Project 3

Data mining

Date: 14/11/2020

Authors:

Andri Ólafsson

Matthías Dan Flemmingsson

Introduction

Project 3 will have us perform cluster analysis on one of three given datasets, as described by the addendum to project 3. The dataset that we Andri Ólafsson and Matthías Dan Flemmingsson have decided to perform cluster analysis on is the Abalone dataset.

Abalone are sea snails, with 56 different accepted species. The dataset is 4117 documented instances of abalone.With 8 different attributes. The attributes documented are:

Name	Data type	Unit/type	Description
Sex	Nominal	Male, female or infant	
Length	Continuous	Millimeter	Longest shell measurement
Diameter	Continuous	Millimeter	Perpendicular to length
Height	Continuous	Millimeter	With meat in shell
Whole weight	Continuous	Grams	Whole abalone
Shucked weight	Continuous	Grams	Weight of meat
Viscera weghts	Continuous	Grams	Gut weight(After bleeding)
Shell Weight	Continuous	Grams	After being dried
Rings	Integer		+1.5 gives the age in years

This dataset has no missing values. But as explained in the description of the dataset, the dataset has had its missing values removed or and replaced. Also the ranges for continuous variables has been scaled, for the use in Artificial neural networks.

Commented [1]: Er þá kannski vitleysa í mér að standardize-a gögnin?

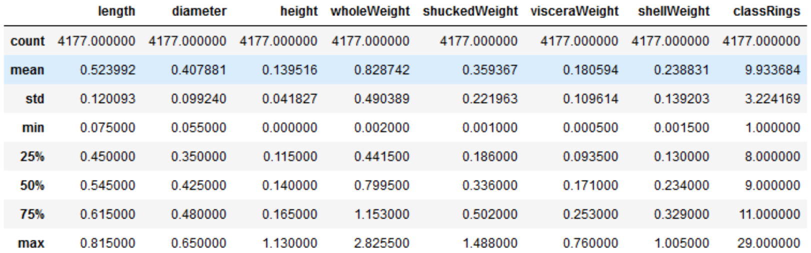
Commented [2]: Myndi ekki segja það. Þar sem við erum að taka normaldreift score sem miðar við fjarlægð frá meðaltali í normaldreifingu. Erum í raun að scala meira.

Held að það ætti að vera í lagi

Potential benefit that could come from this study is the possibility of reducing the work needed to determine the potential age of an abalone and therefore determining. Helping those who determine the kvota for each fishing season, so that abalone won't be overfished, and also which abalone to keep. And with that in mind we could potentially identify which are best to keep in order to preserve the species and who to keep and who to set free to maintain a stable population. The information could also be used in pearl farms.

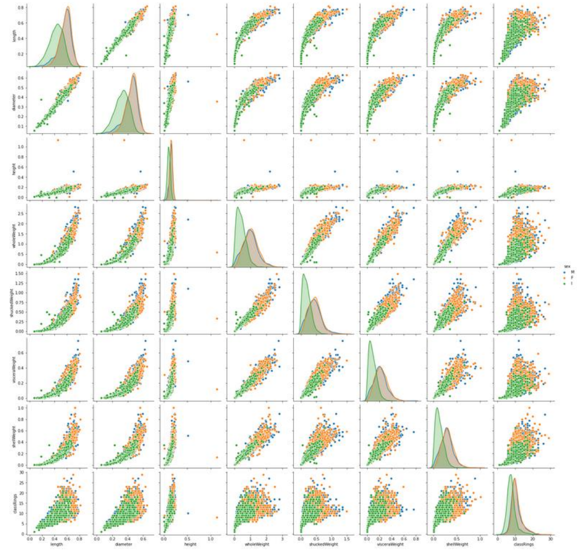
Pre processing

The dataset before preprocessing looks as described.

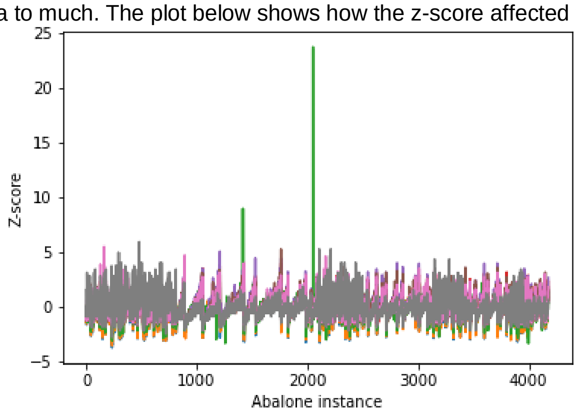


	length	diameter	height	wholeWeight	shuckedWeight	visceraWeight	shellWeight	classRings
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

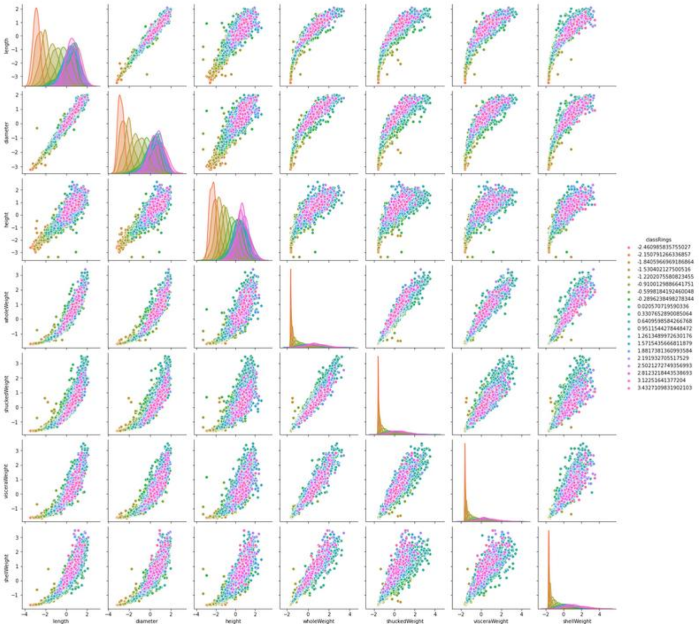
For preprocessing we needed to convert raw data into a csv and add headers. We opted to standardize the dataset instead of going for normalization. The reason being that the dataset contained what we consider weirdly placed outliers, in comparison to the rest of the dataset. The difference being that normalization rescales the values into a defined range of [0,1], deleting the outliers and magically creating the perfect set not preserving the outliers as well as compressing the information into a small interval making the data harder to interpret. Outliers can show some sort of uniqueness in the dataset so we opted to preserve them in order to show the most realistic clustering without the impact that outliers can have. The plot shows a plot of attributes against all attributes before z-score standardization.



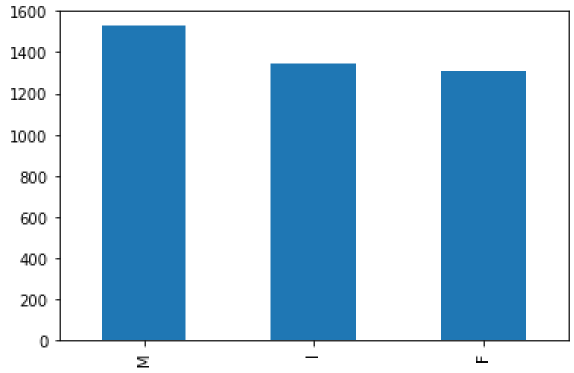
Standardization rescales the data to have mean of 0 and standard deviation of 1. The outliers are still there but their impact on the result is greatly reduced. Giving a more forgiving result to the cluster analysis. To standardize the values we took a z-score of all the values. And then deleted the values who did not enter the scale of inbetween [-3, 3]. The reason we chose the interval -3 to 3 is that around 0.5% of the data lie outside that interval, so more or less all the outliers with a z-score below and beyond those values will not be included and are considered to impact the data to much. The plot below shows how the z-score affected the plotting.



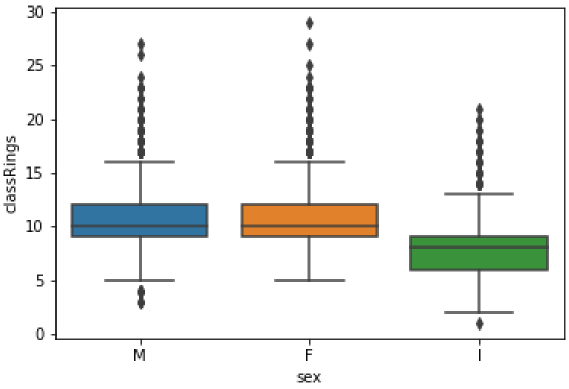
And below are how the plots are effected by z-score standardization.



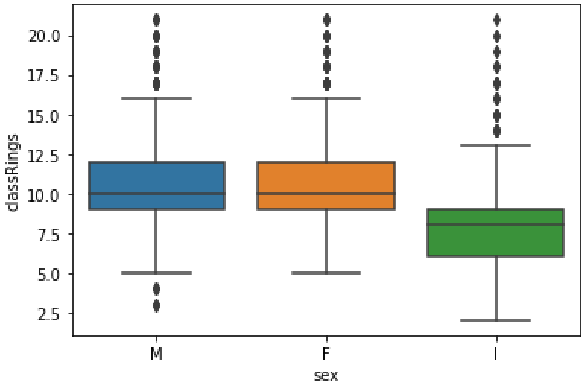
The distribution of the sex, where m = male, I = infant, F = female. Compared to class rings.



The distribution between the sexes are naturally balanced. Since the main goal in the study is to find ways to identify the age or in case of abalones the count of rings in their shell. The box plots below show the before and after the removal of instances via standardization, that fall out of the z-score interval.



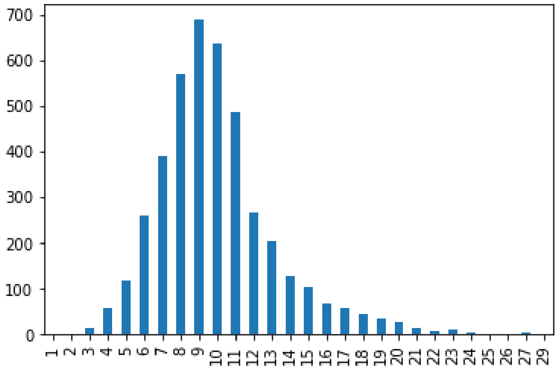
- Before removal



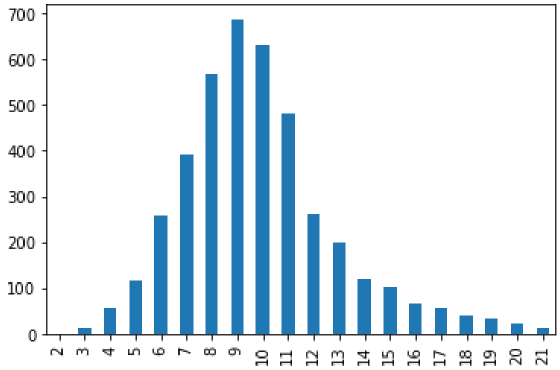
- After removal

We can see that infants contain most of the measurements of the instances.

If we plot the dataset with respect to class rings we can also see a happy sight

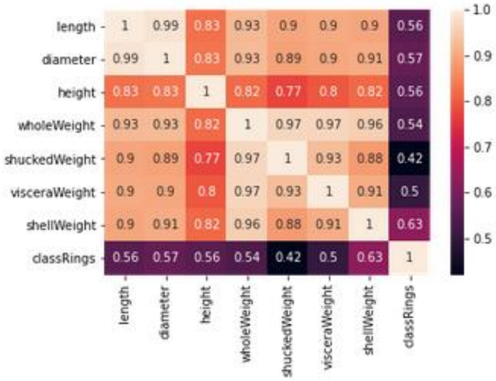


The dataset follows a normal distribution. With z-score in mind the graph looks like

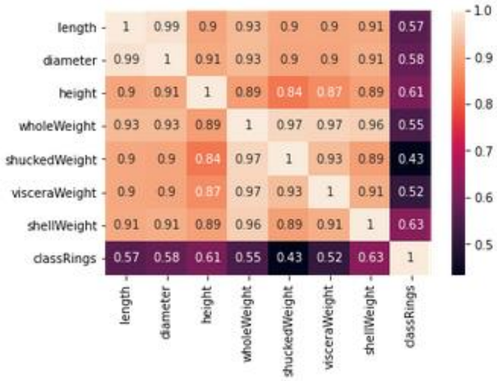


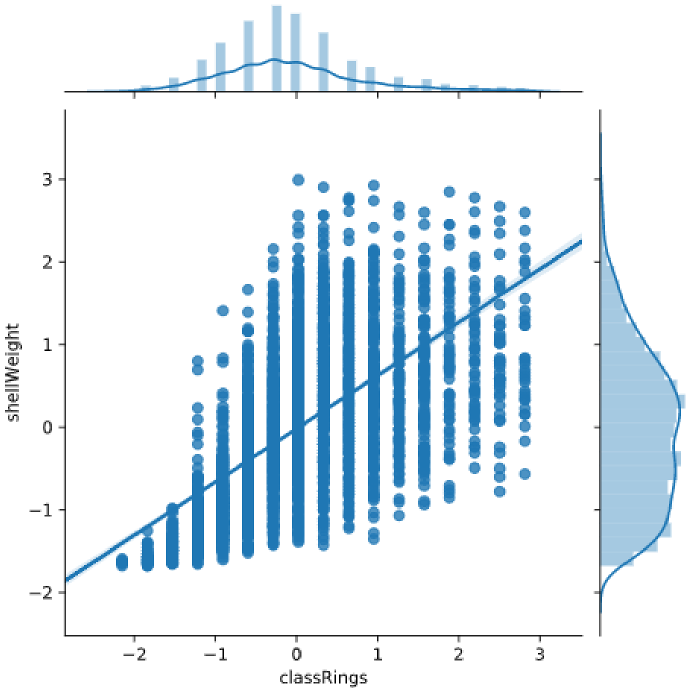
Our highest accuracy for K-means will probably be around a set interval for k between 4 and 18 if we look at the graph. At a glance at least.

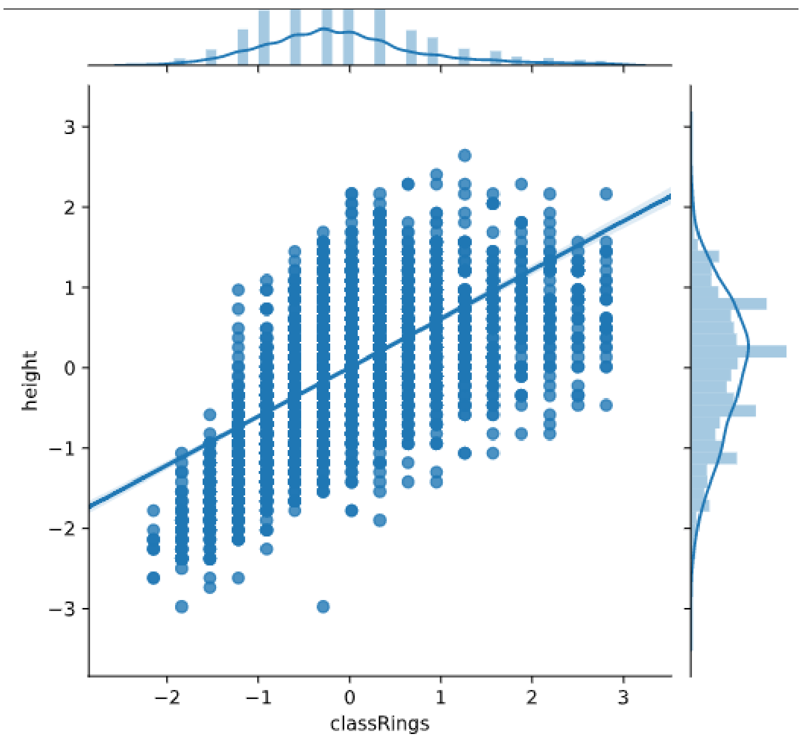
Lets look at the correlation in between attributes with regards to classRings. As we are trying to figure out age. This plot is before z-score standardization

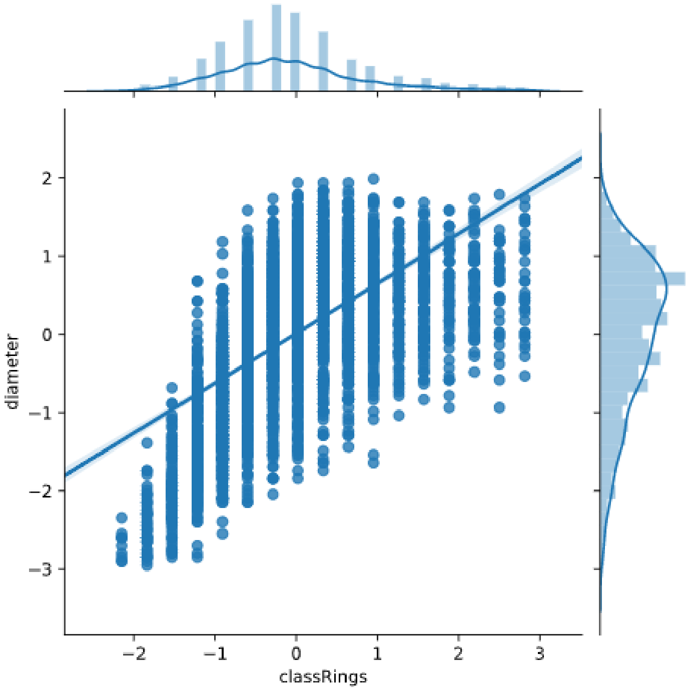


Looking at class rings in the heat map below can see that the weight of the shell, height, and diameter of the abalone are mostly correlated with the age of the abalone.


length
diameter
height
wholeWeight
shuckedWeight
visceraWeight
shellWeight
classRings
1 0.99 0.9 0.93 0.9 0.9 0.91 0.57
0.99 1 0.91 0.93 0.9 0.9 0.91 0.58
0.9 0.91 1 0.89 0.84 0.87 0.89 0.61
0.93 0.93 0.89 1 0.97 0.97 0.96 0.55
0.9 0.9 0.84 0.97 1 0.93 0.89 0.43
0.9 0.9 0.87 0.97 0.93 1 0.91 0.52
0.91 0.91 0.89 0.96 0.89 0.91 1 0.63
0.57 0.58 0.61 0.55 0.43 0.52 0.63 1
1.0
0.9
0.8
0.7
0.6
0.5

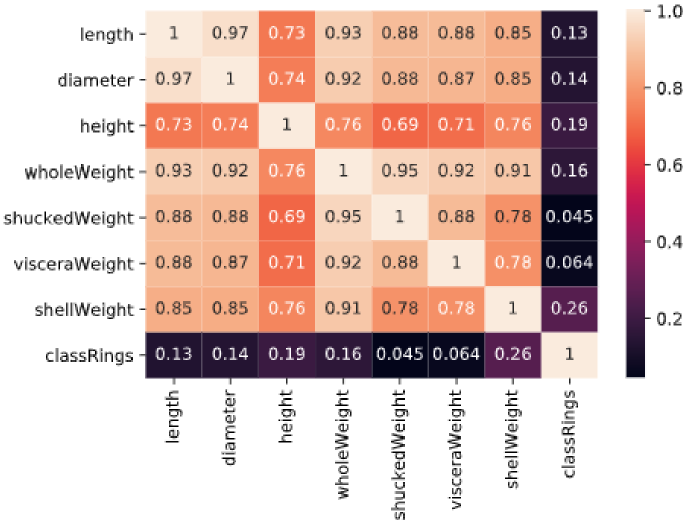

3
2
1
0
−1
−2
shellWeight
−2
−1
0
1
2
3
classRings


3
2
1
0
−1
−2
−3
height
−2
−1
0
1
2
3
classRings



Looking at the correlation when the outliers are removed, we see a slight increase in correlation in height, viscera weight, height and diameter.. Which is a positive sight as the outliers are not affecting the correlation of the attributes as much giving a somewhat more reliable view of the dataset.

Below is a correlation table where we look at only abalones with ring size of 13 and higher, there is little as none correlation of attributes with the target attribute (classRings)



Conclusions :

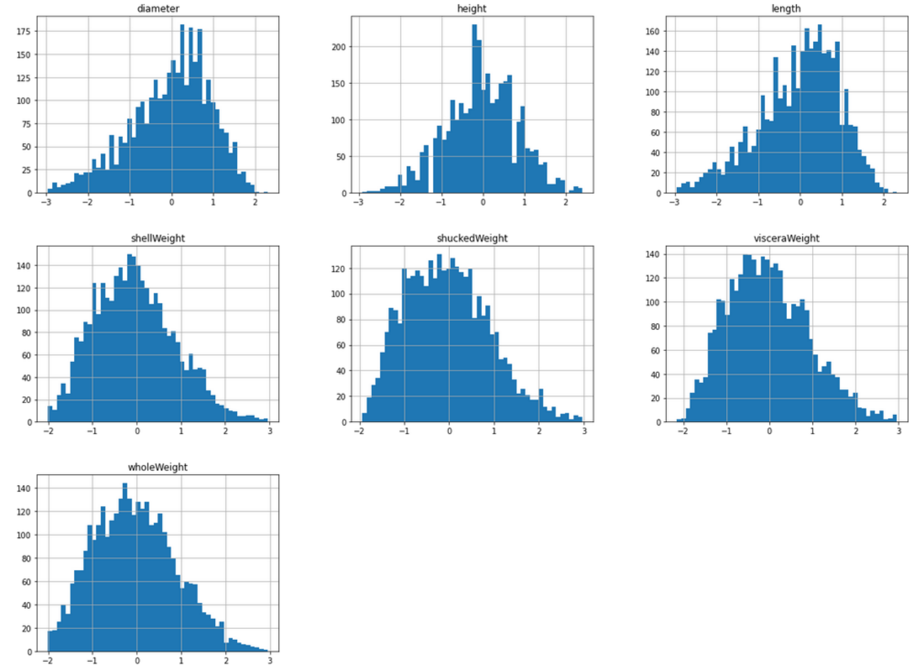
- Weight of abalones is proportional to the size
- By observing correlation of the target variable (rings) with other variables. The size and weight of the abalone varies accordingly to age until the adult age. After reaching adult stage the size and weight stops varying (after about 12 rings (13.5 years old) the measurements have little correlation. That will skew the results as 16% of the dataset consists of older abalone
- No significant differences in size/weight between m/f abalones
- Infant abalone group has lower mean values of size/weight
- Multicolinearity between weight attributes

Clustering

We used the standardized data to perform the clustering. The data used in the clustering process is the data that falls within the z-score interval of [-3, 3] as mentioned in preprocessing.

We split the data into both a testing set and a training set to make sure that the clustering goes well. We are aiming to have the clustering work as a classifier. As the main goal is to cluster in terms of the age of an abalone.

The initial preprocessing resulted in the attributes looking like this:



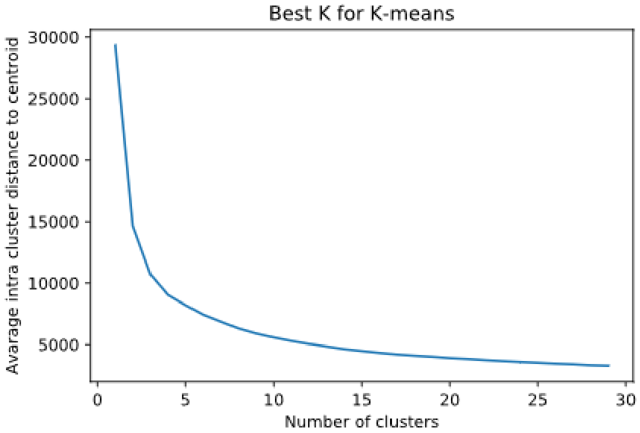
As mentioned in the initial preprocessing, the correlation is skewed and can be seen here on this graph. It will in a way impact the result but it is inevitable due to the skewness of the data itself as most of the data consists of infants, and mid age abalones.

With all that is behind we can now finally start the clustering process. We have some data before hand that was figured out during the preprocessing that can be of use now when we start to cluster.

- We know that the data is skewed, meaning more of the data resides on some end of the dataset. Unbalanced data.
- We also know from the graphs that most data will reside in ages 4-1

The hypothesis is at least valid from before in this case. Where the expected best k will be somewhere between 4-18.

Now we move on to K-means. The best K for K mean lies somewhere between 4-18



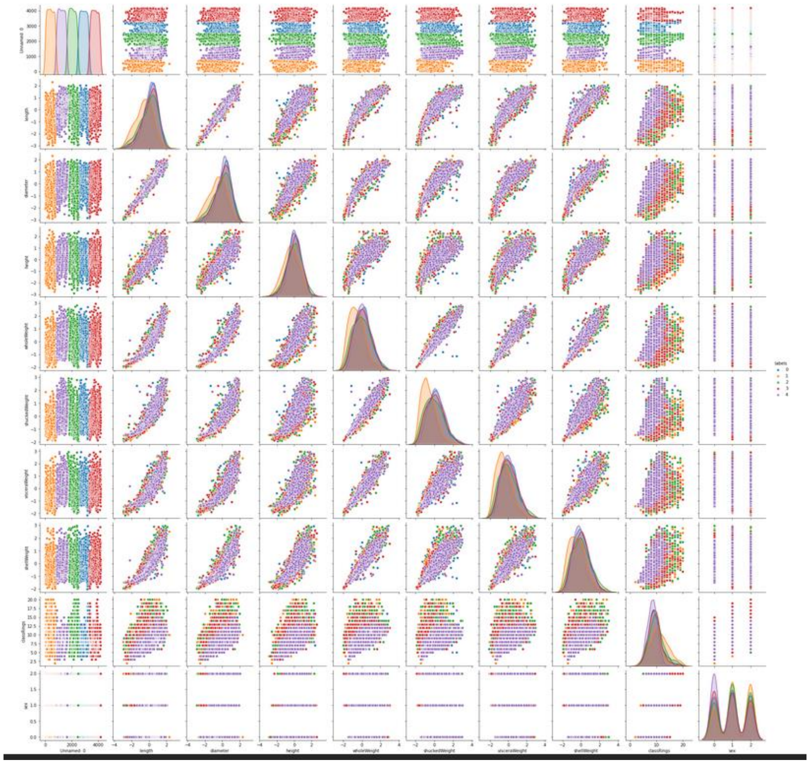
This graph shows us the inertia(SSE) compared to the chosen K for knn. We are going to choose a k that does not underfit nor overfit the data. Our initial guess is that the best K is somewhere between 4-18 clusters. Looking at the graph we can determine that the best K is probably 5. As all values after 5 dont show significant improvement upon the initial SSE. So we choose 5 as our best cluster.

SSE lowers as we progress through the K’s. Leveling once we reach around 15. But it is not to be confused of being the best K. As 15 is not providing enough benefits to be used as the best K. 5 provides enough gains to be used as the best K for clustering, as it is not much higher than 15 in terms of SSE.

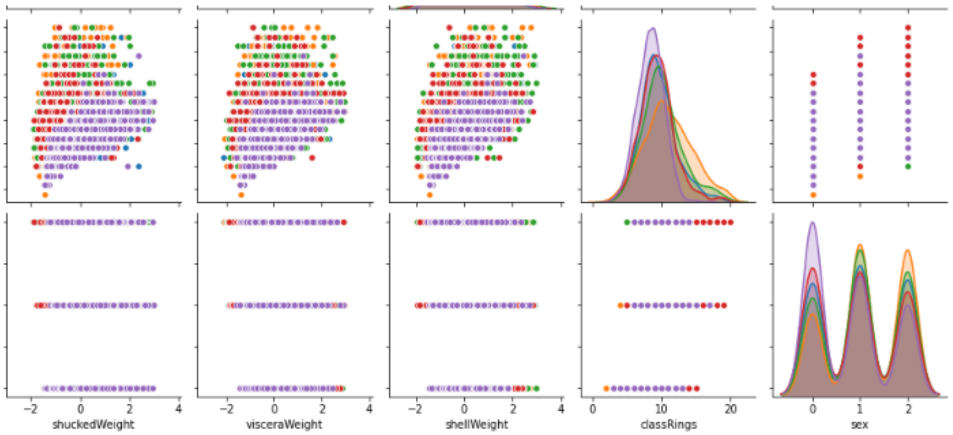
The centroids are positioned at these coordinates:

1. 2900.9516
2. 407.9750
3. 2056.5657
4. 3754.0314
5. 1232.5006

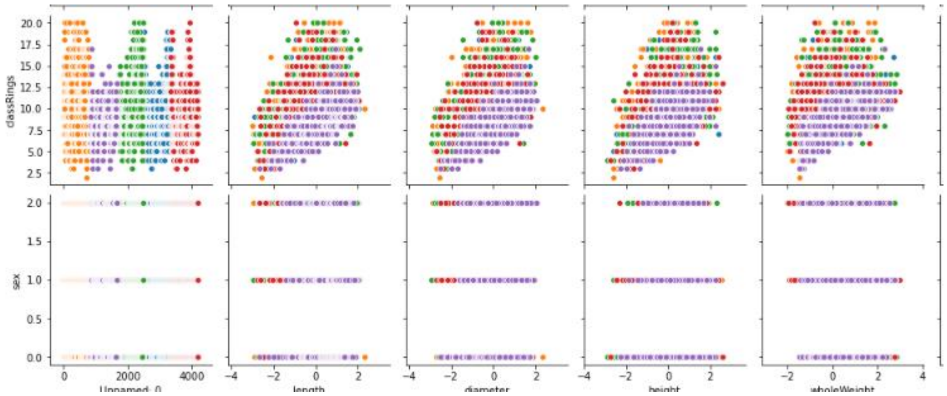
A pairwise plot of the clusters comes out like this:



The one attribute classification that we are most interested in is the classification plot of classRings.

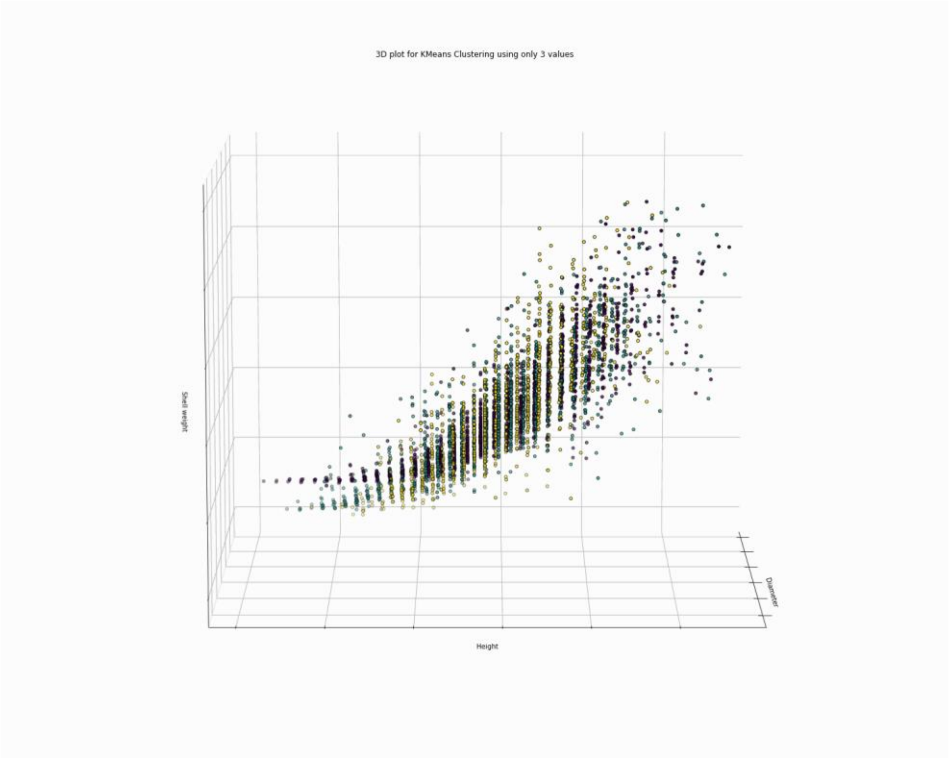


−2
0
2
4
shuckedWeight
visceraWeight
shellWeight
0
10
20
classRings
0
1
2
sex



classRings
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
sex
2.0
1.5
1.0
0.5
0.0
0
2000
4000
Unnamed: 0
−4
−2
0
2
length
diameter
height
wholeWeight
4

The clustering comes out like this in kmeans:



KNN is not a good classifier for this dataset in its raw form. preprocessing and combining data is required to „create“ clusters.

For some unexpected circumstances we could not finish project 3. The work up to this paragraph represents our project 3 until other is announced.

We suspect that the outcome from the classification would be rather split, and those abalones of greater size will be more likely to be older based on what data we have managed to work out.

Conclusions:

- Weight of abalones is proportional to the size
- By observing correlation of the target variable (rings) with other variables. The size and weight of the abalone varies accordingly to age until the adult age. After reaching adult stage the size and weight stops varying (after about 12 rings (13.5 years old) the measurements have little correlation. That will skew the results as 16% of the dataset consists of older abalone
- No significant differences in size/weight between m/f abalones

- Infant abalone group has lower mean values of size/weight
- Multicolinearity between weight attributes
- SSE is dependent on cluster size