G0O02a: Statistical Data Analysis Project

Alexander J. Lindhardt
r0826077

April 2021

1 Exploratory analysis and transformation to normality

1.1

The data we have consists of 600 data points with 17 variables, where 16 of them are continuous and describes the features of a bean. The last one is a categorical variable that describes what kind of bean that data point belongs to. To explore the relationship between the variables, we plot the scatterplot-matrix, see some of the selected scatterplots in figure 1. In the scatterplots, we color the points based on which type of bean it is to make it possible to distinguish them from each other, here we have Dermason in blue, Sira in orange and Seker in green. First thing we notice is that some variables have a strong linear association whereas other variables seem to have none or very little linear association. For example the variable Area has a strong linear correlation with variables: Perimeter, MajorAxisLength, ConvexArea and EquivDiameter, as seen in the first row of figure 1. Meanwhile, some variable pairs with no apparent linear association are plotted in the second row of the same figure. Another interesting property we can retrieve from the scatterplot-matrix is the clustering of the bean types, which is especially evident in the first and last row of figure 1. We can for example see that the Dermason bean has a smaller area than the other two beans and also that Seker has a larger compactness than the other two. We also detect some possible outliers in the plots, for example in the solidity and roundness variables.

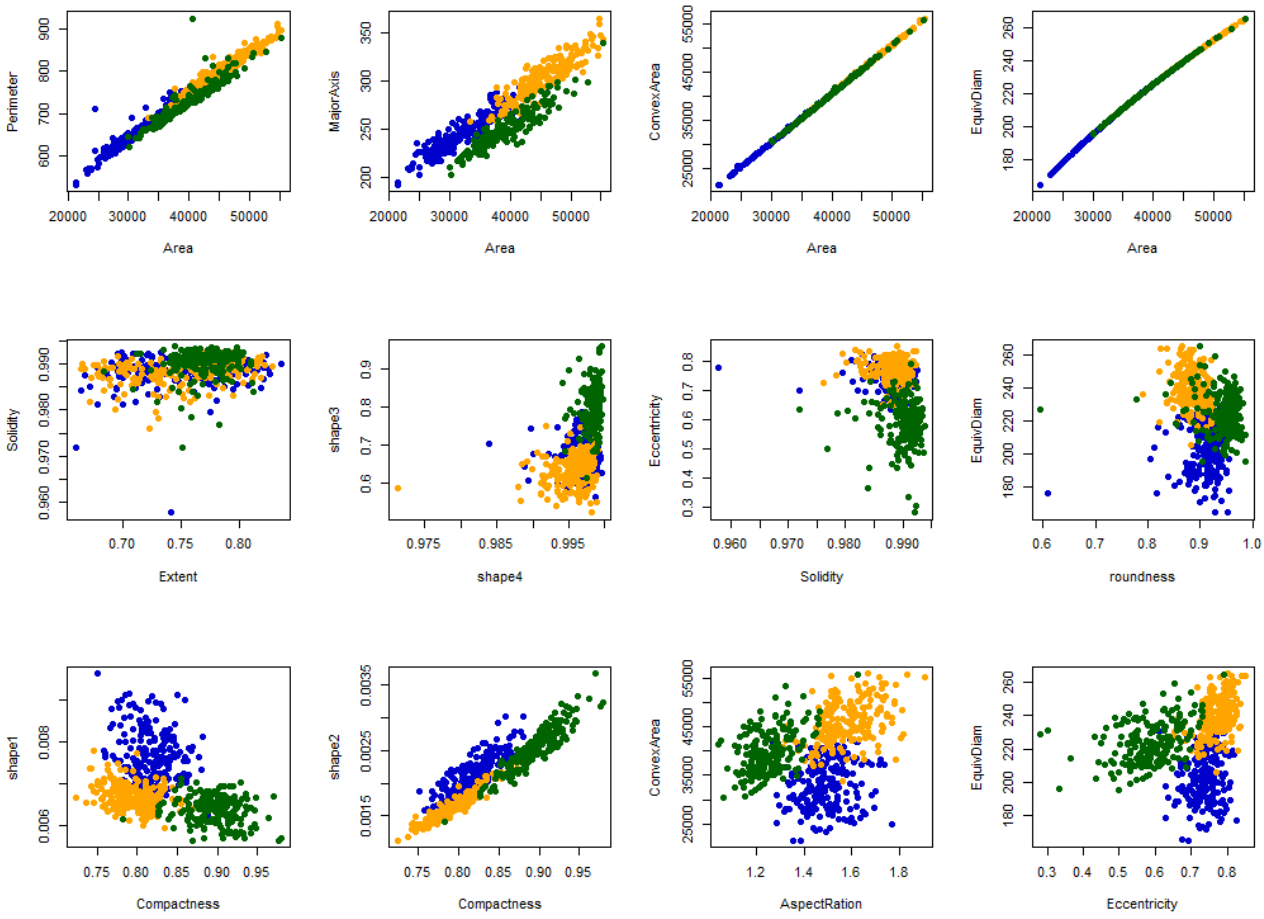


Figure 1: Scatterplots for a few selected pairs of variables. The first row is meant to show the linear association between some variables, the second row to show no linear association and outliers, the third row shows clustering.

To further investigate the correlation between the variables, we visualize the correlation matrix in figure 2. There we can see that some variables have a very strong correlation with many variables, for example Area and AspectRatio meanwhile Solidity and Extent are not very correlated with any other variable. Notice also how AspectRatio and Eccentricity have similar correlation with the same variables, this is also true for Area and Perimeter, ConvexArea and EquivDiam and also roundness and compactness.

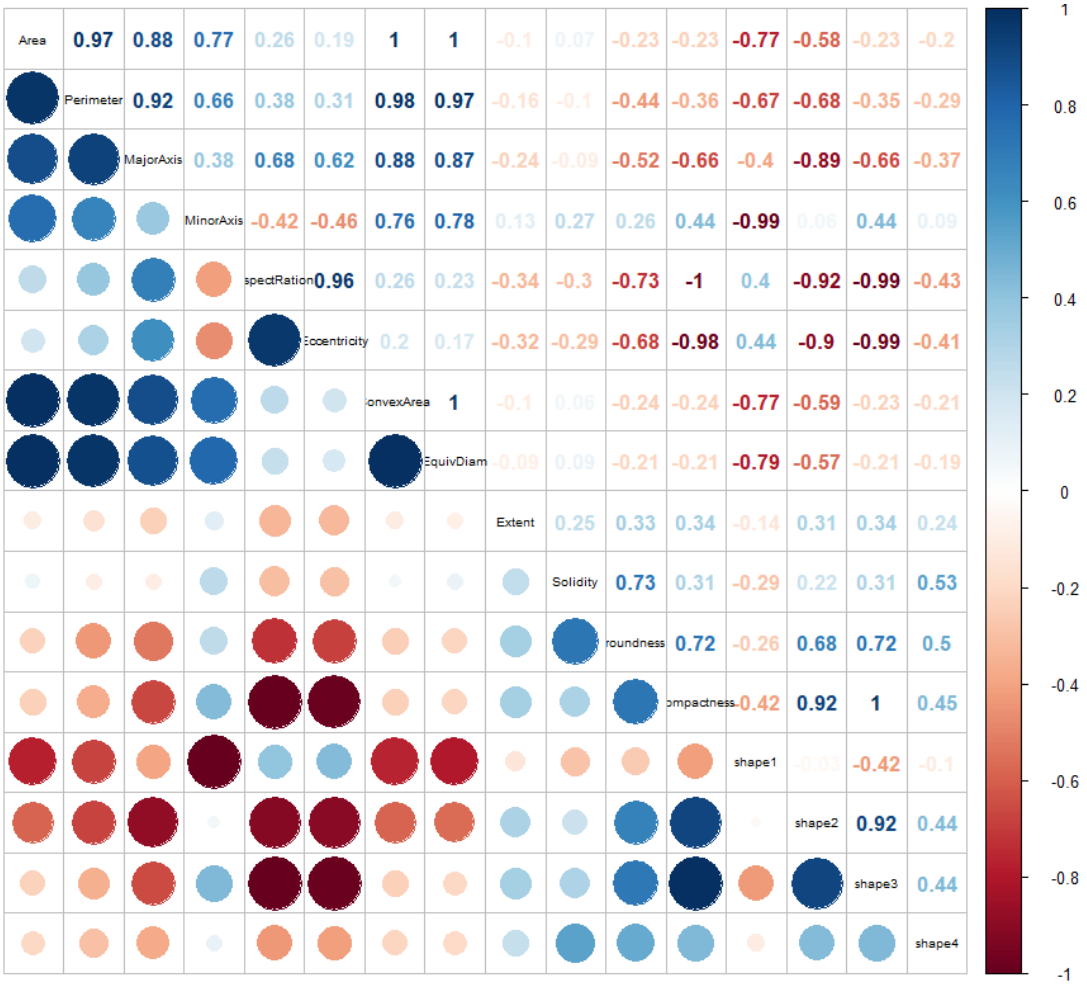
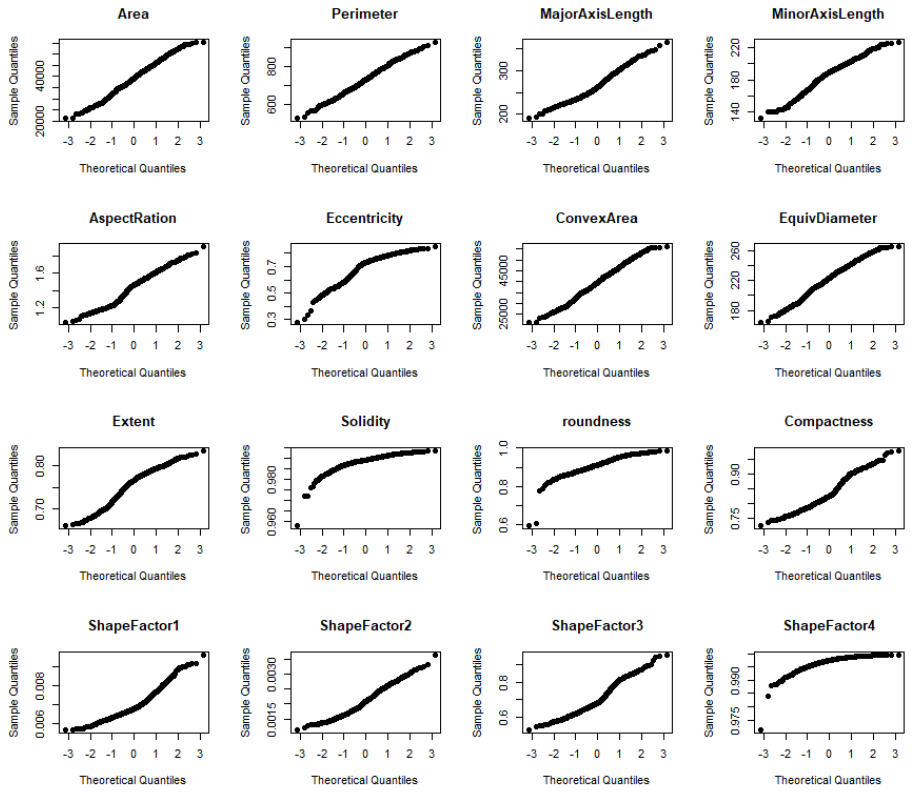


Figure 2: Illustration of the correlation matrix where the colors show if the correlation is positive or negative and the magnitude. The size of the circles and the intensity of the numbers also indicates the magnitude of the correlations.

We also test the data for multivariate normality. We do this by first looking at each continuous variable and see if they are normally distributed, if this is not the case, the multivariate normality is not satisfied. We show the normal QQ-plot for each variable in figure 3a, we also perform a Shapiro-Wilk test and the resulting p-values can be seen in figure 3b. It is very clear by just looking at the normal QQ-plots that some variables are not normally distributed, since they are not following a line. The Shapiro-Wilk test confirms this as we can see that only three variables have a p-value greater than 0.01. We can already draw the conclusion that the data is not normally distributed but we can also see that some of the scatterplots between two variables in figure 1 is not very elliptical and also the χ^2 QQ plot of the squared Mahalanobis distances distinctly shows that the data is not multivariate normal.



(a)

```
Area :  0.03430267
Perimeter :  0.1678205
MajorAxis :  3.250033e-08
MinorAxis :  9.672721e-08
AspectRation :  6.883204e-08
Eccentricity :  1.592077e-18
ConvexArea :  0.0381238
EquivDiam :  0.0004470258
Extent :  2.010661e-12
Solidity :  7.754795e-26
roundness :  1.229859e-18
Compactness :  5.975069e-11
shape1 :  2.316551e-15
shape2 :  4.59718e-07
shape3 :  2.601876e-12
shape4 :  3.405784e-28
```

(b)

Figure 3: (a) The normal QQ-plots for each continuous variables and (b) the p-values after running the Shapiro-Wilk test on each variable.

1.2

We now only consider the continuous variables again, namely the first 16 variables. We will take a look at the distribution for each variable and if their distribution is very different from a normal distribution we will transform them using a Box-Cox transformation. To decide which variables to transform, we again take a look at the normal QQ-plot (figure 3a) and the p-values from the Shapiro-Wilk test (figure 3b) for each variable. We also look at the histogram for each variable, see figure 4 where histograms for the variables furthest from normality are shown.

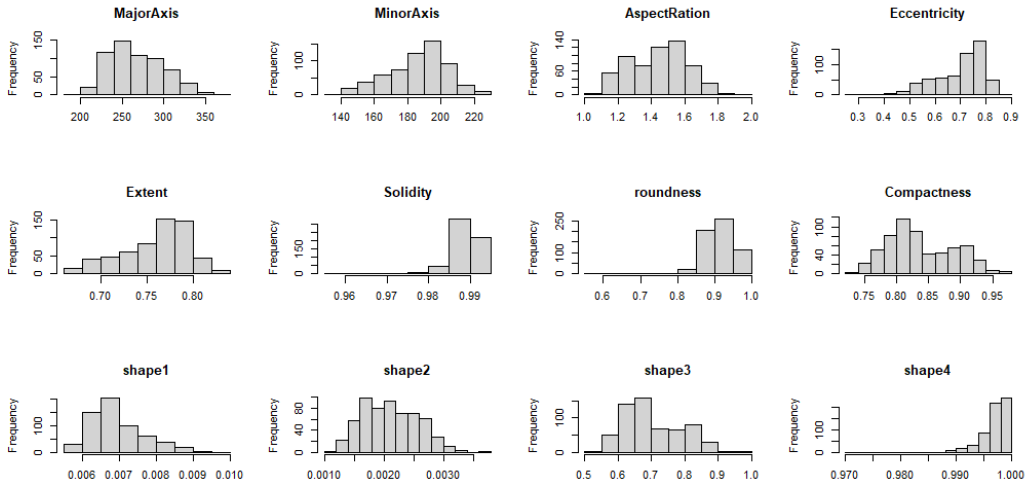
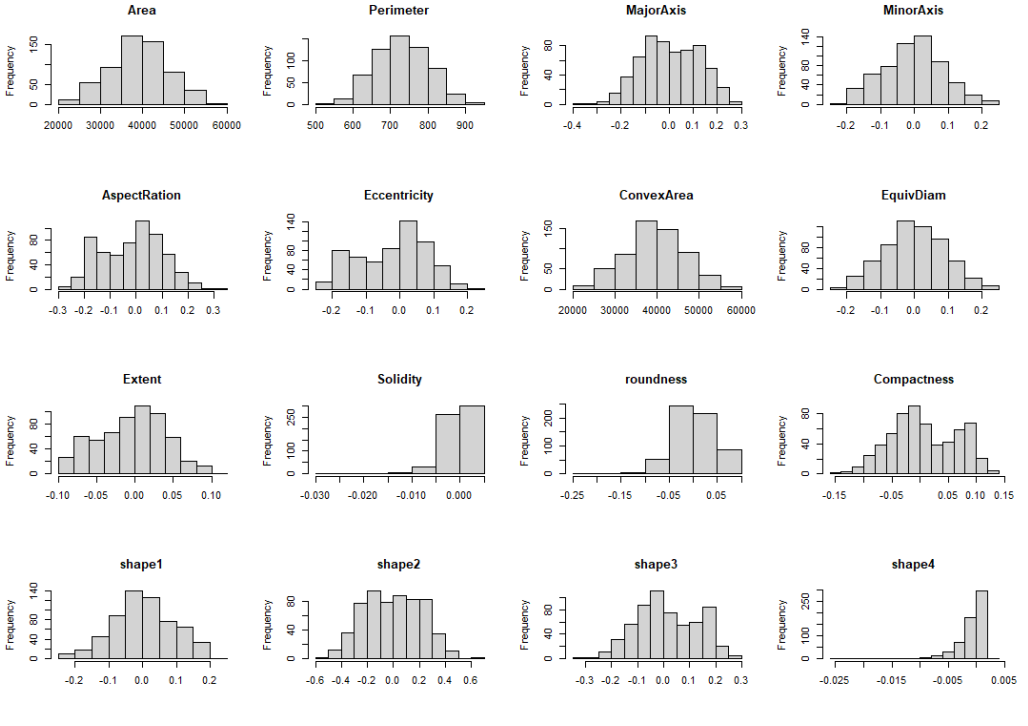


Figure 4: Histograms for variables furthest from normality according to the Shapiro-Wilk test.

By just looking at the QQ-plots and histograms we already see some variables being far from normally distributed, for example the shape factors. The only variables that seems to be normally distributed are Area, Perimeter, ConvexArea and maybe EquivDiameter. The resulting p-values from the Shapiro-Wilk test show that only three variables have a p-value greater than 0.01. So we will use Box-Cox transformation on all variables except for Area, Perimeter and ConvexArea, and since we have some possible outliers, we use a robust transformation to get a more reliable results. After the Box-Cox transformation of the variables that were far from normally distributed we once again plot the histograms, see figure 5a and the p-values from the Shapiro-Wilk test on the transformed data, see figure 5b. We see that some variables are now closer to a normal distribution than before. The skewed variables MajorAxisLength, MinorAxisLength and shapeFactor1 now look normally distributed when looking at the histogram and their p-values are now close to 0.01. Other variables such as shapeFactor4, Solidity and roundness are still very far away from a normal distribution.



(a)

```
Area :  0.03430267
Perimeter :  0.1678205
MajorAxis :  0.0001471291
MinorAxis :  0.003584516
AspectRation :  1.363185e-07
Eccentricity :  4.251434e-11
ConvexArea :  0.0381238
EquivDiam :  0.0410313
Extent :  7.853551e-07
Solidity :  1.138868e-24
roundness :  5.210424e-11
Compactness :  1.021098e-07
shape1 :  0.004213671
shape2 :  4.463957e-05
shape3 :  1.021099e-07
shape4 :  2.013173e-27
```

(b)

Figure 5: (a) Histograms over all variables after Box-Cox transformation and (b) the p-values after performing a Shapiro-Wilk test on all variables.

2 Principal Components Analysis (PCA)

We will now continue to work with the transformed data for the rest of the sections.

2.1

We will perform a PCA based on the correlation matrix since not all variables are on the same scale. Namely the Area variables have a larger variance than the rest of the variables which means that these variables will influence the principal components more than the rest. We visualize this in the left plot in figure 6 where we show the correlation between the variables and

each principal component after performing PCA based on the covariance matrix. We clearly see that the first component is heavily influenced by both Area variables and therefore the first principal component has more than 99% of the variance which is not desirable. If we do the same analysis while using the correlation matrix instead we see on the right plot in figure 6 that the influence from the variables on the principal components are more evenly spread out. We also get that the first component explains 48% of the variance which is more what we want. So this is why we will use the correlation matrix when doing PCA.

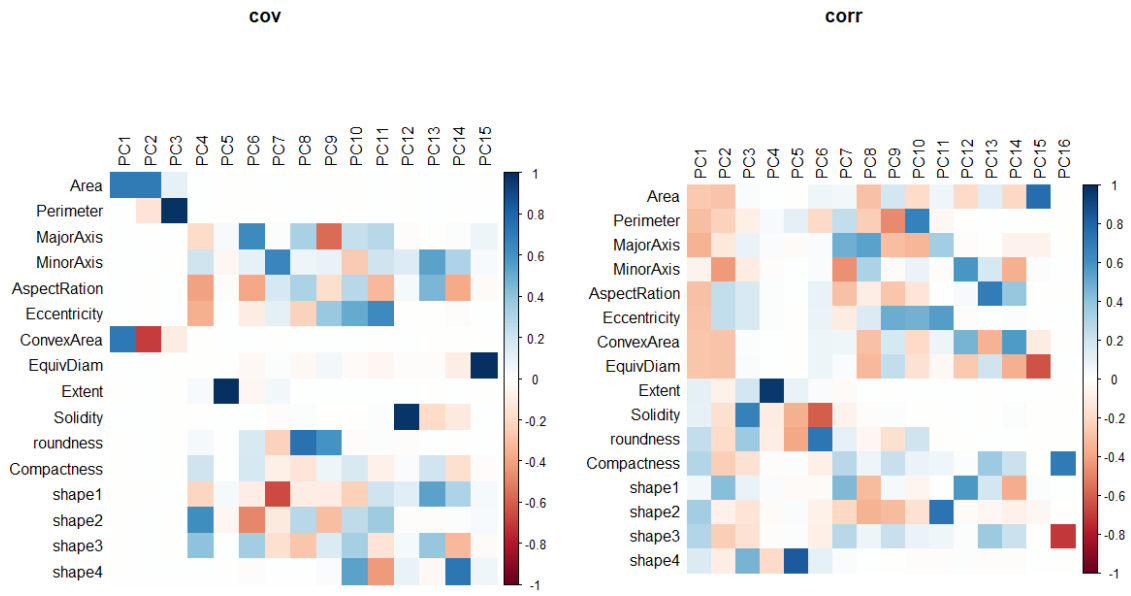


Figure 6: Correlation matrices between the variables and the principal components when doing PCA with the covariance matrix (left) and the correlation matrix (right).

Now when choosing the amount of components, we take a look at the cumulative proportion of explained variance for the principal components and also the scree plot that plots the variance against the number of principal components. The cumulative proportion of the explained variance is shown in figure 7. By looking at these, we see that with two components we already have 80% of the information and with three components we get close to 90%. In the scree plot we see that the "elbow" of the plot is when using three principal components and thereafter it plateaus. So we decide to use three principal components based on this.

```
Call:
PcaClassic(x = mytrainingdata[, 1:16], scale = TRUE)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11     PC12     PC13    PC14     PC15      PC16
Standard deviation     2.7753 2.2878 1.2020 0.93320 0.7663 0.31675 0.18353 0.13824 0.07470 0.04304 0.01933 0.003947 0.002868 0.00228 0.001034 3.431e-10
Proportion of Variance 0.4814 0.3271 0.0903 0.05443 0.0367 0.00627 0.00211 0.00119 0.00035 0.00012 0.00002 0.000000 0.000000 0.00000 0.000000 0.000e+00
Cumulative Proportion  0.4814 0.8085 0.8988 0.95324 0.9899 0.99621 0.99832 0.99951 0.99986 0.99997 1.00000 1.000000 1.000000 1.00000 1.000000 1.000e+00
> |
```

Figure 7: PCA summary on the training set showing the cumulative proportion of variance.

2.2

We make biplots of the first three scores where each data point is colored based on which type of bean it is, see figure 8 for the biplot of the first two components. In this figure we clearly see that the different beans create a cluster, although with some overlap. In the biplot of the

first and the third scores we see that the Sira bean is somewhat grouped by its own whereas the other two types are more mixed together. In the final biplot when we use the second and third component, we can also see clustering with all bean types but not as clear as in the first biplot. We can make the conclusion that the first two principal components contain enough information about the features of the beans to divide the bean types from each other.

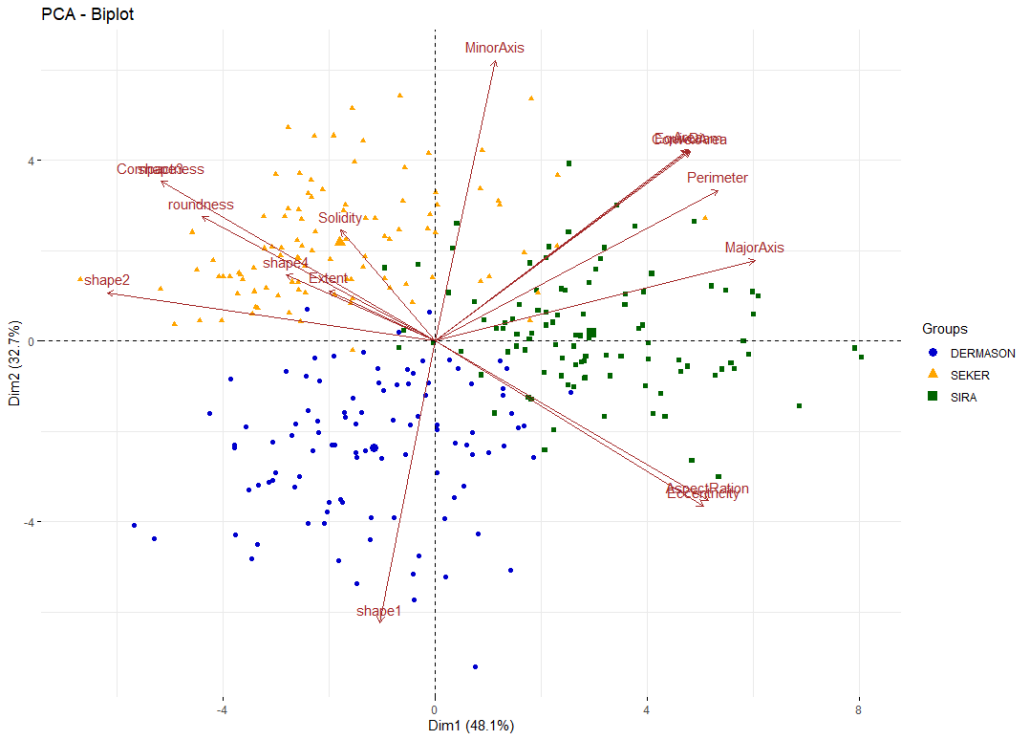


Figure 8: Biplot of the first two scores where the bean type is distinguished by color and symbol.

2.3

We now check if there are any outliers in the training set. First, we look at the classical PCA outlier map in the left plot of figure 9 where we can see that four points in the training set are possible outliers. If we color them specifically and make the scatterplot-matrix for the training set we can see how they compare to other data points. We see that these points are deviating from the rest of the points in some scatter plots and they are therefore likely outliers in the training set. We make the same analysis with robust PCA and we can see the outlier map for that case in the right plot of figure 9. Here it confirms that these four points are PCA outliers and we therefore remove these points from the training set. We also see some mild outliers in the training set but these are not as severe as the others.

2.4

After the outliers have been removed we again have to decide how many principal components we want for the PCA. We check the cumulative proportion of variance and the scree plot and see that they look very similar to before and we therefore choose three components again. We now use the test set to compute their scores and predicted values. We calculate the mean of the predicted Area variable for each bean type and compare it with the mean area for these types, the result can be seen in table 1. We see that the mean of the predicted area between the types are very similar to the actual mean of the area.

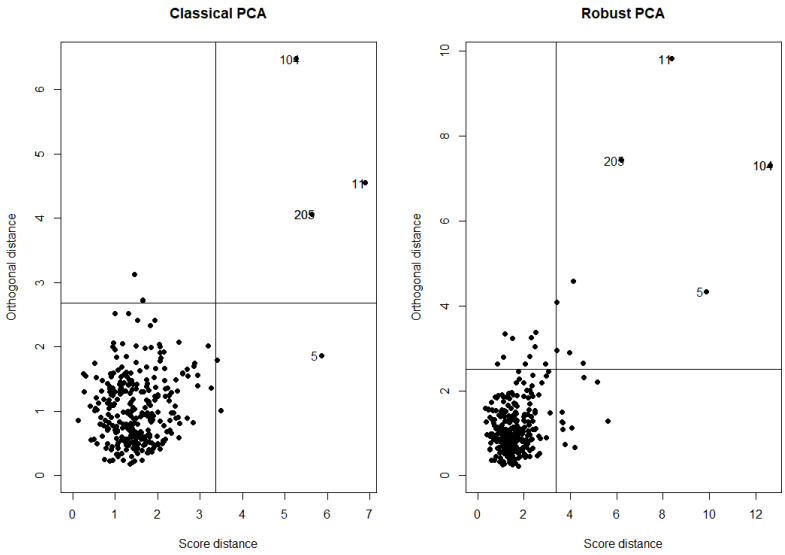


Figure 9: Outlier maps for the training set when applying Classic PCA (left) and Robust PCA (right).

Type	Predicted Area	Area
Dermason	33243.9	32118.7
Sira	45354.9	44729.1
Seker	38828.1	39881.3

Table 1: Table showing the mean of the predicted area and the true area of each bean type.

2.5

Once again, we take a look at an outlier map, but this time we include the test set. We do this with the classical PCA and we visualize this in the left plot of figure 10. We see that the classical PCA doesn't detect any big outliers, if we do the same with robust PCA we see in the right plot of figure 10 that it does detect some outliers but they are not as big as the ones we deleted from before.

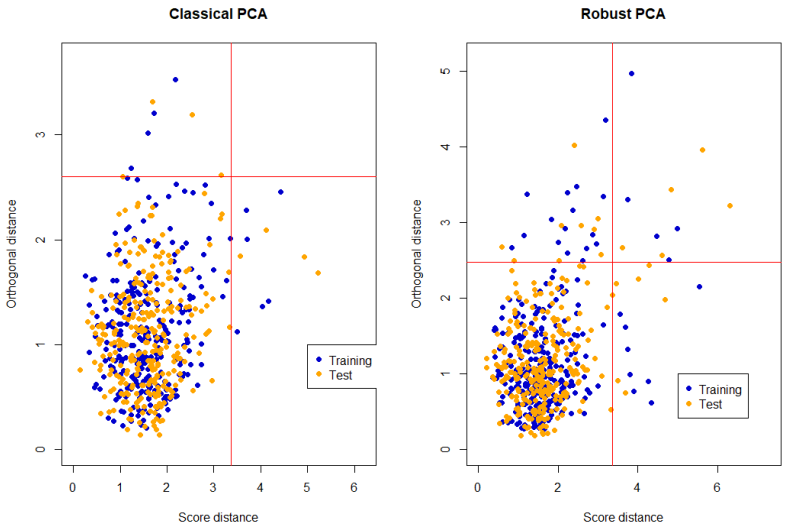


Figure 10: Outlier maps for the training set and test set when applying Classic PCA (left) and Robust PCA (right). The colors indicates which set the data point belongs to.

3 Clustering

3.1

We perform a partitioning clustering analysis on the observations in the training set, by comparing the two clustering methods k-means and k-medoids. First we need to decide how many clusters should be used without using any prior knowledge about the Class variable, meaning we don't know how many different bean types there actually are. To do this, we calculate the average silhouette width for each clustering with different values of k. This tells us how well an object belongs to a cluster it's assigned to on average, meaning that a clustering with a high average width has a more solid structure. The resulting average silhouette widths are shown in table 2 for both k-means and k-medoids clusterings with different k values.

k	2	3	4	5	6
k-medoids	0.289	0.334	0.276	0.252	0.217
k-means	0.315	0.347	0.275	0.250	0.227

Table 2: Average silhouette width for different k for methods k-mean and k-medoids.

We see here that the both methods have the highest average when using three clusters which is very logical since we actually have three different types of beans. It is also worth noting that the k-means clustering has a slightly higher average than k-medoids even though it is more robust against outliers. We could actually see that when including the outliers that we removed from the previous section, the k-medoids clustering have a higher average. Now to choose which method is better fitted for the training data we also look at the clustering plots we get after using both methods, see figure 11. We see that both methods create good clusters but with some overlap in between, however the clusters when using k-means are a bit more tightly together. So for choosing the best partitioning clustering method in our case is not obvious since both k-means and k-medoids returns a very similar result. However, since the resulting clustering plot and average silhouette width are a bit better for k-means, this is the chosen method.

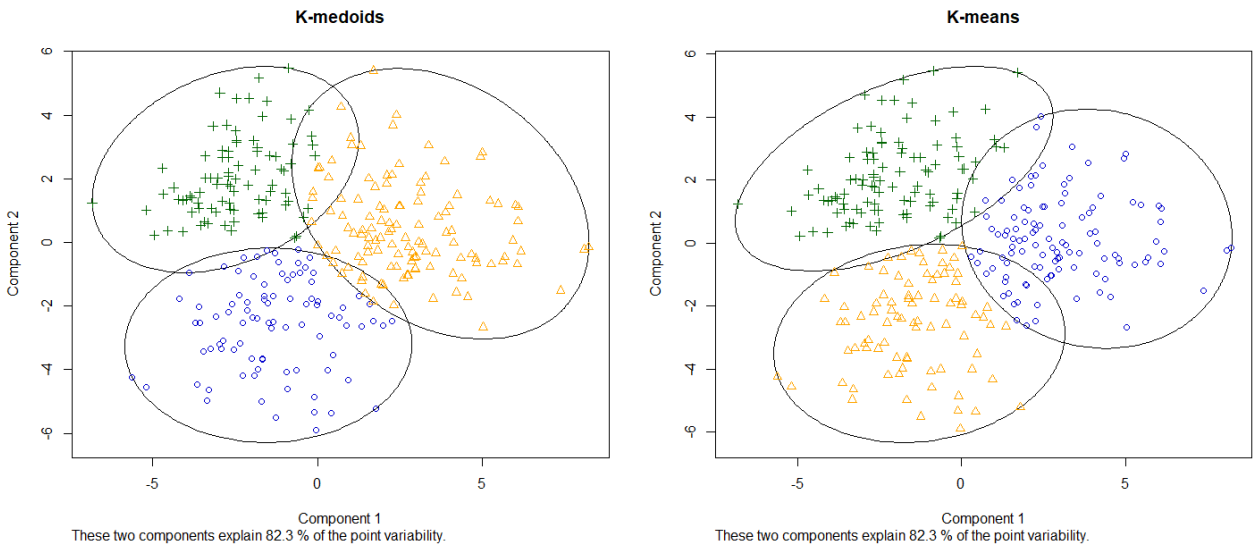


Figure 11: Cluster plots when using clustering methods k-medoids (left) and k-means (right).

3.2

Now we move on to do an hierarchical clustering of the observations in the training set. As in the previous section we will compare and discuss different methods and decide which one is most appropriate for our data set. The first method we use is the agglomerative nesting (AGNES) where we will use single, average and complete linkage and compare internally. The second method is divisive analysis. Now when deciding which agglomerative method is the best, we take a look at the clustering tree for each method. There we can see that single linkage suffers from the chaining effect so we find no distinct clusters. When using average linkage we get some clear clusters together with some very small clusters. Finally with complete linkage we see that we discover three clear clusters and this method seems to be most appropriate for our data set, see the left plot in figure 12 for the complete linkage clustering tree where the three clusters are colored differently.

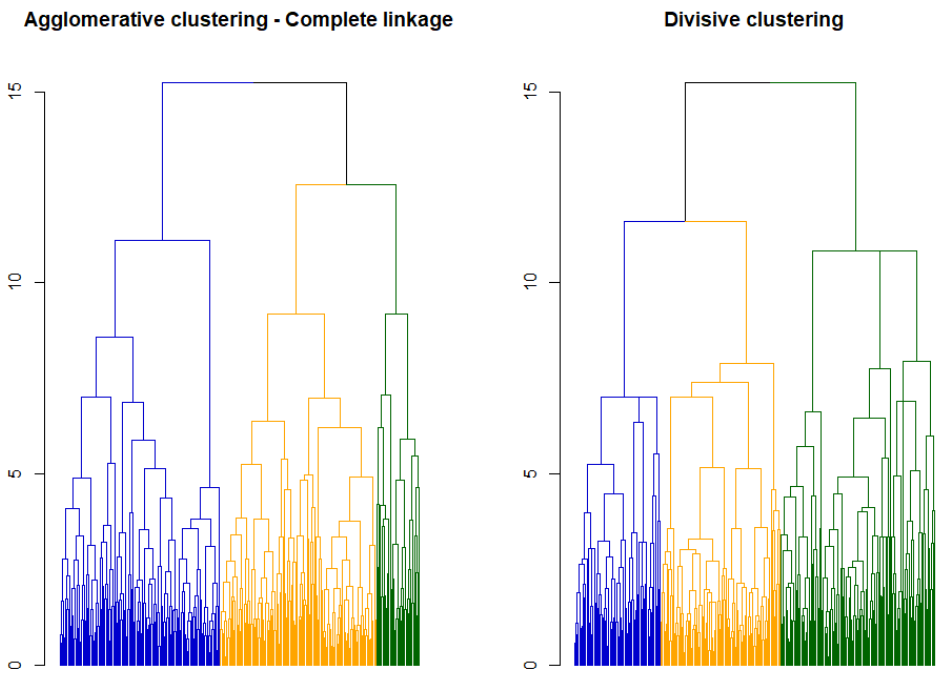


Figure 12: Clustering tree of the observations with three clusters when using agglomerative clustering with complete linkage (left) and divisive clustering (right).

Now we take a look at divisive clustering and the clustering tree that it produces. Once again we see three distinct clusters, see the right plot in figure 12. Now to decide whether the complete linkage agglomerative clustering or the divisive clustering is better fit for this data set we look again at the average silhouette width for the clusterings as well as their respective cluster plots. We see that the silhouette width is 0.33 and 0.19 for the divisive clustering and the agglomerative clustering respectively. Combining this information with the clustering plots in figure 13 we conclude that the divisive method produces a better clustering for the training set.

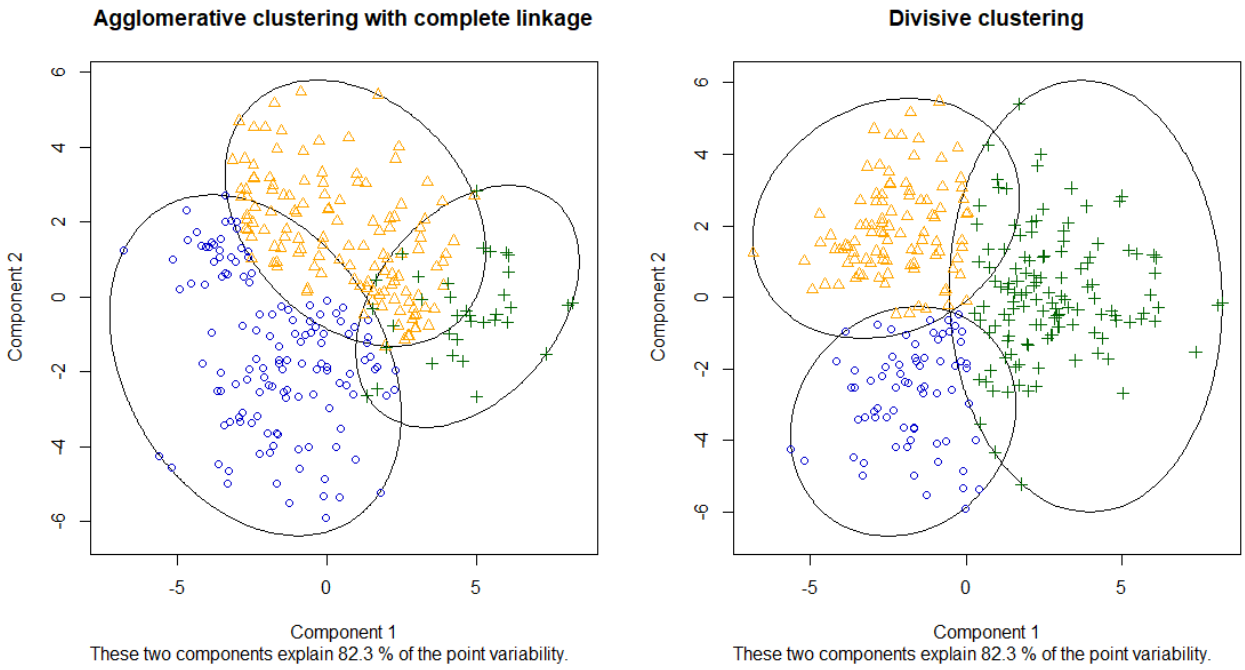


Figure 13: Cluster plots when using clustering methods agglomerative with complete linkage (left) and divisive (right).

3.3

Now instead of looking at the observations, we do a hierarchical clustering on the continuous variables. First, if we again observe the correlation plot in figure 2 we saw that some variables have similar correlation to the same variables. So we already know that some variables are very similar and expect them to belong to the same cluster. For example, by just looking at the correlation matrix we see that some obvious clusters would be: Area-Perimeter-MajorAxis-ConvexArea-EquivDiameter, AspectRatio-Eccentricity and shape4-Solidity. Now performing the clustering using both agglomerative clustering with complete linkage and divisive clustering, see figure 14, we discover that both methods has three clear clusters the only difference being that the variable MinorAxis is in different clusters. Since it is more logical that MinorAxis belongs in the same cluster as MajorAxis and also by looking at the correlation plot we decide to use the agglomerative clustering.

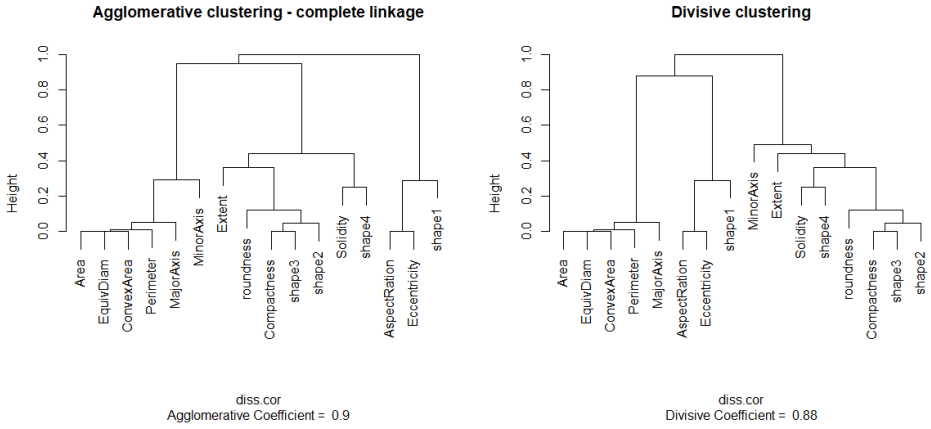


Figure 14: Clustering tree of the variables when using agglomerative clustering with complete linkage (left) and divisive clustering (right).

3.4

To decide which clustering of the observations that is mostly accurate to the true groupings. We take a look at the confusion matrices for both the hierarchical clustering and the partitioning clustering. We see that the accuracy for the k-means clustering had an score of 89% meanwhile the divisive method only had an accuracy of 85%. We now create a heat map where we sort the variables and observations according to the agglomerative clustering for the variables and the k-means clustering for the observations, see figure 15. Here we can see how the different clusters of variables have similar values for the observations inside a certain cluster. For example how the Area cluster is darker green for the second cluster and more yellow in the third cluster. This patterns is the most obvious one and tells us that the clustering is strongly based on the size and geometry of the bean.

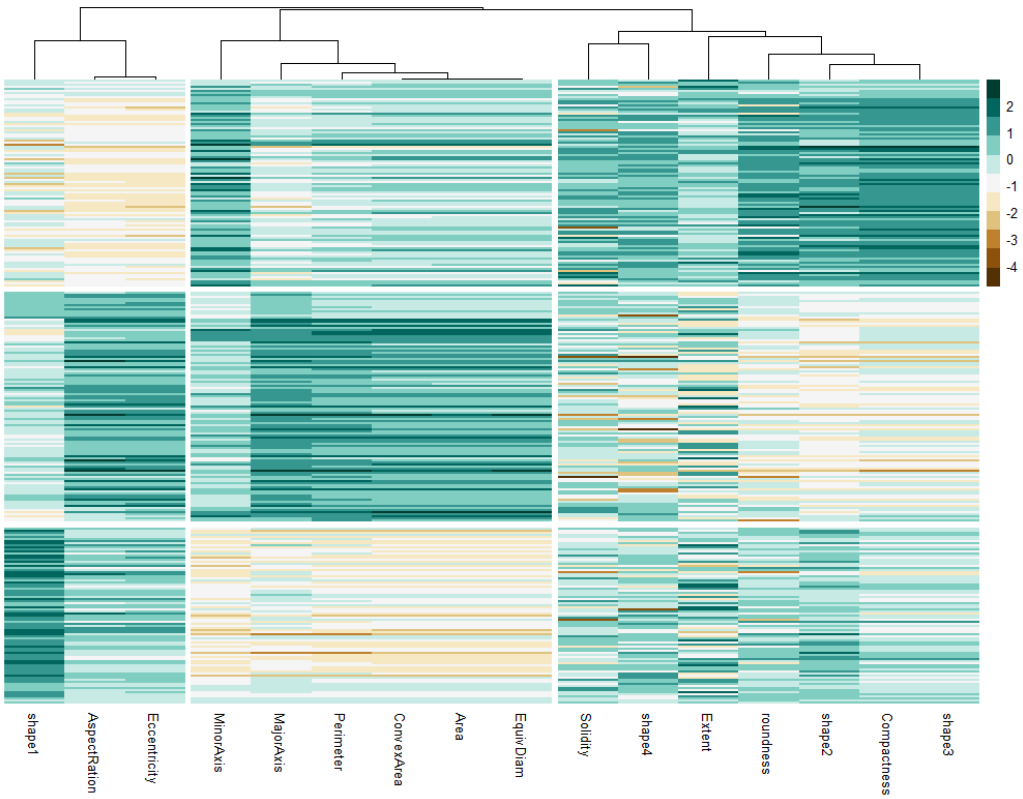


Figure 15: Heat map where variables and observations are sorted according to their clusters. Agglomerative clustering with complete linkage for the variables and k-means clustering for the observations, the gaps distinguishes the clusters.