Clustering of abalone data set

Chris Oosthuizen

21/12/2020

1 Introduction

Clustering is a machine learning technique which is used to groups objects that share underlying patterns together based on similarity measures. These algorithms have countless domain applications and are powerful tools to extract knowledge from complex data structures. Moreover, clustering is unique in the sense that it offer solutions that are most often more interpretable than other machine learning algorithms because of the visualizations they offer. This report aims to illustrate this point by performing a clustering analysis on the abalone data set. The abalone data set contains 9 predictors and a target feature, rings. The description of these features:

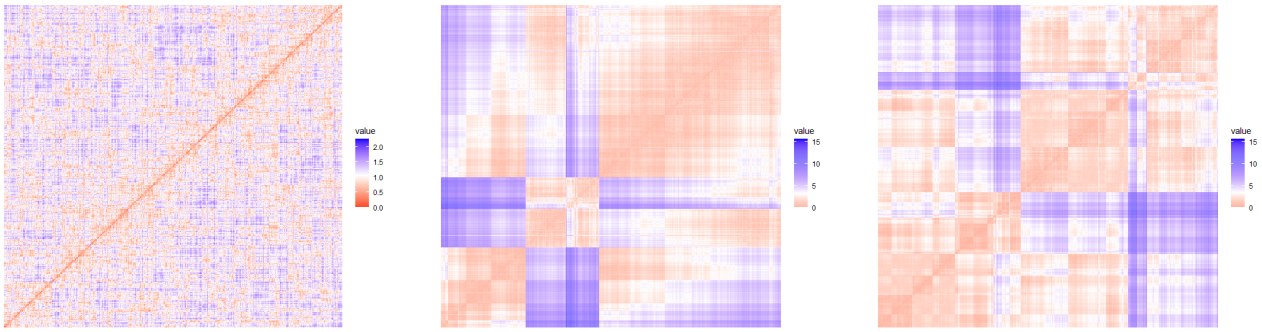
1. Sex, which indicates whether an abalone is male, female, or infant
2. Length, which is the longest dimension of the shell
3. Diameter, which is the dimension of the shell perpendicular to the length
4. Width, which is the thickness of the abalone
5. Height, which is the dimension of the body inside the shell
6. Whole weight, which is the weight of the entire abalone
7. Shucked weight, which is the weight of the abalone itself
8. Viscera weight, which is the weight of the abalone internal organs
9. Shell weight, which is simply the weight of its shell
10. Rings, number of rings on the abalone shell

Generally, clustering is performed as an unsupervised approach to find possible classes which can then be labelled as the target feature. However, in the abalone data set the target feature rings is already provided. Therefore, this report will exclude the target feature rings during clustering, that is, perform unsupervised modeling. The target feature rings will only be compared to the cluster labels after the clustering process and then inspect if there is any similarity between the two. This report will include a discussion on cluster tendency, the clustering process, and clustering analysis with a focus on visualizations for knowledge extraction.

2 Cluster tendency

The feasibility of clustering the abalone data set can be approached by statistical and visual methods. The abalone dataset is mixed because it has a categorical feature (sex), which needs to be considered because both these cluster tendency approaches assume similar data types as input. Therefore, the categorical feature will be one hot encoded to create a mixed data set[1].This is not completely statistically correct, however, it does provide some information with regards to cluster tendency. In order to investigate the effect this may have on the cluster tendency approach, another data set will be created which consist of the numerical features only.

The Hopkins statistic is an indicator of cluster tendency that measures the probability that a data set is generated by a uniform data distribution [1]. The numerical data set and mixed data set produces a Hopkins statistic of 0.945 and 0.939, respectively. The Hopkins statistics are very close to 1, which indicates that both data sets contain meaningful clusters [2]. The visual assessment of cluster tendency (VAT) approach includes generating ordered dissimilarity matrices of the data sets which can then be displayed as an intensity image (ODI). Solid blocks of similar intensity (or colour) along the diagonals of the ODI indicate clusters in the data [3].



(a) Random data (b) Numerical data (c) Mixed data

Figure 1: The ordered dissimilarity images (ODI) for the abalone data

The ODI's of the numerical data set and mixed data set were generated [4] and are shown in fig.1. As expected, the ODI of the randomly generated data set is noisy without any clear intensity blocks. In contrast, both ODI's of the numerical and mixed data sets contains more distinct intensity blocks. However, the intensity blocks of the numerical data set appear to be more apparent than that of the mixed data set. This implies[2] a greater number of underlying clusters that are less distinct. The Hopkins statistic and VAT have strongly indicated that both data sets have underlying clusters.

[1]Column bind of numerical features with binary encoding for each sex level

[2]Potentially, but evident in retrospect

3 Data Clustering

This section includes a detailed discussion on feature selection, because it essentially dictates what possible clustering algorithms are available for use and consequently clustering performance. Furthermore, this section includes a description of the clustering algorithms considered and the final model selected. The clustering algorithm is optimized for finding the best number of clusters and other model hyper-parameters in respect to the validation metrics. Note that emphasis has been placed on the feature selection, the reasoning for this is to rather select the correct feature from the start than removing them after the clustering had been performed. Therefore, the analysis sections do not provide a lot of detail on feature discarding.

3.1 Feature selection

Generally, a principle component analysis (PCA) is useful to investigate the variance between features. However, PCA is only compatible with quantitative data and the abalone data set contains both quantitative and qualitative. Therefore, all the descriptive features are investigated by conducting a factor analysis of mixed data (FAMD). This method considers the associations between both type of variables, and can be seen as a combination of a PCA and a multiple correspondence analysis [5]. The variance explained by the principle components and individual feature contributions to the principal components are generated by the FAMD[4, 6], and is illustrated in fig.2.

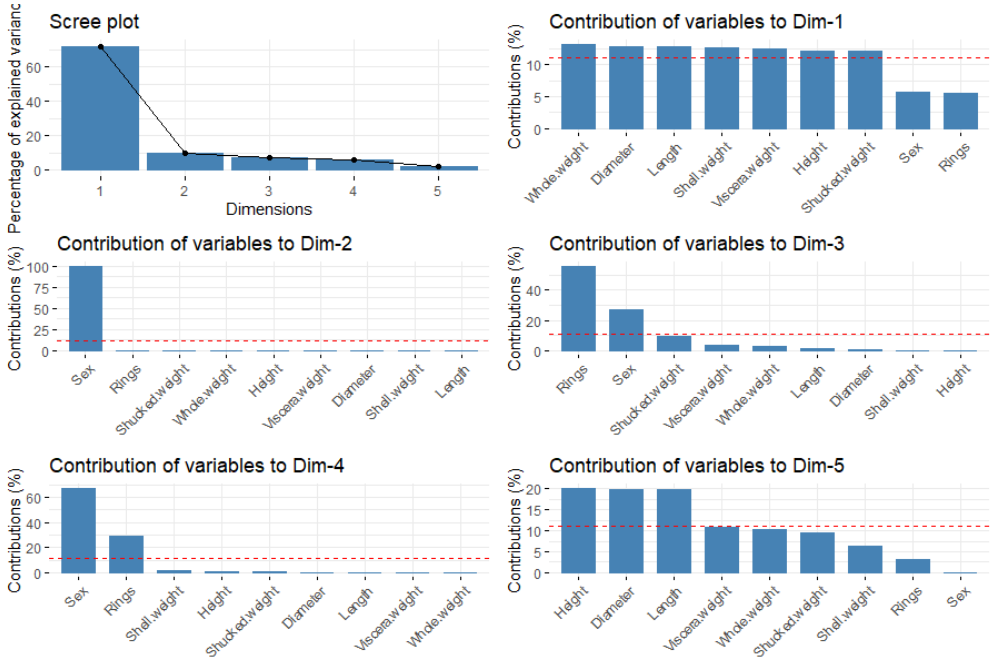


Figure 2: Factor analysis of mixed data scree plot (upper left) and contributions to principal components

The FAMD shows that the first principle component accounts for a large proportion of the variance (72%) compared to the remaining components (less than 10% each). The first

principal component is dominated by the general size of the abalone as demonstrated by the dimensional and weight features contributing the most. The sex feature is less represented on the first component, however, it dominates the second component. Evidently, the sex feature influences the variance and can therefore not be discarded without further consideration. The FAMD does not provide any clear insights as to which numerical features to consider for selection. In combination with an investigation of variance, the correlation between the features are also considered.

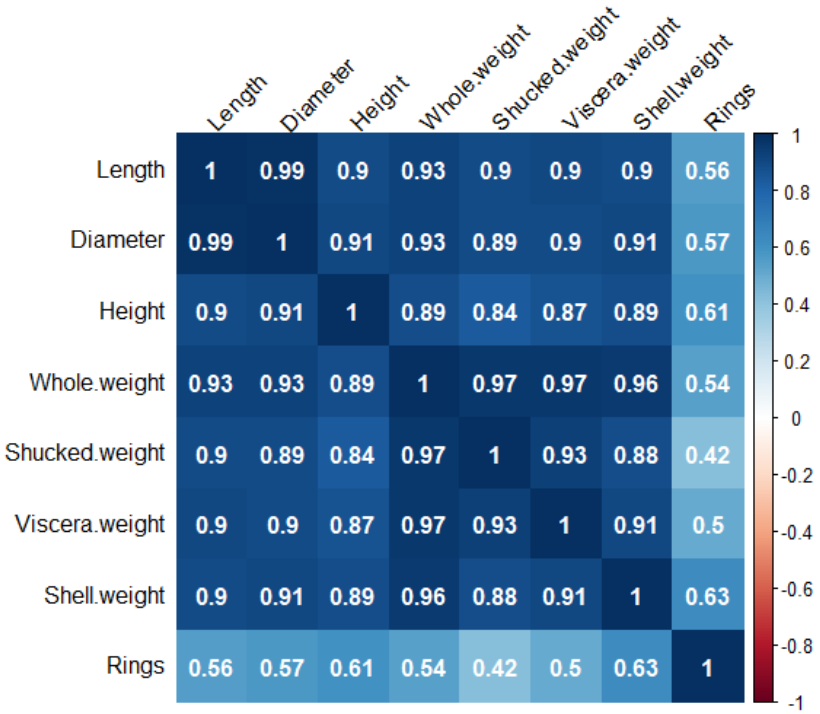


Figure 3: Pearson's correlation matrix for numerical features

The Pearson correlation plot for the numerical predictors is illustrated in fig.3 and shows that the dimensional and weight features are all highly correlated. This indicates possible multicolinearity issues that could potentially affect the clustering process along the line [7]. The matrix also shows that neither numerical features significantly differ in respect to the target feature, which complicates the selection process. Pearson's corrected contingency correlation coefficient [8] for the sex and rings feature is 0.5996, which indicates some correlation and a reason to keep the categorical feature.

The features to be selected are not very clear, and therefore, a more statistically rigorous[3] approach was taken by implementing a hybrid feature search algorithm [9]. The algorithms implements a linear consistency-constrained (LCC) algorithm which firstly evaluates features

[3]maybe too rigorous, and very computationally expensive

individually with the Pearson determination coefficient, and then all the features conjunctively with a relief set measure. The algorithm output has shown that all features (quantitative and qualitative) are to be selected, which is also not very helpful. Including all the features in the clustering may lead to problems because of the multicolinearity. As a last and more intuitive strategy, the top three numerical features with the greatest Pearson correlations coefficients in respect to the target are chosen with the categorical feature. These are:

- Shell weight
- Diameter
- Height
- Sex (male,female,infant)

3.2 Clustering algorithm

The clustering algorithm must be able to handle mixed data in accordance with the features selected. The k-prototypes [10, 11] algorithm [4] was initially considered, however, the model produced an unsatisfactory optimal number of 2 clusters, which most likely corresponds to the adult and infant clusters. The k-mediods algorithm[5] was also considered, however, then the categorical variable needed to be discarded.

The clustering algorithm selected is a model-based approach which has been proposed by McParland and Gormley (2016) [12] and is implemented in the clustMD R package [13]. The algorithm is a Gaussian mixture based model that handle nominal features as categorical manifestations of latent multivariate variables and ordinal features as categorical manifestations of latent univariate Gaussian variables. The mixture of Gaussian distributions enables a latent variable to generate the observed data of mixed type, which can be any combinations of continuous, nominal, or ordinal variables. Under the hood, the clustMD is estimated by an expectation maximization (EM) algorithm and in the case of a nominal variable (sex feature) a Monte Carlo EM algorithm is used.

3.3 Data preparation

The abalone dataset provided has been cleaned in respect to missing values and other entry errors. However, the abalone data set set contains a significant number of outliers, which could impair the clustering process. Among these outliers are 2 observations that have invalid outliers for the height feature with a 28.3 and 12.9 IQR factor[6]. Consequently, these two instances are removed from the data set. The The clustMD algorithm is relatively robust to general data quality issues, however, numerical features are transformed with Z-score

[4] iterates in a manner similar to the k-means algorithm where for the numeric variables the mean and the categorical variables the mode minimizes the total within cluster distance

[5] similar to k-means and robust to outliers

[6] outlier defined by an inter quartile range (IQR) factor greater than 1.5

normalisation or standardization as an additional fail-safe and since the numerical features have different units. Furthermore, the outliers are rather kept than discarded to conserve as much information as possible. As for the categorical feature sex, clustMD algorithm requires the nominal feature to be encoded by integers 1, 2, and 3.

3.4 Clustering and optimisation

The clustMD package provides an additional algorithm, clustMDparallel, that enables the user to run a full factorial design of the clustMD algorithm for different k values and different models[7]. The k-values to be investigated were 15 (which excludes k equal to 1) and 6 different models which adds up to 90 different candidate clustering alternatives. The hyperparameters of clustMDparallel algorithm were set and adjusted according to guidelines [15]. The default algorithm to start the clustMDparallel algorithm in the EM step is hierarchical, however, this is not a preferred method since it is computationally expensive due to the size of the data set. Therefore, a simple k-means clustering algorithm is selected to initialise the clustMD expectation maximisation (MCEM) algorithm. The maximum number of iterations in the MCEM is specified as 100 in combination with an auto-stop feature, that is based on a moving average of the approximated log likelihood values, which terminates the program when converged. The algorithm took approximately 18 min to converge.

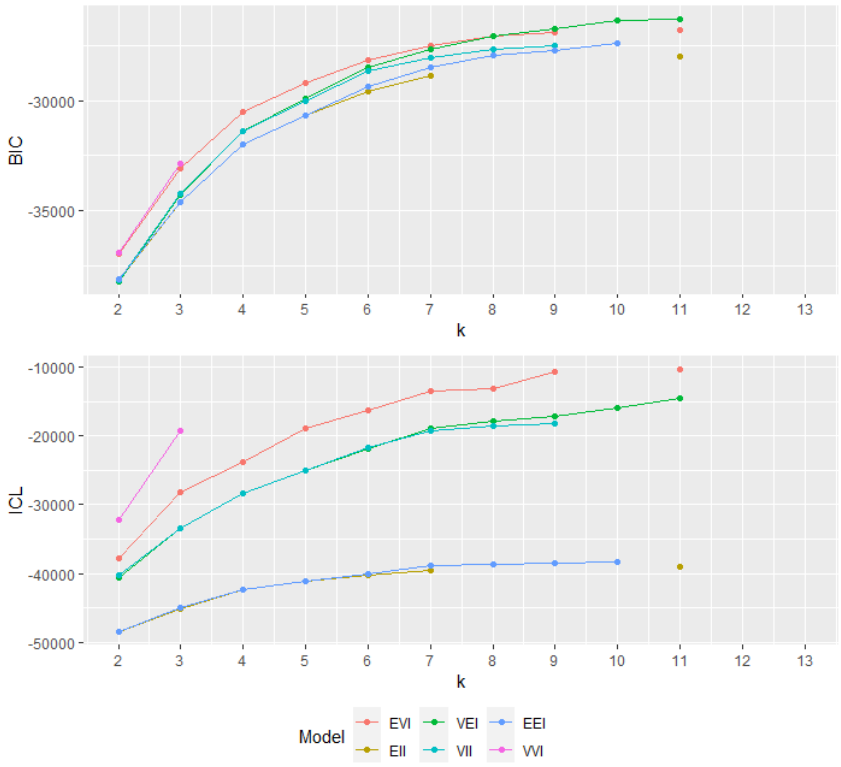


Figure 4: Clustering models for different numbers of k and hyperparameters

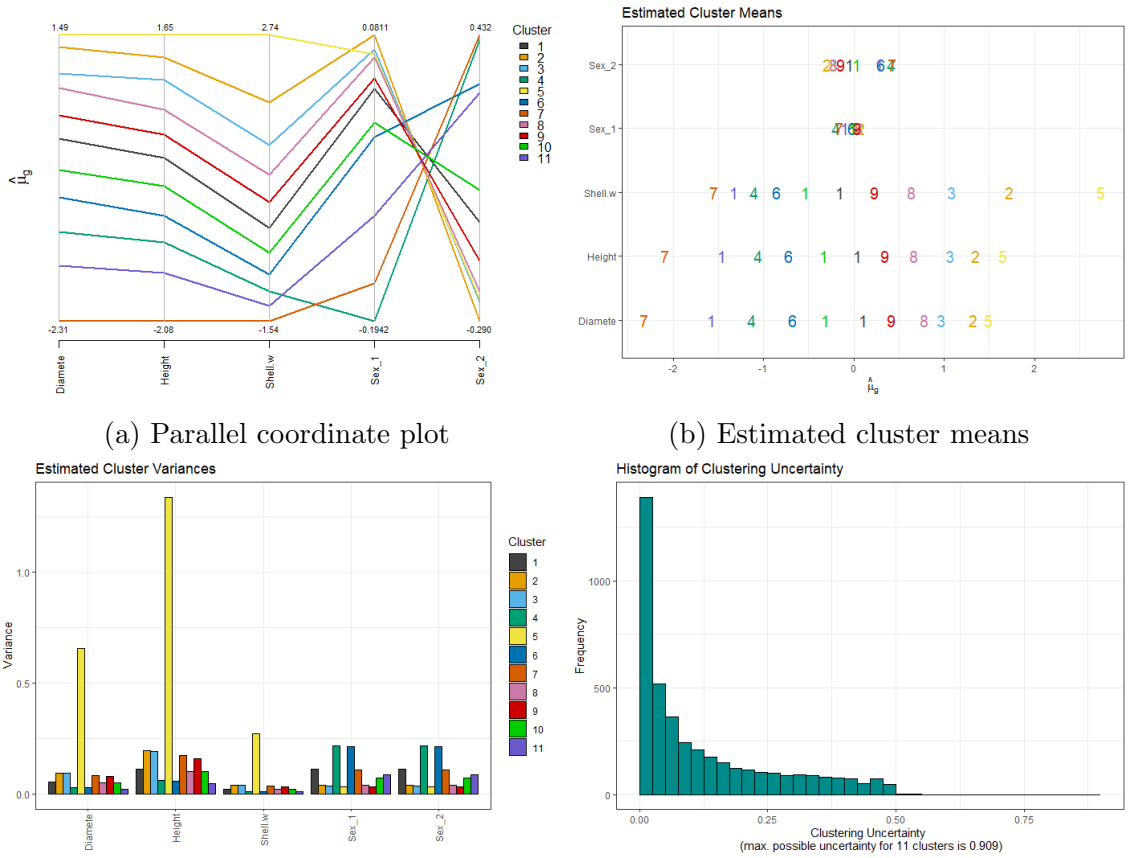
The full-factorial design is shown in fig.4 above and the hyperparameters are provided in

[7]Covariance matrix structure with different degrees of parsimony or hyperparameters[14]

respect to the Bayesian information criterion (BIC) and the integrated complete-data likelihood (ICL). Gaussian mixture methods (GMM) can fit clusters with arbitrarily shaped blobs or ellipses, and this in therory, allows the GMM to fit any number of complex clusters since they are superimposed. Therefore, the BIC is imposed as a regularisation technique to penalise the model for fitting too many Gaussians. The user should favour the ICL over the BIC into finding well separated classes, and conversely the BIC over the ICL when overlapping clusters are not a major concern [16]. In this case, the BIC criterion is favored over the ICL because the data is clearly not well separated in the abalone dataset. The optimal model is identified with a VEI covariance matrix structure with 11 clusters.

3.5 Clustering results

The output plots of the clustMD algorithm for the optimal model is shown below in fig.5, which provides information regarding the separation, cluster means, cluster variance, and the clustering uncertainty. Note that these results will be discussed in the following sections.



(a) Parallel coordinate plot

(b) Estimated cluster means

(c) Estimated cluster variance

(d) Clustering uncertainty

Figure 5: clustMD algorithm clustering output

3.6 Descriptive statistics

The descriptive statistics, which includes tables for the numerical and categorical features in table 1 to table 22, the cluster centroids in table 23, and cluster modes in table 24. The number of instances that belong to each cluster is specified in the categorical tables under the column N for each cluster. Note that, these results will be discussed by using visualizations in the following sections.

Table 1: Cluster 1 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.34	0.405	0.418	0.42	0.435	0.47	0	0	6
Height	0.095	0.135	0.141	0.14	0.15	0.185	0	0	13
Shell.weight	0.15	0.205	0.218	0.218	0.234	0.262	0	0	2

Table 2: Cluster 1 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	F	494	169	34.21	1
Sex	M	494	168	34.01	2
Sex	I	494	157	31.78	3

Table 3: Cluster 2 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.465	0.525	0.541	0.545	0.56	0.605	0	0	3
Height	0.155	0.18	0.191	0.19	0.2	0.24	0	0	1
Shell.weight	0.415	0.455	0.481	0.476	0.502	0.555	0	0	0

Table 4: Cluster 2 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	F	260	130	50	1
Sex	M	260	128	49.23	2
Sex	I	260	2	0.77	3

Table 5: Cluster 3 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.37	0.49	0.505	0.51	0.525	0.605	0	0	16
Height	0.125	0.17	0.181	0.18	0.19	0.235	0	0	14
Shell.weight	0.305	0.372	0.392	0.39	0.41	0.452	0	0	1

Table 6: Cluster 3 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	F	432	212	49.07	1
Sex	M	432	210	48.61	2
Sex	I	432	10	2.31	3

Table 7: Cluster 4 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.24	0.285	0.295	0.295	0.31	0.33	0	0	2
Height	0.065	0.09	0.099	0.1	0.105	0.14	0	0	6
Shell.weight	0.06	0.075	0.085	0.085	0.093	0.15	0	0	3

Table 8: Cluster 4 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	I	308	226	73.38	1
Sex	M	308	57	18.51	2
Sex	F	308	25	8.12	3

Table 9: Cluster 5 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.23	0.536	0.557	0.565	0.585	0.65	0	0	2
Height	0	0.195	0.204	0.21	0.22	0.25	1	0	2
Shell.weight	0.288	0.578	0.628	0.607	0.664	1.005	0	0	10

Table 10: Cluster 5 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	M	94	47	50	1
Sex	F	94	45	47.87	2
Sex	I	94	2	2.13	3

Table 11: Cluster 6 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.28	0.33	0.34	0.34	0.35	0.38	0	0	5
Height	0.08	0.105	0.111	0.11	0.115	0.14	0	0	11
Shell.weight	0.088	0.11	0.12	0.12	0.13	0.155	0	0	0

Table 12: Cluster 6 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	I	349	217	62.18	1
Sex	F	349	69	19.77	2
Sex	M	349	63	18.05	3

Table 13: Cluster 7 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.055	0.15	0.178	0.19	0.205	0.34	0	0	2
Height	0	0.05	0.059	0.06	0.07	0.1	1	0	3
Shell.weight	0.002	0.014	0.024	0.025	0.032	0.115	0	0	2

Table 14: Cluster 7 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	I	217	183	84.33	1
Sex	M	217	32	14.75	2
Sex	F	217	2	0.92	3

Table 15: Cluster 8 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.42	0.475	0.486	0.485	0.5	0.55	0	0	11
Height	0.125	0.155	0.165	0.165	0.175	0.205	0	0	1
Shell.weight	0.286	0.315	0.329	0.33	0.345	0.37	0	0	0

Table 16: Cluster 8 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	M	488	239	48.98	1
Sex	F	488	220	45.08	2
Sex	I	488	29	5.94	3

Table 17: Cluster 9 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.345	0.435	0.45	0.45	0.469	0.565	0	0	14
Height	0.08	0.145	0.153	0.15	0.16	0.215	0	0	48
Shell.weight	0.162	0.255	0.272	0.273	0.288	0.355	0	0	9

Table 18: Cluster 9 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	M	714	352	49.3	1
Sex	F	714	281	39.36	2
Sex	I	714	81	11.34	3

Table 19: Cluster 10 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.27	0.365	0.377	0.38	0.39	0.465	0	0	14
Height	0.08	0.12	0.126	0.125	0.135	0.18	0	0	12
Shell.weight	0.105	0.15	0.165	0.165	0.179	0.211	0	0	1

Table 20: Cluster 10 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	I	572	250	43.71	1
Sex	M	572	186	32.52	2
Sex	F	572	136	23.78	3

Table 21: Cluster 11 descriptive statistics for numerical predictors

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Diameter	0.215	0.24	0.252	0.255	0.265	0.3	0	0	0
Height	0.065	0.075	0.083	0.085	0.09	0.105	0	0	0
Shell.weight	0.032	0.046	0.055	0.055	0.062	0.09	0	0	2

Table 22: Cluster 11 descriptive statistics for categorical predictors

variables	levels	N	freq	ratio	rank
Sex	I	247	185	74.9	1
Sex	M	247	45	18.22	2
Sex	F	247	17	6.88	3

Table 23: Cluster centroids (standardized mean)

Cluster	Diameter	Height	Shell.weight	Male-Female	Female-Infant
1	0.112	0.048	-0.147	0.029	-0.041
2	1.332	1.353	1.724	0.081	-0.29
3	0.971	1.066	1.088	0.067	-0.241
4	-1.129	-1.055	-1.102	-0.194	0.419
5	1.493	1.653	2.739	0.062	-0.234
6	-0.674	-0.716	-0.848	-0.018	0.308
7	-2.314	-2.085	-1.543	-0.158	0.432
8	0.782	0.67	0.644	0.059	-0.216
9	0.424	0.35	0.236	0.039	-0.139
10	-0.306	-0.329	-0.529	-0.003	0.041
11	-1.574	-1.456	-1.318	-0.093	0.285

Table 24: modes

Cluster	Mode
1	9
2	11
3	11
4	7
5	12
6	7
7	5
8	10
9	9
10	8
11	6

4 Cluster visualization and analysis

This section includes visualizations pertaining to the clustering solution with an analysis thereof. The visualizations and analytics are provided for a cluster PCA plot, descriptive plots, a scatter plot matrix (SPLOM), feature histograms, and a parallel coordinate plot. Note that, the analytics are provided with the visualizations to spare the reader from referring back to previous sections.

4.1 Cluster plot

The cluster plots in respect to the principal components are provided in fig.6 below and illustrates the clusters formed. Furthermore, the plots are separated in terms of the categorical feature, and shows the cluster plots for the male, female, and infant features. The surrounding shaded areas are ellipses[8] drawn around the individual clusters and assumes a multivariate normal distribution.

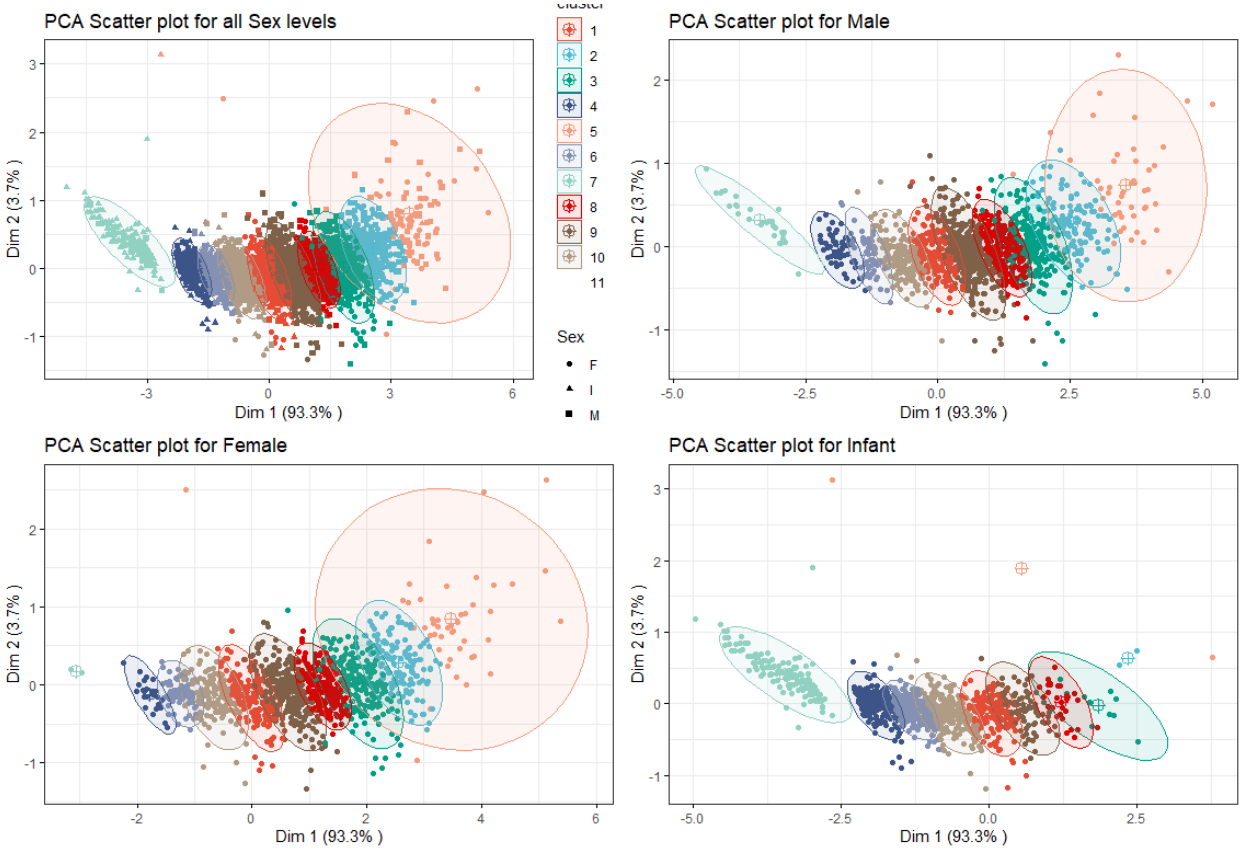


Figure 6: Cluster scatter plot on principal components for all categorical predictors

These highlighted ellipses seem to be fuzzy, however, underneath the highlighted ellipses the clusters are decently separated. The overlapping cluster 5 region and size thereof, is most likely due to the outliers in the data. This cluster includes mostly adult abalone, because it

[8]which are generally characterized by Gaussian mixture models

is practically non-existent in the infant plot. The initial abalone EDA report did not find any major distinctions between the male and the female classes. It is therefore interesting that cluster 7, which mainly shares properties with infant abalone, is very small in the female plot but of similar size in the male plot. Furthermore, cluster 2 on the infant plot is practically non-existent, which indicates that this cluster also pertains to the adult abalone. As expected, the range on the first principal component of the infant plot is less than that for the male and female plot. Most of the ellipses are orientated vertically mainly because of the dominating first principle component.

4.2 Descriptive statistics visualization

This section includes a discussion on the rules that can be extracted from the descriptive statistics for both numerical and categorical features with the aid of visualizations.

4.2.1 Numerical features

The relationship between the numerical descriptors and the target feature is compared with the clusters in fig.7 below. The order of the rankings[9] are similar for all plots, which makes it very difficult to extract specific rules. Cluster 7 is distinct from the other clusters as shown in the two upper plots, and this cluster has the lowest mean of rings as seen on the bottom right plot. The shell weight cluster ranges[10] are much smaller than the other features, and is most likely due to the outliers in cluster 5.

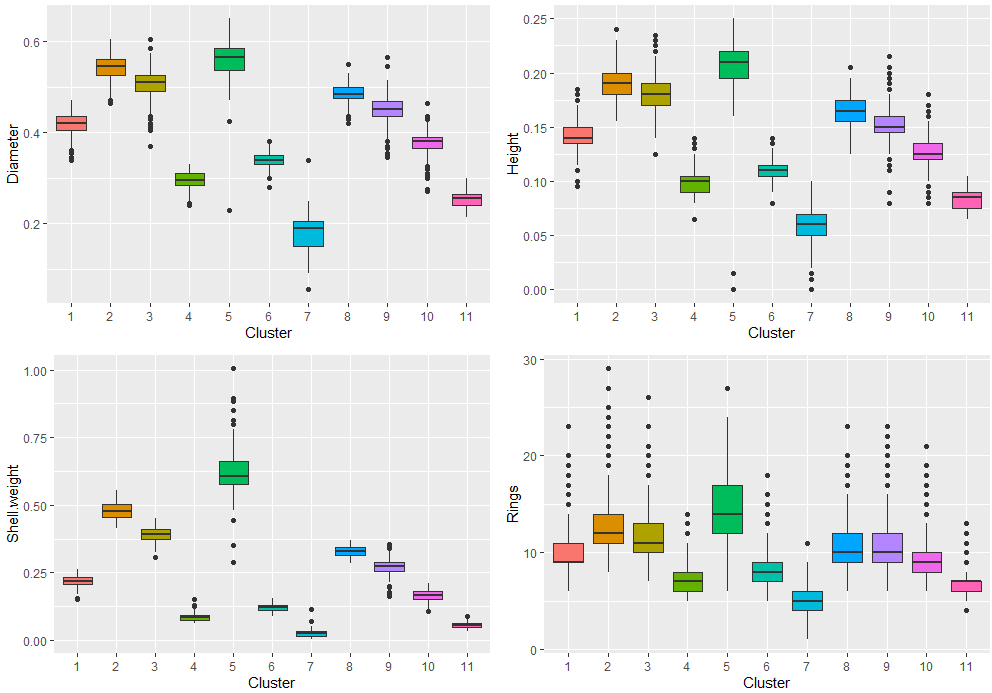


Figure 7: Box plots for numerical predictors and target feature

[9] measure on the y-axis for the respective plot
[10] from quartile 1 to 3

4.2.2 Categorical features

The categorical feature and the target feature is compared to the clusters in fig.8 below with a jitter plot and offers somewhat useful insights than that of the previous plot. Cluster 7 has a high concentration of infants, with a small number of rings indicated by the orange colour (1-5 rings), compared with the corresponding male and female. Moreover, this cluster contains less females than male, which indicate that the female and male features are not equivalent, that is, they can not necessarily be grouped together as adult. Cluster 2 and 3 practically does not contain any infants and is therefore and adult group. Cluster 4 also predominantly consist of infants, as the region is more concentrated than that of the adults. This cluster has more rings (7-11) than the other infant cluster 7. Cluster 5 is another adult cluster with the same number of rings as cluster 2, but the adults are less densely populated.

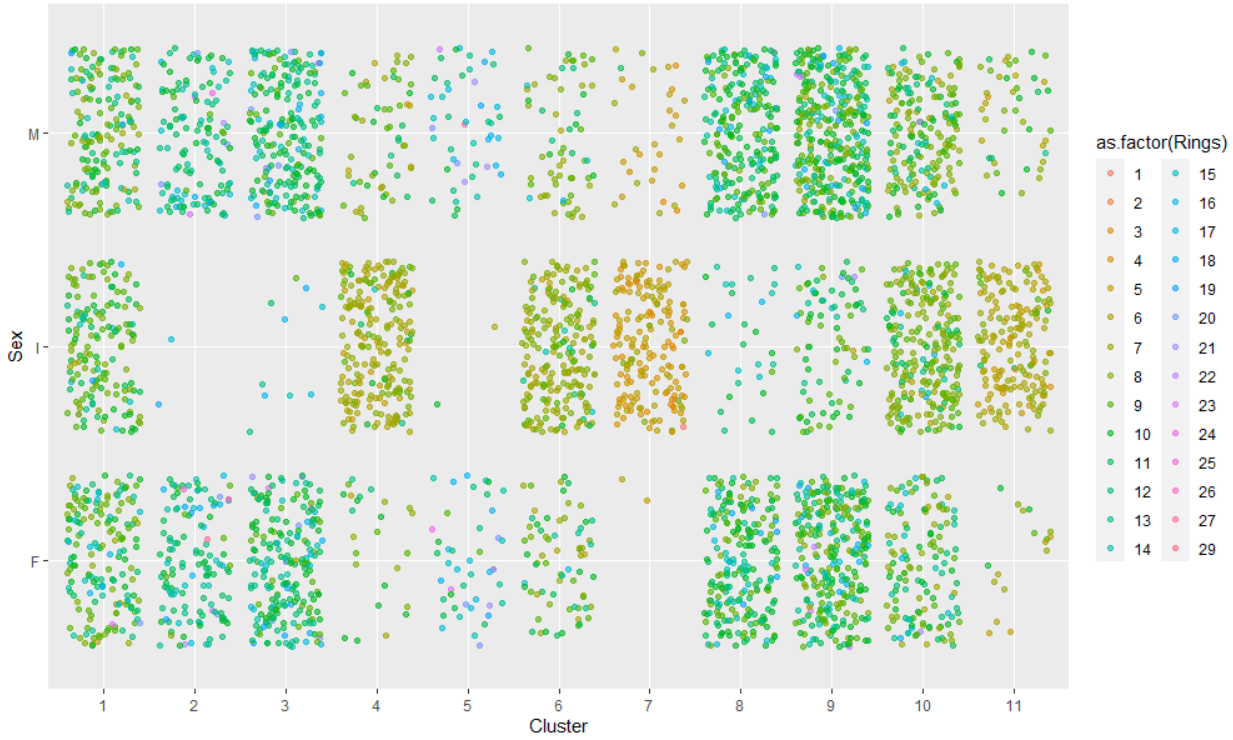


Figure 8: Jitter plot for categorical predictors and target feature

Overall the female and male axis are nearly symmetrical to each other, except in cluster 7 and cluster 11. Similarly, the sex for all three categories are nearly symmetrical in clusters 1 and 10. As mentioned above, it is still difficult to discern rules in regards to the number of rings, because they all with the target variable. This will be a recurring issue in the remainder of this report.

4.2.3 Cluster centroids

The cluster centroids on standardized scale [11] is provided below in fig.23 and illustrates the ranking of the numerical features. For the categorical feature centroid plot, the top left

[11] same as input to the clustering model

cluster centroids represent the infant groups, the bottom right represents the adult groups, and those in the middle represent abalone adults and infants who have similar numerical features. These can be verified by referring back to the previous section. It is noteworthy that cluster 6 is further away from the middle line, which indicates some distinct aspect of the cluster.

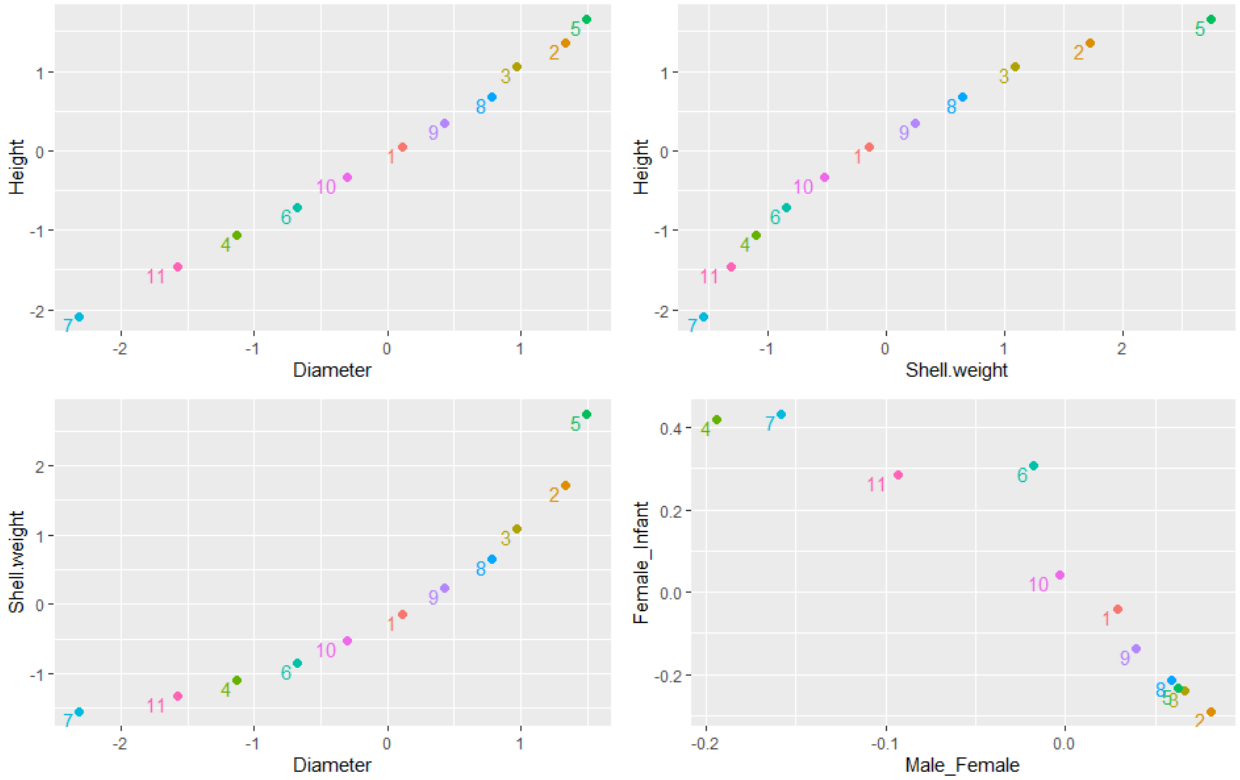


Figure 9: Cluster centroids scatter plot matrix

4.3 Scatter plot matrix

The scatter plot matrix (SPLOM) is illustrated below in fig.10 and provides the relationship between all descriptors, the target variable, and the cluster it belongs to. The SPLOM shows that the clusters are mainly formed along increasing along the numerical features of the data. The smallest dimensions and weights belong to cluster 7 and 11, which are also dominated by infant abalones. In contrast the greatest dimensions and weights belong to clusters 5 and 2, which are also dominated by the infant abalone. The SPLOM does, however, provide a new insight into how the clusters are separated. The clusters are separated diagonally[12] on the dimensional (height and diameter) plots, whereas the clusters are separated more perpendicular to the axis of the shell weight in respect to the dimensional features. This indicates that the clustering favours the shell weight feature more than that of the dimensional ones. However, it is not very clear whether one of the features can be discarded, possibly only one of the two dimensional features, but not the rest.

[12]perpendicular to the a linear regression fit of the data

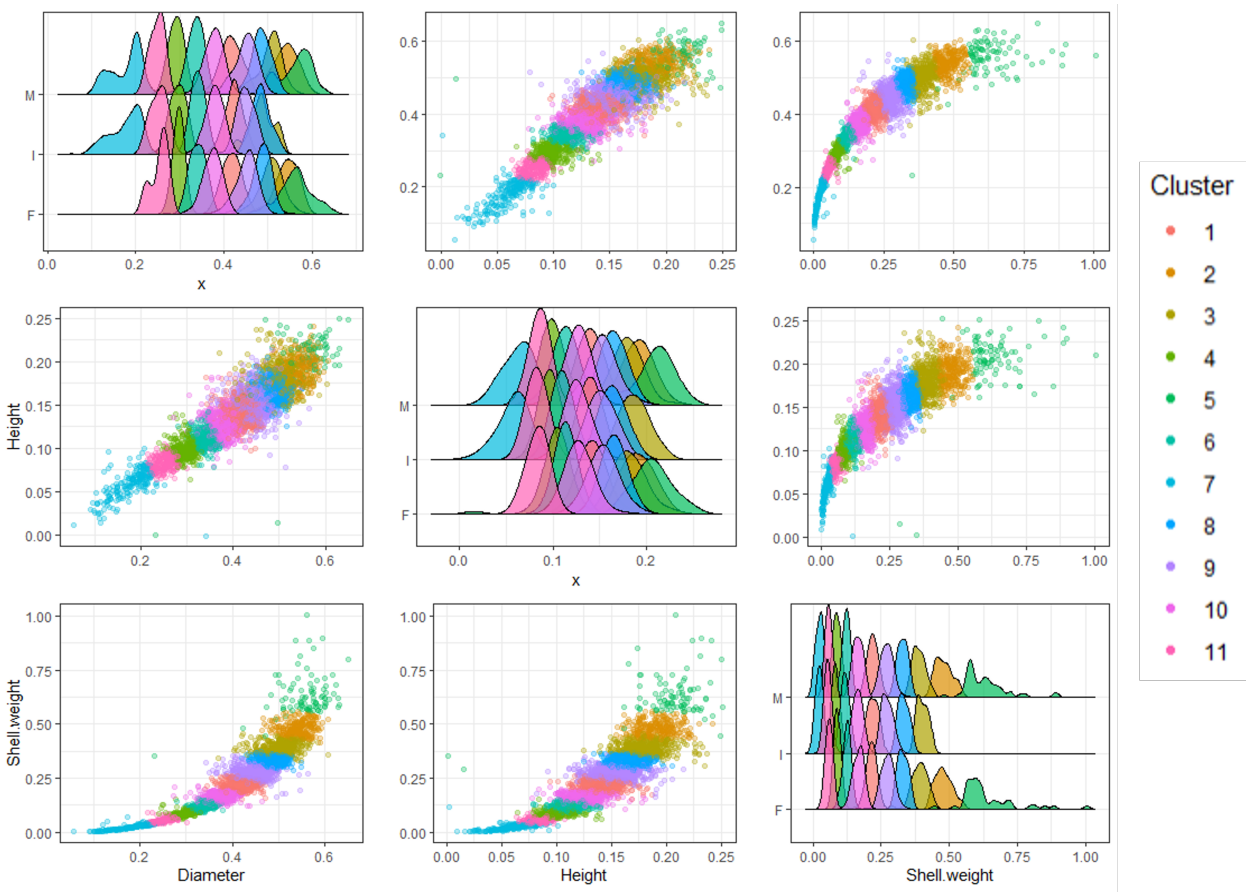


Figure 10: Scatter plot matrix for predictors

It is difficult to extract rules, however in general:

- From the upper left plot, if the diameter is less than 0.20, then it belongs to cluster 7 and is either a male or a infant, which has 1-5 rings.
- From the middle plot, if the height is less than 0.05, then it belongs to cluster 7 and is either a male or a infant, which has 1-5 rings.
- From the lower right plot, if the shell weight is greater than 0.5, it either belongs to cluster 2 or 5 and is an adult, which has a wide range of rings, but generally 10 - 15.

4.4 Feature histograms

This section provides an extensive number of histogram plots pertaining to the descriptive and target features in respect to the clusters they belong to. The composition of the rings feature with respect to the clusters is illustrated in the histogram in fig.11 below. The plot gives an indication of what proportion of number rings falls within a given cluster. For example, the most densely populated class with a number of 10 rings, is mainly composed of cluster 9, 8 and 3. Furthermore, the plot allows to quickly check which number of rings that occur the most in each cluster. These modes are concentrated in the middle between a number 5 and 12 rings (also showed in table 24.

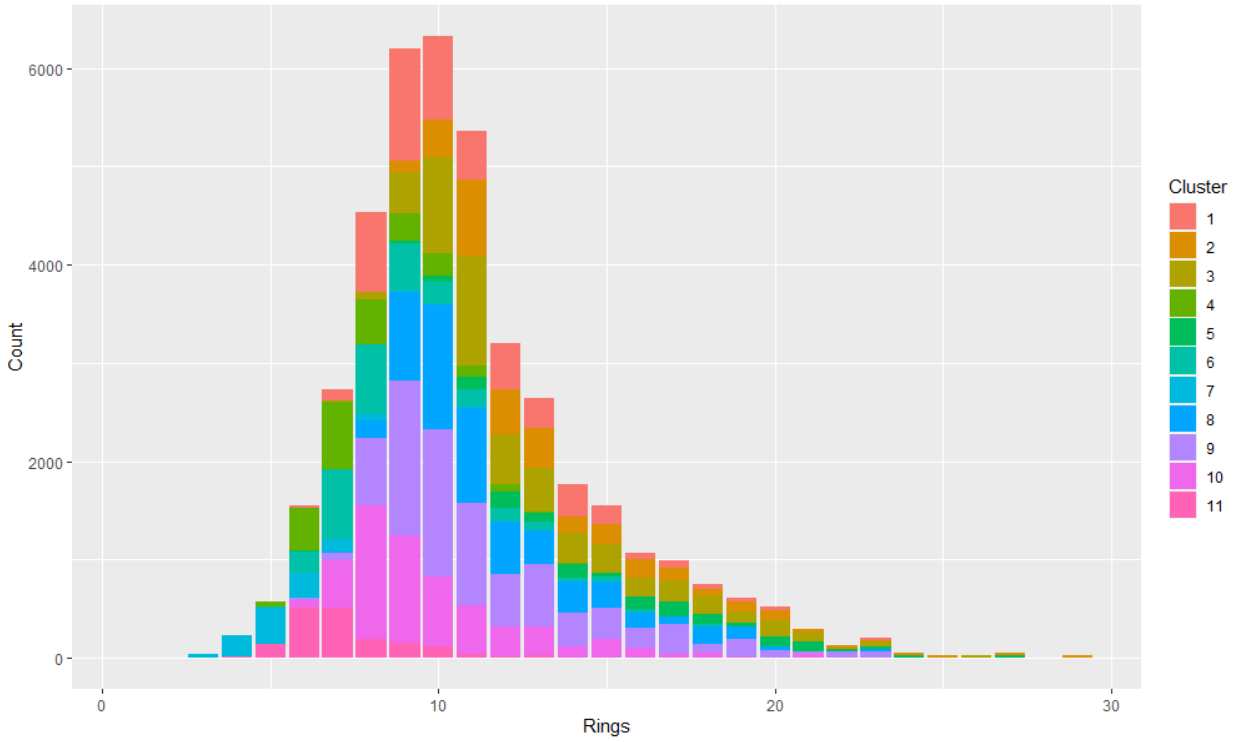


Figure 11: Stacked histogram grouped by cluster

The histograms for the height feature for each individual cluster with respect to the composition of the number of rings are provided below. The plot for all clusters in fig.12 colour scale demonstrates the growth of the abalone where it starts at orange (small number of rings) and develops to dark green (greater number of rings). Cluster 1 occupies the centre of the height feature with a relatively low variance and composed of large range of rings. Cluster 2 and 3 in fig.13 is similar range of rings and variance, although cluster 3 contains more instances. Cluster 4 in fig.14 has a relatively small range of rings, whereas cluster 5 has a large range of rings and high variance. Cluster 5 is also has the biggest height dimensions compared to other clusters. Cluster 6 in 15 has relatively high variance with a medium sized ring range. Cluster 7 has the lowest height dimensions compared to other clusters and a degree with concentrated number of rings between 4 and 6. Cluster 10 in fig.17 has a relatively low variance with a concentration of 6 to 10 rings, which indicate infant abalone.

The histogram plots for the diameter feature, shown in fig.18 to 23, are very similar to that of the height feature just discussed. Therefore, removing one of these features is a possibility as mentioned before. The histogram plots for the shell weight feature, shown in fig.24 to 29. Cluster 1 has a very low variance around 0.24 and a mostly dominated by a smaller number of rings. Cluster 2 25 only consists of adult abalone as indicated by the heavy weight, however, it still has range containing a low number of rings. Similarly, cluster 5 26 is dominated by adult abalone, but they have a greater average number of rings. In contrast, cluster 7 and 11, illustrated in fig.27 and 29, respectively, belongs to the infant abalone with lighter weights. However, the infants in cluster 7 have somewhat more rings than those in cluster 11 as illustrated by the darker green blue colour scale.

4.4.1 Height predictor

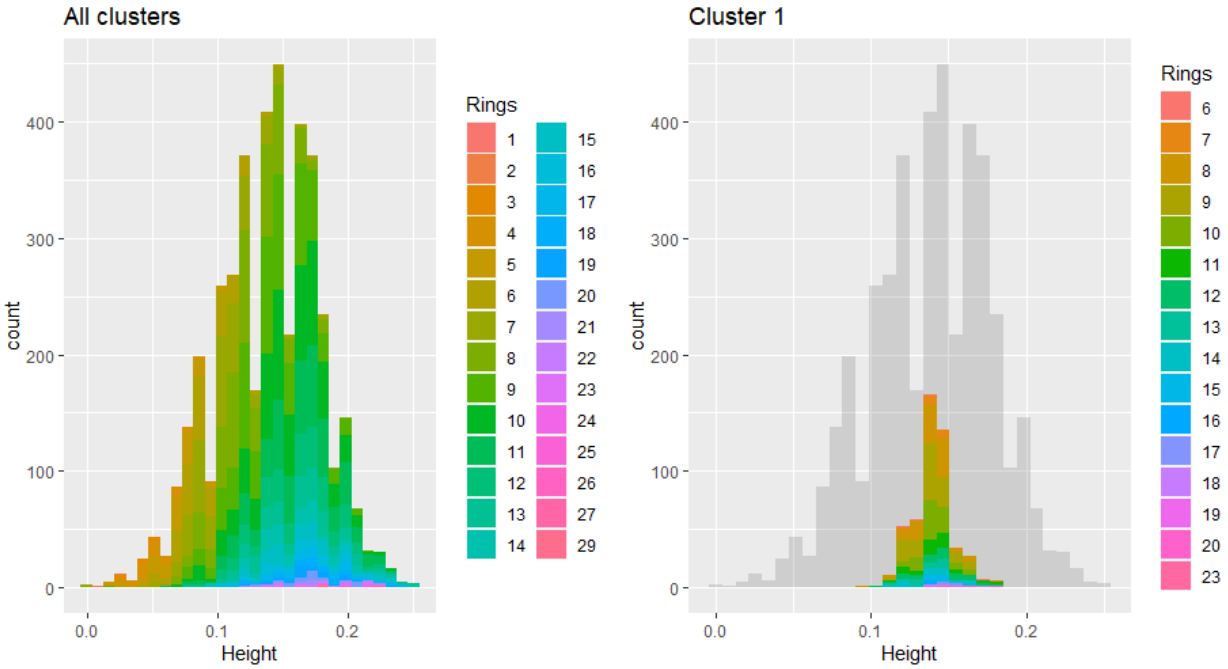


Figure 12: Histogram for height feature of all clusters and cluster 1

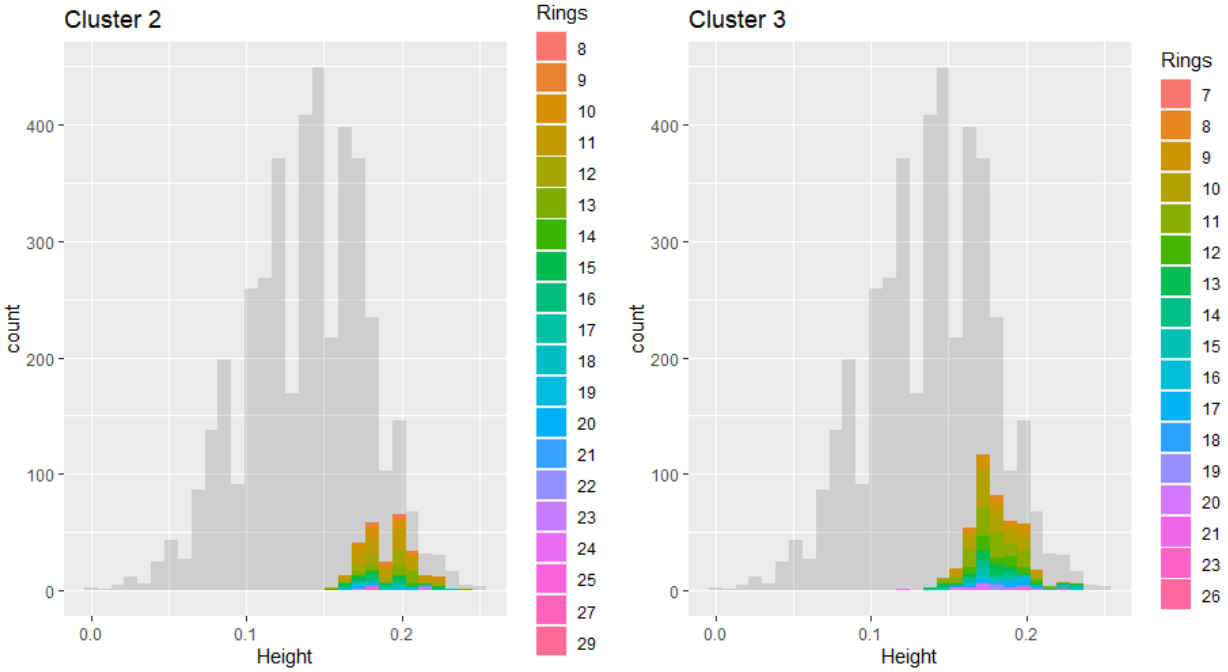


Figure 13: Histogram for height feature of cluster 2 and 3

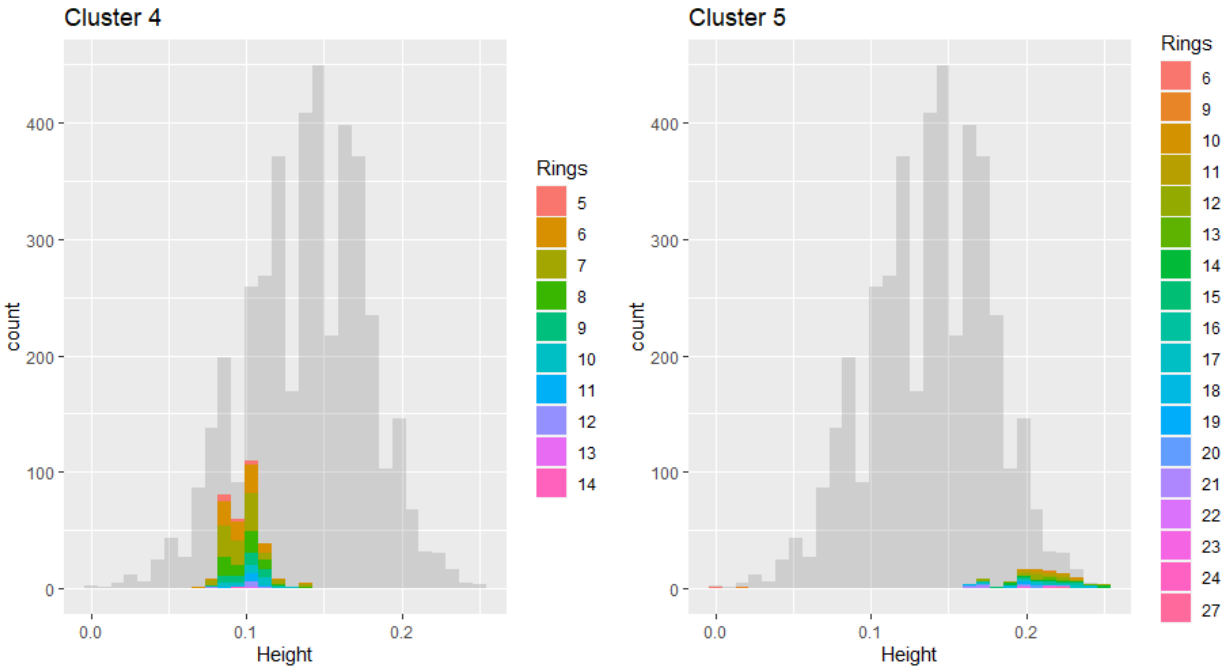


Figure 14: Histogram for height feature of cluster 4 and 5

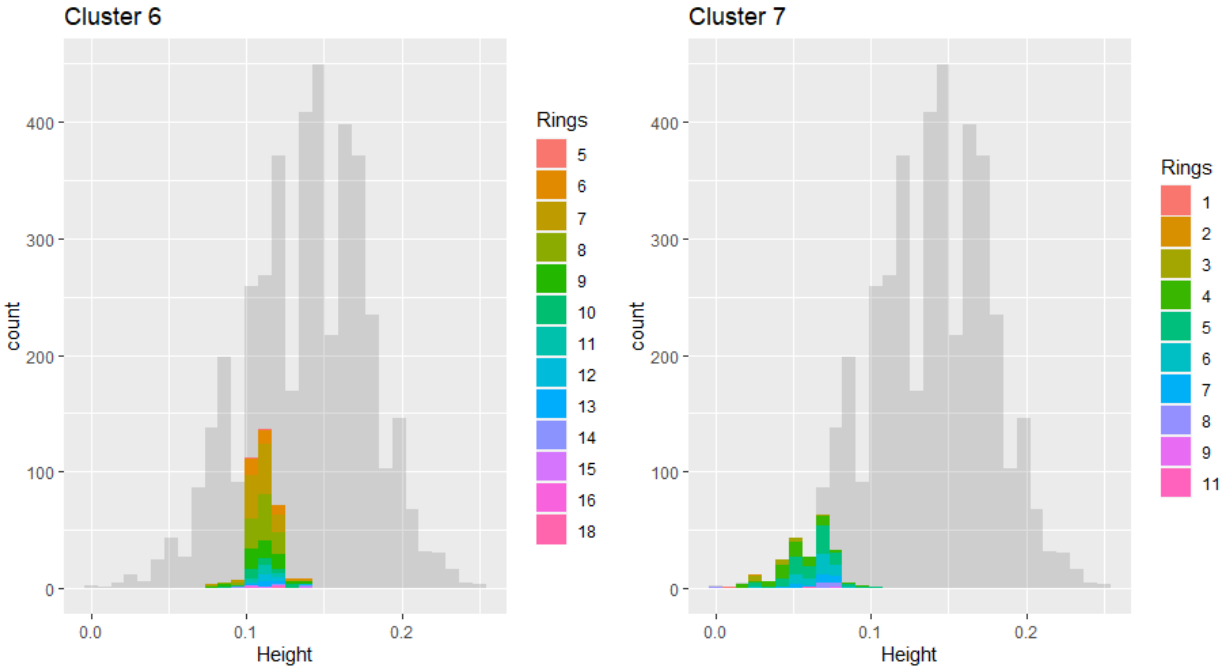


Figure 15: Histogram for height feature of cluster 6 and 7

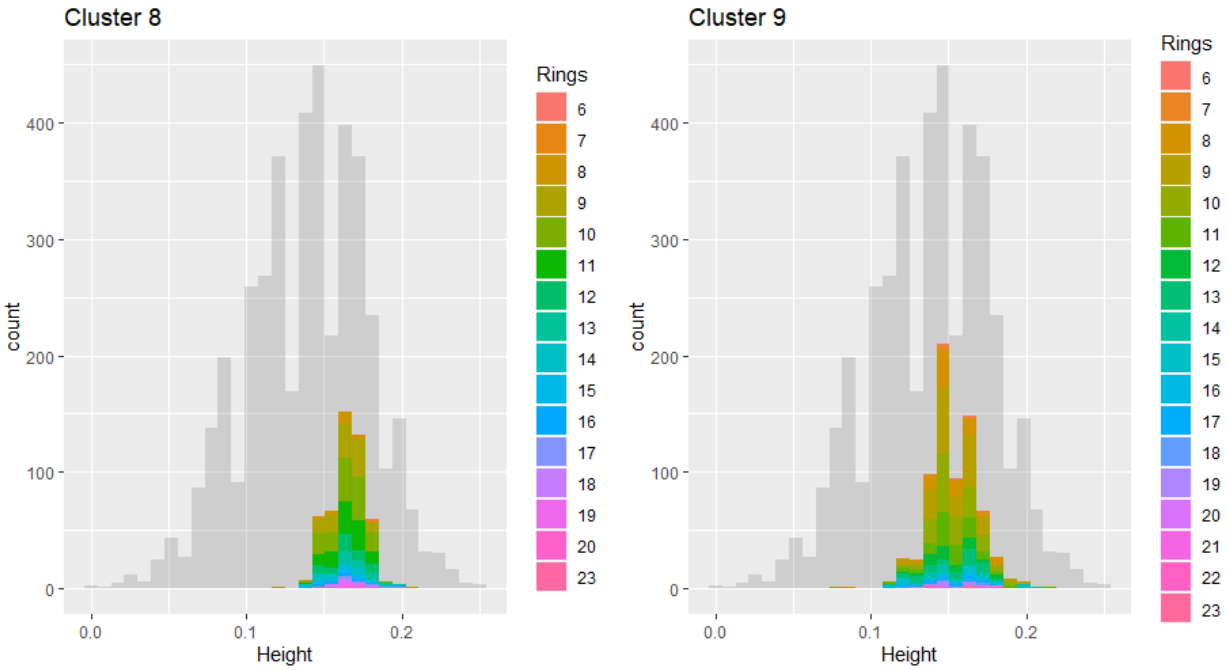


Figure 16: Histogram for height feature of cluster 8 and 9

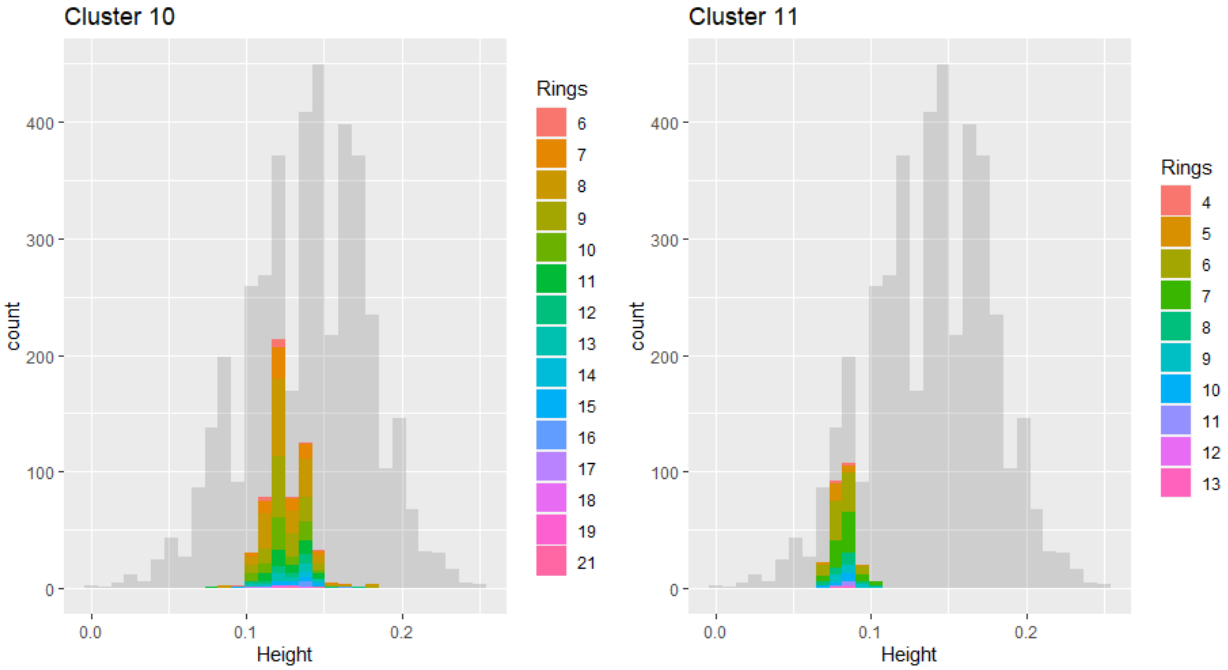


Figure 17: Histogram for height feature of cluster 10 and 11

4.4.2 Diameter predictor

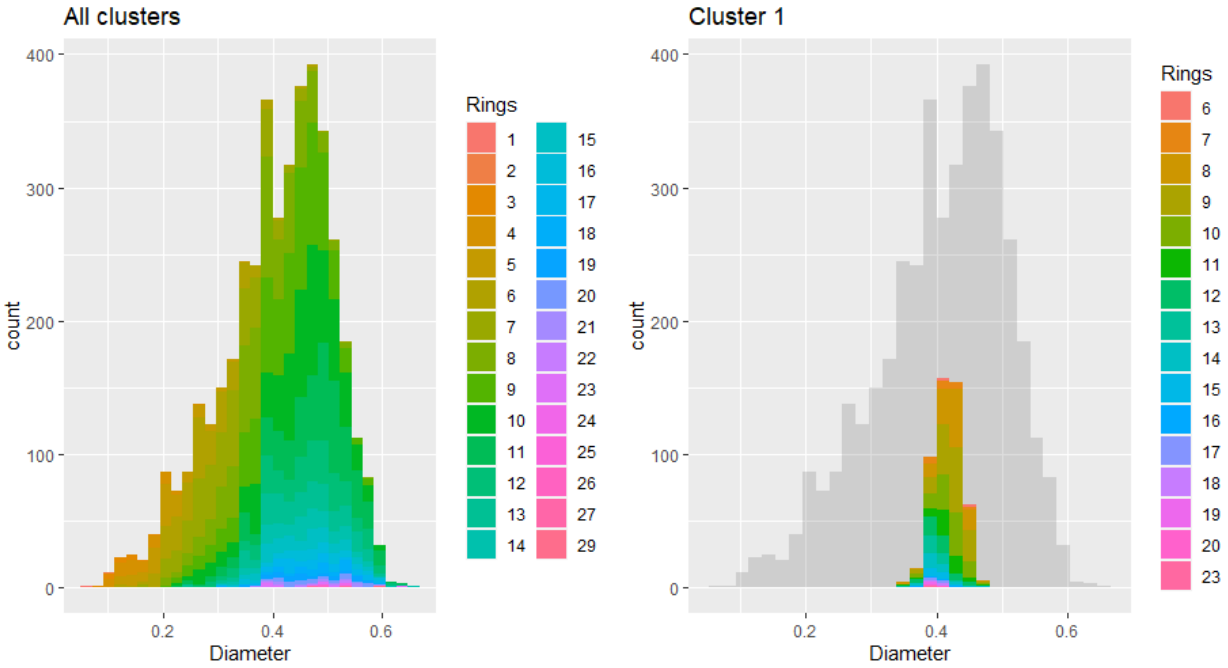


Figure 18: Histogram for diameter feature of all clusters and cluster 1

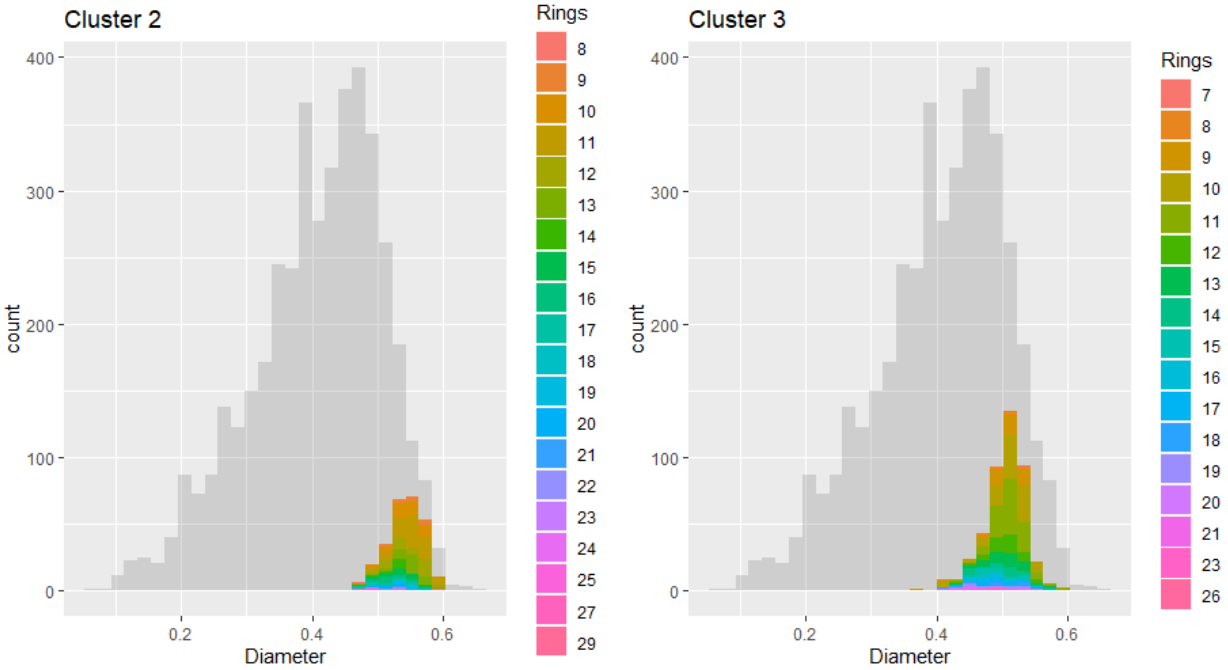


Figure 19: Histogram for diameter feature of cluster 2 and 3

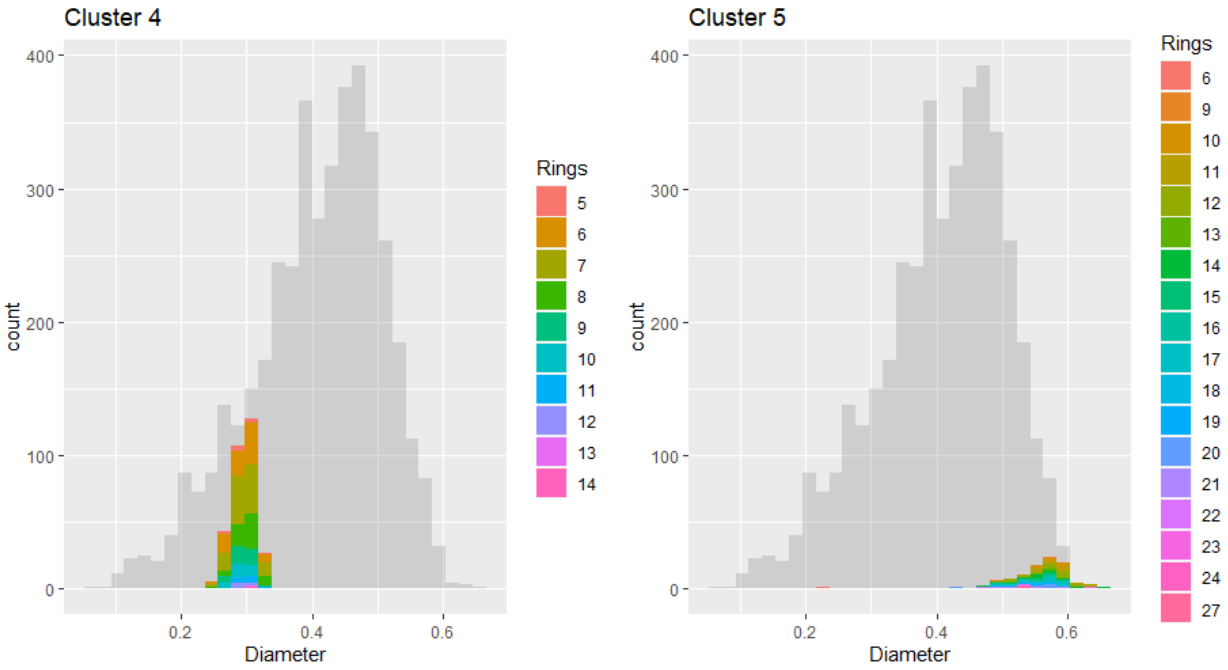


Figure 20: Histogram for diameter feature of cluster 4 and 5

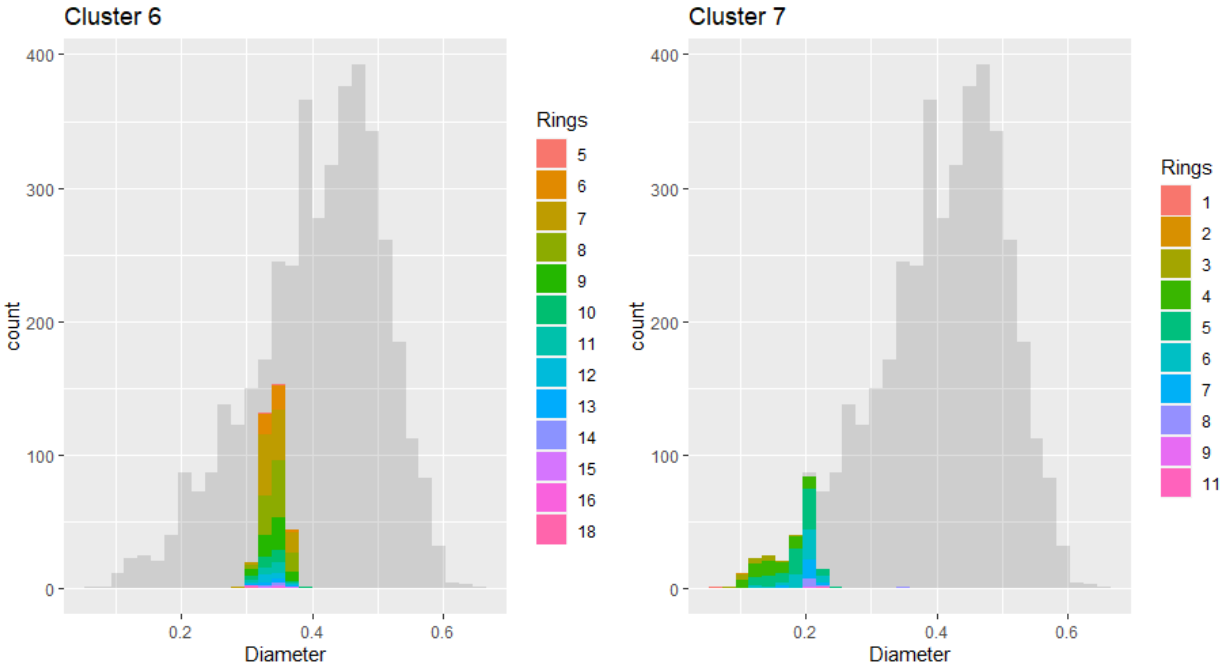


Figure 21: Histogram for diameter feature of cluster 6 and 7

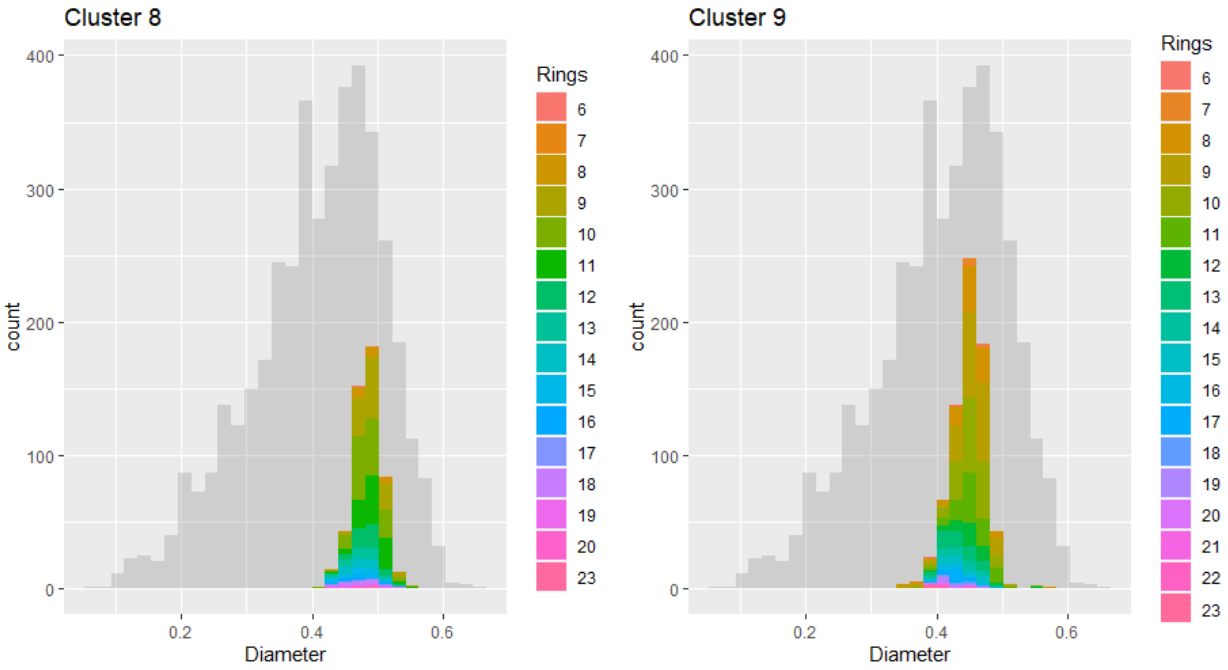


Figure 22: Histogram for diameter feature of cluster 8 and 9

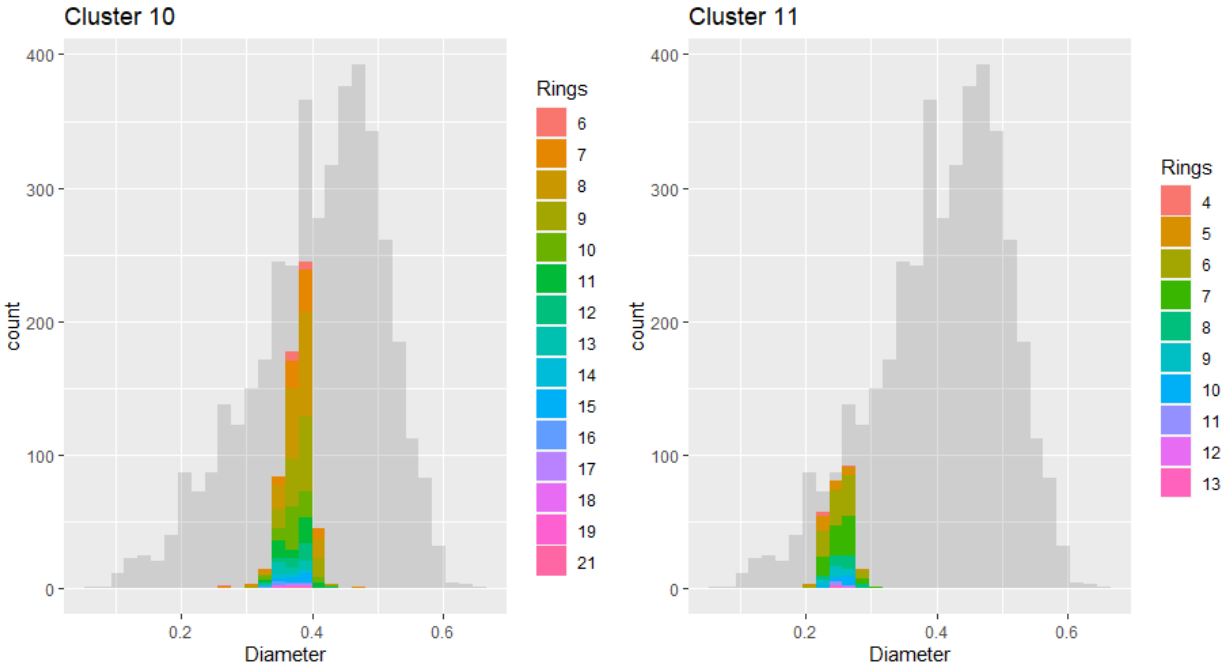


Figure 23: Histogram for diameter feature of cluster 10 and 11

4.4.3 Shell weight predictor

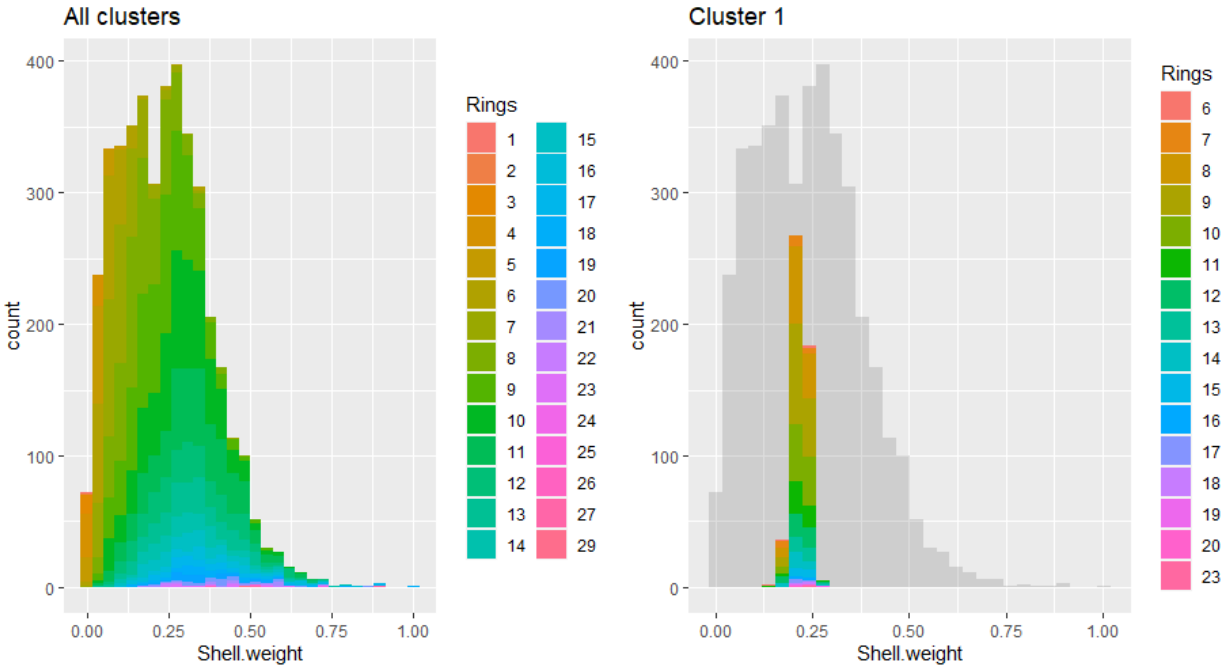


Figure 24: Histogram for shell feature of all clusters and cluster 1

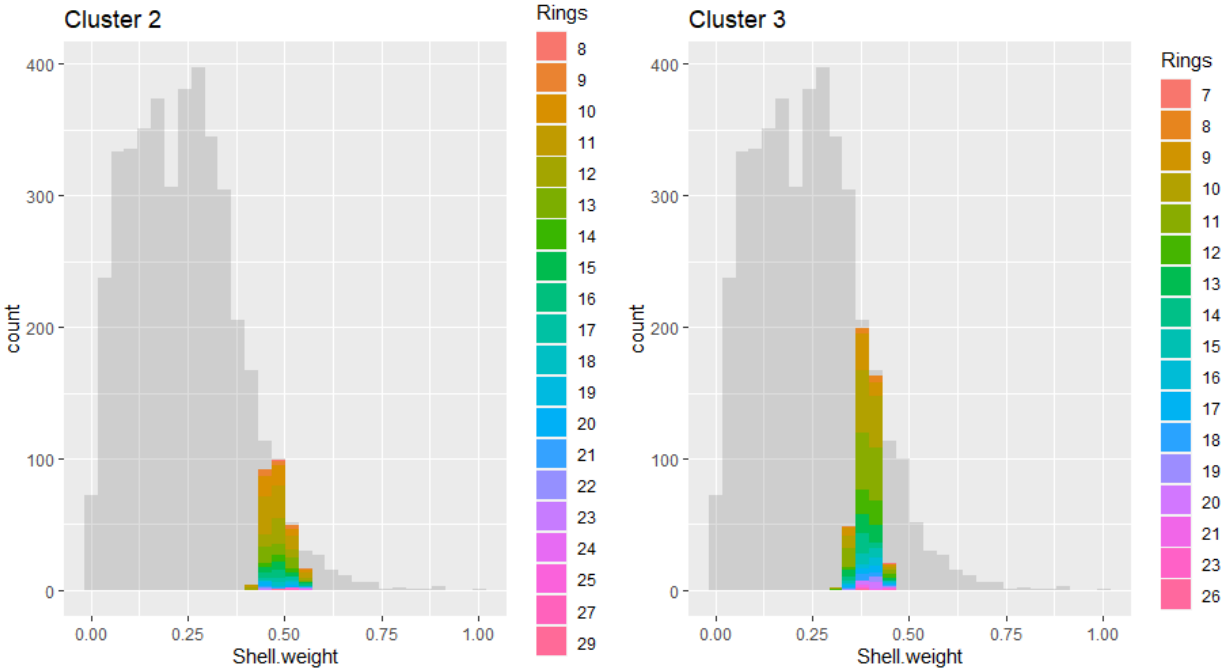


Figure 25: Histogram for shell feature of cluster 2 and 3

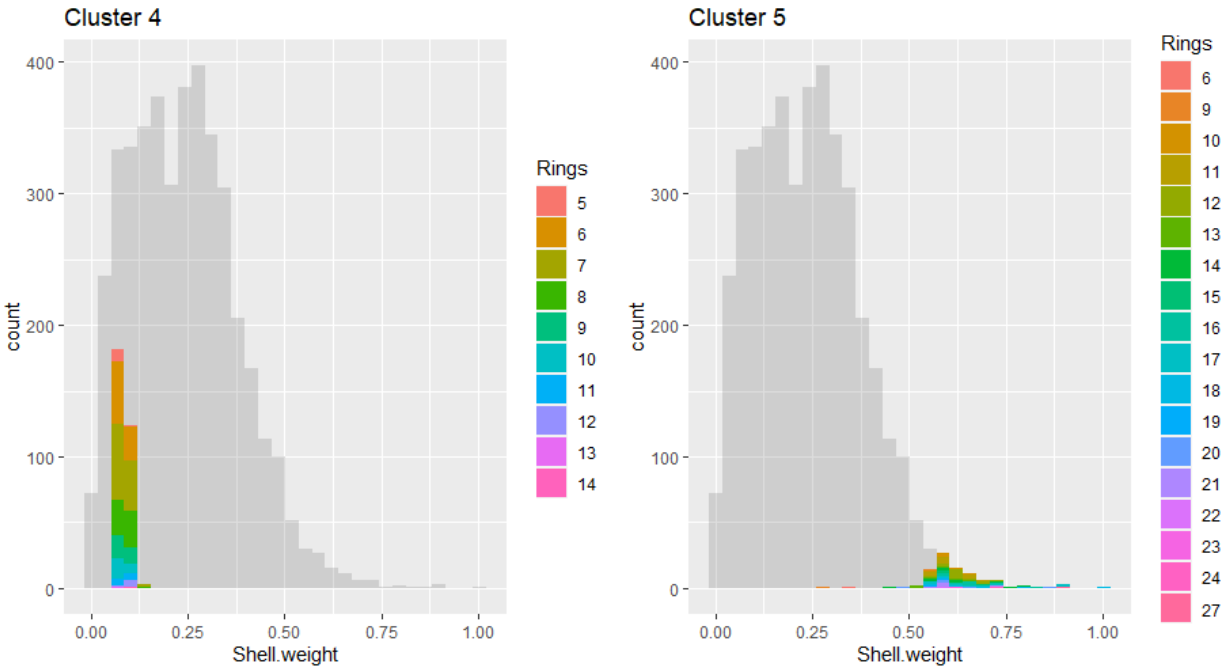


Figure 26: Histogram for shell feature of cluster 4 and 5

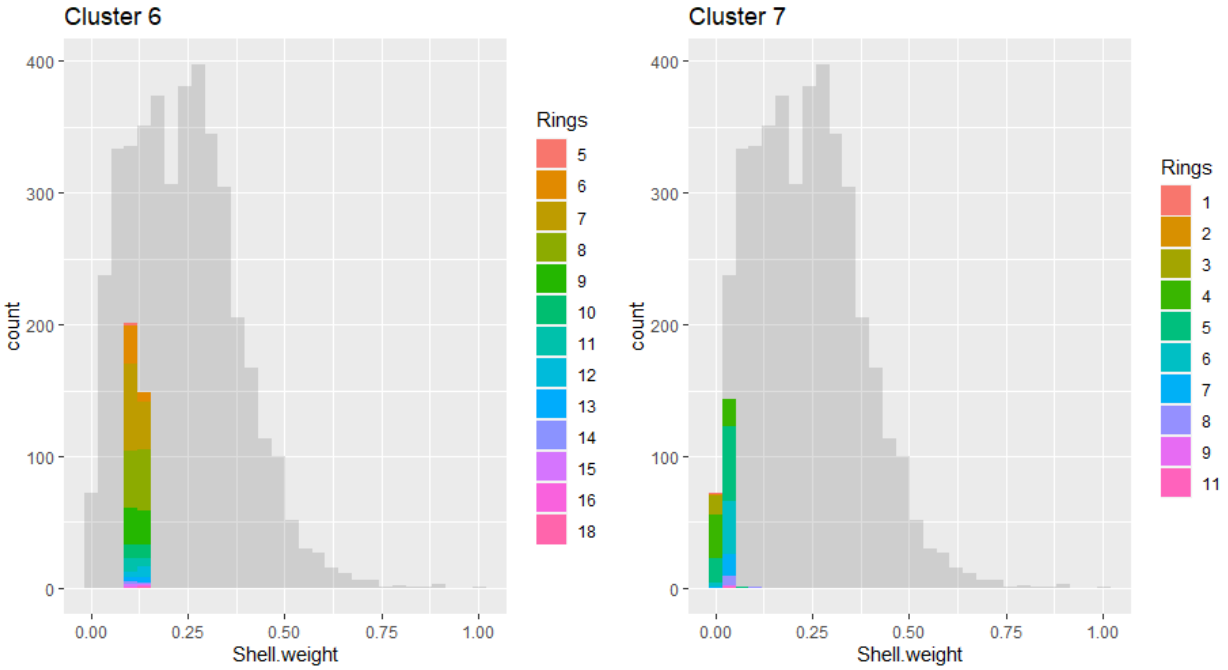


Figure 27: Histogram for shell feature of cluster 6 and 7

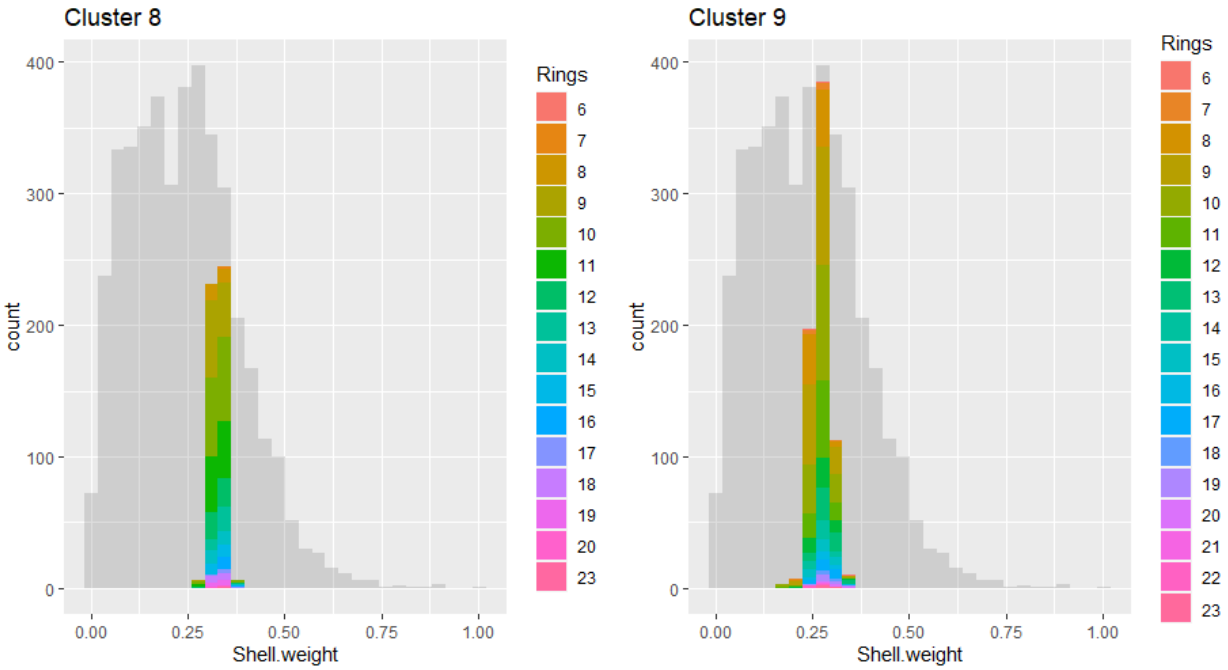


Figure 28: Histogram for shell feature of cluster 8 and 9

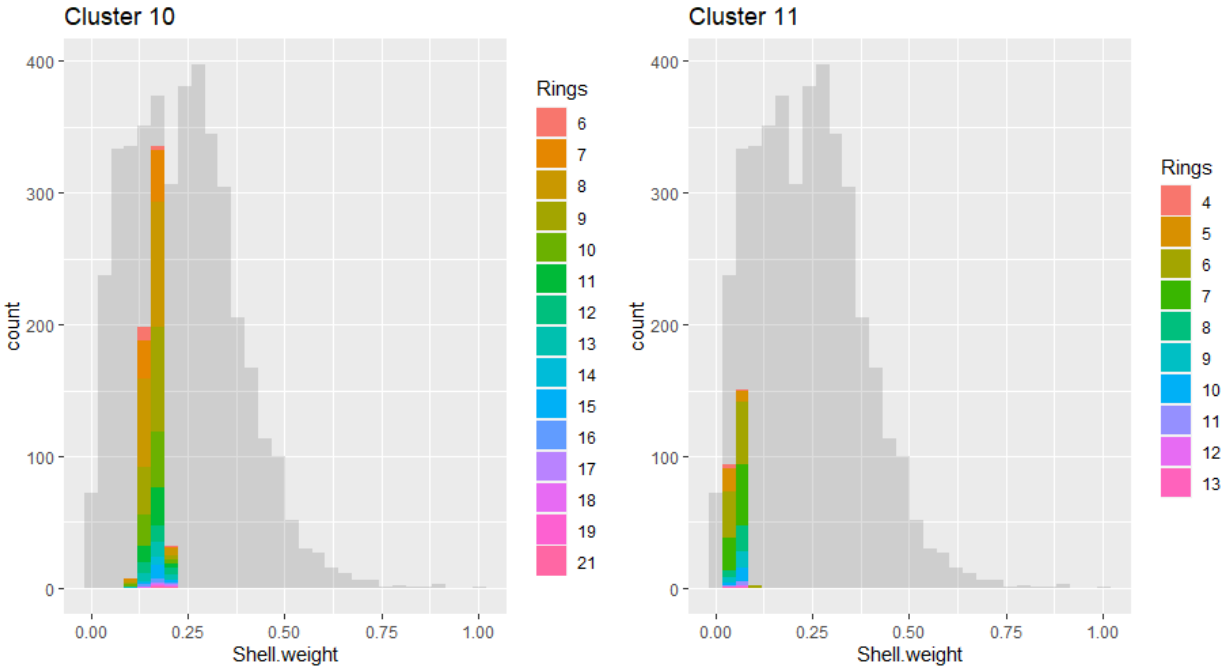


Figure 29: Histogram for shell feature of cluster 10 and 11

4.5 Parallel coordinate plot

The parallel coordinate plot is provided in fig.30 below and shows all of the original descriptive features in respect to the clusters. The first three elements on the x-axis are the

descriptors used for clustering, and the remaining were discarded during feature selection. The solid colour line segments at the intersections of the model descriptors are brighter and more clear than those that were discarded, because the model was trained to fit these features. The shell weight axis does, however, have a few outliers indicated by the thin green lines belonging to cluster 5. The order of the clusters line groups also correspond to that of the centroids shown in fig.9. From this plot, it is not that clear whether any of the descriptors used in modeling can be discarded.

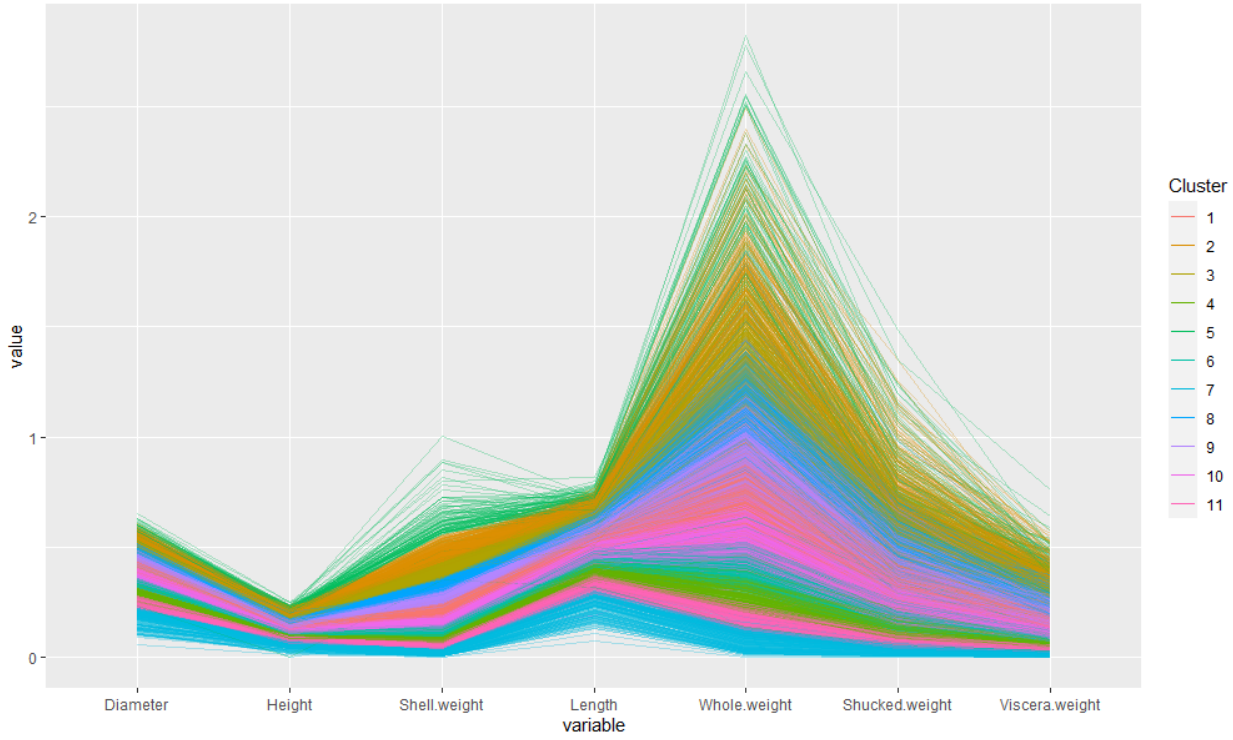


Figure 30: Parallel coordinate plot for all features in abalone data set grouped by cluster

5 Conclusion

The aim of this report was to perform a cluster analysis on the abalone dataset and demonstrate understanding by using various visualization techniques. In this regard, the report has successfully completed the requirement. However, this report was also set out to extract value from the dataset, which in this case, the result was not satisfactory. The reason for this is due to several reasons, the most prominent being the fact that unsupervised clustering[13] was performed and the model could not fully capture the underlying relationship to the rings feature. Moreover, this analysis could potentially have been improved by employing more strict outlier response strategies, dealing with multicolinearity and consider not including the categorical variable into the clustering process.

[13]see next section for supervised implementation

References

[1] A. Banerjee and R. N. Dave. Validating clusters using the hopkins statistic. In *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542)*, volume 1, pages 149–153 vol.1, 2004.

[2] Charu Aggarwal. *Data mining : the textbook p. 158*. Springer, Cham, 2015.

[3] James C Bezdek and Richard J Hathaway. Vat: A tool for visual assessment of (cluster) tendency. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, volume 3, pages 2225–2230. IEEE, 2002.

[4] Alboukadel Kassambara and Fabian Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2019. R package version 1.0.6.

[5] A. KASSAMBARA. *Practical Guide To Principal Component Methods in R: PCA, M(CA), FAMD, MFA, HCPC, factoextra*. Multivariate Analysis. Sthda.com, 2017.

[6] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[7] Rajan Sambandam. Cluster analysis gets complicated. *Marketing Research: A Magazine of Management Applications*, 2003.

[8] Salvatore Mangiafico. *rcompanion: Functions to Support Extension Education Program Evaluation*, 2020. R package version 2.3.26.

[9] Francisco Aragón-Royón, Alfonso Jiménez-Vílchez, Antonio Arauzo-Azofra, and José Manuel Benitez. Fsinr: an exhaustive package for feature selection. *arXiv e-prints*, page arXiv:2002.10330, feb 2020.

[10] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. 1997.

[11] Gero Szepannek. clustmixtype: User-friendly clustering of mixed-type data in r. *The R Journal*, pages 200–208, 2018.

[12] Damien McParland and Isobel Claire Gormley. Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2):155–169, June 2016.

[13] Gero Szepannek. clustmixtype: User-friendly clustering of mixed-type data in r. *The R Journal*, pages 200–208, 2018.

[14] Damien Mcparland and Isobel Gormley. Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, 10, 11 2015.

[15] Paolo Giordani. *An introduction to clustering with R*. Springer, Singapore, 2020.

[16] Jean-Patrick Baudry. Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics*, page 1064, 2015.