Assignment III Submission

Task 1: k-Means

- **Corresponds (more or less) to the three expected species?** Two species are "merged" and one outlier gets its own cluster, but i would say more or less it's accurate

Number of records in each cluster:

Cluster	Records
1	193
2	1
3	106

Task 2: Preprocessing

- **Is it better to rescale before or after detecting and filtering out the outliers?**
  For chosen method it was better to remove outliers before rescaling. Otherwise, you don't get your full "range of movement" since the outliers shift the rescaling.
  When using LocalOutlierFactor we found that normalization before removing outlier worked better as we got a more reasonable distance between Max, Min values and the Mean value in two of the attributes compared to removing outliers and then normalizing. The disadvantage of doing this is that the data then have to be rescaled after removing the outliers.
- **Corresponds (more or less) to the three expected species? YES/NO**
  Yes

Number of records in each cluster:

Cluster	Records
1	97
2	105
3	95

Coordinates of the three centroids:

Cluster	PW	PL	SW	SL
1	0.64291899	0.75259259	0.20938578	0.11684982
2	0.83568011	0.65052632	0.78443936	0.8048583
3	0.77212478	0.59908362	0.62109667	0.5222046

Task 3: Choice of K

- **Which K corresponds to the best clustering? (using the Davies-Bouldin index).**
  2

Task 4: Hierarchical Clustering

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

Cluster sizes:

Cluster	Records
1	192

Cluster	Records
2	105

- **Which approaches producing (more or less) correct clustering corresponding to the three species, if any?**

Method	Correct?
SingleLink	No
CompleteLink	Only 2 clusters. Setting clusters = 3 gives correct clusters and the dendrogram has wide separation into 3 clusters
AverageLink	Only 2 clusters. Setting clusters = 3 gives correct clusters but the dendrogram has shallow separation for 3 clusters

Task 5: DB-Scan

- **How many clusters does DB-SCAN find with eps=1, min_samples=5?**
  1
- **Can you give a value for epsilon leading to two clusters (plus noise)?**
  0.25

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?**
  0.6 - 0.8 because it is curving the most there