Module D3 - Project - Version C

General information:

1. The project comprises tasks for both parts of the module. You have to write a short report.
2. For the report you have to fill out a markdown (*.Rmd*) file in R and convert it into a .html file. In order to use markdown install the package "rmarkdown" (see:rmarkdown). Use the function "knit to html" to produce the final file.
3. The report consists of a short introduction, the codes and the results for the tasks and a short interpretation/discussion for each task. For further explanations see "evaluation criteria" at the end of this document.
4. Please upload the report as .html-file on Moodle. **Deadline is December 19, 23:59**.
5. The report file is named such that the project, the version and the author are contained in the name. E.g., proj_vC_ThomasOtt.html

In the project you will work with the following data (as appendix in the email):

- *Wineparameters.csv*: This file contains the measurements of several parameters from 68 bottles of wine. The wine ID is unique for each bottle.
- *Winespect.csv*: This file contains infrared (IR) spectra of the wines. The wine ID corresponds to the ones in *Wineparameters.csv*. The spectrometer's wavelength range is from 1 to just over 4000 cm^{-1}. Within this range, there are contributions from how the IR light interacts with many different chemical compounds. IR light is absorbed differently depending on the masses of the molecules and the types and strengths of the bonds between them. Some of the recognized chemical compounds are important for identifying different types of wines. Others provide only irrelevant, or even distracting information and has to be filtered. Here, the range from 600 to 2000 cm^{-1} has been selected. This frequency range contains 1280 measurements.

You shall first perform an exploratory data analysis for a selection of wines. Then you will investigate how well certain wine parameters can be predicted from the spectra or other parameters using linear regression. The list below will guide you through the steps to be taken.

General tasks

- **Task 1:** Fill out your name etc. and write a short abstract summarising the overall purpose of the study, the basic methods used and the major findings. This task may be completed at the end.

- **Task 2:** For all the tasks, consider the wine spectra of the grape varieties *Blauburgunder, Merlot 2009, Garanoir_Gamaret, Pinot noir, Zweigelt, Chardonay, Rheinriesling, Sauvignon Blanc*. Store the selection in the variables *selection_spectra* and *selection_parameters*.

Tasks about Part 1 (Exploratory Data Analysis)

- **Task 3:** Illustrate the spectra as well as possible with a suitable plot of the basic plots described in chapter 2 of the lecture notes (Part 1). Colour the red wines and wines differently such that they can potentially be identified in the plot. Add a legend. Give a short description/interpretation.

- **Task 4:** Perform a MDS and plot the data in 2D. Colour the data points such that all the red wines and all the white wines have the same colour. On top, label the data points with the name of the variety (e.g. Blauburgunder).Give a short description/interpretation.

- **Task 5:** Perform a PCA and check how much variance is explained by the first 8 principal components. Illustrate your results in a plot and report/comment on the results, i.e. give an interpretation/conclusion.

- **Task 6:** For tasks 6-8, you have to use as many principal components as needed to explain at least 97% of the variance. Perform a Ward's clustering. Label the dendrogram using the wine varieties. In order to assess the quality of the Ward's clustering, calculate a confusion matrix, where the actual classes correspond to the wine grape varieties (choose as many clusters as varieties). Give an interpretation of the results.

- **Task 7:** In order to further assess the quality of Ward's clustering, calculate a confusion matrix, where you distinguish between red and white wines with two clusters. Calculate the Rand index (Rand index). Give an interpretation of it and discuss the quality of the clustering (here we expect you to google for finding a way to calculate this).

- **Task 8:** Calculate the Rand index for the results in Task 6. Compare Ward's method and single linkage by calculating the Rand index.

Tasks about Part 2 (Regression)

- **Task 9:** For now, let us focus only on the relationship between the first principal component pc1 (see Task 4) as explanatory variable for the dependent variable alcohol volume in the wine. Investigate graphically the nature of their relationship by plotting the data and overlying both a linear regression line and a loess non-parametric regression line. Discuss whether the linear regression describes the data well enough.
- **Task 10:** Fit a linear regression model with outcome alcohol volume and explanatory variable pc1. Plot the residuals against the fitted values. What do you observe? Is there a connection to Task 9?
- **Task 11:** Fit a multiple regression model with outcome alcohol volume and explanatory variables pc1, pc2, Glucose and pH without any interactions or non-linear effects. Assuming that there are not influential observations or outliers and that the model is correctly specified, assess the model assumptions. If some assumptions are violated, briefly describe how you would adress these issues (no analysis required!).
- **Task 12:** Assuming that regression diagnostic did not flag any issue for the multiple regression model, are all estimated parameters significant? Interpret the significant parameters (explain in your own words).
- **Task 13:** Could you remove non-significant parameters from the multiple regression model via a model selection approach? How? Why do you think that non-significant parameters are not necessary?
- **Task 14:** Fit a linear regression model with outcome alcohol volume and a main effect for pc2, as well as main effects and interactions for pc1 and color (coded as a factor). Extract the coefficients and translate them into one regression equation for each color (see an example of what is expected on the beginning of page 53 in the script, Part 2).

Evaluation Criteria

The reported will be evaluated and graded according to the following criteria:

Report overall (failed if not)	Format correct? Submitted in time?
Work overall (failed if not)	Is the work original (obviously produced by author, no plagiarism)?
Abstract (0-4 points)	Does it describe the problem/methods/results adequately? Language adequate and correct?
Tasks 1-14 (0-2 points per task)	Code, results and interpretation (correct and concise)?
Submission (0 or 1 point)	Is the file correctly named?