Assignment 2
PCA peer-review

Getting started

This is the template for peer-reviewing your colleagues' work. Engaging in peer reviewing is not only important but also incredibly beneficial for your growth as a programmer and data analyst. By participating in this collaborative process, you'll receive valuable feedback that helps you improve your data analysis skills and coding skills. The purpose of the peer feedback is not to give a quantitative assessment (*i.e.* a numerical grade), but rather an indication (supported by constructive criticism) of the quality of your peers' work. Additionally, as a reviewer, you'll sharpen your critical thinking abilities and gain exposure to different styles, fostering a supportive learning environment that accelerates your progress in advanced programming and biomedical data analysis.

How to use this template:

- Download a local copy of this document for you to fill out.
- It is mandatory to fill out all of the parts of this form. Non-mandatory questions are present, and they will be labeled as such.
- Remove all the blue text and replace it with your answers.
- Be kind and constructive with others. Use this peer review to help them improve.
- Once you're done, convert the file into a pdf before submission.

Peer review results

Peer-reviewer (optional): Martijn Stoorvogel

Topic 1: Data analysis

Aspect 1.1: Formal analysis. The aim of this section is to evaluate whether the logical steps followed to answer the questions are convincing.
Use the dedicated column ("1. Formal analysis") to provide comments on the following aspects:

- Methodology. Evaluate the soundness and appropriateness of the analytical methods and insights used to answer the question. Consider both the textual and graphical elements of the response.
- Assess whether the answer is comprehensible and clearly stated.

- Assess whether the provided answer is overall convincing.

Aspect 1.2: Clarity of the figures. The aim of this part is to evaluate how clearly the figures can convey relevant information.
Use the dedicated column ("2. Quality of the figures") to comment on:

- *Figure formatting*: The figure should use appropriate labeling, titles, and legends to guide the reader and provide context. Colors should be easily distinguishable and font sizes should be appropriate for interpreting the content.
- *Clarity and Information*: A high-quality figure should be clear and easy to interpret. The information presented should be concise, well-organized, and effectively convey the intended message.
- *Visual Presentation*: The visual presentation of a figure plays a crucial role in its quality. A good figure should have aesthetically pleasing design elements, including appropriate colors, fonts, and line styles. The size and scaling of the elements within the figure should be visually appealing and proportional, avoiding overcrowding or excessive whitespace.

You can write your answers on a question basis below:

Original question	**1. Formal analysis**	**2. Quality of the figures**
Q1.1. Are there any highly correlated variables (i.e., correlation larger than 0.90)? Yes/no Visualize this answer with a single figure and report it below.	Clear function to find the correlation between variables. However, the incorrect variables such as subject ID, acitivity ID and timestamp are included in the correlation plot.	Figure has clear labels and title (only advise is to use capitals in title name). However, the coloring could be a bit more clear by having negative values a different color than positive values. The 0 point could be white and the other colors. Also scale from -1 to 1 would be more clear.
Q1.2. Are there any missing data? □ Yes □ No	Yes, however the method could be more explicit by checking whether each patient is in the data. Each timeframe and each activity per patient.	Not applicable.
Q1.3. How many duplicated rows are present in your dataset, if any? Specify the number below.	Yes, i did exactly the same	Not applicable.

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	I agree with the answer, however in the code I only see that the boxplots for minmax scaling are analyzed. I would also have investigated the others and shown the results for those to comment on. In the code I cannot find the code to create the plot	Figure is clear for its purpose and well labeled. It is clear from the explanation and the plot that minmax or standard scaling is the most appropriate method. The figure supports the explanation which is what it should do.
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	Clear method and to the point. Correct answer is found as far as I know.	Figure is clear and well commented. Only the cumulative figure would have been sufficient and maybe more clear in this case.
Q2.2 What variable has the largest effect on PC1?	Clear method, same as mine and to the point just find the maximum loading. However the method just gives the table rather than returning the max value.	Not applicable.
Q2.3 What variable has the largest effect on PC2?	Clear method, same as mine and to the point just find the maximum loading. However the method just gives the table rather than returning the max value.	Not applicable.
Q2.4 Is there a variable that is not relevant to compute PC2?	The table was printed with all loadings but not a line of code to display those with the lowest loading.	Not applicable.
Q2.5 Do variables coming from the same sensor contribute similarly to PC1 and PC2?	I interpreted the question differently and as there is no explanation provided but just the figure I cannot tell what this person means?	Figure is well commented and labelled.
Q.2.6 What does each point in your PCA score plot represent?	Yes correct answer!	Not applicable.

Q.2.7. Do your samples form somewhat distinct groups in the space of PC1 and PC2?	Yes, as is clear from the figure as well	Not applicable.
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?	PCA is able to use a reduced amount of variables to capture the groups of the various activity IDs. Actually in this case from the 12 available parameters it was reduced to just 2 using PCA which already is able to distinguish the 4 groups from each other.	PCA is able to use a reduced amount of variables to capture the groups of the various activity IDs. Actually in this case from the 12 available parameters it was reduced to just 2 using PCA which already is able to distinguish the 4 groups from each other. I would advise the use colors not as closely related as it is difficult to see the difference in this case.
Q.3.2. Are there features that clearly separate class 1 from the other ones?	Accidentily the same plot as in the next question is displayed and there is no further explanation so I must disagree with this one.	Figure formatting and labelling is correct.
Q.3.3. Are there features that clearly separate class 2 from the other ones?	I agree with the answer completely. As we can see the loadings from class 2 are very different as the ones from the other group which shows that there is a clear difference between those groups.	Figure formatting and labelling is correct. However, it would have been usefull to highlight the class of interest in one color and all other classes in a different color
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?	The answer seems correct. Explanation is well thought of in this case and explains the variance for both sensors and why the one is better than the other. We can see that the figure with acceleration is unable to find clear groups which explains that acceleration is not a good sensor type to analyse this.	Figure was well formatted and labelled. Just the colors could have been a bit more distinctive to be more clear.
Q.3.5. If you had to attempt an interpretation of what you just noted with your data analysis, what would it be? (max 300 words)	Not provided.	Not applicable.

Topic 2: Code.

Aspect 2.1. **Code clarity**. Code clarity can be judged by assessing how easily the code can be understood and interpreted by other developers. Clear code should have clear and meaningful variable and function names, concise and well-organized code structure, appropriate comments, and adherence to coding conventions and best practices.
(Score each aspect from 0 to 10, with 10 being the best).

Aspect 2.2: Code organization. Have a look at the code submitted by your colleague as part of the assignment, based on the aspects below. Use the occasion to provide feedback on each of the following points.
(Score each aspect from 0 to 10, with 10 being the best)

Like you did for assignment 1, reflect on the points mentioned below.

How much does the code follow aspects of good programming style (e.g., but not limited to PEP8 style)?	• Greatly! Good job. 😀 • Most times. 😊 • Almost never. 😕 • Never. 😫
You can comment on specific style aspects, like the ones below. If you're unsure about the score, look at the PEP8 style guideline (as an example of style and usage).	
• Correct indentation.	Score 8
• Naming conventions (clarity, consistency, etc.)	Score 8
Use this space to provide comments on the scores assigned above.	Functions were used for each step/question and were well commented on.
As you know, code documentation is essential as it provides valuable insights into the purpose, functionality, and usage of code, making it easier for other developers to understand, maintain, and collaborate on the codebase. In this section, you will have the opportunity to help your peers improve their documentation. Have a look at this link if you're struggling with your review.	

• *Variable, functions, and Constant Documentation*: Document the purpose and usage of variables, functions and constants, especially if their names alone are not self-explanatory.	Score 8
• *Code Comments*: Include comments to explain complex logic, important decisions, or any non-obvious code segments. Comments should focus on the why rather than the how, as the latter should ideally be expressed through clear code.	Score 8
Use this space to provide comments on the scores assigned above.	The code is just well commented on and reading the code is easy to understand
• *Modularity*: Assess how well the code is organized into logical modules or functions, with clear responsibilities and separation of concerns.	Score 8
• *Code Duplication*: Identify any redundant or duplicated code blocks.	Score 8
Topic 3: Other reflections. Use this opportunity to reflect on the data analysis and solutions you have seen and compare it with your own work. Provide your comments below.	
How different was the adopted solution from yours? Please provide a comment.	The plot of the gyroscopic data pca against the acceleration data pca was way more clear to convince that the gyroscopic data should be used as it was clear that only with the gyroscopic data we could still find groups.
What are your main recommendations for improving the analysis workflow?	Explain the figures more as sometimes the figure needs to be self-explanatory which it is not and it is not clear whether you understood the question

What are your main recommendations for improving the figures?	Think more about color usage in the figures (correlation, boxplots) to make the figures standout more and convey the message more clear
Is there something that you particularly liked about the provided work? If so, use the occasion to inform your peer about it!	The usage of functions to break down the code into multiple parts was well done!
(Optional) Use this space to write additional comments you might have that we forgot to ask.	Write your comments here.

Document creator: Francesca Grisoni, f.grisoni@tue.nl. Reach out via Slack in case of questions.