1 Who-is-who of food

A journalist (Malcom Gladwell) once wrote that when two products it makes sense that they do not taste the same. In the data set, sandwiches dataset, we have a breakdown of several sandwich food categories—the following are recorded:

1. the Brand (A – H)
2. the name of the product
3. Category it falls into (e.g. Chicken, Fish, Beef, etc.).

(a) describe how Brand is related to the category of food?

The dataset also has information about 2 more sets of variables, (a) nutritional variables, including:

4. total amount of fat (TFat),
5. protein (Protein),
6. carbohydrates (Carb),
7. fiber (Fiber),
8. sodium (Sodium)

and (b) physical variables:

9. total number of calories (Calories),
10. weight (Weight).

(b) describe how the 5 nutritional variables relate to the 2 physical variables.

2 Pottery

The *pottery.csv* data consists of the results of chemical analysis on Romano-British pottery (45 pots) made in three different regions (region 1 contains kiln 1, region 2 contains kilns 2 and 3, and region 3 contains kilns 4 and 5).

Using the *pottery.csv* data set, standardize the relevant numeric variables (i.e. ignore *kiln*).

a. Create a distance matrix using Euclidean distances and use this to do a hierarchical agglomerative cluster analysis using the centroid criterion. Insert the resultant dendogram. How many clusters would you use? Justify your answer.

b. Use an MDS to visualize the data. Use the MDS results to determine which regions or kilns appear to be distinctive.

3 Consumer Credit analysis

A bank seeks to create an updated risk model to make future credit decisions. Credit bureau data describing individuals (at the time of application) was recorded. The final outcome [i.e. TARGET] of the loan was also determined (as either 'paid-off' or 'bad debt').

The dataset is ("Ass2Credit.csv"). The variable details are described below:

Variable	**Description**
TARGET	1=Bad Debt, 0=Paid-off
CollectCnt	Number of Times a debtor has been called
InqFinanceCnt24	Number Finance Inquiries in the 24 Months

InqTimeLast	Time Since Last Credit Inquiry
TLTimeFirst	Time Since First 'Trade Line'[1]
TLBalHCPct	Percent Trade Line Balance to 'High Credit'[2]
TLSatPct	Ratio of 'Satisfactory Trade Lines'[3] to 'Total Trade Lines'
TLSum	Total Balance All Trade Lines
TLOpenPct	Percent Trade Lines Open
TLDel60Cnt24	Number of trade lines 60 days+ in last 24 months [late payments]

a. Do a Canonical Discriminant (Function) Analysis (DFA) to find a linear function of the variables that best discriminates between those people who pay off their debt and those that don't (i.e. TARGET = 0 and 1 respectively).

1 How many discriminant functions are there? Plot the results of the discriminant analysis.

2 Which variables are most important in discriminating the bad and good (i.e. 'paid-off') debtors?

3 Can you reduce the number of variables required? Justify your answer

b. Using only those variables that you deem to give good predictive power, do a Fisher Discriminant Analysis and estimate the mis-classification error rate that you would expect to get with new data. **Justify your choice of either linear or quadratic discriminant analysis**.