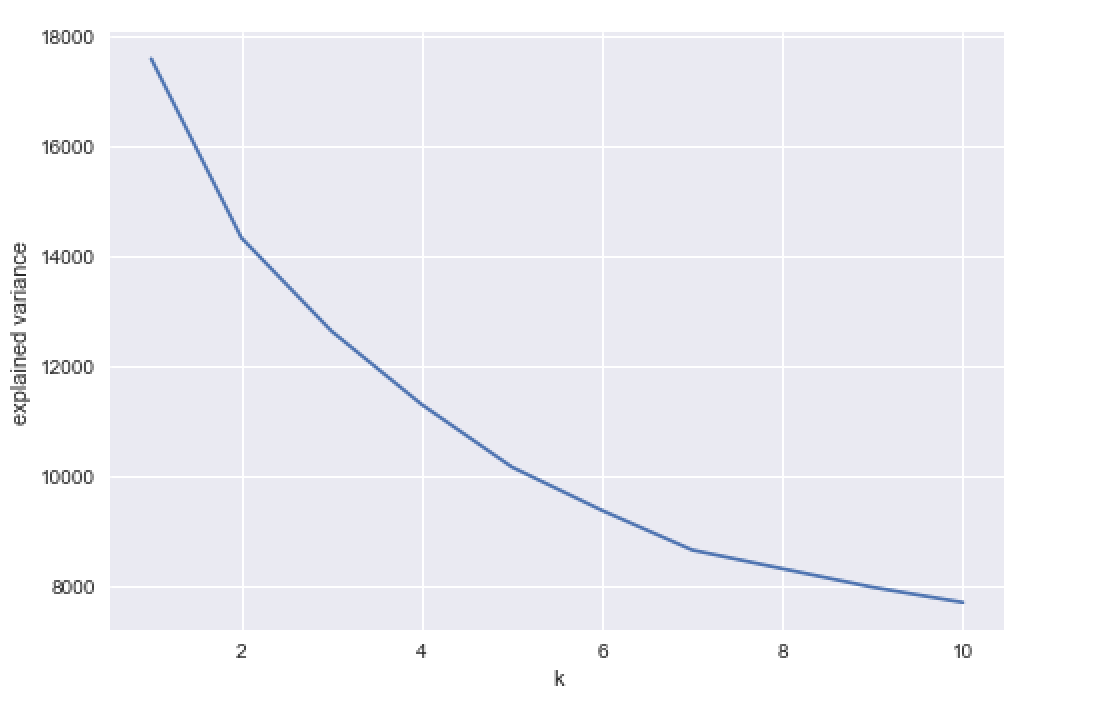
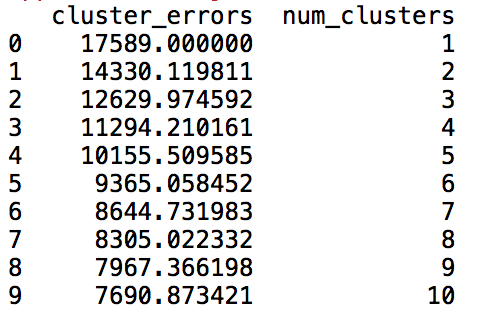
* For this milestone, we do clustering the features of dataset without labels. With obtained k, we do k-means clustering on 11 feature dimensions which are “fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol”. Then we evaluate the performance of K-means clustering through visualization and graph.
* **K-means clustering**: Given a set of n observations xi, which is d-dimensional real vector, K-means clustering n observations into k partitions as to minimize the within-cluster sum of squares.
* Preprocessing: since we have high-dimensional feature spaces and various ranges of values involved, we normalize the data first.
* Choosing optimal number of clusters
* Elbow’s method: it uses the percentage of variance that is explained as a function of the number of clusters. (explained variance) – explained variance measures the proportion to which model with k accounts for the variance of a given dataset. The first clusters will add much information that would explain a lot of variance but at some point the marginal gain of information will decrease and this point is the optimal k we choose.
* We choose the optimal K from the point where the marginal gain in explained variance drops.
* For extension of k-means clustering, we used GMM for in-depth analysis
* Disadvantage of K-means clustering
  + Although it is simple and relatively easy to understand, the non-probabilistic nature of k-means and its use of simple distance-from-cluster-center to assign cluster membership leads to poor performance for many real-world situations.
  + These two disadvantages of k-means—its lack of flexibility in cluster shape and lack of probabilistic cluster assignment—can be generalized with GMM
* **GMM(Gaussian Mixture Models)**
  + This is also know as EM(Expectation-Maximization)
    - E-step: for each point, find weights encoding the probability of membership in each cluster
    - M-step: for each cluster, update its location, normalization, and shape based on all data points, making use of the weights
  + It attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset.
* # of Components(clusters)
  + Means of correcting for over-fitting is to adjust the model likelihoods using some analytic criterion such as the [Akaike information criterion (AIC)](https://en.wikipedia.org/wiki/Akaike_information_criterion) or the [Bayesian information criterion (BIC)](https://en.wikipedia.org/wiki/Bayesian_information_criterion).
  + The optimal number of clusters is the value that minimizes the AIC or BIC,
* Model selection for GMM
  + 'full' (each component has its own general covariance matrix),
  + 'tied' (all components share the same general covariance matrix),
  + 'diag' (each component has its own diagonal covariance matrix),
  + 'spherical' (each component has its own single variance).
  + A more complicated and computationally expensive model (especially as the number of dimensions grows) is to use covariance\_type="full", which allows each cluster to be modeled as an ellipse with arbitrary orientation.

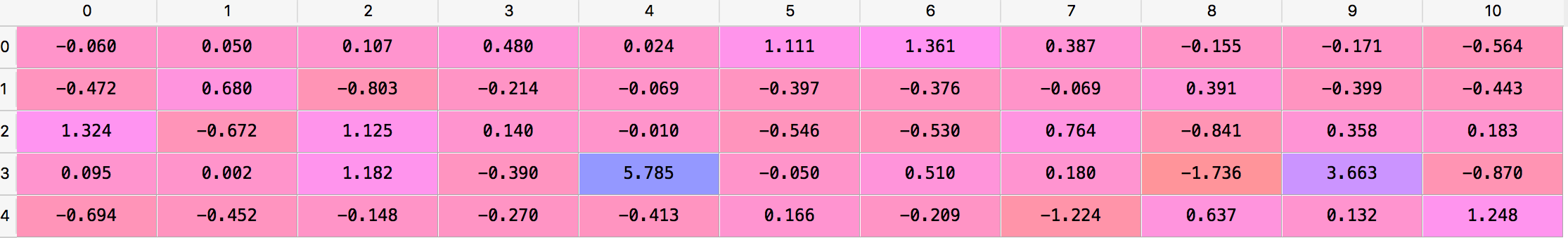
**K-means**



* The optimal k can be picked at 5, in which marginal gain in explained variance has remarkably decreased.
* Showing explained variance in chart

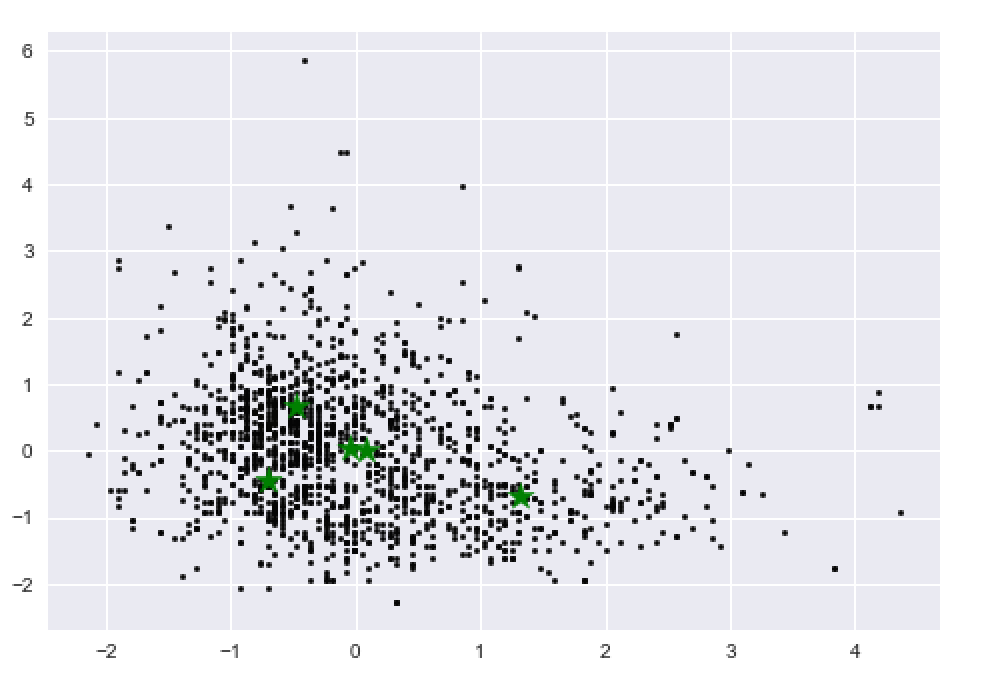


* Cluster centers for k=3

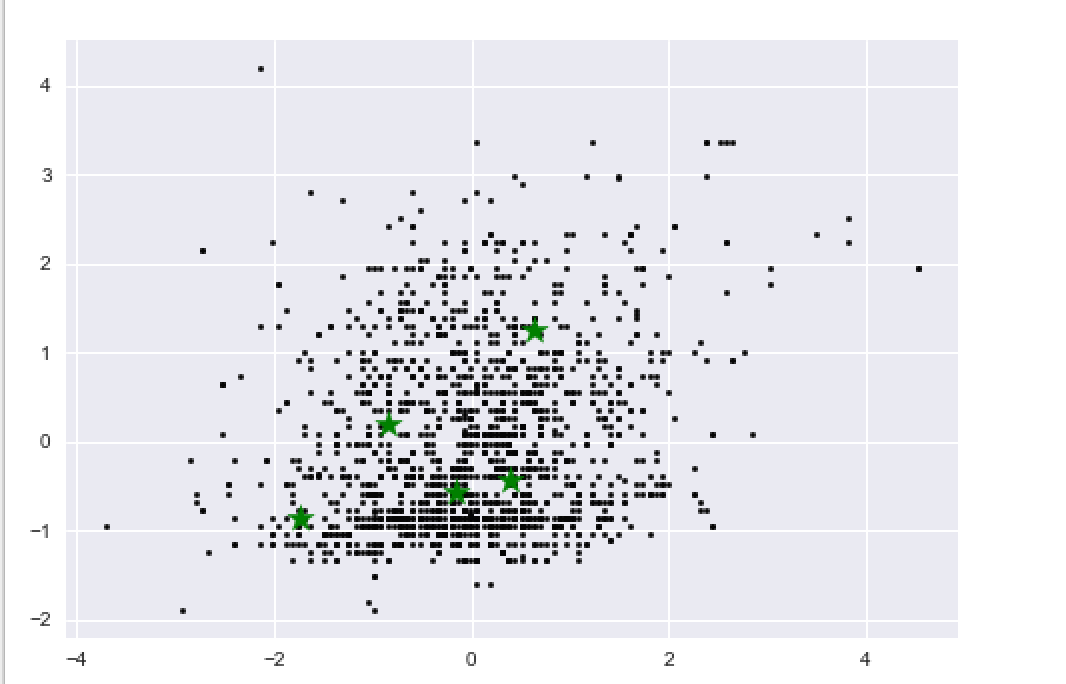


* Several visualization of feature vs. feature graphs as in terms of clusters

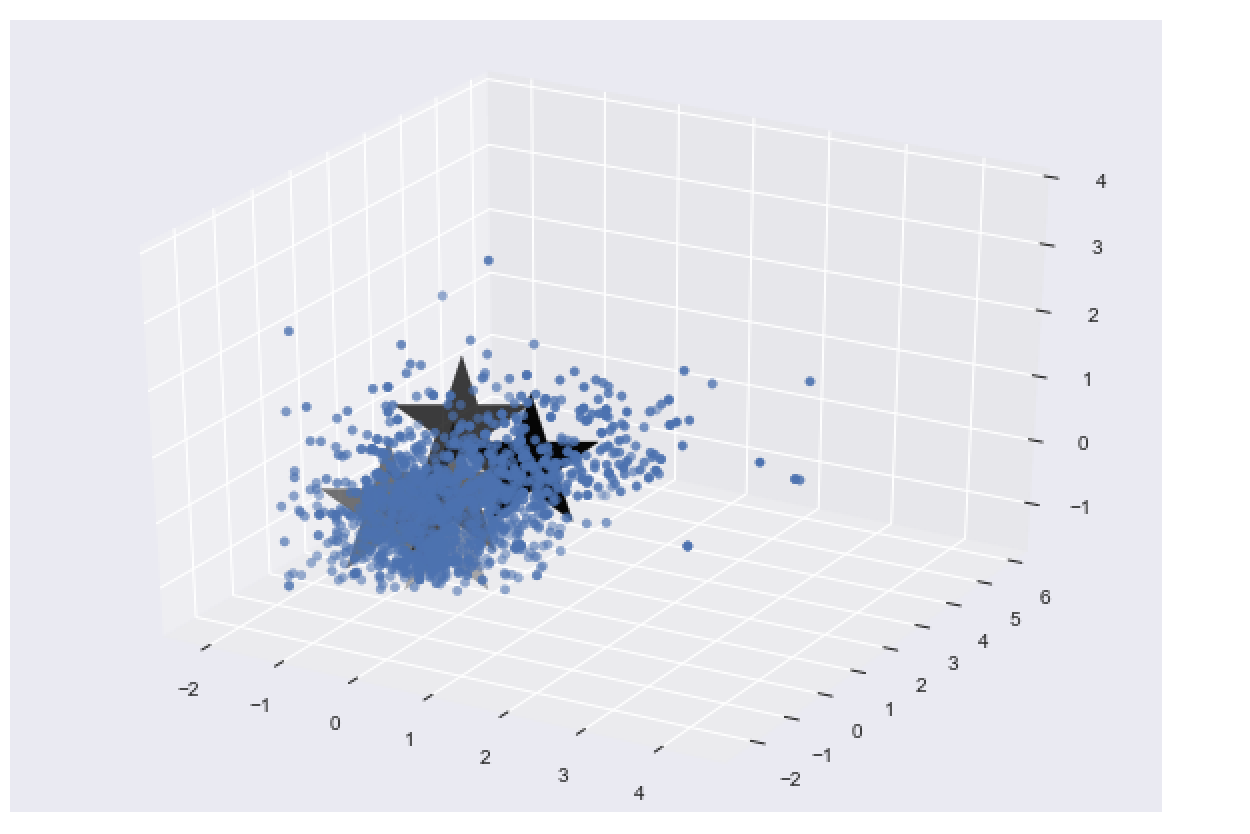
1. Volatile acidity vs. citric acid



1. pH vs. alcohol



* In 3D visualization with three features: fixed acidity, volatile acidity, and citric acid



With normalized data for offsetting constraints by different features, we can analyze through visualization that k=5 clustering is clustering the dataset into 5 partitions; however, since the feature space is high-dimensional and with large dataset, 5 means clustering cannot perfectly partition dataset into 5 clusters. We can also see that data points are somewhat concentrated in that it is hard to manipulate the sparse solution with k-means clustering. But, still it is optimal way of clustering for this dataset since with elbow’s method, we attained k=5 and data points are sparsely distributed among these centroids.

* Evaluation through predicting the labels is not appropriately applied for k-means clustering since the labels are made corresponding to the number of clusters and the spherical clusters from clustering helplessly lets clusters to possibly contain multiple overlapped data points.
* We can still use the labels we attained from 5-means clustering to predict the label and compare with the actual labels: 0.550969355847, but we cannot conclude this accuracy as our precision of K-means clustering as reasoning above.

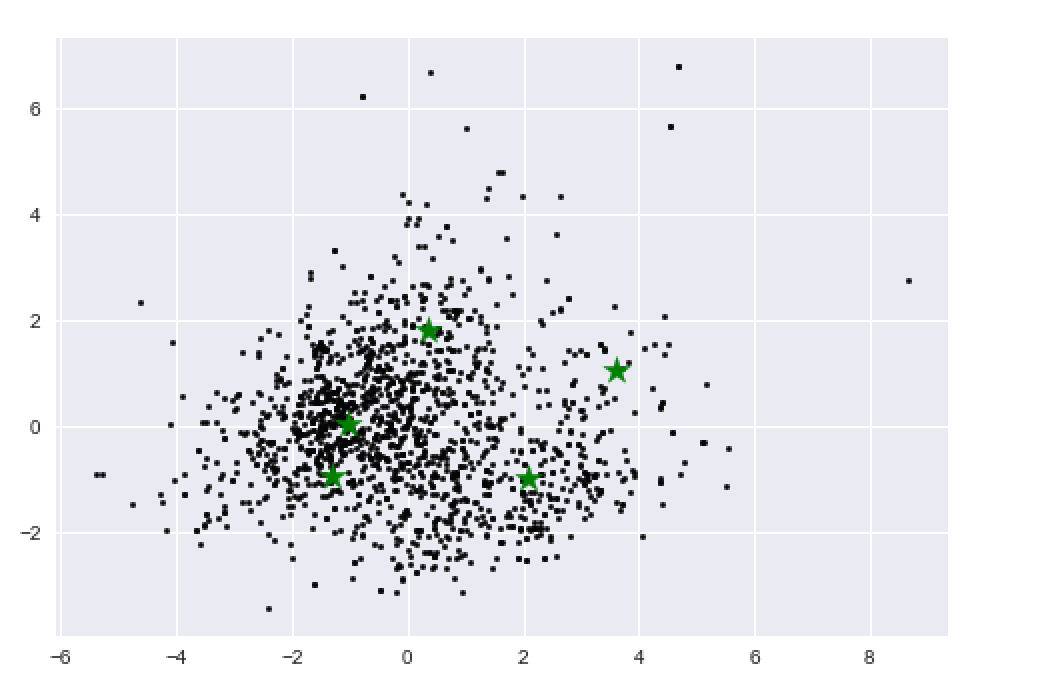
**PCA**

* PCA(Principal Component Analysis) version of K-means clustering with reduction in feature dimension
* PCA: It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated values into a set of values of linearly uncorrelated variables.
  + Choosing the number of principal components

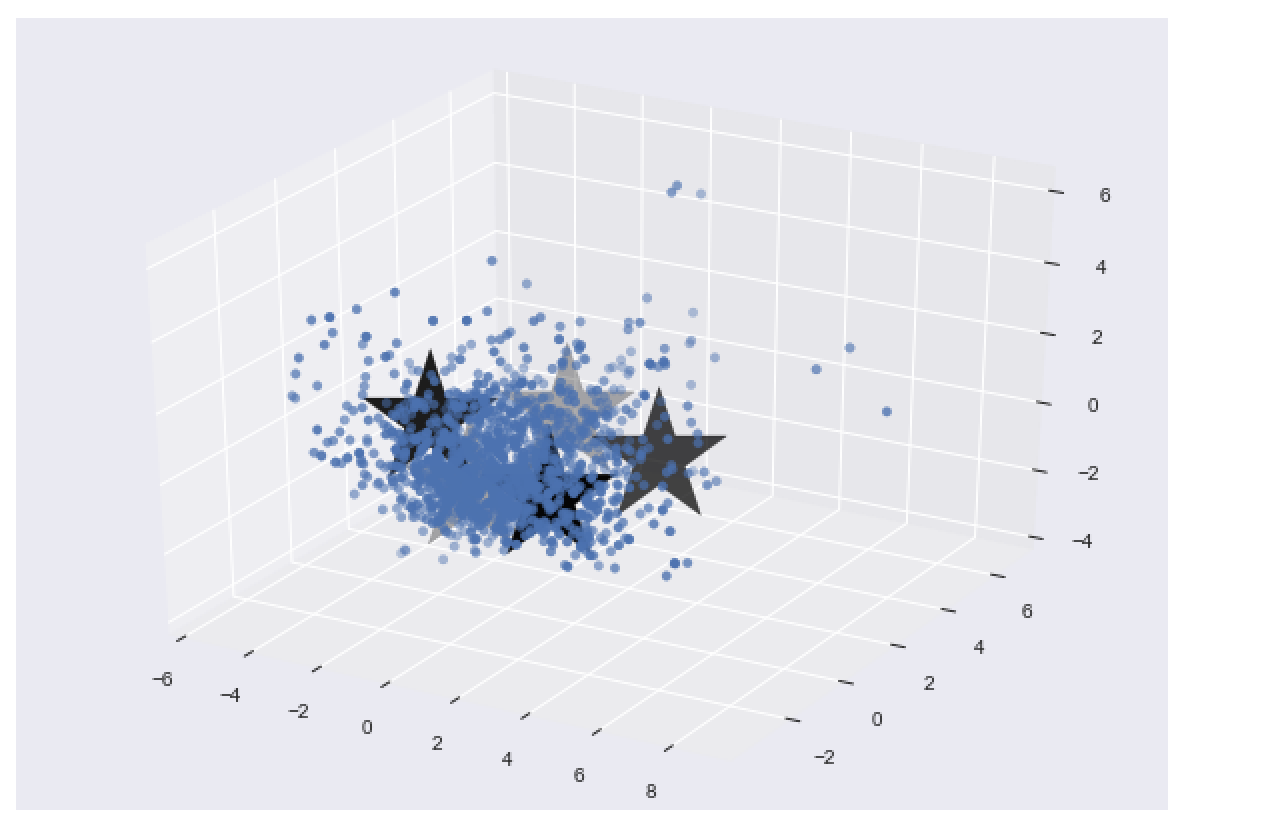


Here, we see 95% and above cumulative explained variance is around 8.

* Now, we know that we can reduce the feature dimension to 8 and use k-means clustering on these reduced dimension.
* The result for first feature and second feature from reduced 8 vectors

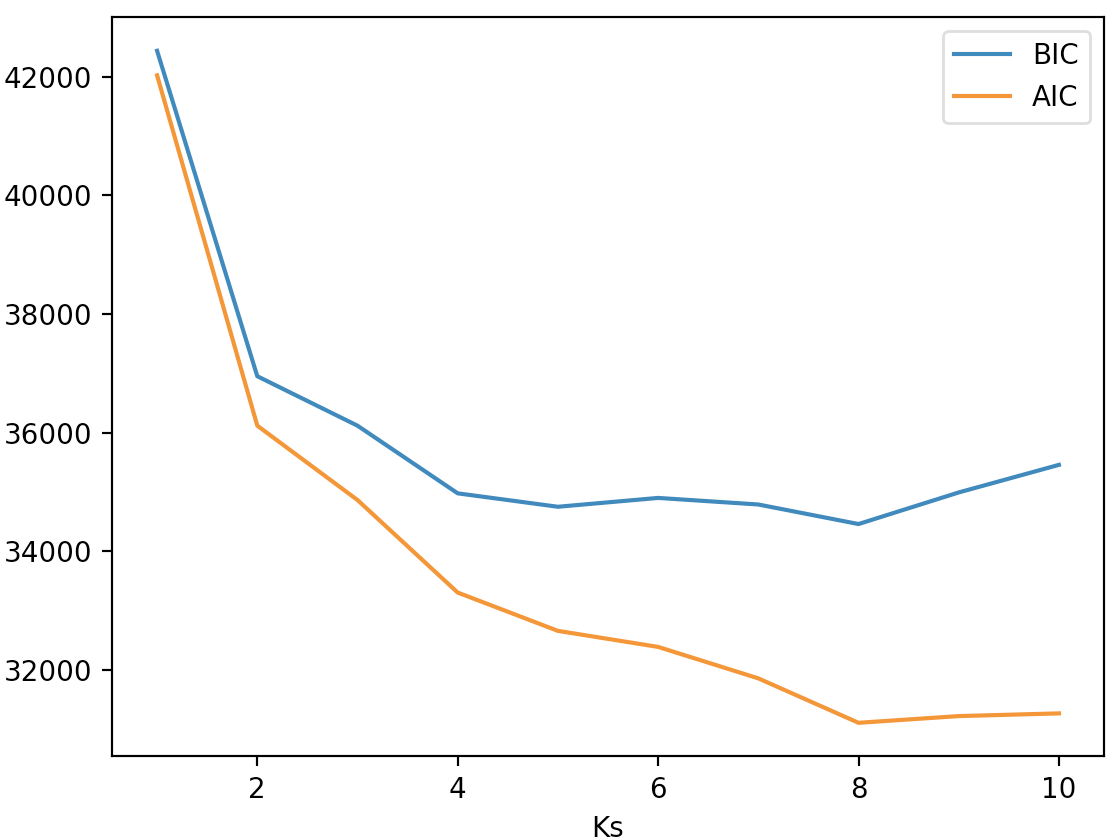


* The result in 3D space with first, second and third generated features



* From these two visualizations, we can achieve that K-means have been more spread in the data points and tend to achieve in more sparse way.

**GMM**



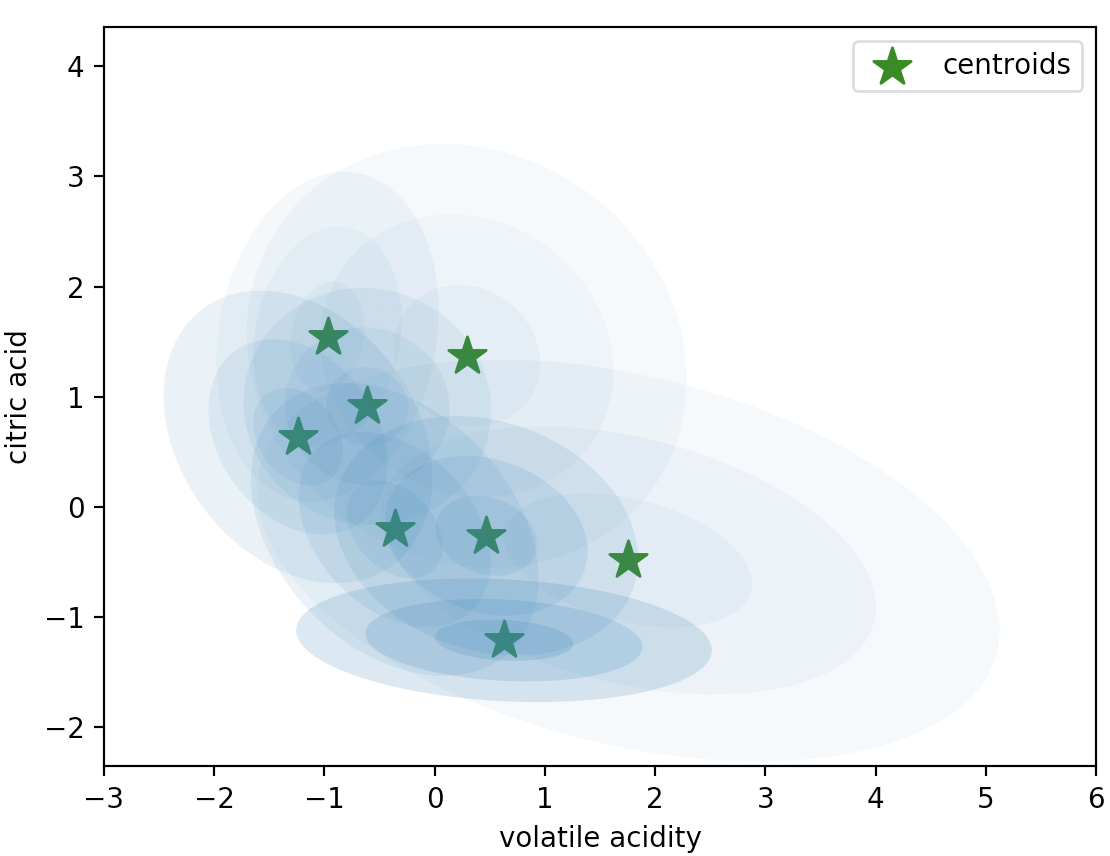
* The optimal # of components can be picked at 8, where AIC and BIC have minimum values.

By the research, we have found that feature 2,3, and 5 shows a pattern.

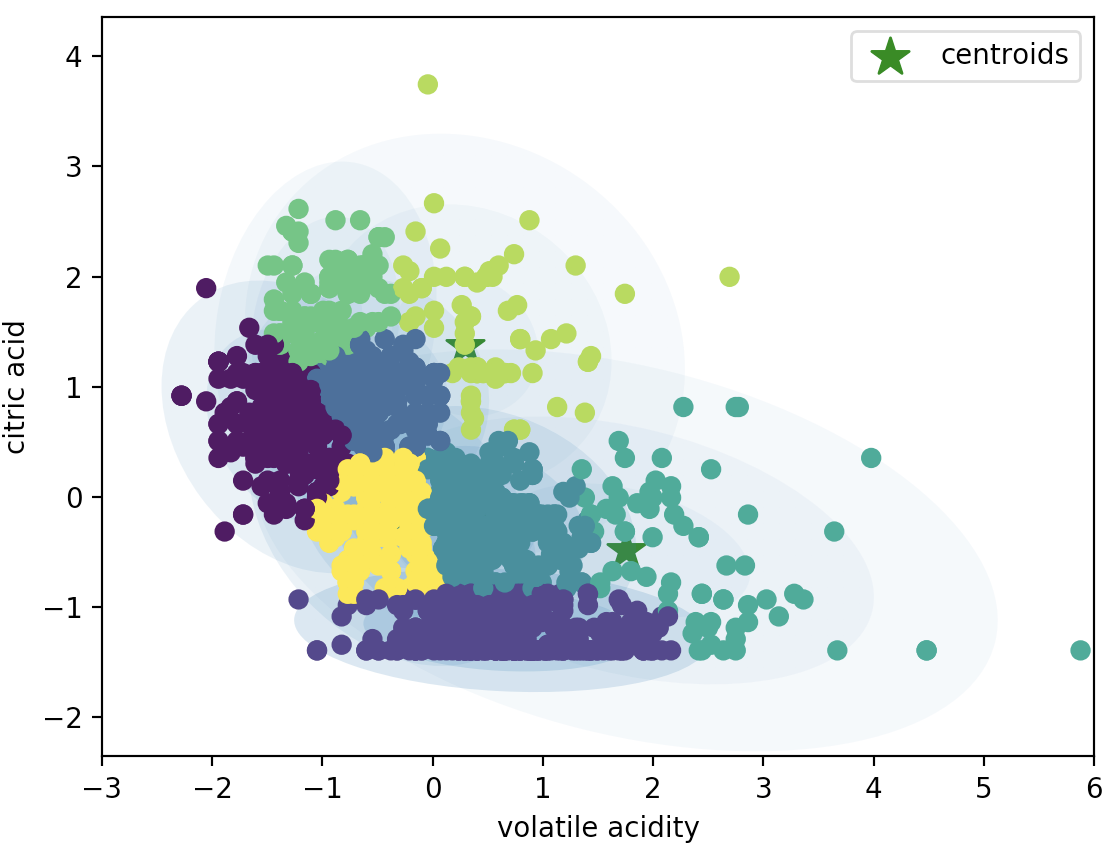
2 - volatile acidity   
3 - citric acid   
5 - chlorides

RED

1. Volatile acidity vs. citric acid

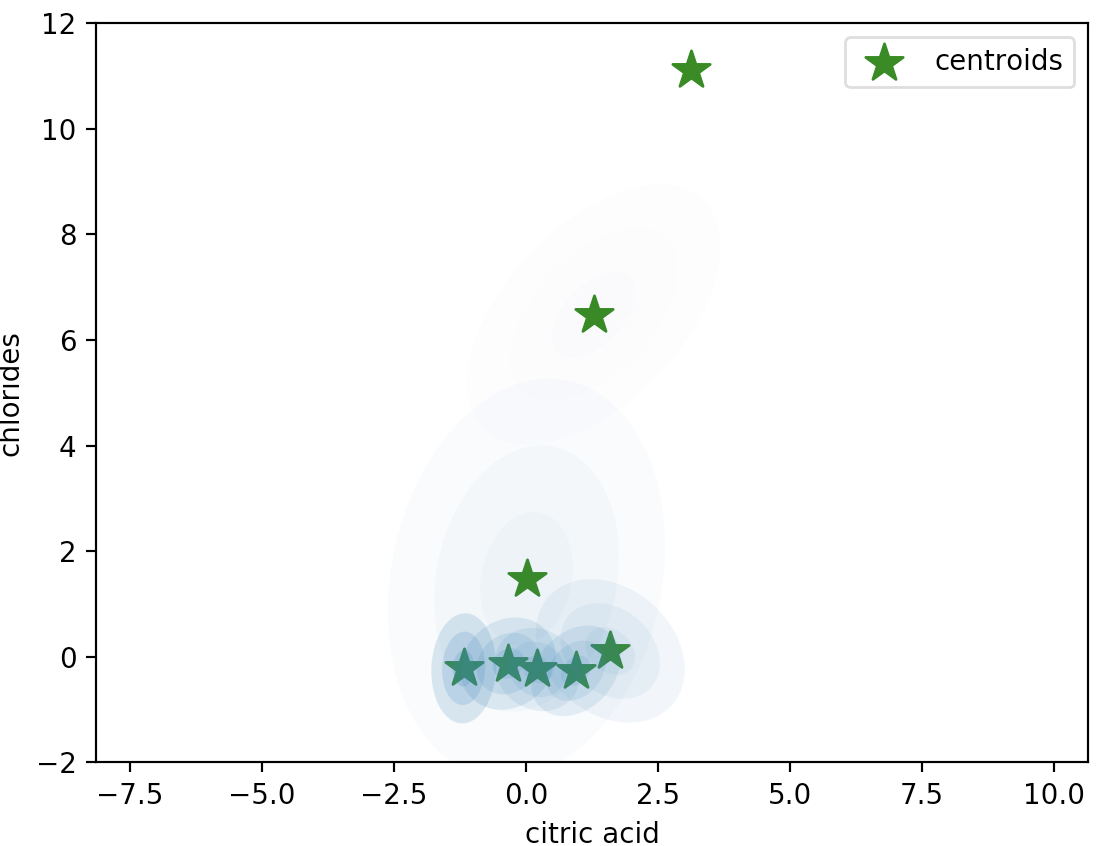


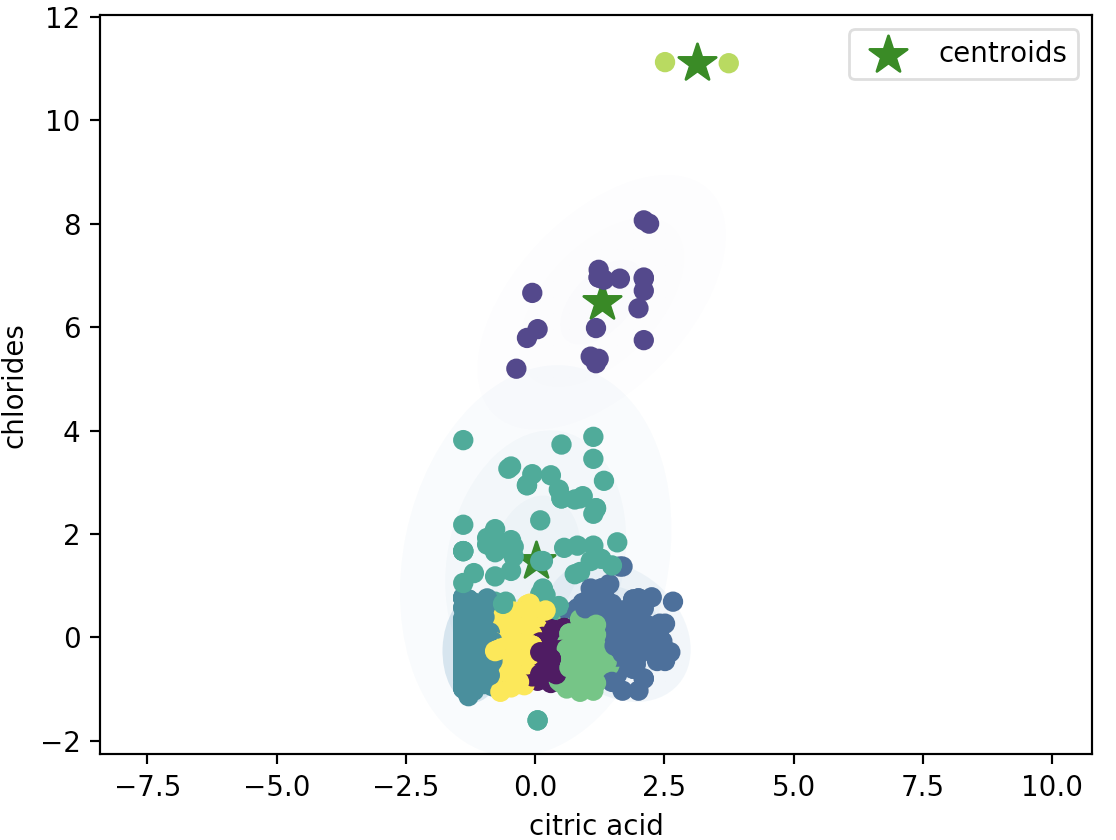
* Ellipses(covariance) around centroids



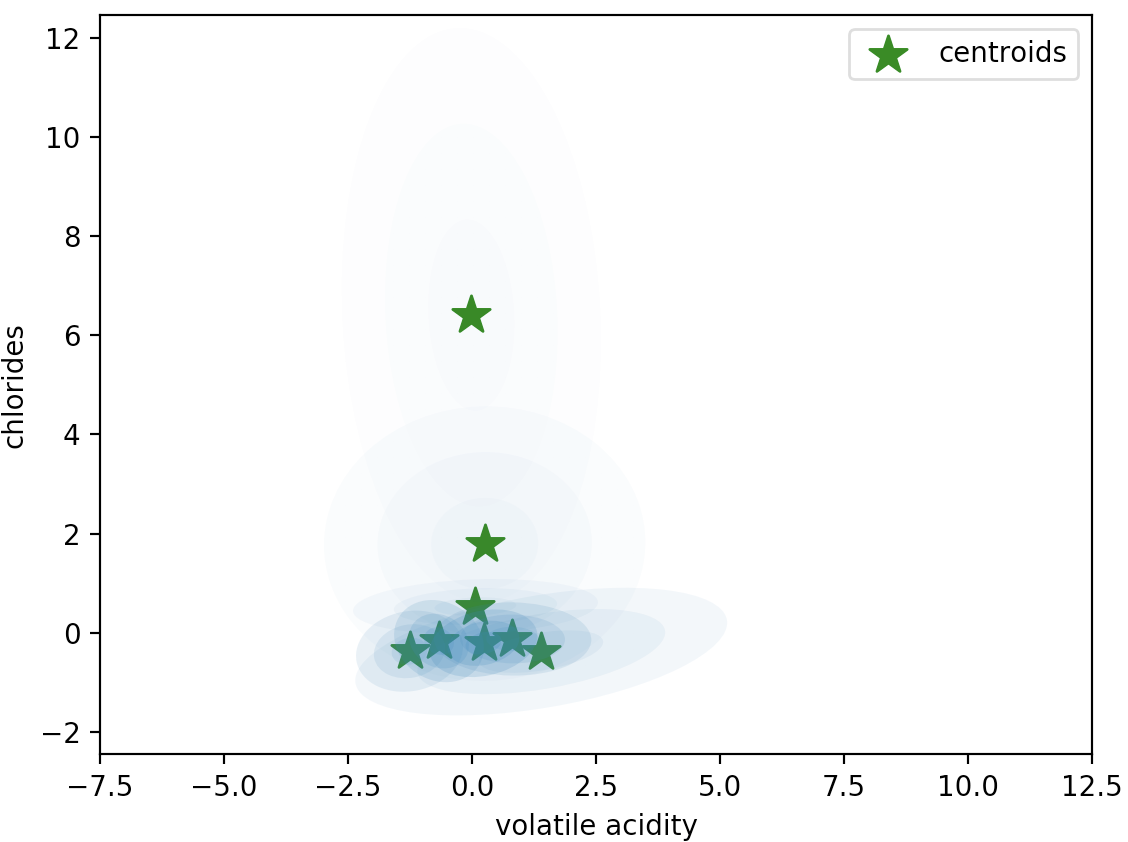
* Ellipses(covariance) around centroids with 8 clustered labels

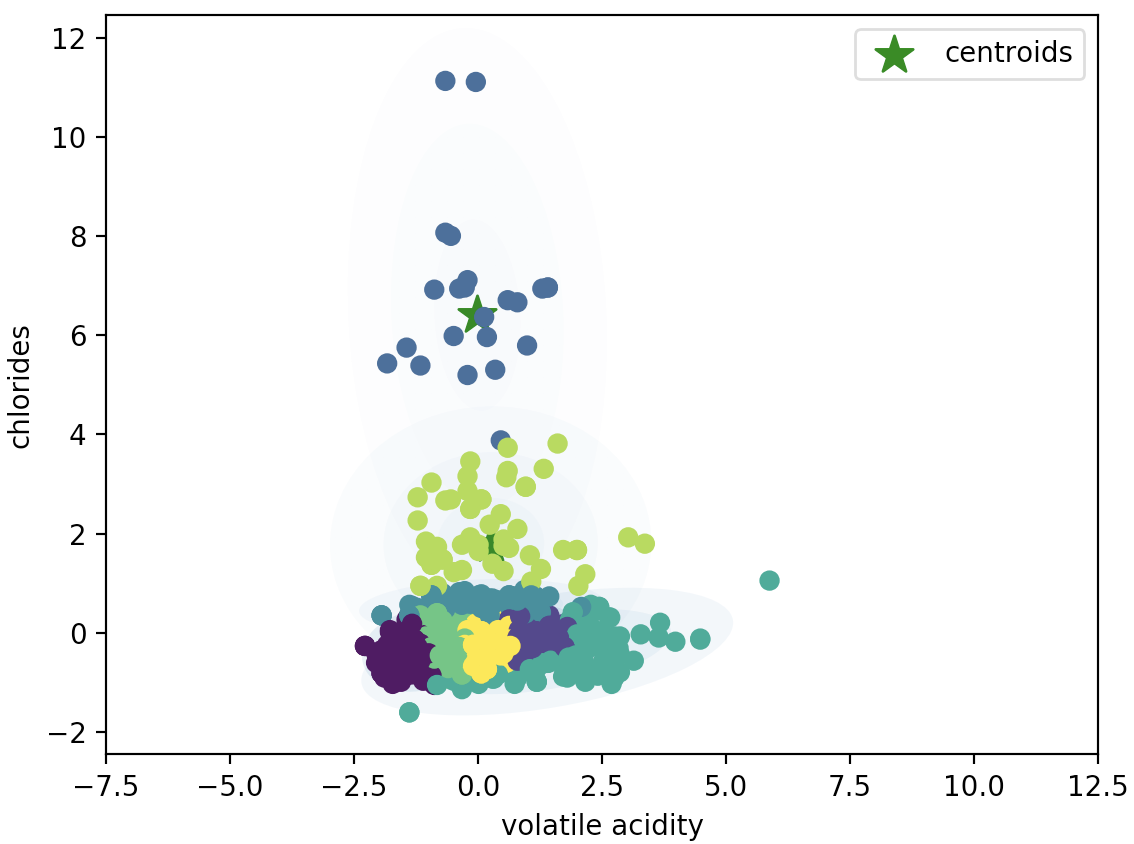
1. citric acid vs. chlorides





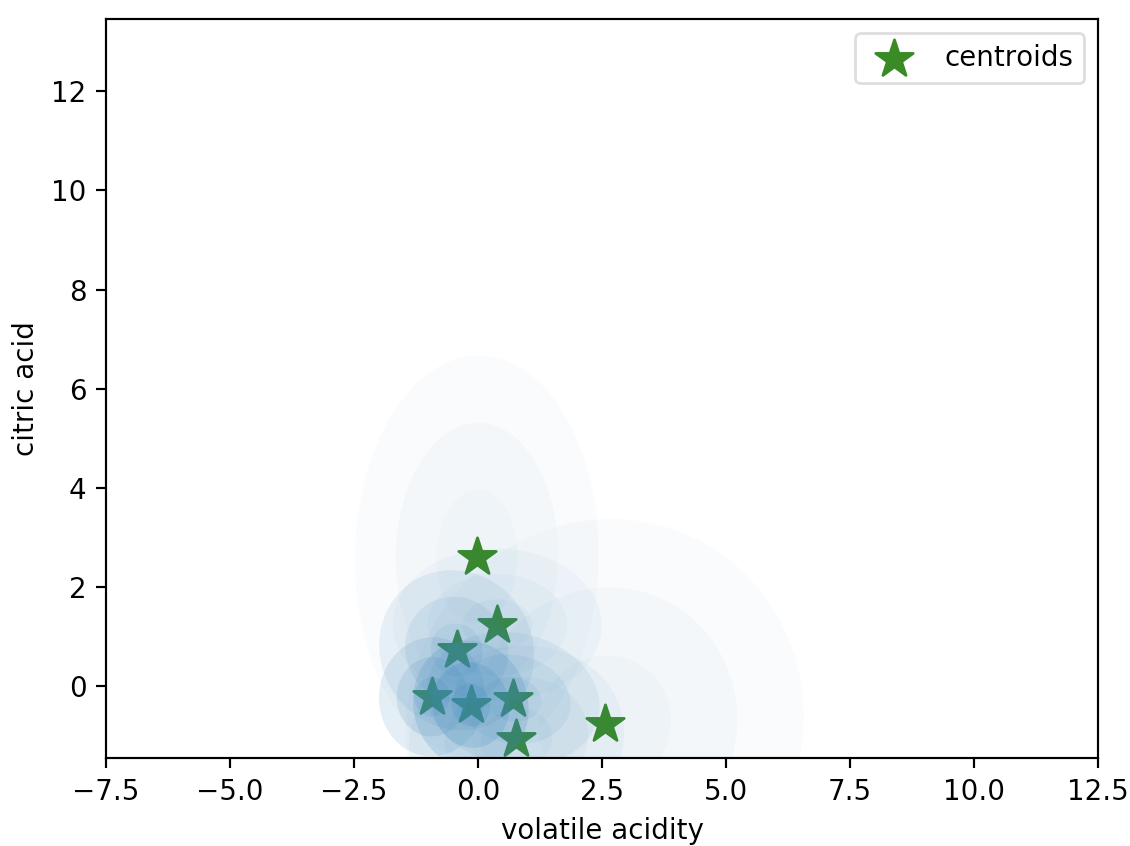
1. volatile acidity vs. chlorides

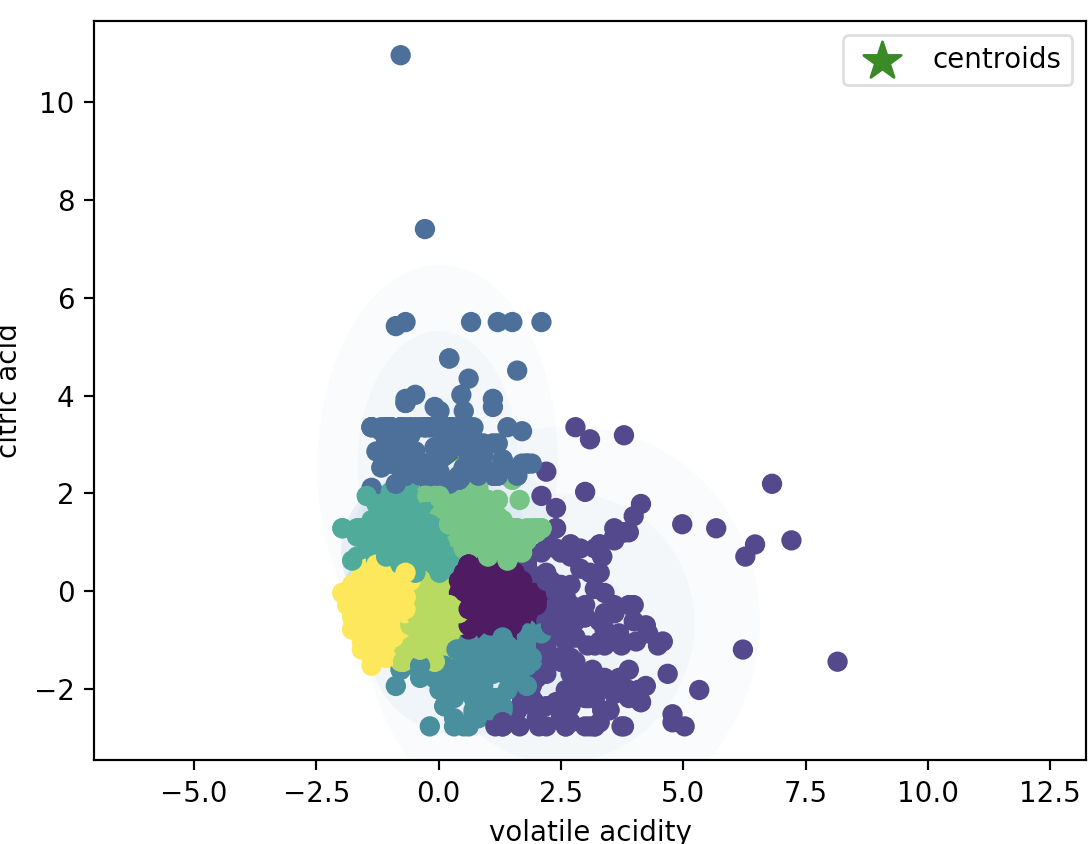




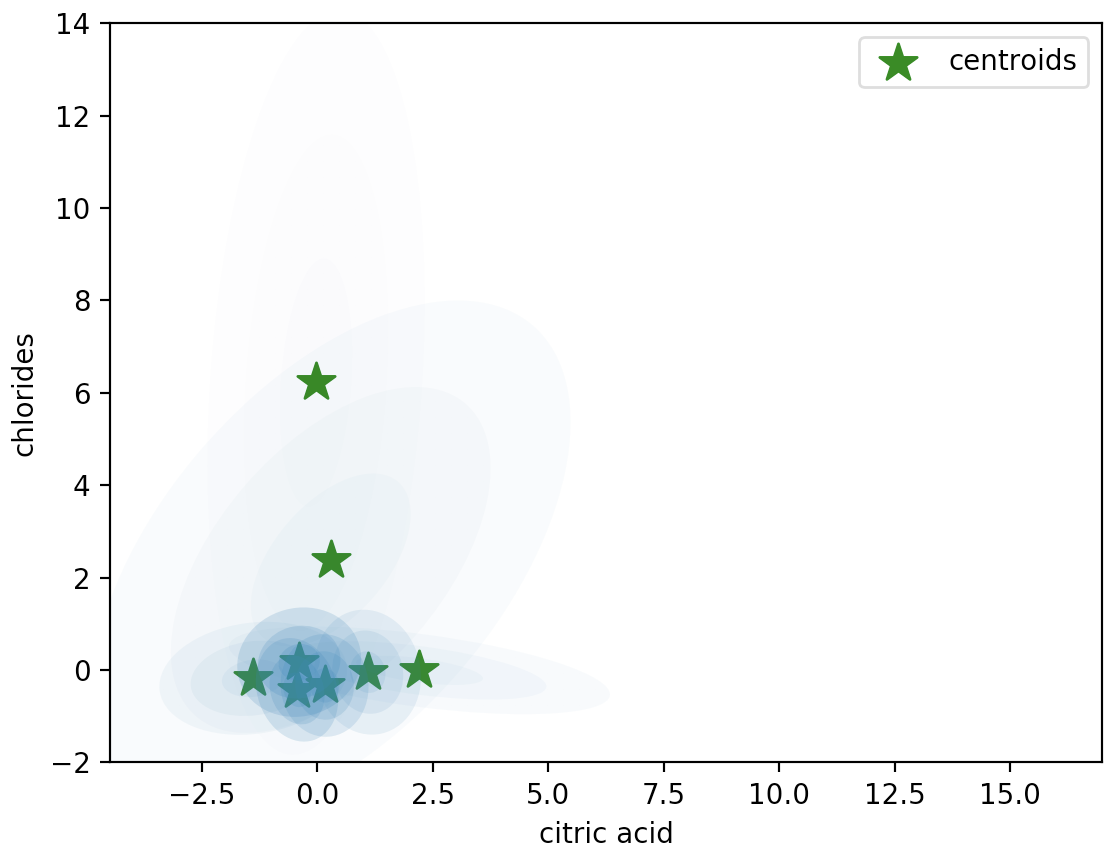
WHITE

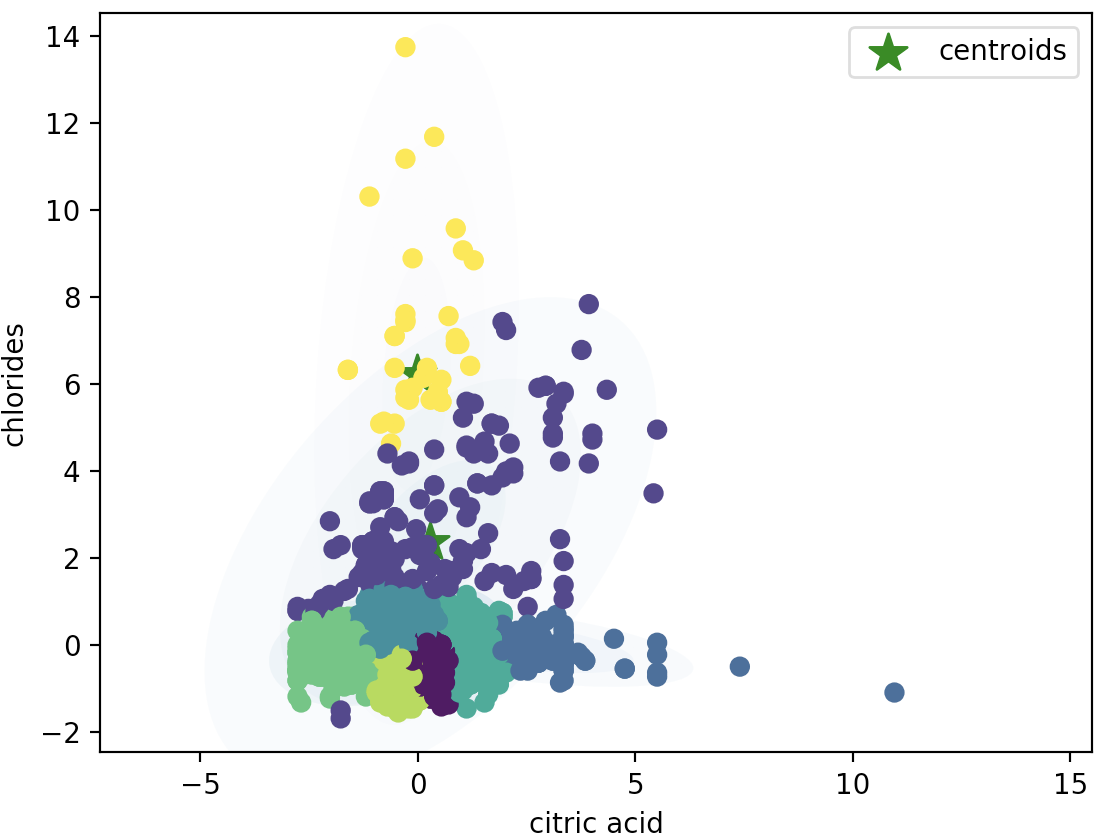
1. Volatile acidity vs. citric acid



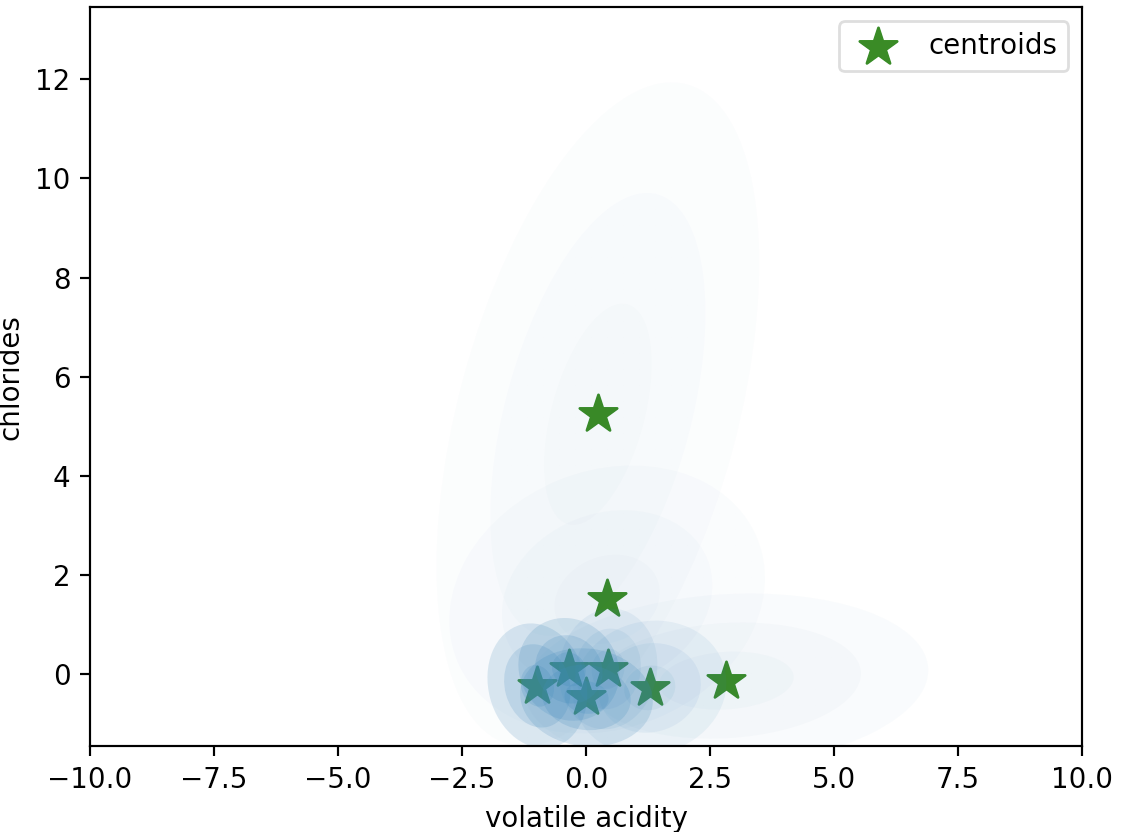


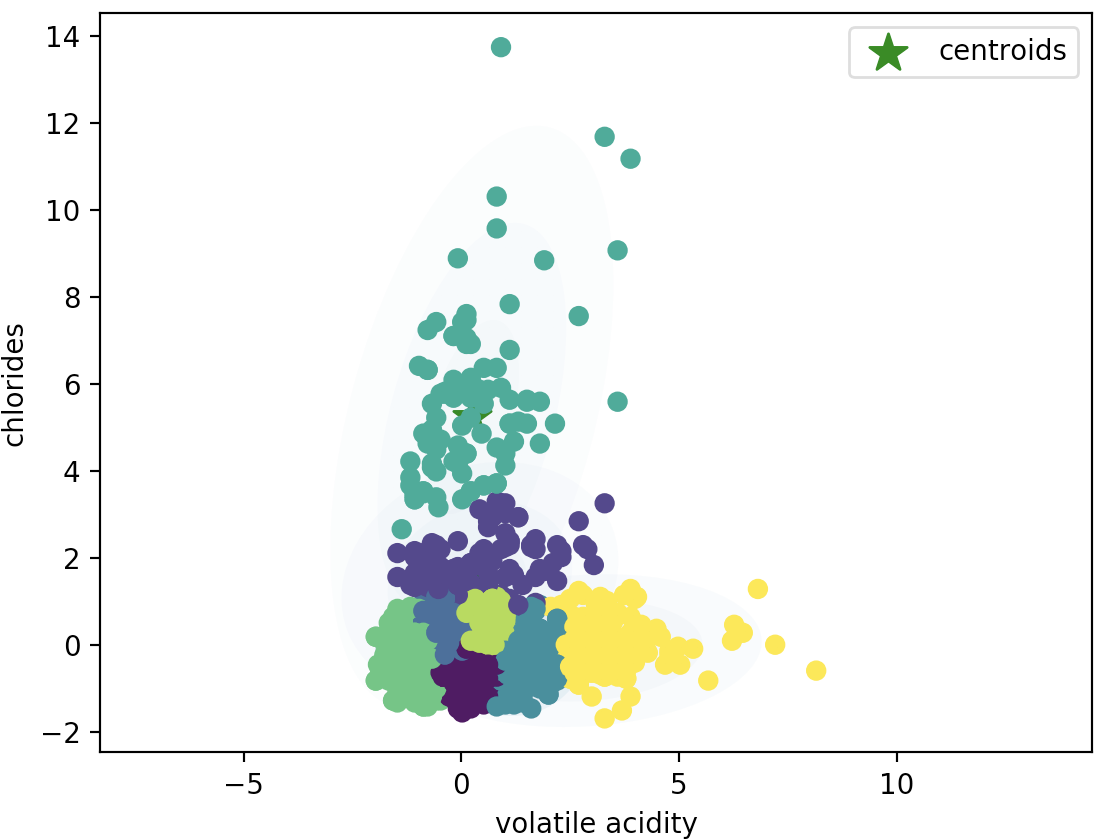
1. citric acid vs. chlorides



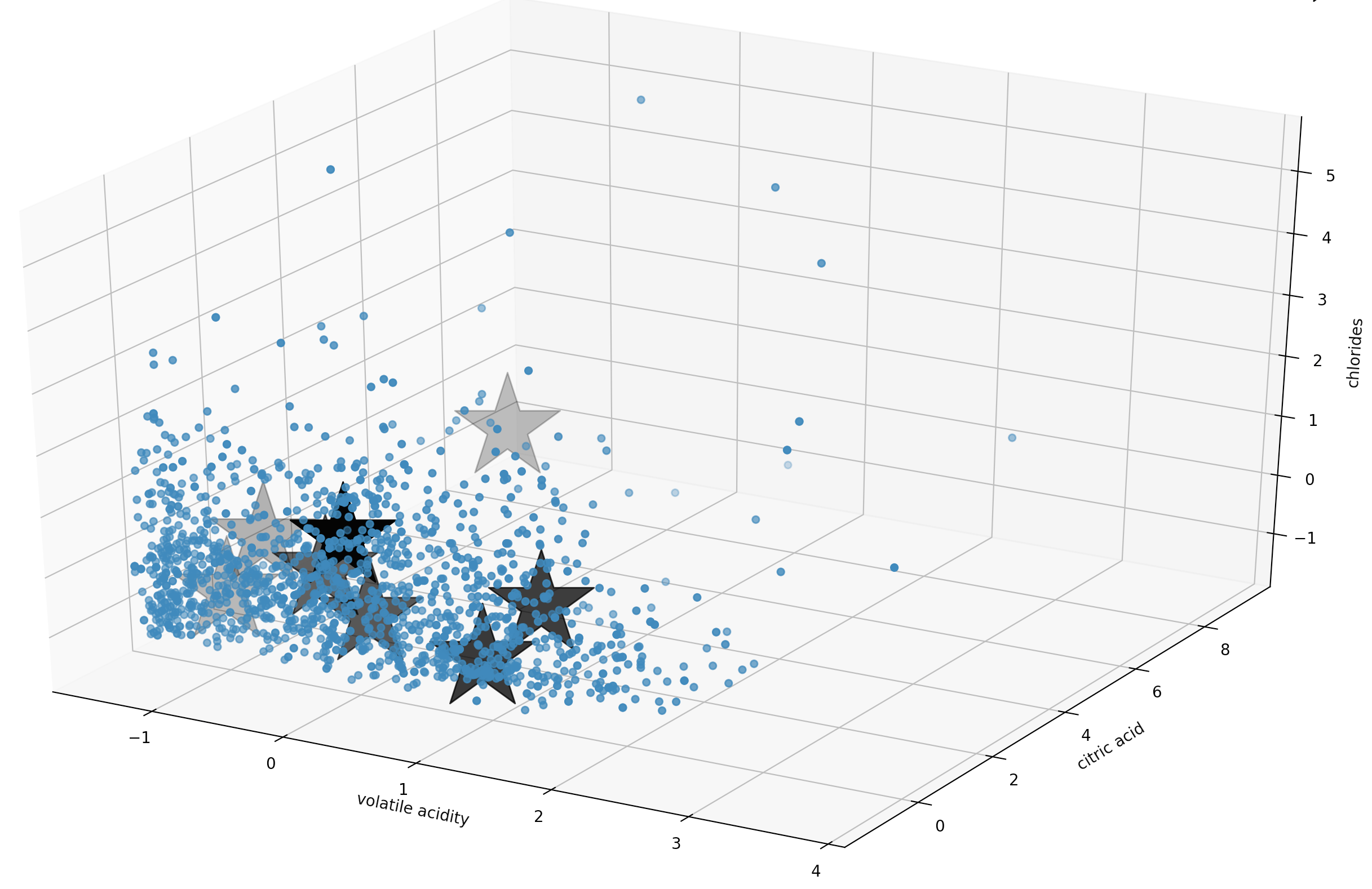
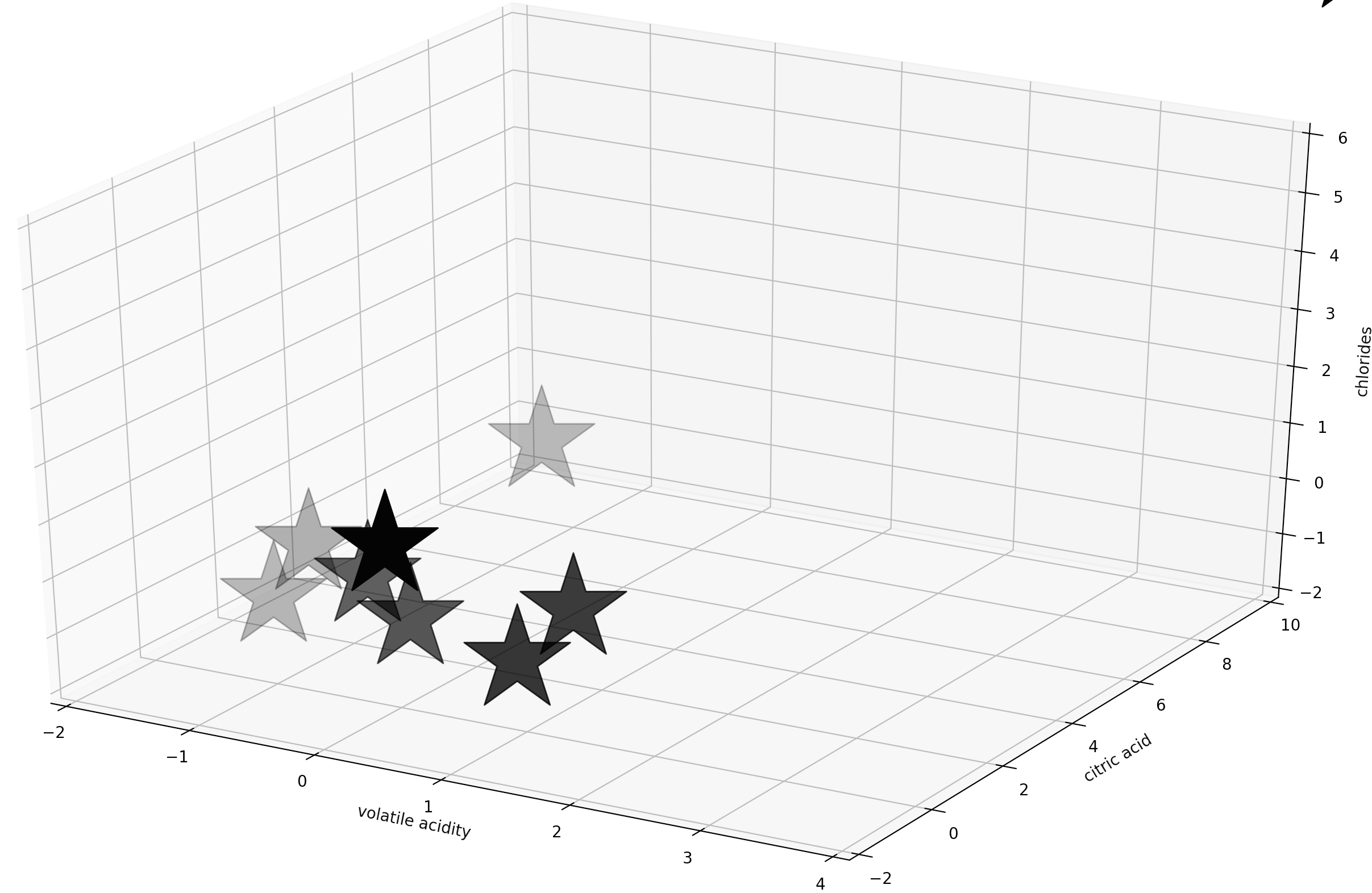


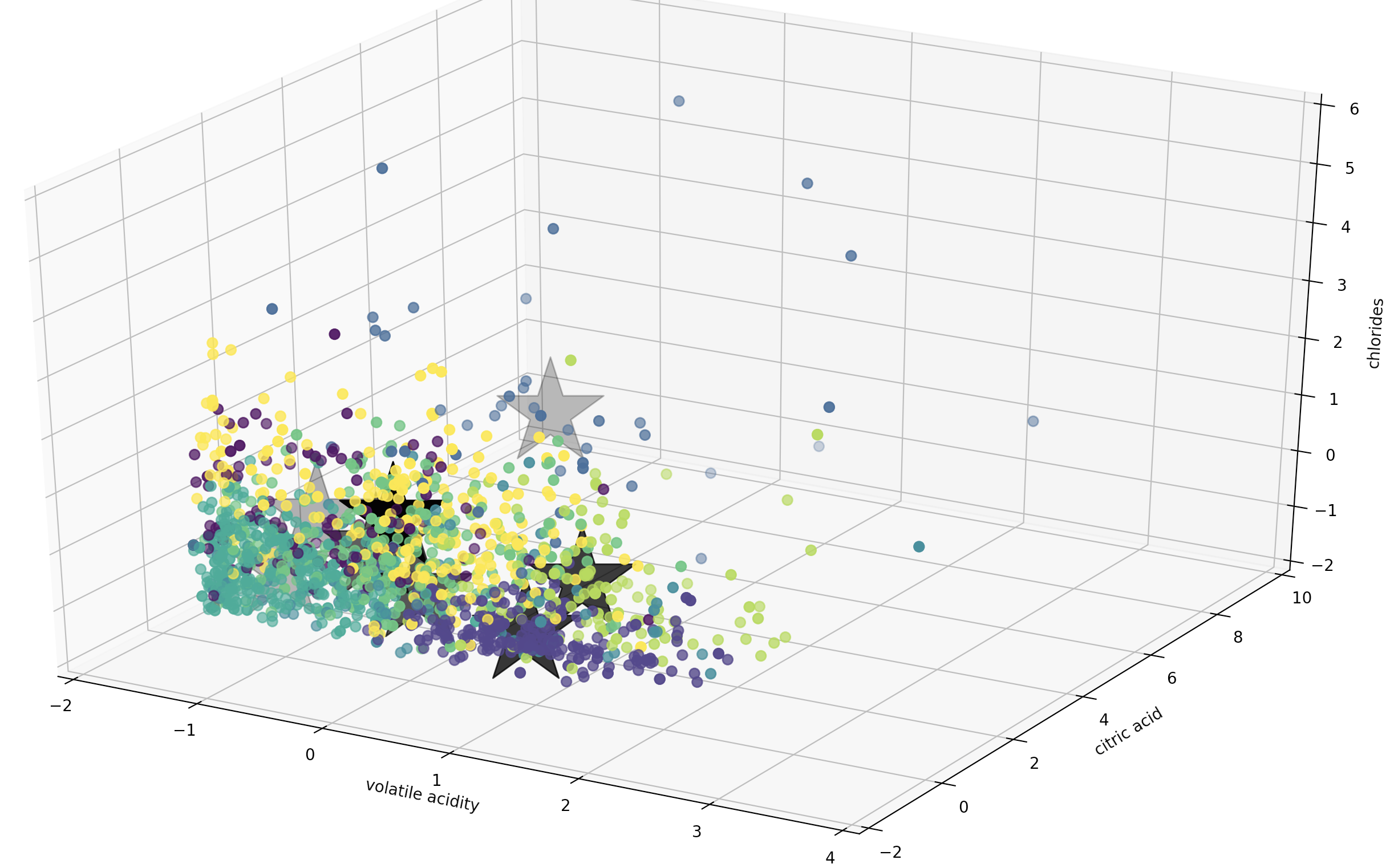
1. volatile acidity vs. chlorides





In 3D visualization with three features: fixed acidity, volatile acidity, and citric acid





With normalized data for offsetting constraints by different features, we can analyze through visualization by clustering the dataset into 8 partitions; however, since the feature space is high-dimensional and with large dataset, It is hard to see the clear distinction between each clusters. But, still it is optimal way of clustering for this dataset since with high dimension, we should use full covariance with 8 being the cluster size, which is obtained by AIC-BIC test.

Prediction accuracy for red wine dataset - 20.575359599749843%

Prediction accuracy for white wine dataset - 9.514087382605145%