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:

Group Number 51

Group Members: Jonas Blum, Xin Tian, Sanjatul Islam, Abdulfattah Morad, Riccardo Rebecchi

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?

Average length: -5.7055
Standard Deviation: 303.7889

TASK 2: database preprocessing

How many instances are there for each class?

Virginica

3000

Setosa

3000

Versicolor

500

TASK 3: data cleaning

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

If the system doesn't know which values are missing,
the missing values are interpreted as valid values which will most likely lead to statistical bias.

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

Average: 3.52759
Standard deviation: 2.10249

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

Average: 3.5206
Standard deviation: 2.01850

Code used:

```
numerical_features = data[['pl', 'pw', 'sl', 'sw']]
threshold = 3
z_scores = np.abs(stats.zscore(numerical_features))
data = data[(z_scores < threshold).all(axis=1)]
```

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

Since we used a threshold of 3, the removed data points are extreme outliers so they are probably wrongly measured
One of the removed points had a sepal length of 51.0 which is extremely unlikely

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

First remove the missing data because the outliers need proper average and standard deviation values in order to calculate the z-score and remove them if their z-score is too high.
If we didn't remove the missing data first, the average and std values would be wrong.

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?

In that case we would not remove outliers because there could be certain people which have many more connections than others -> for example influencers which would then be removed even though they are valid people.

TASK 4: data transformation

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

Average sepal length: 0.05433
Standard deviation: 0.041879

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

Average sepal length: 2.1007731023396865e-16
Standard deviation: 1

How many components have been selected after 4.3?

To retain at least 95% of the variance we need the first retain the first two principal components:
0.91329008 + 0.04017737 = 0.95346745 = 95.34%

How much variance is captured by the first two components?

95.34%

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

First principal component as a combination of the original attributes:
[0.33847865 -0.0766534   0.86707949 0.35739281]

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

Only 1 as the first principal component already explains 99% of the variance

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

Only 1 as the first principal component already explains 99% of the variance

TASK 5:

	Simple sampling	Bootstrapping	Stratified (5.3)	Stratified (5.4)
Number of iris versicolor	21	12	249	50
Number of iris setosa	64	69	1496	50

Number of iris virginica	65	69	1498	50
Are there repeated identifiers?	No	Yes	No	No
Does the number of iris versicolor included in the sample change if you change the local random seed?	Yes	Yes	No	No