Data Mining Project, Part 2
Breast Cancer

Tymoteusz Cieślik, Jakub Błażejewski

24.01.2022

1 Introduction

After descriptive analysis of data and classification along with detailed accuracy assessment, in the second part of our Data Mining project we focus on cluster analysis with quality assessments and dimensionality reduction.

We would like to remind you that we are analysing the data set provided by UCI Machine Learning called Breast Cancer Wisconsin (Diagnostic). The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

2 Theory and used methods

Most of the methods we used for our analysis are those learnt on the course Data Mining. Specifically they were:

- cluster analysis
  - k-means algorithm
  - PAM algorithm
  - AGNES algorithm
  - DIANA algorithm
- quality assessment of cluster analysis
  - external validation
  - internal validation
- dimensional reduction
  - PCA
  - MDS

2.1 k-means algorithm

The k-means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number k of clusters to be specified. This algorithm divides a set of N samples into k disjoint clusters C, each described by the mean μ_j of the samples in the cluster. The means are commonly called the cluster "centroids".

The k-means algorithm can be interpreted as an iterative solution of the minimization of the the inertia criterion, or within-cluster sum-of-squares criterion:

$$\tilde{W}(C) = \sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

It scales well to large number of samples and has been used across a large range of application areas in many different fields. All known variants of k-means algorithms converge. What is more, this algorithm has low computational complexity. Nevertheless, the convergence to the optimal solution is unfortunately not guaranteed and this algorithm has some issues with outliers.

2.2 PAM algorithm

The PAM (Partition Around Medoids) algorithm is a generalization of k-means algorithm. While k-means tries to minimize the within cluster sum-of-squares, this algorithm tries to minimize the sum of distances between each point and the "medoid" of its cluster, which is a data point with the least total distance to the other members of its cluster.

We search for k representative medoids by minimizing the following criterion:

$$W(C) = \sum_{i=1}^{n} d(x_i, m_{j(i)}),$$

where $m_{j(i)}$ denotes medoid (cluster center) closest for observation x_i and $j = 1, ..., k$. Unlike the k-means method, this algorithm can be used for features of any type (including qualitative). However, in case of large data sets, it is better to use modification of this algorithm called CLARA.

2.3 AGNES algorithm

The AGNES (AGLomerative NESting) algorithm is the most common type of agglomerative hierarchical clustering used to group objects in clusters based on their similarity. Here at the begging each observation is a separate cluster. We search for two closest clusters and merge them until one large cluster with all the objects is created. The result is a tree-based representation of the objects, named dendrogram.

- single linkage - the nearest point algorithm, where for all points i in cluster u and v in cluster:

$$d(u, v) = \min dist(u[i], v[j])$$

- complete linkage - the farthest point algorithm, where for all points i in cluster u and v in cluster:

$$d(u, v) = \max dist(u[i], v[j])$$

- average linkage - the average inter-cluster distance algorithm, where for all points i and j in cluster where $|u|$ and $|v|$ are the cardinalities of clusters u and v, respectively:

$$d(u, v) = \sum_{i,j} \frac{d(u[i], v[j])}{(|u||v|)}$$

2.4 DIANA algorithm

The DIANA (DIvisive ANAlysis) algorithm is type of divisive hierarchical clustering, where we start with all observations in one cluster. At each time step, the largest available cluster splits into two smaller clusters until finally all clusters, comprise of single objects. The algorithm finds an object, which has the greatest average distance to the rest of the objects in a set.

The results can be shown also on a tree-based representation named dendrogram. This algorithm is computationally expensive and hence not suitable for large data sets.

2.5 Quality assessment of cluster analysis

A good clustering method will produce high quality clusters. It is always good to evaluate the clustering results, as they can be significantly affected by the algorithm or parameters choice.

First type of quality assessment are external indices and they refer to external information. What is important, this information cannot be used before in the process of partitioning into clusters. We can distinguish different external indices that we use:

- partition agreement measure - comparing to what extent two partitions agree.

Second type of quality assessment are internal indices and they only refer to the properties contained in the data. Also it is calculated on the basis of the same data that were used to find clusters. We can distinguish different internal indices such as:

- silhouette index - for assessing compactness and spatial separation of clusters
- Dunn index - for identifying compact and well separated clusters
- connectivity - for reflecting the degree of connectedness of the clusters
- stability - for assessment of stability of clusters.

2.6 Dimensionality reduction

When the dimensionality of the data increases, meaning that the data becoming increasingly sparse in the input space, analysis becomes significantly more difficult. Many clustering may result in reduced classification accuracy or poor quality clusters because of not enough training objects to build an accurate predictive model or less meaningful fundamental definitions of density and the distance between points in higher dimensions. We should use the preprocessing methods to clear the data by feature selection or feature extraction.

In our analysis we use two methods:

- PCA - Principal Components Analysis, where we select the first few main components that explain a certain fraction of total data variability. Here each component needs to be a linear combination of the original variables, all components are orthogonal and ordered.
- MDS - Multidimensional Scaling, where we create a map displaying the relative positions of a number of objects having a table of the distances between them. Here we try to preserve another important property of that is original distances between objects by finding a projection of the data to a lower-dimensional space.

3 Results

3.1 Clustering

In this part of the project we will deal with the problem of clustering the data in different manners. We will be using methods such as k-means, PAM, AGNES etc. in order to group objects according to their similarity. In addition to that we would like to assess the quality of results using the tools that are intended for it. In order for the results to be more viable we introduce the new categorical variable to our data. It deals with the size of the tumor and assigns one of the expressions ("small", "medium", "big", "huge") in the accordance to the mean area. Having that and the division of the tumors into malignant and benign we generally would like to focus on the results of clustering our data into 2 and 4 parts to observe how accurate are they. Besides that we will naturally check which number of clusters for each method is the most optimal using the comparison of average silhouette index values.

3.1.1 k-means algorithm

The first method for clustering is k-means algorithm that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

If we set the parameter k to 2 and impose the algorithm on the data set we can draw the plot of *concavity* vs *area* where there are two clusters of data labeled originally as malignant (M) or benign (B).

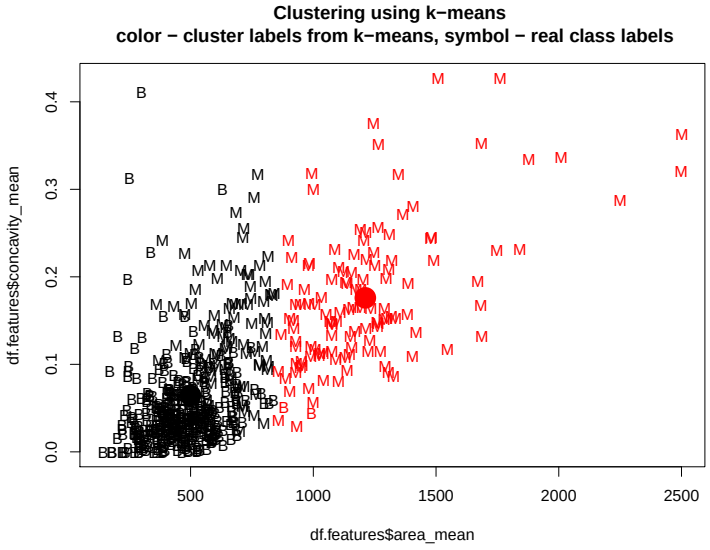

Figure 1: Scatter plot of the k-means clustering

As we see on the Figure 1 the majority of the points in the clusters are of the same value representing the danger of the cancer. In particular we can check the percent of the data that are precisely assessed to the correct part. It turns out that 84% of the values were assigned to the correct clusters what is a relatively good score. In addition we can take look at the Figure 2 that represents the cluster plot and we can observe that the areas of the clusters almost do not overlap and there is a visible border between them.

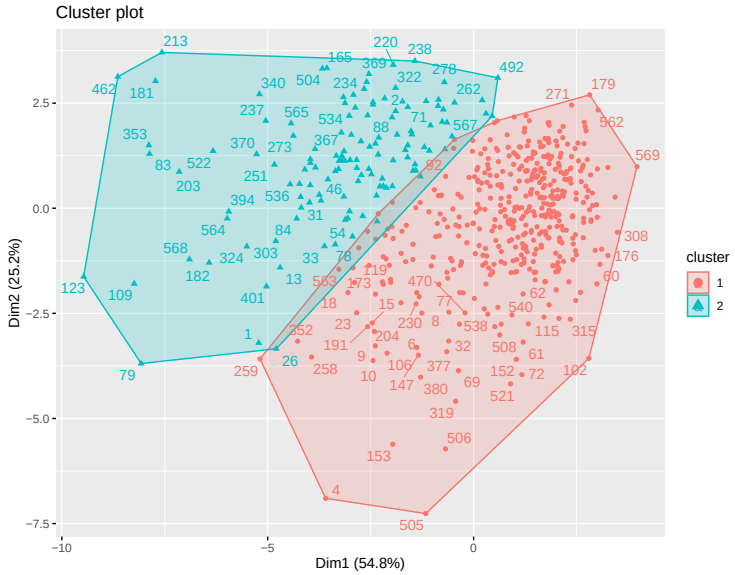

Figure 2: Cluster plot of the k-means clustering

When it comes to the results of the algorithm used for $k = 4$ we can find the same dependence on the scatter plot, where all the clusters consist in majority on one type of factor what can be seen on Figure 3 with the centers of the clusters marked as filled circles.

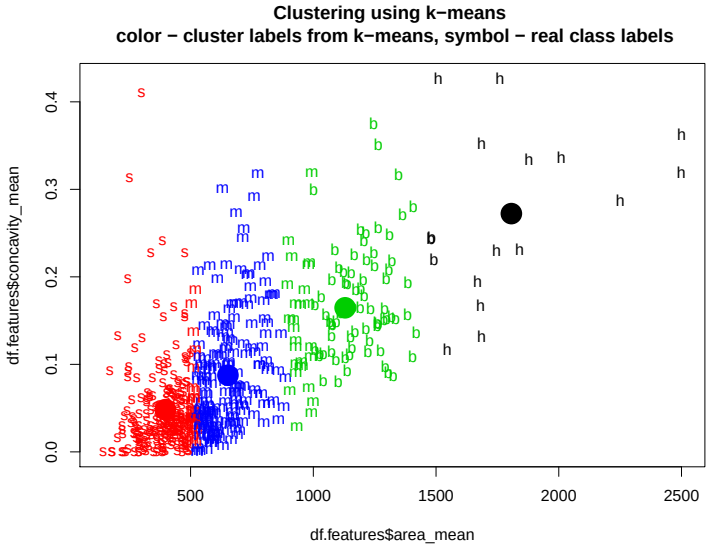

Figure 3: Scatter plot of the k-means clustering

This time the accuracy of assignment reaches about 89% what can be considered as a very good result. On top of it we can also present the cluster plot of the algorithm.

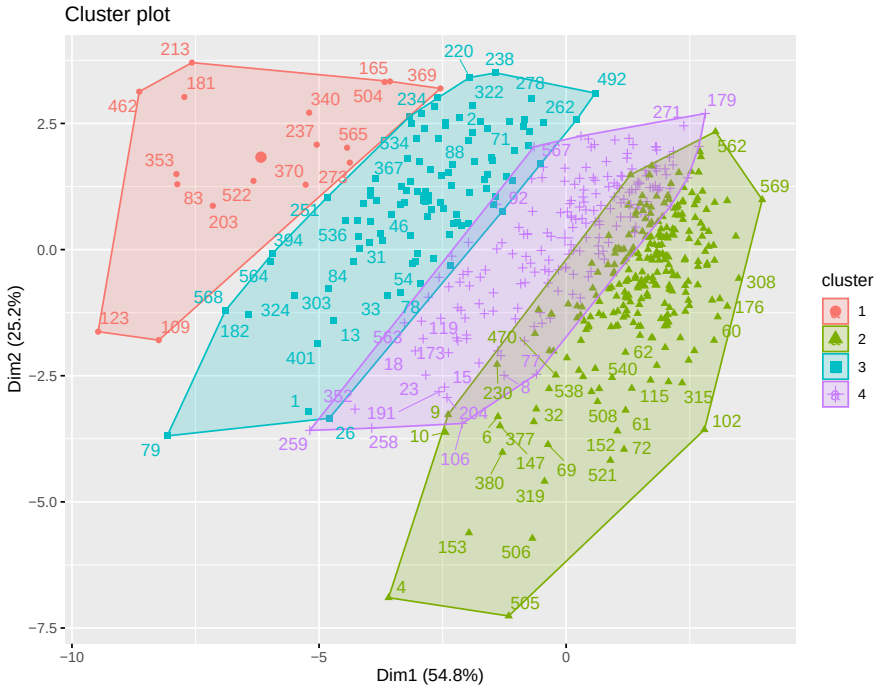

Figure 4: Cluster plot of the k-means clustering

On the Figure 4 we find out that despite the partial overlapping of the cluster areas the general assessment is quite proper.

Let us now focus on the checking what number of clusters would be optimal for k-means algorithm and our data. In order to do that we can explore the mean silhouette index values for each of the values of parameter $k \in [2, 10]$. These values are presented in the following table 5.

	silhouette_index
2	0.678054884702848
3	0.533378861058905
4	0.524045578918413
5	0.519950656101521
6	0.518292032067742
7	0.53997858715577
8	0.538466331200889
9	0.526413109577823
10	0.532508180818543

Figure 5: Table of mean silhouette values

It indicates that the value of k equal to 2 is clearly the best among others, but it is not enough to state that we have already chosen the best parameter for this method. The next things that we would like to analyse are the within-cluster and between-cluster dispersion. The value of k would be the one that is located the closest to the crossing of the lines on the chart of the dispersions presented on Figure 6.

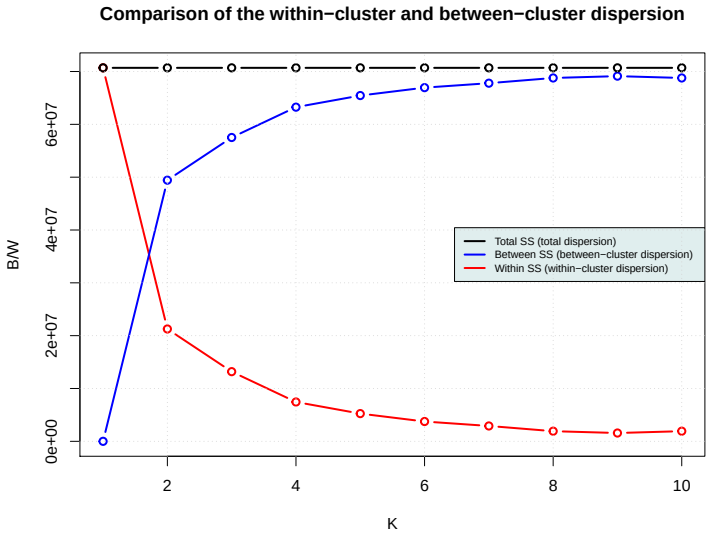

Figure 6: Plot of the dispersions in the clusters

It seems that also for $k = 2$ we can obtain the most optimal results, as the crossing point is just before the values of 2 on x axis. The last method of checking the quality of assessment will be the so called *elbowmethod* in which we are looking for a strong bend in the chart and try to indicate k, for which the within-cluster dispersion stops rapidly decreasing, and increasing the number of clusters does not yield much improvement what is shown on Figure 7, where we used *wss* method.

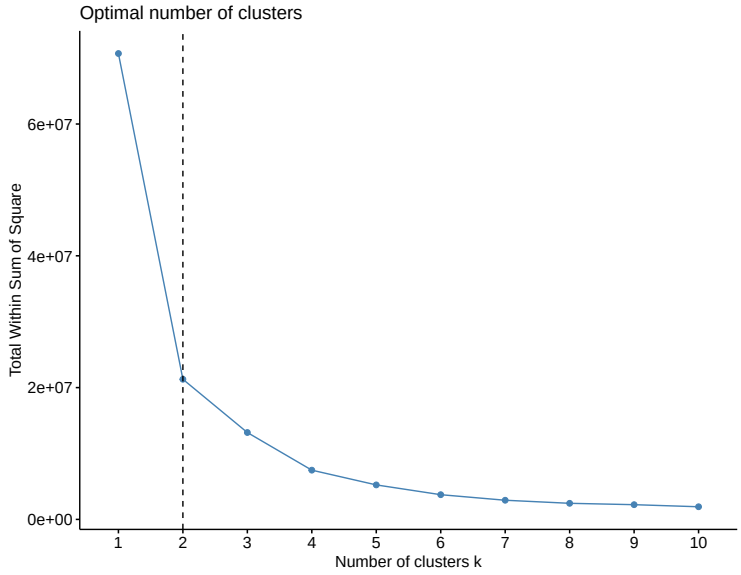

Figure 7: Plot of the elbow method

We can clearly see that the biggest bend is achieved at point 2, what makes this value the most optimal. Additionally we can also present the other methods – silhouette on Figure 8 and gap on Figure 9.

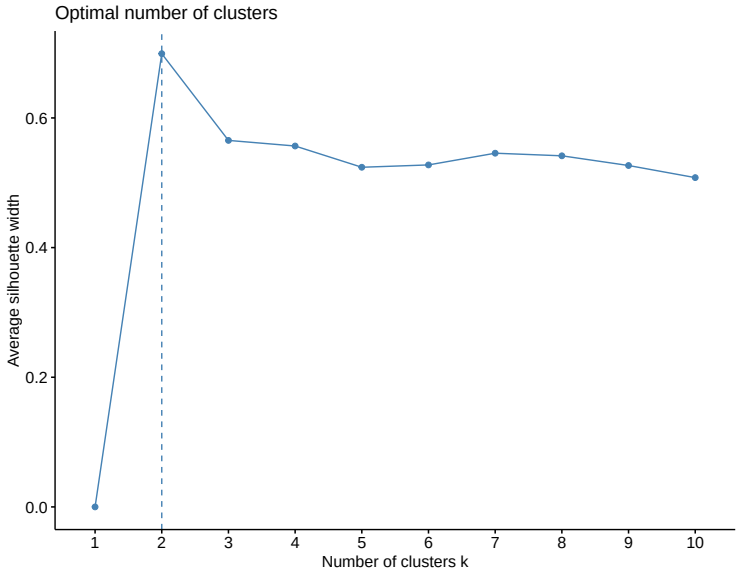

Figure 8: Plot of the silhouette method

It also indicates that the kest value of parameter k is 2.

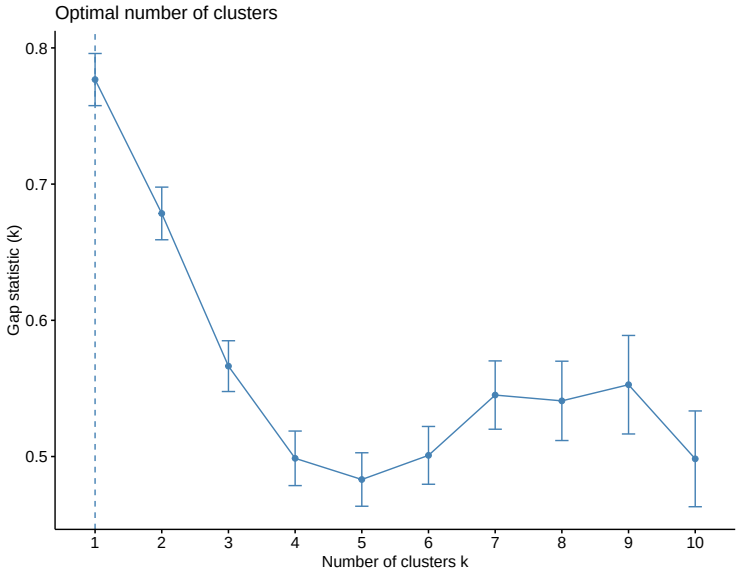

Figure 9: Plot of the silhouette method

In this case it says that one cluster is the best for the data but we can treat it as an insignificant observation, because all other methods indicated $k = 2$ as best. Finally, after performing the analysis and the quality assessment we can claim that the value of k equal to 2 is the most optimal for the k-means method and provides the best final results. Now we also have to check the other clustering methods to inspect whether they would return similar outcome.

3.1.2 PAM algorithm

Next method for clustering that we want to present is PAM algorithm (Partitioning Around Medoids) that is generalization of k-means method. Having the set of n objects we can select sequentially K representative objects (medoids), which are the initial centers of clusters, than we can minimize the sum of the distances of objects from corresponding medoids by replacing the current medoids by other (not yet selected) observations.

Let us set the parameter k to 2 and impose the algorithm on the analyzed data set. Now we can draw the plot of *concavity* vs *area* where the clusters of data are named originally as malignant (M) or benign (B).

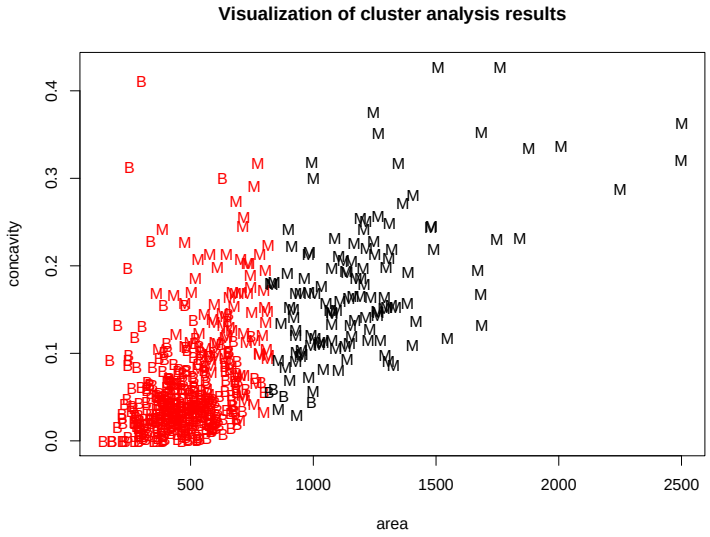

Figure 10: Scatter plot of the PAM clustering

On the Figure 10 we can see that the majority of the points in the clusters are of the same value representing the danger of the cancer. Again we can check the percentage of the data that are precisely assessed to the correct cluster. It turns out that it is also 84% of the values that were assigned correctly and it is a relatively good score. Let us also look at the Figure 11 that shows the cluster plot. We can here observe that the areas of the clusters overlap quite more than for the k-means method, but it is still quite good.

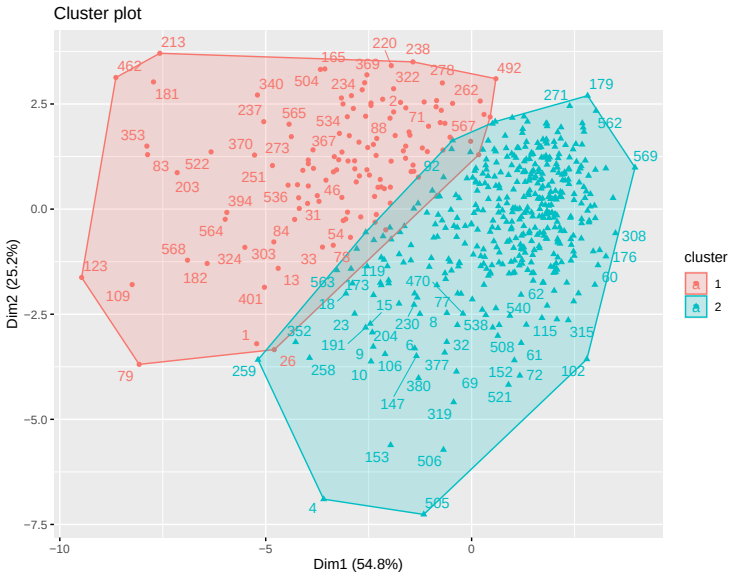

Figure 11: Cluster plot of the PAM clustering

When it comes to the results of the algorithm used for $k = 4$ we can find the same dependence on the scatter plot, where all the clusters consist in majority on one type of factor what can be seen on Figure 12 with the centers of the clusters marked as filled circles.

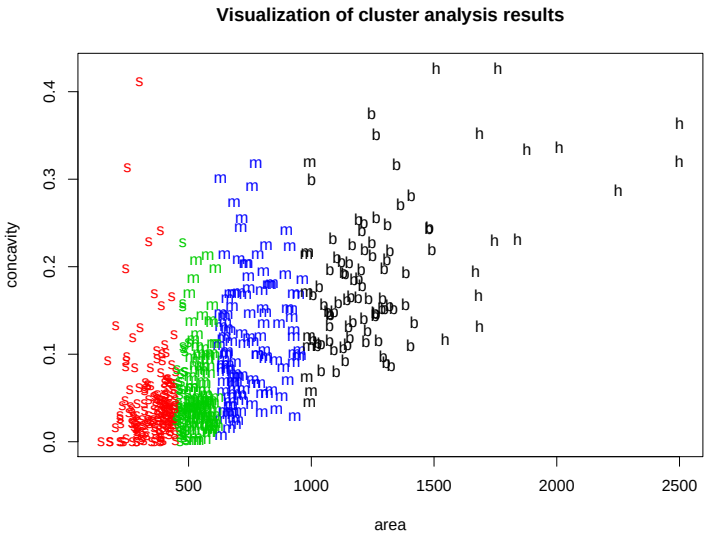

Figure 12: Scatter plot of the k-means clustering

This time the accuracy of assignment reaches about 86% what can be considered as a very good result. We also present the cluster plot of the algorithm. On the Figure 13 we find out that despite the partial overlapping of the cluster areas the general assessment is quite proper, but there are visible intersecting areas.

Figure 13: Cluster plot of the PAM clustering

Now let us focus on checking what number of clusters would be optimal for PAM algorithm used for our data. In order to do that we explore the mean silhouette index values for each of the values of parameter $k \in [2, 10]$. These values are presented in the following table 14.

	silhouette_index
2	0.694598960186602
3	0.539662727303898
4	0.499486243151599
5	0.521996442545453
6	0.527055946235206
7	0.546579238022712
8	0.523702253877569
9	0.517615793473133
10	0.538696407487823

Figure 14: Table of mean silhouette values

In this case it also indicates that the value of k equal to 2 is the best among others.
Another method of quality assessment would be the silhouette plot in the clusters, which can be seen in 15.

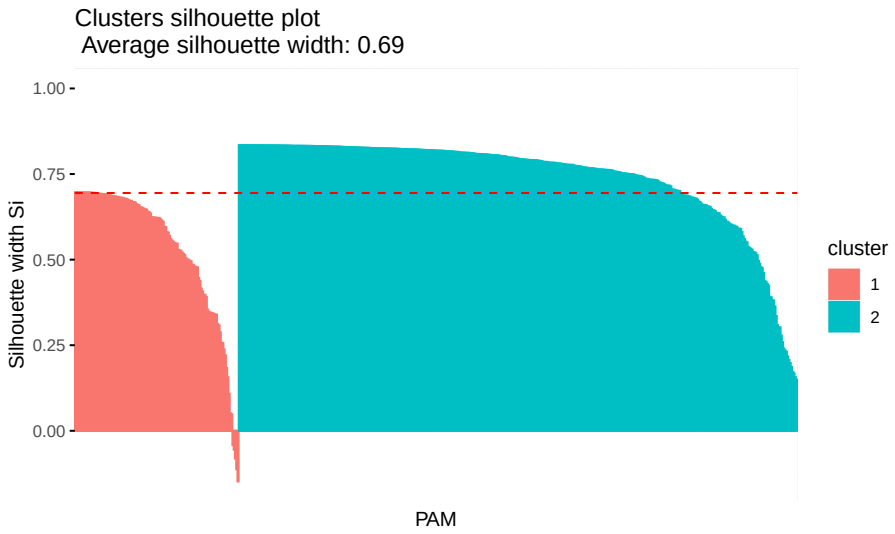

Figure 15: Table of mean silhouette values

We can observe that the silhouette values for almost every observation is positive what means that the clustering algorithm is working properly. The same situation can be visualized for $k = 4$ on 16.

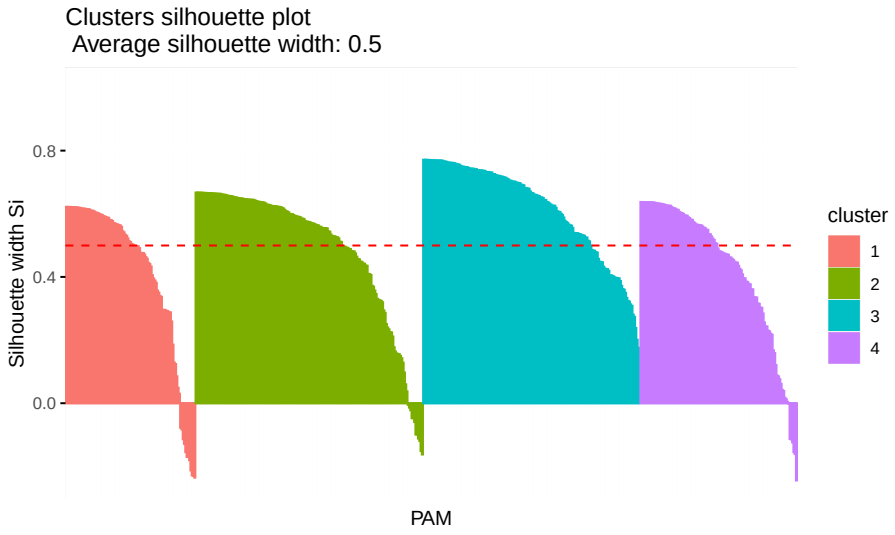

Figure 16: Table of mean silhouette values

This time the number of observations, for which the silhouette index is negative, is slightly bigger and we can suspect that it is more optimal for the k to be smaller than 4. Besides these outcomes it is not enough to state that we have already chosen the best parameter for this method. Checking the quality of assessment we will use again the so called *elbowmethod* in which we are looking for a strong bend in the chart and try to indicate k, for which the within-cluster dispersion stops rapidly decreasing, and increasing the number of clusters does not yield much improvement what is shown on Figure 17, where we used wss method.

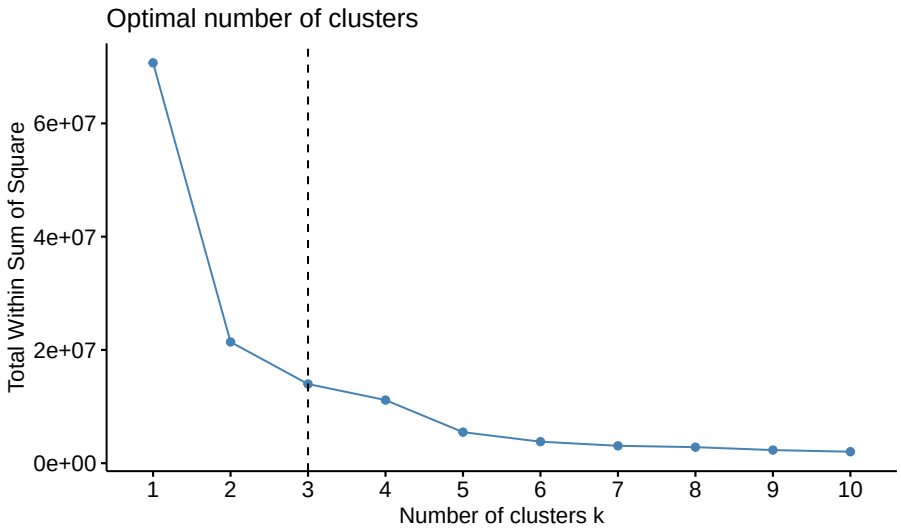

Figure 17: Plot of the elbow method

We can see that the biggest bend is achieved at point 3 and this is different from the previous one. Let us check and present other methods – silhouette on Figure 18 and gap on Figure 19.

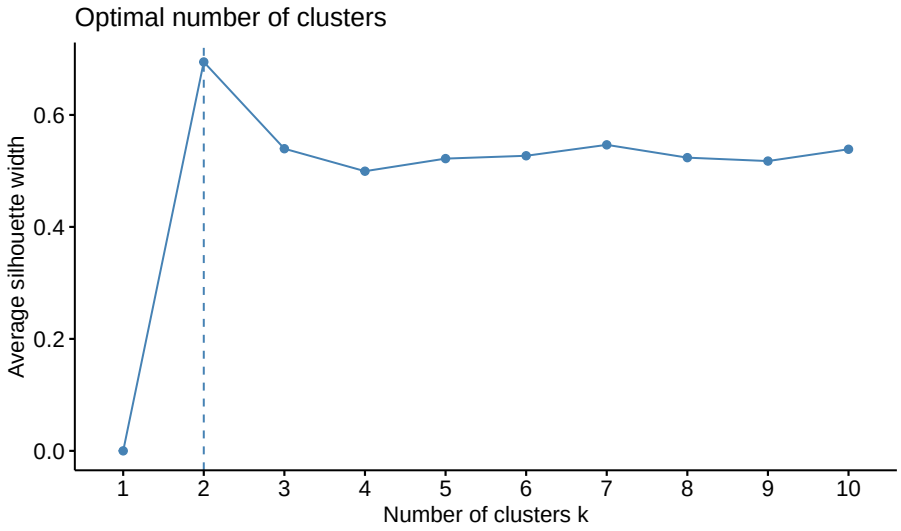

Figure 18: Plot of the silhouette method

It indicates that the best value of parameter k is 2.

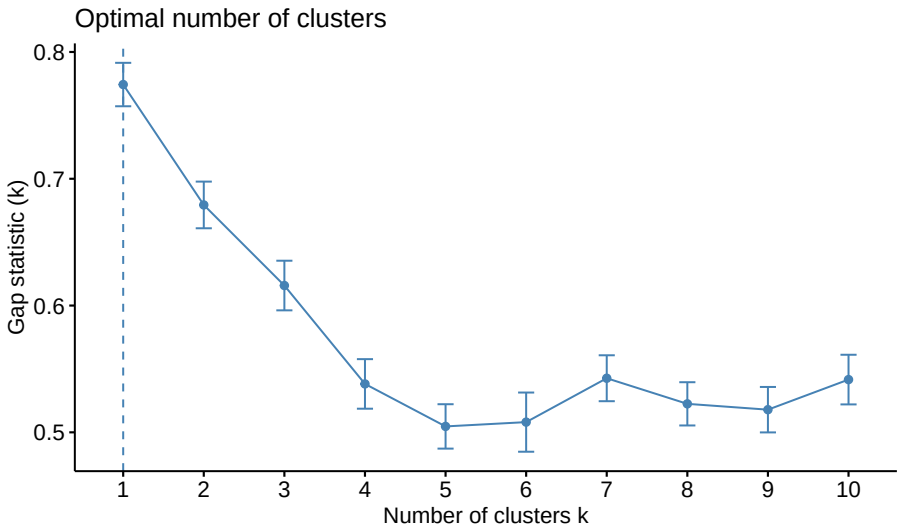

Figure 19: Plot of the silhouette method

Here we can see that only one cluster is the best for the data, so this is third different value we obtained for this method. Nevertheless almost all other methods indicated $k = 2$ as best and in average we can say it is 2 that is optimal number. Hence, after performing the analysis and the quality assessment we can claim that the value of k equal to 2 is the most optimal for the PAM method and provides the best final results, but it was not that clear as for the k-means method.

3.1.3 Agglomerative nesting – AGNES algorithm

In this part we will be describing the AGNES method where at first, each observation is a small cluster by itself. Clusters are merged until only one large cluster remains which contains all the observations. At each stage the two nearest clusters are combined to form one larger cluster. During this analysis we will focus on the three types of the linkage in this algorithm: average, simple and complete

Again, similarly as in the previous methods we can present the correctness of matching the observations into clusters in comparison with their type. The results are in the table 20

	k=2	k=4
average	0.632688927943761	0.630931458699473
single	0.743409490333919	0.557117750439367
complete	0.437609841827768	0.896309314586995

Figure 20: Table of the percent of clustered observations matched correctly with their labels

We notice that for average and single linkage the parameter $k = 2$ performs better, but if we consider complete type, the bigger value of k gives substantially better results.

If it comes to the cluster plots for AGNES method, they are presented in 21 and 22.

Figure 21: Cluster plot of AGNES method

It shows a little bit different way of clustering for this method as one of the collections is clearly smaller and contains much less observations, but the areas of the clusters hardly overlap.

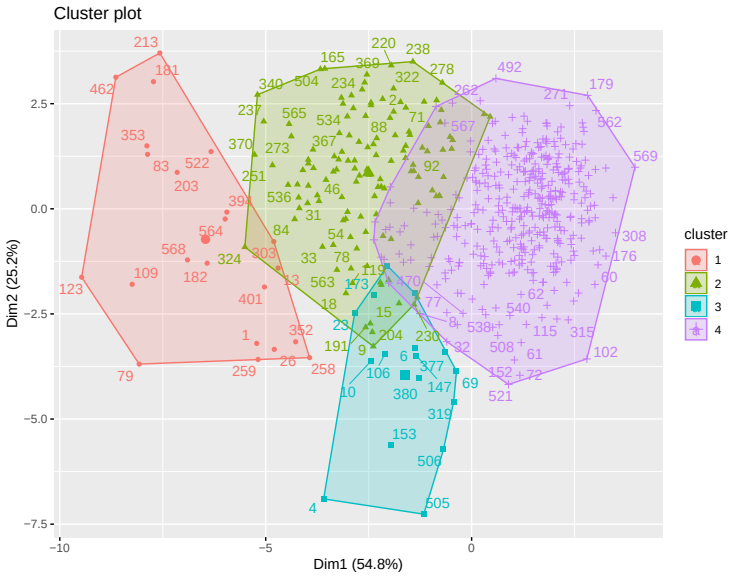

Figure 22: Cluster plot of AGNES method

When we set the number of clusters to 4, the sets of values are also slightly different in comparison to the previously used methods but clearly they almost do not intersect and each group is well separated.

We would also like to analyse which value of the parameter k is the most optimal. In order to check that we created a table with the average silhouette index for each k presented in 23.

	average	simple	complete
2	0.837269350667102	0.888074589845863	0.670709545051377
3	0.71895324606742	0.570694661959725	0.615756222021903
4	0.68333016143821	0.415780775084566	0.528397532150154
5	0.618486214261303	0.437207500379894	0.604865748356035
6	0.580451885723989	0.496521030399422	0.597495666176837
7	0.52660260278379	0.544455587351673	0.571573010765557
8	0.538202706249977	0.568360507902826	0.554227028914229
9	0.539976327998952	0.487960619671986	0.543086744182456
10	0.544388209879956	0.375626614078193	0.542502541666403

Figure 23: Table of mean silhouette values

Each of the columns in the table indicates that the parameter k equal to 2 is the most optimal and gives the best results. Now we can look into each of the linkage method separately, beginning with the average one.

- Average linkage As the first quality assessment we will take into consideration the silhouette plot for $k = 2$, which is shown in 24

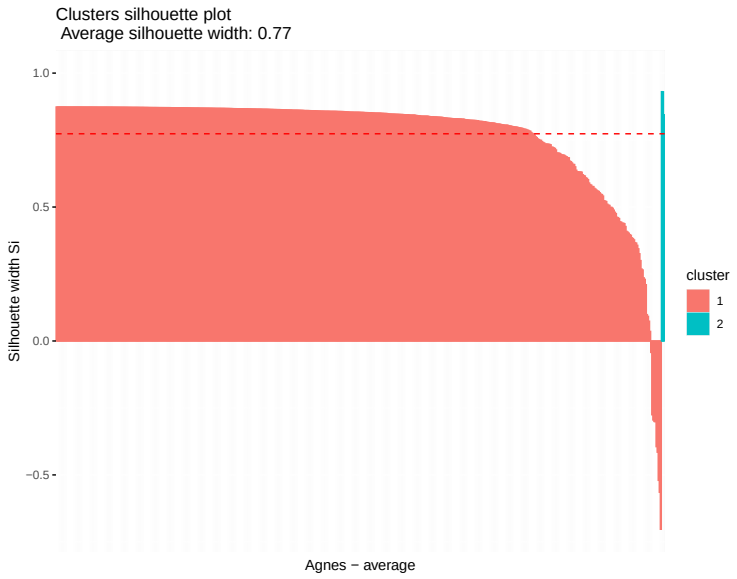

Figure 24: Silhouette plot of AGNES average linkage

We can see that almost every observation is included in one cluster and the small rest makes up the second one. Despite that the values are mainly positive.

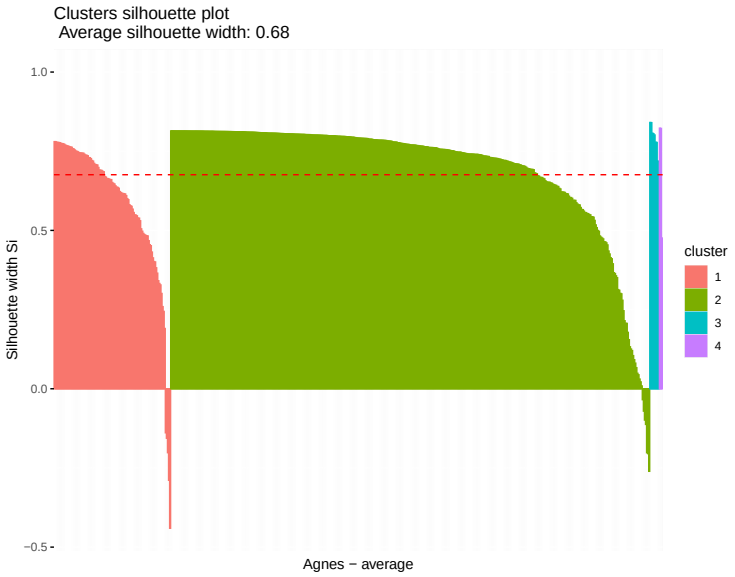

Figure 25: Silhouette plot of AGNES average linkage

Similar situation occurs when we deal with $k = 4$. On figure 25 there are 2 substantial clusters and 2 very small ones, but it does not change the fact that the values of silhouette indices are relatively high what makes it a reliable tool.

Additionally we will present the dendrogram of this method, what could let us find out how the clusters are being created. The figures 26 and 27 shows the division into 2 and 4 clusters respectively.

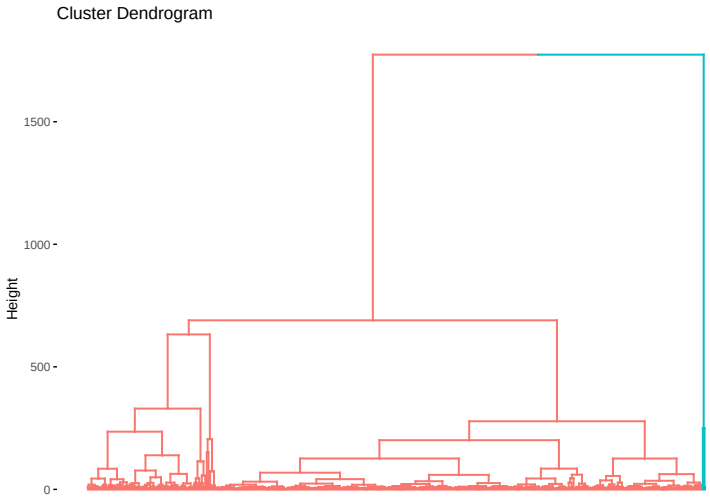

Figure 26: Dendrogram of AGNES average linkage

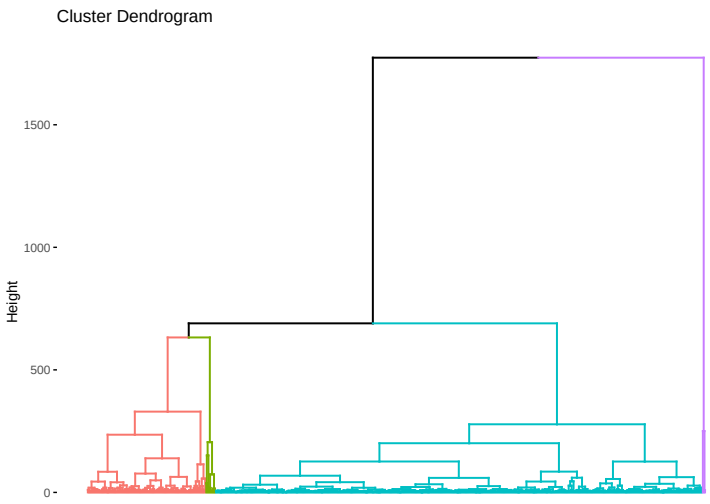

Figure 27: Dendrogram of AGNES average linkage

As it was already indicated in the silhouette plots, some clusters are being created with very few observations what may have caused some problems with the correct matching to the labels. Despite that the division makes sense because in the type of average linkage the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance what makes some clusters smaller than others.

- Single linkage

  In similar fashion as before, we will present the silhouette plots for $k = 2$ and $k = 4$. They are shown in figures 28 and 29.

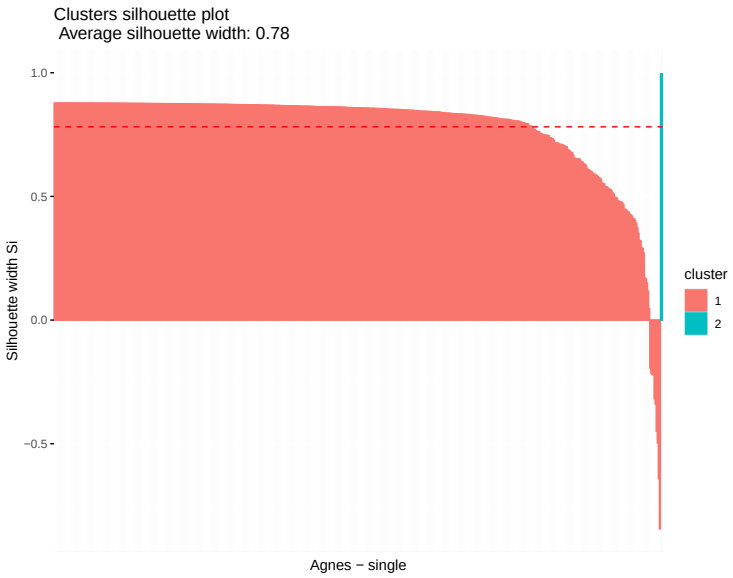

Figure 28: Silhouette plot of AGNES average linkage

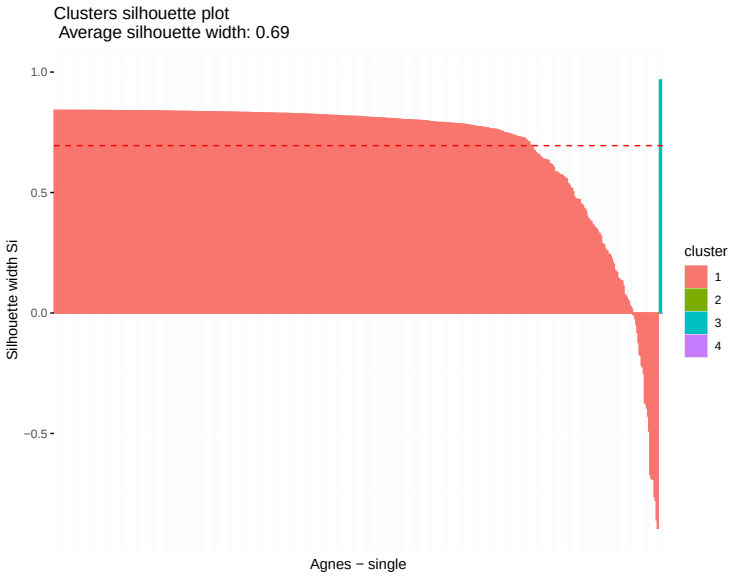

Figure 29: Silhouette plot of AGNES average linkage

What is interesting, there is only one observation in the two of the clusters when the parameter k is set to 4. We can assume that this method is not a reliable tool for our data as it distorts the outcomes.

In addition to that we could also analyze the dendrograms but almost every branch of it will be colored in the same tone as the observations are included mainly in one cluster. The dendrogram for $k = 2$ is presented on figure 30 .

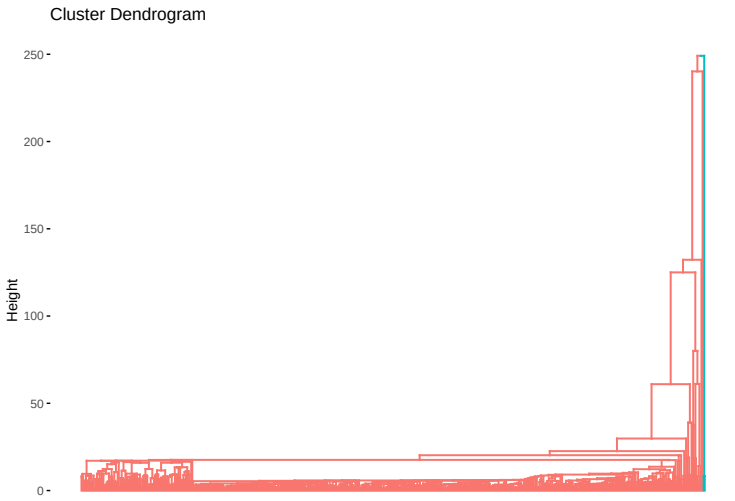

Figure 30: Dendrogram of AGNES single linkage

The same situation happens when we analyse the dendrogram with $k = 4$, what is presented on figure 31.

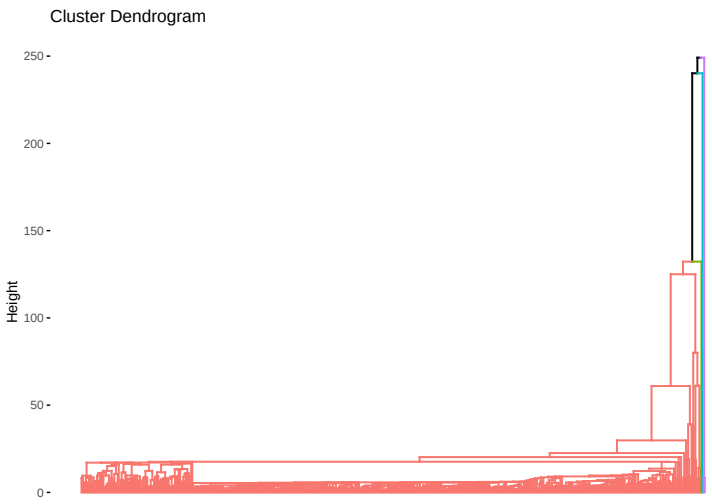

Figure 31: Dendrogram of AGNES single linkage

As the single type works very badly for our data set it also cannot be recognized as a proper tool for quality assessment. Let us find out how the last type of linkage behaves.

- Complete linkage

  This time we will also present the silhouette plots visualized on figures 32 and 33

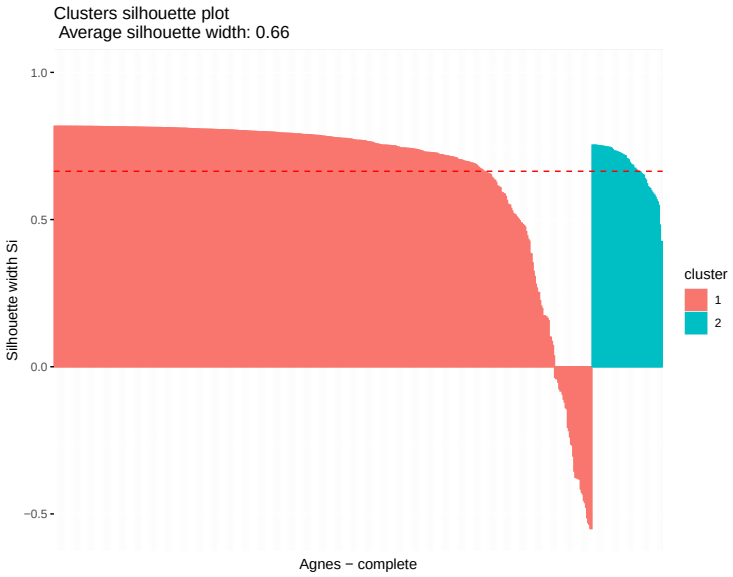

Figure 32: Silhouette plot of AGNES complete linkage

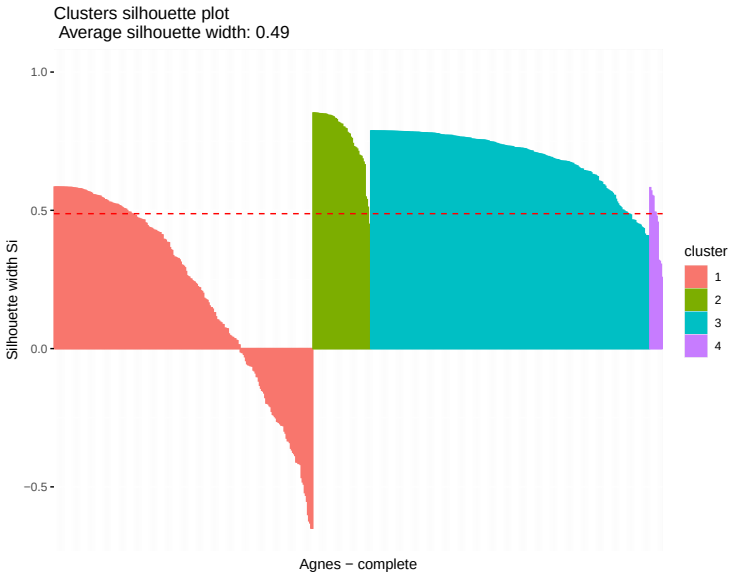

Figure 33: Silhouette plot of AGNES complete linkage

As we can notice, the clusters are more balanced than before, but for $k = 2$ and particularly for $k = 4$ there are many observations for which the silhouette indices take really small values what decreases the average silhouette value substantially.

For the analysis of the dendrograms we expect that they will not be as one-sided as before and the different branches will belong to different clusters. The answer can be seen on figures 34 and **??**.

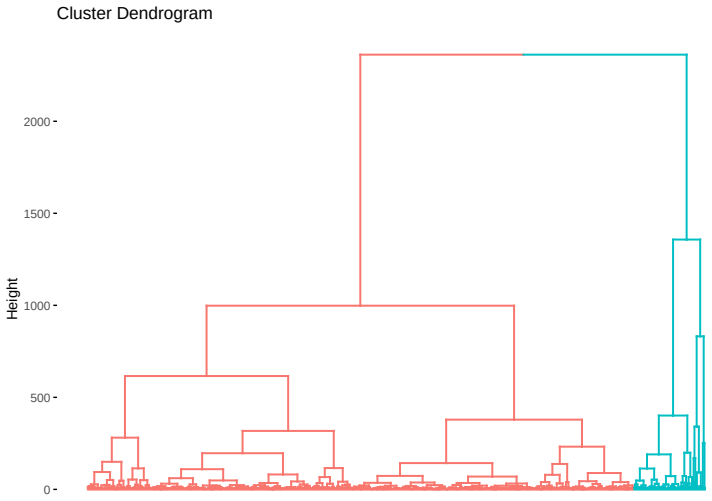

Figure 34: Dendrogram of AGNES complete linkage

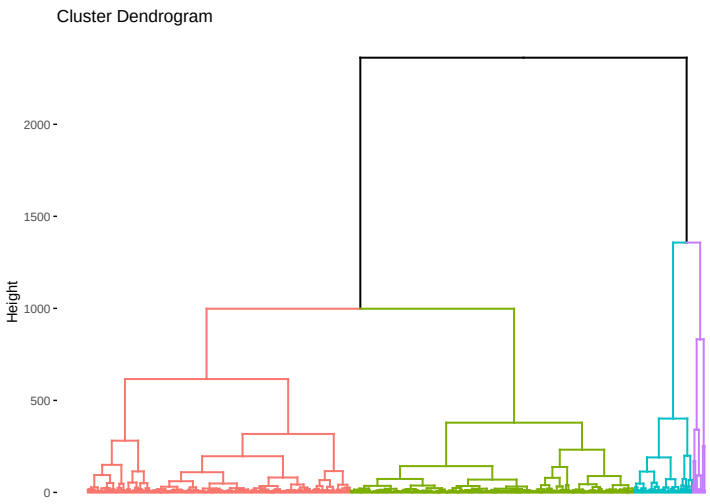

Figure 35: Dendrogram of AGNES complete linkage

As we expected, this time the dendrograms are way more balanced and we can clearly notice the boundaries between clusters what was not too obvious for previous linkage types.

Finally we can assume that the best value of parameter k would be 2, as the mean value of silhouette indices for every type of AGNES method indicate it.

3.1.4 Divisive Analysis – DIANA

In the last method that we are going to analyse the DIANA algorithm, where all data points are initially assigned a single cluster. Further, the clusters are split into two least similar clusters. This is done recursively until clusters groups are formed which are distinct to each other. Similarly as in the previous methods we will introduce the correctness of fit of the clusters to the labels for $k = 2$ and $k = 4$. It appears that for smaller k it is approximately 83% and for bigger 95% what gives us extremely good matching. Next up, we can present the cluster plots drawn on figures 36 and 37 for k equal to 2 and 4 respectively.

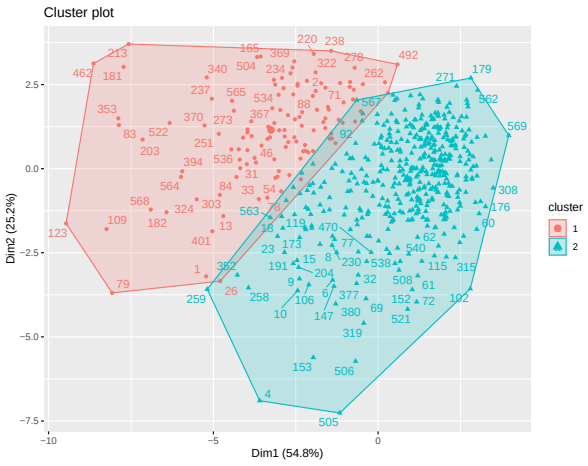

Figure 36: Silhouette plot of DIANA method

Figure 37: Dendrogram of of DIANA method

In case where there are 2 clusters, the division is clear and the matching with labels reaches 84%, the same occurs when k is equal to 4, as the clusters' areas do not overlap, but the correctness of fit is at the level of 56%. What is interesting, if we analyse the mean silhouette indices for each value of k, it seems that it achieves the highest value at $k = 4$ despite the low level of matching correctness, what can be seen in table 38.

	average
2	0.651079861696977
3	0.620731848627635
4	0.673717330670941
5	0.604949099142416
6	0.589571672583484
7	0.553794019784841
8	0.542470778953705
9	0.52338860840281
10	0.520407423959724

Figure 38: Table of mean silhouette values

Now, we can present the silhouette plots and dendrograms in order to observe how the silhouette indices look like for each observation and how does it affect the branches of the 'tree'. As first, we look at the case where k is equal to 2. Figure 39 shows that there are no observations or which the silhouette index would be negative and on 40 we see the division into clusters. It seems that the size of both groups are reasonably similar.

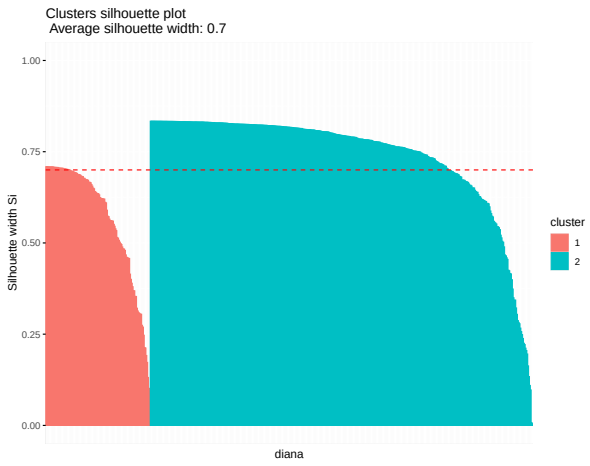

Figure 39: Silhouette plot of DIANA method

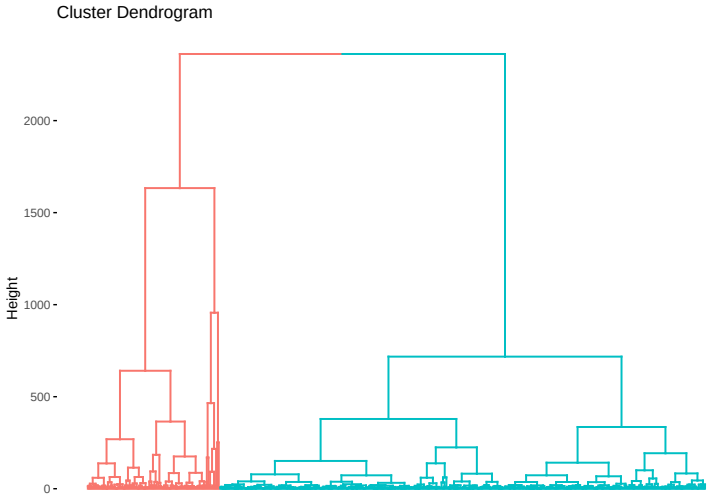

Figure 40: Dendrogram of of DIANA method

If it comes to the $k = 4$, the silhouette plot and dendrogram show that the sizes of the clusters are not equal – two of them are of negligible size. On figure 41 we notice that only small fraction of the observations have negative index. The dendrogram 42 shows the littleness of the two smallest clusters and gives us an information of how these clusters were created.

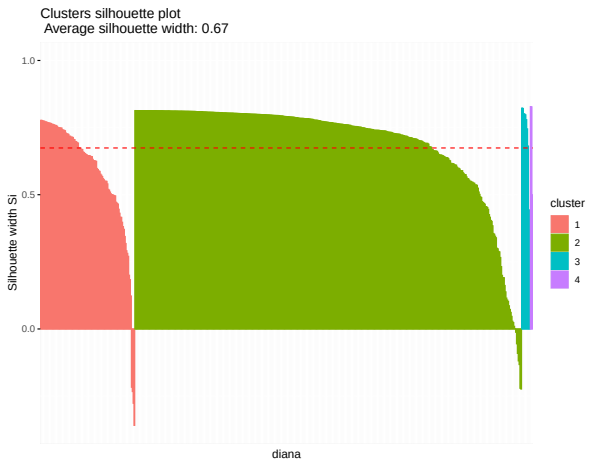

Figure 41: Silhouette plot of DIANA method

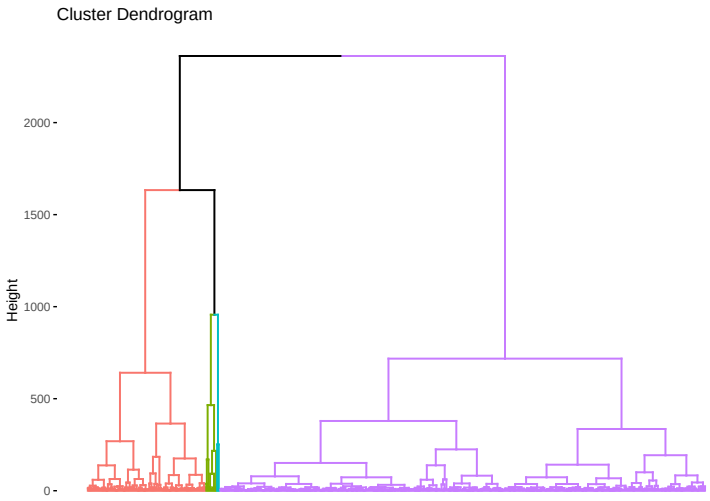

Figure 42: Dendrogram of of DIANA method

All in all, the results for DIANA algorithm are quite diverse, and we cannot clearly state that one of the values of parameter k is the most optimal. The choice would balance between the values of 2 and 4 depending on the initial criteria of judgment.

3.1.5 Quality assessment of cluster analysis

In this section we would like to focus on the advanced quality assessment of the different methods of clustering. We will start with the analysis of the partition agreement between each of the four methods of clustering that we presented before. On the 43 we can see the percent of the equally assigned observations to the clusters for each value of parameter k for the combination of each two algorithms.

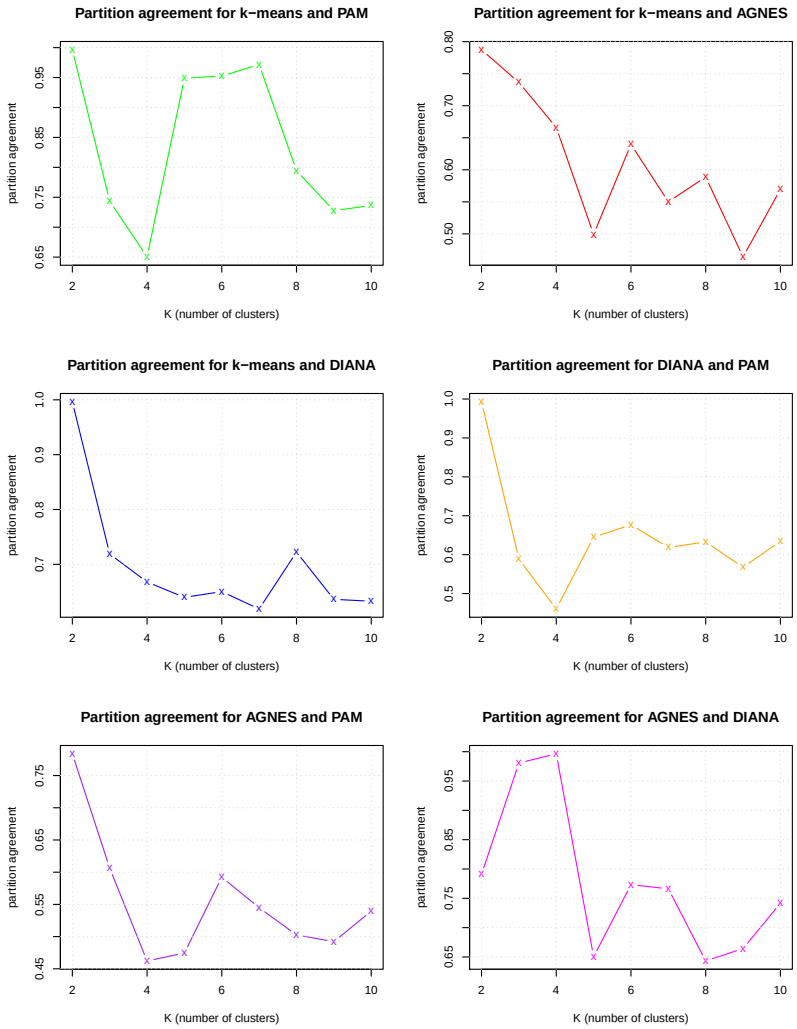

Figure 43: Partition agreement plots

- K-means and PAM – the highest level partition agreement is reached when k is equal to 2, but we also achieve quite good results for $k \in \{5, 6, 7\}$,
- K-means and AGNES – The highest values of this index are achieved with smaller values of k, e.g $k = 2$,
- K-means and DIANA – similarly, the best agreement is at $k = 2$,
- DIANA and PAM – same situation as before, $k = 2$,
- AGNES and PAM – the highest value for partition agreement is reached when $k = 2$,
- AGNES and DIANA – this time the situation is a little bit different, as the best value of partition agreement is achieved when $k = 4$.

From this we can draw conclusions that the methods are consistent witch each other when parameter k is equal to 2 and they differ more and more while we increase it.

Other way to determine the best number of clusters is the use of the NBCluster package in R that provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. Its appliance on the data frame of observations resulted in such an outcome presented on 44.

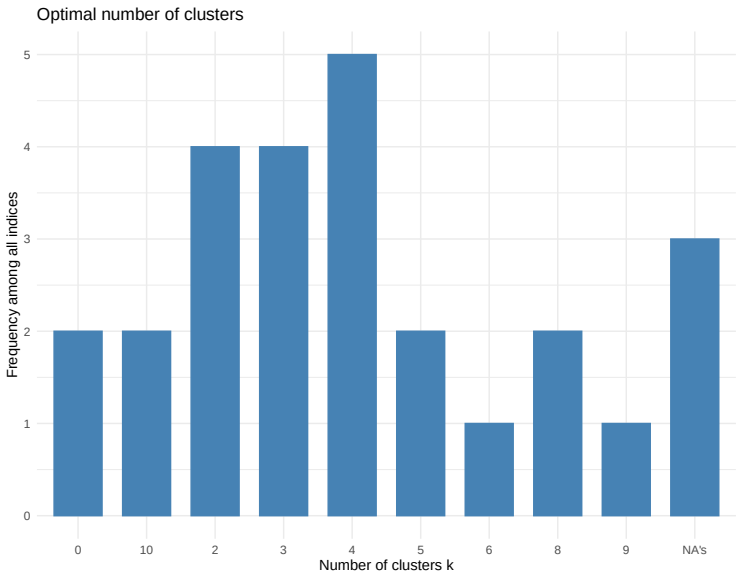

Figure 44: Frequency of occurence of values k

It seems that this method has decided that if $k \in \{2, 3, 4\}$ the clustering returns the most desired outcome. Let us also present the same bar plot 45 but in this case as the argument of the function we use the dissimilarity matrix.

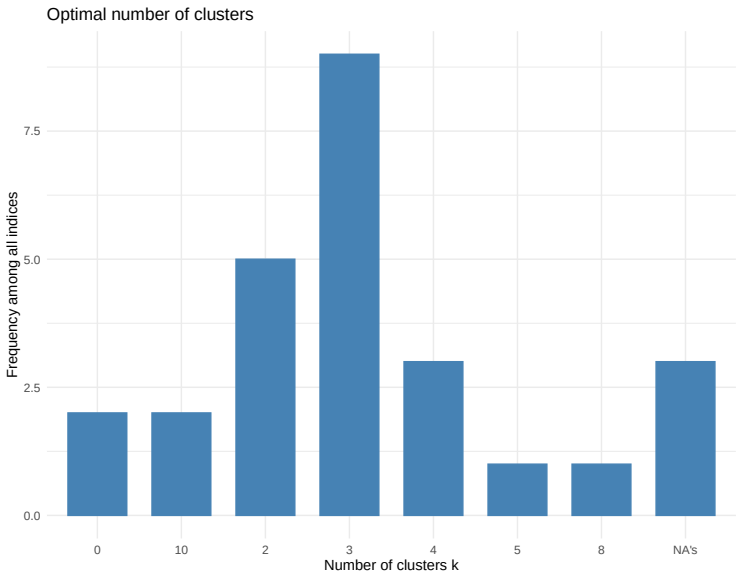

Figure 45: Frequency of occurence of values k

It appears that for this type of algorithm, clearly the best number of clusters is 3. It seems quite unnatural, as for all the methods that were shown previously, $k = 3$ did not give the most optimal results.

Another approach to assess the quality of the clustering methods is the analysis of the internal validation plots, where we can compare the Connectivity and Silhouette and Dunn indices of different methods. From the figure 46 we can state that the optimal value of k is 2, as for this point the indices for Connectivity are the lowest and for Silhouette – the highest.

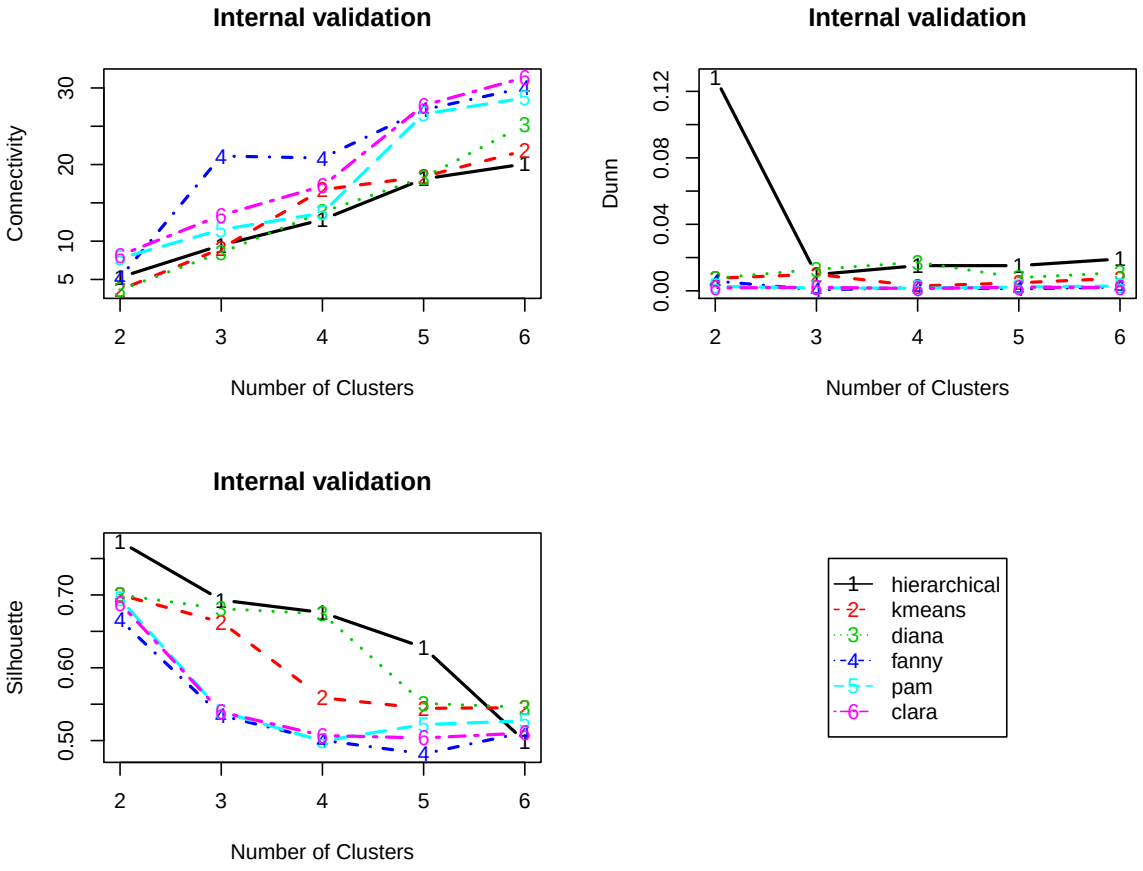

Figure 46: Internal validation plots

As well as the internal validation, we can use the stability validation in order to also pick the best number of clusters. The plots on figure 47 show us that for this measure, the most optimal k's would be the bigger values like 4 or 6.

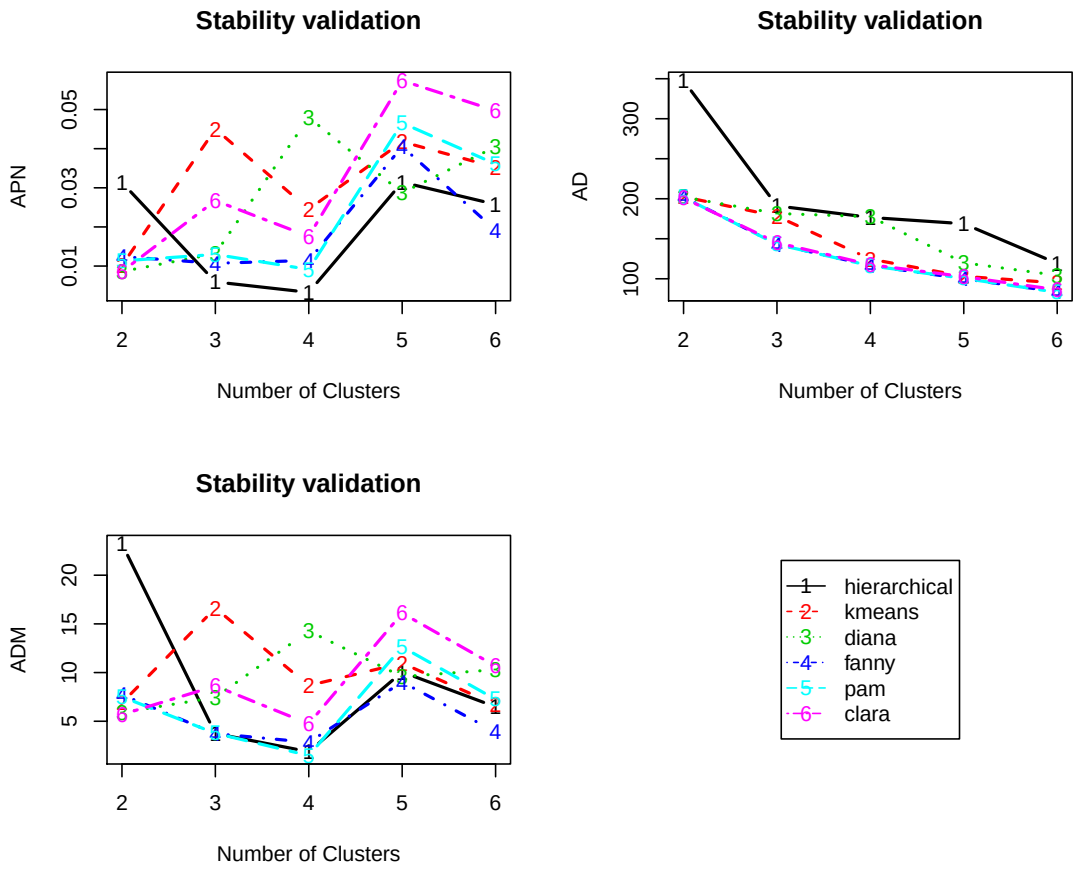

Figure 47: Stability validation plots

To sum up, we used different clustering methods, which gave us really valuable results and according to them, in general the best and most optimal of parameter k is 2, although it was not chosen unanimously. In some cases, the clustering algorithms were not efficient, considering only a single observation as a separate cluster. Despite that, if we take a look at the percentage of correctness of matching the observations' clusters and labels, for each algorithm we obtained results of over 60% what excludes total randomness. The analysis of Silhouette indices also indicate that the optimal value of k is 2, and other validation methods in general confirm it.

3.2 Dimensional reduction

This problem consists of the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. In this part we would like to use methods that allow to perform such analysis and investigate their general impact on the data.

3.2.1 Principal Component Analysis – PCA

It is the main linear technique for dimensionality reduction. It performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. This is also the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. If we impose it on our data we can at first compute the cumulative variance of each subsequent component, what is shown on 48. For each component the variance increases by a smaller amount reaching about the value of 1 – the first component has the biggest variance and last – the smallest.

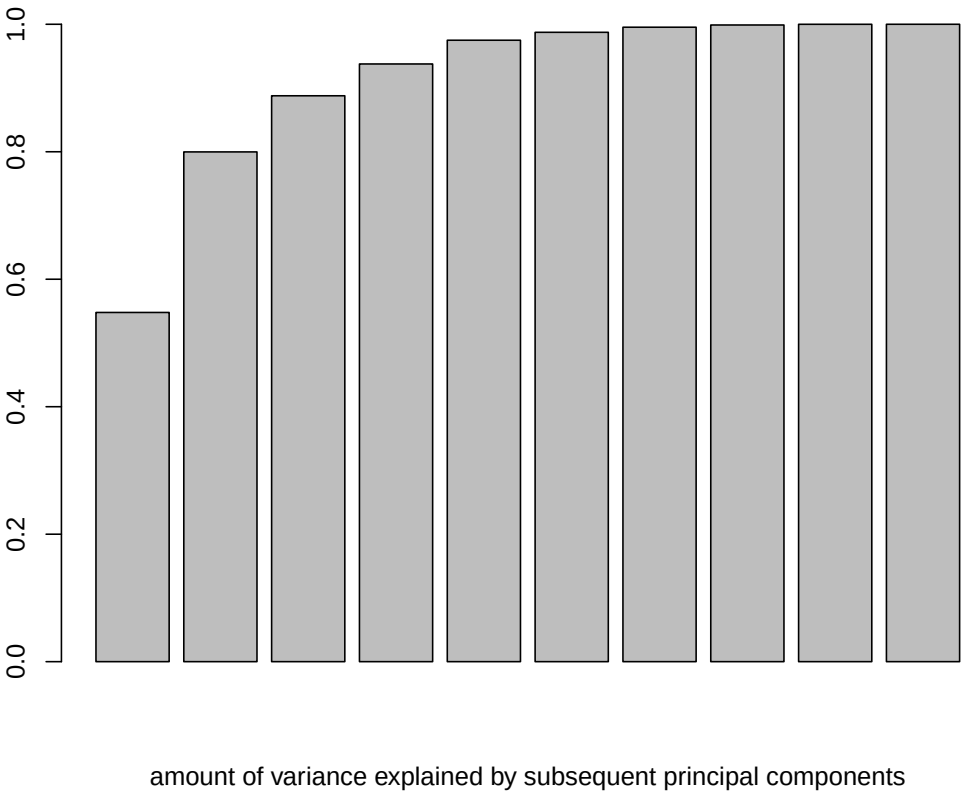

Figure 48: Plot of cumulative variance

It is explained in a better way if we take a look at 49. It shows the contribution to the cumulative variance of each component. From it we could state that the optimal number of components could be somewhere between 3 and 5.

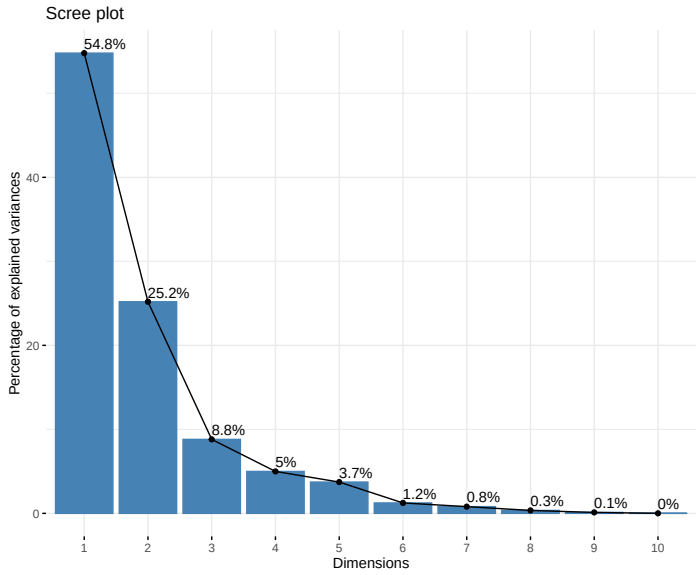

Figure 49: Plot of variance contribution

Next up, we can present the contribution of each variable to the first 3 components. For the primary one it looks as on 50. We can clearly notice that six of the features have much bigger impact than the rest and can be considered as main contributors.

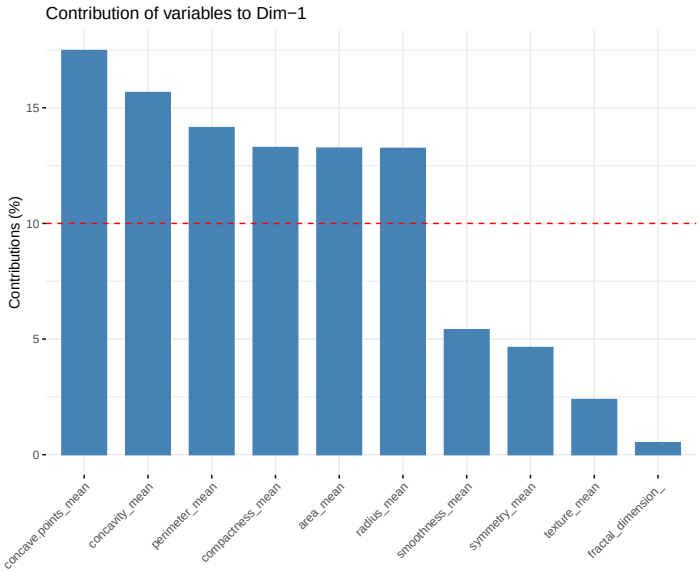

Figure 50: Plot of the contribution of variables

For the second one presented on 51 we find out that the name of main contributor can be assigned to the 'fractal dimension' variable, as the rest do not affect the result in a significant way.

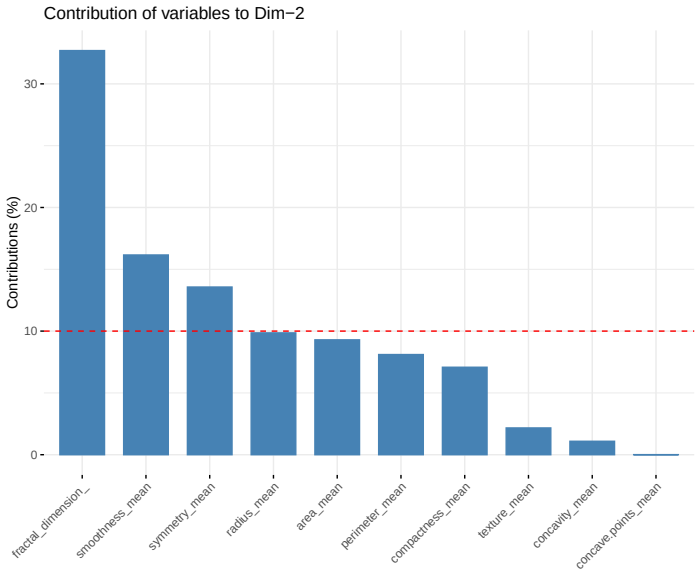

Figure 51: Plot of the contribution of variables

The third one shows us that almost only one factor contributes to this component – texture, having over 90%.

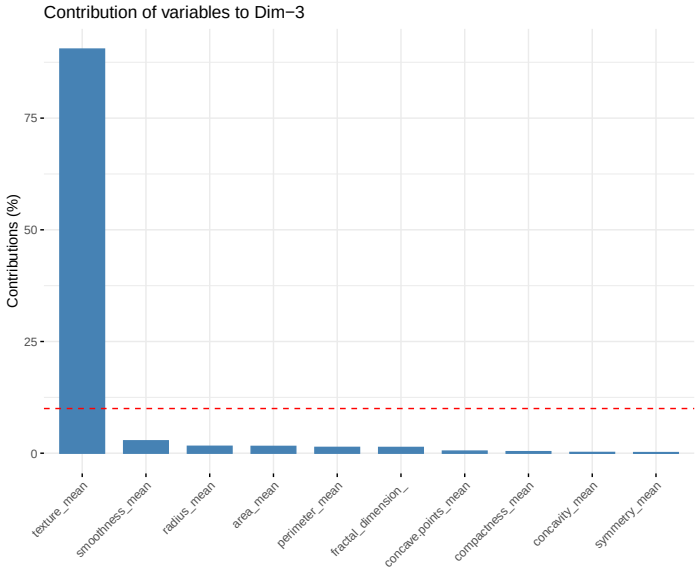

Figure 52: Plot of the contribution of variables

If it comes to the results of PCA algorithm, we can present the plot of the individuals showing the group that they belong to where there are 2 groups considered (53) and 4 (54). It turns out that the observations are quite well distributed and we can clearly notice the dependence between clusters.

Figure 53: Plot of PCA individuals

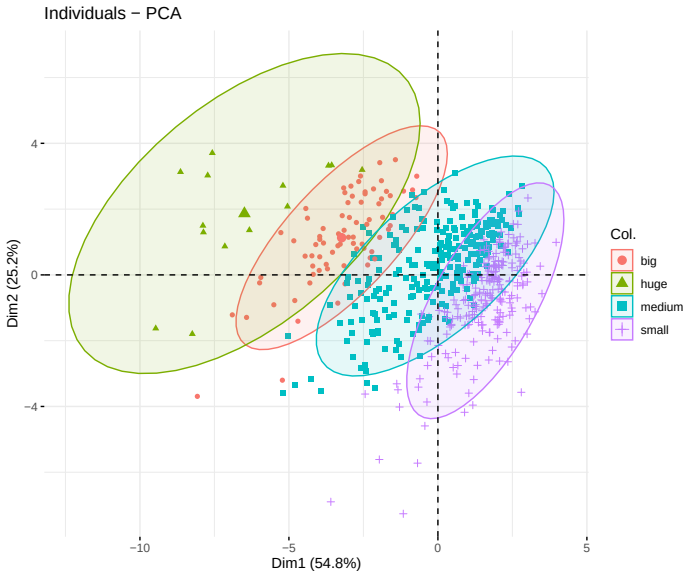

Figure 54: Plot of PCA individuals

Additionally we can introduce the biplot, where we could see the observations after performing PCA and the graph of variables and their eigenvectors & correlation.

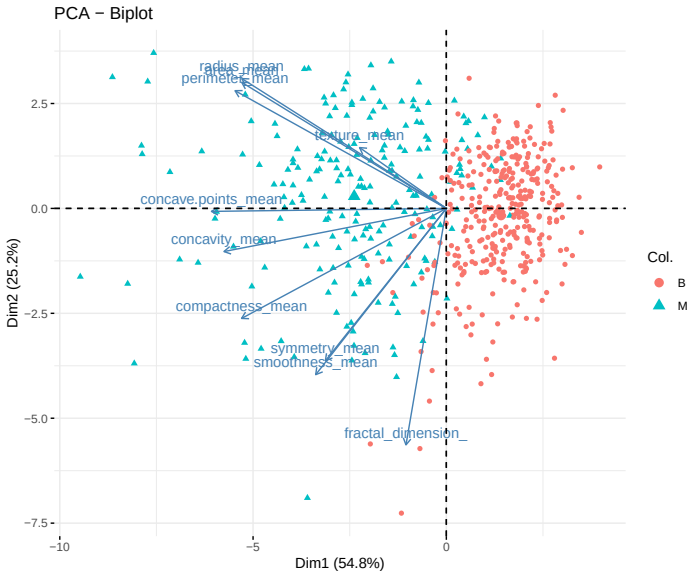

Figure 55: Plot of PCA individuals and variables

We can notice that on figure 55 the most significant factors would be area radius and perimeter as they mostly indicate whether the cancer is malignant. The variables such as fractal dimension or smoothness do not really affect the judgement. Besides that, we can state that in general the higher the values of an observation, the higher chance that the tumor is malignant.

Finally we can claim that the PCA method works correctly and we can omit the variables in our data that do not really have a significant influence on the diagnosis of the cancer. Such factors as fractal dimension or smoothness are generally describing the non malicious features. On the other hand, radius and perimeter are highly correlated with the diagnosis, and are corresponding to the type of cancer.

3.2.2 Multidimensional Scaling – MDS

Multidimensional scaling (MDS) tries to find a low-dimensional representation of the data in which the distances respect well the distances in the original high-dimensional space. What is more, MDS is a technique used for analyzing similarity or dissimilarity data. You can see the plot of the data after using MDS on 56. Objects that are more similar (or have shorter distances) are closer together on the chart than objects that are less similar (or have longer distances). We can see one large center of the points while the rest is proportionally distributed.

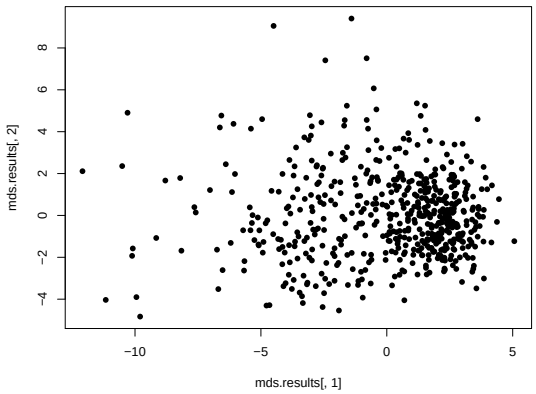

Figure 56: Plot of the MDS algorithm results

It is important to check the quality of an assessment of used method. MDS is an iterative process in which the accuracy of the statistical model is quantified by the output parameter *stress*, which is a numerical measure of the incompatibility between the current configuration and the input data. The MDS process is repeated until the difference between the calculated coordinates and the input data is minimized according to a predetermined stop criterion, such as a minimum *stress* level. Stress can also be used to determine the number of dimensions of the solution. Let us see the results of comparison *stress* criterion with dimension on 57 and analyse the behaviour.

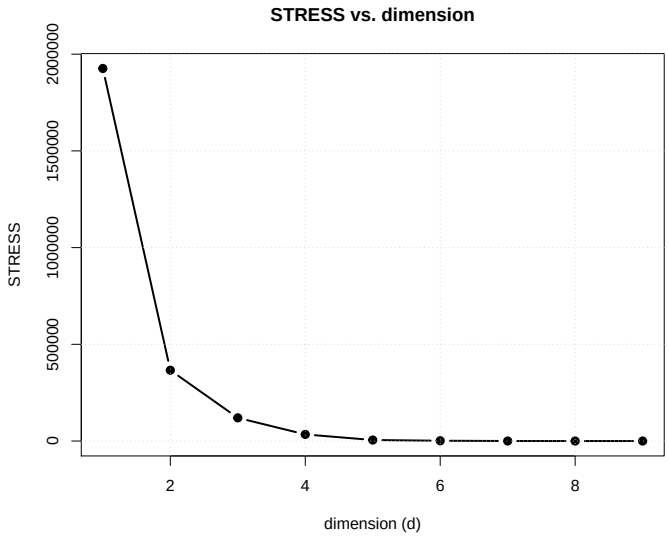

Figure 57: *stress* vs dimension

We can see that increasing the dimension decreases the *stress* value at the beginning and than the curve levels off. There is almost no difference between the last dimensions, but there is a huge difference between 1 and 2 dimension. The quality of the MDS mapping improves as the dimension increases.

For quality assessment we also want to check the Shepard diagram in which we compare the original distances with distances obtained after usage of MDS method. The Shepard diagram is a scatter plot of input against output distances for every pair of items scaled. The results for different dimension are shown on 58.

A really accurate dimension reduction will produce a straight line. We can observe it for the high dimensions (7-9). We can see that the quality of the MDS mapping improves as the dimension increases.

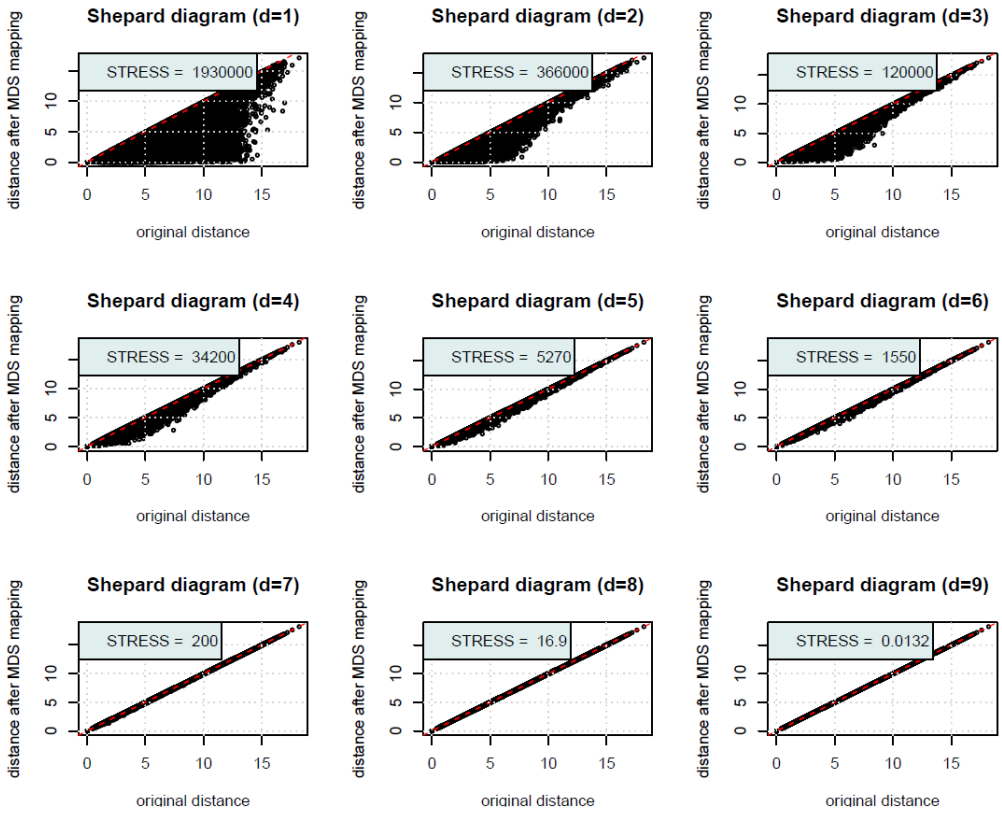

Figure 58: Different shepards

4 Discussion of the results and conclusions

The clustering of the data allowed us to notice the pattern and similarity between the observations and could help us dig deeper into the data structure. Different methods were of use to determine the optimal parameter k that defined the number of separate clusters, which in most cases was equal to 2. That is the result that satisfies us, because the data consists of the observations are of type M or B, which is the most influential part in analysis of the danger of the tumor.

It clearly indicates that there is a division between those two types and those methods allowed us to notice that. Thanks to that, if we had more cases for which we would have to determine the kind of cancer not knowing the diagnosis, it would be much easier to correctly assign them into their categories. Even though some of the methods did not emphatically confirm the mentioned result, each one of them helped us in some way to find out the desired outcome.

Validating the scores reassured us that our analysis was performed correctly and removed any remaining doubts. The appliance of dimensional reduction methods allowed us to better understand the dependencies between variables and their influence. We have also examined which of these factors were the most important in the diagnosing whether the cancer is malignant or benign. Thanks to all of the performed analysis we can surely claim that omitting some of the variables would not change much the outcome of how algorithms work and the decision-making processes.