UNSUPERVISED MACHINE LEARNING

Class and Clustering Project

NEW YORK UNIVERSITY, MACHINE LEARNING GRADUATE COURSE PAPER

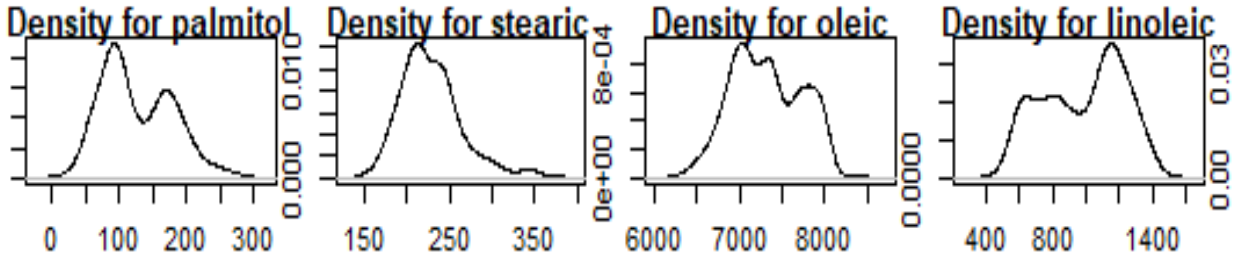
Unsupervised Machine Learning Class and Clustering Project

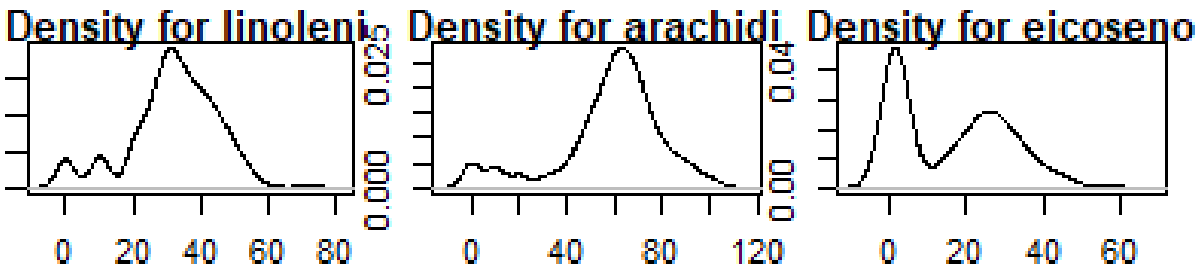
The data set I utilized for my project is called the olive oil dataset. This dataset includes eight types of fatty acids. The eight types of fatty acids are: (1) palmitic, (2) palmitoleic, (3) stearic, (4) oleic, (5) linoleic, (6), linolenic, (7), arachidic, and (8), eicosenoic. The data details the percentage composition of each of the fatty acids; these fatty acids are found in the lipid fraction concerning 572 differing Italian olive oils. In addition there are 9 areas of collection including 4 from southern Italy (Calabria, Sicily, as well as North and South Apulia), 2 from Sardinia (Coastal and Inland), and 3 from the area of Northern Italy (West and East Liguria and Umbria).

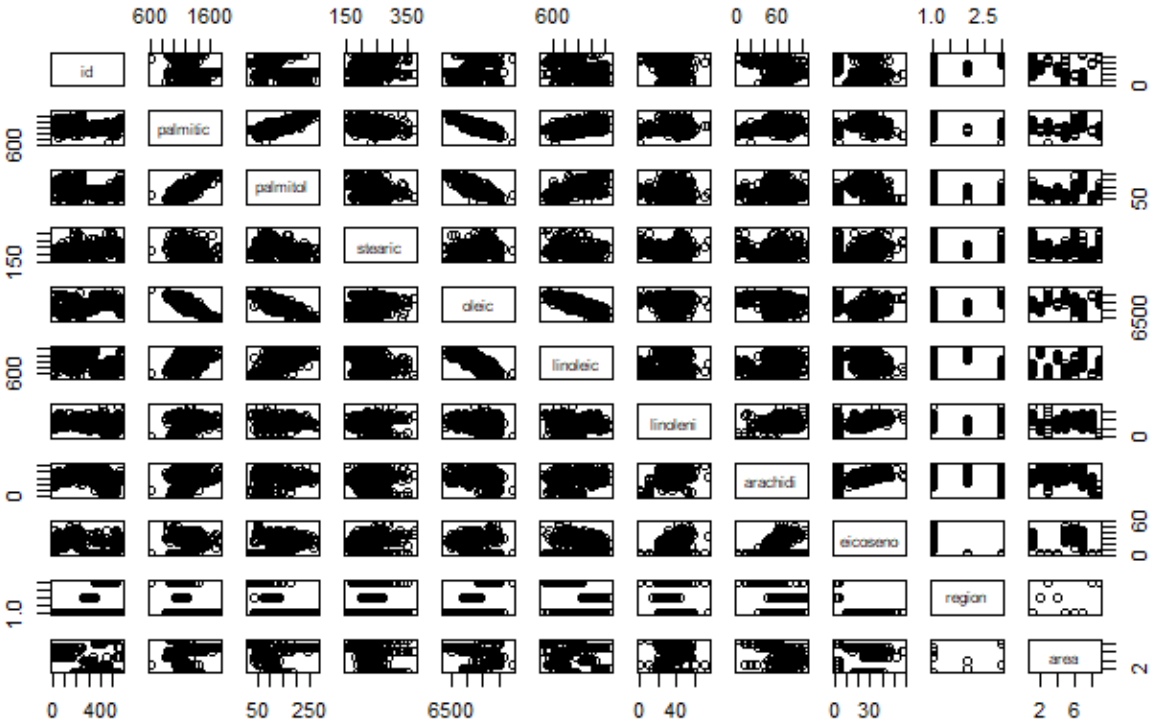
This data set has 572 rows and each corresponding row is related to a specific olive oil specimen. There are 10 columns. Columns 1 – 2 relate to the area that is macro, meaning the geographical area (partial reference is related to administrative borders of geographical areas in Italy) of origin of the olive oils. Additionally, in columns 3-10 there is information about the eight types of fatty acids and the chemical measurements on those eight fatty acids. Additional variables that have been measured that are no apart of the clustering atre the macro-areas and region of origins of the olive oils. I will make comparisons using these variables towards the end of this project.

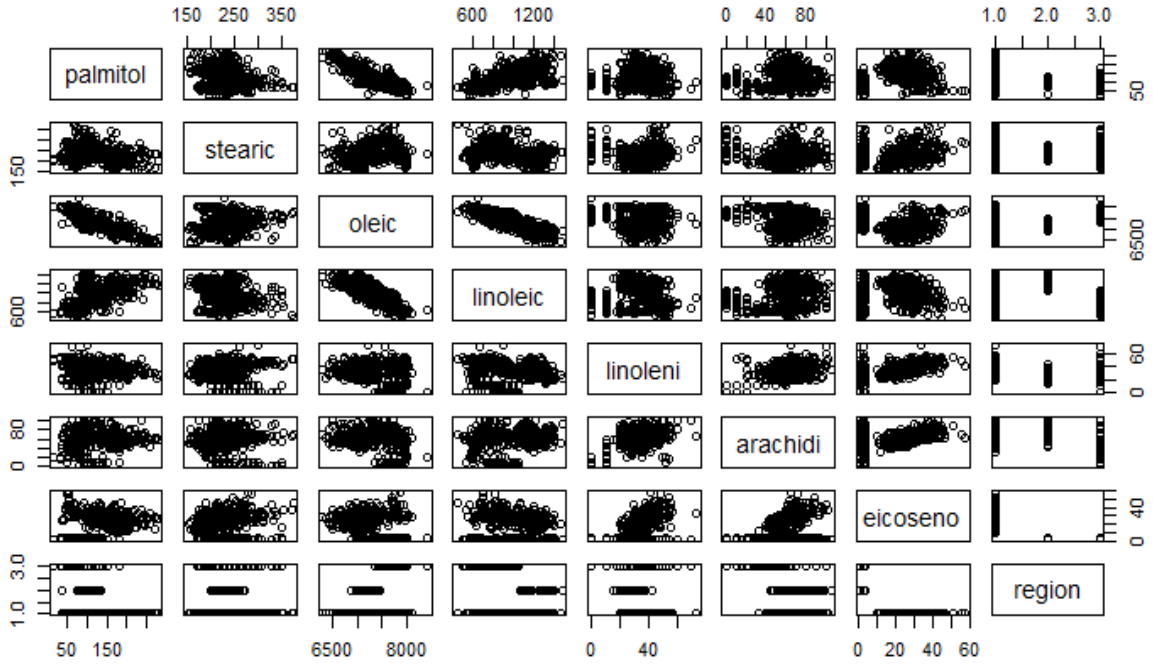
My first step included exploring the need for transformations/taking the logs or the need to rescale the measurements via analysing the univariate densities. The clustering was not carried out via principle components, instead I opted to use the original scale for clustering.

The data are fairly symmetrically distributed – not proposing a problem for our goal of clustering the data. Had the graphs displayed severe skews, then that would create a problematic condition. However, it is clear from the graphs below that transformation of the olive oil data set for the purposes of this project are unnecessary. Standardizing is optimal because it will accomplish the needed task of changing all the scale of the measures to be all the same in regards to units of standard deviation (sd).




Density for linoleni
Density for arachidi
Density for eicoseno
0.000
0.020
0.00
0.04
0 20 40 60 80
0 40 80 120
0 20 40 60


id
palmitic
palmitol
stearic
oleic
linoleic
linoleni
arachidi
eicoseno
region
area
600 1600
150 350
600
0 60
1.0 2.5
0 400
50 250
6500
0 40
0 30
2 6



(The graph above is the standardized graph)

There is evidence of bimodality but there is not enough of extreme skews in the above graphs hence, no transformation will be needed.

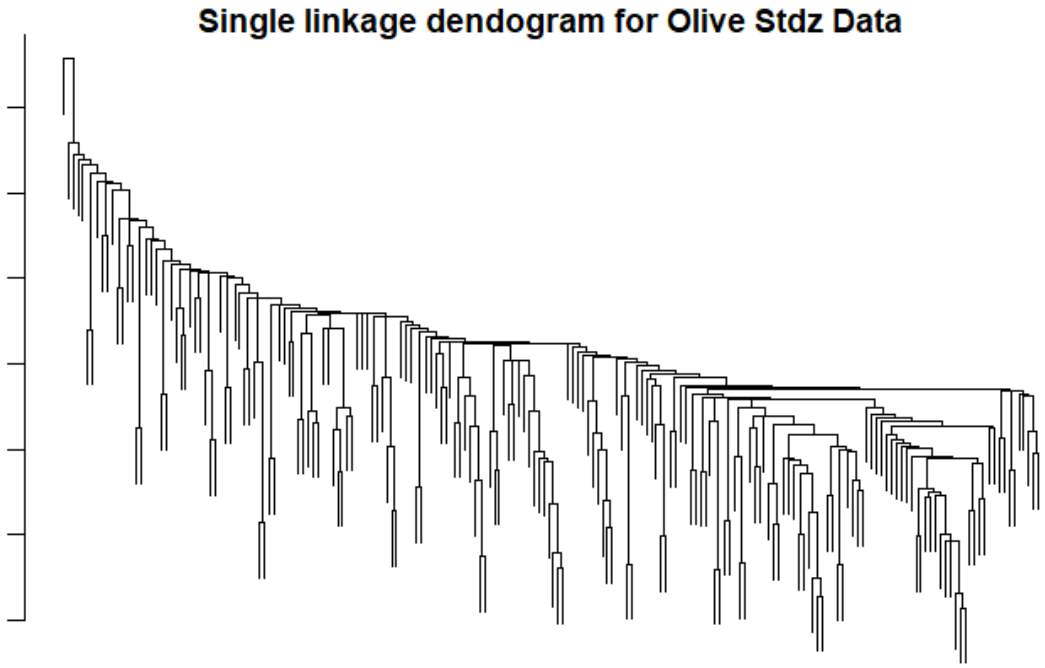
Furthermore, the bivariate plots are not as revealing or forward as I'd like. However, I can see via the axes' range that some of the bivariate plots possess features that have a greater amount of visability than other bivariate plots and their features. Given this, and the fact that there is so much obscurity amongst the clusters, I may need to adjust for this. Hence, I discern that standardizing this data set is warranted in regards to the raw data set. It is not certain given the above bivarite plots that standardizing using the PC is necessary since it may have very little consequence/alterations to the above bivariate plots. I explored the need for transformationvs via analyzing univariate densities while also replotting bivariate plots composed of new measurements. I also analyzed bivariate plots using the principal compoents using the standardized version of the olive oil data set. I utilized the princomp function to create the principle compenents and used columns 3:10 because they contained the needed features. The clustering were taken on the raw (and standardized) measures, and not the principal components.

For the purposes of this project I did standardize the data (using the z-form) via the following r-code:

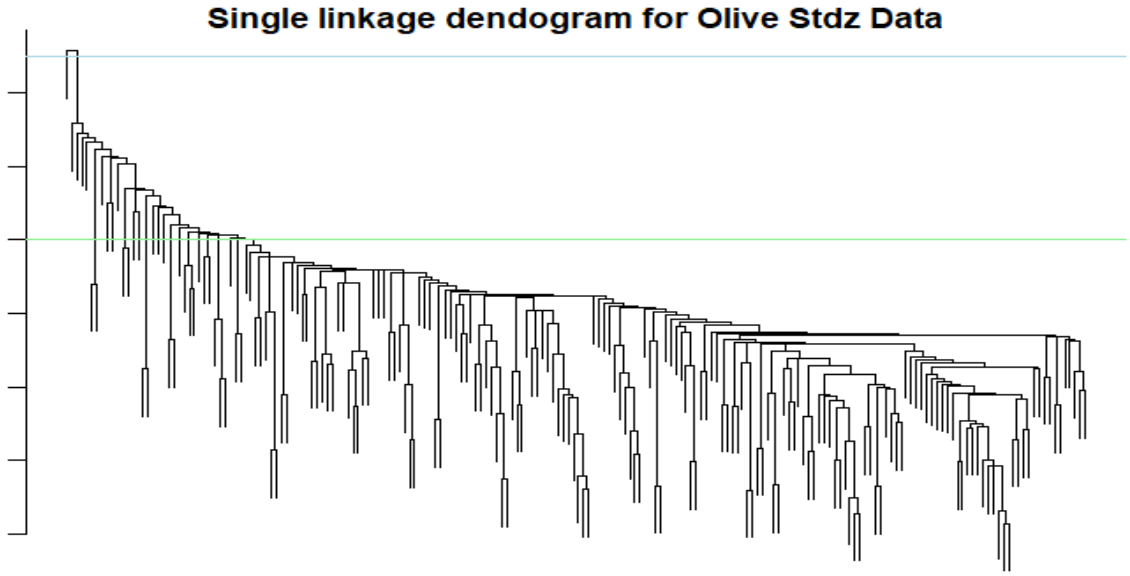
crabs.stdz <- crabs;

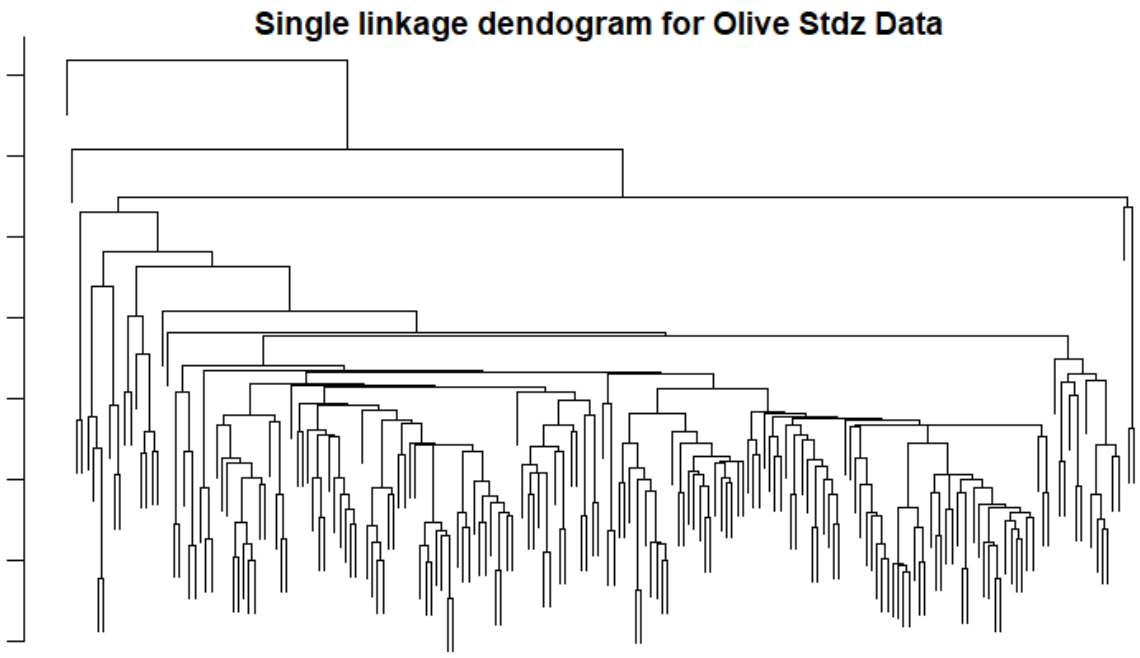
crabs.stdz[,4:8] <- scale(crabs[,4:8])

For the purposes of this project, the PC's pair plots were deferred until analysis of different clusterings occurred.



Given that I do see clearly defined clusters from the above graph, I will proceed with the single linkage hierarchical clustering. I will use the single and assume that the Euclidean distance is sufficient for the olive oil data set. I used single linkage in order to gain a clearer picture of the clusters that I am unable to view with other methods. In contrast, the centroid linkage would require movement towards solutions that are very tight instead of what we see above in which the single linkage produces very "stringy" looking clusters.





This is best for centroid method because it appears that the appropriate amounts of clusters may be 3 or 4 clusters. I prefer 3 clusters as the fourth one appears to not hold any clusters. I will now visualize the difference between 3 and 4 cluster solutions to gain more accurate information.

I prefer using the single linkage because it can give a clearer picture of the clusters and can be accomplished with ease. It allows for the easy understanding of cluster distances. On the other hand, Centroid linkage is also less sensitive in regards to outliers compared to other options such as single linkage and ultimately was not chosen because it is less sensitive to outliers. Hence, it appears that single linkage was the better choice given the clarity of viewing the clusters and because it's more sensitive to outliers. However, a downside is that centroid linkage usually doesn't perform as well as the Ward method or average linkage.

Clustering method

a. Choice of clustering method (and rationale)

The chosen clustering method is agglomerative hierchical clustering. I chose this method over other methods because this method did not require a notation of the number of clusters. The number of clusters can even be selected based on data visualization metrics for cluster analysis. Furthermore, I chose this method because this algorithm is not dependent upon choosing a particular distance metric. Given the complexity of the data set (i.e., the data set being spread out geographically) and well as based on acid type, committting to a distance metric seemed unnecessary because an option was available to use an algorithm that does not require establishing a distance metric. Given the data set and variables, it seems most optimal to use hierarchical clustering and R's hclust function. The R's hclust function will produce an automated results chart of the hierarchical clustering algorithm.

b. How you will deterimine the number of clusters (C(g) is only one approach, there is 1-Wilk's Lambda, BIC, e.g.)

The single linkage approach was utilized to determine the number of clusters. Using the hierchical clustering method, the number of clusters can be selected based upon what looks like the best number of clusters to choose.