This document contains questions to help you reflect about the operations applied to the data during this assignment. You have to fill it in and submit it on Studium (one sheet per group).

Group number and group members:

TASK 1: reading the data

What data type have you assigned to attribute *id?*

What do you think is the practical consequence of setting this data type?

What are the average length of sepals (sl) and their standard deviation?

TASK 2: database preprocessing

How many instances are there for each class?

Virginica

Setosa

Versicolor

TASK 3: data cleaning

Why is it important to let the system know which values are missing?

What are the average length of sepals (sl) and their standard deviation after declaring missing values (3.1)?

What are the average length of sepals (sl) and their standard deviation after removing outliers (3.2)?

Do you think the outliers you have removed were noise (that is, wrong measurements) or unusual but correct observations?

Would you first handle missing data and then remove outliers, or the other way round? Why?

Assume your observations (records) represent people in a social network, and one variable stores their degree centrality. Would you remove outliers in this case? why?

TASK 4: data transformation

What are the average length and standard deviation of sepals after min-max normalization?

What are the average length and standard deviation of sepals after standardization?

How many components have been selected after 4.3?

How much variance is captured by the first two components?

How is the first component defined as a combination of the original attributes?

How many components would have been selected after 4.4 (that is, with an attribute expressed on a larger range)?

How many components would have been selected after 4.5 (that is, with an outlier)?

TASK 5:

	Simple sampling	Bootstrapping	Stratified (5.3)	Stratified (5.4)
Number of iris versicolor				
Number of iris setosa				

Number of iris virginica				
Are there repeated identifiers?				
Does the number of iris versicolor included in the sample change if you change the local random seed?				