Analisi dei dati per il marketing (compito 13 dic 2024)

Caricare in memoria il dataset USA tramite l'istruzione

load USA

Questo dataset contiene 3 indicatori di criminalità dei 50 stati americani (Murder, Assault, Rape)

Murder = Assassini

Assault = Aggressioni

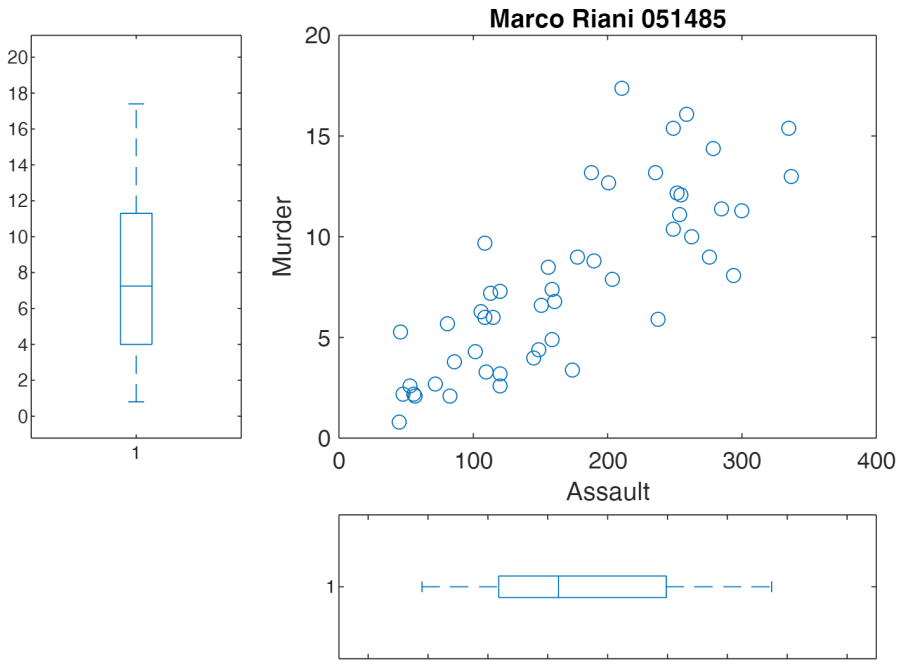
Rape = Stupri

ed un indicatore legato alla densità della popolazione urbana (UrbanPop).

Costruire il diagramma di dispersione con i boxplot ai margini tra le variabili "Assault" e"Murder" aggiungendo i titoli sull'asse x e sull'asse y

Aggiungere come titolo del grafico il proprio nome, cognome e numero di matricola (**punti 3**).

```
load USA.mat
Xtable=USA;
xlab="Assault";
ylab="Murder";
scatterboxplot(Xtable{:,xlab},Xtable{:,ylab});
xlabel(xlab); ylabel(ylab)
% Aggiungere come titolo del grafico il proprio cognome e numero di
matricola.
title('Marco Riani 051485')
```



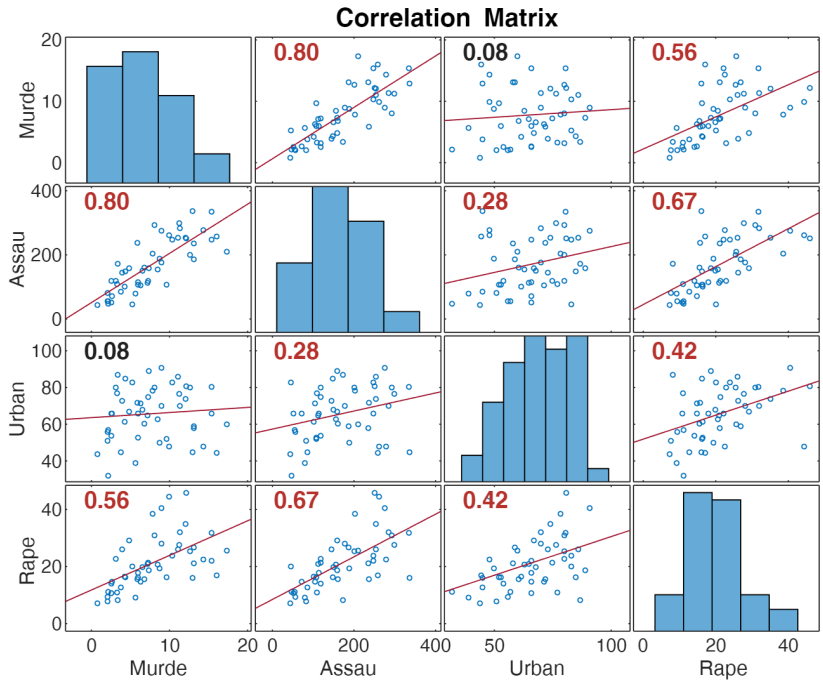
Commentare il diagramma di dispersione creato al punto precedente e la presenza di outliers univariati (**punti 2**)

```
% Forte relazione diretta tra le due variabili. Non ci sono valori anomali
% univariati
```

Costruire la matrice di correlazione e calcolare i relativi p-values in formato table .

Costruire la matrice dei diagrammi di dispersione sovrapponendo ad ogni pannello la retta di regressione. Inserire il valore del coefficiente di correlazione lineare tra ogni coppia di variabili in ogni pannello e colorarlo in rosso oppure in nero a seconda della sua significatività (**punti 5**)

```
[Rtable,Pvaltable]=corrplot(Xtable,'TestR','on');
```



Commentare la significatività delle correlazioni tra le variabili utilizzando il livello di significatività 0.05 **(punti 3)**

```
disp('p values del test di assenza di correlazione')
```

```
p values del test di assenza di correlazione
```

```
disp(Pvaltable)
```

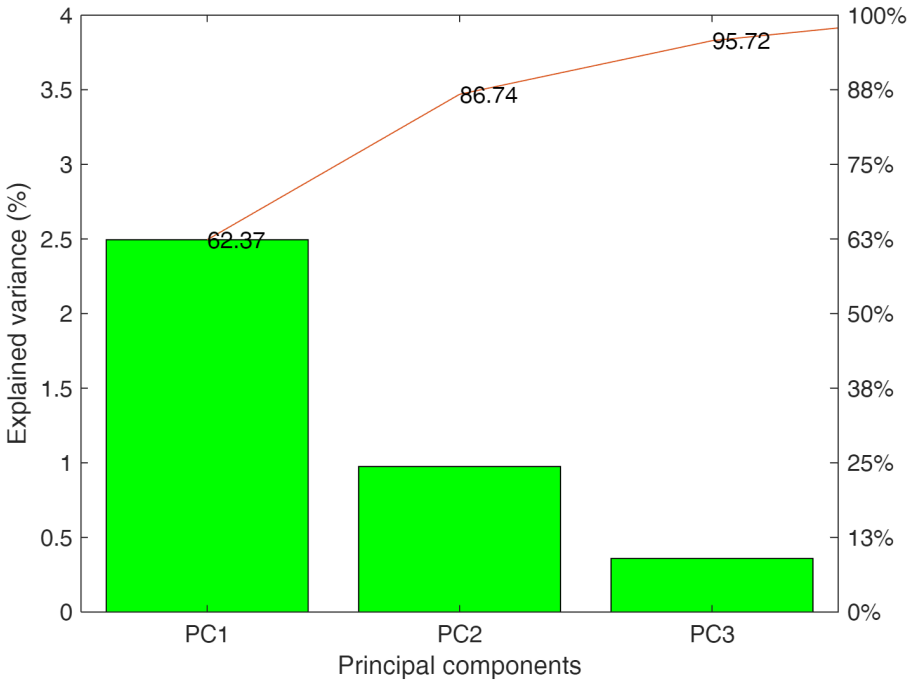
	Murder	Assault	UrbanPop	Rape
Murder	1	2.5958e-12	0.56541	2.0308e-05
Assault	2.5958e-12	1	0.047465	1.3638e-07
UrbanPop	0.56541	0.047465	1	0.0023512
Rape	2.0308e-05	1.3638e-07	0.0023512	1

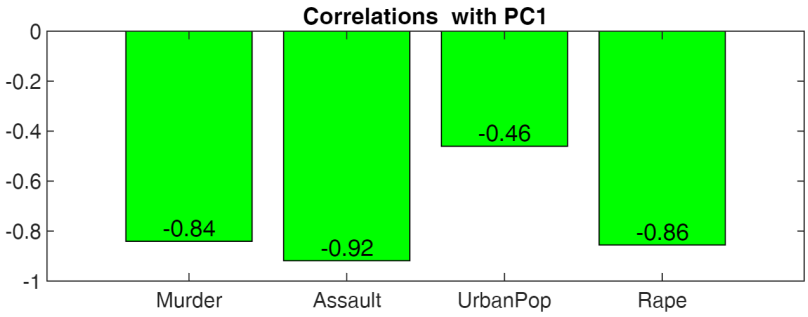
```
% Tutte le correlazioni risultano significative tranne quella tra Murder e
UrbanPop.
```

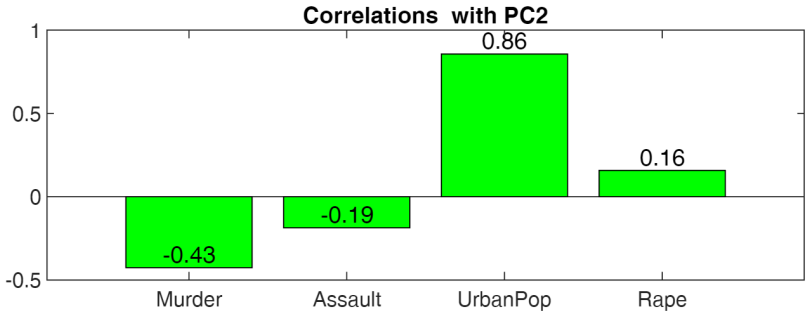
Utilizzando la tecnica delle componenti principali ridurre le dimensioni **(punti 2).**

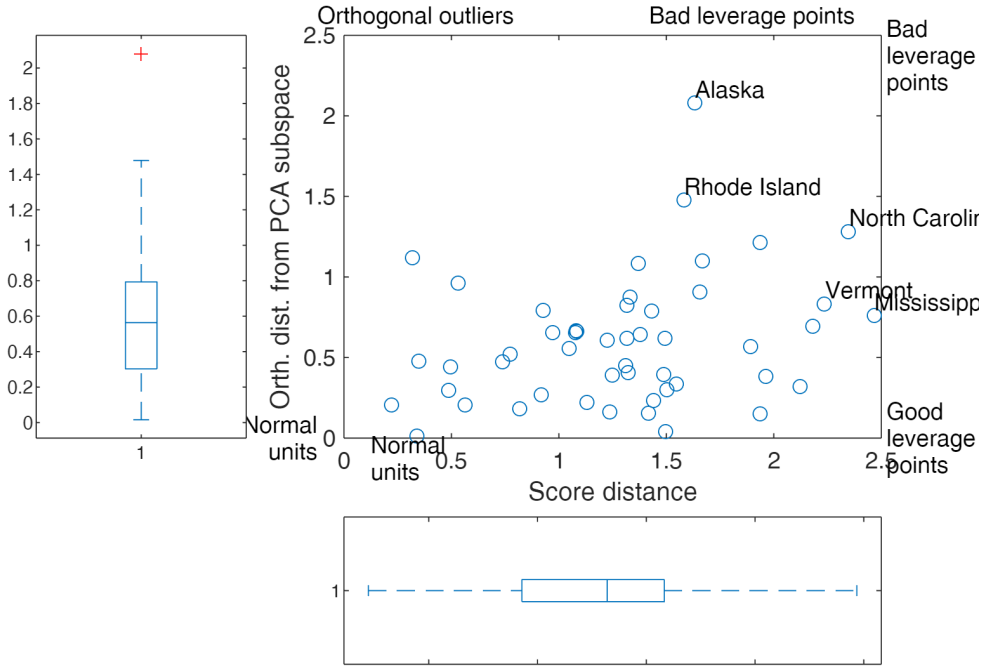
```
outPCA=pcaFS(Xtable);
```

```
Initial correlation matrix
               Murder    Assault    UrbanPop    Rape
               ------    -------    --------    ----

    Murder      1.00      0.80        0.08      0.56
    Assault     0.80      1.00        0.28      0.67
    UrbanPop    0.08      0.28        1.00      0.42
    Rape        0.56      0.67        0.42      1.00

Explained variance by PCs
           Eigenvalues    Explained_Variance    Explained_Variance_cum
           -----------    ------------------    ----------------------

    PC1       2.49              62.37                   62.37
    PC2       0.97              24.37                   86.74
    PC3       0.36               8.97                   95.72
    PC4       0.17               4.28                  100.00

Loadings = correlations between variables and PCs
                PC1      PC2
                -----    -----

    Murder      -0.84    -0.43
    Assault     -0.92    -0.19
    UrbanPop    -0.46     0.86
    Rape        -0.86     0.16

Communalities
               PC1     PC2     PC1-PC2
               ----    ----    -------

    Murder     0.71    0.18     0.89
    Assault    0.84    0.03     0.88
    UrbanPop   0.21    0.73     0.95
    Rape       0.73    0.02     0.76

Units with the 5 largest values of (combined) score and orthogonal distance
     2    33    24    45    39
```

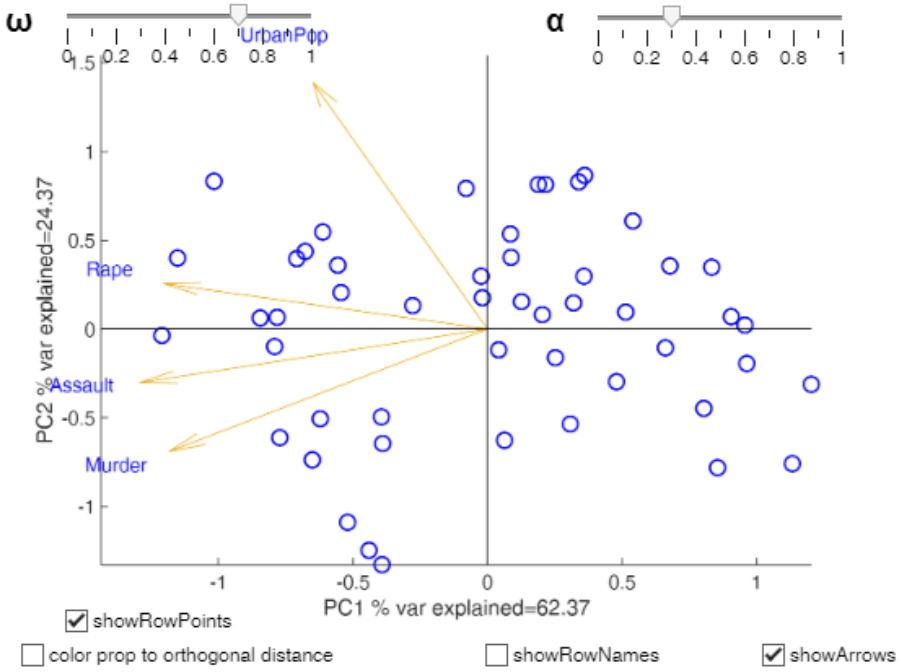

Explained variance (%)
Principal components
PC1
PC2
PC3
62.37
86.74
95.72
0%
13%
25%
38%
50%
63%
75%
88%
100%


Correlations with PC1
-0.84
-0.92
-0.46
-0.86
Murder
Assault
UrbanPop
Rape


Correlations with PC2
-0.43
-0.19
0.86
0.16
Murder
Assault
UrbanPop
Rape



$$\textit{Row points} : (\sqrt{n-1})^{\omega} U \Gamma^{\alpha} \qquad \textit{Arrow coordinates} : (\sqrt{n-1})^{1-\omega} V \Gamma^{1-\alpha}$$



Commentare la varianza spiegata dalle diverse componenti e il numero di componenti da tenere utilizzando i primi due criteri riportati a p. 449 del libro di testo **(punti 3)**

```
% Secondo il primo criterio le CP devono spiegaare almeno il 70 per cento
% di variabilità totale. Dato che le prime due componenti spiegano circa
l'87 per
% cento della varianza totale è lecito tenere solo le prime 2 componenti
```

```
% Secondo il criterio 0.95^p occorre spiegare almeno il 0.95^4=0.8145 di
% variabilità totale. Dato che le prime due componenti spiegano circa l'87
per
% cento della varianza totale è lecito tenere solo le prime 2 componenti
```

Commentare il significato della prima componente principale **(punti 3)**

```
% La prima componente principale è un indicatore di "Sicurezza". Tanto più
% ci si sposta da sinistra verso destra tanto più Murder, Assault e Rape
% diminuiscono. Le città più sicure sono caratterizzate da valori bassi di
% UrbanPop
```

Commentare la posizione dello Stato della Florida nello spazio delle prime due componenti principali **(punti 3).**

```
% La Florida presenta il valore più basso della prima PC.
% E' nella direzione delle variabili Rape e Assault, di conseguenza
% presenta valori molto elevati di questi due indicatori.
```

Commentare l'angolo nel biplot tra la variabile UrbanPop e Murder e quello tra Assault e Murder **(punti 3)**

```
% L'angolo tra UrbanPop e Murder è vicino a 90 gradi e sta ad indicare
l'assenza di correlazione
% tra queste due variabili. Questo conferma quello che avevamo visto in
% precedenza (correlazione non significativa).
% L'angolo tra Assault e Murder è molto basso e sta ad indicare una forte
% correlazione diretta tra le due variabili
```

Estrarre le 3 unità che presentano la distanza ortogonale più grande (orthDist).

Mostrare i record associati a questre 3 unità nella Command Window **(punti 3)**

```
[~,indsor]=sort(outPCA.orthDist,'descend');
disp(Xtable(indsor(1:3),:))
```

	Murder	Assault	UrbanPop	Rape
Alaska	10	263	48	44.5
Rhode Island	3.4	174	87	8.3
North Carolina	13	337	45	16.1