EDA OLI EXTRAVERGINE D'OLIVA

Relazione di Francesco Malferrari

Il set di dati preso in analisi è proveniente da oli extravergine d'oliva di diversa provenienza, ciascuno rappresentato nel dataset con acronimi indicati di seguito:

- NA = Nord della Puglia
- SA = Sud della Puglia
- U = Umbria
- WL = Liguria ovest (si usa l'oliva taggiasca)
- EL = Liguria est

In particolare, è stata determinata la concentrazione di sette acidi grassi attraverso la gascromatografia (GC):

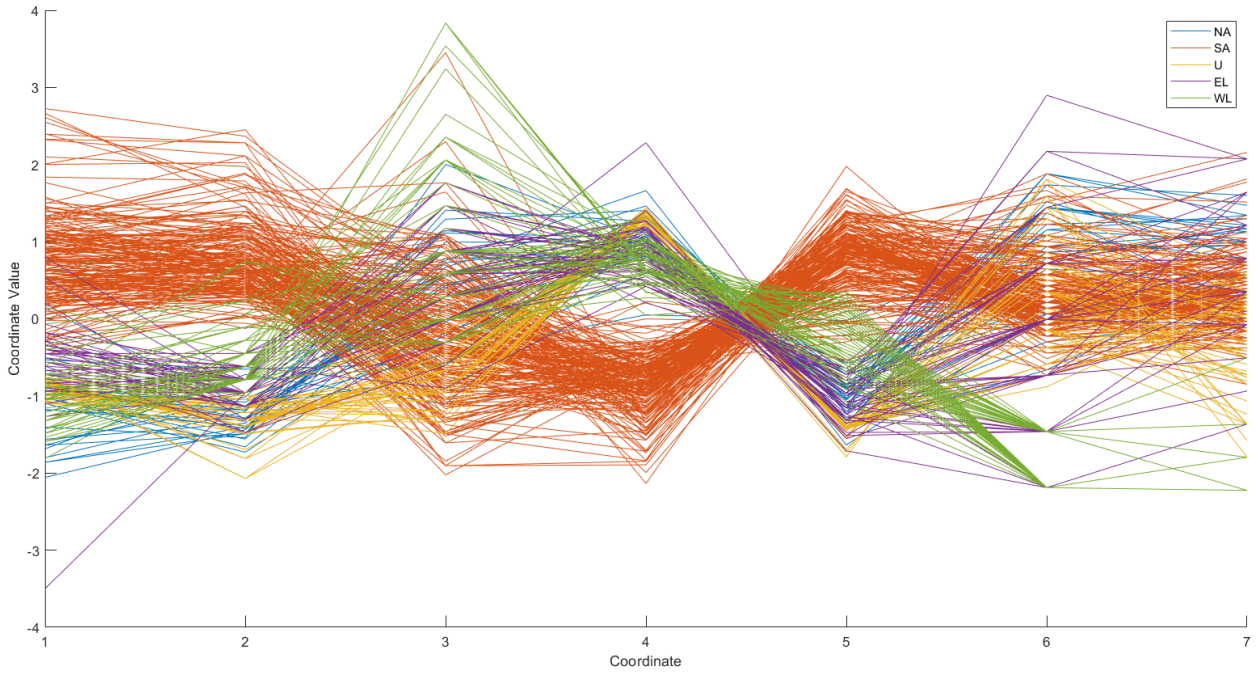
- Palmitico $CH_3(CH_2)_{14}COOH$
- Palmitoleico $C_{16}H_{30}O_2$
- Stearico $C_{18}H_{36}O_2$
- Oleico $CH_3(CH_2)_7CHCH(CH_2)_7COOH$
- Linoleico $C_{18}H_{32}O_2$
- Eicosanoico $C_{20}H_{40}O_2$
- Linolenico $C_{18}H_{30}O_2$

La tabella che colleziona queste concentrazioni per ciascun olio è composta da 7 variabili e 382 campioni (25 per NA, 206 per SA, 51 per U, 50 per WL e 50 per EL).

Quello che si vuole fare con questo dataset è l'analisi esplorativa dei dati (EDA) che consiste in diversi metodi per mettere in evidenza fenomeni come pattern, trend, correlazioni e dati anomali. Inoltre, lo strumento grafico permette più intuitivamente rispetto ai dati grezzi su tabella di vedere tendenze a raggruppamenti o distribuzioni in maniera indiretta.

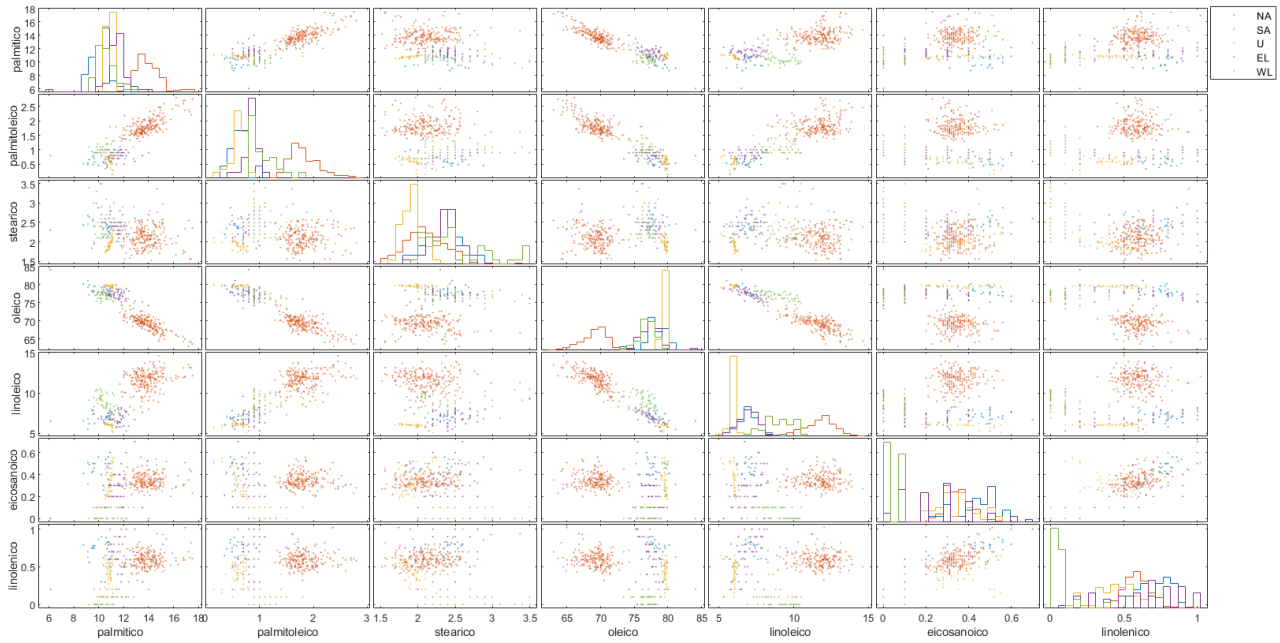
Attraverso l'utilizzo di Matlab sono state generate delle tabelle dalla quale si possono osservare i comportamenti dei dati.

Il primo grafico si chiama “Parallel coordinates” ed è utilizzato per mostrare, attraverso serie di punti corrispondenti a tipi di olio diversi, gli andamenti rispetto alle variabili, queste indicate sull’asse x (“coordinates”).



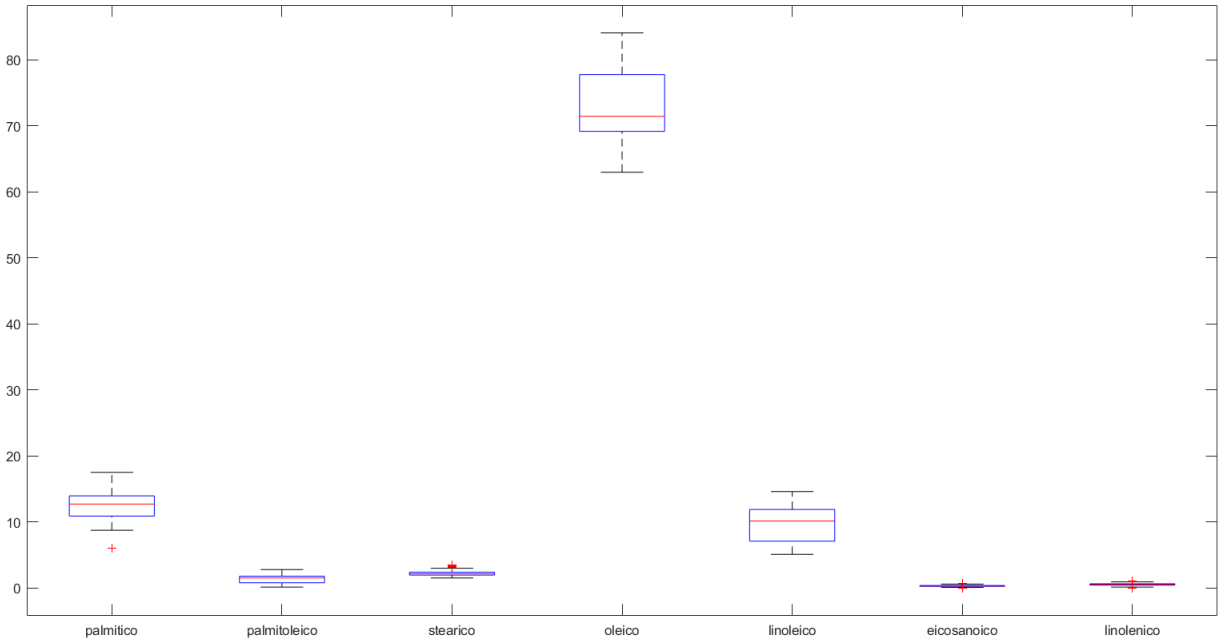
La considerazione che si può fare è che l’olio SA mostra dei valori più alti quando gli altri hanno concentrazioni più basse e viceversa, il che può già mostrare dei raggruppamenti dove SA è separato dagli altri. Questi ultimi hanno valori molto simili e sono difficili da distinguere (con l’eccezione del WL che è più visibile rispetto agli altri).

Il secondo strumento è il gplotmatrix che genera una matrice di scatter plot per ogni combinazione di variabili del dataset. Nella diagonale sono presenti istogrammi che misurano la distribuzione di ciascuna variabile.



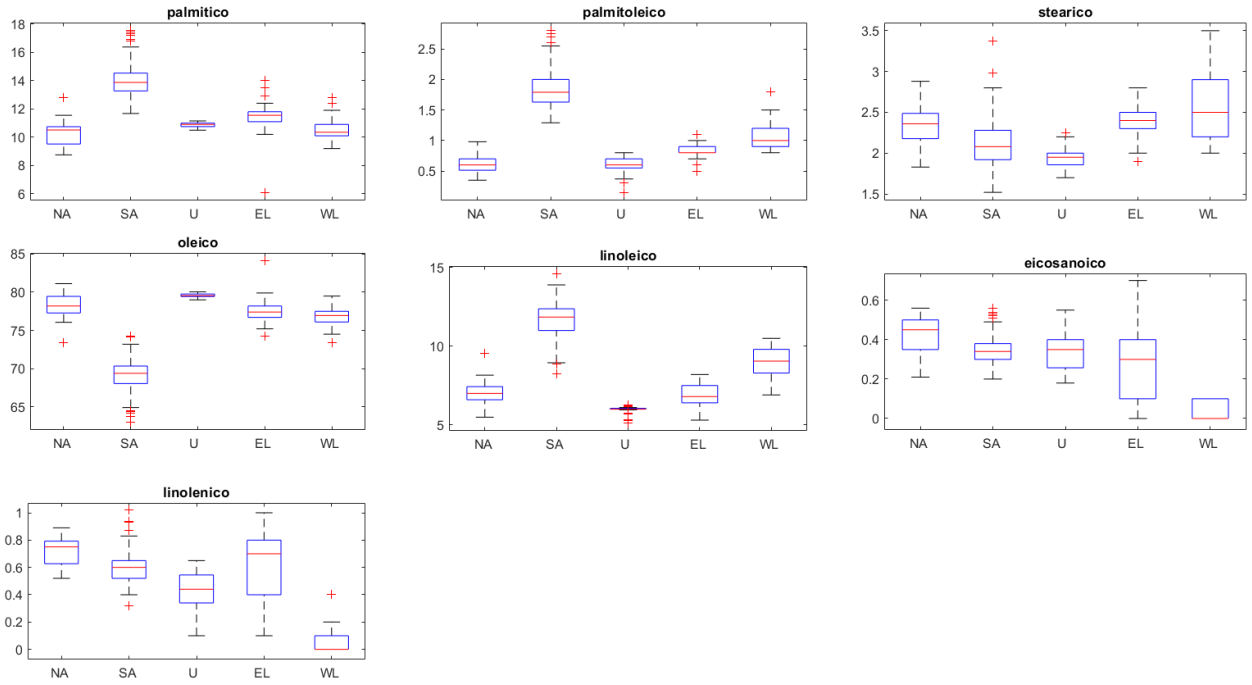
La cosa che si può notare è che il SA è più distinguibile e separabile nella maggior parte dei casi rispetto ad altri oli, mentre quest'ultimi sono inseparabili fatta eccezione del WL che è possibile isolarlo in maniera approssimativa nella maggior parte dei casi. Le combinazioni che non permettono a SA di essere separabile da opportune rette sono stearico-linolenico, eicosanoico-stearico ed eicosanoico-linolenico. Invece le combinazioni che non permettono di separare in maniera accettabile il WL da opportune rette sono oleico-palmitico, oleico-palmitoleico, palmitico-stearico e oleico-stearico.

Il terzo strumento è il box plot che serve per mostrare la distribuzione dei valori per ciascuna variabile. Su ogni casella raffigurata, la linea centrale indica la mediana, mentre i bordi inferiore e superiore della casella indicano rispettivamente il 25° e il 75° percentile. I baffi si estendono ai punti dati più estremi non considerati gli outliers. Quest'ultimi vengono raffigurati individualmente utilizzando il simbolo del marcatore "+".



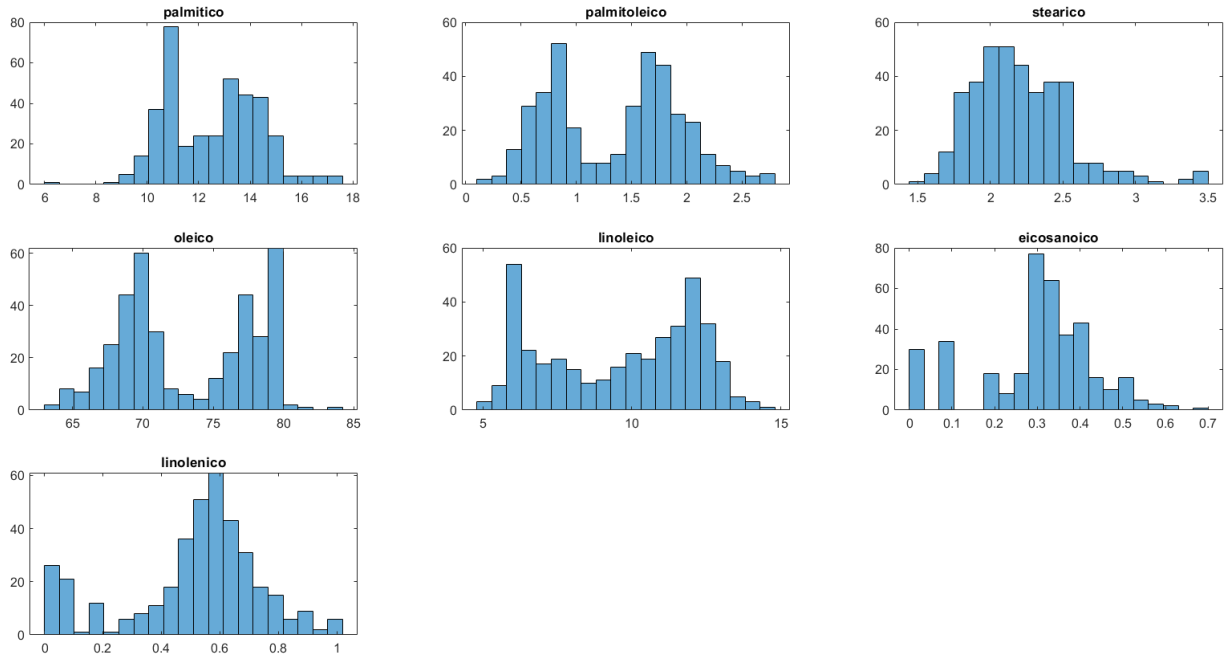
Da questi dati non si possono estrapolare troppe informazioni: si può affermare che i valori dell'acido oleico sono più alti rispetto agli altri. Le altre variabili tendono ad avere range di valori più piccoli, specialmente gli acidi palmitoleico, stearico, eicosanoico e linolenico che risultano essere molto sottili. In questi ultimi casi sono presenti anche degli outliers.

Per una migliore rappresentazione si è deciso di realizzare sette grafici, dove ciascuno di questi rappresenta il boxplot di un acido e sull'asse x sono presenti i vari tipi di olio.



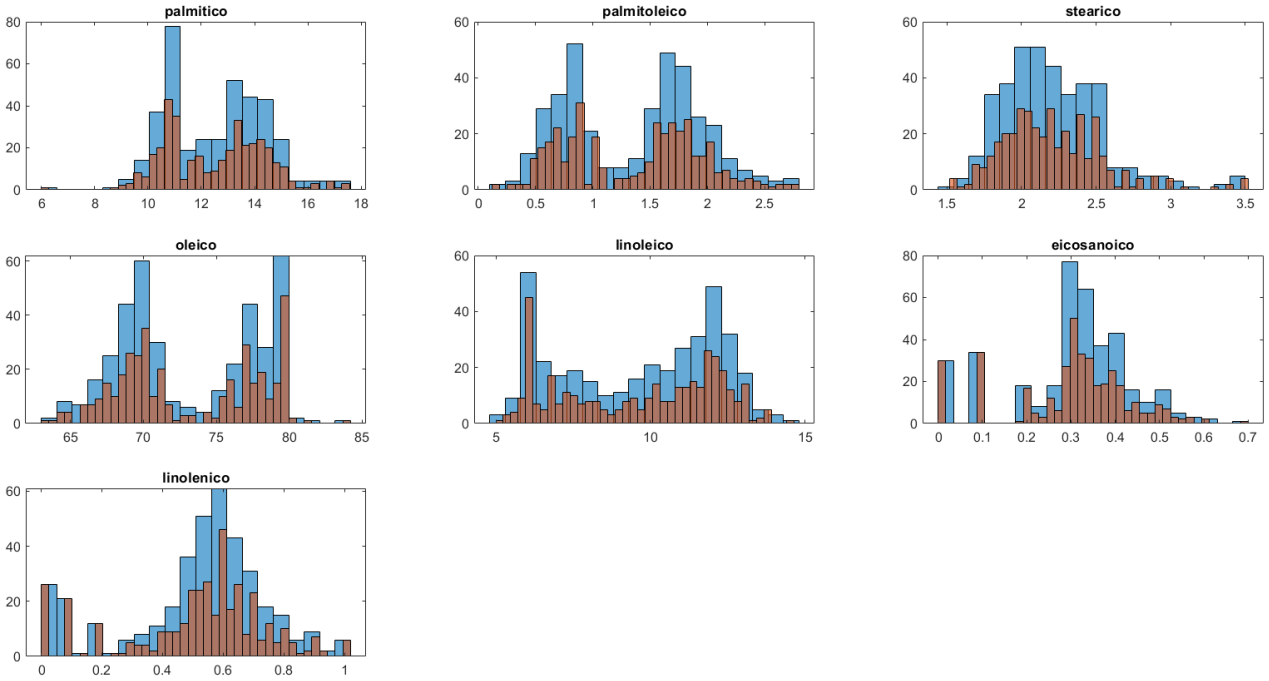
Questi mostrano già qualcosa di più interessante: per quanto riguarda gli acidi stearico, eicosanoico e linolenico i range di tutti gli oli si intersecano l'uno con l'altro (tranne un po' il WL negli ultimi due oli). Mentre per quanto riguarda gli altri acidi, l'olio SA non si interseca quasi per nulla (a differenza degli altri), rendendo questi acidi abbastanza determinanti per identificare il tipo di olio tra il SA e tutti gli altri. Quindi si possono identificare due gruppi. Un altro aspetto che si può notare è la presenza di diversi valori anomali, in grande percentuale per l'olio SA.

Il quarto strumento è l'istrogramma di frequenza, per la precisione sette, che corrisponde al numero di variabili. Questo tipo di grafico rappresenta in istogramma uno scatter plot dividendo i valori in intervalli (detti "bins") e l'altezza di ciascuna barra corrisponde al numero di oggetti in un determinato bin.

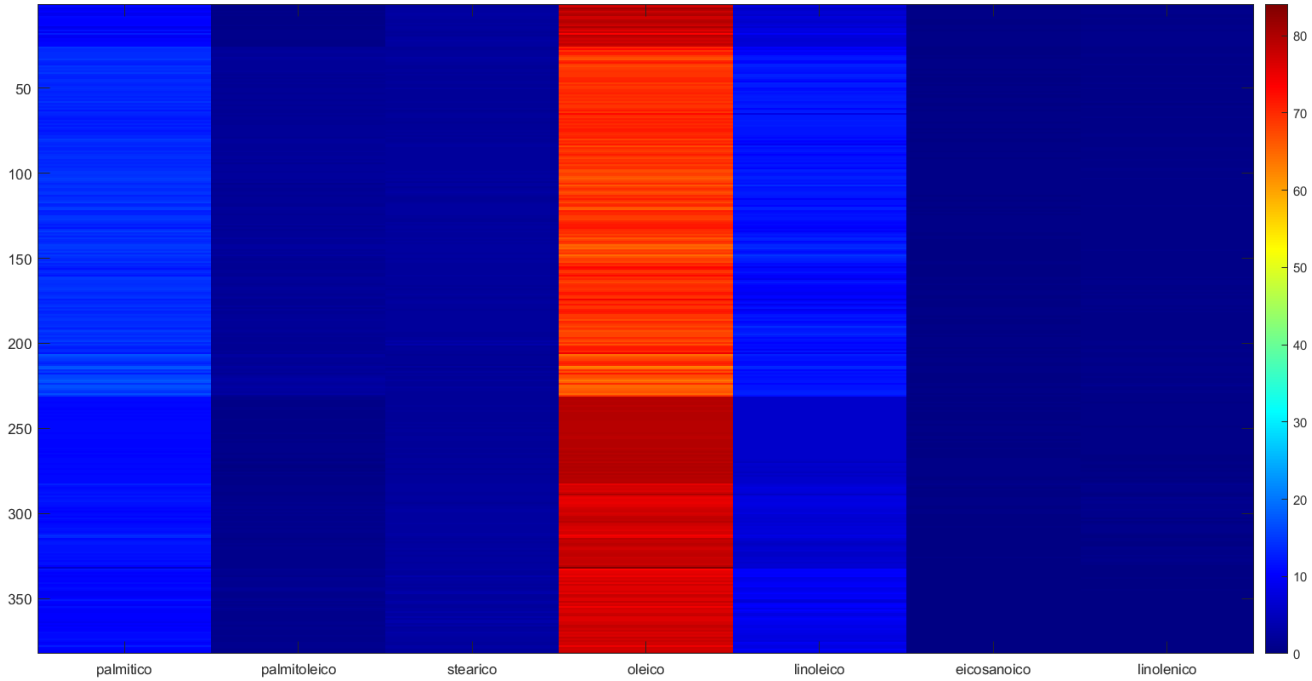


Come già accennato per i boxplot, sono visibili due raggruppamenti che assomigliano in maniera molto approssimativa a gaussiane. Le eccezioni sono gli acidi linolenico ed eicosanoico, che presentano un gruppo molto grande e uno (o più) molto piccolo quasi inesistente, e lo stearico che è un gruppo unico. Questi mini gruppi individuati, confrontando con il blox pot, si può presumere che siano dovuti ai valori associati a WL.

Quanto affermato si vede ancora meglio raddoppiando il numero di bins.

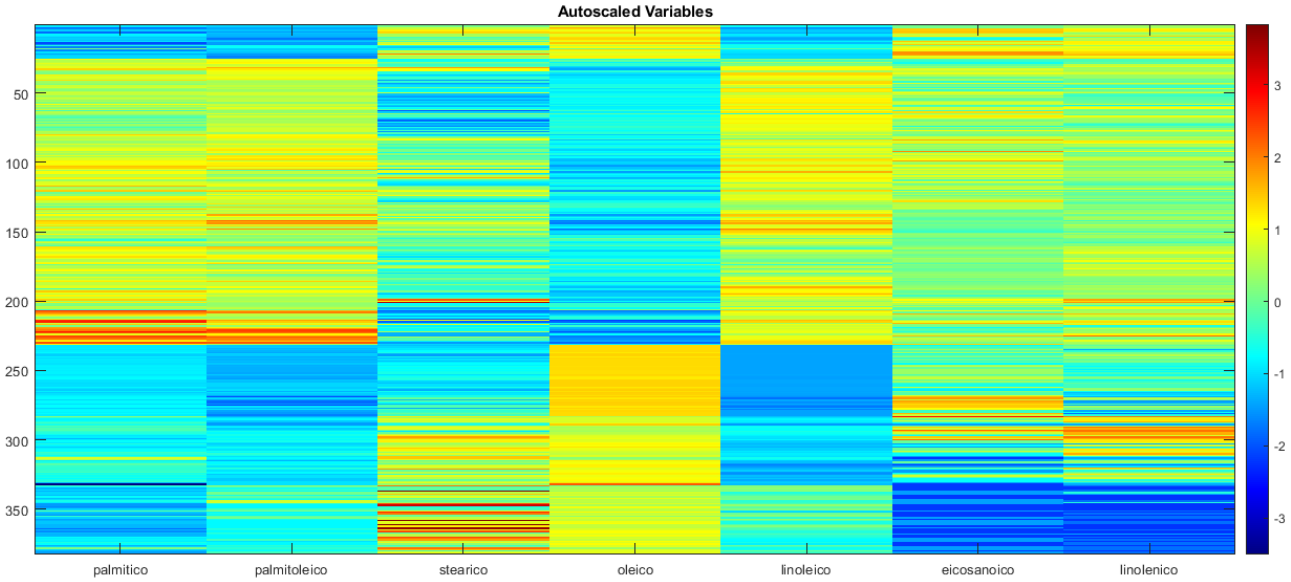


Il quinto strumento è la rappresentazione tramite immagine con bande colorate per vedere la distribuzione dei valori per ogni variabile nella sua colonna.



Come si può notare ogni colonna presenta bande di colori molto simili tra loro rendendo quasi un colore unico omogeneo.

È quindi necessario autoscalare le variabili passando da un range di 0/80+ a -3/3 indipendente per ogni colonna.



Si può vedere come ogni colonna presenta bande di colore coerenti con il box plot. L'unico olio che si può distinguere in mezzo ai campioni è il SA, gli altri non sono sufficientemente differenziabili.

Il sesto strumento è la matrice di correlazione quadrata che consiste in un'immagine molto simile alla precedente ma dove il colore delle bande (sia verticali che orizzontali) dipende dai coefficienti di correlazione per ogni coppia di campioni.

Il risultato è un rettangolo interamente rosso, perché come accennato prima i valori delle variabili risultano essere molto simili tra loro. Quindi è necessario ridurre il range da -1/1 a 0.955/1.

Da qua si posso individuare i due gruppi già noti, cioè SA e tutti gli altri oli. Inoltre si può vedere un sottogruppo composto da solo WL, che presenta un colore leggermente diverso rispetto agli altri.

Traendo delle conclusioni, i gruppi individuati nella combinazione delle varie concentrazioni sono due: SA e gli altri oli. Un aspetto interessante è che tuttavia esistono due sottogruppi, uno dei quali composto solo da WL che presenta qualche dissimilarità rimenendo tuttavia abbastanza legato.

Inoltre l'olio SA è presente con concentrazioni più alte di acido palmitico, linoleico e palmitoleico e più basse di acido oleico. Mentre è più probabile avere l'olio WL con valori più bassi di acido eicosanoico e linolenico.