Submission sheet
Assignment III

TASK 1: k-Means

Corresponds (more or less) to the three expected species? ☐YES ☐NO No as there is only 1 sample in one cluster

Number of records in each cluster: 1) 193 2) 106 3) 1

TASK 2: preprocessing

Is it better to rescale before or after detecting and filtering out the outliers? Rescale after filtering out outliers

Corresponds (more or less) to the three expected species? ☐YES ☐ NO -> Yes

Number of records in each cluster: 1) 118 2) 95 3) 57

Coordinates of the three centroids:

	PW	PL	SW	SL
1)	3.424	4.987	0.297	1.466
2)	2.878	6.388	1.914	5.19
3)	2.688	5.705	1.312	4.082

TASK 3: choice of k

Which K corresponds to the best clustering? (using the Davies-Boulding index).
k=2 -> score: 0.353

TASK 4: Hierarchical clustering

Using SingleLink, how many records are included in each of the two top clusters?

Cluster 1: 175

Cluster 2: 95

Which approaches produce a (more or less) correct clustering corresponding to the three species, if any?

SingleLink: No -> since individual points get added to the big clusters at later stages

CompleteLink: Yes, there are 3 distinct clusters which are merged in the last 2 steps

AverageLink: Yes, there are 3 distinct clusters which are merged in the last 2 steps

TASK 5: DB-Scan

How many clusters does DB-SCAN find with eps=1, min_samples=5? 2

Can you give a value for epsilon leading to two clusters (plus noise)? 0.7

K-DISTANCES:

Which K did you use? 5

According to the k-distances plot, what value(s) of epsilon would you consider as a parameter to DB-Scan and why?

I would consider e=0.25 as a parameter to DB scan, since after wards there is a steep incline i the plot