# G0O02a: Statistical Data Analysis: Project 2

Prof. Mia Hubert, Ruicong Yao

May 2024

This project consists of a **clustering part** and a **regression part**.

You answer the questions by performing an appropriate analysis with `R`. Use the provided R Markdown template to add your code and interpretation.

Your R Markdown file should be named "LastName_FirstName_Project2.Rmd" and uploaded on Toledo on **June 15, 2024** (23h) at the latest. This project is graded on 6 points.

## Part 1: cluster analysis

The clustering part is based on the Rice dataset (from the first project) on 5 different varieties of rice. For this assignement we only consider the variables `Area, Perimeter, Major Axis, Minor Axis, Solidity` and `Class`. Here's a brief overview of the variables:

| | |
|---|---|
| **Area (A)** | The area of a rice grain and the number of pixels within its boundaries. |
| **Perimeter (P)** | Rice circumference is defined as the length of its border. |
| **Major Axis (L)** | The distance between the ends of the longest line that can be drawn on a rice. |
| **Minor Axis** ($\ell$) | The longest line that can be drawn from the rice while standing perpendicular to the main axis. |
| **Solidity (S)** | The ratio of the pixels in the convex area to those found in the rice grain. |
| **Class (Cl)** | 1 (Arborio), 2 (Basmati), 3 (Ipsala), 4 (Jasmine), 5 (Karacadag) |

The goal is to cluster the data based on the 5 continuous variables **ignoring the given class**, and to investigate to which extent this clustering is in agreement with the given class.

Each student randomly draws an individual data set of 400 observations. You use the following code, where you change 0012345 by your student number.

```
set.seed(0012345)
rice_dataset <- read.csv("Rice_Dataset_5.csv")
mysample <- sample(1:nrow(rice_dataset), 400)
mydata <- data.frame(rice_dataset[mysample, c(1:4, 7, 11)])
```

1. Argue whether the data should be rescaled or not, and standardize your data if you find this appropriate.

2. Make a scatterplot matrix of the first 5 variables where you color the observations according to their given class (6th variable). What do you observe? What do you expect when you would cluster the data into 5 clusters?

3. Order the observations in your dataset according to their given class. Construct an appropriate dissimilarity matrix (using the first 5 variables) and visualise it. What do you observe? Is it in line with your findings on the scatterplots?

4. Perform $K$-medoids with 5 clusters. Don't use the class information! Investigate to which extent the resulting clustering is in agreement with the given class.

5. Assume now that you have no a priori idea about the number of clusters in your data. What value of $K$ would you then select? Perform $K$-medoids with that value of $K$ and discuss how well it classifies the observations into $K$ clusters. If your chosen $K$ is different from 5, describe to which extent the resulting clustering is in agreement with the given class.

6. Make a heatmap of the observations, ordered according to their assigned cluster (based on your chosen $K$). Interpret the result.

# Part 2: Regression

The regression part is based on the cars dataset:

| Variable | Description |
|---|---|
| manufacturer | Car manufacturer or importer. |
| model | Car model. |
| description | Further details on the car model. |
| euro_standard | Euro Standard to which the record applies. |
| transmission_type | Transmission type. Either Automatic or Manual. |
| engine_capacity | Engine capacity in cubic centimeters (cc). |
| fuel_type | Fuel type this car uses, Diesel, Petrol or Hybrid. |
| urban_metric | Fuel consumption in urban conditions in liters per 100 Kilometers (l/100 Km). |
| extra_urban_metric | Fuel consumption in extra-urban conditions in liters per 100 Kilometers (l/100 Km). |
| combined_metric | Combined fuel consumption: average of the urban and extra-urban tests, weighted by the distances covered in each part, in liters per 100 Kilometers (l/100 Km). |
| noise_level | External noise emitted by a car shown in decibels. |
| co2 | $CO_2$ emissions in grammes per kilometer (g/km). |
| co_emissions | Carbon monoxide emissions in milligrammes per kilometer (mg/km). |
| nox_emissions | Nitrogen oxides emissions in milligrammes per kilometer (mg/km). |

The goal is to predict the $CO_2$ emissions of the cars from the other variables, except `manufacturer`, `model` and `description`. Each student randomly draws an individual data set of 500 observations. You use the following code, where you change 0012345 by your student number.

```
cars_data <- read.table("cars_data.txt", sep = "", header = T)
cars_data$euro_standard <- as.factor(cars_data$euro_standard)
cars_data$transmission_type <- as.factor(cars_data$transmission_type)
cars_data$fuel_type <- as.factor(cars_data$fuel_type)
set.seed(0012345)
data_ind <- sample.int(n = nrow(cars_data), size = 500, replace = F)
mydata <- cars_data[data_ind, -c(1,2,3)]
```

1. Consider the linear regression model containing all observed predictor variables. Do not transform the predictors nor the response. Interaction terms or higher order terms do not need to be included. We call this model $M_1$. Does this model suffer from multicollinearity?

2. Perform stepwise regression starting from the model which only includes the intercept. This makes the model $M_2$. Does model $M_2$ suffer from multicollinearity?

3. If appropriate, remove one or more variables from $M_2$ to reduce the degree of multicollinearity. The resulting model is $M_3$ (this might be equal to $M_2$).

4. Investigate whether $M_3$ satisfies the assumptions of the normal linear regression model.

5. Are all predictor variables in $M_3$ significant? If not, remove those that are not significant. This makes model $M_4$.

6. Compute the following statistics for all considered models ($M_1$, $M_2$, $M_3$ and $M_4$): $R^2$, $R^2_{adj}$, AIC, RMSE on the full dataset, RM-PRESS and $\text{RMSE}_{cv}$ with 5 folds. Discuss the difference of the models on these metrics and argue which model you select as the best one.