Assignment 2

Part 1: preliminary data analysis and pre-treatment

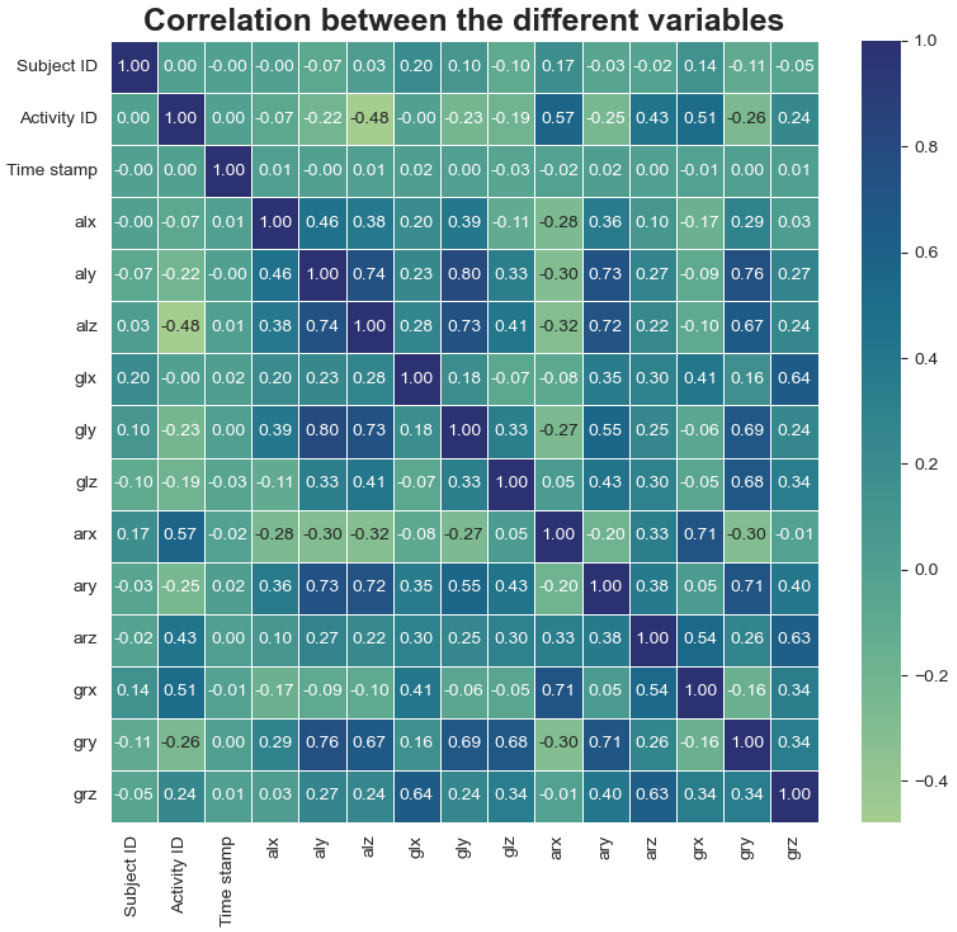
In this part of the assignment, you will reflect on what is contained in your data and how you can pretreat your data.

1.1. Preliminary analysis

Q1.1. Are there any highly correlated variables (*i.e.*, correlation larger than 0.90)? [2 points]

- □ Yes
- ✓ No

Visualize this answer with a *single* figure and report it below.



Q1.2. Are there any missing data? [3 points]

- □ Yes
- ✓ No

Q1.3. How many duplicated rows are present in your dataset, if any? Specify the number below: [5 points]
____0____

If you have replied yes to Q1.1 remove one of the highly correlated variables from your dataset. If you have found missing data or duplicates, proceed to remove the corresponding rows. If you have not found correlated variables, missing data, or duplicates, you can proceed to the next parts of the questionnaire and use your pretreated data for the assignment.

1.2. Data scaling

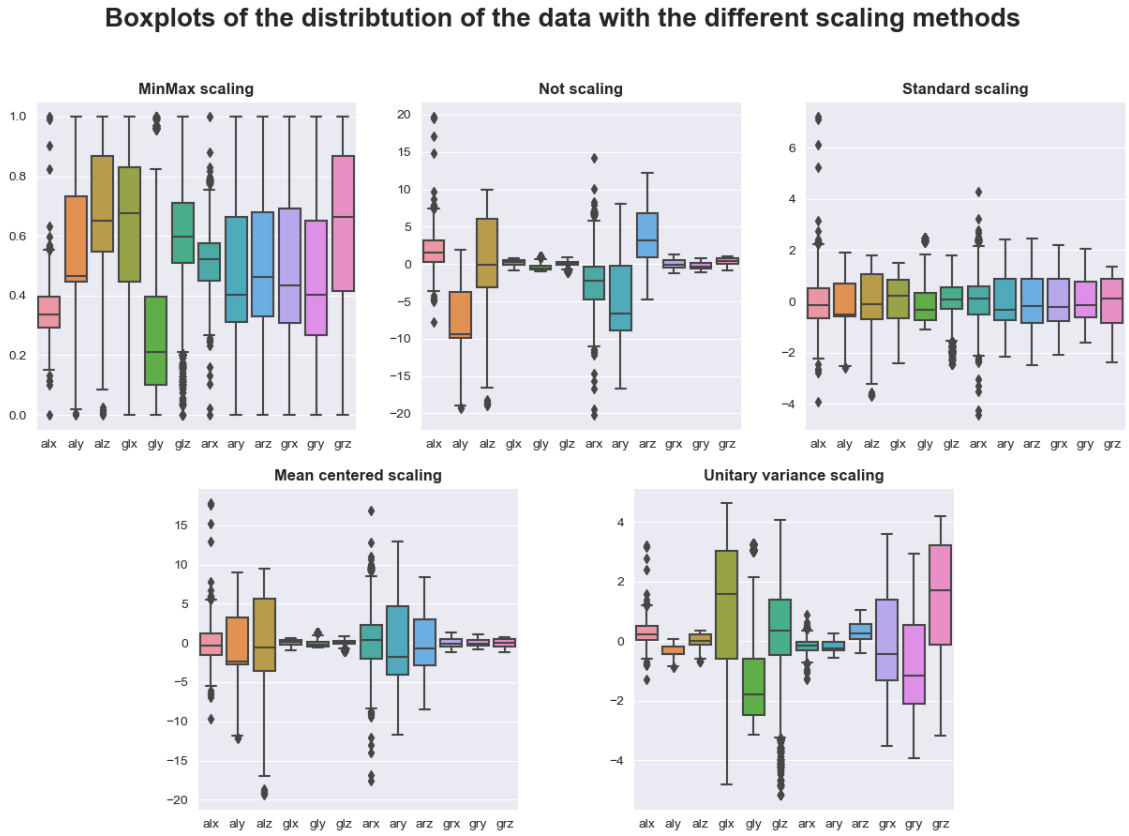
When choosing whether to scale the data and what scaling procedure to use, one should (a) analyze how data distribute, and (b) reflect on the meaning of each feature.

Q1.4. What type(s) of scaling procedures among the ones reported below would be in principle correct for the dataset under analysis and why? Select the answer(s) you consider correct, and explain why you made your choices. [multiple choices possible! 10 points]

Type of scaling (check if correct)	**Reason for your answer**
No scaling (data used as they are)	No scaling shows a wide distribution. The acceleration data is totally different from the gyroscopic data. The range of the acceleration data is much higher than the range of the gyroscopic data, so it is not a good method for scaling on this dataset.
Mean centering (mean equal to 0)	Mean centering scaling still shows a very different distribution between the gyroscopic data and the acceleration data. The range of the acceleration data is much higher than the range of the gyroscopic data, so it is not a good method for scaling on this dataset. The distribution looks similar to no scaling.
Unitary variance (variance equal to 1)	Unitary variance scaling also shows a very different distribution between the gyroscopic data and the acceleration data. The range of the gyroscopic data is much higher than the range of the acceleration data, so it is not a good method for scaling on this dataset.
MinMax scaling (minimum and maximum equal to 0 and 1, respectively) ✓	MinMax scaling shows a similar distribution between all the different variables. The IQR of all the variables is very similar, while the range is also the same. Therefore, for this dataset, it is a good way for scaling the data

Standard scaling (mean equal to 0 and variance equal to 1) ✓	Standard scaling also shows a similar distribution between all variables, with similar IQR's. Unlike MinMax scaling, standard scaling shows more outliers. However, it is still a good way to scale this data

You can report *one* figure below if this helps you explain your reasoning.



Part 2: Principal Component Analysis

In this part of the assignment, you will start performing the steps of a classical PCA analysis.

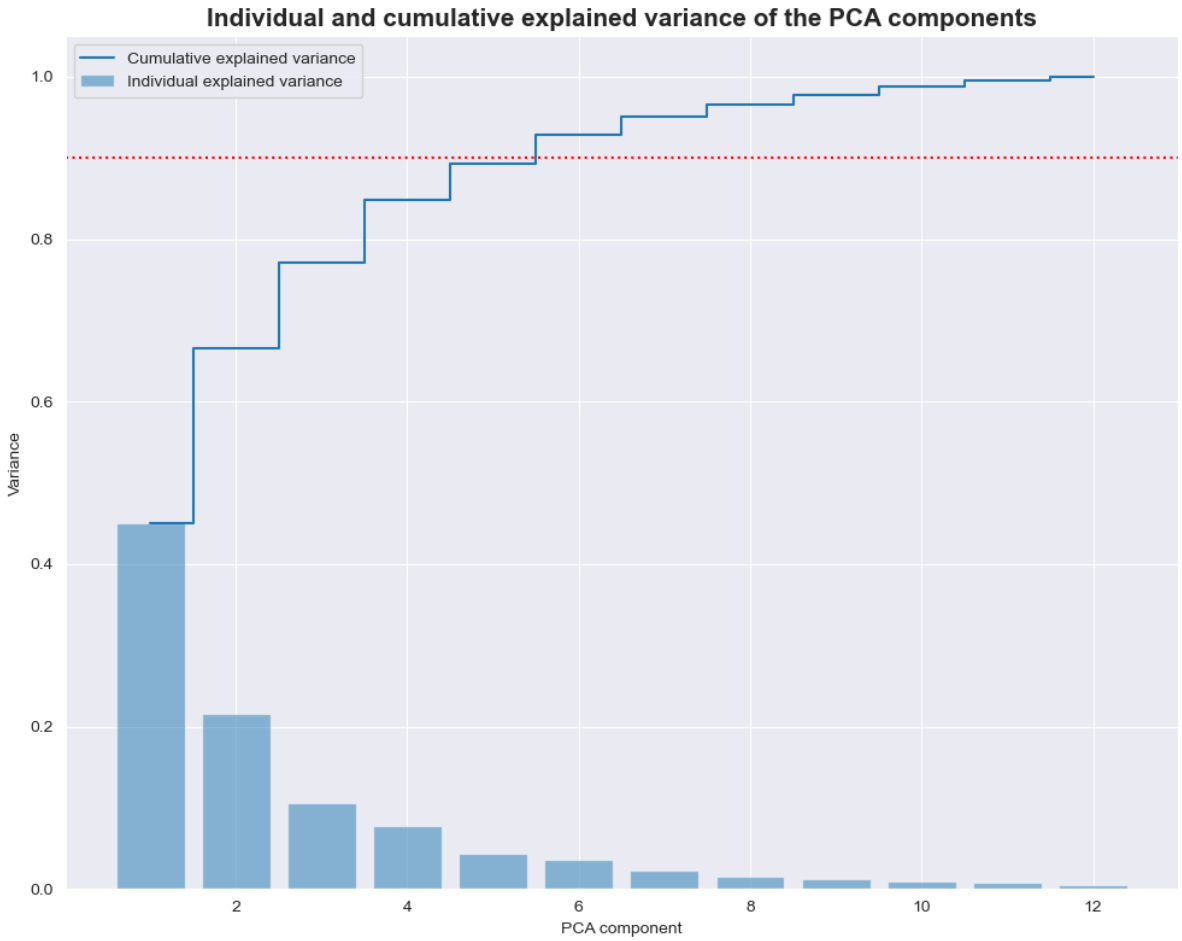
2.1. Selection of the number of principal components

For this exercise, take your pretreated dataset as in Part 1. To allow for comparability of the results across all of you, you will use MinMax scaling.

Q2.1 How many variables are necessary to capture at least 90% of your dataset variance (given the steps explained above)? Insert your answer below: [5 points]

6

Visualize your answers with a figure and report it below.



For the rest of the exercise, we will be using the number of principal components you have reported in your answer. Be mindful: if your components are too many to solve this exercise swiftly, something might have gone wrong in the pipeline ;)

2.2. Analysis of the variables

Q2.2 What variable has the largest effect on PC1? [5 points]

gry

Motivate how you reached this conclusion below (maximum 100 words).
If you compute the highest loadings for the different PC values, you can see which variable has the largest effect on the desired PC values. The variable containing the highest absolute loading has the largest effect on the PC value. The highest absolute loading of PC1 is 0.465804 which corresponds to gry.

Q2.3 What variable has the largest effect on PC2? [5 points]
grx

Motivate how you reached this conclusion below (maximum 100 words).
If you compute the highest loadings for the different PC values, you can see which variable has the largest effect on the desired PC values. The variable containing the highest absolute

loading has the largest effect on the PC value. The highest absolute loading of PC2 is 0.520583 which corresponds to grx.

Q2.4 Is there a variable that is not relevant to compute PC2? [5 points]

- ✓ Yes
- □ No

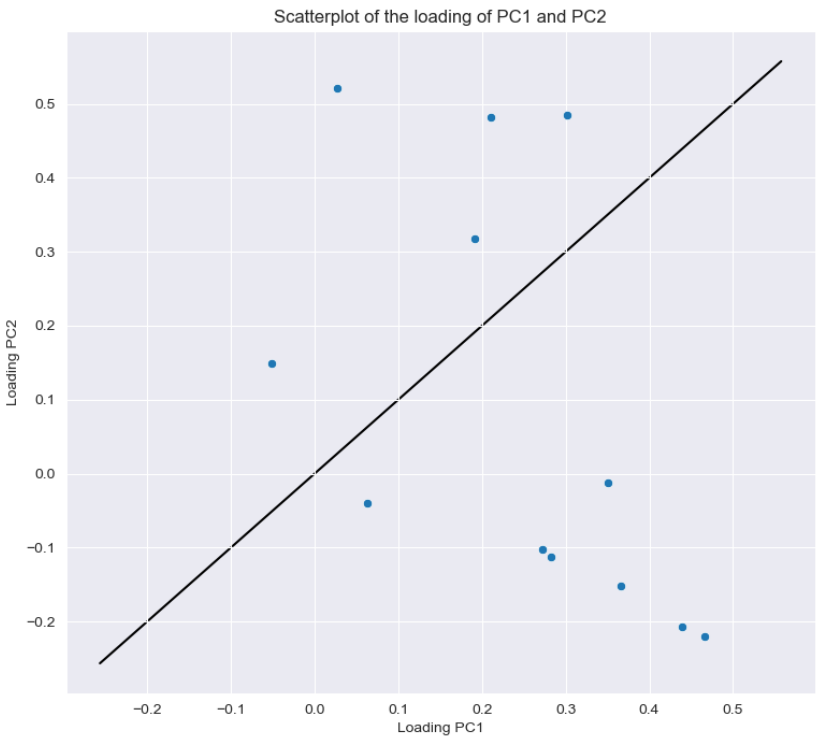
Motivate your answer (max 50 words)

The loading of the acceleration sensor ary is -0.012029. This is a considerable lower amount than the other variables. So this variable is not relevant to compute, because it almost doesn't contribute anything to PC2

Q2.5 Do variables coming from the same sensor contribute similarly to PC1 and PC2? [10 points]

- □ Yes, in all cases
- □ Yes, in some cases
- ✓ No, with few exceptions
- □ No, never

Visualize your answers with a *single* figure and report it below.



2.3. Analysis of the samples

Have a look at the score plots obtained in the first two principal components, and answer to the questions below.

Q.2.6 What does each point in your PCA score plot represent? [5 points]

- □ The measurements of a single sensor across subjects, activities, and timeframes
- □ A single timeframe across subjects
- ✓ A snapshot of a subject in a timeframe, while performing a certain activity
- □ A subject across timeframes, while performing a certain activity
- □ A snapshot of a subject in a timeframe, while performing multiple activities
- □ Other, specify: ___

Q.2.7. Do your samples form somewhat distinct groups in the space of PC1 and PC2? [5 points]

- ✓ Yes
- □ No

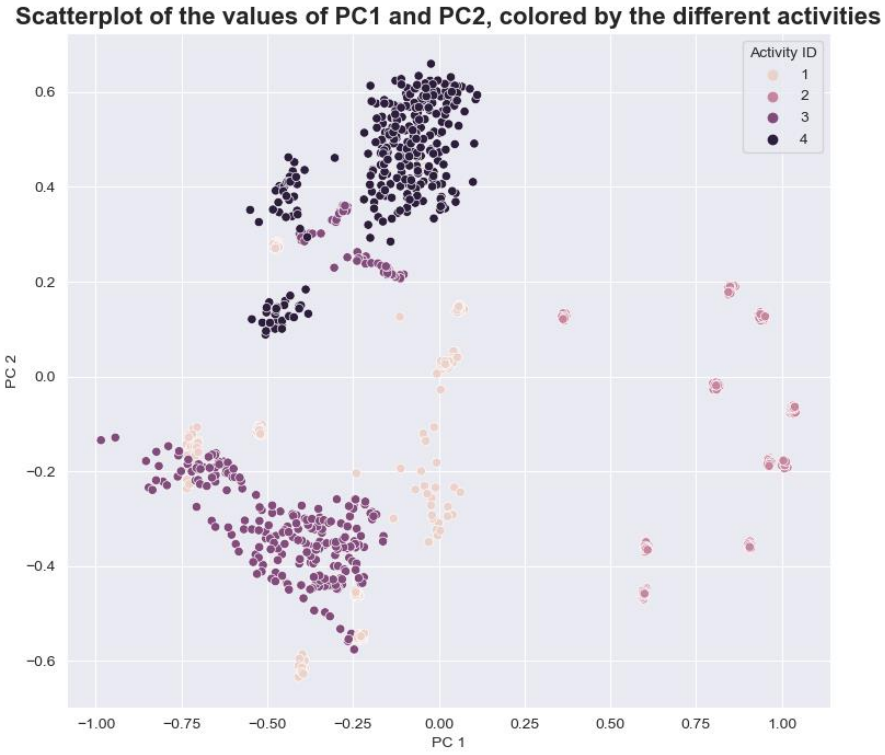
Part 3: Interpretation

Here, you will use the insights you gathered so far, to interpret the data.

Q.3.1. If you have answered *yes* to the previous question (Q.2.7), do these groups correspond to some information you have available, and if so, what information? [10 points] (Hint: consider colouring the points in your score plot using different types of information)

- □ There are no evident groups.
- □ They do not correspond to any information available.
- □ The sample group according to the row ordering in the original dataset.
- □ The groups correspond to the difference between study subjects.
- ✓ The groups correspond to certain activity types.
- □ The groups capture the passing of measuring time.

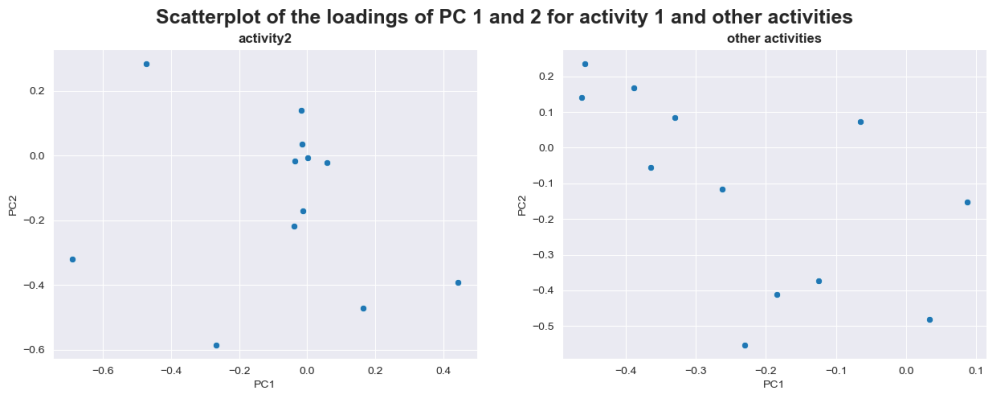
Motivate your answers to **Q.2.9** and **Q.3.1** with a *single* figure and report it below.



Q.3.2. Are there features that clearly separate class 1 from the other ones? [10 points]

- ✓ Yes
- □ No
- □ Impossible to say using only these two components

Write some text and use a figure to explain your reasoning. If you answered 'yes', provide a list of the features you identified to support your reasoning (max 100 words).

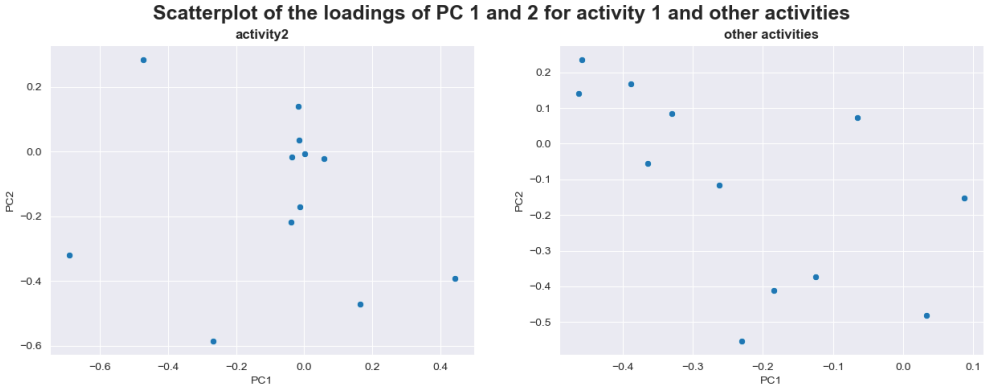


Q.3.3. Are there features that clearly separate class 2 from the other ones? [10 points]

- ✓Yes

- □ No
- □ Impossible to say using only these two components

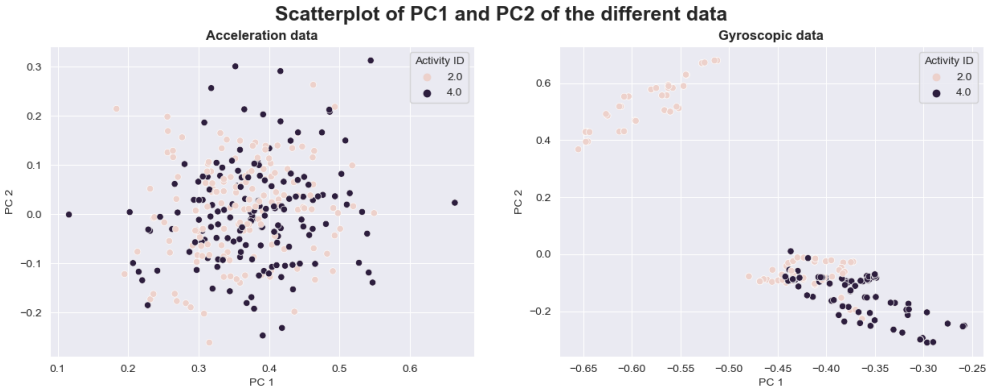
Write some text and use a figure to explain your reasoning. If you answered ‘yes’, provide a list of the features you identified to support your reasoning (max 100 words).



Q.3.4. If one were to develop an approach to distinguish only between Walking (activity no. 2) and Jogging (activity no. 4) only, and had to choose between measuring only acceleration or gyroscopic information, what would they choose? [10 points]

- ✓ gyroscopic data
- □ acceleration data

Write some text and use a figure (optional) to explain your reasoning (max 100 words).
If you separate the acceleration and gyroscopic data after the scaling and perform PCA on both individually, you get an indication of the principal components of the acceleration and



gyroscopic data. If you plot PC 1 and PC 2 of both the acceleration and gyroscopic data, you can see that acceleration information shows you a cloud of points of both the activities . Looking at the gyroscopic data, you can clearly see two different groups. One containing activity 2, and the other activity 4. So to distinguish Walking and Jogging you should measure the gyroscopic data.

If you had to attempt an interpretation of what you just noted with your data analysis, what would it be? (max 300 words) [bonus question, it adds up to 10 extra points if your answer is interesting and well-thought of]

[figure here]