University of Applied Sciences

HOCHSCHULE
EMDEN·LEER

Information on the assignment

Prof. Dr. Joachim Schwarz

General framework

- A data set is provided for the homework. This is to be analyzed using the statistical software package R on the basis of predefined tasks.
- The wine dataset contains chemical and sensory information on a total of 6497 Portuguese red and white wines. A description of the variables can be found on the following pages.
- Guideline for the length of the entire term paper: 15-20 pages (incl. graphics) excluding title page and appendices or indexes. Please note: This is a guideline. If the term paper is longer or shorter, this does not automatically lead to a fail. The decisive factor for the assignment is whether the content is complete and correct.
- A **maximum of two people** can write the assignment together and will be graded together. Write your names on the title page!
- **Deadline: 30th June 2025, 23:59.**

General framework

- As analyses are to be carried out in R, the evaluation script must be included in the appendix. If the evaluation script is missing or not executable, the term paper will be graded as “failed”.
- If the R script is not executable for individual tasks, or if the results in the term paper deviate from the results of R, the corresponding task is graded as “failed” (see the following grading scheme).
- Some of the tasks also include research on the Internet or in literature sources. The sources used must be referred to according to the usual rules for academic papers, i.e. they must be cited in the text and a literature section must be provided in the appendix. The only exception: The lecture notes and the R scripts developed in the lecture do not need to be cited. If the references are missing, the corresponding parts of the paper will be graded as failed. This may result in a fail for the whole assignment.

General framework

- No screenshots or copied output without further preparation in the text! Analysis results must be prepared appropriately; failure to do so will result in a deduction from the grade. Screenshots can be included in the appendix. Graphics produced in R, on the other hand, can be included in the text.
- Follow the formal requirements for writing academic papers (Guidelines for Writing Academic Papers), with the following exceptions: Table of contents, list of figures and list of tables are not required. The example structures presented in the guidelines are also not relevant here.
- The sources and other tools used and the statuary declaration mentioned in the Guidelines for Writing Academic Papers must be included in the term paper. See also the examples for documenting the use of AI tools on the following pages. For the statuary declaration, see also the following link: https://www.hs-emden-leer.de/en/university-of-applied-sciences/organization/departments-a-z/admission-and-examination-office/forms-downloads.

Data description for the wine data set (1/2)

Variable	Description
X	Row number, not used, can be ignored.
fixed acidity (g/l)	Most acids involved with wine or fixed or nonvolatile (do not evaporate readily).
volatile acidity (g/l)	The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.
citric acid (g/l)	Found in small quantities, citric acid can add 'freshness' and flavor to wines.
residual sugar (g/l)	The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.
chlorides (g/l)	The amount of salt in the wine.
free sulfur dioxide (mg/l)	The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.
total sulfur dioxide (mg/l)	Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.

Data description for the wine data set (2/2)

Variable	Description
density (g/ml)	The density of water is close to that of water depending on the percent alcohol and sugar content.
pH	Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.
sulphates (g/l)	A wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.
alcohol (% vol)	Does not need any explanations ☺.
quality	Score between 0 and 10. The higher the quality, the higher the score.
variety	A factor with two levels: red, white.

Guidelines for the use of AI tools (large language models, etc.)

Topic	Activity	Permission	Citation
Generate ideas	Brainstorming, structural ideas, first drafts, optimization of research questions	Yes	As a tool
Literature research	Research and initial references to literature and sources	Yes	As a tool
Contents	Summary of the current state of research	No	
Contents	Generate texts	No	
Contents	Answering research questions	No	
Contents	Revise and optimize texts linguistically	Yes	As a tool
Meda, graphics	Generate graphics, presentations videos	Yes	As a tool
Data	Data analyses (observe data protection when uploading data!)	Only for data protection compliance!	As a tool, compliance with data protection laws must be stated!
Data and results	Interpretation of analysis results	No	
Academic work	Tools for the organization of scientific work (e.g. translation, literature management)	Yes	As a tool
Source code	Revision, new implementation	Yes	Must be specified in the code header. Prompts must be specified.
Source code	Troubleshooting and debugging	Yes	Must be specified in the code header. Prompts must be specified.

Examples for documenting the use of AI tools

Used tool	Type of use	Affected parts of the work	Remarks
DeepL Translator	Translation of text passages from your own text	Whole paper or thesis	
ChatGPT (OpenAI)	Preparation of text suggestions	Chapter 1, p. 3, section 2	Has been marked in the text or in footnotes.
ChatGPT (OpenAI)	Asking ChatGPT on the topic of the work, comparison of the results with own research	Chapter 2, p. 5-7	Further information and complete chat history as screenshots in the appendix.
ChatGPT (OpenAI)	Use of ChatGPT to create the structure	Structure, table of contents	Draft outline was revised and supplemented.
Typeset.io / SCISPACE	Literature research	Whole paper or thesis	
DeepL Write	Rephrasing of text passages	Whole paper or thesis	
Midjourney	Creation of visualizations	Fig. 4, p. 8 and fig. 6, p. 13	Fig. 6, p. 13 has been significantly revised.
ChatGPT (OpenAI)	Creation of codes for data analysis	Chapter 5	Created codes were tested and modified if necessary. The entire code is included in the appendix.
ChatGPT (OpenAI)	Carrying out data analyses	Chapter 5	Further information and complete chat history as screenshots in the appendix. Personal information was removed from the data in advance.

Grading notes

The following criteria are used to grade the term paper:

- Correctness of content and methodology in description, execution and interpretation.
- Consistently scientific, concise formulation.
- Reproducibility of the results from the given data using the executable R script, comments on understanding if necessary.
- Reference to appropriate textbook or internet sources for the method (if no method from the lecture has been used).

Note:

Check the reproducibility by running your R code again with the data provided: Do you get your documented results again?

Grading scheme

- The grading scheme serves as a guide; deviations may be made in individual cases. The deviation is then justified.
- If everything is fulfilled, the grade is 1.0. The grade is reduced by 0.3 for each partially fulfilled item and by 1 for each unfulfilled item.

Evaluation	Fulfilled	Partially fulfilled	Unful-filled
Task 1			
Task 2			
Task 3			
Task 4			
Task 5			
Task 6			
Elegant / skillful use of R			
Formal aspects and literature			

Tasks

1. a) Read in the wine data set with e.g. read.csv(“yourpath/wine.csv”). Display the distribution parameters (mean, standard deviation, minimum, lower quartile, median, upper quartile, maximum) for all metric variables in tabular form. Show the frequency distributions for all categorical variables. Please also include missing values for all variables.

   b) Create suitable graphics for all variables. Also calculate the skewness coefficient, if possible (can be included in the table generated above). Use the graphics and the key figure to assess: Are there any outliers? Is the respective distribution left-skewed, right-skewed or symmetrical? Note: All graphics can be included in the appendix; a summary of the results is sufficient for the text.
2. Research how the t test can be carried out in R and answer the question whether the alcohol content of red and white wines differs significantly. Also check the application requirements of the t test!
3. For this task, consider only the red wines. Investigate whether the quality of the wine depends on its chemical and sensory properties. To do this, carry out a linear regression for all chemical and sensory variables. Take quality as the target variable. Also check the regression requirements. Dealing with violated regression requirements, if any, is not necessary, documentation is sufficient.

Tasks

4. Really good wines have a quality of 8 or more, while bad wines have a quality of 4 or less. Use a suitable method to determine whether a wine is good or bad based on its chemical and sensory properties.
5. In data mining, it is common practice to divide data sets into a training data set and a validation data set. The model is developed on the training data set, while the validation data set is used to determine the model quality indicators.

   Analyze whether the color of the wine (red or white) can be determined based on its chemical and sensory properties. To do this, divide the data set into a training data set and a validation data set. Develop a model for predicting the color of the wine based on the training data set. Then use the validation data set to first predict the color of the wine. Finally, check the model quality with the confusion matrix and the corresponding quality figures and the AUC value.

   Note: The logistic regression requires a 0/1 target value. It may be necessary to transform the wine variety (red or white) into 0 and 1.
6. Investigate whether the chemical and sensory properties of the wines can be condensed into a few factors.

   Note: It may be necessary to reduce the number of variables until the correlation matrix is suitable for a factor analysis. The MSA can be used for this.