Post-Block Assignment 1: Option 1

Francois van Zyl 18620426

21 December, 2020

1.1 Does Clustering Make Sense?

To check whether this data set contains any underlying clusters, as opposed to simply being random points, I will consider two different measures. One of the most well-known measures of testing whether a data set contains underlying clusters is the Hopkins statistic. This statistic measures the probability that a data set was generated from a uniform data distribution, implying that no meaningful clusters exist within this data set. Quoting the provided textbook by Alboukadel Kassambara, it is calculated as the mean nearest neighbour distance in a random data set, divided by the sum of the mean nearest neighbour distances of our provided dataset and that of the randomly generated data set. The formula is displayed below, and a Hopkins statistic of approximately 0.5 means that the user-provided and randomly generated data sets are close together, implying that the provided data set is close to uniformly distributed and does not have any meaningful clusters. If the Hopkins statistic is close to zero, the null hypothesis that the provided data set is uniformly distributed can be rejected, and it is possible to conclude that the data set is clusterable.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Before applying this test to the provided data set, I first standardized the numeric variables within the data set to ensure that the features are comparable such that no feature will skew any future distance measures. To this end, I applied z-score normalization to the data set. This entailed subtracting the respective features' means and dividing these features by their respective standard deviations.

$$x' = \frac{x_i - \mu(x)}{\sigma(x)}$$

A summary of the scaled data used for the two clustering tendency tests are displayed below. Note that the feature Gender is removed from the data set, as I implemented the euclidean distance measure which requires numeric data.

Table 1: Descriptive Statistics [z-score normalized data]

Length	Diameter	Height	Whole.weight
Min. :-3.7387	Min. :-3.5558	Min. :-3.33555	Min. :-1.68589
1st Qu.:-0.6161	1st Qu.:-0.5832	1st Qu.:-0.58614	1st Qu.:-0.78966
Median : 0.1749	Median : 0.1725	Median : 0.01156	Median :-0.05963
Mean : 0.0000	Mean : 0.0000	Mean : 0.00000	Mean : 0.00000
3rd Qu.: 0.7578	3rd Qu.: 0.7267	3rd Qu.: 0.60926	3rd Qu.: 0.66123
Max. : 2.4232	Max. : 2.4397	Max. :23.68045	Max. : 4.07178

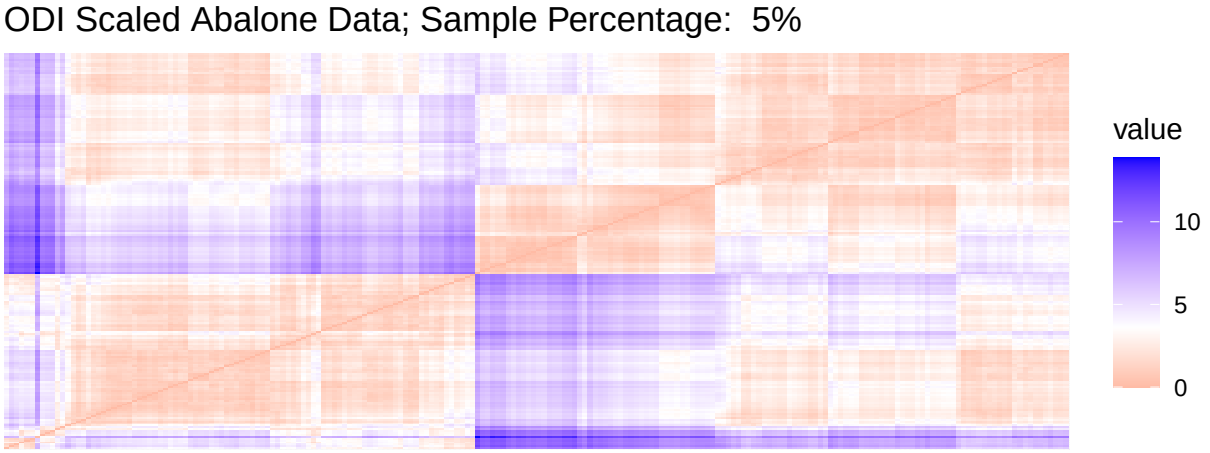
Shucked.weight	Viscera.weight	Shell.weight	Rings
Min. :-1.6145	Min. :-1.64298	Min. :-1.7049	Min. :-2.7708
1st Qu.:-0.7811	1st Qu.:-0.79455	1st Qu.:-0.7818	1st Qu.:-0.5997
Median :-0.1053	Median :-0.08752	Median :-0.0347	Median :-0.2896
Mean : 0.0000	Mean : 0.00000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 0.6426	3rd Qu.: 0.66056	3rd Qu.: 0.6478	3rd Qu.: 0.3307
Max. : 5.0848	Max. : 5.28587	Max. : 5.5040	Max. : 5.9136

To calculate the Hopkins statistic, I implemented a random sample size of 5% of the entries within the data set. The calculated Hopkins statistic on the standardized abalone data is displayed below, and when we recall that a value close to zero implies that the null hypothesis of the data being uniformly distributed can be rejected, it is possible to conclude that the data set is clusterable.

Table 3: Calculated Hopkins Statistic; Sample Percentage: 5%

H
0.0501788

After inspecting a statistical method of assessing the clustering tendency, I proceeded to investigate a visual method, namely the Visual Assessment of Cluster Tendency Algorithm (VAT). This algorithm computes the dissimilarity matrix of the entire data set by a specified distance measure, and then reorders this dissimilarity matrix into a heatmap called an ordered dissimilarity image (ODI). This ODI allows for a convenient visual investigation of the data set's clustering tendencies by looking for areas which are defined by borders of high dissimilarity. To construct this ODI, I implemented the euclidean distance measure with a random sample size of 5% of the entries within the data set. In this ODI the color red represents areas with high similarity, or low dissimilarity, and the color blue represents areas with low similarity, or high dissimilarity. When inspecting the ODI's diagonal, we can see that there is some underlying cluster structure to this data set as there are two distinguished clusters of high similarity at the lower-left and upper-right ends of the diagonal. These clusters are separated by borders of high dissimilarity. After seeing this ODI, my initial thoughts lead me to believe that this data set contains two meaningful clusters. However, further tests will be performed to test the optimal number of clusters at a later stage. The combined results of these two measures indicate that cluster analysis should be continued on this data set. Finally, this might seem redundant to mention, but I also noted that when inspecting the levels of similarity this data set does not contain perfectly separated boundaries of dissimilarity, especially for the upper-right cluster, and concluded one can expect this for most data sets.



1.2 Data Clustering

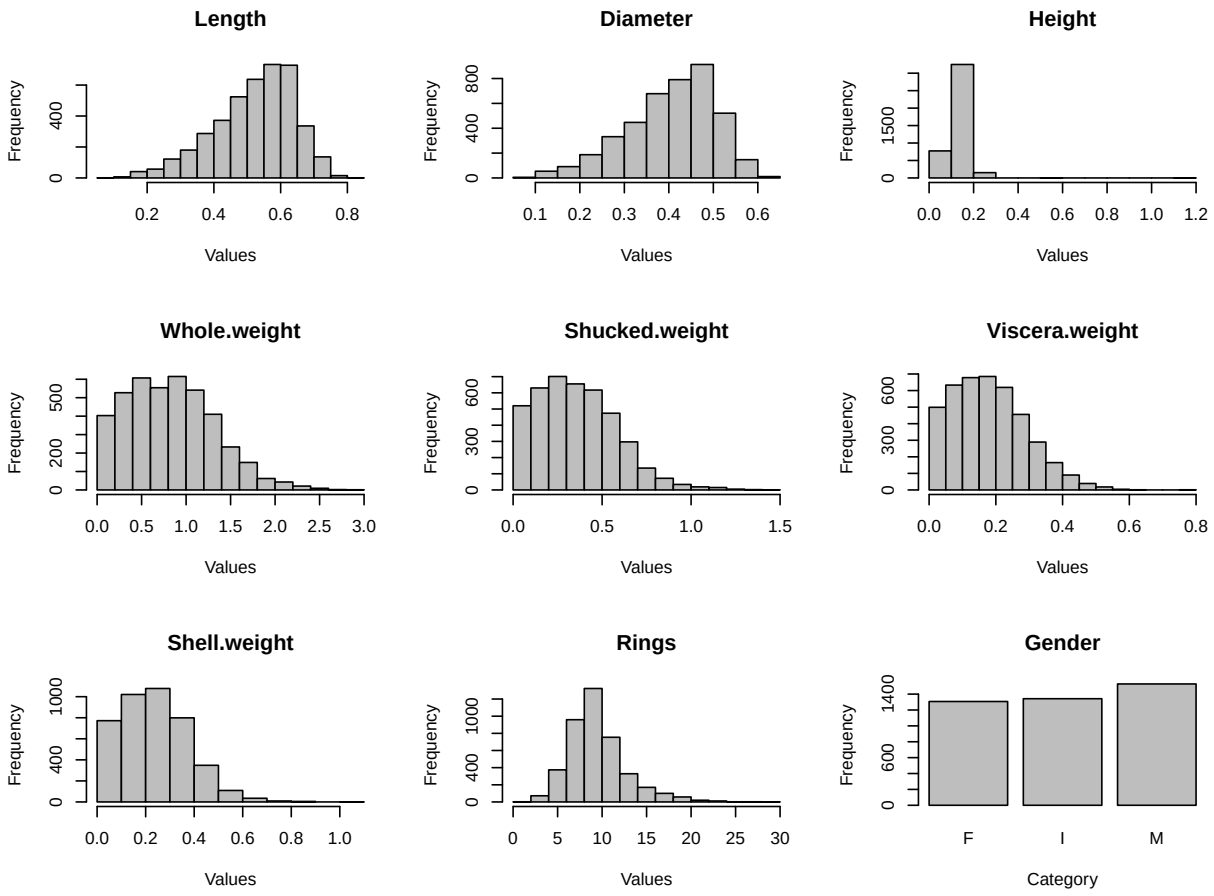


Table 4: Descriptive Statistics [Original data]

Gender	Length	Diameter	Height	Whole.weight
F:1307	Min. :0.075	Min. :0.0550	Min. :0.0000	Min. :0.0020
I:1342	1st Qu.:0.450	1st Qu.:0.3500	1st Qu.:0.1150	1st Qu.:0.4415
M:1528	Median :0.545	Median :0.4250	Median :0.1400	Median :0.7995
NA	Mean :0.524	Mean :0.4079	Mean :0.1395	Mean :0.8287
NA	3rd Qu.:0.615	3rd Qu.:0.4800	3rd Qu.:0.1650	3rd Qu.:1.1530
NA	Max. :0.815	Max. :0.6500	Max. :1.1300	Max. :2.8255

Shucked.weight	Viscera.weight	Shell.weight	Rings
Min. :0.0010	Min. :0.0005	Min. :0.0015	Min. : 1.000
1st Qu.:0.1860	1st Qu.:0.0935	1st Qu.:0.1300	1st Qu.: 8.000
Median :0.3360	Median :0.1710	Median :0.2340	Median : 9.000
Mean :0.3594	Mean :0.1806	Mean :0.2388	Mean : 9.934
3rd Qu.:0.5020	3rd Qu.:0.2530	3rd Qu.:0.3290	3rd Qu.:11.000
Max. :1.4880	Max. :0.7600	Max. :1.0050	Max. :29.000

Before deciding on a clustering algorithm, I wanted to briefly investigate the data set to get a better understanding of the characteristics of the data set. To this end, a brief data visualization and feature descriptive statistics are displayed on the previous page. Please note that these visualizations and statistics describe the original data set, and therefore it differs from the data used for the Section *1.1*.

1.2.1 Clustering Algorithm

As we are looking to identify k groups based on their similarity, partitional clustering methods seemed to be the most intuitive choice. The two most well-known partitional clustering algorithms are K-means and K-mediods. Both K-means and K-mediods initialize k centroids, then repeatedly assign the data points to the closest centroid, after which it recomputes the centroid of the clusters. This assignment and centroid recomputation phase repeats until no further change is visible within the centroids of the clusters. The centroids are initially randomly selected, and the closest centroid is defined as the centroid with the smallest distance measure towards it. The distinction between k-means and k-mediods is in the way that these cluster centroids are recalculated.

In k-means, the cluster centroids are recalculated as the arithmetic mean value of all the data points belonging to a cluster. In k-mediods, the cluster centroids, or cluster mediods, are recalculated as the most centrally located points which improve the cost function. Due to this distinction, k-mediod clustering algorithms are less sensitive to outliers, as they do not operate on arithmetic means, but instead the actual data points. When inspecting the visualizations and statistics on the previous page, we can see that there are multiple skewed distributions and outliers within the data set. The feature Height seems to have the worst outliers, and most of the other features seem to have skewed distributions. For this reason, I believe k-mediods would be a suitable clustering algorithm choice and proceeded to select the PAM algorithm (Partitioning Around Medoids) as my choice of a clustering algorithm.

Quoting the provided textbook by Alboukadel Kassambara, the PAM algorithm searches for k mediods within a dataset, constructing initial clusters by assigning observations to each initial mediod, after which each mediod is swapped with non-mediod points and the objective function is calculated for each potential swap. The objective function relates to the sum of the dissimilarities of all data points to the closest mediod, and the swap that minimizes the objective function the most is selected as the new mediod. This swapping phase is continued until the objective function cannot be decreased any further. A more structured explanation of this algorithm is displayed below.

1. Initialize k data points to become mediods, either randomly or from user-defined values.
2. Calculate dissimilarity matrix using specified distance metric [typically: euclidean, manhattan]
3. Using this dissimilarity matrix, assign data points to the closest mediod
4. For all k clusters, search for data points that decrease the average dissimilarity coefficient and, if found, accept the data point with the largest decrease in the average dissimilarity coefficient as the new mediod.
5. If any mediod has been updated, return to step 3. Otherwise, terminate the algorithm.

1.2.2 Features Used and Data Transformation

1.2.2.1 Features Used Before normalizing the data, I discarded the Gender feature from the data set, as categorical variables are not applicable with the euclidean distance measures. I then investigated the correlation matrix as seen in Table 6 to identify any remaining redundant features in the data set. I found no uncorrelated features, and I left the data set as is. No further tests were performed to remove any features.

Table 6: Correlation Matrix

	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
Length	1.00	0.99	0.83	0.93	0.90	0.90	0.90	0.56
Diameter	0.99	1.00	0.83	0.93	0.89	0.90	0.91	0.57
Height	0.83	0.83	1.00	0.82	0.77	0.80	0.82	0.56
Whole Weight	0.93	0.93	0.82	1.00	0.97	0.97	0.96	0.54
Shucked Weight	0.90	0.89	0.77	0.97	1.00	0.93	0.88	0.42
Viscera Weight	0.90	0.90	0.80	0.97	0.93	1.00	0.91	0.50
Shell Weight	0.90	0.91	0.82	0.96	0.88	0.91	1.00	0.63
Rings	0.56	0.57	0.56	0.54	0.42	0.50	0.63	1.00

1.2.2.2 Data Transformation Despite the PAM algorithm being more robust to outliers than k-means, it is not immune to being skewed by outliers. For this reason, either a robust normalization technique should be used or the outliers should be identified and removed from the data set. All considered normalization techniques will ensure that the implemented euclidean distance measure will not be skewed by different units and scales between the data set's features. I considered four normalization techniques, and investigated the resulting clustering performance across $k \in [2, 6]$. The clustering performance will be evaluated by three measures: the resulting clusters' connectivity, Dunn index, and average silhouette width. The meaning of these three measures are discussed in the next section, *1.2.2*, and for now it should only be noted that we are attempting to minimize the connectivity, whilst simultaneously maximizing the average silhouette width and Dunn index. The formula used and the relevant scores achieved across these three performance measures are displayed for each normalization technique, as well as the highest obtained scores in each measure across $k \in [2, 6]$. I first considered the z-score and robust scalar normalization techniques. Both these methods are considered robust to outliers. For these two methods, I did not apply any outlier removal technique to the original data set. The results for the z-score normalized clustering are displayed in Table 7 and 8.

$$x' = \frac{x_i - \mu(x)}{\sigma(x)}$$

Table 7: PAM internal measures over varying k [z-score]

	k = 2	k = 3	k = 4	k = 5	k = 6
Connectivity	191.5944444	346.2059524	537.1492063	634.7051587	707.6642857
Dunn	0.0062444	0.0049566	0.0048445	0.0040054	0.0045203
Silhouette	0.4698651	0.3834767	0.3293280	0.2921374	0.2717377

Table 8: PAM best measures obtained [z-score]

	Score	Method	Clusters
Connectivity	191.5944444	pam	2
Dunn	0.0062444	pam	2
Silhouette	0.4698651	pam	2

It can be seen that the clustering with the z-score normalized data set performed the best at $k = 2$ for all considered measures. I proceeded to investigate the robust scalar normalization technique, and the results are displayed in Table 9 and 10.

$$x' = \frac{x - median(x)}{Q_{3rd}(x) - Q_{1st}(x)}$$

Table 9: PAM internal measures over varying k [robust scalar]

	k = 2	k = 3	k = 4	k = 5	k = 6
Connectivity	187.3341270	358.7234127	502.6297619	589.4063492	665.7773810
Dunn	0.0047286	0.0041465	0.0047366	0.0051734	0.0054986
Silhouette	0.4454762	0.3502236	0.3527365	0.3199095	0.2919754

Table 10: PAM best measures obtained [robust scalar]

	Score	Method	Clusters
Connectivity	187.3341270	pam	2
Dunn	0.0054986	pam	6
Silhouette	0.4454762	pam	2

As seen above, the robust scalar normalization technique performed worse than the z-score normalization technique in every measure except the connectivity. After investigating these two techniques, I proceeded to investigate min-max scaling. This normalization technique is sensitive to outliers, and prior to applying it to the data set I needed to remove the outliers from the data set. I therefore applied an outlier removal technique to the original data set. I elected to use the interquartile method to identify the outliers, in which any points falling outside 1.5 times the interquartile range are identified as an outlier and removed from the data set. This outlier removal technique was applied to all the features in the data set. The results are displayed on the next page in Table 11 and 12. This method can be seen to perform the best thus far in all three considered internal measures.

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

Table 11: PAM internal measures over varying k [min-max]

	k = 2	k = 3	k = 4	k = 5	k = 6
Connectivity	177.4769841	337.8650794	565.8468254	575.6301587	656.6650794
Dunn	0.0276022	0.0282454	0.0274791	0.0320790	0.0180350
Silhouette	0.4820180	0.3917339	0.3310703	0.3066578	0.2723959

Table 12: PAM best measures obtained [min-max]

	Score	Method	Clusters
Connectivity	177.476984	pam	2
Dunn	0.032079	pam	5
Silhouette	0.482018	pam	2

The final normalization technique considered was that of unit scaling. This normalization technique is also sensitive to outliers, and I applied the interquartile outlier detection method in the same manner as that of the min-max normalization technique. The results are displayed below in Table 13 and 14. The unit scaling normalization technique performed better than the min-max normalization technique in the connectivity and average silhouette width measures and slightly worse in the Dunn index measure.

$$x' = \frac{x}{||x||}$$

Table 13: PAM internal measures over varying k [unit]

	k = 2	k = 3	k = 4	k = 5	k = 6
Connectivity	168.9750000	279.3972222	460.2619048	514.3071429	694.5714286
Dunn	0.0219818	0.0222986	0.0208781	0.0279870	0.0224759
Silhouette	0.5164396	0.4298652	0.3759114	0.3578434	0.3129747

Table 14: PAM best measures obtained [unit]

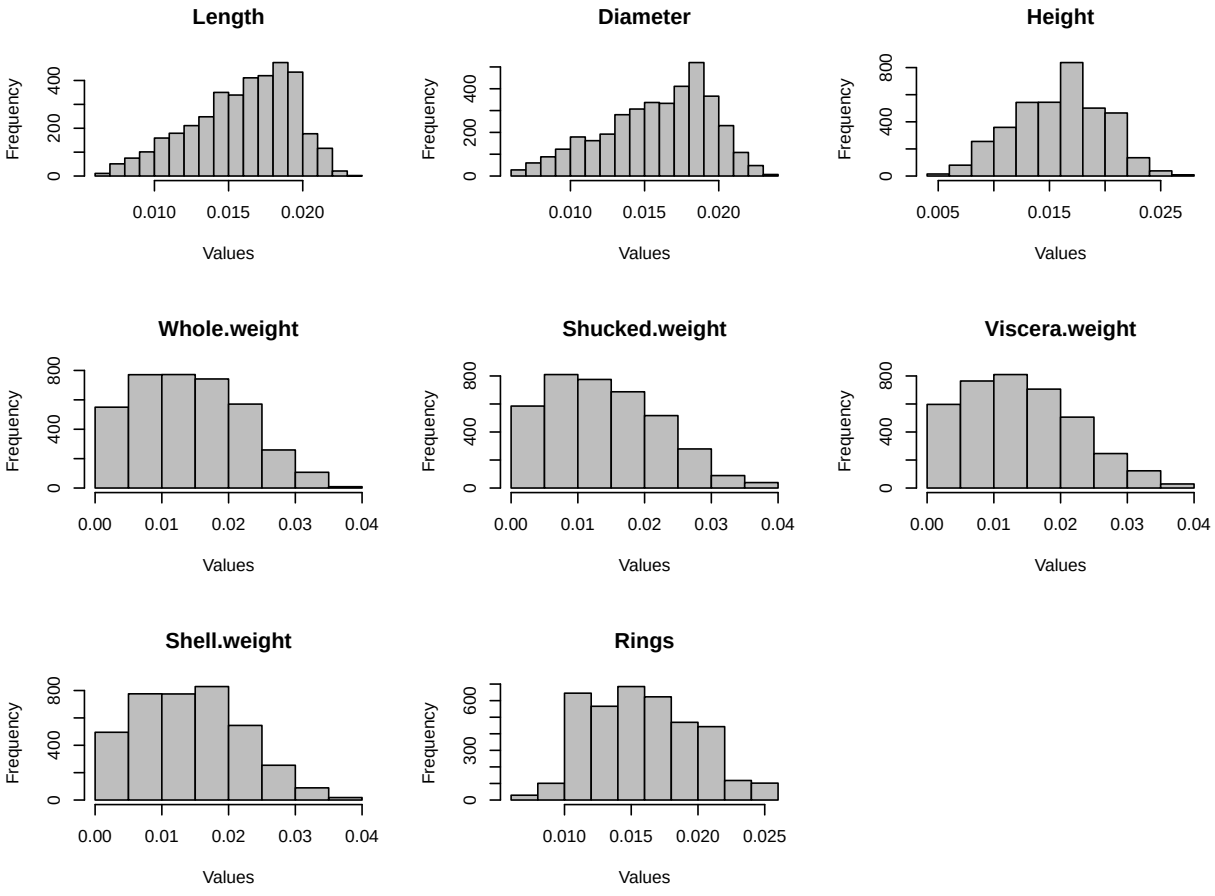
	Score	Method	Clusters
Connectivity	168.9750000	pam	2
Dunn	0.0279870	pam	5
Silhouette	0.5164396	pam	2

To conclude, despite robust scaling and z-score normalization being robust to outliers, the unit scaled clustering approach ultimately performed the best of all considered normalization techniques. I therefore selected unit scaling as the appropriate normalization technique, and a brief visualization and descriptive statistics of the data to be used is displayed on the next page. Once again, the outliers were removed from the dataset by the interquartile method before applying unit length scaling.

Table 15: Descriptive Statistics [Unit-length normalized data]

Length	Diameter	Height	Whole.weight
Min. :0.006257	Min. :0.006069	Min. :0.004588	Min. :0.0007609
1st Qu.:0.013735	1st Qu.:0.013508	1st Qu.:0.012617	1st Qu.:0.0077520
Median :0.016329	Median :0.016444	Median :0.016058	Median :0.0137226
Mean :0.015903	Mean :0.015857	Mean :0.015751	Mean :0.0141814
3rd Qu.:0.018618	3rd Qu.:0.018598	3rd Qu.:0.018925	3rd Qu.:0.0200066
Max. :0.023196	Max. :0.023492	Max. :0.027527	Max. :0.0380886

Shucked.weight	Viscera.weight	Shell.weight	Rings
Min. :0.0006866	Min. :4.041e-05	Min. :0.0008217	Min. :0.006697
1st Qu.:0.0073100	1st Qu.:7.314e-03	1st Qu.:0.0079013	1st Qu.:0.013393
Median :0.0132064	Median :1.325e-02	Median :0.0139063	Median :0.015067
Mean :0.0140338	Mean :1.405e-02	Mean :0.0142906	Mean :0.015788
3rd Qu.:0.0198702	3rd Qu.:1.972e-02	3rd Qu.:0.0198797	3rd Qu.:0.018415
Max. :0.0387712	Max. :3.976e-02	Max. :0.0395066	Max. :0.025112



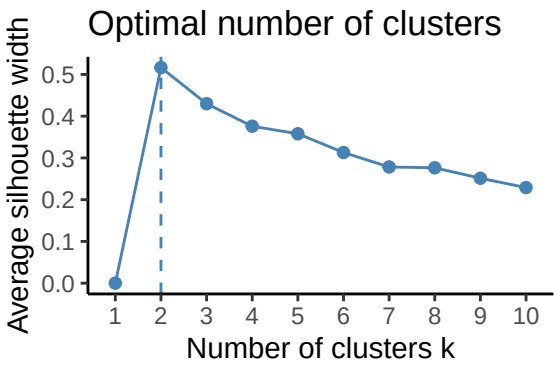
1.2.2 Optimal number of clusters

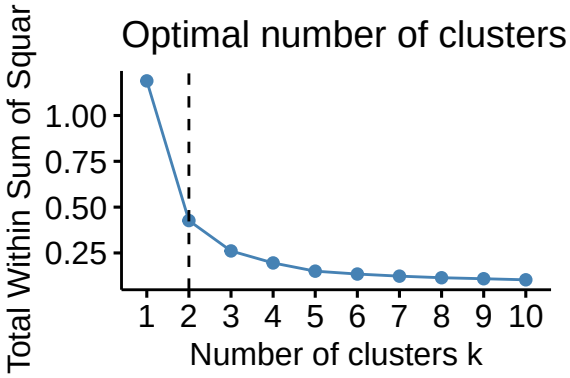
To find the optimal number of clusters for this application, I considered two direct methods and one statistical method. The direct methods I considered were the average silhouette and elbow methods, and both of these methods are concerned with optimizing a global clustering criterion. The average silhouette method seeks to maximize the average silhouette width of the clustering across different values of k. This so-called silhouette width coefficient measures the distances between the points in the clusters, and therefore this method aims to maximize the separation between the different cluster centers by maximizing the average silhouette width. A larger average silhouette width implies a better separation between the clusters, and therefore higher inter-cluster distances.

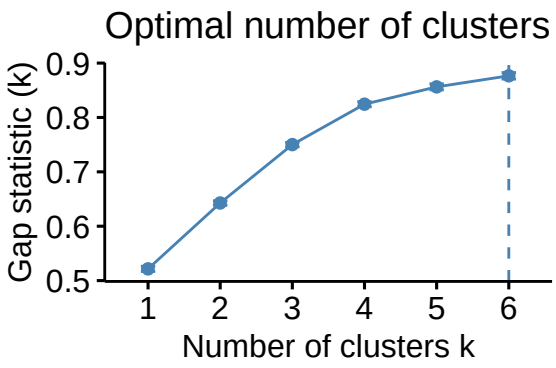
However, we can recall that the objective function of partitional clustering algorithms typically aim to minimize intra-cluster distances (i.e. find compact clusters) as well as to maximize inter-cluster distances (i.e. find well-separated clusters). Therefore, the average silhouette method is not sufficient by itself as it does not consider the minimization of the intra-cluster distances. To this end, I will consider the elbow method. This elbow method considers the minimization of the total intra-cluster variation, or total within-cluster sum of squares (WSS) across different values of k. Therefore, to achieve compact clusters, the so-called bend/knee in the WSS plot is considered as an indication of the optimal number of clusters.

The results for the two direct methods are displayed below. The average silhouette width can be seen to be at a maximum of $k = 2$, and the WSS is at a minimum at $k = 2$. These methods therefore suggest two clusters as the optimal number of clusters for maximizing inter-cluster distances and minimizing intra-cluster distances.

I also considered the gap statistic method and, quoting the class slides provided to us, this statistical approach compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data, and therefore the value of k that maximizes this gap statistic is the estimate of the optimal number of clusters. The results for this statistical method is displayed below, and it can be seen to be at a maximum of $k = 6$, which disagrees with what was found in the previous two methods.







As two final internal measures for cluster validation, I investigated the cluster connectivity and Dunn index. The cluster connectivity is a minimization criterion that measures to what extent data points are placed in the same cluster as their nearest neighbors, and the Dunn index is a maximization criterion that represents the ratio of the minimum inter-cluster distance and the maximum intra-cluster distance. The cluster connectivity and average silhouette width are at their respective minimums and maximums at $k = 2$, yet for the Dunn Index the clustering approach performed slightly better at $k = 5$. However, by majority rule, I selected $k = 2$ as the optimal number of clusters and proceeded to investigate the cluster descriptive statistics.

Table 17: PAM internal measures over varying k

	k = 2	k = 3	k = 4	k = 5	k = 6
Connectivity	168.9750000	279.3972222	460.2619048	514.3071429	694.5714286
Dunn	0.0219818	0.0222986	0.0208781	0.0279870	0.0224759
Silhouette	0.5164396	0.4298652	0.3759114	0.3578434	0.3129747

Table 18: Best achieved PAM internal measures

	Score	Method	Clusters
Connectivity	168.9750000	pam	2
Dunn	0.0279870	pam	5
Silhouette	0.5164396	pam	2

1.2.3 Cluster Descriptive Statistics

Table 19: Entries in different clusters

Cluster	Entries
1	1929
2	1852

Table 20: Cluster centroids [Scaled]

Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
0.0140398	0.0137037	0.0131902	0.0078773	0.0076735	0.0082836	0.0082174	0.0133931
0.0187705	0.0189895	0.0183516	0.0207227	0.0202135	0.0201635	0.0199113	0.0167414

Table 21: Cluster centroids [Unscaled]

Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
0.460	0.350	0.115	0.4400	0.1900	0.1025	0.130	8
0.615	0.485	0.160	1.1575	0.5005	0.2495	0.315	10

Table 19 illustrates how many entries are within each cluster, and Table 20 and 21 illustrate the cluster centroids in their respective scaled and unscaled forms. We can note that the centroid values for length, diameter, whole weight, shucked weight, viscera weight, and shell weight are quite different. To further investigate the types of ranges the features can take, I investigated some descriptive statistics of the features within the two different clusters.

Table 22: Descriptive Statistics: Cluster 1 [Unscaled]

Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
Min. :0.205	Min. :0.155	Min. :0.0400	Min. :0.0425	Min. :0.0170	Min. :0.0005	Min. :0.0130	Min. : 4.000
1st Qu.:0.375	1st Qu.:0.285	1st Qu.:0.0950	1st Qu.:0.2505	1st Qu.:0.1060	1st Qu.:0.0535	1st Qu.:0.0765	1st Qu.: 7.000
Median :0.450	Median :0.345	Median :0.1150	Median :0.4390	Median :0.1855	Median :0.0925	Median :0.1290	Median : 8.000
Mean :0.434	Mean :0.333	Mean :0.1112	Mean :0.4281	Mean :0.1852	Mean :0.0931	Mean :0.1282	Mean : 8.355
3rd Qu.:0.500	3rd Qu.:0.385	3rd Qu.:0.1300	3rd Qu.:0.5970	3rd Qu.:0.2555	3rd Qu.:0.1310	3rd Qu.:0.1750	3rd Qu.:10.000
Max. :0.600	Max. :0.465	Max. :0.1900	Max. :0.9100	Max. :0.4950	Max. :0.2270	Max. :0.3500	Max. :15.000

Table 23: Descriptive Statistics: Cluster 2 [Unscaled]

Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
Min. :0.4850	Min. :0.345	Min. :0.0800	Min. :0.6855	Min. :0.2405	Min. :0.1120	Min. :0.1210	Min. : 6.00
1st Qu.:0.5750	1st Qu.:0.450	1st Qu.:0.1500	1st Qu.:0.9407	1st Qu.:0.4115	1st Qu.:0.2040	1st Qu.:0.2690	1st Qu.: 9.00
Median :0.6100	Median :0.475	Median :0.1650	Median :1.1255	Median :0.4950	Median :0.2460	Median :0.3150	Median :10.00
Mean :0.6117	Mean :0.480	Mean :0.1645	Mean :1.1712	Mean :0.5165	Mean :0.2580	Mean :0.3281	Mean :10.55
3rd Qu.:0.6450	3rd Qu.:0.505	3rd Qu.:0.1750	3rd Qu.:1.3435	3rd Qu.:0.6026	3rd Qu.:0.3011	3rd Qu.:0.3750	3rd Qu.:12.00
Max. :0.7600	Max. :0.600	Max. :0.2400	Max. :2.1275	Max. :0.9600	Max. :0.4920	Max. :0.6250	Max. :15.00

	Cluster 1	Cluster 2
F	344	816
I	1110	151
M	475	885

When inspecting the above tables, my initial thoughts lead me to believe that this clustering solution will be distinguished by the length, diameter, whole weight, shucked weight, viscera weight, and shell weight features. This is because their centroid locations and ranges that these features assume within the two clusters are quite different with some distinct boundaries that will be discussed later. I noticed that the target feature, Rings, does not seem to be as informative for the clustering as the other features since it's range of values are quite similar across both clusters. However, this will all be investigated at a later stage. The gender feature provides a great deal of distinguishment between the clusters, and when inspecting the distribution of genders across the clusters we can note that these clusters seem to be characterised mostly by adult and infant abalones as the clusters are dominated by the two separate age groups. Once again, Gender was reinserted into the dataset, and was not used in the actual clustering solution.

Table 25: Ring Modes for Clusters [Unscaled]

	Cluster: 1	Cluster: 2
Rings:	8	10

Table 25 illustrates the modes that were calculated for the target feature in each cluster. I noted that the calculated modes were exactly the same as the cluster centroids, as seen in Table 21. Table 26 shows some descriptive statistics of the clusters. Note that these statistics are still in the respective scaled form. The maximum, average and median within cluster distances are shown, and it can be seen that the second cluster is larger than the first cluster. However, when referring back to Table 19 we can note that there are actually less data points in this cluster, implying it is less compact than the first cluster. Similarly, the minimum and average distances between points in the separate clusters are displayed. It can be seen that at least one of the cluster points are quite close to each other, as the clusterwise minimum distance between the separate clusters are very low. However, the clusterwise average distance between the separate clusters indicate that most of the points seem to be situated significantly further from each other than this separate clusterwise minimum distance.

Table 26: Descriptive statistics: Part 1 [Scaled data]

Measure	Cluster 1	Cluster 2
Maximum within cluster distances/diameters	0.0357043	0.0502994
Clusterwise within cluster average distances	0.0119635	0.0144911
Clusterwise within cluster median distances	0.0107634	0.0127746
Clusterwise minimum distances of the separate clusters	0.0011057	0.0011057
Clusterwise average distances of the separate clusters	0.0295378	0.0295378
Cluster average silhouette widths	0.5600442	0.4710221
Widest within cluster gaps	0.0083200	0.0080269

Furthermore, the cluster average silhouette widths are displayed and when recalling the meaning of silhouette widths from Section 1.2.2 it can be seen that the first cluster has a better separation than the second cluster. The widest cluster gaps are also shown, and are quite similar across both clusters.

Table 27: Descriptive statistics: Part 2 [Scaled data]

nms	vls
Dunn	0.0219818
Dunn 2	2.0383414
Entropy of Cluster Membership	0.6929398
Average within/between ratio	0.4469383
Calinski and Harabasz index	6751.0750424
Separation Index	0.0028239

Table 15 displays some more detailed cluster statistics, the first of which is the Dunn index, calculated firstly as previously mentioned.

$$\frac{MinimumSeparation}{MaximumDiameter}$$

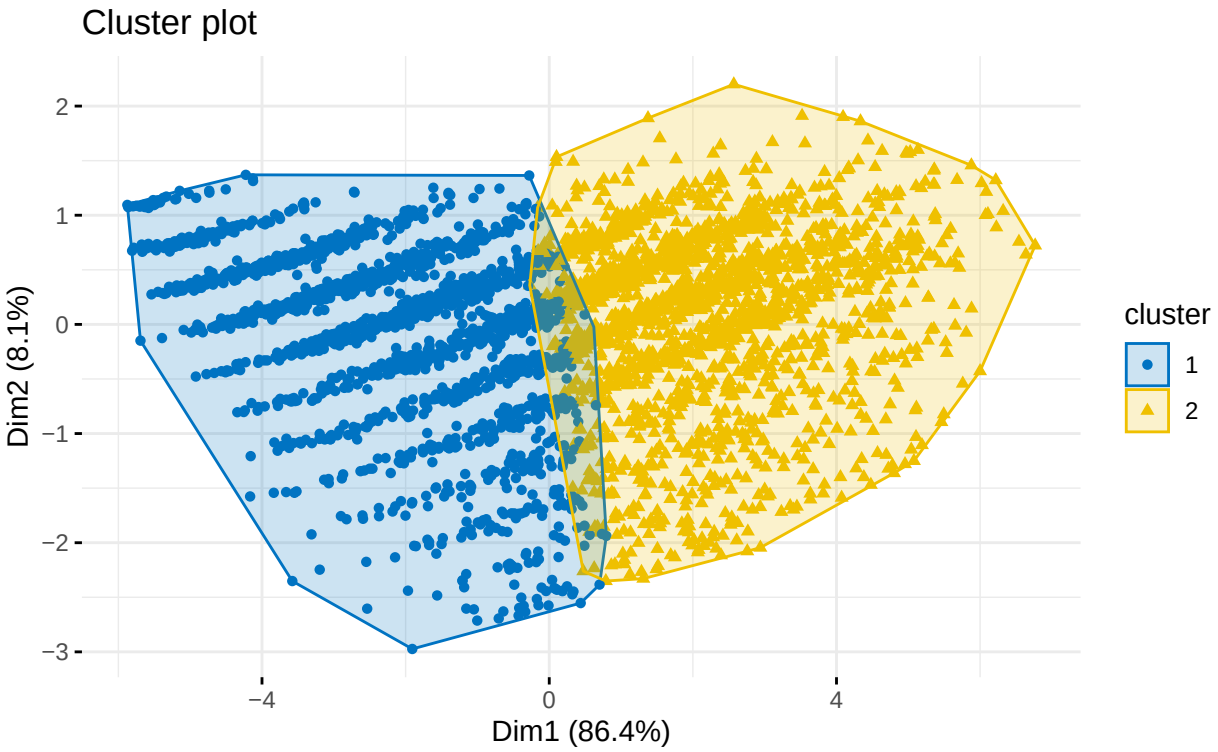
This measure is followed by a lesser known dunn index variant calculated as follows.

$$\frac{MinimumAverageDissimilarity}{MaximumAverageDissimilarity}$$

In this variant, the minimum dissimilarity is calculated between the two clusters, and the average dissimilarity is calculated within the separate clusters. Furthermore, the entropy of the distribution of cluster memberships, the average within/between ratios, the Calinski and Harabasz index, and the separation index are also displayed.

1.3 Cluster Visualization

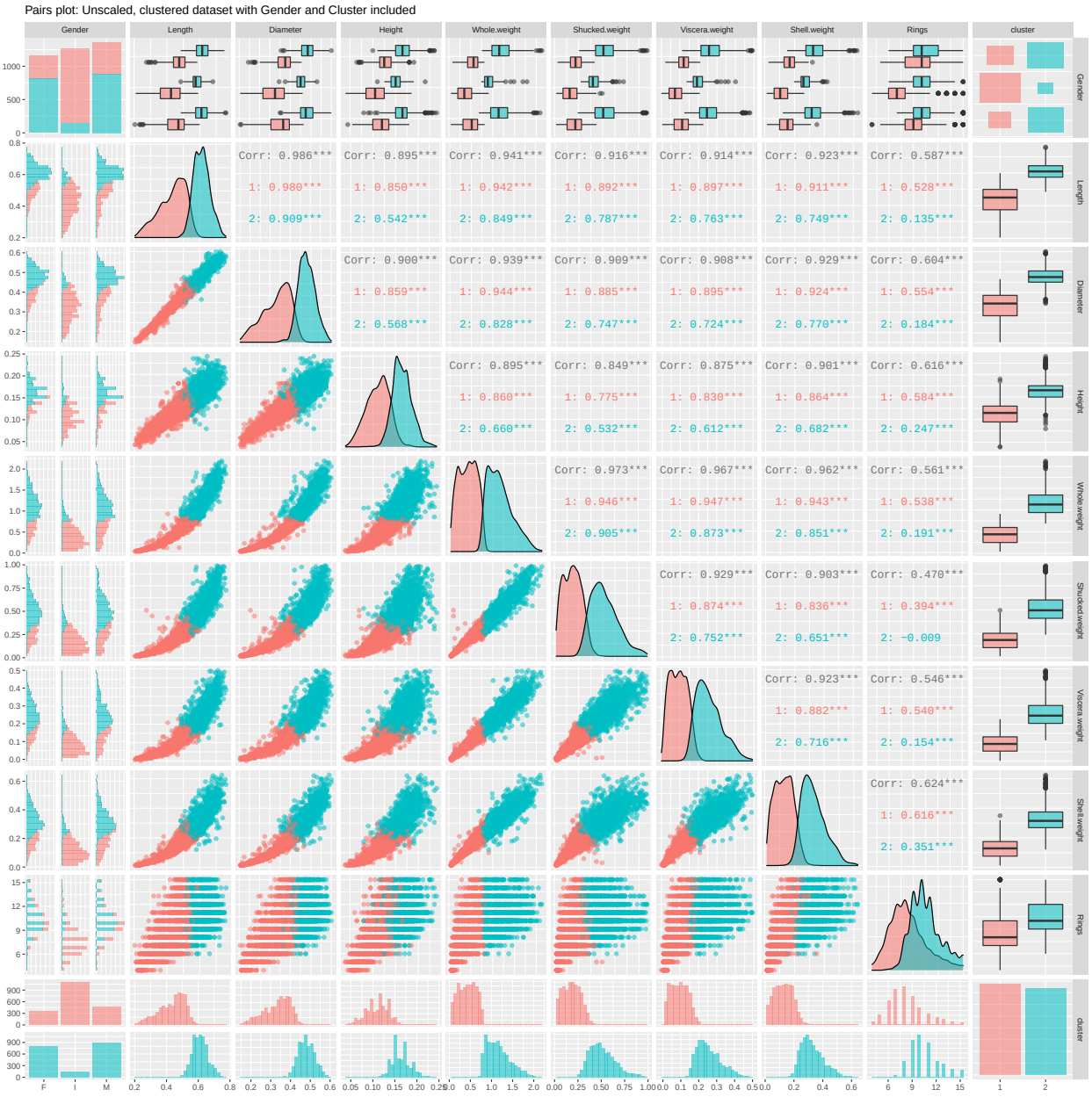
1.3.1 Cluster Plot



A visualization of the cluster is displayed above. The function used, *factoextra::fviz_cluster()*, performs PCA to reduce the dimensionality of the data such that it can be visualized in a 2D plot by using the first two principal dimensions. It can be seen that the first principal dimensions accounts by far the most of the data set's variability with 86.4% explanation. There seems to be some overlap which will be considered in Section *1.4*.

1.3.2 Cluster Scatterplot Matrix

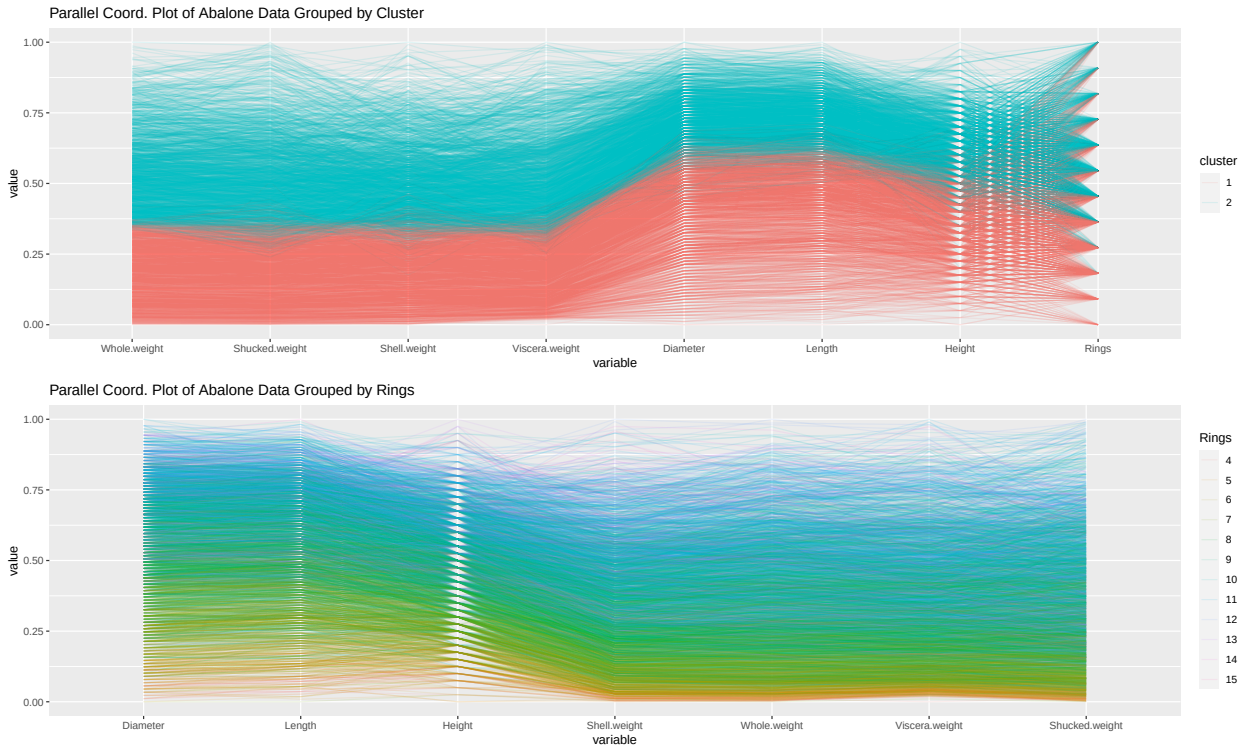
The cluster scatterplot matrix is displayed below where Cluster 1 is highlighted in red, and Cluster 2 is highlighted in blue. In addition to illustrating the clusters formed across the descriptive features used in the clustering algorithm, I re-inserted Gender into the data set and included a newly created cluster feature which simply represented the allocated cluster. These two features were inserted into the data set to aid with visualization and rule extraction in Section 1.4, and their respective columns can be seen at the left columns and bottom rows. The descriptive features actually used in the clustering are Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight, and Rings.



1.3.3 Cluster Histogram Plot

The descriptive feature histograms are displayed above. The first set of histograms display all the descriptive features used in the clustering process, with the exception of the target feature, in which a color scheme was applied to distinguish between the different cluster allocations. This color scheme was kept the same as in the cluster scatterplot matrix, and Cluster 1 is highlighted in red, and Cluster 2 is highlighted in blue. This set of histograms is technically not what the question asked for, but I included it for completion and my own investigation. To address the assignment's question, a second plot containing an unstacked set of histograms is included, and in this set of histograms the effect of varying the target feature, or the number of rings, is displayed across the different features' distributions. This second histogram will allow for investigation of the most informative features in classifying an abalone's number of rings.

1.3.4 Cluster Parallel Coordinate Plot



The parallel coordinate plot is displayed above. The color of the plot assumes one of two discrete values, which represent with the cluster that the data entries were identified as belonging in. This color scheme was kept the same as in the cluster scatterplot matrix, and Cluster 1 is highlighted in red, and Cluster 2 is highlighted in blue. A second parallel coordinate plot is also displayed, to aid in later investigation, where the number of Rings determine the color of the data entries. It is important to note, that for both of these figures, I used a function *GGally::ggparcoord* which allowed me to reorder the features on the x-label, such that the entries are displayed in a smooth and continuous manner. This will ultimately allow for easier investigation, as the variables with the smoothest and most homogenous connections will be placed on the left-hand side of the graph, and the most turbulent variables will be placed on the right-hand side of the graph. Finally, all features were scaled accordingly by min-max scaling to further aid in visualizing the parallel coordinate plot.

1.4 Cluster Analysis

1.4.1 Cluster Rules From Descriptive Statistics

Referring back to page 11 and 12, I will consider the tabulated descriptive statistics for this part of the assignment. The first rule can be deduced when inspecting the distribution of the Gender feature across the clusters. The distribution is displayed below, and there is a clear distinction between the age groups placed in the different clusters, in which the first cluster contains the majority of the infant abalones and the second cluster contains the majority of the adult abalones.

	Cluster 1	Cluster 2
F	344	816
I	1110	151
M	475	885

When investigating Table 22 and 23, we can see that the length, diameter, whole weight, shucked weight, viscera weight and shell weight features are also quite different across the clusters. These features are summarized in the tables below.

Length	Diameter	Whole.weight	Shucked.weight	Viscera.weight	Shell.weight
Min. :0.205	Min. :0.155	Min. :0.0425	Min. :0.0170	Min. :0.0005	Min. :0.0130
1st Qu.:0.375	1st Qu.:0.285	1st Qu.:0.2505	1st Qu.:0.1060	1st Qu.:0.0535	1st Qu.:0.0765
Median :0.450	Median :0.345	Median :0.4390	Median :0.1855	Median :0.0925	Median :0.1290
Mean :0.434	Mean :0.333	Mean :0.4281	Mean :0.1852	Mean :0.0931	Mean :0.1282
3rd Qu.:0.500	3rd Qu.:0.385	3rd Qu.:0.5970	3rd Qu.:0.2555	3rd Qu.:0.1310	3rd Qu.:0.1750
Max. :0.600	Max. :0.465	Max. :0.9100	Max. :0.4950	Max. :0.2270	Max. :0.3500

Length	Diameter	Whole.weight	Shucked.weight	Viscera.weight	Shell.weight
Min. :0.4850	Min. :0.345	Min. :0.6855	Min. :0.2405	Min. :0.1120	Min. :0.1210
1st Qu.:0.5750	1st Qu.:0.450	1st Qu.:0.9407	1st Qu.:0.4115	1st Qu.:0.2040	1st Qu.:0.2690
Median :0.6100	Median :0.475	Median :1.1255	Median :0.4950	Median :0.2460	Median :0.3150
Mean :0.6117	Mean :0.480	Mean :1.1712	Mean :0.5165	Mean :0.2580	Mean :0.3281
3rd Qu.:0.6450	3rd Qu.:0.505	3rd Qu.:1.3435	3rd Qu.:0.6026	3rd Qu.:0.3011	3rd Qu.:0.3750
Max. :0.7600	Max. :0.600	Max. :2.1275	Max. :0.9600	Max. :0.4920	Max. :0.6250

When inspecting the above two tables, we can deduce the following 7 general, and somewhat vague rules.

1. Infant abalones are more likely to be placed in the first cluster, and adult abalones are more likely to be placed in the second cluster.
2. Abalones with lower lengths are more likely to be placed in the first cluster, and abalones with higher lengths are more likely to be placed in the second cluster.
3. Abalones with lower diameters are more likely to be placed in the first cluster, and abalones with higher diameters are more likely to be placed in the second cluster.

4. Abalones with lower whole weights are more likely to be placed in the first cluster, and abalones with higher whole weights are more likely to be placed in the second cluster.
5. Abalones with lower shucked weights are more likely to be placed in the first cluster, and abalones with higher shucked weights are more likely to be placed in the second cluster.
6. Abalones with lower viscera weights are more likely to be placed in the first cluster, and abalones with higher viscera weights are more likely to be placed in the second cluster.
7. Abalones with lower shell weights are more likely to be placed in the first cluster, and abalones with higher shell weights are more likely to be placed in the second cluster.

However, in an attempt to deduce more specific rules, I investigated the minimums, maximums, and quantiles of the relevant features and saw that in my clustering approach the following rules hold.

IF: $Length > 0.6$ || $Diameter > 0.465$ || $Whole.Weight > 0.91$: Cluster 2

ELSE IF: $Shucked.Weight > 0.495$ || $Viscera.Weight > 0.227$ || $Shell.Weight > 0.35$: Cluster 2

ELSE IF: $Length < 0.485$ || $Diameter < 0.345$ || $Whole.Weight < 0.6855$: Cluster 1

ELSE IF: $Shucked.Weight < 0.2405$ || $Viscera.Weight < 0.1120$ || $Shell.Weight < 0.1210$: Cluster 1

Table 31: Median Values of Features across Clusters

	Cluster 1	Cluster 2
Length	0.4500	0.6100
Diameter	0.3450	0.4750
Whole.weight	0.4390	1.1255
Shucked.weight	0.1855	0.4950
Viscera.weight	0.0925	0.2460
Shell.weight	0.1290	0.3150

These rules are particularly informative as it can be seen that the values used in the IF statements are approximately around the median values of the features, and do not operate on small margins outside the 1st and 3rd quantiles. This can be confirmed when inspecting the quantiles from the descriptive statistics above.

1.4.2 Cluster Scatterplot Matrix Feature Separation & Rule Identification

1.4.2.1 Feature Cluster Separation Identification

When inspecting the SPLOM on page 15, we can note that the features Rings and Height are the worst at separating the clusters and the whole weight feature is the best at separating the clusters. This is particularly evident when investigating the density plot on the diagonal. The whole weight feature has the least overlap between clusters, and rings and height have the most overlap. The remaining descriptive features, not considering Gender or Cluster, provide a decent separation of clusters. Therefore, to conclude, the descriptive features that result in good separation of clusters are as follows, in no particular order: Length, Diameter, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight.

1.4.2.1 Feature Cluster Rule Separation Identification

I struggled with this question, and was uncertain how to deduce any new rules from the scatterplot matrix. The assignment detail distinguished the rules between section *1.4.1* and *1.4.2* as rules to discern among clusters, and rules to characterize clusters. I might have misunderstood this, but I believe this to be the same task. Therefore if I assume I'm allowed to repeat rules from the previous question, I deduced the following rules. Note that all values used are approximate, as it is difficult to read a precise value from the SPLOM graph.

1. Infant abalones are more likely to be placed in the first cluster, and adult abalones are more likely to be placed in the second cluster.
2. Abalones with a $Length < 0.5$ are definitely placed in the first cluster, and abalones with a $Length > 0.6$ are definitely placed in the second cluster.
3. Abalones with a $Diameter < 0.4$ are definitely placed in the first cluster, and abalones with a $Diameter > 0.5$ are definitely placed in the second cluster.
4. Abalones with a $Whole.Weight < 0.8$ are definitely placed in the first cluster, and abalones with a $Whole.Weight > 1.0$ are definitely placed in the second cluster.
5. Abalones with a $Shucked.Weight < 0.25$ are definitely placed in the first cluster, and abalones with a $Shucked.Weight > 0.5$ are definitely placed in the second cluster
6. Abalones with a $Viscera.weight < 0.1$ are definitely placed in the first cluster, and abalones with a $Viscera.weight > 0.2$ are definitely placed in the second cluster
7. Abalones with a $Shell.Weight < 0.1$ are definitely placed in the first cluster, and abalones with a $Shell.Weight > 0.4$ are definitely placed in the second cluster.

This process ultimately resulted in me re-doing the previous rules, and a very similar if-statement can be used to characterize the clusters.

IF: $Length > 0.6$ || $Diameter > 0.5$ || $Whole.Weight > 1$: Cluster 2

ELSE IF: $Shucked.Weight > 0.5$ || $Viscera.Weight > 0.2$ || $Shell.Weight > 0.4$: Cluster 2

ELSE IF: $Length < 0.5$ || $Diameter < 0.4$ || $Whole.Weight < 0.8$: Cluster 1

ELSE IF: $Shucked.Weight < 0.25$ || $Viscera.Weight < 0.1$ || $Shell.Weight < 0.1$: Cluster 1

1.4.3 Cluster Histogram Feature Importance

On page 16, two sets of histograms are displayed. The first set of histograms displays the separation of the clusters for each feature, and the second set of histograms displays unstacked histograms with reference to each of the number of rings for each feature. The first set of histograms is technically not necessary for identifying which features are most important for the classification of the abalone, as it instead gives insight into the separation between the two clusters and I included it for the sake of completion and my own investigation. The second, unstacked set of histograms gives insight into the change of each feature for varying numbers of rings. Features that display changes in distribution as the target feature is varied will be considered as having a higher importance in classifying the abalones.

When inspecting the set of histograms, it is immediately obvious which features' distributions change the most with varying rings. The left three features in the unstacked set of histograms share a general increasing trend in their respective values as the number of rings increases. These three left features are the features length, diameter and height and are the three most informative features in classifying the abalone. It is hard to say which features are specifically the most informative, but it is clear that of these three most informative features, height is the least informative as its distribution changes less than length and diameter's respective distributions do.

Simiarly, the four right features' distributions vary the least across varying numbers of rings. These features are whole weight, shucked weight, viscera weight and shell weight, and they are the four least informative features in this data set. Once again, it is hard to rank these least informative features as their distributions seem quite similar, but I concluded that shell weight was the most informative from these four least informative features as its' distribution varies just a slight bit more than the other least informative features.

1.4.4 Cluster Parallel Coordinate Plot Feature Importance

On page 17, two parallel coordinate plots are displayed. The first parallel coordinate plot displays the formed clusters across the dataset, and the left-most variables on the x-axis represent the most informative features in clustering the data set. The second parallel coordinate plot displays the number of rings across the data set, and the left-most variables on the x-axis represent the most informative features in classifying the number of rings. These left-most features in the coordinate plots are considered the most informative features with respect to either the clustering or the number of rings. This is because the color of the parallel coordinate plots are set to vary according to either the cluster, or the number of rings, and by setting this color, the function I implemented allowed for re-ordering the variables across these plots such that the most contiguous sets of lines across features are displayed in a decreasing sense of similarity from left to right. We can therefore deduce that the areas of the parallel plot with the most distinct and clean separations of lines are the most informative features. This is somewhat different from the class implementation, but I found it to be quite a convenient and handy tool. I then proceeded to investigate the assignment question, which specified that the parallel coordinate plot should be used support the most important features in classifying the abalone.

In terms of classifying the abalones' number of rings, the second parallel coordinate plot illustrates a smooth set of lines flowing through the diameter, length, and height features, after which the lines become a bit more entangled. This is especially clear when we focus on the top blue and middle green chunks of the lines flowing through these three features. When we trace these two sections of lines from the left-most to the right-most features we can see that at the left-most side these chunks of lines are an ordered and almost separable set of lines, yet after passing through the height feature and moving into the shell weight feature, the previously separable set of lines start to get mixed and the borders of the parallel coordinate plot become much less clear. Therefore, the three most informative features are diameter, length, height and the four least informative features are shell weight, whole weight, viscera weight and shucked weight. In terms of classifying the number of rings, I could not identify any specific feature should be removed as none of the least informative features seemed redundant enough to justify its' removal.

In terms of clustering the dataset, the first parallel coordinate plot illustrates a smooth and highly separable set of lines. The blue and red chunks of lines only start to become entangled after the lines have passed

the length feature and enter the height and ring features. These features provide the least information with regards to clustering the data set, and can potentially be removed from the clustering approach.

References

Kassambara, A. (2017). Practical guide to cluster analysis in R: unsupervised machine learning. Journal of Computational and Graphical Statistics.

Engelbrecht, AP. (2019). Topic 4: Data Clustering and Analysis, Data Analytics 344, Stellenbosch University, Department of Industrial Enigneering, and Division of Computer Science

Notes

Unscaling of medoids in *Section 1.2.3* and remapping in *Section 1.3.2* was performed by finding the row indices in the scaled data set, then matching those indices in the data set right before it was scaled. A similar process was performed to reinsert Gender after clustering has been performed.

|| was used to indicate the OR operator

The full dataset was not used in the ODI, as the resulting graph was excessively large [>100MB] and did not differ much from a sampled ODI. The Hopkins statistic did also not differ much when implemented on a smaller sample of the data set.

```
## Finding R package dependencies ... Done!
```

```
##  [1] "rmarkdown"   "knitr"       "cluster"     "dplyr"       "factoextra"
##  [6] "knitr"       "readxl"      "clustertend" "factoextra"  "ggplot2"
## [11] "knitr"       "knitr"       "knitr"       "knitr"       "clValid"
## [16] "knitr"       "knitr"       "clValid"     "knitr"       "factoextra"
## [21] "fpc"         "knitr"       "prodlim"     "knitr"       "knitr"
## [26] "knitr"       "knitr"       "pracma"      "knitr"       "knitr"
## [31] "factoextra"  "GGally"      "grid"        "ggplot2"     "gridExtra"
## [36] "GGally"      "knitr"       "knitr"       "knitr"       "renv"
## [41] "cluster"     "dplyr"       "factoextra"  "fpc"         "NbClust"
## [46] "readxl"      "webshot"
```