'Data Analysis' — Computer Lab Session Exam
Master 1 MLDM
Saint-Étienne, France

Fabrice Muhlenbach & Damien Fourure

Preliminary Remarks

Exam in 'Data Analysis' using R *Project for Statistical Computing*. Documents (lecture and lab sessions notes) are allowed, R codes too, but not Internet searches, forums, e-mail, web chat or other kind of communication. The total score of all the exercises is equal to 100 points. Write your answers to the questions in your R code as a comment, in a sentence preceded by a sharp symbol (#). Use your family name for naming your file (so your file will be FamilyName_FirstName.R) and send your file via *Claroline* platform.

1 *Quatre-Quarts*, *Kouglof*, *Crêpes*, *Îles flottantes* and *Beignets* (30 points)

The pastry chef of a boarding school became ill and has to be replaced for the next week. The apprentice pastry who replace him will prepare the afternoon snacks for the pupils for everyday of the week (5 days, Monday to Friday) but he does not have the exact number of pupils (he was just told that there are 5 classes of pupils, so he imagines there must be more than a hundred of pupils). The two information available to him are: (1) the recipes of the 5 desserts or pastries originally planned by the pastry chef for the week and (2) the exact ingredients quantities necessary for the preparation of these five desserts or pastries.

In the fridge and the stocks, the apprentice pastry found:

- butter: 7.75 kg (= 7,750 g)
- (white) sugar: 7,630 g
- (all-purpose) flour: 31.5 kg
- (medium) eggs: 333
- milk: 27.6 liter (= 27,600 ml)

There are also salt, baking powder (for the crêpes, the kouglof and the beignets), honey, jam, marmalade, chocolate spreadable paste, almonds, icing sugar, raisins (in French: « raisins secs »), oil for deep frying and other things in undefined quantities.

The recipes of the chef found by the apprentice give the following quantities:

- pound cake (in French: « quatre-quarts »), serves 6: 125 g butter, 125 g sugar, 125 g flour, and 2 eggs (no milk)
- *kouglof* (a pastry from Alsace), serves 8: 150 g butter, 70 g sugar, 400 g flour, 3 egg, and 80 ml milk

- *crêpes* (a kind of pancakes), serves 10 (=20 crêpes): 50 g butter, 20 g sugar, 250 g flour, 4 eggs and 500 ml milk
- floating island (eggs in snow, in French « îles flottantes » or « œufs à la neige »), serves 5: 60 g sugar, 5 eggs, and 600 ml milk (no butter and no flour)
- *beignets* (a kind of doughnuts), serves 6 (=18 midsize beignets), 120 g butter, 120 g sugar, 1000 g flour, 4 eggs, and 300 ml milk.

Questions

20 pts Solve the problem with R in order to find the number of people (the pupils of the boarding school).

10 pts After that, indicate the number of pound cakes, *kouglof*, *crêpes*, floating islands and *beignets* to be prepared during the week.

2 Offshore Fishing (40 points)

A new underwater detection system is able to identify different schools of fish. In most cases, fishes are grouped together by members of the same species (called "fish schools"). For fishermen who throw large nets into the sea, it is important to have an idea of the number of species present in order to avoid catching uninteresting fish species. The file fishes.csv (found on *Claroline*) indicates the presence of fishes on an underwater area under the fishing boat.

Questions

5 pts Plot the pairwise representation (a matrix of scatterplots) of this dataset (R function pairs). How many clusters do you think to find in this dataset?

10 pts Find two different methods for better representing the dataset. Plot the data with these other methods. Now, how many clusters do you think to find in this dataset?

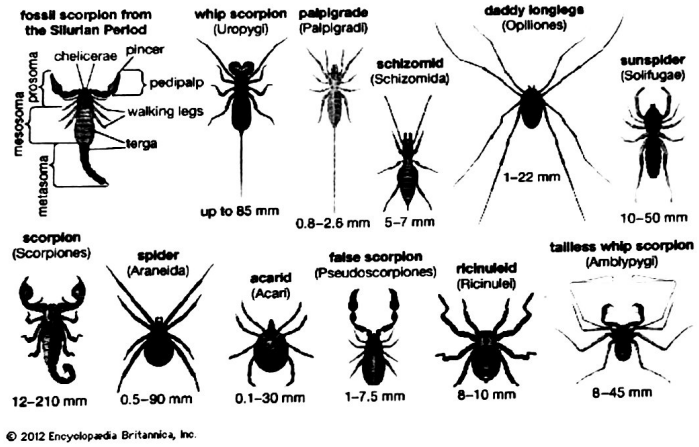
15 pts
- Use the k-means algorithm (function kmeans) from k = 2 to 10 on the dataset. For each value of k, run the k-means algorithm 5 times (the clustering results can change due to the random choice of the initial k centroids), print the value of the within-cluster sum of squares (WSS), the between-cluster sum of squares (BSS), and compute the Calinski and Harabasz clustering quality index (obtained by $C\_H\_index = \frac{BSS/k-1}{WSS/n-k}$, with k the number of clusters and n the number of observations).
- Consider for each k-means try with a given number k the best value (= the maximal value) obtained for the 5 tries for the Calinski and Harabasz clustering quality index. For which value of k this index is maximal, i.e., how many clusters this index proposes to find in this dataset?
- Run k-means algorithm another time with k equals to the number suggested by Calinski and Harabasz index. Plot the dataset in 3 dimensions (by using rgl package) and with a different color for each cluster (e.g., by using col=kmeans.result$cluster). What will be your final suggestion of the number of schools of fishes?

10 pts Install the package cluster and load this library. Compute the Euclidian distance on the dataset (function dist). Run k-means algorithm with the number of k clusters that you have discovered, as well as k-means algorithm with the number $k+1$ clusters and $k-1$ clusters, and save the results in 3 variables. For each of the 3 k-means results, plot the *silhouette* index:

```
sk <- silhouette(kmeansresults$cl, fishDistance)
plot(sk)
```

The *silhouette* is a method of interpretation and validation of consistency within clusters of data. The *silhouette* ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high *silhouette* value, then the clustering configuration is appropriate. Does the *silhouette* graphic result seem to be consistent with the result given by Calinski and Harabasz index? Explain why.

3 Arachnid (30 points)



A biologist sent you a file recording the information concerning 33 arachnids[1] he studied (file arachnids.csv found on *Claroline* platform). In this file, they are 3 numerical attributes (V1, V2 and V3), the sex (male or female) and the age (in month). The biologist wants to know if it is possible to predict the age of an unknown arachnid of this species with the 3 variables which are easily measurable numerical features.

Questions

5 pts Download the file from the *Claroline* platform and load the CSV file using the function read.csv or with "Import Dataset" in RStudio. Print the summary and plot the data pairs representation. Print the correlation coefficient of the numerical variables (all attributes but not the sex which is a factor). Is one variable (or more) well correlated with the age? Is it possible to obtain easily a linear regression model for predicting the age with the other variables?

10 pts Plot the pairwise representation of all the numerical attributes with a different color for the sex. Do you see a difference? Divide the dataset in two subsets: the males and the females. Print the correlation coefficients for each subsets. Do you think that it is possible now to obtain a good regression model for predicting the age with one ore more other variables?

15 pts What will be the most interesting variable for this model? For each sex subset of arachnid observations, compute a linear model for predicting the age. With these models, can you predict the age of a female arachnid with the parameters V1=30, V2=16, and V3=23?

[1] Arachnid is the class of joint-legged invertebrate animals (e.g., the spiders or the scorpions, as shown on the figure)