# Data and AI assignment 2: Combine and refine

## Team members:

1. Astrid Piot
2. Catarina Rilho
3. Cristina Marques
4. Greg François Reynders
5. Robbe Van de Velde
6. Seppe Haems
7. Rocío Pérez Holgado

Overall score: 3/3

## Summary

We can see that you put a lot of effort into this assignment. You demonstrate great ability to analyze data and work with both the tools and concepts covered in class. We look forward to see how you will perform in the next assignments.

## Defining goals 1/1

1. Unable to sufficiently identify business or data mining goals
2. Able to briefly identify business and/or data mining goals
3. **Able to explain interesting business goals and data mining goals in a reasonable and compelling way.**

**Motivation**

Your group did what we were looking for here which was defining a business goal and translating it into a clear data mining goal. Adding the sub questions on the top of your document and using that to structure your analysis is a good addition too.

## Handling null values and outliers 1/1

1. Unable to identify
2. Able to identify null values
3. **Able to demonstrate appropriate ways to handle null values and outliers**

**Motivation**

The null values in the date column of the assessments column were added by us at random. For these it's harder to do a concrete action. Simply ignoring them for now is fine as well.

Changing the score of students that did not submit to -1 works, however it would have been equally good to set these students scores to 0.

We introduced a second challenge by making a lot of values of students based in Scotland missing. It is important that you indeed picked up on this fact. As you wrote in your report the correct course of action is most likely to contact business and potentially just exclude students from Scotland out of your analysis.

Making a separate category for missing values is something you did correctly as well. As we saw in class, the missingness of certain values provides information as well.

You checked whether the sum of the non-exam assessments was 100 % and each had a weight of 100 %. Strictly speaking this is not correct, because some modules had two exams such that the total weight for the course was 300 % instead of the usual 200 %. You can check this with `assessments.groupby(["code_module", "code_presentation"]).sum("weight")`. As you did check whether both assumptions made by the data dictionary in detail we believe it is harsh to subtract points for this.

## Appropriate joins 1/1

1. Only attemps to analyze data from separate tables
2. Able to perform basic analysis from at least two joined tables
3. **Able to merge data appriopriately from several data sources to perform complex analysis**

**Motivation**

You completed a detailed analysis and already identified a number of key drivers of student failure such as imd_band, disability status, region, previous education and more. In conjunction with what you alreayd having a clear data mining goal we believev you're in a good position to start the next assignments.