Exploratory Data Analysis - olive oil dataset

Fabio Bove - 216219@studenti.unimore.it

Python libraries:

scikit-learn: *Utilizzata per misurazioni statistiche e normalizzare i dati*

seaborn / **matplotlib**: *Utilizzate per la generazione e visualizzazione di grafici dei dati*

pandas: *Utilizzata per la manipolazione del dataset e misurazioni statistiche*

** Per visualizzare tutti grafici in maniera più chiara, anche quelli non inclusi nel documento, le aggiungo il link alla cartella drive che li contiene:*
https://drive.google.com/drive/folders/1v4Lz--eu5WMPLI_DLNn9gubtN_kCTfba?usp=sharing

** Ho inoltre incluso lo script python utilizzato per la generazione del report. (Mi scuso per la mancanza di commenti, l'ho realizzato molto rapidamente)*

Dataset

Utilizzando **pandas** diamo una prima occhiata alla struttura del nostro dataset.

	category	palmitico	palmitoleico	stearico	oleico	linoleico	eicosanoico	linolenico
0	NA	10.750000	0.75	2.26	78.230011	6.720000	0.36	0.60
1	NA	10.880000	0.73	2.24	77.089996	7.810000	0.31	0.61
2	NA	9.109999	0.54	2.46	81.129997	5.490000	0.31	0.63
3	NA	9.660000	0.57	2.40	79.519997	6.190000	0.50	0.78
4	NA	10.510000	0.67	2.59	77.709999	6.720000	0.50	0.80
..	...	...	...	...	...	...	...	...
377	WL	12.800000	1.10	2.90	74.900002	7.900000	0.10	0.10
378	WL	10.600000	1.00	2.70	77.400002	8.100001	0.10	0.10
379	WL	10.100000	0.90	2.10	77.199997	9.700000	0.00	0.00
380	WL	9.899999	1.20	2.50	77.500000	8.700000	0.10	0.10
381	WL	9.600001	0.80	2.40	79.500000	7.400000	0.10	0.20

[382 rows x 8 columns]

Visualizzazione del DataFrame con i dati di "olive_oil.csv"

```
SA     206
U       51
EL      50
WL      50
NA      25
Name: category, dtype: int64
```

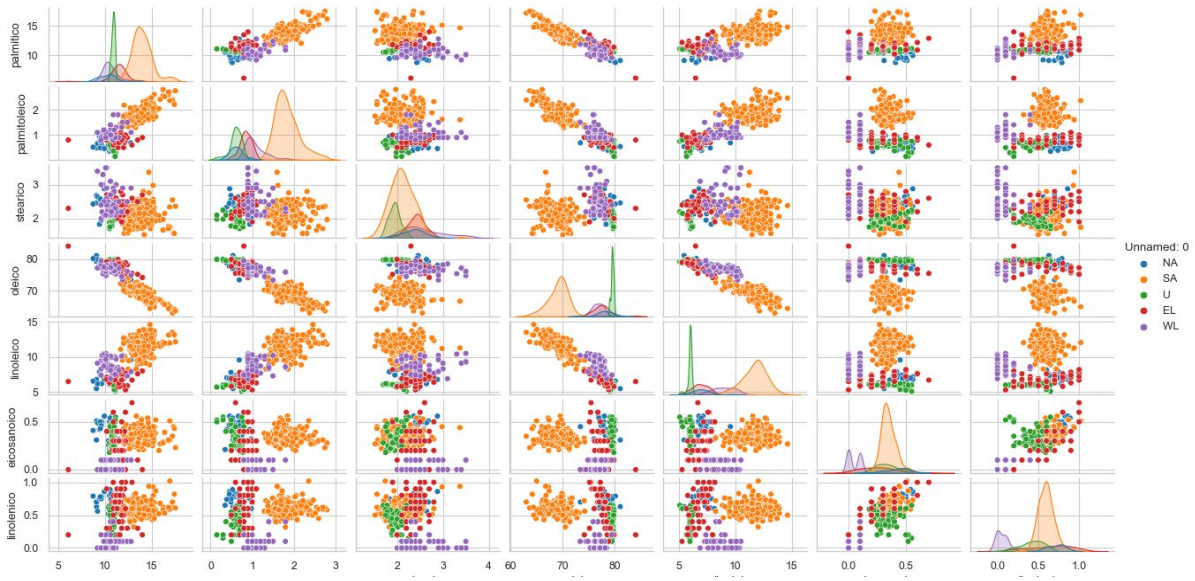
Numero di dati/righe per ogni categoria

```
category        0
palmitico       0
palmitoleico    0
stearico        0
oleico          0
linoleico       0
eicosanoico     0
linolenico      0
```

Numero di valori nulli per ogni colonna del dataset

Pairplot - Gplotmatrix

python sns.pairplot() = Matlab gplotmatrix()

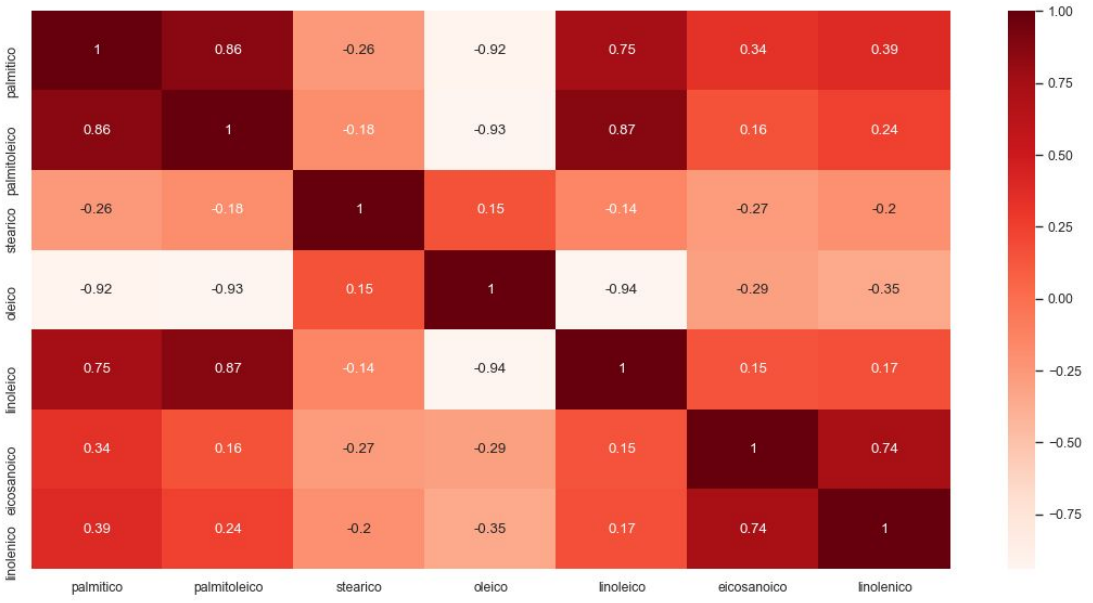


In generale i campioni d'olio con provenienza SA sono quasi sempre ben distinguibili, mentre è molto più difficile distinguere gli altri campioni (WL, EL, U, NA).

Ad una prima vista le features "*palmitoleico*", "*palmitico*" e "*stearico*" potrebbero essere quelle che meglio dividono i campioni in cluster.

Heatmap - Colormap

Essendo complicato vedere relazioni fra così tante variabili in un singolo grafico utilizzo la matrice di correlazione per identificarle.
python sns.heatmap() = Matlab colormap()



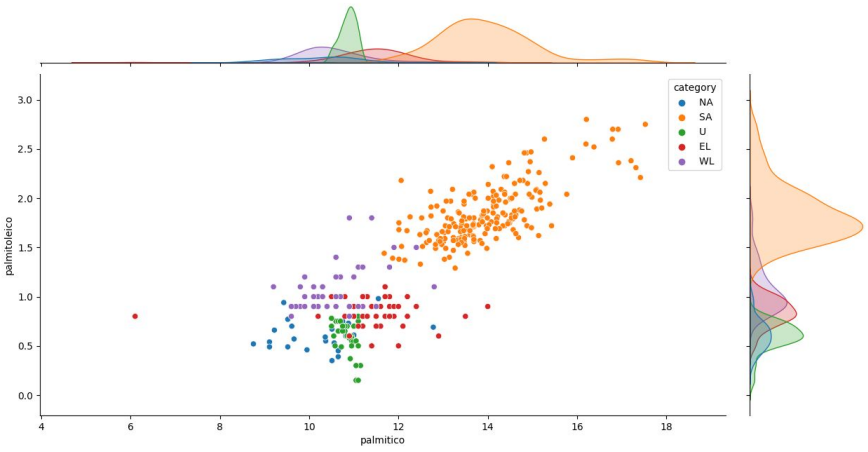
Possiamo vedere una serie di relazioni fra le seguenti coppie di variabili:

- linoleico, palmitoleico [0.86]
- palmitico, palmitoleico[0.87]
- eicosenoico, linoleico [0.74]

La variabile “stearico” sembrerebbe essere l’unica indipendente dalle altre.

Scatter Hist

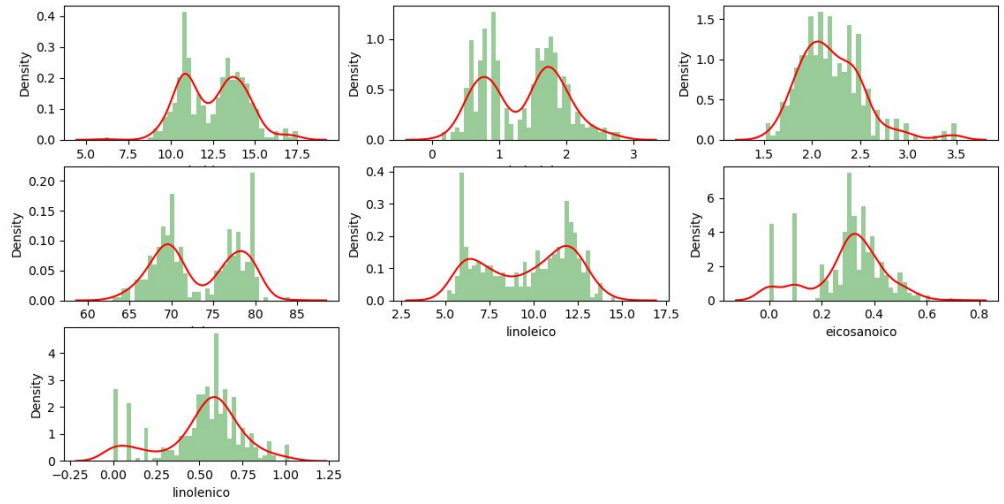
Scatter histogram fra le variabili: “palmitoleico” vs “palmitico”
python sns.jointplot() = Matlab scatterhist()



Histograms

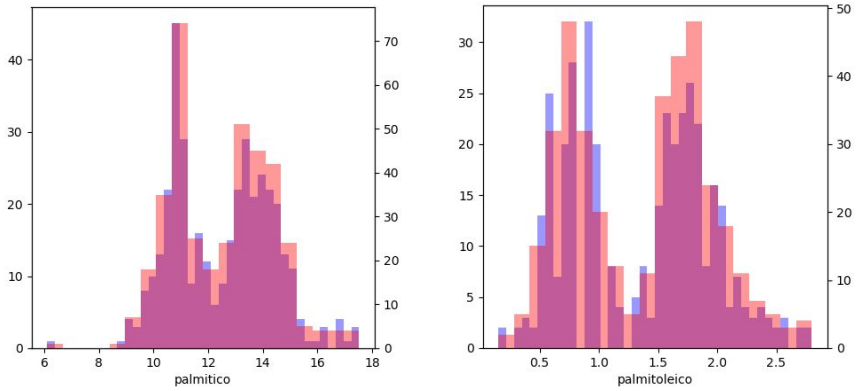
Osserviamo la distribuzione dei valori di ogni variabile per le 5 categorie utilizzando gli istogrammi.
python sns.distplot() = Matlab histfit()

Parametri importanti:
bins = 2 * int(round(math.sqrt(olive_oil.shape[0]))) # numero di righe del dataset
kde = True



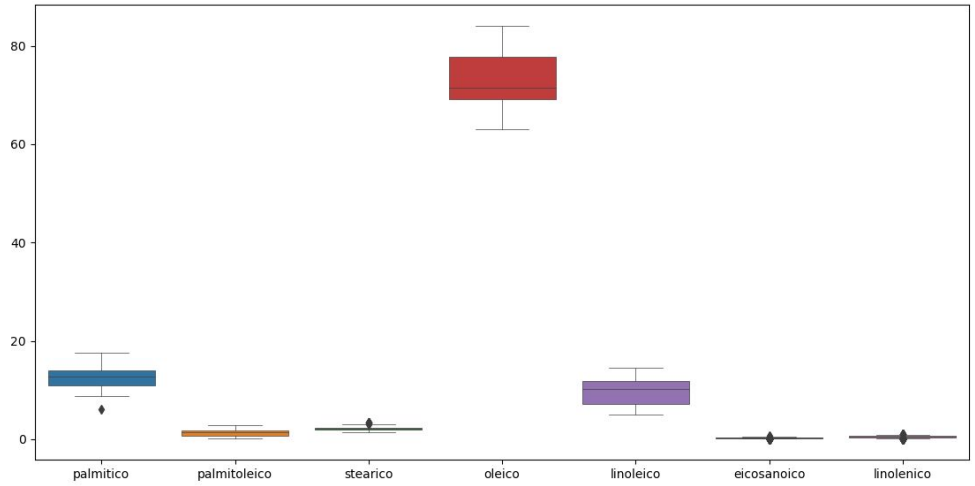
Istogrammi sovrapposti con diversi bins per la features “palmitico” e “palmitoleico”.

Nell’esempio sottostante sono stati sovrapposti gli istogrammi generati con numero di bins differente (rispettivamente 20 e 40).per la rappresentazione delle features delle variabili palmitico e palmitoleico.

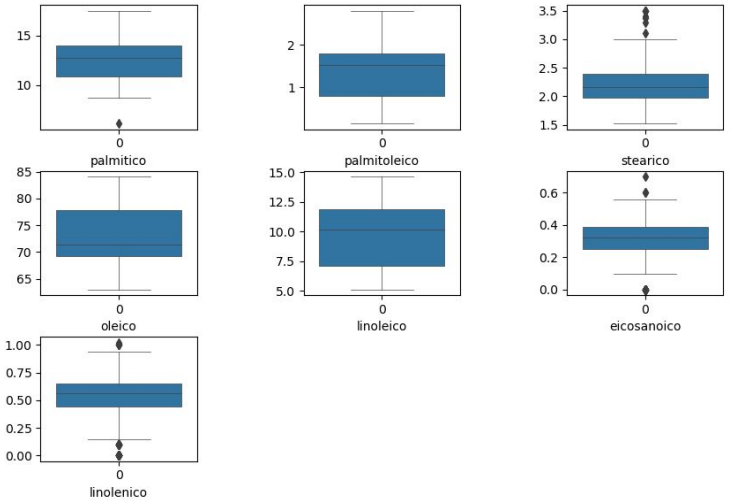


Box-plot

Per osservare le distribuzioni delle features.
python sns.boxplot() = Matlab boxplot()



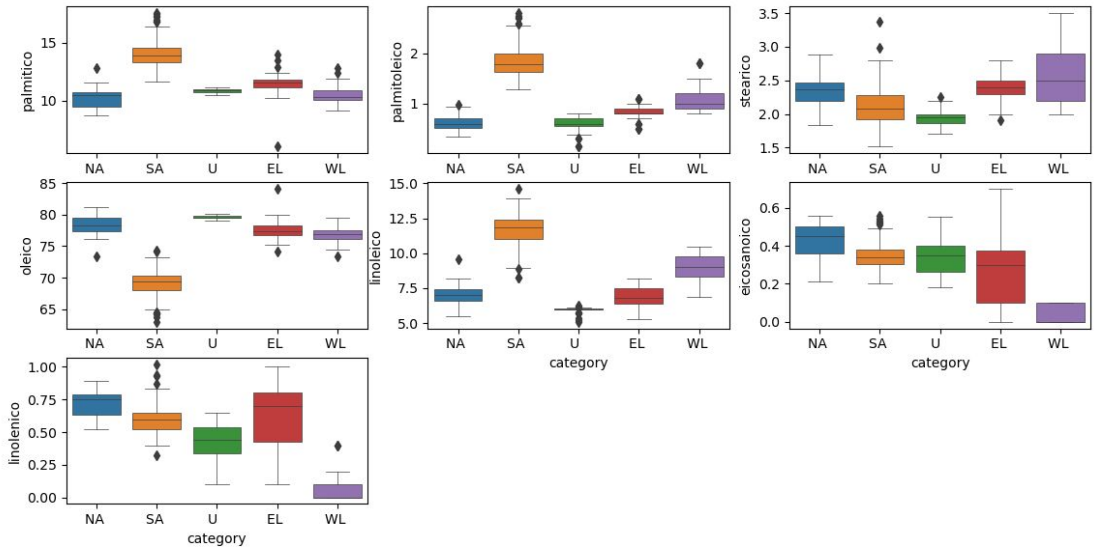
Stesso grafico di quello precedente ma separando le features per osservare meglio la loro distribuzione:



Sembrerebbe che la distribuzione dei valori delle nostre variabili fra i vari campioni sia piuttosto normale, seppur alcune features (in particolare “oleico”) hanno valori decisamente più elevati rispetto agli altri.

Possiamo inoltre identificare la presenza di outliers per le variabili linoleico, eicosenoico, stearico e palmitico.

Box-plot per la rappresentazione delle distribuzioni dei valori delle nostre variabili in ogni categoria di campioni.



Parallel Coordinates Plot

Rappresentazione dei valori di tutte le features del dataset per tutti i campioni.
python pd.parallel_coordinates() = Matlab parallelIcoor()