




‘Data Analysis’ — Computer Lab Session Exam

Master 1 MLDM / COSI / CIMET / 3DMT

Saint-Étienne, France

Fabrice Muhlenbach & Damien Fourure

Preliminary Remarks

Exam in ‘Data Analysis’ 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 sharp symbol (#). Use your family name for naming your file (so your file will be family_name.R) and send your file via *Claroline* platform.

1 Rush hour at the fast-food restaurant (30 points)



Your friend was hired at a famous brand of fast food restaurant. After his first day he explains to you that the restaurant organization is awful and he asks for your help.

He gives you the `fast_food.csv` file (found on *Claroline* platform). The file contains the order time for each customer of the day. The time is in decimal format (for example 10:30 AM = 10.50).

Questions

- 10 pts Download the file from the *Claroline* platform and load the CSV file using the function `read.csv`. Print the summary and plot the histogram. Assuming that the data are drawn from a Gaussian distribution, compute the mean and the standard deviation and plot the Gaussian density onto the same plot than the histogram (be careful with the X-axis limits).
- 5 pts Your friend asked you how many customers he should expect between 01:30 PM and 06:00 PM (= `]13.5; 18]`). Compute this estimation according to the Gaussian density. Write a function to count the real value to compare it with your estimation. Do you think that the estimation is correct?
- 5 pts To get a better estimation, assume that the data are drawn from two Gaussian distributions. Split the data into two subsets: the first one with all customers who came before 03:00 PM (`<= 15`) and the second one with all customers who came after 03:00 PM (`> 15`).
- 5 pts Compute the means and standard deviations of the two Gaussian distributions (from the subsets) and plot them on the same graph than the histogram.
- 5 pts Compute the estimate value according to the two Gaussian distributions. Is this estimate better than the previous one? If so, or if not, why?

2 Sold your car (30 points)



You own a light van car but you are not using it anymore so you want to sell it. The problem is that you do not have any idea of the price that you can ask. Fortunately you have the file `light_van_car.csv` (found on *Claroline* platform) who contains information about 67 cars. For each observations/car you have 3 values: the price at which the car has been sold, the number of kilometers of the car and its age (in years, for example a value of 2 means that there is model of 2014).

Questions

- 5 pts Download the file from the *Claroline* platform and load the CSV file using the function `read.csv`. Print the summary and plot the data pairs representation. Do you see an outlier? If so, why it is an outlier?
- 5 pts Intuitively, the price of a car decrease in a logarithmic way in function of its age and kilometers (you can see this onto the pairs representation). Preprocess data by applying a logarithmic function onto the data price and age (not onto the cost). Print the new data pairs representation. Do you see a new outlier? If so, why it is an outlier?
- 5 pts Remove all the outliers. For that you can create a new list by assigning the values of an old one minus the data at a specific index like that:

```
new_list <- old_list[-index,]
```

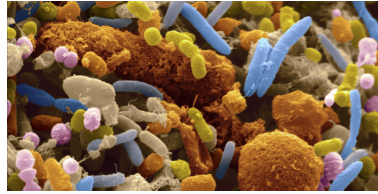
- 5 pts Plot the cost as a function of the number of kilometers. Compute a simple linear regression of the cost in function of the number of kilometers by using the `lm` function. Plot the fitted line and observations with a confident interval of 90% for the prediction.
- 5 pts Your car has a lot of kilometers but it is not very old so you want to take also the age of the car into account. Compute a multiple linear regression following the model:

$$Cost = \beta_0 + \beta_1 \times \log_Nb.km + \beta_2 \times \log_Years$$

with `log_Nb.km` and `log_Years` the logarithmic values of the `Nb.km` and the `Years` respectively.

- 5 pts Your car has 86,900 km and is 2 years old. What is the estimate price that you can ask for it?

3 Bacteria (40 points)




A doctor friend doing his research in the field of bacterial epidemics contact you for helping him to better understand a biological phenomenon. He sent you a data file containing the results obtained on bacterial tests conducted in 10,000 biological samples. As he wants to develop antibiotics specifically dedicated to a particular population of bacteria, he wants you to help him to understand how many different bacteria groups were found in the studied samples.

Load the CSV file `bacteria.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), 10,000 observations (the biological samples) and 5 variables (X_1 , X_2 , X_3 , X_4 and X_5 representing the results obtained on the different biological tests). The line number is also present in the data file, it can be used as the row name and then removed for not interfering with the analysis.

```
row.names(bacteria) <- bacteria[,1]
bacteria[,1] <- NULL
```

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?
- 5 pts Find another method for better representing the dataset. Plot the data with this other method. Now, how many clusters do you think to find in this dataset?
- 5 pts Before analyzing the data, it is necessary to prepare them. Check if the variables are correlated and, if so, remove one of the variables that would be a duplicate of the other one.
- 5 pts To continue the data preparation, display the summary of the data and transform the data so that all the variables will have the same scale with a z-score standardization.
- 20 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 groups of bacteria?