**Unsupervised learning – clustering analysis**

In our project we decide to use cluster analysis to group together attributes of white wine.

Find a similarity between attributes that’s a good step to build a good model.

It is a form of unsupervised learning – that’s mean machine is learning from a raw data. All target attributes has been removed for the moment of training[[1]](#footnote-1).

In that particular case we decide to use Partitioning approach. We used k-means and x-means and a variation of k-means to compare these two methods. K-means is the best known method in Data Science. We also choose x-means as in x-means you don’t need to specify number of clusters.

K-means calculating again the centroid after every assignment and repeat that step until no change. That’s why it is sensitive for changes.

In clustering we have be very careful with outliers. That’s why we made double careful when making data pre-processing.

1. Correlations for clustering purposes



Correlations – rapidminer – White wines

The above picture shows us correlations between attributes of our dataset. If the correlation is higher then the purple colour is more intensive.

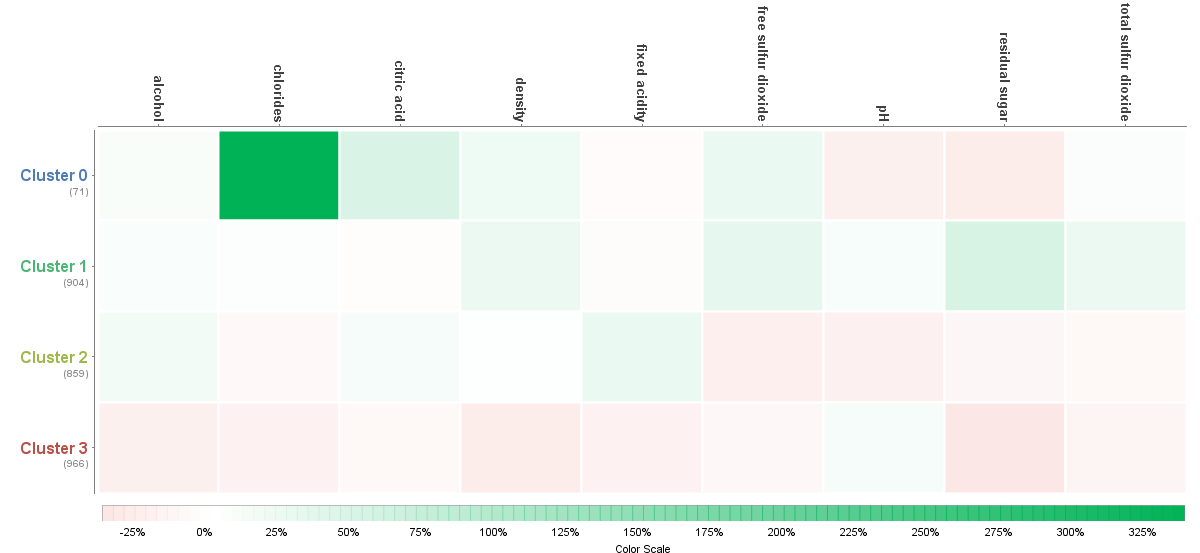
The biggest correlation we can observe between:

* Total sulfur dioxide and free sulfur dioxide 0.585
* Fixed acid and PH -0.48
* Residual sugar and density 0.443

This information can be very helpful in later stage of project when building a prediction model. This is also a basis for clustering analysis. That is also our basis for choosing the right number of clusters in k-means cluster analysis.

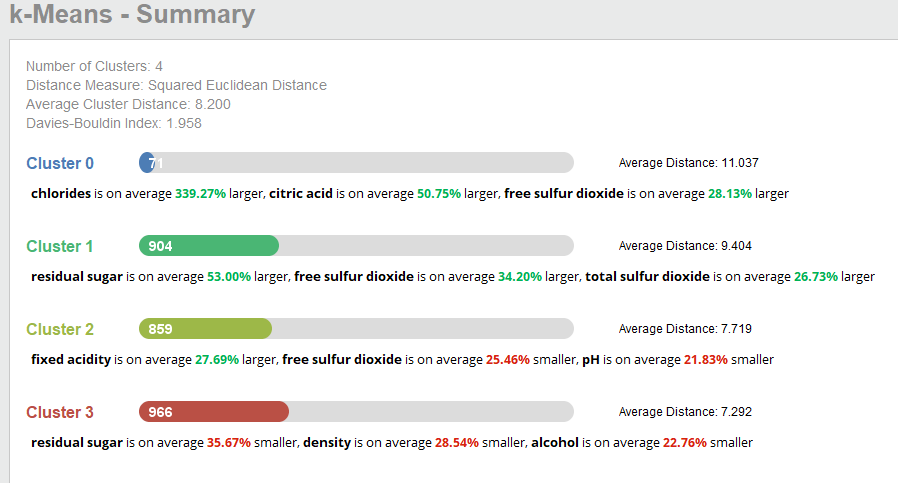
1. K-means clustering for white wine.

We start that part by examining the scatter plots from data exploration part of project, and we decide to use 4 clusters. We did analysis in Rapidminer studio:



1. Heat map 1 – Rapidminer studio k-means– White wines attributes

Chlorides could look here as outliers as we can see very intensive green. But in reality there chlorides in wine depends from the kind of grapes. There are kind with very high volume of chlorides.



Summary 1 – K-means Rapidminer studio – White wines

In above pictures it is clearly visible that 4 clusters were good choice. Although it is still not perfect distribution.

By Davis Bouldin Index we got only 1.958. It is an index for evaluation of cluster algorithms. As closer to zero then the clusters has been better assigned and they are closer to each other.

We can see that we have 4 clusters.

In Cluster zero we can see that Chlorides are on average 339.27% larger than others, citric acid is on average 50.75% larger and free sulfur dioxide is on average 28.13% larger then in others. Ther eis only 71 entries in this category but average distance here is 11.

In Cluster 1 we can see that residual sugar is on average 53% larger, free sulfur dioxide is on average 34.20% larger and total sulfur dioxide is on average 26.73% larger. There is 904 entries in this cluster but Average distance here is 9.4.

In Cluster 2 fixed acidity is on average 27.69% larger free sulfur dioxide is on average 25.46% smaller and PH is on average 21.83% smaller. There is 859 entries and average distance here is 7.7

In Cluster 3 residual sugar is on average 35.67% smaller, density is on average 28.54% smaller, alcohol is on average 22.76% smaller. That is the biggest cluster with 966 entries and average distance here is 7.2.

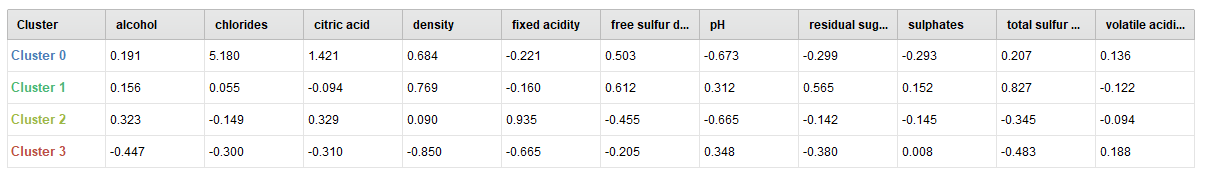
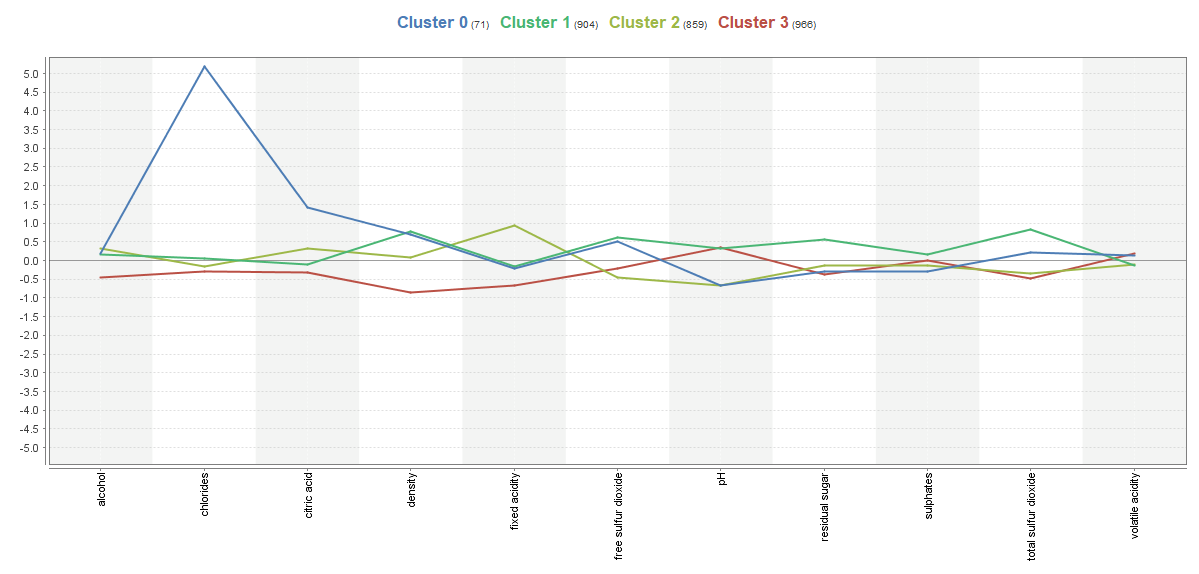
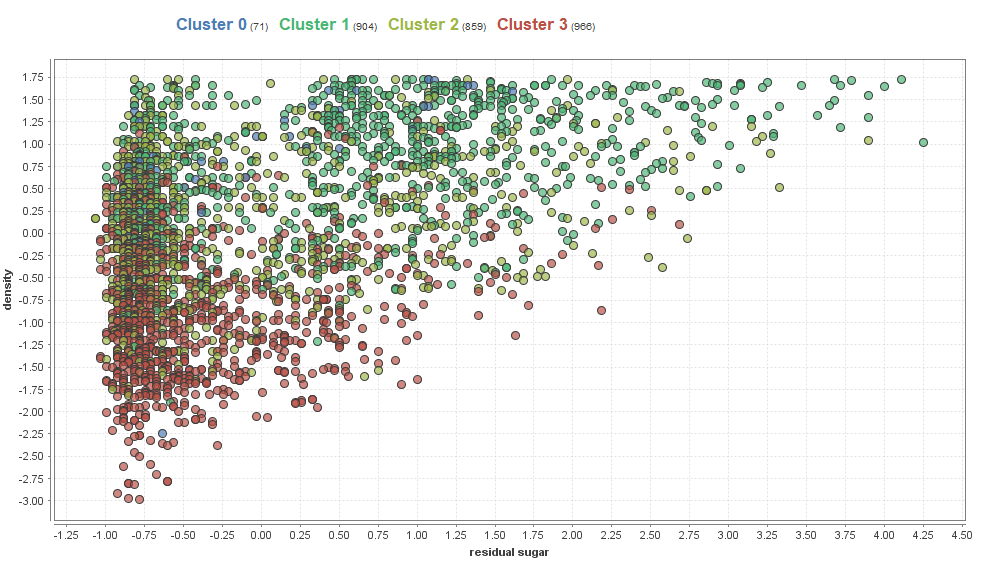
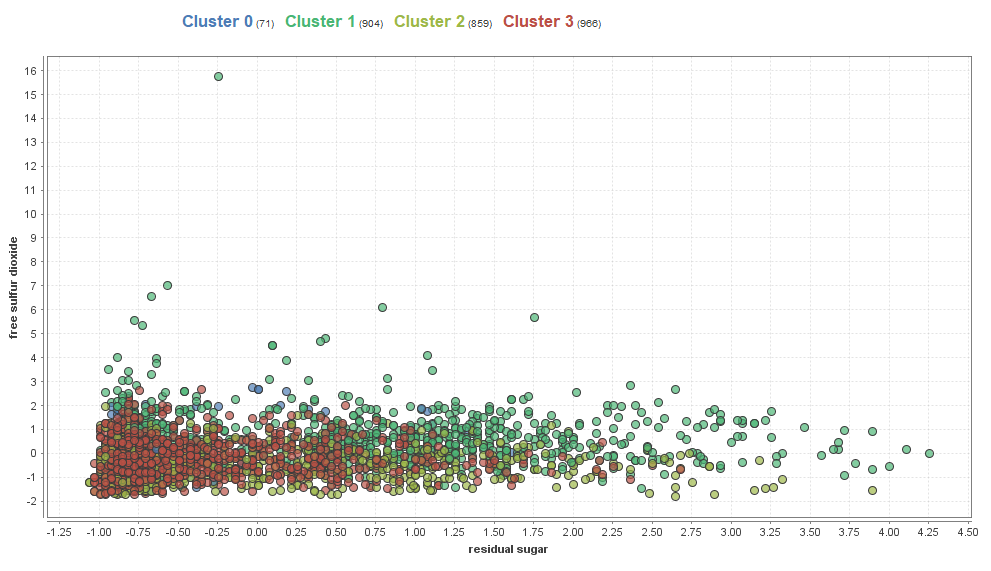
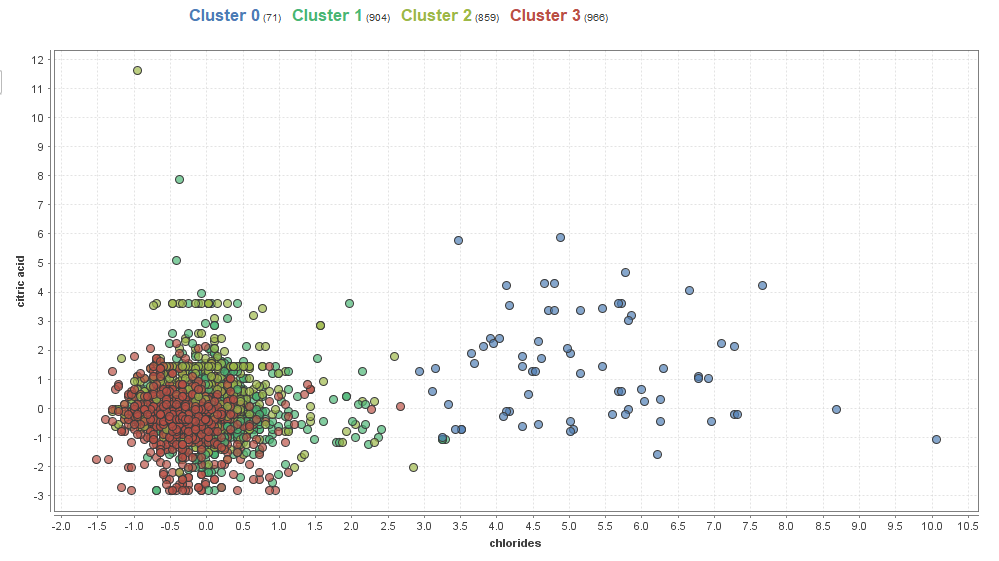
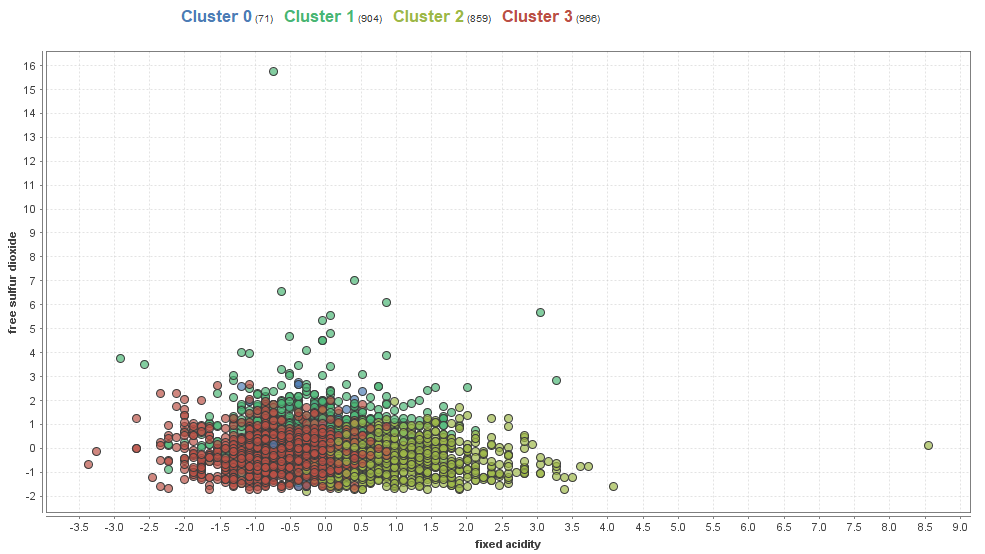


Table 1 – Rapidminer, k-means white wine

In tables we can see the differences between clusters. Below there is line plot with 4 clusters and scatter plots for all 4 clusters separately.

Line plot – Rapid miner k-means – White wine

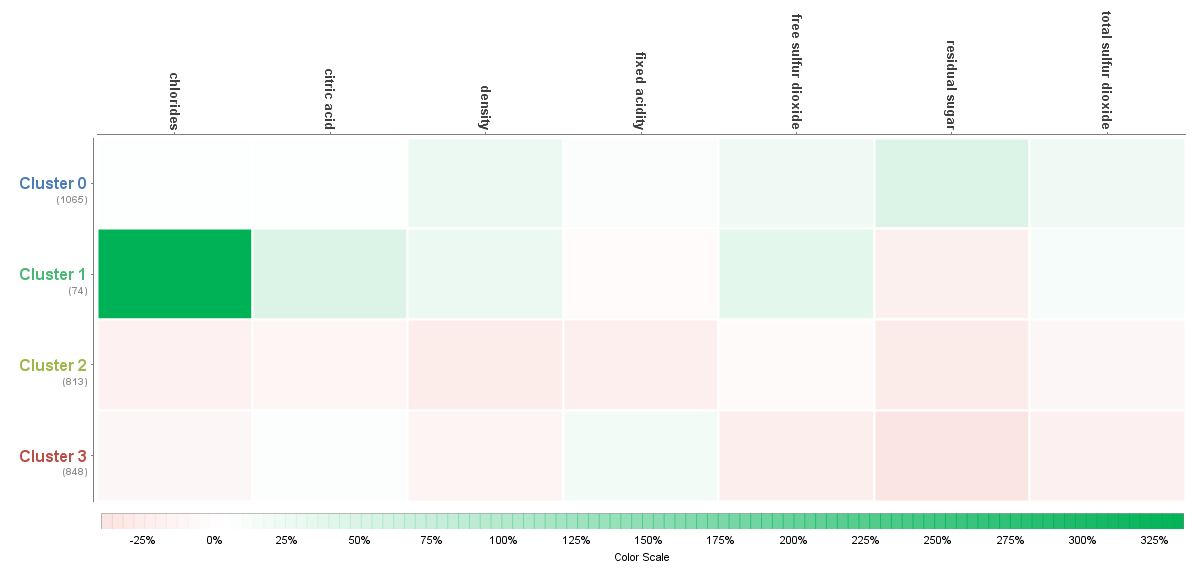


Scatterplot Rapidminer X-mean white wine

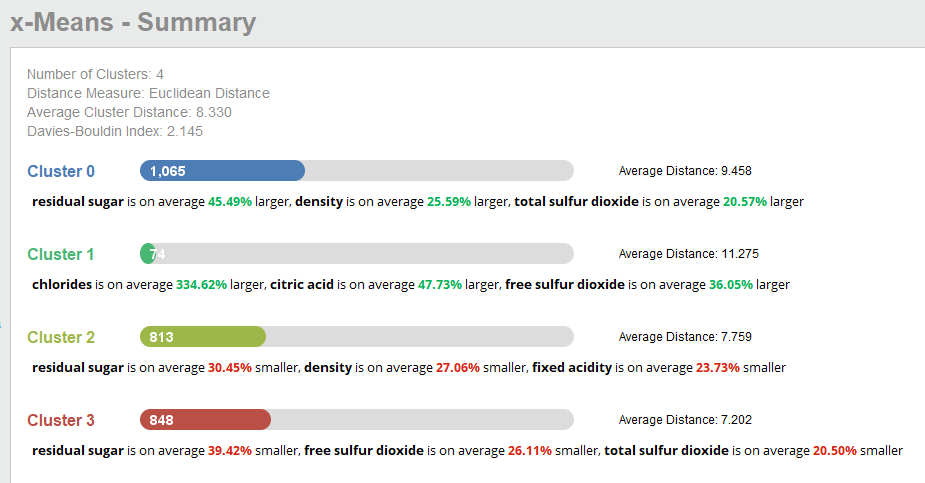
1. X- means clustering for white wine.

In this analysis we didn’t need to choose number of clusters. The algorithm do it itself.

X-means it is a variation of K-means. So let’s see the results:

**Heatmap 2 – Rapidminer X-means White wines

In this algorithm we can see that chlorides have been associates in different cluster than the previous k-means algorithm. Interesting fact is those algorithms also choose doing 4 clusters. That’s mean we made similar assumptions.



Summary Rapidminer X-means – white wines

By Davis Bouldin Index we got only 2.145. So its means that k-means was better choice in that case.

We can see that we have 4 clusters too.

In Cluster zero we can see that residual sugar are on average 45.49% larger than others, density is on average 25.59% larger and total sulfur dioxide is on average 20.57% larger then in others. It has large number of entries 1065 and average distance here is 9.45.

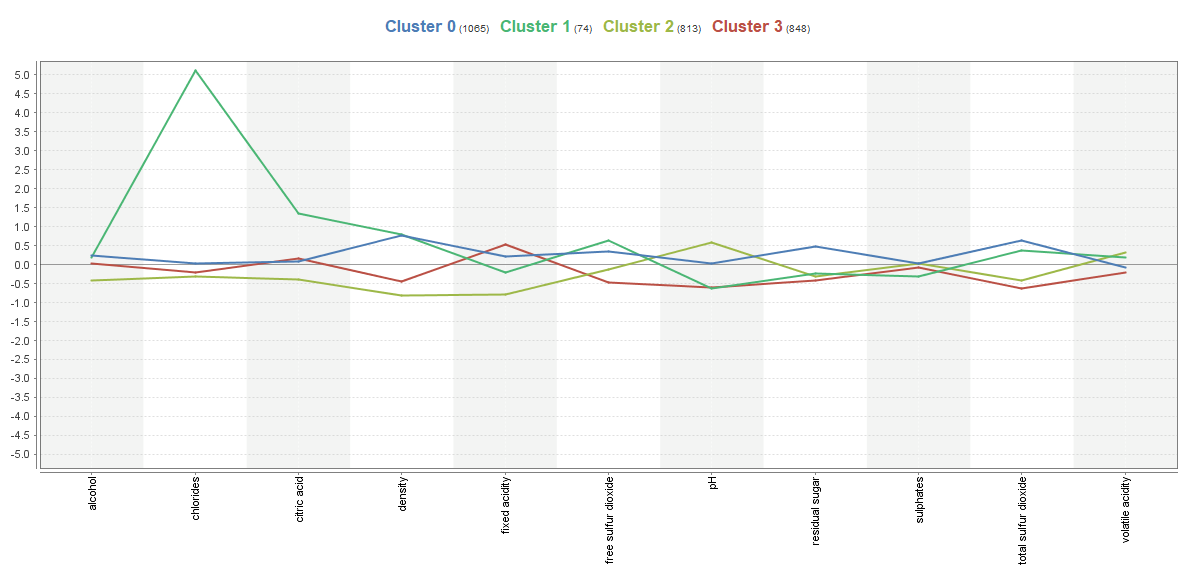
In Cluster 1 we can see that chlorides is on average 334.62% larger, citric acid is on average 47.73% larger and free sulfur dioxide is on average 36.05% larger. There is only 74 entries here but Average distance here is 11.27.

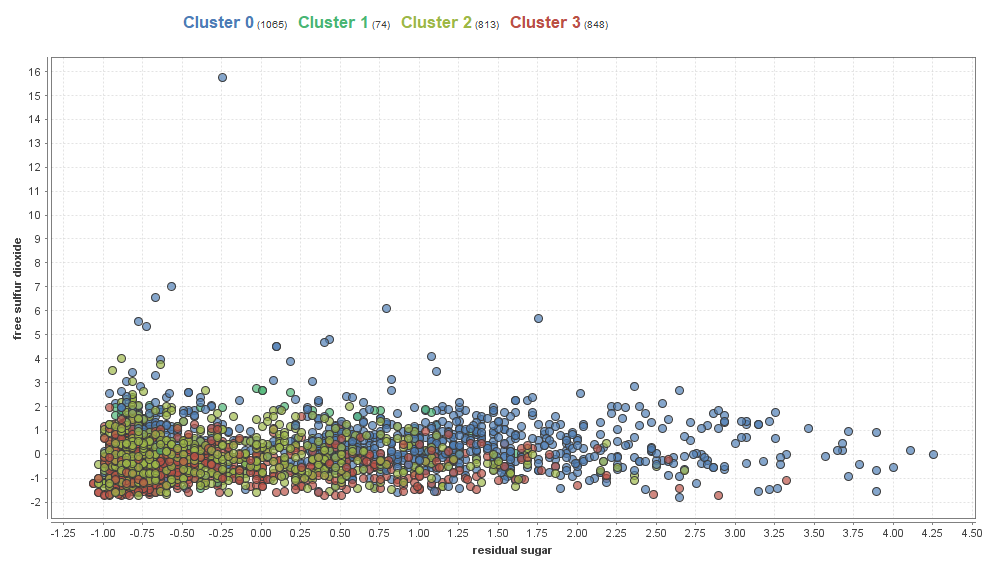
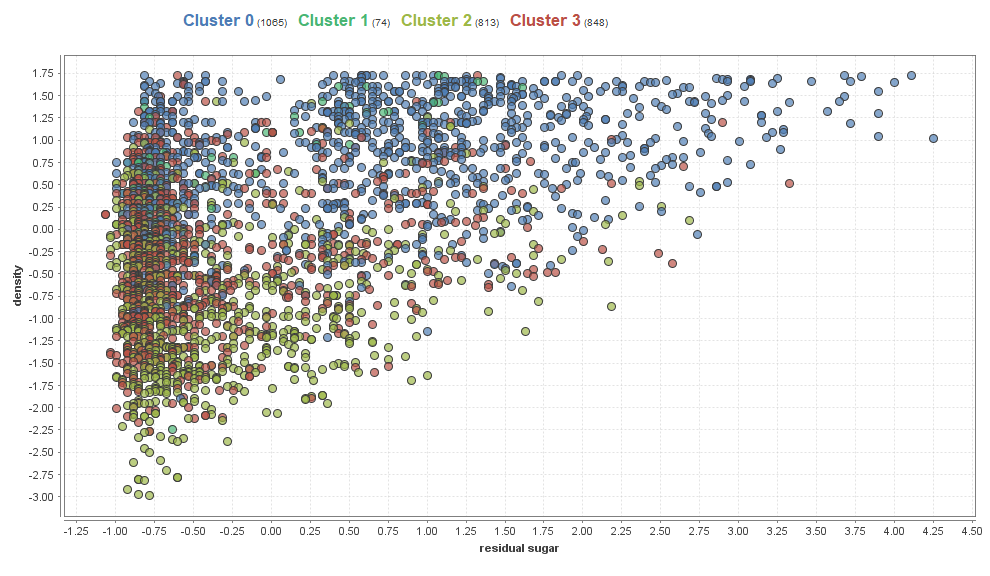
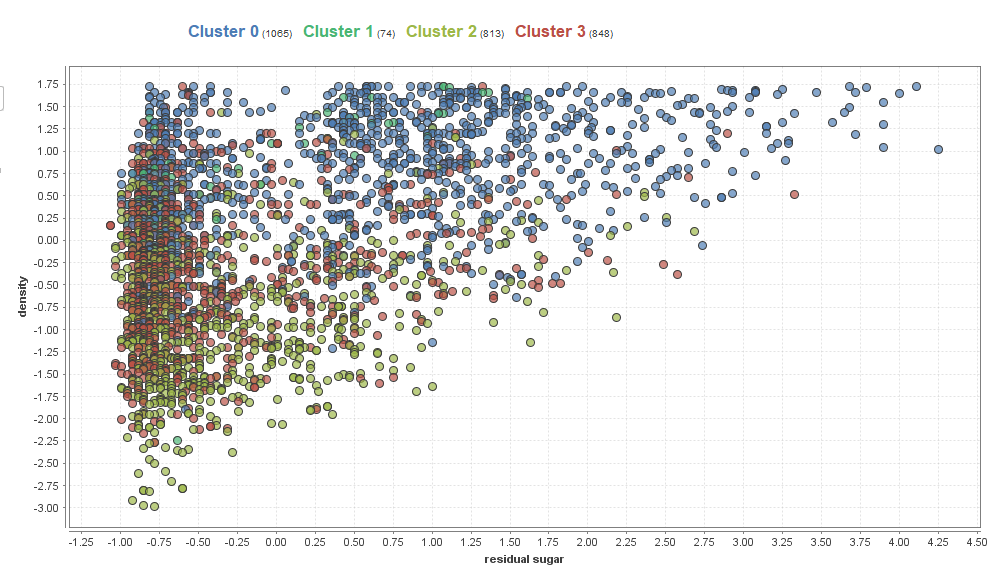
In Cluster 2 residual sugar is on average 30.45% smaller density is on average 27.06% smaller and fixed acidity is on average 23.73% smaller. There is 813 entries in this cluster and average distance here is 7.7

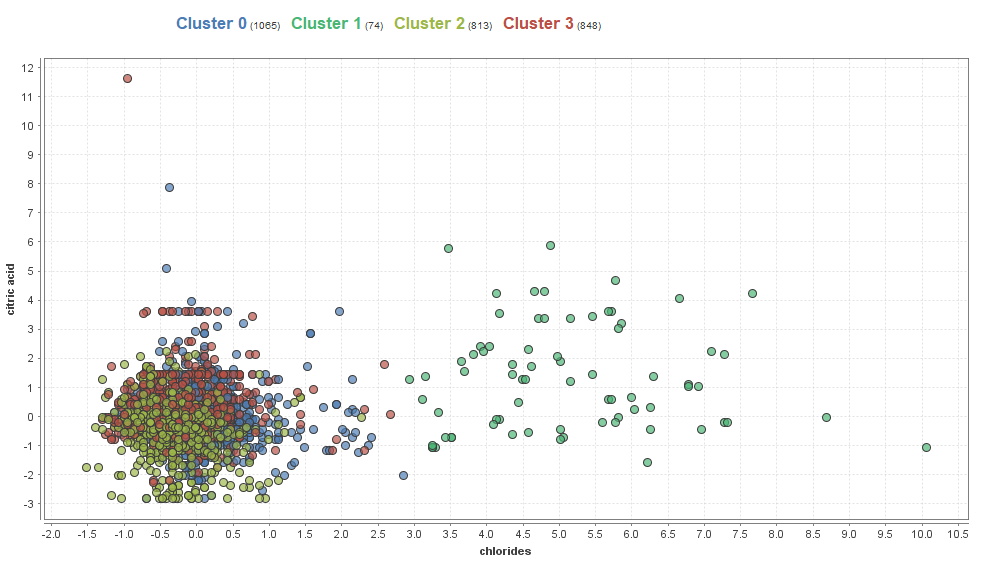
In Cluster 3 residual sugar is on average 39.42% smaller, free sulfur dioxide is on average 26.11% smaller, total sulfur dioxide is on average 20.5% smaller. There is 848 entries and average distance here is 7.2.



In tables we can see the differences between clusters. Below there is line plot with 4 clusters and scatter plots for all 4 clusters separately







Scatter plots of 4 clusters of white wines x-means - Rapidminer

1. Weaknesses of clustering

* We cannot use it for categorical data as it applicable for continuous attributes only.
* In k-means we need to specify number of clusters
* You need to remove outliers.
* Only standard shape of clusters
* X-means seems to be variation of k-means but as in our example k-means was better choice in this particular case.

1. Han J, Kamber M., Pei J. 2012 *Data mining Concepts and Techniques,* p. 19 ISBN 978-0-12-381479-1 [↑](#footnote-ref-1)