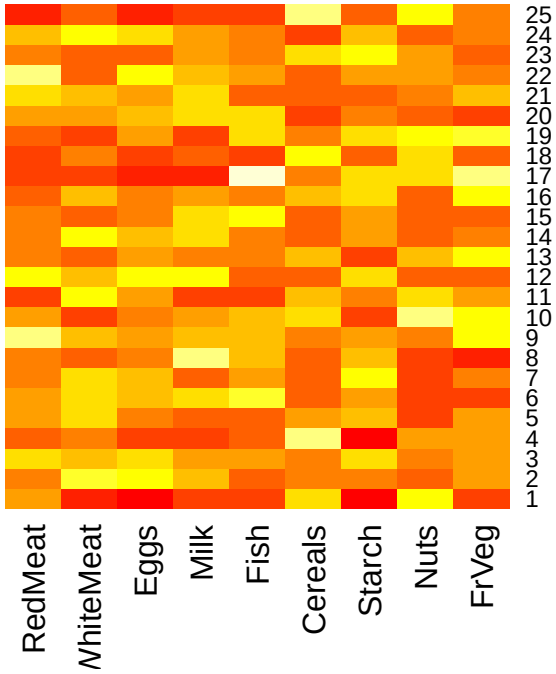
The scatterplot matrix with the variables from the data set *protein* is presented below. For each pair of variables is a linear regression line plotted.



The relationship between the variables is interpreted by looking at the slope of the regression curves. Some relatively strong relationships can be seen, for example between eggs and white meat and between eggs and red meat. Milk and eggs and nuts and starch are two other examples of variables that has a positive relationship. Examples of negative relationships can be seen between starch and cereals and between nuts and white meat.

2.

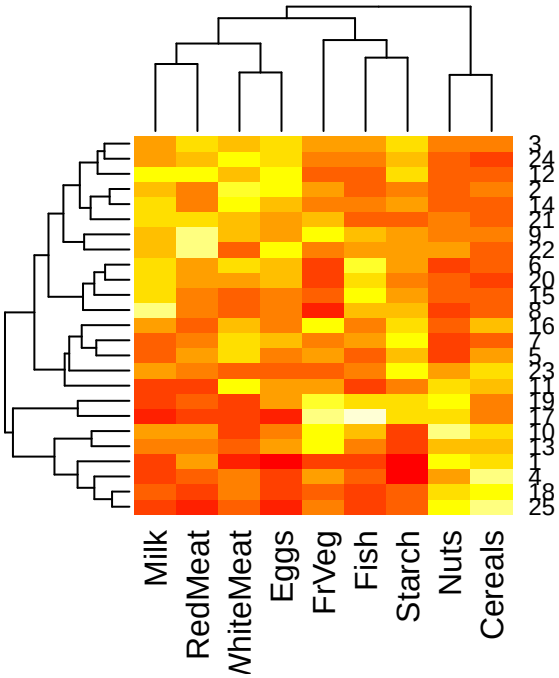
A heatmap for the data with no permutations over the rows or the columns is plotted.



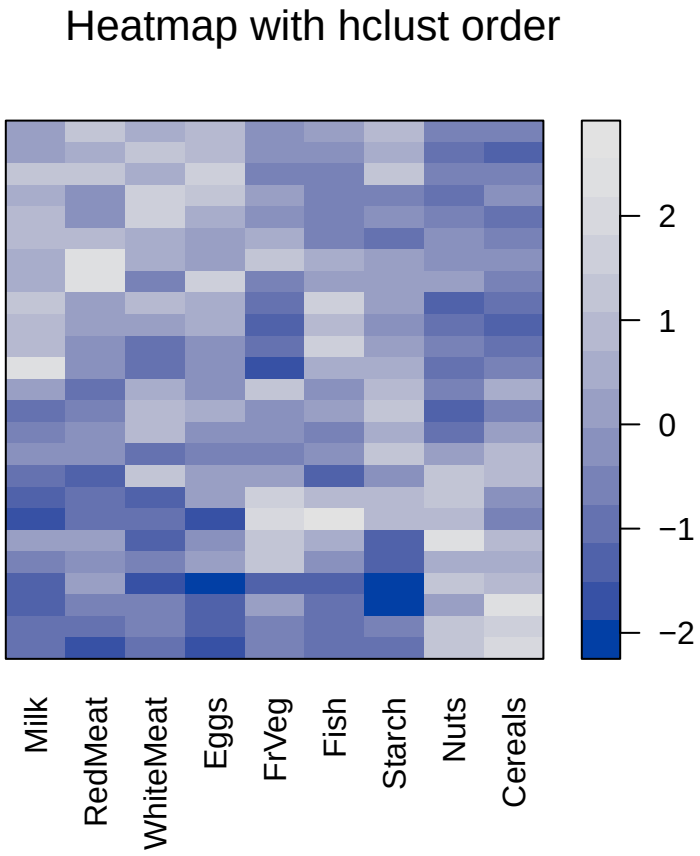
It is hard to find any clear clusters or outliers in the heatmap. Perhaps a correaltion can be noted for red meat and white meat where low values for the former coincides with low values for the latter. The same pattern seem to be present also for high values. Eggs and milk seem to have a fairly similar correlation.

3.

Next, a heatmap given by a two-way hierarchical clustering is created.



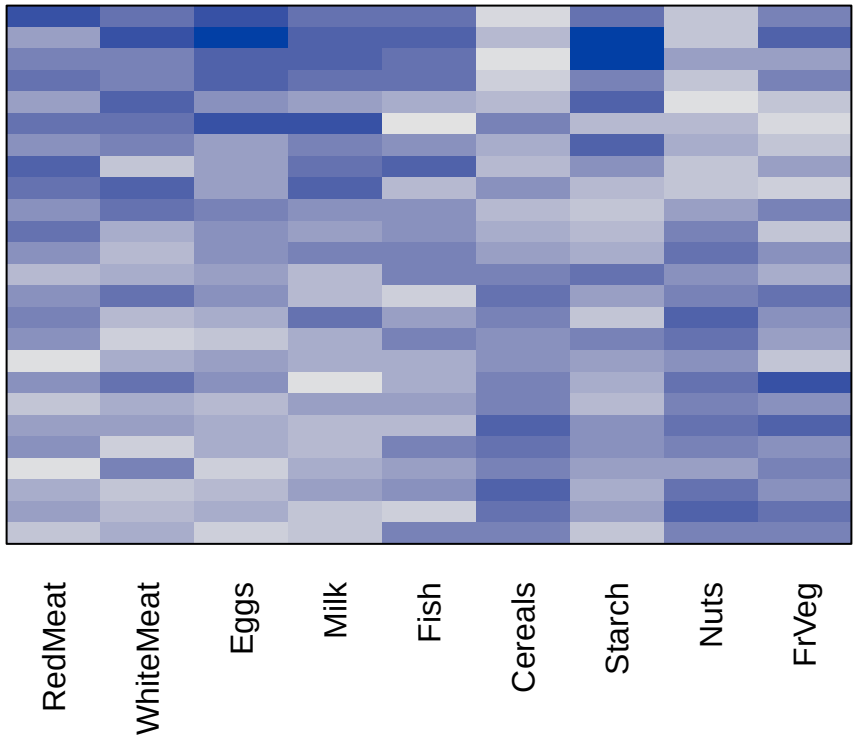
This plot is modified so that another color palette is used and the dendrograms are removed.



Two different clusters are thougt to be seen in the new heatmap, one at the bottom and one at the top. The first cluster, at the bottom, contains the countries with low values for milk, red meat, white meat and eggs and high values for nuts and cereals. The second cluster has the opposite pattern with high values for milk, red meat, white meat and eggs and low values for nuts and cereals. One possible outlier is the observation in the middle of the heatmap which has a high value for milk and a very low value for fruits and vegetables.
The first cluster contains the countries Yugoslavia, Romania, Bulgaria, Albania, Italy, Greece, Portugal and Spain. This is countries from eastern Europe and southern Europe. The second cluster contains Belgium, W Germany, Ireland, Austria, Netherlands, France, UK and Denmark. This is mostly countries from western Europe. It is then 8 countries left, which depending on how many clusters you want either could be splitted into two more clusters or assigned into the most fitting one of the two mentioned clusters. Sweden, Norway, Finland and Poland could either be a third cluster or added to the second. East Germany, Czechoslovakia, USSR and Hungary could be a fourth cluster or added to the first.
In general are countries in the same region clustered together. A reasonable result since it is fairly logical that people from the same region in general consumes the same type of products.

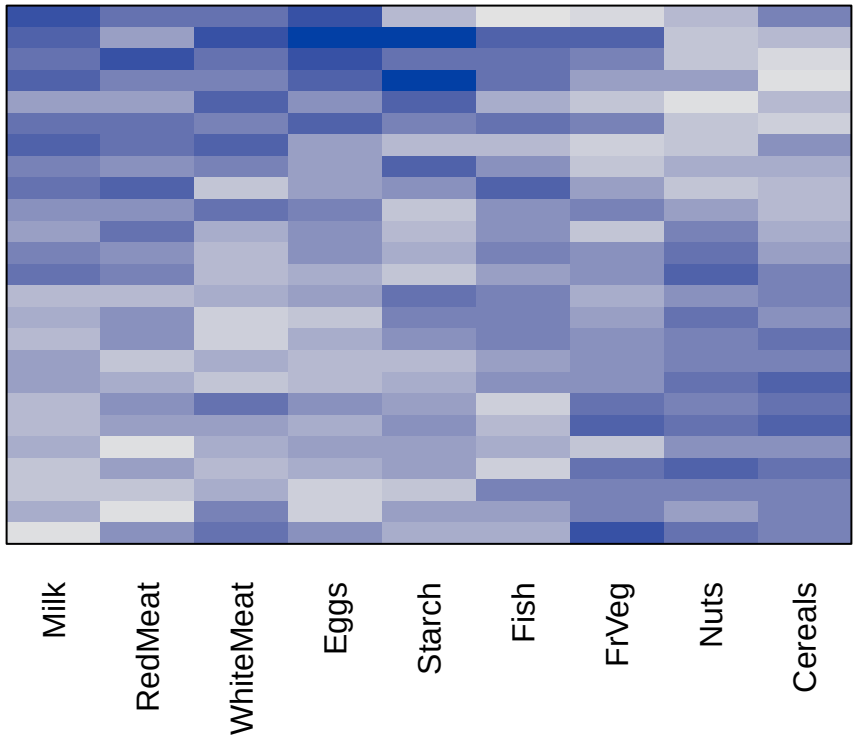
4.

The heatmap for the PCA seriations is plotted below.



Almost the same variables as before (red meat, white meat, eggs, milk, nuts, cereals) are the most important. The only difference is that the impact of the variable for fruits and vegetables is a bit clearer. Regarding the clusters can the same general pattern be seen. One cluster consists of countries from eastern and southern Europe and one of countries from western and northern Europe.

The corresponding heatmap for the BBURCG seriation.



Again, the order of the rows and columns has changed a little but the same overall pattern is present. Perhaps this is the heatmap that is easiest to analyze. The observations at the top has low values for milk, red meat, white meat and eggs and low values for fruits and vegetables, nuts and cereals. The opposite general pattern is seen for the countries at the bottom half of the heatmap. Just as before, the clusters consists of western/northern countries and eastern/southern countries.

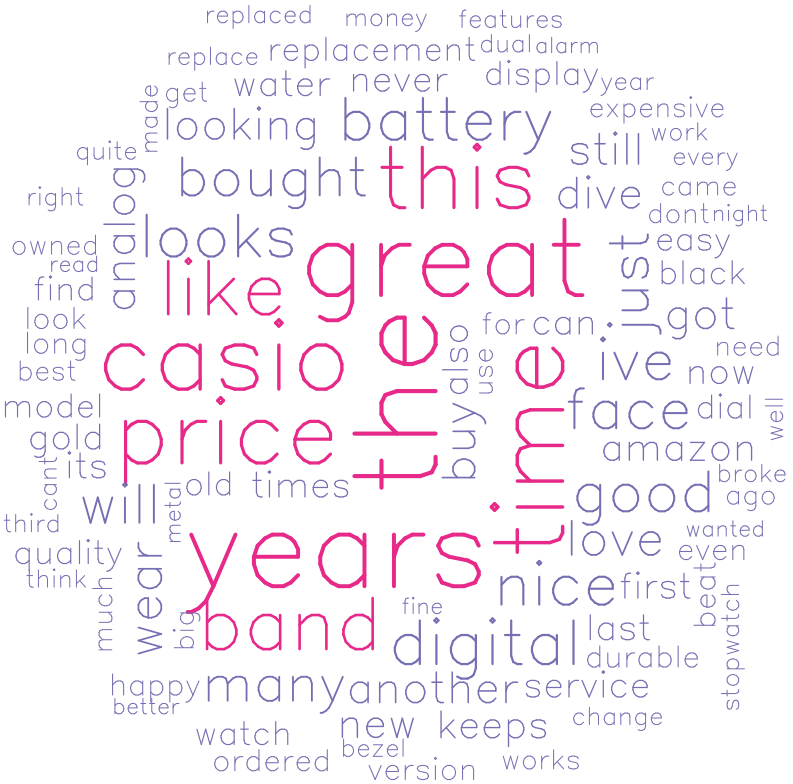
5.

Common between the scatterplot matrix and the heatmaps is the information regarding the relationships between variables. The heatmaps clearly showed the relationship between white meat, red meat, eggs and milk. All of these relationships were apparent already during the analysis of the scatterplot matrix.

Assignment 2

1.

The wordcloud for the feedback from the pleased customers.



Large words in the wordcloud are the name of the brand, Casio, and affection words like “great” and “like”. Other, assumed, positives like “price”, “nice”, “good” and “looks” can be seen among the most frequently used words.

The wordcloud created from the feedback given by the unpleased customers.

Frequently used words by unpleased customers are Casio, “time” and Amazon which is the website where the clock was bought. Some other, relatively, frequent words are “back”, “great” and “like”.

2.

3.