




‘From Statistics to Data Mining’ — Computer Lab Session Exam

Master 1 COSI / CPS2

Saint-Étienne, France

Fabrice Muhlenbach

Preliminary Remarks

‘From Statistics to Data Mining’ exam using  *Project for Statistical Computing*. Documents (lecture and lab sessions notes) are allowed,  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  code as a comment, in a sentence preceded by a number sign or hash symbol (#). Use your family name for naming your file (so your file will be `FamilyName_FirstName.R`) and send your file via *Claroline* platform. Exam duration: 90 mn (1:30).

1 *Gaufres, Quatre-Quarts, Crêpes, Îles flottantes and Beignets* (40 points)

The pastry chef of an infant school of 100 pupils became ill and has to be replaced for the next week. The apprentice pastry who replace him will prepare the desserts for the pupils for everyday of the week (5 days, from Monday to Friday) but he does not know the exact number of pupils who are eating at the canteen (of course 100 at the most). He was just told that most children have lunch in the canteen but there are usually fewer people on Wednesdays because the children do not have class in the afternoon. 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: 4.935 kg (= 4,935 g)
- (white) sugar: 6,075 g
- (all-purpose) flour: 21.375 kg
- (medium) eggs: 252
- milk: 22.5 liter (= 22,500 ml)


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

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

- for Monday: waffles (in French: *gaufres*), for waffle dough serves 8: 80 g butter, 100 g sugar, 250 g flour, 4 eggs, and 350 ml milk
- for Tuesday: pound cake (in French: « quatre-quarts »), serves 6: 125 g butter, 125 g sugar, 125 g flour, and 2 eggs (no milk)

- for Wednesday: *crêpes* (a kind of pancakes), serves 10 (=20 crêpes): 50 g butter, 20 g sugar, 250 g flour, 4 eggs and 500 ml milk
- for Thursday: 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)
- for Friday: *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


- 30 pts Solve the problem with  in order to find the number of people (the number of pupils eating in the canteen for each day of the week).
- 10 pts After that, indicate the number of waffles, *quatre-quarts*, *crêpes*, *îles flottantes* and *beignets* to be prepared during the week.

2 Space Battle (60 points)

In a video game, you participate to the strategic team of the Earth Union whose mission is to protect the planet Earth and its inhabitants against an invasion of alien spaceships. A probe spy from the Earth Union has identified the spatial positions of the enemy warships. You have to identify how these enemy warships are grouped so that the Earth Union Space Force will be able to make the best use of its resources in defense spaceships.

Load the CSV file `space_battle.csv` (found on *Claroline* platform) with the `read.csv` function or with “Import Dataset” in RStudio. The dataset contains one header (with the name of each variable), 100 observations (the alien war spaceships) and 3 variables (X_1 , X_2 and X_3 representing the space coordinates of the alien spaceships).

Questions

- 5 pts Plot the pairwise representation (a matrix of scatterplots) of this dataset ( function `pairs`). How many clusters do you think to find in this dataset?
- 20 pts Find 2 other methods for better representing the dataset. Plot the data for each method. Now, how many clusters do you think to find in this dataset?
- 30 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?
- 5 pts 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`). Do the same (a plot with a different color for each cluster) with the pairwise representation. What will be your final suggestion of the number of groups of Earth Union defense spaceships to send for protecting your planet against the alien invasion?