Data Mining project - part II

Katarzyna Macioszek, Ada Majchrzak

January 13, 2024

1 Introduction

In the second part of the project we will use clustering analysis in connection with dimension reduction methods on *Spambase* dataset. Our goal is to asses clustering algorithms performance in differentiation between *spam* and *nonspam* types. For that we shall use *k-means*, PAM and AGNES clustering algorithms. We will also check separation measures of clusters built using original data and data after dimension reduction done with PCA algorithm. Additionally we will also compare clustering results with classification methods performed on the former part of the project.

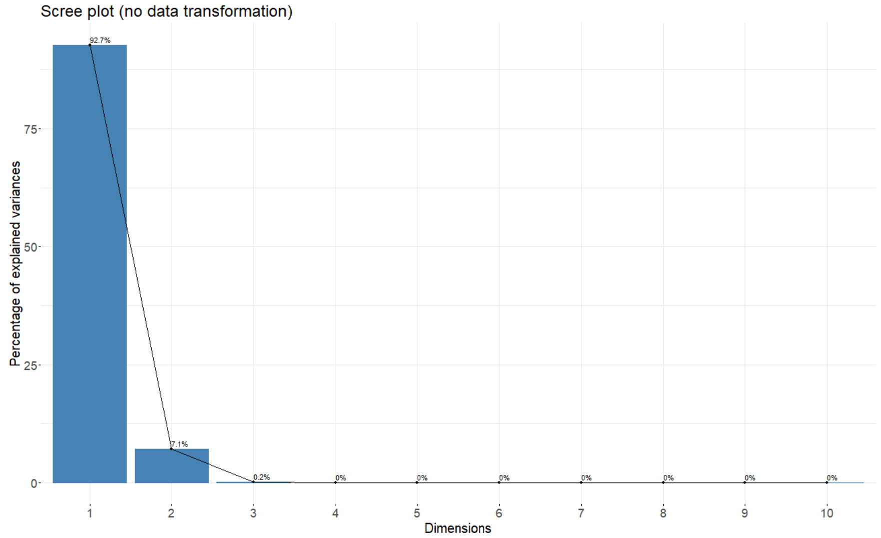
2 Data preparation

To be consistent with analysis done in previous part of the project we again exclude columns that exhibit high correlation with each other. In this eliminated features are *num415* and *num857*. Furthermore we also use the same data transformations: z-score standardization and $\log(x + 0.1)$ transformation to make the explanatory variables have comparable values, as it is crucial for clustering and dimension reduction algorithms.

3 PCA

To be able to visualize the multidimensional data easily, we start with performing the principal component analysis (PCA). As mentioned before we use two data transformations, however to display an interesting behavior of the method let us first present the outputs for non-transformed data. To visualize PCA results we will use scree plot and contributions of variables in first two principal components.

Figure 1: PCA results for original data

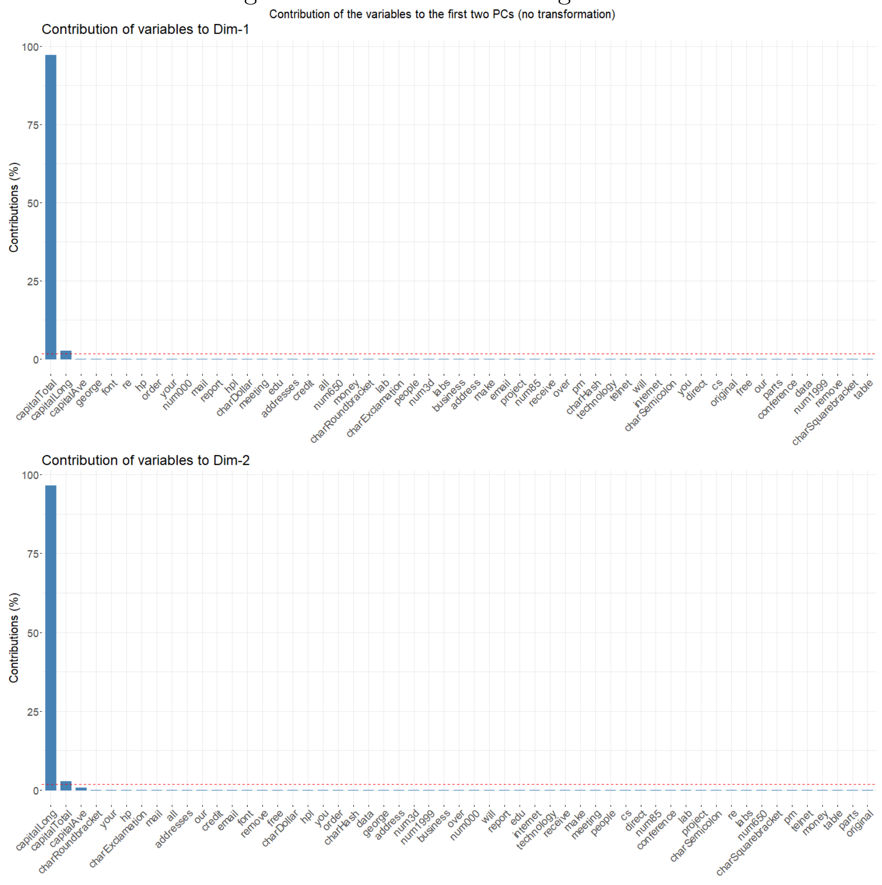


As we expect the algorithm has poor quality as only variables with highest values are taken into consideration. In our data set most of the variables are frequencies with values in interval $[0, 100]$ and only capital letters characteristics have higher values. That causes the observed behavior of PCA method, which results with only 2 principal components, so in the following analyses we will use only transformed data.

Let us now check the principal components for data after z-score standardization.

As we anticipated in this case we get more principal components created and also increased number of features is taken into account.

Figure 2: PCA results for original data



Here again we observe that contribution of many variables is higher than for original data. Also we see that in comparison to z-score transformed data we will probably be able to eliminate more principal components for the latter analysis.

Add part about cutting PCs and method used, clean up plots

4 Clustering

For clustering analysis we will use k-means, PAM and AGNES algorithms for transformed *Spambase* data set and also for data set build using principal components extracted while performing PCA.

We are lucky to have the knowledge about number of classes in the examined data set (*spam* and *nonspam*), however let us check anyway if the number of clusters determined using `fviz_nbclust` method is the same. On Figures 7, 8, 9 and 10 we present the results for k-means and PAM algorithms.

In all cases the estimated number of clusters is equal to or one more that the actual number of clusters. It means that we should assume that number of clusters is 2 or 3, but considering the knowledge about used data set, we can assume two clusters in each algorithm.

4.1 k-means

4.2 PAM

4.3 AGNES

Figure 3: PCA results for standardized data

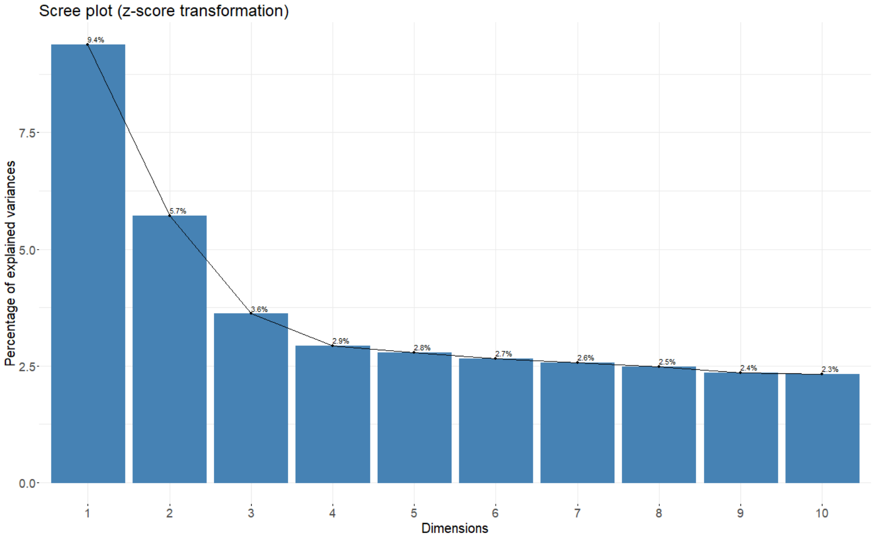


Figure 4: PCA results for standardized data

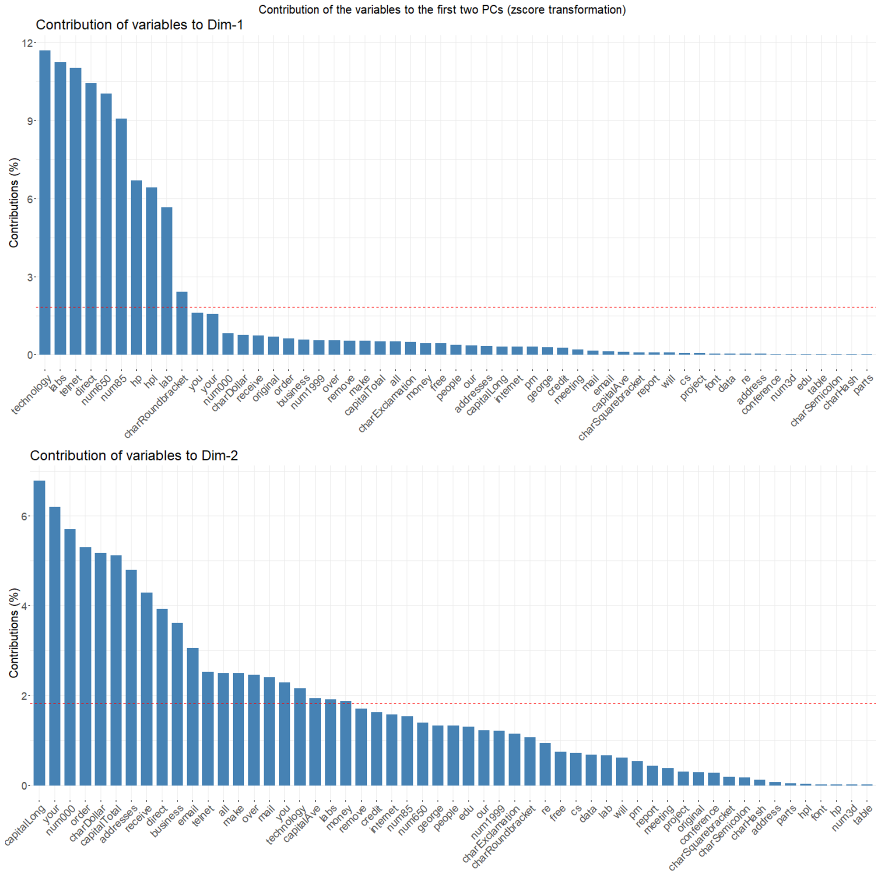


Figure 5: PCA results for log-transformed data

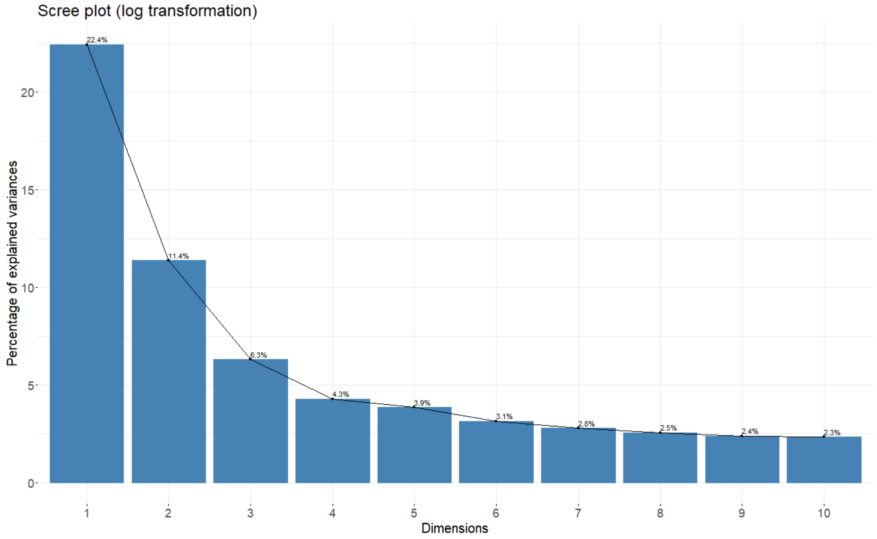


Figure 6: PCA results for log-transformed data

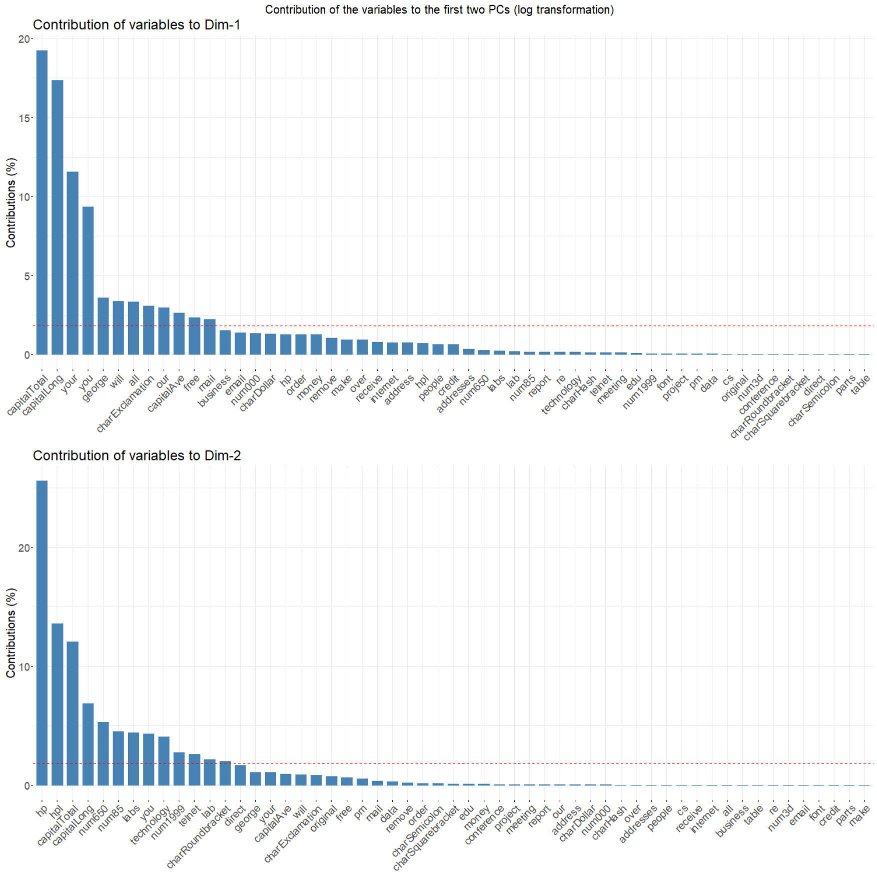


Figure 7: Number of clusters for k-means (no PCA)

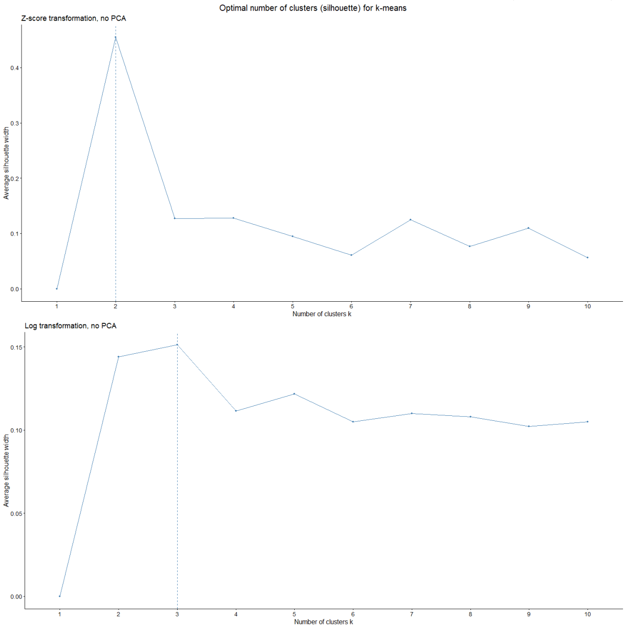


Figure 8: Number of clusters for pam (no PCA)

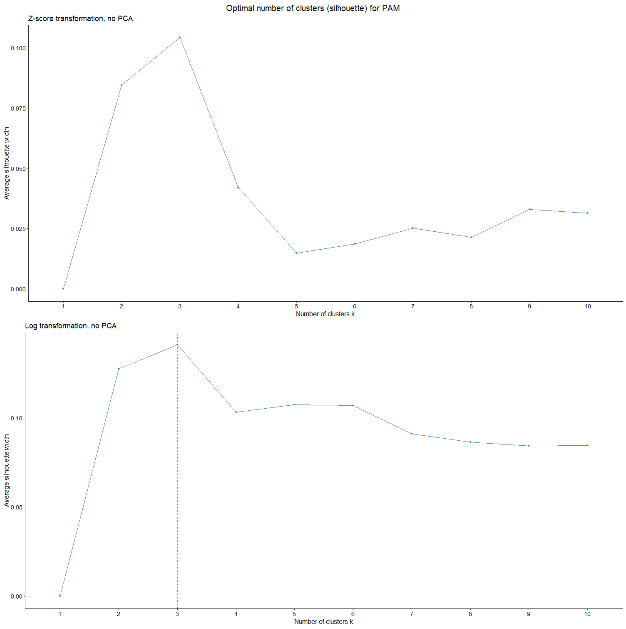


Figure 9: Number of clusters for k-means (with PCA)

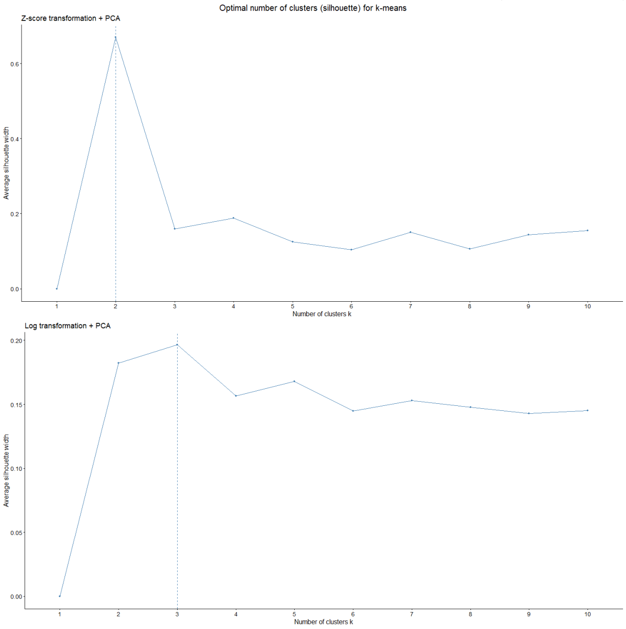
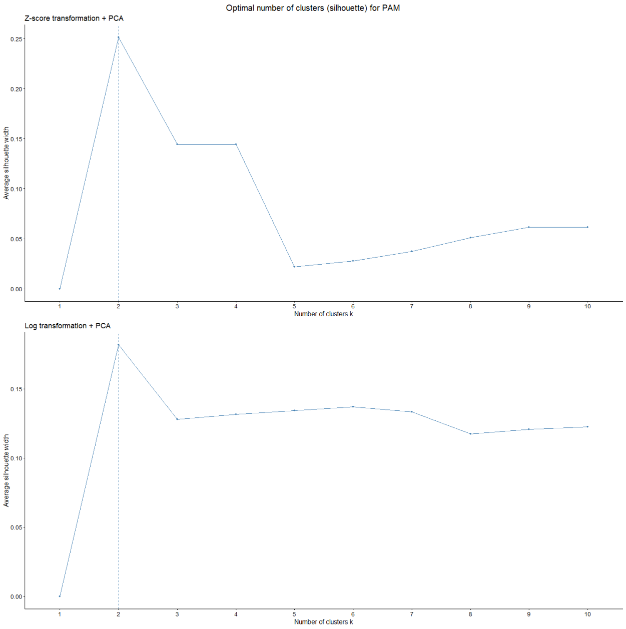


Figure 10: Number of clusters for pam (with PCA)


Optimal number of clusters (silhouette) for PAM
Z-score transformation + PCA
Average silhouette width
Number of clusters k
Log transformation + PCA
Average silhouette width
Number of clusters k