

EUROPA: SCIENZA, TECNOLOGIA E INNOVAZIONE

Analisi statistica



Corso di Metodi e Tecniche per l'Analisi dei Dati

Docente: Prof.ssa A.G Nobile

Studente: Gino Farisano

1.EUROPA: SCIENZA, TECNOLOGIA E INNOVAZIONE	5
1.1 Introduzione	5
2. DATASET UTILIZZATO	5
3. ANALISI STATISTICA	7
3.1 Creazione del Data Frame	7
3.2 Analisi statistica mediante rappresentazione grafica	9
3.2.1 Distribuzione di frequenza della spesa totale per ricerca e sviluppo (% rispetto al PIL)	10
3.2.2 Distribuzione di frequenza del numero di brevetti rilasciati (per milione di abitanti).	12
3.2.3 Distribuzione di frequenza degli addetti nel settore Ricerca e Sviluppo nei paesi dell'UE (% per 1000 abitanti)	13
3.2.4 Distribuzione di frequenza delle imprese innovatrici nei paesi dell'UE	15
3.2.5 Distribuzione di frequenza dei laureati in discipline tecnico-scientifiche nei paesi dell'UE	16
3.3 Confronto tra Europa settentrionale, occidentale, orientale e meridionale	19
3.3.1 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - spesa totale per ricerca e sviluppo	21
3.3.2 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - numero di brevetti	22
3.3.3 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - addetti alla ricerca e sviluppo	23
3.3.4 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - imprese innovatrici	24
3.3.5 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - laureati in discipline tecnico-scientifiche	25
3.4 Minimo e massimo per colonne	26

3.4.1 Minimo e massimo nazioni	27
3.4.2 Minimo e massimo macroregioni	28
3.5 Visione globale dei migliore e dei peggiori	28
3.5.1 Minimi	29
3.5.2 Massimi e minimi a confronto	30
3.5.3 Massimi macroregioni	31
3.5.4 Visione globale macroregioni	31
4. Indici di posizione e dispersione	32
4.1 Boxplot relativo alla distribuzione della spesa nel settore Ricerca e Sviluppo dei paesi dell'UE	35
4.2 Boxplot relativo alla distribuzione dei brevetti nei paesi dell'UE	36
4.3 Boxplot relativo alla distribuzione degli addetti alla ricerca ricerca e sviluppo dei paesi della UE	37
4.5 Boxplot relativo alla distribuzione dei laureati in discipline tecniche-scientifiche dei paesi dell'UE.	39
5.Correlazione tra variabili	40
5.1 Covarianza	40
5.2. Coefficiente di correlazione	40
5.3. Scatterplot	42
5.3.1 Scatterplot relativo alla spesa per il settore Ricerca e Sviluppo e al numero di addetti nel settore Ricerca e Sviluppo.	42
6.Analisi dei cluster	44
6.1. Metodi gerarchici	47
6.1.1 Metodo del legame singolo	48
6.1.2 Metodo del legame complete	49
6.1.3. Metodo del legame medio	50
6.1.4. Metodo del centroide	51

6.1.5 Metodo della mediana	52
6.2 Screeplot metodo legame complete	55
6.2.1 Metodi non gerarchici	56
6.2.2. Scelta casuale dei punti di riferimento	57
6.2.3. Scelta dei centroidi come punti di riferimento	58

1.EUROPA: SCIENZA, TECNOLOGIA E INNOVAZIONE

1.1 *Introduzione*

L'attività di ricerca e l'accesso alle tecnologie dell'informazione sono riconosciuti come motori fondamentali dell'economia della conoscenza e assumono un ruolo basilare nelle strategie di sviluppo europee. Gli indicatori che misurano questi fenomeni riguardano sia l'input sia l'output delle attività innovative e contribuiscono a migliorare la comprensione del livello di progresso di un paese. In questo elaborato a tal proposito, a partire dai dati Istat degli stati dell'UE, realizzeremo un'analisi statistica al fine di comprendere quanto i diversi paesi dell'area euro "spendono" in ricerca e sviluppo. Ad esempio nel nostro Paese la spesa per ricerca e sviluppo incide per l'1,26 per cento del Pil (2012). Tale valore non è lontano dall'obiettivo Europa 2020 fissato per l'Italia (1,53 per cento), ma è ancora distante dall'obiettivo comune dei paesi Ue, fissato al 3 per cento e superato solo dai paesi scandinavi.

Per l'analisi dei dati utilizzeremo R (<http://www.r-project.org>), un software open source e multipiattaforma che fornisce un'ambiente dedicato per l'elaborazione statistica e grafica.

2. DATASET UTILIZZATO

Vi sono diversi indicatori che permettono di comprendere le politiche di ricerca e sviluppo di un dato paese. In questo lavoro abbiamo deciso di utilizzarne cinque, in particolare:

- 1.** Spesa totale per ricerca e sviluppo (% rispetto al PIL): il conseguimento di un adeguato rapporto tra spesa per ricerca e sviluppo (R&S) e Pil è uno dei cinque obiettivi cardine stabiliti nell'ambito della strategia "Europa 2020" per accrescere i livelli di produttività, di occupazione e di benessere sociale.
- 2.** Numero di brevetti (in media per milione di abitanti): uno dei principali indicatori di output con cui viene misurata l'attività innovativa di un paese è dato dal numero di brevetti registrati. Questi vengono desunti da fonti amministrative o grazie alla presenza di uffici internazionali di brevetti, quali l'ufficio europeo dei brevetti (European Patent Office, Epo).
- 3.** Addetti alla ricerca e sviluppo (% per 1000 abitanti): per valutare l'apporto delle risorse umane all'economia della conoscenza si fa riferimento al numero di addetti impegnati nelle attività di ricerca e sviluppo (R&S). Considerati in rapporto all'occupazione, alla popolazione attiva o a quella residente, forniscono un

indicatore dell'intensità dell'attività scientifica e tecnologica di un paese in termini di risorse umane utilizzate.

- 4.** Imprese innovative (%): sebbene l'innovazione sia un fenomeno complesso e ancora poco indagato nelle sue relazioni con la crescita economica e l'occupazione, essa rappresenta un obiettivo comune delle politiche di sviluppo economico nazionali ed europee.
- 5.** Laureati in discipline tecnico-scientifiche (%), età 20-29 anni): la quota di giovani che hanno conseguito un titolo accademico nell'area S&T (Science and Technology) rappresenta una buona approssimazione del flusso annuale di persone altamente qualificate, potenzialmente disponibili a operare nel campo della ricerca e sviluppo. Uno scarso numero di laureati in S&T si traduce, per i paesi, in una perdita complessiva di competitività internazionale nel campo dell'alta tecnologia perché rende difficile il reclutamento di ricercatori e tecnici ad alta qualificazione scientifica da parte delle imprese.

Di seguito viene riportato lo screenshot del dataset utilizzato:

Nazioni	Spesa totale per ricerca e sviluppo (% rispetto al PIL)	Numero di brevetti	Addetti alla ricerca e sviluppo (% per 1000 abitanti)	Imprese innovative (%)	Laureati in discipline tecnico-scientifiche (%)
Austria	2,81	208	7,4	39,3	16,4
Belgio	2,24	138	5,8	46,5	13
Bulgaria	0,62	2,3	2,3	16,9	13,3
Cipro	0,43	9,4	1,4	29,8	9
Croazia	0,75	6,8	2,4	25	17,4
Danimarca	3,03	226,7	10,5	38,1	18,8
Estonia	2,16	28,1	4,4	38,4	13,2
Finlandia	3,43	255,8	10	44,6	21,7
Francia	2,23	130	6,3	36,7	22,1
Germania	2,88	282,6	7,3	55	16,2
Grecia	0,69	5,8	3,4	34,3	13,9
Irlanda	1,58	68,5	4,9	42,3	22,5
Italia	1,26	74,9	4	41,5	13,2
Lettonia	0,66	7,4	2,7	19,5	13,5
Lituania	0,9	5,1	3,5	18,9	23
Lussemburgo	1,16	150,2	9,3	48,4	2,8
Malta	0,87	8,5	3,5	35,9	11,1
Paesi Bassi	1,97	181,6	7,3	44,5	10,7
Polonia	0,89	9,3	2,4	16,1	17,9
Portogallo	1,37	8,9	4,5	41,3	19,4
Regno Unito	1,63	83,5	5,6	34	19,8
Repubblica Ceca	1,79	18,3	5,7	35,6	16,7
Romania	0,48	1,7	1,5	6,3	18,7
Slovacchia	0,81	8,4	3,4	19,7	17,9
Slovenia	2,58	50,8	7,7	32,7	17,9
Spagna	1,27	32	4,5	23,2	15,6
Svezia	3,28	294,9	8,6	45,2	15,9
Ungheria	1,27	19,2	3,6	16,4	9,5

3. ANALISI STATISTICA

L'attività di analisi svolta si è articolata nel calcolo e nella valutazione di:

- Distribuzioni di frequenza
- Indici di posizione e dispersione
- Correlazione tra variabili
- Cluster

Innanzitutto, prima di procedere all'importazione del file “csv” nell'ambiente R, per una questione di leggibilità, si è ritenuto opportuno rinominare le colonne del dataset nel modo seguente:

1. STRS: spesa totale per ricerca e sviluppo (% rispetto al PIL).
2. NB: numero di brevetti (per milione di abitanti).
3. ARS: addetti alla ricerca e sviluppo (% per 1000 abitanti).
4. II: imprese innovative (%).
5. LDTS: laureati in discipline tecnico-scientifiche (%).

3.1 Creazione del Data Frame

I dati oggetto dell'analisi sono stati importati nell'ambiente R utilizzando il seguente comando:

In particolare, `read.csv()` permette di leggere un file in formato “csv” creando un data frame.

```
> mytable=read.csv(file="MTADDatasetCSVAbbreviazione.csv",header=TRUE,sep=";",dec=",")
```

Oltre al nome del file è stato necessario utilizzare “header=TRUE” per indicare che la prima riga del file contiene i nomi delle variabili, `sep=";"` per indicare il carattere separatore tra i campi e `dec=","` per indicare il carattere utilizzato come separatore per i decimali.

Di seguito viene mostrato il data frame ottenuto:

	Nazioni	STRS	NB	ARS	II	L
1	Austria	2.81	208.0	7.4	39.3	16.4
2	Belgio	2.24	138.0	5.8	46.5	13.0
3	Bulgaria	0.62	2.3	2.3	16.9	13.3
4	Cipro	0.43	9.4	1.4	29.8	9.0
5	Croazia	0.75	6.8	2.4	25.0	17.4
6	Danimarca	3.03	226.7	10.5	38.1	18.8
7	Estonia	2.16	28.1	4.4	38.4	13.2
8	Finlandia	3.43	255.8	10.0	44.6	21.7
9	Francia	2.23	130.0	6.3	36.7	22.1
10	Germania	2.88	282.6	7.3	55.0	16.2
11	Grecia	0.69	5.8	3.4	34.3	13.9
12	Irlanda	1.58	68.5	4.9	42.3	22.5
13	Italia	1.26	74.9	4.0	41.5	13.2
14	Lettonia	0.66	7.4	2.7	19.5	13.5
15	Lituania	0.90	5.1	3.5	18.9	23.0
16	Lussemburgo	1.16	150.2	9.3	48.4	2.8
17	Malta	0.87	8.5	3.5	35.9	11.1
18	Paesi Bassi	1.97	181.6	7.3	44.5	10.7
19	Polonia	0.89	9.3	2.4	16.1	17.9
20	Portogallo	1.37	8.9	4.5	41.3	19.4
21	Regno Unito	1.63	83.5	5.6	34.0	19.8
22	Repubblica Ceca	1.79	18.3	5.7	35.6	16.7
23	Romania	0.48	1.7	1.5	6.3	18.7
24	Slovacchia	0.81	8.4	3.4	19.7	17.9
25	Slovenia	2.58	50.8	7.7	32.7	17.9
26	Spagna	1.27	32.0	4.5	23.2	15.6
27	Svezia	3.28	294.9	8.6	45.2	15.9
28	Ungheria	1.27	19.2	3.6	16.4	9.5

In particolare, in R un data frame è un oggetto di tipo lista che si presenta in forma di tabella (matrice di dati) ed è costituito da righe e colonne; ogni riga del data frame individua un'osservazione e ad ogni colonna corrisponde una variabile. Come per le matrici, le colonne di un data frame hanno tutte la stessa lunghezza e i valori contenuti in ogni singola colonna sono omogenei (numeri, caratteri, valori logici, etc). Invece, a differenza delle matrici, un data frame può avere colonne di tipo diverso. Le righe e le colonne possono avere delle etichette costituite da nomi o da valori numerici. Nella fattispecie, il nostro data frame presenta come etichetta delle righe un numero quindi, per una questione di comodità si è ritenuto opportuno rinominare quest'ultime con il nome delle nazioni; a tal proposito la colonna “Nazioni” è stata eliminata.

Di seguito i comandi utilizzati e il risultato ottenuto:

```

> rownames(mytable)<-c("AU","BE","BU","CI","CR","DA","ES","FI","FR","GE","GR","IR","IT","LE","LI","LU","MA","PA","PL","PR","RU","RC","RO","SC","SV","SP","SZ","UN")
> mytable$Nazioni <- NULL
> mytable
   STRS    NB   ARS   II    L
AU 2.81 208.0 7.4 39.3 16.4
BE 2.24 138.0 5.8 46.5 13.0
BU 0.62  2.3 2.3 16.9 13.3
CI 0.43  9.4 1.4 29.8  9.0
CR 0.75  6.8 2.4 25.0 17.4
DA 3.03 226.7 10.5 38.1 18.8
ES 2.16 28.1  4.4 38.4 13.2
FI 3.43 255.8 10.0 44.6 21.7
FR 2.23 130.0 6.3 36.7 22.1
GE 2.88 282.6 7.3 55.0 16.2
GR 0.69  5.8 3.4 34.3 13.9
IR 1.58 68.5 4.9 42.3 22.5
IT 1.26 74.9  4.0 41.5 13.2
LE 0.66  7.4 2.7 19.5 13.5
LI 0.90  5.1 3.5 18.9 23.0
LU 1.16 150.2 9.3 48.4  2.8
MA 0.87  8.5 3.5 35.9 11.1
PA 1.97 181.6 7.3 44.5 10.7
PL 0.89  9.3 2.4 16.1 17.9
PR 1.37  8.9 4.5 41.3 19.4
RU 1.63 83.5 5.6 34.0 19.8
RC 1.79 18.3 5.7 35.6 16.7
RO 0.48  1.7 1.5 6.3 18.7
SC 0.81  8.4 3.4 19.7 17.9
SV 2.58 50.8  7.7 32.7 17.9
SP 1.27 32.0 4.5 23.2 15.6
SZ 3.28 294.9 8.6 45.2 15.9
UN 1.27 19.2 3.6 16.4  9.5

```

3.2 Analisi statistica mediante rappresentazione grafica

In questo paragrafo vedremo come sia possibile in R costruire tabelle di frequenza e grafici di alta qualità. Il sistema R è anche dotato di un sofisticato ambiente grafico che permette di creare grafici per illustrare i risultati di elaborazioni statistiche. Con R è possibile impostare parametri per modificare ogni aspetto di un grafico, simboli, colori ed esportare nei più comuni formati, sia vettoriali (quali .ps, .pdf) che bitmap (.bmp, .jpg). Per avere un'idea delle potenzialità offerte dall'ambiente grafico di R, basta visualizzare la dimostrazione riassuntiva che può essere ottenuta digitando il comando `demo(graphics)` e digitando ripetutamente il tasto “Invio” per procedere alla visualizzazione delle immagini in successione. In particolare, in questa sezione mostreremo il calcolo della distribuzione di frequenza semplice; a tal proposito introduciamo in prima istanza i concetti teorici utili alla comprensione delle analisi effettuate:

Sia X una variabile e $x_1, x_2, x_3, \dots, x_k$ le distinte modalità che essa assume e consideriamo un insieme n di osservazioni della variabile X. Se indichiamo con n_i il numero di volte in cui ogni valore x_i compare nel campione, ovvero la frequenza assoluta con cui esso appare nel campione, l'insieme $\{(x_i, n_i), i = 1, 2, \dots, k\}$ prende il nome di distribuzione di frequenza. Se non ci sono dati mancanti, la somma delle frequenze assolute risulta essere uguale al numero di osservazioni effettuate, ovvero $n = n_1, n_2, \dots, n_k$. La frequenza relativa invece risulta essere il rapporto tra la frequenza assoluta e la numerosità del campione ovvero n_i/n . Se i dati non mancano la somma delle frequenze relative è sempre unitaria, ossia $f_1 + f_2 + \dots + f_k = 1$. Nei seguenti sotto paragrafi verranno analizzate le frequenze assolute e relative dei dati

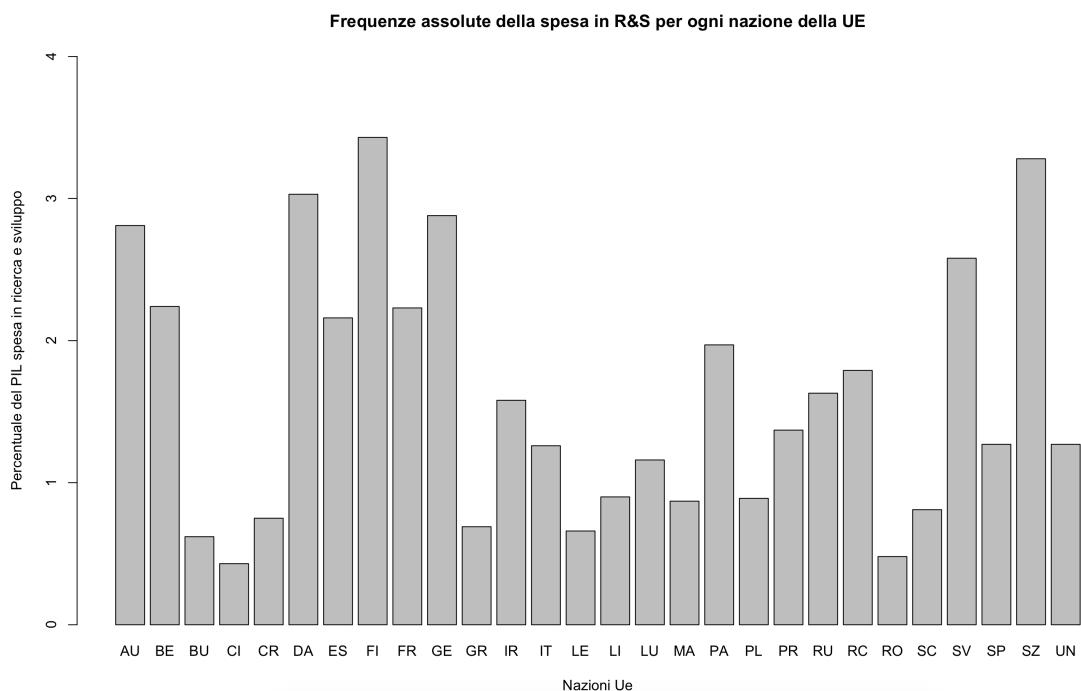
attraverso l'utilizzo di diversi tipi di grafico come ad esempio il grafico a torta e quello a barre.

3.2.1 Distribuzione di frequenza della spesa totale per ricerca e sviluppo (% rispetto al PIL)

Come si evince dalla tabella “mytable” la spesa nel settore ricerca e sviluppo per ogni paese è espressa come percentuale sul PIL. Utilizziamo quindi, per le frequenze assolute un grafico a barre mentre per le frequenze relative (come denominatore utilizzeremo la somma delle percentuali) un grafico a bastoncini.

Di seguito i comandi utilizzati e i rispettivi grafici ottenuti:

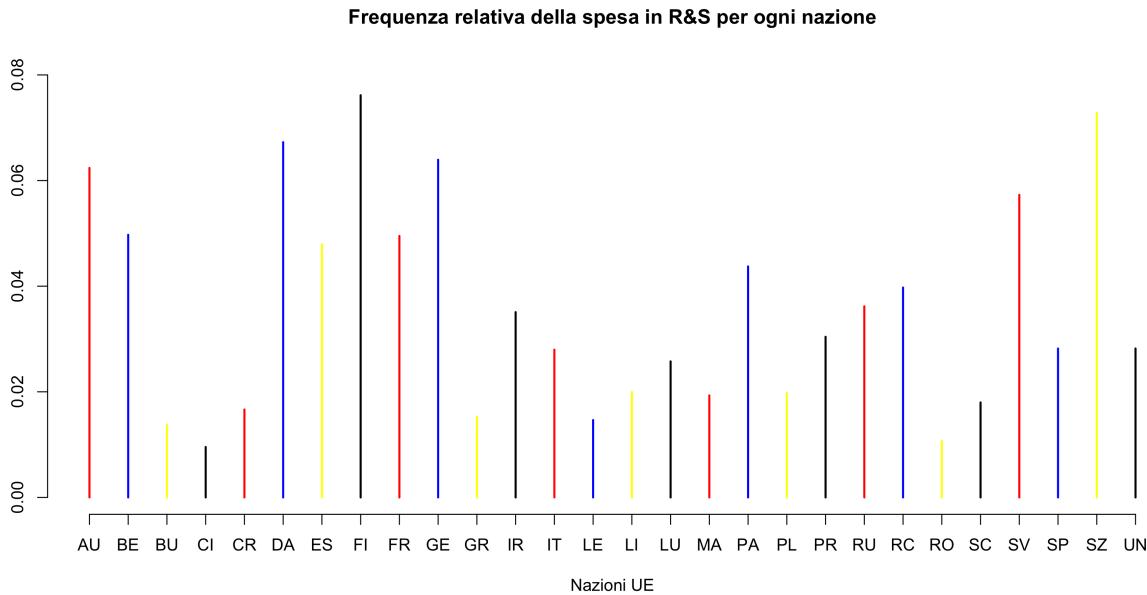
```
> barplot(mytable[,1],names.arg=rownames(mytable),ylim=c(0,4),xlab="Nazioni Ue",ylab="Percentuale del PIL spesa in ricerca e sviluppo",main="Frequenze assolute della spesa in R&S per ogni nazione della UE")
```



```

> totPercentuale <- sum(mytable[,1])
> freqrel <- mytable[,1]/totPercentuale
> freqrel
 [1] 0.062388988 0.049733570 0.013765542 0.009547069 0.016651865 0.067273535 0.047957371 0.076154529 0.049511545 0.063943162
[11] 0.015319716 0.035079929 0.027975133 0.014653641 0.019982238 0.025754885 0.019316163 0.043738899 0.019760213 0.030417407
[21] 0.036190053 0.039742451 0.010657194 0.017984014 0.057282416 0.028197158 0.072824156 0.028197158
> sum(freqrel)
[1] 1
> plot(as.table(freqrel),main="Frequenza relativa della spesa in R&S per ogni nazione",ylim=c(0,0.08),xlab="Nazioni UE",ylab=
",xaxt="n",col=c("red","blue","yellow","black"))
> axis(1,at=1:28,labels=rownames(mytable))

```



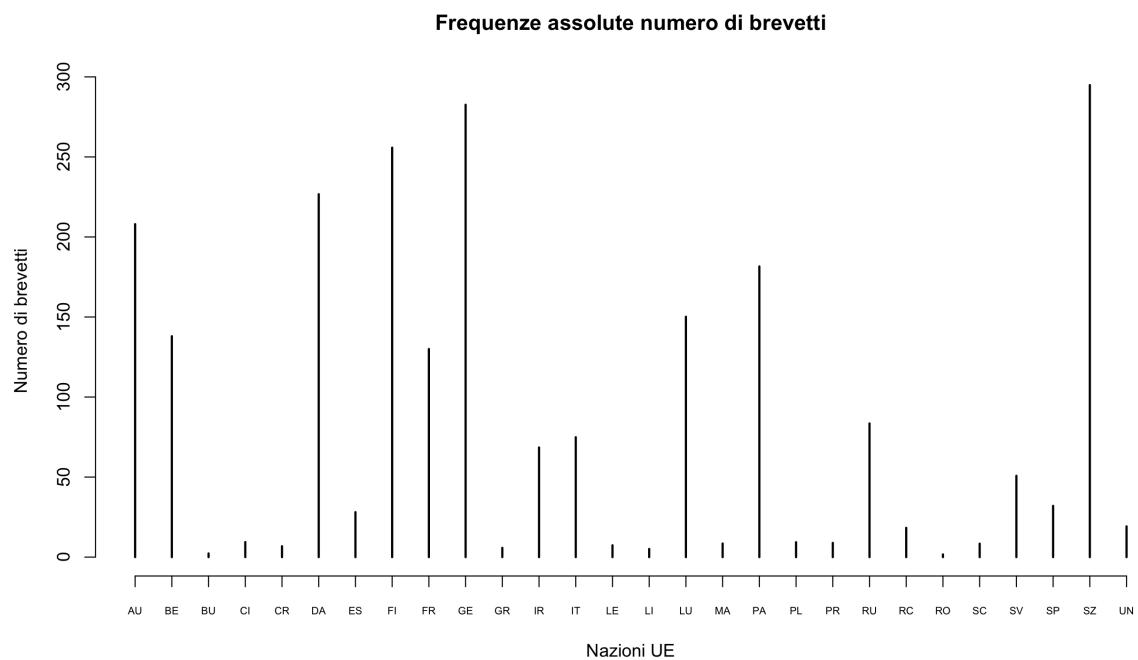
Dai grafici si evince come i paesi maggiormente propensi ad investire in ricerca e sviluppo siano quelli scandinavi (Finlandia, Svezia e Danimarca), gli unici a superare la soglia del 3 per cento del PIL, fissato come obiettivo comune dei paesi UE. Germania e Austria investono rispettivamente il 2,88 e il 2,81 per cento del Pil, ben al di sopra di Francia (2,23 per cento) e Regno Unito (1,63 per cento). Come è noto, i bilanci fortemente positivi di questi paesi sono determinati dal numero di imprese operanti in settori a forte intensità di R&S (Svezia: industria farmaceutica, automobilistica e delle apparecchiature delle comunicazioni; Finlandia: apparecchiature delle telecomunicazioni; Germania: veicoli a motore; Danimarca: industria farmaceutica/bio-tecnologie e servizi ICT). L'Italia è stabilmente al di sotto del Portogallo (1,37 per cento) e della Spagna (1,27 per cento), paesi che nell'ultimo anno hanno visto scendere loro intensità di R&S sul Pil. All'ultimo posto troviamo, invece, Cipro.

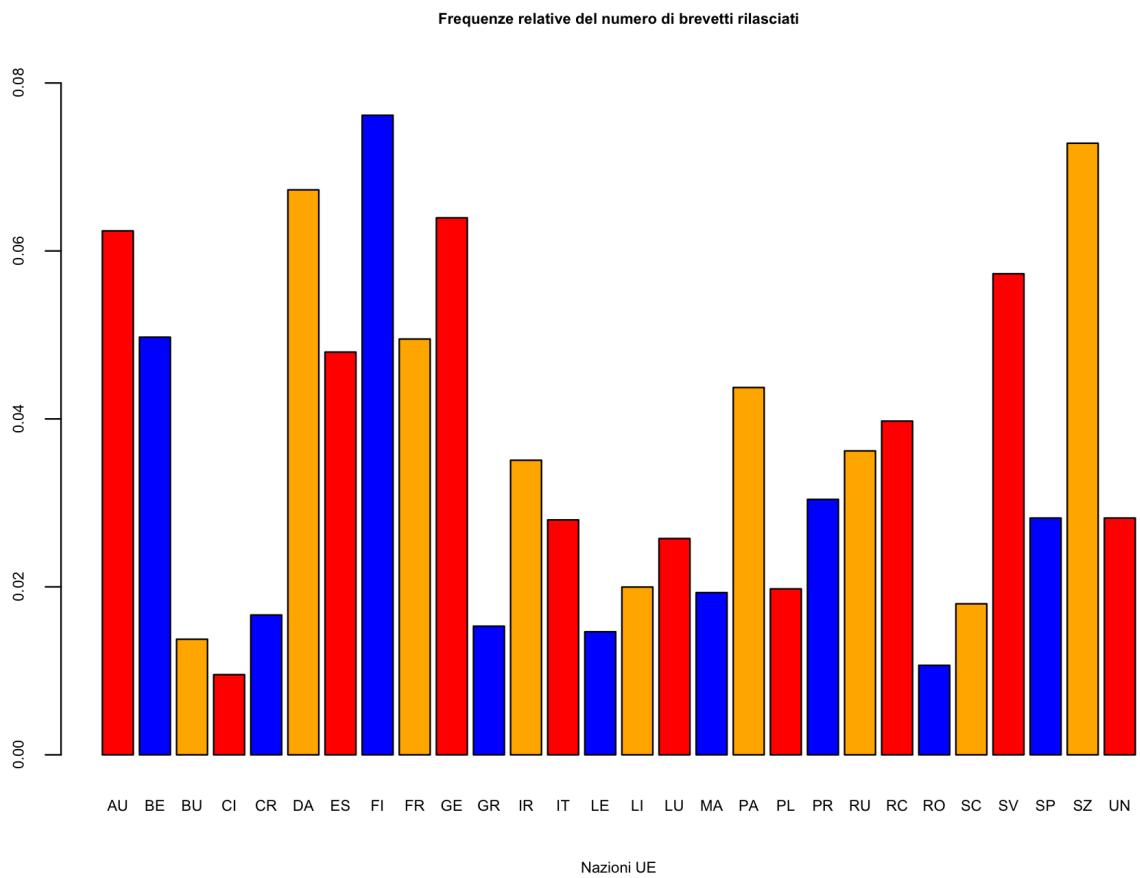
3.2.2 Distribuzione di frequenza del numero di brevetti rilasciati (per milione di abitanti).

Per quanto riguarda il numero di brevetti rilasciati i valori numerici riportati nella tabella “mytable” non sono espressi in percentuale come per la colonna STRS analizzata nel paragrafo precedente. Tuttavia procediamo allo stesso modo per calcolare le frequenze assolute e relative.

Di seguito i frammenti di codice utilizzati e i rispettivi grafici

```
> plot(as.table(mytable[,2]),xlab="Nazioni UE",ylab="Numero di brevetti",main="Frequenze assolute numero di brevetti",xaxt="n")
> axis(1,at=1:28,labels=rownames(mytable),par(ps = 7, cex = 1, cex.main = 1))
```





L'indice di intensità brevettuale mostra una variabilità elevatissima (da 1,7 della Romania a 294,9 della Svezia) da una parte con un gruppo di paesi nord europei tutti al disopra della media Ue28 e dall'altra un variegato gruppo di nazioni formato dai paesi mediterranei e da quelli di recente ingresso nell'Unione che presentano valori decisamente inferiori alla media Ue28.

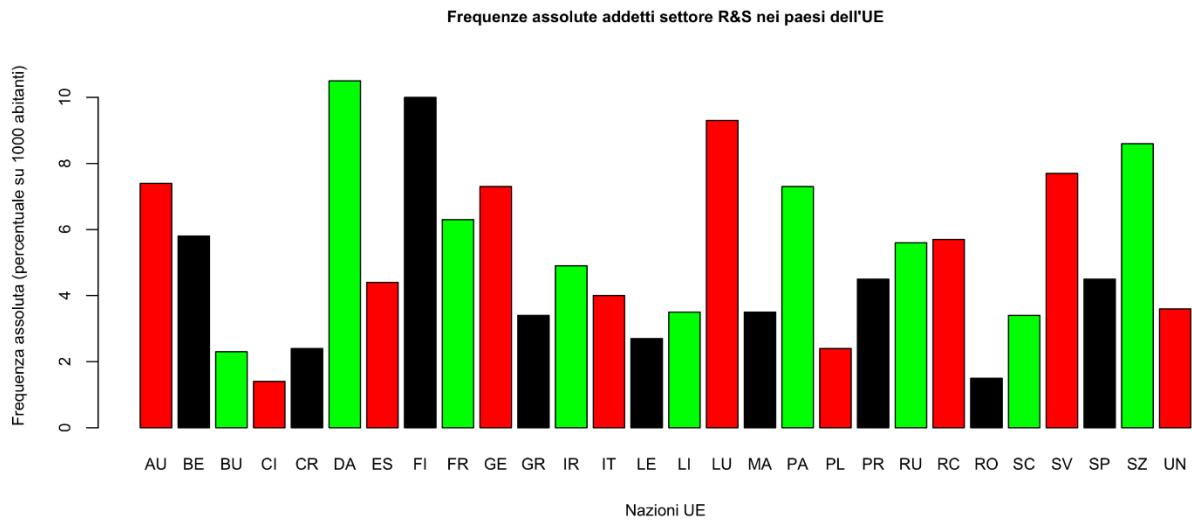
Anche nel 2010 la Svezia e la Germania si confermano per i valori più elevati dell'indice di intensità dei brevetti, con il Regno Unito e l'Italia che si collocano al di sotto della media Ue28. Cipro e Romania sono invece i paesi che presentano un tasso di indice brevettuale più basso.

3.2.3 Distribuzione di frequenza degli addetti nel settore Ricerca e Sviluppo nei paesi dell'UE (% per 1000 abitanti)

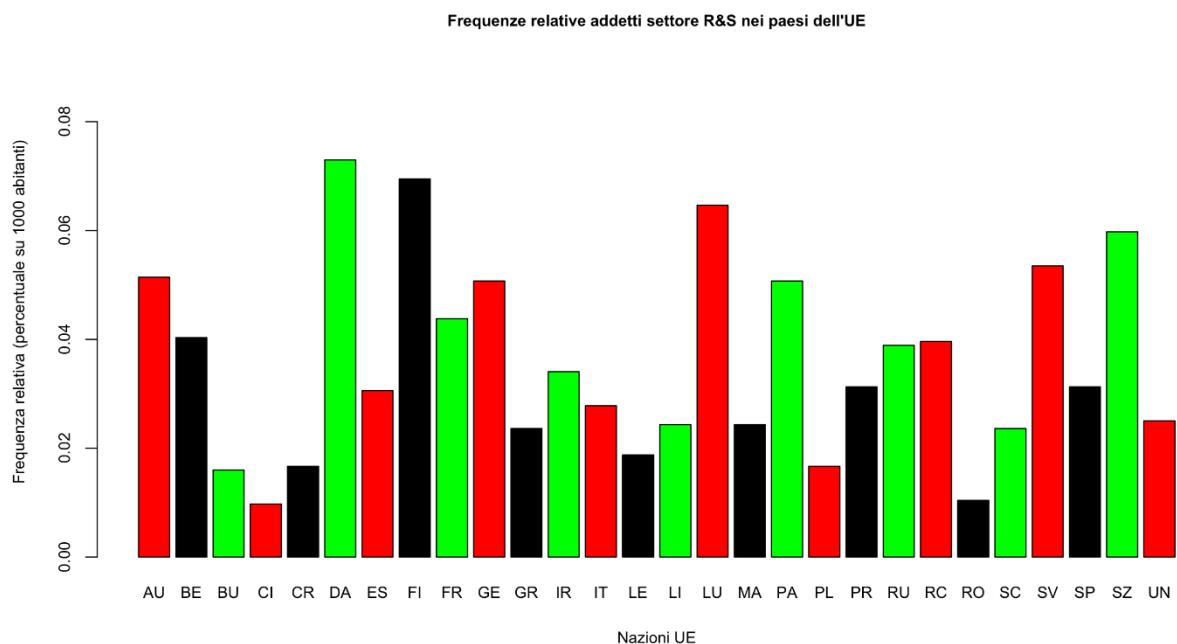
Nel data frame viene riportata la percentuale di addetti nel settore Ricerca e Sviluppo (% per 1000 abitanti) di ogni paese dell'Unione Europea. La distribuzione di

frequenza assoluta e relativa vengono rappresentate mediante un grafico a barre ottenuto nel seguente modo:

```
> barplot(mytable[,3],xlab="Nazioni UE",ylim=c(0,11),names.arg=rownames(mytable),ylab="Frequenza assoluta (percentuale su 1000 abitanti)",main="Frequenze assolute addetti settore R&S nei paesi dell'UE",col=c("red","black","green"))
```



```
> totARS <- sum(mytable[,3])
> freqRelARS<-mytable[,3]/totARS
> freqRelARS
[1] 0.051424600 0.040305768 0.015983322 0.009728978 0.016678249 0.072967338 0.030576789 0.069492703 0.043780403 0.050729673 0.023627519 0.034051425 0.027797081 0.018763030
[15] 0.024322446 0.064628214 0.024322446 0.050729673 0.016678249 0.031271716 0.038915914 0.039610841 0.010423905 0.023627519 0.053509382 0.031271716 0.059763725 0.025017373
> sum(freqRelARS)
[1] 1
> barplot(freqRelARS,xlab="Nazioni UE",ylim=c(0,0.09),names.arg=rownames(mytable),ylab="Frequenza relativa (percentuale su 1000 abitanti)",main="Frequenze relative addetti settore R&S nei paesi dell'UE",col=c("red","black","green"))
```



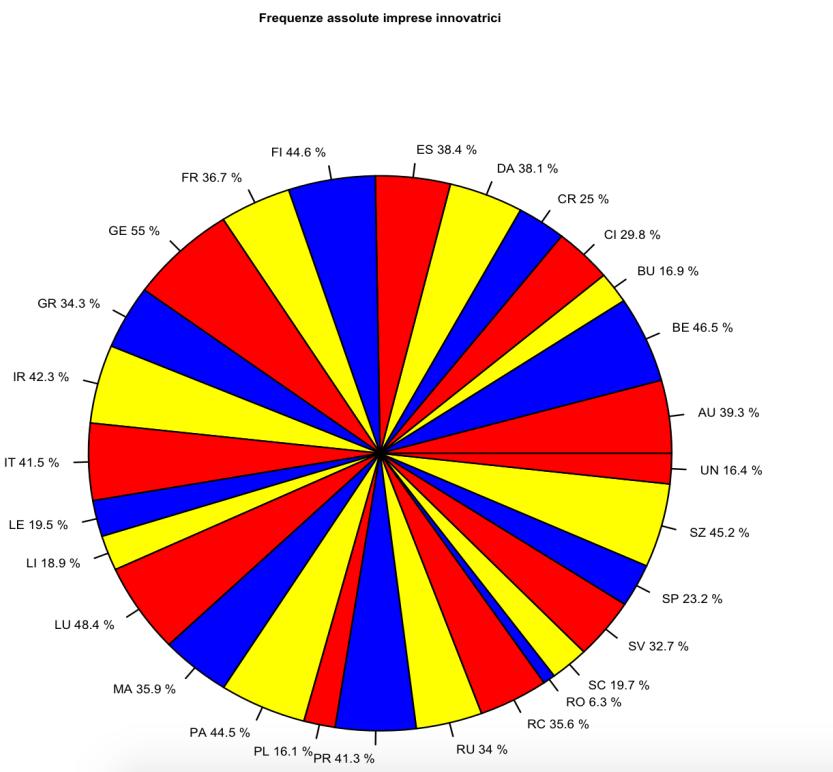
Nell’Ue28, nel 2012, gli addetti alla R&S (unità equivalenti a tempo pieno) sono mediamente 5,3 ogni mille abitanti. Il valore dell’indicatore varia da 10,5 (Danimarca) a 1,4 (Cipro). I primi posti della graduatoria europea sono occupati da paesi dell’Ue15; l’Italia, con 4,0 addetti per mille abitanti, si colloca al di sotto di Portogallo e Spagna (4,5) ma al di sopra della Grecia (3,4).

3.2.4 Distribuzione di frequenza delle imprese innovative nei paesi dell’UE

Come mostrato per le precedenti colonne del dataframe, anche in questo caso mostriamo la distribuzione delle frequenze assolute e relative per i diversi paesi dell’UE. Utilizziamo a tal proposito un grafico a torta per la rappresentazione delle frequenze assolute ed uno a barre per le frequenze relative. Per una questione di leggibilità si è ritenuto opportuno inserire le frequenze accanto al nome delle nazioni.

Di seguito gli script e i grafici ottenuti:

```
> pie(mytable[,4],labels=paste(crownames(mytable),mytable[,4],"%"),main="Frequenze assolute imprese innovative",col=c("red","blue","yellow"))
```

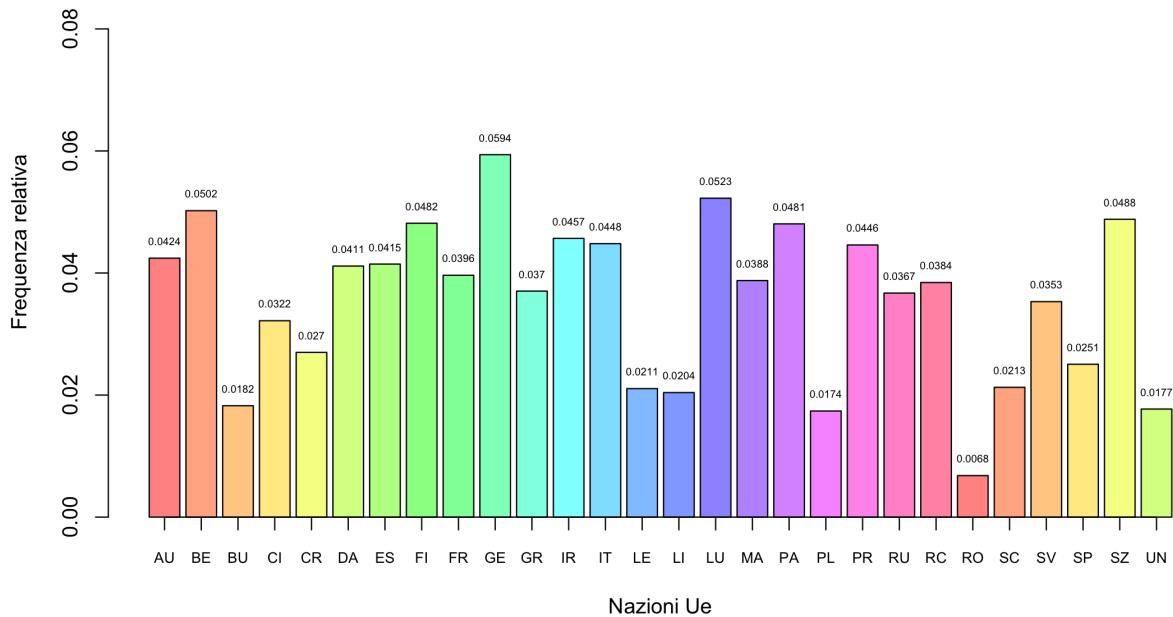


```

> tot<-sum(mytable[,4])
> freqRel<-mytable[,4]/tot
> bp<-barplot(freqRel,xlab="Nazioni Ue", ylim=c(0,0.09),ylab="Frequenza relativa",col=rainbow(22,s=0.5),main="Frequenze relative imprese innovatrici dell'UE")
> axis(1, at = bp, labels = name, cex.axis = 0.7)
> text(x=bp,freqRel,labels=round(freqRel,4),pos=3,xpd=NA,cex=0.5)

```

Frequenze relative imprese innovatrici dell'UE



Anche se nella lettura dei risultati dell'indagine sull'innovazione nelle imprese occorre considerare la diversità delle strutture economiche e produttive dei vari paesi, l'indicatore sul numero di imprese che hanno svolto attività innovative di prodotto o processo consente un primo confronto sulla propensione a innovare nei paesi dell'Ue. Nel triennio 2010-2012, l'Italia, con il 41,5 per cento di imprese innovative, si colloca al di sopra della media europea (36,0). Si conferma il ruolo trainante della Germania (55,0 per cento). Tra i paesi leader nell'innovazione continuano a primeggiare i paesi dell'Europa settentrionale, ma al di sopra della media europea si posizionano anche il Portogallo (41,3) e l'Estonia (38,4). Una bassa propensione all'innovazione si registra, invece, nei paesi dell'Europa orientale e in Spagna. Sotto la media europea è anche il Regno Unito.

3.2.5 Distribuzione di frequenza dei laureati in discipline tecniche-scientifiche nei paesi dell'UE

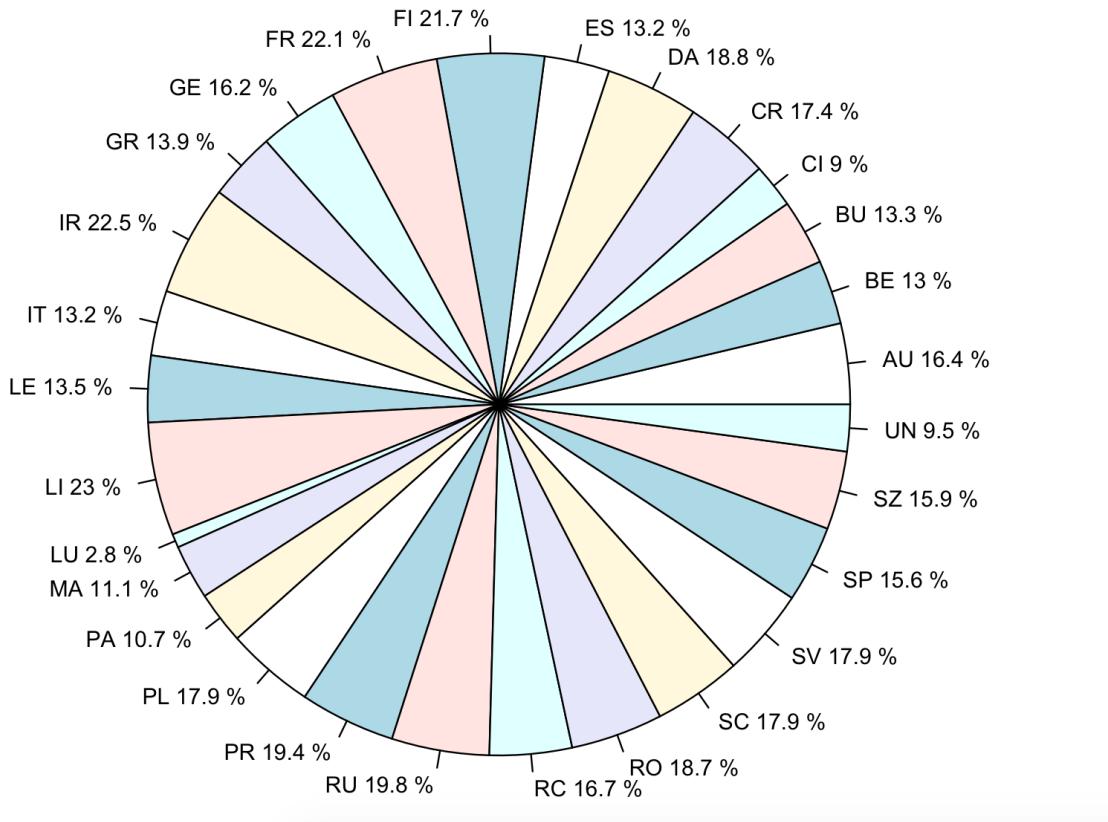
Concludiamo con questo paragrafo l'analisi per colonne dei dati oggetto di studio. A tal proposito mostriamo con un grafico a torta la distribuzione delle frequenze assolute e con un grafico a barre la distribuzione delle frequenze relative. Da notare come nel secondo grafico si sia scelto di ordinare i dati in ordine crescente per consentire al

lettore di comprendere “a primo impatto” in quali paesi vi sono più laureati in discipline tecnico-scientifiche.

Di seguito gli script e i grafici ottenuti:

```
> pie(mytable[,5],labels=paste(rownames(mytable),mytable[,5],"%"), cex=0.8,main="Frequenze assolute dei laureati in discipline tecnico-scientifiche")
```

Frequenze assolute dei laureati in discipline tecnico-scientifiche



```
> order.pop<-order(mytable$L)
> order.pop
[1] 16  4 28 18 17  2  7 13  3 14 11 26 27 10  1 22  5 19 24 25 23  6 20 21  8  9 12 15
```

Da notare come order.pop contenga gli indici della colonna “L” di mytable. Avremo quindi che il sedicesimo elemento della colonna “L” è il più piccolo, segue il quarto e così via.

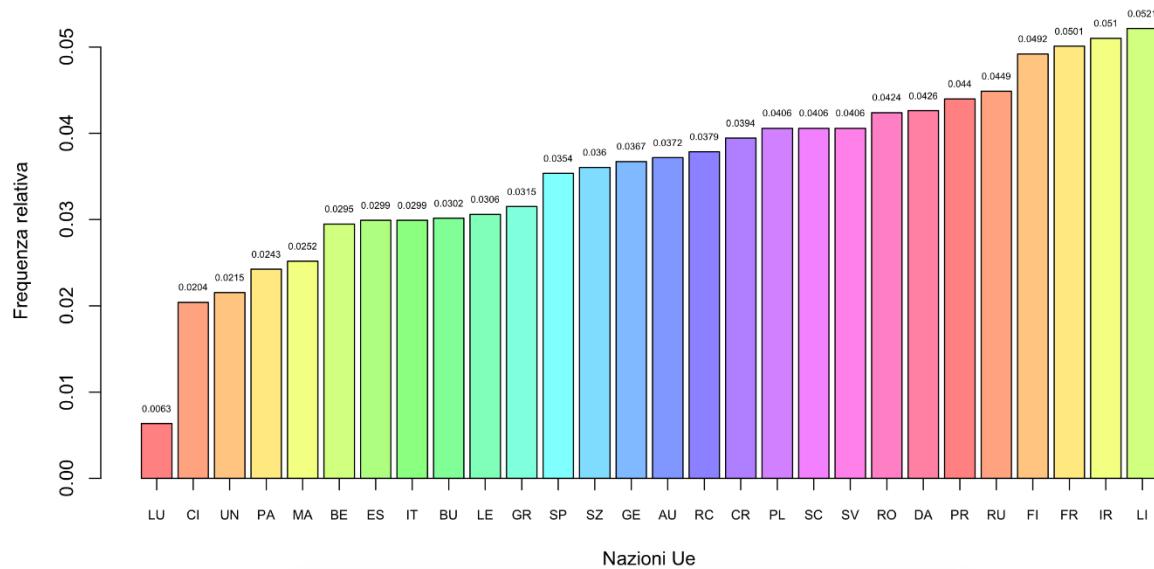
Ordinando gli elementi della tabella “mytable” per gli indici ottenuti la tabella risultante sarà:

	STRS	NB	ARS	II	L
LU	1.16	150.2	9.3	48.4	2.8
CI	0.43	9.4	1.4	29.8	9.0
UN	1.27	19.2	3.6	16.4	9.5
PA	1.97	181.6	7.3	44.5	10.7
MA	0.87	8.5	3.5	35.9	11.1
BE	2.24	138.0	5.8	46.5	13.0
ES	2.16	28.1	4.4	38.4	13.2
IT	1.26	74.9	4.0	41.5	13.2
BU	0.62	2.3	2.3	16.9	13.3
LE	0.66	7.4	2.7	19.5	13.5
GR	0.69	5.8	3.4	34.3	13.9
SP	1.27	32.0	4.5	23.2	15.6
SZ	3.28	294.9	8.6	45.2	15.9
GE	2.88	282.6	7.3	55.0	16.2
AU	2.81	208.0	7.4	39.3	16.4
RC	1.79	18.3	5.7	35.6	16.7
CR	0.75	6.8	2.4	25.0	17.4
PL	0.89	9.3	2.4	16.1	17.9
SC	0.81	8.4	3.4	19.7	17.9
SV	2.58	50.8	7.7	32.7	17.9
RO	0.48	1.7	1.5	6.3	18.7
DA	3.03	226.7	10.5	38.1	18.8
PR	1.37	8.9	4.5	41.3	19.4
RU	1.63	83.5	5.6	34.0	19.8
FI	3.43	255.8	10.0	44.6	21.7
FR	2.23	130.0	6.3	36.7	22.1
IR	1.58	68.5	4.9	42.3	22.5
LI	0.90	5.1	3.5	18.9	23.0

Di seguito lo script e il grafico ottenuto:

```
> mytableOrdered<-mytable[order.pop,]
> tot<-sum(mytableOrdered[,5])
> freqRel<-mytableOrdered[,5]/tot
> max(freqRel)
[1] 0.05214237
> br<-barplot(freqRel,xlab="Nazioni Ue",ylim=c(0,0.055),ylab="Frequenza relativa",col=rainbow(22,s=0.5),main="Frequenze relative laureati discipline tecniche-scientifiche nei paesi dell'UE")
> axis(1,at = bp, labels=rownames(mytableOrdered),cex.axis=0.7)
> text(x=bp,freqRel,labels=round(freqRel,4),pos=3,xpd=NA,cex=0.5)
```

Frequenze relative laureati discipline tecniche-scientifiche nei paesi dell'UE



La media dei paesi Ue28 è pari a 17,1 laureati ogni mille 20-29enni. I divari all'interno dell'unione europea sono rilevanti: le quote dei laureati in S&T superano il 20 per mille in Lituania, Irlanda, Francia e Finlandia; anche Regno Unito, Portogallo, Danimarca e Romania registrano incidenze elevate, ben al di sopra della media europea. L'Italia, con una quota pari al 13,2 per mille, si colloca al ventunesimo posto nella graduatoria dei paesi europei, al pari dell'Estonia, con uno scarto in negativo di quasi 4 punti percentuali dalla media comunitaria.

3.3 Confronto tra Europa settentrionale, occidentale, orientale e meridionale

Grazie ai dati in nostro possesso è stato possibile effettuare uno studio per aree geografiche diverse, ossia Europa settentrionale, occidentale, orientale e meridionale. In particolare si è utilizzata seguente suddivisione:

- Europa settentrionale: Svezia, Finlandia, Danimarca, Estonia, Lettonia, Lituania, Irlanda e Regno Unito
- Europa occidentale: Francia, Belgio, Paesi Bassi, Lussemburgo, Germania, Austria
- Europa orientale: Romania, Polonia, Rep. Ceca, Slovacchia, Ungheria, Bulgaria
- Europa meridionale: Croazia, Cipro, Grecia, Italia, Malta, Portogallo, Slovenia, Spagna

In base a tale suddivisione si è quindi proceduto al fusione delle nazioni. In particolare:

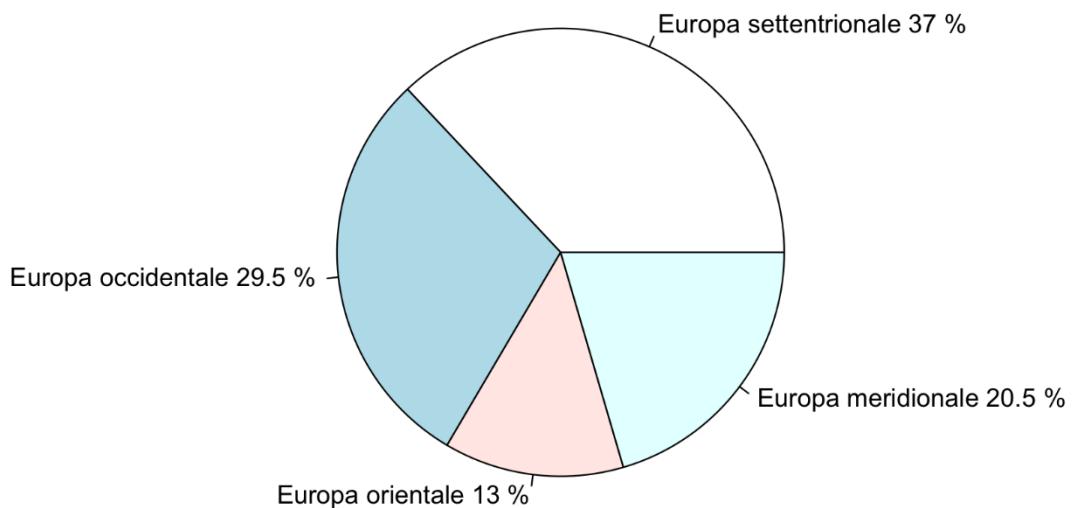
```
> sett<-mytable[27,]+mytable[8,]+mytable[6,]+mytable[7,]+mytable[14,]+mytable[15,]+mytable[12,]+mytable[21,]
> occi<-mytable[9,]+mytable[2,]+mytable[18,]+mytable[16,]+mytable[10,]+mytable[1,]
> orie<-mytable[23,]+mytable[19,]+mytable[22,]+mytable[24,]+mytable[28,]+mytable[3,]
> meri<-mytable[5,]+mytable[4,]+mytable[11,]+mytable[13,]+mytable[17,]+mytable[20,]+mytable[25,]+mytable[26,]
> row.names(sett)=c("Europa settentrionale")
> row.names(occi)=c("Europa occidentale")
> row.names(orie)=c("Europa orientale")
> row.names(meri)=c("Europa meridionale")
> #merging in un solo dataframe
> mytableDivision<-sett
> mytableDivision = rbind(mytableDivision,occi)
> mytableDivision = rbind(mytableDivision,orie)
> mytableDivision = rbind(mytableDivision,meri)
> mytableDivision
   STRS     NB    ARS     II      L
Europa settentrionale 16.67 970.0  50.2 281.0 148.4
Europa occidentale    13.29 1090.4 43.4 270.4  81.2
Europa orientale      5.86  59.2 18.9 111.0  94.0
Europa meridionale    9.22 197.1 31.4 263.7 117.5
```

Dal momento che le categorie in analisi sono solamente quattro, si è scelto di rappresentare tali dati utilizzando dei grafici a torta, ottenuti tramite il comando pie. La funzione definita ed utilizzata per generare questo tipo di grafici è riportata di seguito:

```
plot.pie <-function(column,title){
  freqRel <- round((column/sum(column))*100,1)
  pie(column,labels= paste(DIVISION.LABELS,freqRel,"%"),main=title)
}
```

3.3.1 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - spesa totale per ricerca e sviluppo

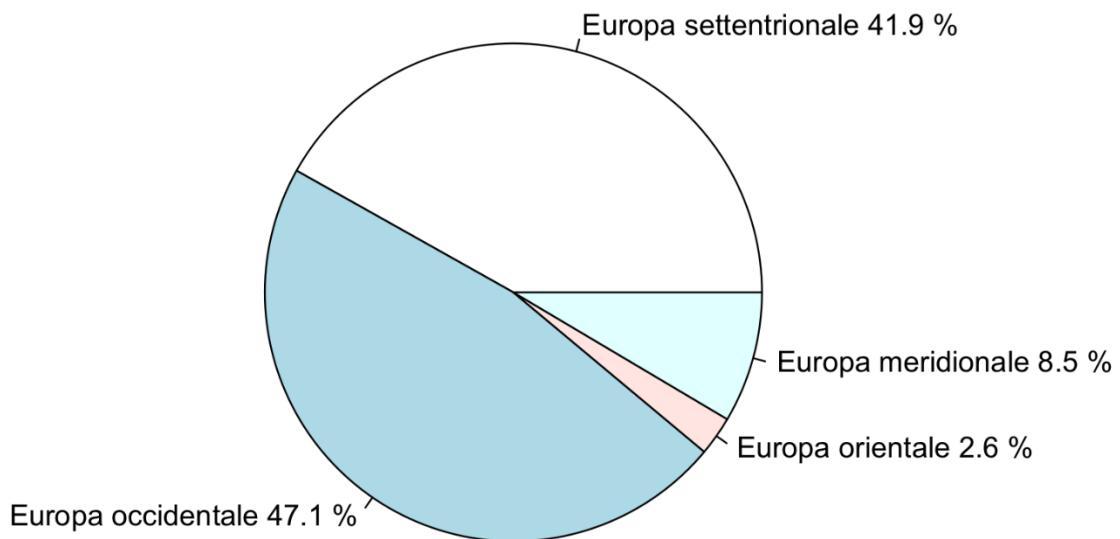
Frequenza relativa spesa totale per ricerca e sviluppo



In accordo ai risultati riportati nella sezione 3.2.1, dal grafico sopra riportato si evince come l'Europa settentrionale sia la macroregione a spendere di più in ricerca e sviluppo. Segue l'Europa occidentale (Germania e Austria contribuiscono significativamente al raggiungimento di tale risultato). Infine abbiamo l'Europa meridionale (20.5%) e l'Europa orientale (13%).

3.3.2 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - numero di brevetti

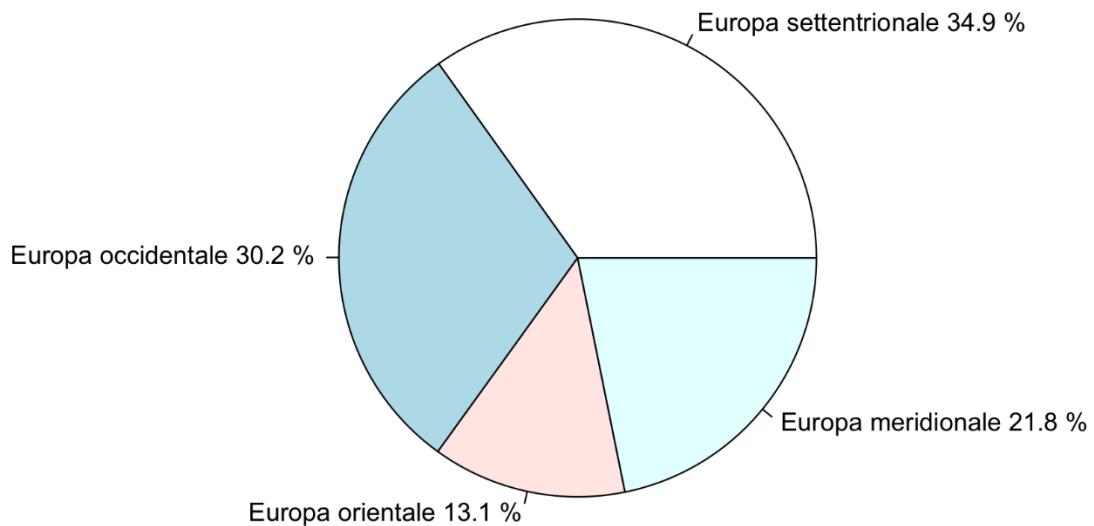
Frequenza relativa numero di brevetti



Dal grafico sopra riportato si evince come i paesi dell'Europa occidentale e settentrionale abbiano un indice di attività brevettuale maggiore mentre i paesi dell'Europa meridionale e orientale riportino un indice decisamente più basso.

3.3.3 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - addetti alla ricerca e sviluppo

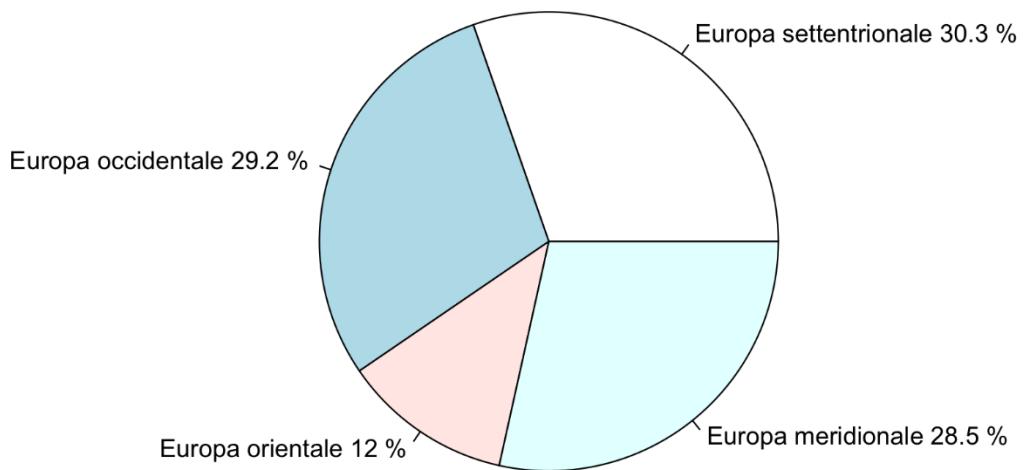
Frequenza relativa addetti alla ricerca e sviluppo



Anche in questo caso, la macroregione con più addetti nel settore ricerca e sviluppo è l'Europa settentrionale. Segue l'Europa occidentale (con uno stacco di quasi 5 punti percentuali) dalla prima classificata. Fanalino di coda l'Europa meridionale (21.8 %) e l'Europa orientale (13.1%).

3.3.4 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - imprese innovatrici

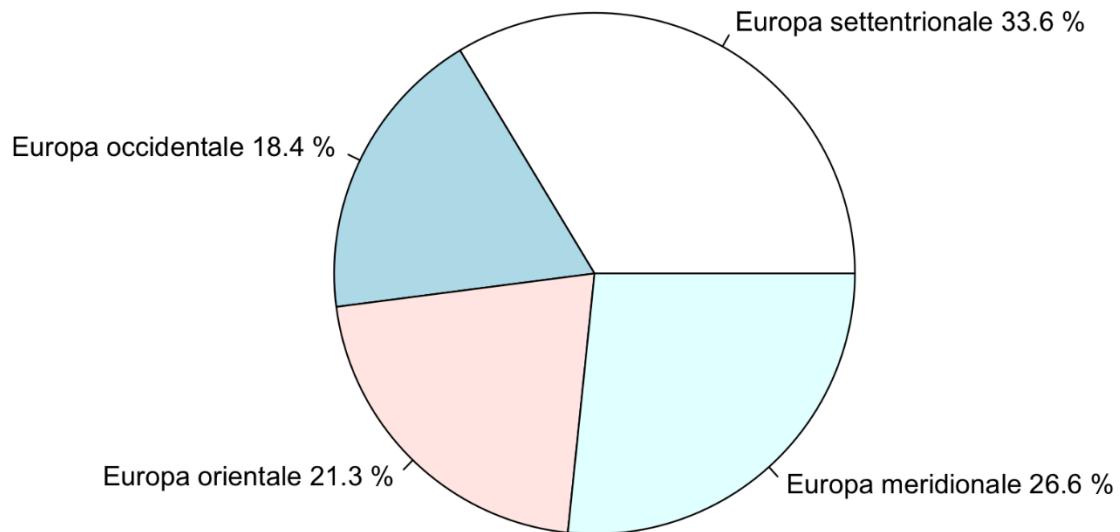
Frequenza relativa imprese innovatrici



Dal grafico sopra riportato si evince come, ad eccezione dell'Europa orientale (12%), tutte le macroregioni europee presentino un egual numero di imprese innovatrici.

3.3.5 Confronto tra Europa settentrionale, occidentale, orientale e meridionale - laureati in discipline tecnico-scientifiche

Frequenza relativa laureati in discipline tecnico-scientifiche



Dal grafico sopra riportato si evince come, anche per quanto riguarda i laureati in discipline tecnico scientifiche, la prima in classifica sia l'Europa settentrionale. È interessante notare come l'Europa orientale, a differenza dei risultati presentati nelle sezioni precedenti, presenti un valore superiore a quello dell'Europa occidentale e non troppo inferiore rispetto a quello dell'Europa meridionale.

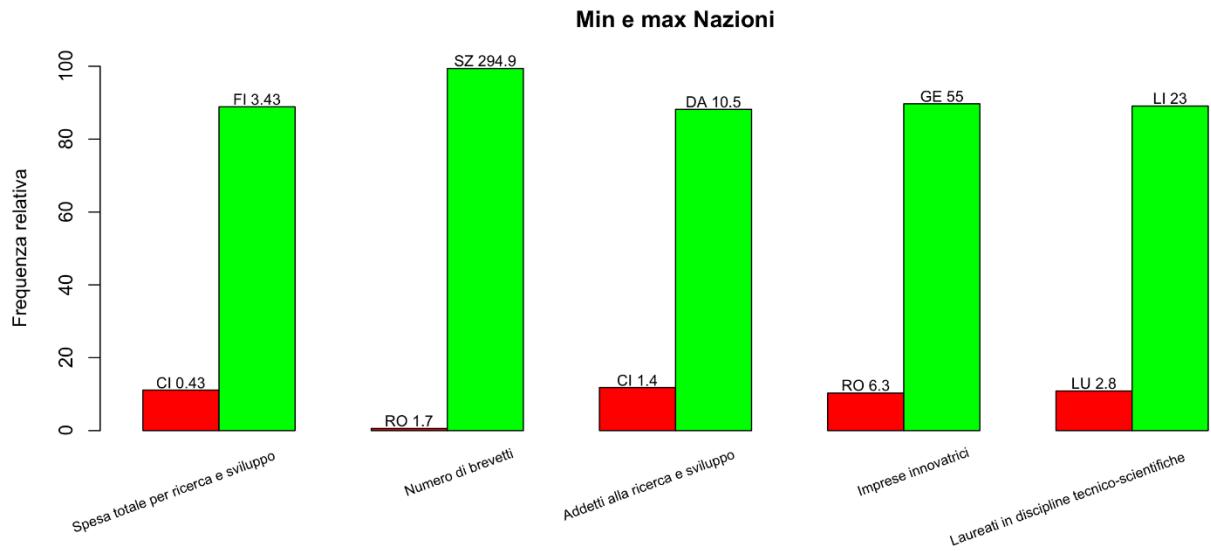
3.4 Minimo e massimo per colonne

In questa sezione mostreremo colonna per colonna qual è la nazione/macroregione che ottiene il risultato migliore e quale quello peggiore.

La funzione utilizzata per ottenere i grafici di seguito mostrati è la seguente:

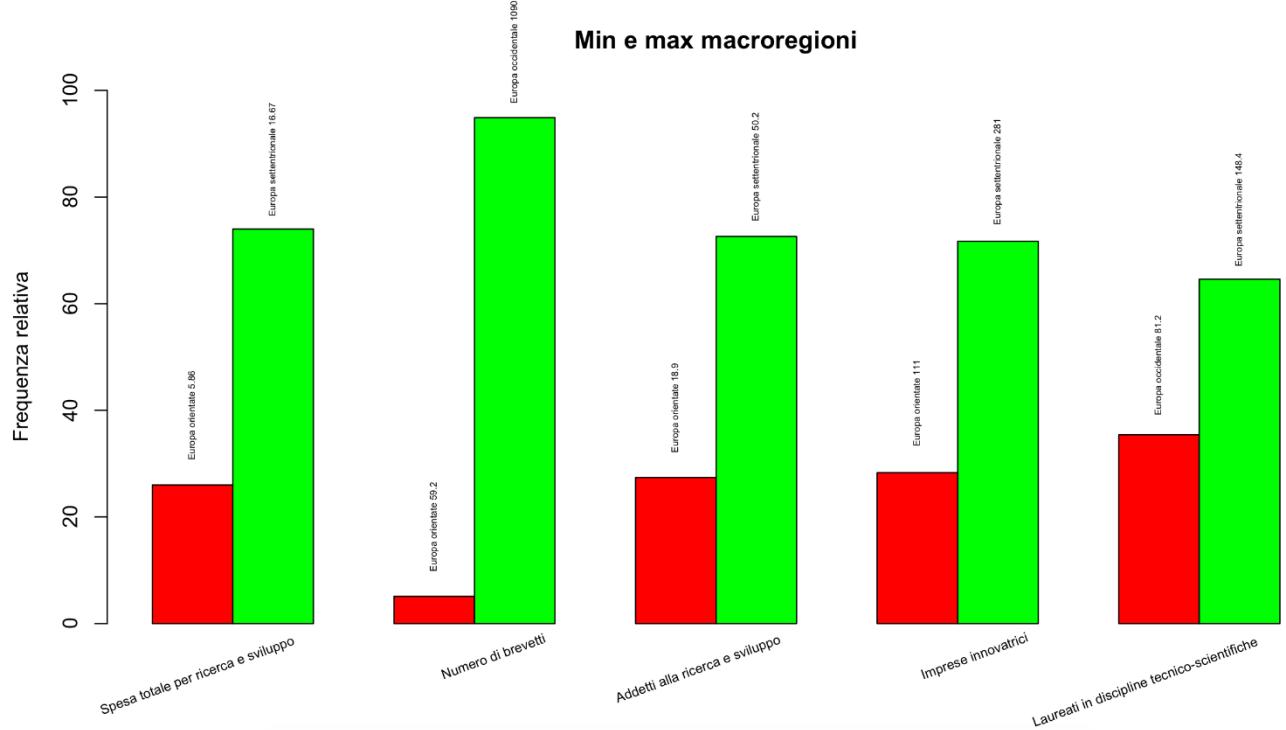
```
plot.barplot<-function(dataframe,title){  
  
min <- min(dataframe[,1])  
minName <- row.names(dataframe)[(which(dataframe[,1]==min(dataframe[,1])))]  
  
max <- max(dataframe[,1])  
maxName <- row.names(dataframe)[(which(dataframe[,1]==max(dataframe[,1])))]  
  
tot<-min+max  
  
percMin<-round((min/tot)*100,1)  
percMax<-round((max/tot)*100,1)  
  
tabMinMax <- c(min,max)  
tabMinMaxPerc <- c(percMin,percMax)  
tabMinMaxPercForPrint <- c(percMin,percMax)  
tabName <- c(minName,maxName)  
  
for(i in (2:(dim(dataframe)[2]))){  
  
min <- min(dataframe[,i])  
minName <- row.names(dataframe)[(which(dataframe[,i]==min(dataframe[,i])))]  
  
max <- max(dataframe[,i])  
maxName <- row.names(dataframe)[(which(dataframe[,i]==max(dataframe[,i])))]  
  
tot<-min+max  
  
percMin<-round((min/tot)*100,1)  
percMax<-round((max/tot)*100,1)  
  
tabMinMax <- c(tabMinMax,c(min,max))  
tabName <- c(tabName,c(minName,maxName))  
tabMinMaxPercForPrint <-c(tabMinMaxPercForPrint,c(percMin,percMax))  
  
tabMinMaxPerc = rbind(tabMinMaxPerc,c(percMin,percMax))  
  
}  
  
bp<-barplot(t(as.matrix(tabMinMaxPerc)) , beside=TRUE, main=title, ylab="Frequenza relativa", ylim=c(0,100), xaxt="n", col=c("red","green"))  
text(seq(3,70,by=3), par("usr")[3]-(5.0), srt = 20, adj = 1.1, xpd =TRUE , labels =COLUMNS,cex=0.73)  
text(x=bp,tabMinMaxPercForPrint+2,labels=paste(tabName,tabMinMax),xpd=NA,cex=0.8)  
}
```

3.4.1 Minimo e massimo nazioni



Come si evince dal grafico ottenuto, è presente una variabilità elevatissima tra le diverse nazioni per ognuna delle colonne analizzate. In particolare fanalino di coda è spesso Cipro mentre al primo posto (barre verdi) troviamo spesso uno dei paesi dell'Europa settentrionale. Da notare come per quanto riguarda la colonna "Laureati indiscipline tecnico-scientifiche" il valore più basso ottenuto sia quello del Lussemburgo.

3.4.2 Minimo e massimo macroregioni



Alla luce dei risultati presentati nel paragrafo 3.3.1 (i primi posti in classifica assegnati quasi sempre a nazioni dell’Europa settentrionale) anche dal grafico in alto si evince come i migliori risultati siano dell’Europa settentrionale, eccezion fatta per quanto riguarda la colonna “Numero di brevetti” in cui ottiene il miglior risultato l’Europa Occidentale. Fanalino di coda, invece, sempre l’Europa Orientale.

3.5 Visione globale dei migliori e dei peggiori

Mostriamo con dei grafici a barre la distribuzione dei fattori relativi al progresso ed all’impegno nel settore Ricerca e Sviluppo per i paesi che nel grafico 3.3.1 hanno ottenuto le prime ed ultime posizioni. La funzione utilizzata è la seguente:

```

#ex. rowsNumber: c(1,5,7), ex color: c(red,blue,violet)
plot.barplotGlobal<-function(dataframe,rowsNumber,columnStart,columnEnd,color){

  par(mfcol=c(columnStart,columnEnd))

  namesNations<-c()

  #take the nations' name
  for (i in 1:dim(dataframe)[1]){
    if(i %in% rowsNumber){
      namesNations<-c(namesNations,rownames(dataframe)[i])
    }
  }

  for(i in columnStart:columnEnd){

    rows<-NULL

    for(j in rowsNumber){

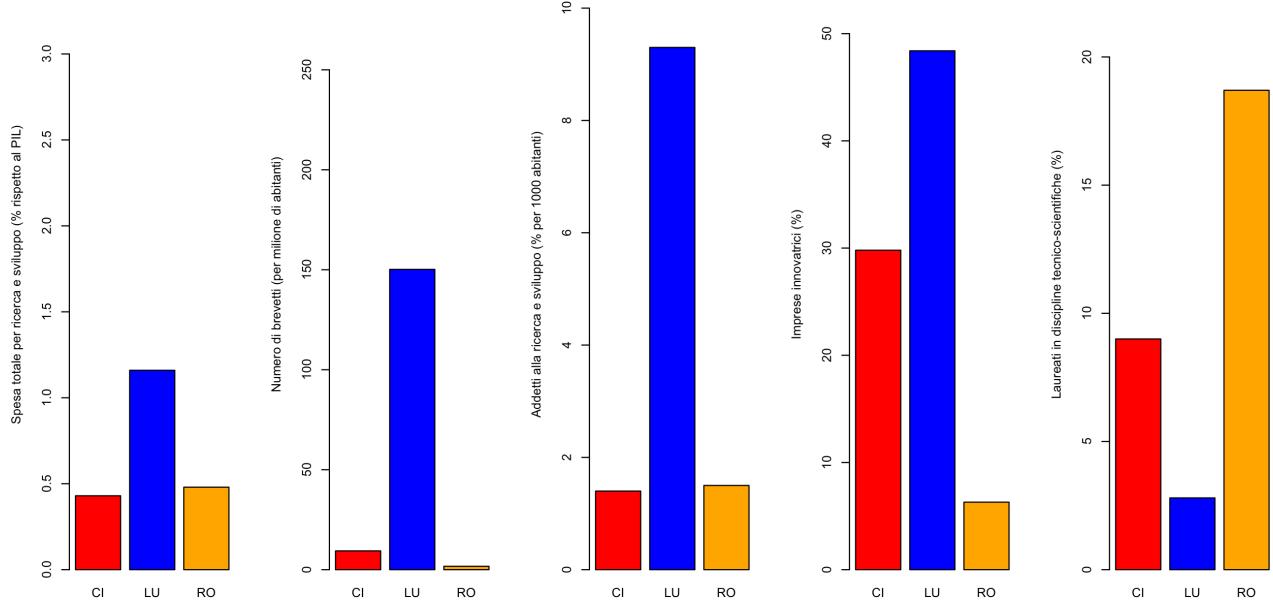
      rows<-c(rows,dataframe[j,i])
    }

    print(rows)

    br<-barplot(rows,names.arg=namesNations,ylim=c(0,max(dataframe[,i])),ylab=COLUMNS[i],col=color)
  }
}

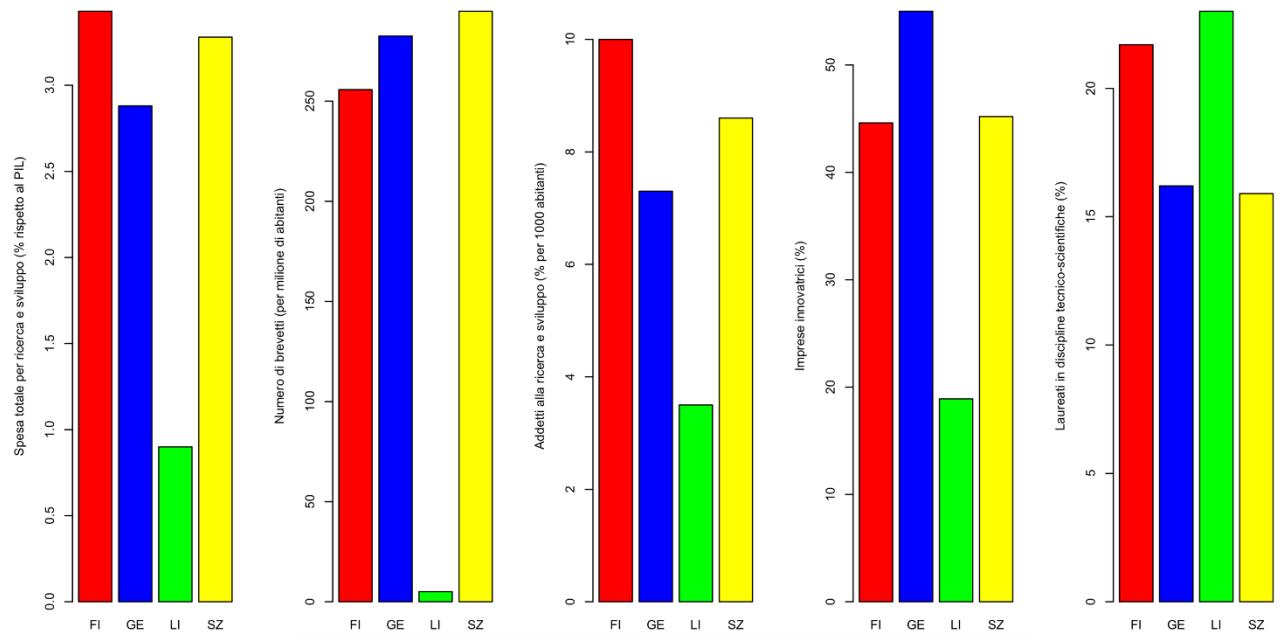
```

3.5.1 Minimi



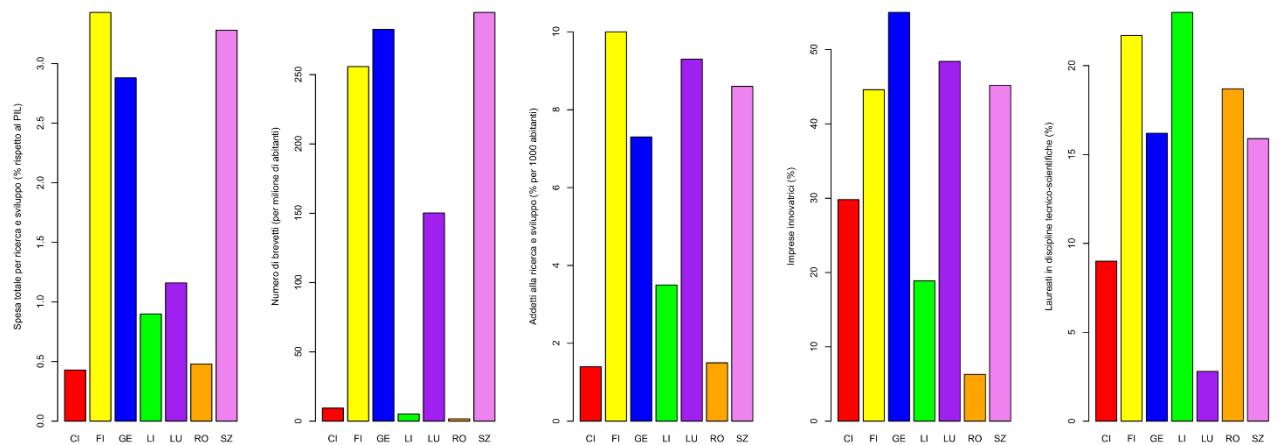
Dal grafico in alto, è interessante notare come il Lussemburgo (peggiore risultato per “Laureati in discipline tecnico-scientifiche (%)”) per i rimanenti indicatori si comporti decisamente meglio delle altre nazioni che hanno ottenuto almeno un minimo.

3.4.2 Massimi

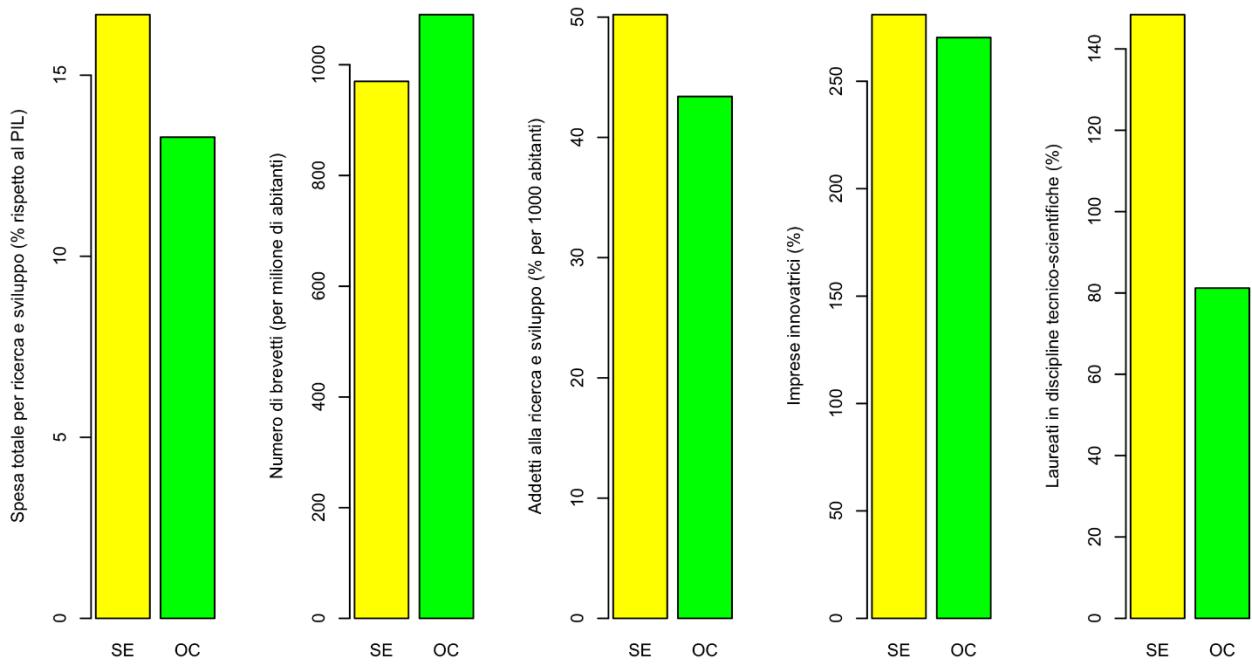


È evidente dal grafico in alto come la Lituania che guadagna il primo posto per il numero di laureati in discipline tecnico scientifiche, ottenga in realtà risultati decisamente più bassi per gli altri indicatori oggetto di studio. Per quanto riguarda, invece, Finlandia, Svezia e Germania quest’ultime ottengono dei buoni punteggi per tutti gli indicatori.

3.5.2 Massimi e minimi a confronto

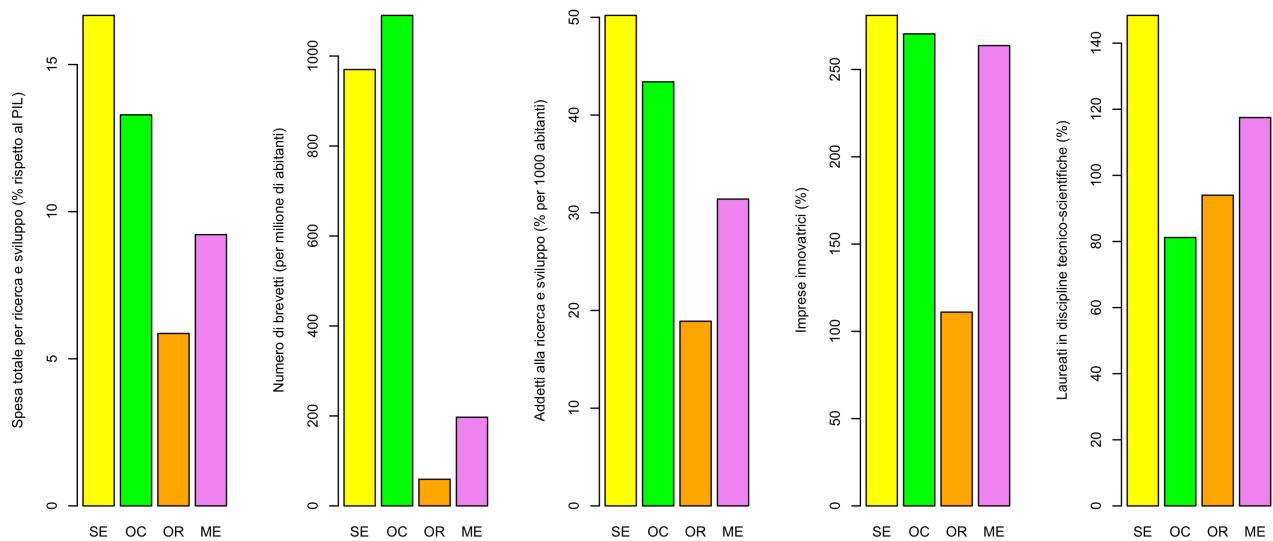


3.5.3 Massimi macroregioni



3.5.4 Visione globale macroregioni

Poiché i risultati peggiori sono stati ottenuti solamente dall'Europa orientale (eccezione fatta per “Laureati in discipline tecnico scientifiche (%)” - Europa occidentale) mostriamo nel grafico successivo tutte le macroregioni a confronto.



4. INDICI DI POSIZIONE E DISPERSIONE

In questo paragrafo verrà compiuta un'analisi dettagliata dei dati in possesso mediante indici di sintesi, al fine di confermare ciò che era stato visto precedentemente mediante l'utilizzo di strumenti grafici. Gli indici utilizzati sono:

- La **media** e la **mediana**. Sono indici di posizione che descrivono intorno a quali valori è centrato l'insieme dei dati.
- I **quartili**. Sono indici di posizione che si ottengono dividendo l'insieme dei dati ordinati in quattro parti uguali: il primo quartile Q_1 è un valore tale che il 25% dei dati ordinati è minore o uguale di Q_1 ; il secondo quartile, Q_2 , è un valore tale che il 50% dei dati ordinati risulta minore o uguale di Q_2 , ed in particolare tale valore coincide con la mediana; il terzo quartile, invece, Q_3 , è un valore tale che il 75% dei dati ordinati è minore o uguale a Q_3 .
- La **varianza**, la **deviazione standart** e il **coefficiente di variazione** sono indici di dispersione dei dati. Tali indici misurano la dispersione dei dati attorno alla media.

Mostriamo adesso una definizione formale degli indici appena elencati.

Supponiamo di avere un insieme x_1, x_2, \dots, x_n di n dati, detto campione di ampiezza o numerosità pari a n . La **media campionaria** risulta essere la media aritmetica di questi valori. Fondamentalmente è uguale a:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Ordinando in maniera crescente l'insieme dei dati di dimensione n possiamo definire la **mediana** nel seguente modo:

- Se n è dispari, la **mediana** è il valore in posizione $(n + 1)/2$.
- Se n è pari, la **mediana** è definita come la media aritmetica dei valori che occupano le posizioni $n/2$ e $\left(\frac{n}{2}\right) + 1$.

La media e la mediana risultano essere utili per descrivere i valori centrali dei dati. La media utilizza tutti i dati dell'insieme ed è influenzata da valori molto alti o molto bassi, mentre invece la mediana considera soltanto uno o due valori della distribuzione dei dati senza considerare i valori estremi.

La **varianza campionaria**, denotata con s^2 (s in caso si parli di deviazione

standard), è definita nel seguente modo:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Il **coefficiente di variazione** è definito come il rapporto tra la deviazione standard (la deviazione standard ha la stessa unità di misura dei valori osservati al contrario della varianza che ha come unità di misura il quadrato dell'unità di misura dei valori di riferimento) e il modulo della media campionaria, formalmente:

$$CV = \frac{s}{|\bar{x}|}$$

Dopo aver definito gli indici che verranno utilizzati nell'analisi compiuta, viene calcolato per ogni variabile:

- Il minimo
- Il primo quartile
- La mediana
- La media
- Il terzo quartile
- Il massimo
- La varianza
- La deviazione standard
- Il coefficiente di variazione

Inoltre verrò mostrato anche il relativo boxplot. Il boxplot, detto anche scatola con baffi, è il disegno di una scatola i cui estremi sono $Q1$ e $Q3$, tagliata da una linea orizzontale in corrispondenza di $Q2$, ossia della mediana. In basso e in alto sono presenti altre due linee orizzontali, dette baffi. Il baffo inferiore corrisponde anche al valore più piccolo tra le osservazioni che risulta maggiore o uguale di $Q1 - 1.5 * (Q3 - Q1)$, mentre il baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale di $Q3 + 1.5 * (Q3 - Q1)$. La distanza tra il primo e il terzo quartile è detta intervallo interquartile o scarto interquartile. Se tutti i dati rientrano nell'intervallo $Q1 - 1.5 * (Q3 - Q1)$, $Q3 + 1.5 * (Q3 - Q1)$ i baffi sono posti in corrispondenza del minimo valore e del massimo valore del campione. Gli eventuali valori al di fuori dell'intervallo $Q1 - 1.5 * (Q3 - Q1)$, $Q3 + 1.5 * (Q3 - Q1)$ sono detti valori anomali e vengono mostrati sotto forma di punti. Le caratteristiche della distribuzione di frequenza che si evincono tramite un boxplot sono: la centralità, la forma, la dispersione e la presenza di eventuali valori anomali, detti *outlier*. La centralità è espressa dalla mediana. La forma simmetrica o asimmetrica può essere dedotta esaminando le distanze del primo

e del terzo quartile dalla linea mediana. I baffi, superiore e inferiore, forniscono informazioni sulla dispersione e sulla forma della distribuzione, ed anche sulle code della distribuzione. Infatti la dispersione è deducibile esaminando le distanze del baffo superiore da $Q3$ e del baffo inferiore da $Q1$.

Le funzioni utilizzate per ottenere le informazioni complessive di una variabile sono:

- **summary()**: calcola per la variabile fornita in input: il minimo, il massimo, i quartili e la media campionaria.
- **var()**: calcola la varianza della variabile fornita in input.
- **sd()**: calcola la deviazione standard della variabile input.
- **boxplot()**: disegna il boxplot relativo alla variabile data in input.

Lo script di seguito mostrato utilizza queste funzioni per stampare a video gli indici fin ora discussi. Da notare come questo crei anche il boxplot in cui sono inserite automaticamente gli outlier, il massimo, il minimo e la mediana (i nomi delle nazioni sono aggiunti automaticamente senza che l'utente li inserisca manualmente con il comando `text`).

```
plot.boxplot <- function(dataframe,columnNumber,adjustsMin=0,adjustsMax=0,color="green"){
  print(summary(dataframe[,columnNumber]))

  print(paste("Varianza:",var(dataframe[,columnNumber])))
  standardDeviation <- sd(dataframe[,columnNumber])
  print(paste("Deviazione standard:", standardDeviation))

  print(paste("Coefficiente di variazione:", standardDeviation/abs(mean(dataframe[,columnNumber]))))

  myboxplot <- boxplot(dataframe[,columnNumber],main=COLUMNS[columnNumber],col=color)

  #print all the information in the boxplot
  info <- myboxplot$stats
  outlier <- myboxplot$out
  #row
  numberOfRows <- dim(dataframe)[1]

  textMin <-c()
  textMax <-c()

  #min and max
  for(i in 1:numberOfRow){

    if(dataframe[i,columnNumber] %in% info){
      #the min
      if(dataframe[i,columnNumber]== info[1]){
        #more min
        textMin <- c(textMin,rownames(dataframe)[i])
      }
      #the max
      } else if(dataframe[i,columnNumber]== info[5]){
        #more max
        textMax <- c(textMax,rownames(dataframe)[i])
      }
    }

    } else if(dataframe[i,columnNumber] %in% outlier){
      #right
      text(1,dataframe[i,columnNumber],rownames(dataframe)[i],cex=0.6,pos=4)
    }

  }

  #below
  text(1,info[1]+ adjustMin, textMin,cex=0.6, pos=1)
  #above
  text(1,info[5]+ adjustMax, textMax,cex=0.6, pos=3)

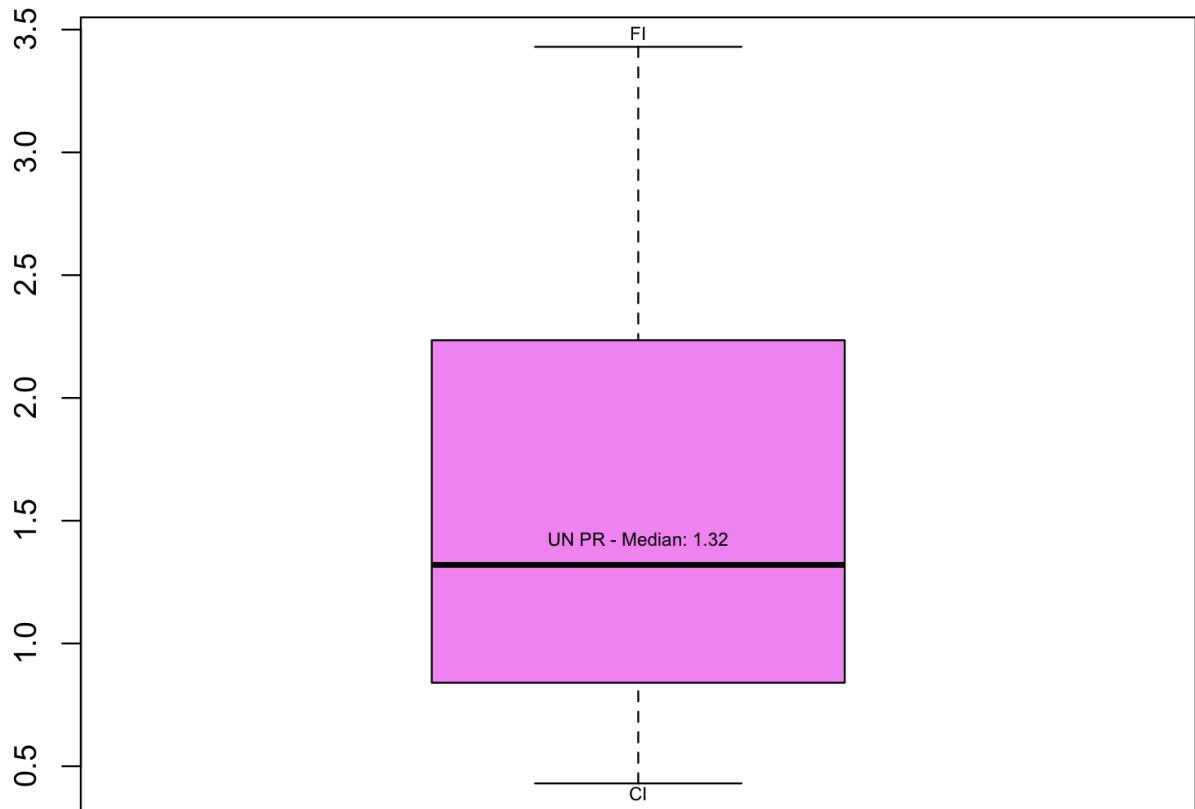
  myOrder <- order(dataFrame[,columnNumber])

  if((numberOfRow %% 2)==0){
    text(1,info[3], paste(rownames(dataframe)[myOrder[numberOfRow/2]],rownames(dataframe)[myOrder[(numberOfRow/2)+1]],"- Median:",info[3]),cex=0.6, pos=3)
  } else {
    text(1,info[3], paste(rownames(dataframe)[myOrder[round(numberOfRow/2)]],"- Median:",info[3]),cex=0.6, pos=3)
  }
}
```

4.1 Boxplot relativo alla distribuzione della spesa nel settore Ricerca e Sviluppo dei paesi dell'UE

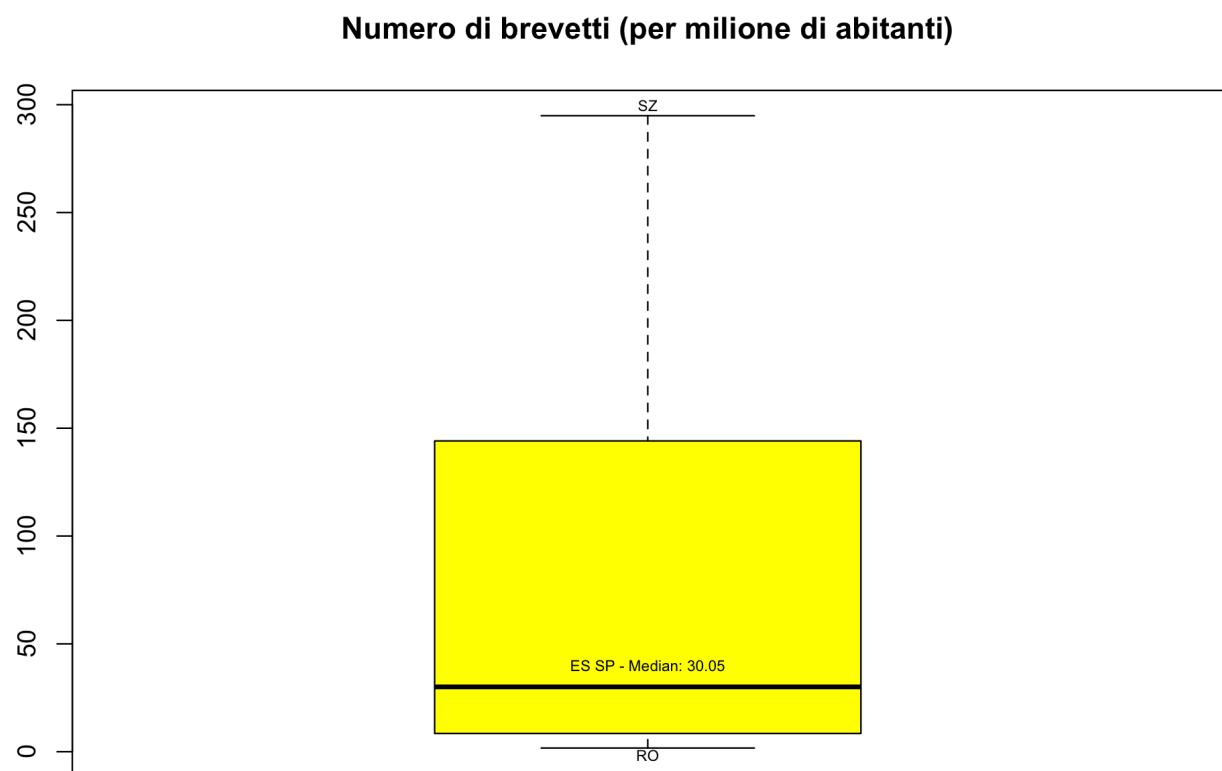
```
> plot.boxplot(mytable,1,+0.05,-0.05,color="violet")
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   0.430  0.855  1.320  1.609  2.232  3.430
[1] "Varianza: 0.826679365079365"
[1] "Deviazione standard: 0.909219096301527"
[1] "Coefficiente di variazione: 0.565233896457432"
```

Spesa totale per ricerca e sviluppo (% rispetto al PIL)



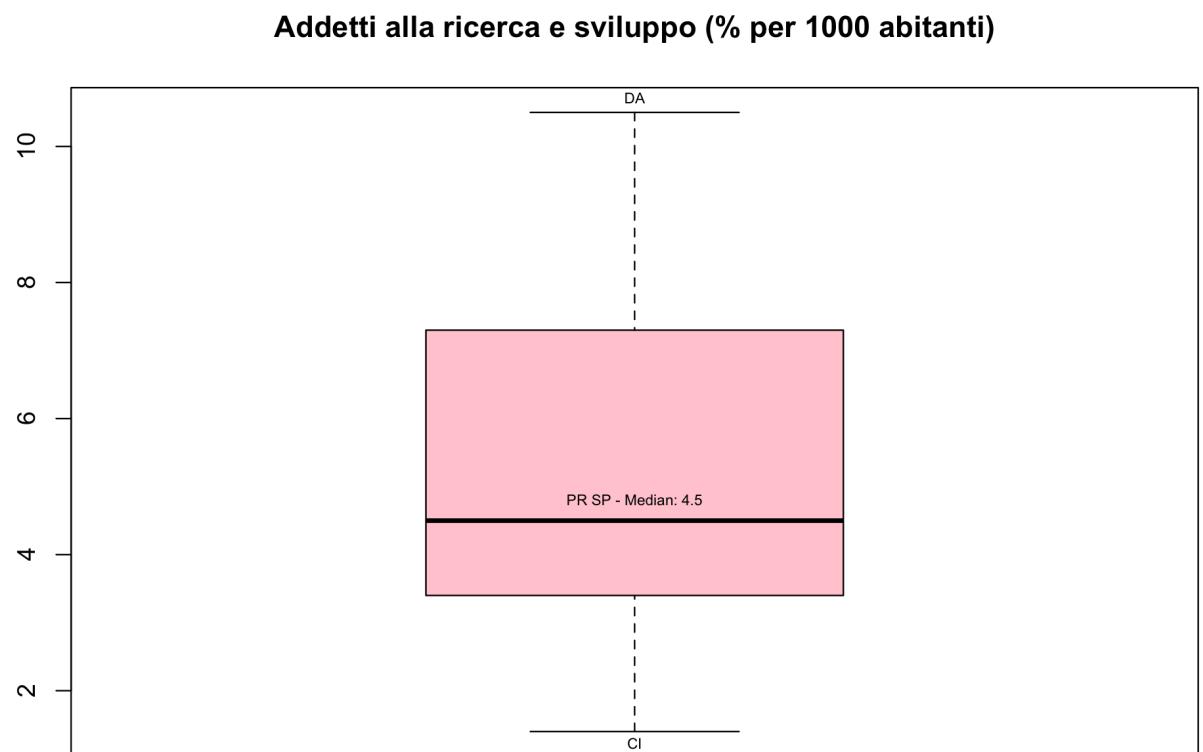
4.2 Boxplot relativo alla distribuzione dei brevetti nei paesi dell'UE

```
> plot.boxplot(mytable,2,+5,-5,color="yellow")
   Min. 1st Qu. Median Mean 3rd Qu. Max.
1.700 8.475 30.050 82.740 141.000 294.900
[1] "Varianza: 9246.5846957672"
[1] "Deviazione standard: 96.1591633478952"
[1] "Coefficiente di variazione: 1.162194748453"
```



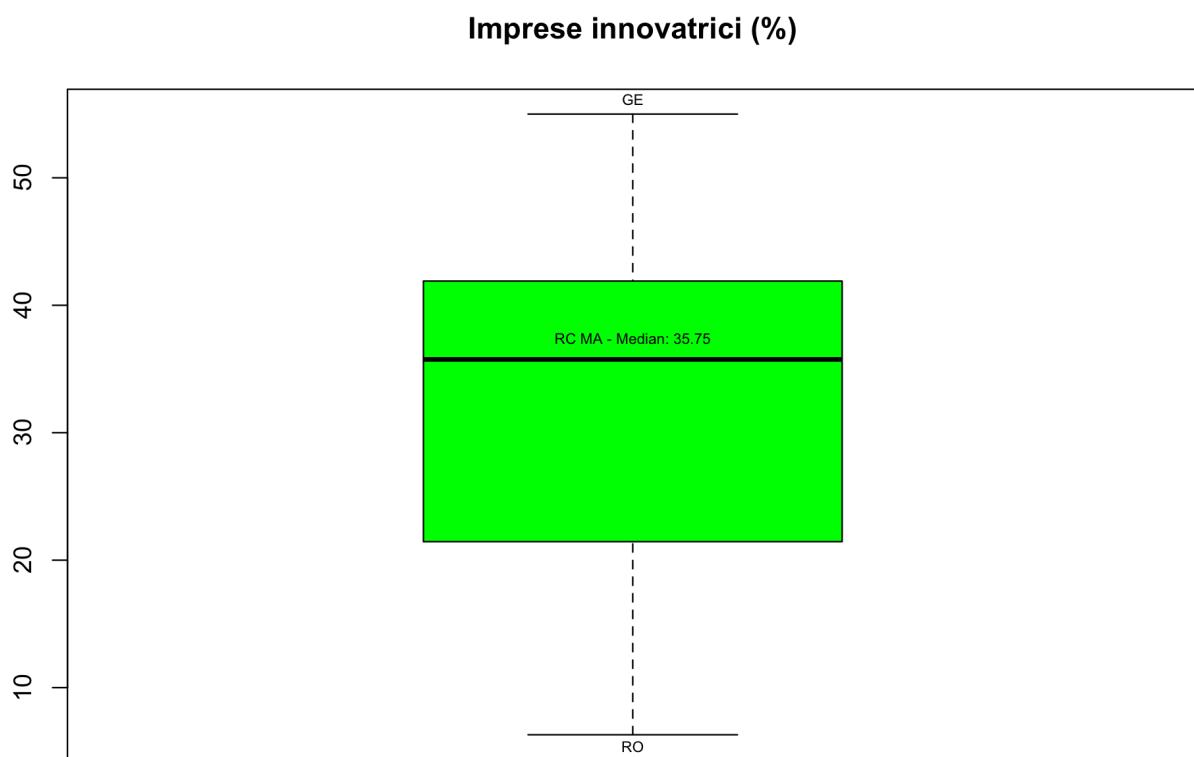
4.3 Boxplot relativo alla distribuzione degli addetti alla ricerca ricerca e sviluppo dei paesi della UE

```
> plot.boxplot(mytable,3,+0.09,-0.09,color="pink")
   Min. 1st Qu. Median Mean 3rd Qu. Max.
1.400 3.400 4.500 5.139 7.300 10.500
[1] "Varianza: 6.58247354497354"
[1] "Deviazione standard: 2.56563316648611"
[1] "Coefficiente di variazione: 0.499219796119604"
```



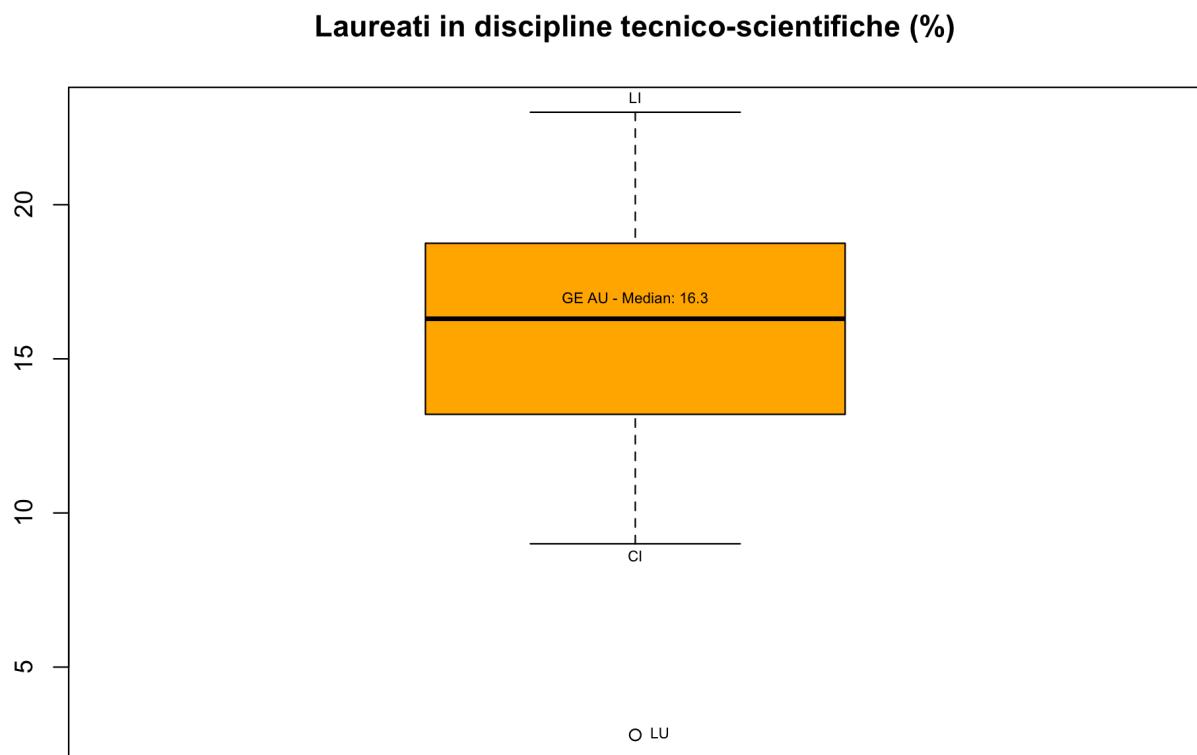
4.4 Boxplot relativo alla distribuzione delle imprese innovative dei paesi dell'UE.

```
> plot.boxplot(mytable,4,+0.5,-0.5,color="green")
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   6.30  22.32  35.75 33.08  41.70  55.00
[1] "Varianza: 145.723425925926"
[1] "Deviazione standard: 12.0715958317832"
[1] "Coefficiente di variazione: 0.364976442381957"
```



4.5 Boxplot relativo alla distribuzione dei laureati in discipline tecniche-scientifiche dei paesi dell'UE.

```
> plot.boxplot(mytable,5,+0.2,-0.2,color="orange")
   Min. 1st Qu. Median Mean 3rd Qu. Max.
2.80 13.20 16.30 15.75 18.72 23.00
[1] "Varianza: 21.2685052910053"
[1] "Deviazione standard: 4.61177897247963"
[1] "Coefficiente di variazione: 0.292744981250124"
```



5. CORRELAZIONE TRA VARIABILI

In questo paragrafo viene analizzata la correlazione tra le variabili al fine di mettere in evidenza le relazioni che intercorrono tra esse. La dipendenza tra due variabili X e Y può essere verificata mediante l'utilizzo di diagrammi di dispersione o scatterplot. Per fare ciò, si pongono in ascissa i dati relativi ad una delle due variabili e in ordinata quelli relativi all'altra variabile. Le singole osservazioni sono rappresentate mediante punti o cerchietti. Lo scatterplot mette in evidenza se i punti risultano essere sparsi senza regolarità o se, invece, risultano essere sparsi seguendo una regolarità, ovvero se le variabili sono connesse mediante una relazione lineare ad esempio. Una misura quantitativa della correlazione tra variabili può essere ottenuta mediante:

- Covarianza campionaria
- Coefficiente di correlazione campionaria

5.1 Covarianza

Date due variabili X e Y, la covarianza campionaria tra le due variabili è definita nel seguente modo:

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Dove x_i e y_i per $i = 1, 2, \dots, n$ sono le osservazioni delle singole variabili e \bar{x} e \bar{y} le medie campionarie delle variabili. Se $C_{xy} > 0$ le due variabili risultano essere correlate positivamente; se $C_{xy} < 0$ significa che la correlazione è negativa; se, invece, $C_{xy} = 0$ si dice che le due variabili non sono correlate. La covarianza campionaria delle variabili prese in considerazione è stata calcolata nel seguente modo:

```
> covarianza <- cov(mytable)
> covarianza
      STRS          NB          ARS          II          L
STRS  0.8266794  74.89410  2.0212434  7.340926  1.1260794
NB    74.8940952 9246.58470 209.7428439 823.258426 23.0767063
ARS   2.0212434  209.74284  6.5824735 22.105093  0.7115212
II    7.3409259  823.25843  22.1050926 145.723426 -6.8745370
L     1.1260794  23.07671  0.7115212 -6.874537  21.2685053
```

5.2. Coefficiente di correlazione

Date due variabili X e Y, il coefficiente di correlazione tra le due variabili è definito come:

$$r_{xy} = \frac{C_{xy}}{s_x s_y}$$

Il coefficiente di correlazione r_{xy} risulta, quindi, essere il rapporto tra la covarianza tra X e Y ed il prodotto delle deviazioni standard delle singole variabili indicate con s_x e s_y . Se $r_{xy} > 0$ allora le variabili risultano essere correlate positivamente; se $r_{xy} < 0$ le variabili sono correlate negativamente; infine, se $r_{xy} = 0$ ciò indica che non c'è correlazione tra le variabili.

Inoltre per la correlazione diretta (e analogamente per quella inversa) si distingue:

- se $0 < r_{xy} < 0,3$ si ha correlazione debole;
- se $0,3 < r_{xy} < 0,7$ si ha correlazione moderata;
- se $r_{xy} > 0,7$ si ha correlazione forte.

L'indice di correlazione vale 0 se le due variabili sono indipendenti. Non vale la conclusione opposta: in altri termini, l'incorrelazione è condizione necessaria ma non sufficiente per l'indipendenza.

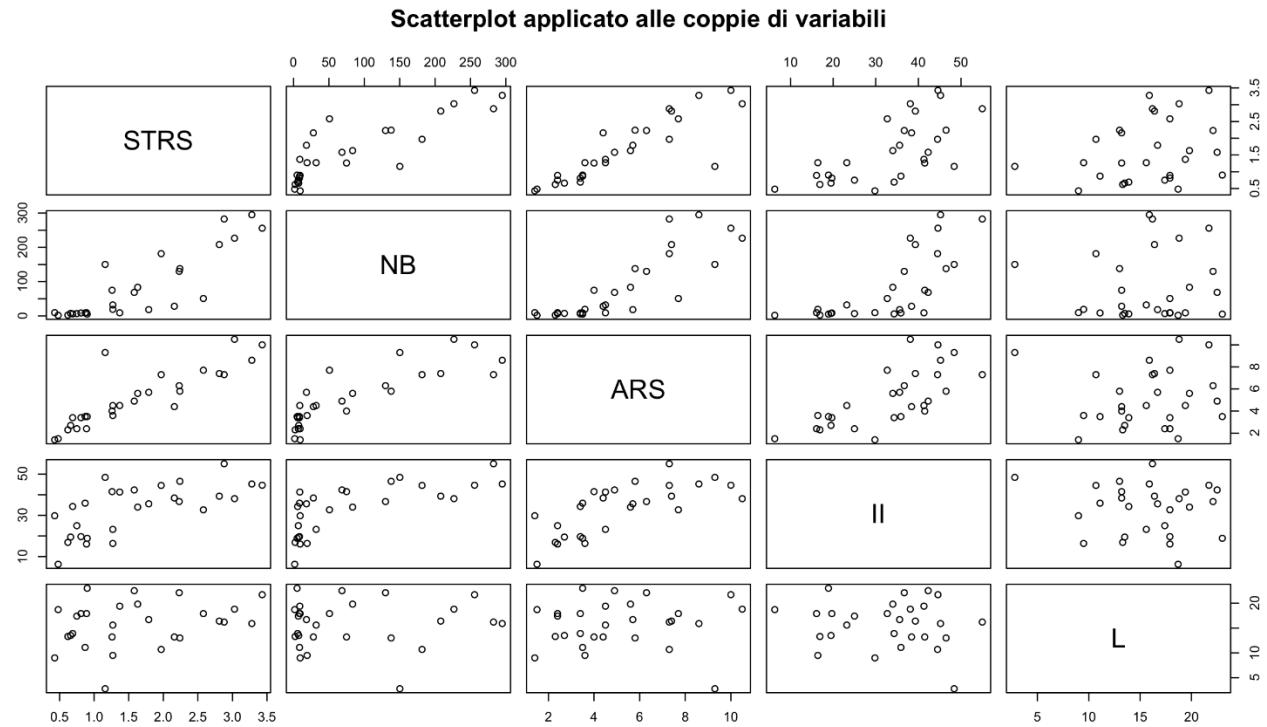
Il coefficiente di correlazione delle variabili utilizzate per l'analisi statistica è stato calcolato come segue:

```
> correlazionie <- cor(mytable)
> correlazione <- cor(mytable)
> correlazione
      STRS          NB          ARS          II          L
STRS 1.0000000 0.85662028 0.86647393 0.6688329 0.26855421
NB    0.8566203 1.00000000 0.85016244 0.7092198 0.05203729
ARS   0.8664739 0.85016244 1.00000000 0.7137286 0.06013465
II    0.6688329 0.70921975 0.71372859 1.0000000 -0.12348389
L     0.2685542 0.05203729 0.06013465 -0.1234839 1.00000000
```

5.3. Scatterplot

Come detto precedentemente, le relazioni tra le variabili possono essere messe a confronto mediante l'utilizzo di diagrammi di dispersione o scatterplot. Il risultato sarà una nuvola di punti in cui ogni punto corrisponde alla coppia di variabili (x_i, y_i) . Di seguito andiamo a vedere, tramite uno scatterplot, se esiste una qualche relazione tra le variabili prese in considerazione. Lo scatterplot è stato ottenuto nel seguente modo:

```
> pairs(mytable,main="Scatterplot applicato alle coppie di variabili")
```



5.3.1 Scatterplot relativo alla spesa per il settore Ricerca e Sviluppo e al numero di addetti nel settore Ricerca e Sviluppo.

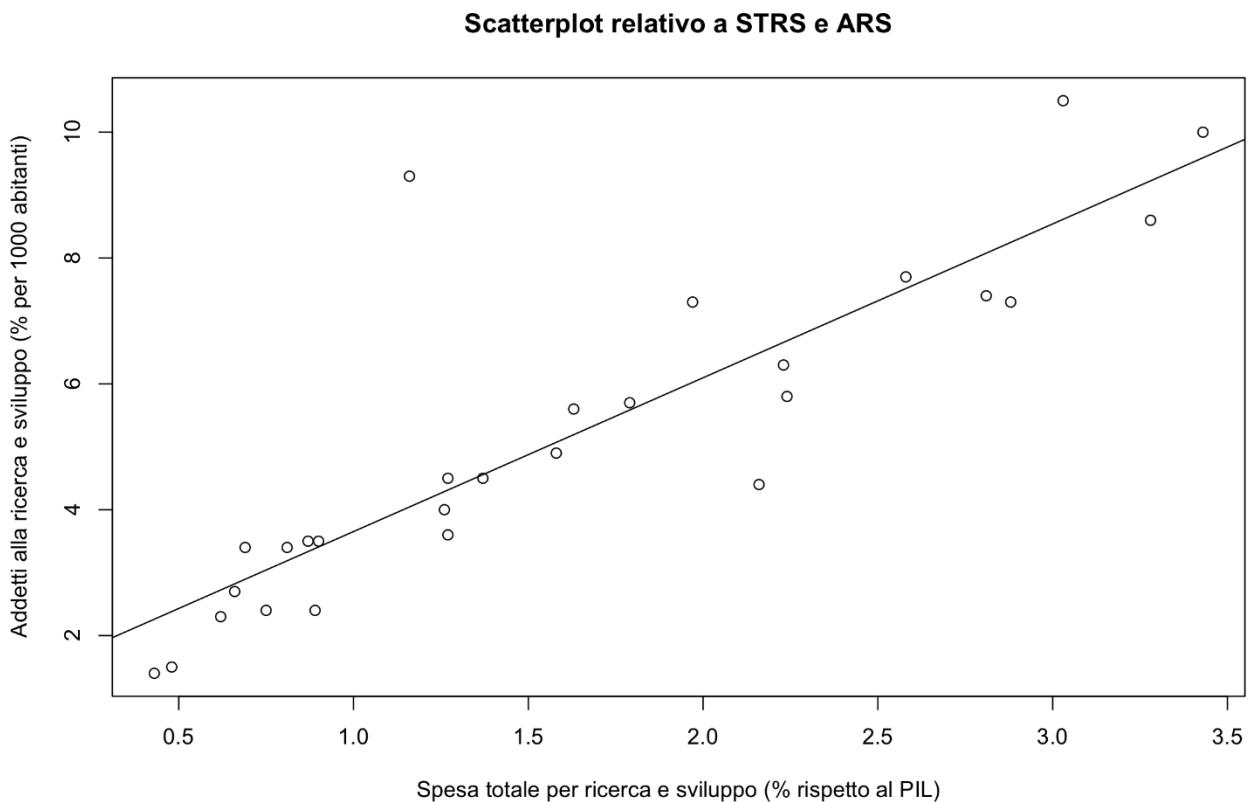
Com'è possibile notare dallo scatterplot precedente esiste una relazione tra Spesa totale per ricerca e sviluppo (% rispetto al PIL) e il numero di addetti nel settore Ricerca e Sviluppo (% per 1000 abitanti). Le variabili, come possiamo vedere nelle tabelle relative alla covarianze e al coefficiente di correlazione, sono correlate positivamente, infatti in corrispondenza della spesa per il settore Ricerca e Sviluppo e addetti nel settore Ricerca e Sviluppo abbiamo una covarianza di 2.0212434 ed un

coefficiente di correlazione pari a 0.86647393. Lo scatterplot relativo alle due variabili è ottenuto mediante l'esecuzione dei seguenti comandi:

```
plot(mytable[,1],mytable[,3],main="Scatterplot relativo a STRS e ARS", xlab=COLUMNS[1],ylab=COLUMNS[3])
abline(lm(mytable[,3]~mytable[,1]))
```

In particolare abline è utilizzato per aggiungere, in questa situazione una linea retta al grafico, la tilde per creare la formula (numero di addetti in ricerca e sviluppo in funzione della spesa in ricerca e sviluppo) ed infine lm per adattare modelli lineari (per effettuare la regressione).

Più formalmente, in statistica la regressione lineare rappresenta un metodo di stima del valore atteso condizionato di una variabile dipendente, Y , dati i valori di altre variabili indipendenti, X_1, X_2, \dots, X_k : $E[Y | X_1, X_2, \dots, X_k]$.



Dal grafico evince come i valori delle variabili STRS e ARS siano molto vicini alla retta d'interpolazione, proprio per la forte correlazione presente tra le due variabili.

6.ANALISI DEI CLUSTER

Questo paragrafo sarà concentrato sull'analisi dei cluster. L'analisi dei cluster è una metodologia che permette di raggruppare degli elementi in sottoinsiemi, detti appunto cluster, appartenenti ad un insieme più ampio. Lo scopo generale dei vari metodi attraverso i quali si attua l'analisi dei cluster è quello di ottenere raggruppamenti in base alla somiglianza, in modo che gli elementi appartenenti allo stesso gruppo siano più simili possibile, mentre quelli appartenenti a gruppi diversi differiscano il più possibile. In altre parole lo scopo è quello di ottenere un'alta omogeneità all'interno dei gruppi e un'alta eterogeneità tra gruppi distinti. Formalmente, il problema dell'analisi dei cluster, consiste nell'individuare m sottoinsiemi, detti cluster, di individui in $I = I_1, I_2, \dots, I_n$ con $m < n$, tali che l'individuo I_i appartenga ad un unico sottoinsieme. Gli individui assegnati allo stesso cluster sono detti simili, mentre quelli appartenenti a cluster differenti sono detti dissimili. La similarità tra due elementi I_i, I_j con $i \neq j$, può essere quantificata mediante l'utilizzo di un coefficiente di similarità $s_{ij} = s(X_i, X_j)$, oppure mediante l'utilizzo di una misura di distanza $d_{ij} = d(X_i, X_j)$. Un criterio per risolvere il problema del clustering è quello di: assegnare, ad esempio, due individui allo stesso cluster se il coefficiente di similarità delle caratteristiche è prossimo a 1, oppure se la distanza tra i punti degli individui risulta essere molto piccola; assegnare due individui a cluster differenti se, ad esempio, il coefficiente di similarità è prossimo a 0, oppure la distanza tra i punti degli individui risulta essere elevata. In questo paragrafo elencheremo le diverse misure di distanza che possono essere utilizzate. Per determinare i cluster e gli elementi appartenenti ad ognuno di essi, viene prodotta una matrice delle distanze a seconda della metrica scelta. Le opzioni disponibili per il calcolo della matrice delle distanze sono:

- Metrica euclidea: è la metrica più familiare ed è definita nel seguente modo

$$d_2(X_i, X_j) = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}}$$

essa è fortemente influenzata dall'unità di misura in base alla quale è valutata ciascuna delle p caratteristiche.

- Metrica di Manhattan:

$$d_1(X_i, X_j) = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

- Metrica di Chebycev:

$$d_\infty(X_i, X_j) = \max_{k=1,2,\dots,p} |x_{ik} - x_{jk}|$$

- Metrica di Minkowski:

$$d_r(X_i, X_j) = \left[\sum_{k=1}^p |x_{ik} - x_{jk}|^r \right]^{\frac{1}{r}}$$

include come caso particolare la metrica euclidea, quella di Manhattan e quella di Chebycev, a seconda del valore di r. Nello specifico: se r=1 otteniamo la metrica di Manhattan, se r=∞ quella di Chebycev e, infine, se r=2 abbiamo la metrica euclidea.

- Metrica di Canberra:

$$d_c(X_i, X_j) = \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{|x_{ik} + x_{jk}|}$$

La metrica più usata risulta essere quella euclidea che però è la più fortemente influenzata dall'unità di misura in base alla quale è valutata ciascuna delle p caratteristiche. Per ovviare a questo si ricorre alla standardizzazione dei dati, ossia da questi si sottrae la media campionaria relativa a quel carattere e si divide il valore ottenuto per la deviazione standard. La matrice dei dati standardizzati si ottiene in questo modo:

```

> #example value 1,1
> (mytable[1,1]-mean(mytable[,1]))/sd(mytable[,1])
[1] 1.321385
> mytableScaled <- scale(mytable)
> mytableScaled
      STRS        NB       ARS       II        L
AU  1.32138511  1.302639394  0.88115258  0.51567333  0.14016903
BE  0.69447350  0.574679649  0.25752485  1.11211477 -0.59707359
BU -1.08727526 -0.836522313 -1.10666082 -1.33992226 -0.53202277
CI -1.29624579 -0.762686396 -1.45745142 -0.27129802 -1.46441785
CR -0.94429542 -0.789724901 -1.06768409 -0.66892564  0.35700509
DA  1.56335099  1.497108640  2.08943132  0.41626642  0.66057558
ES  0.60648591 -0.568217150 -0.28814942  0.44111815 -0.55370638
FI  2.00328895  1.799731906  1.89454765  0.95472050  1.28940017
FR  0.68347505  0.491484250  0.45240851  0.30029170  1.37613459
GE  1.39837425  2.078436493  0.84217585  1.81624702  0.09680181
GR -1.01028612 -0.800124325 -0.67791676  0.10147788 -0.40192113
IR -0.03142414 -0.148080383 -0.09326575  0.76419059  1.46286902
IT -0.38337451 -0.081524063 -0.44405636  0.69791932 -0.55370638
LE -1.04328146 -0.783485246 -0.95075389 -1.12454063 -0.48865556
LI -0.77931868 -0.807403923 -0.63894002 -1.17424408  1.57128705
LU -0.49335900  0.701552634  1.62171052  1.26950904 -2.80880144
MA -0.81231403 -0.772045878 -0.63894002  0.23402043 -1.00906211
PA  0.39751538  1.028094576  0.84217585  0.94643659 -1.09579654
PL -0.79031713 -0.763726338 -1.06768409 -1.40619353  0.46542312
PR -0.26239157 -0.767886108 -0.24917269  0.68135151  0.79067722
RU  0.02356811  0.007910991  0.17957138  0.07662616  0.87741164
RC  0.19954329 -0.670131514  0.21854811  0.20916870  0.20521985
RO -1.24125355 -0.842761968 -1.41847469 -2.21801660  0.63889197
SC -0.87830473 -0.773085821 -0.67791676 -1.10797281  0.46542312
SV  1.06842077 -0.332150204  0.99808278 -0.03106466  0.46542312
SP -0.37237606 -0.527659392 -0.24917269 -0.81803600 -0.03329982
SZ  1.83831222  2.206349420  1.34887338  1.00442395  0.03175100
UN -0.37237606 -0.660772032 -0.59996329 -1.38134181 -1.35599981
attr("scaled:center")
      STRS        NB       ARS       II        L
1.608571 82.739286  5.139286 33.075000 15.753571
attr("scaled:scale")
      STRS        NB       ARS       II        L
0.9092191 96.1591633 2.5656332 12.0715958 4.6117790

```

Successivamente viene calcolata la matrice delle distanze dando in input la matrice dei dati scalata alla funzione dist nel seguente modo:

```
> distCityableScaled.method="euclidean")
   AU    BE    BU    CI    CR    DA    ES    FE    FR    GE    GR    IR    IT    LE    LT    LU    MA    PA    PL    PR    RU    RC    RO    SC    SV    SP    SZ
BE 1.469849
BU 4.620259 3.6116256
CI 4.475447 3.3782777 1.4776647
CR 3.841814 3.2226873 1.1281762 0.9737617
DA 3.683088 3.2226873 1.1281762 0.9737617
ES 2.4218748 1.4366945 2.5223889 2.0634698 3.5084393
FE 2.4218748 1.4366945 2.5223889 2.0634698 3.5084393
FR 1.6380658 3.1442882 2.8751198 2.0521134 2.2348052 3.3278393 2.4431508
GE 1.5174585 2.018771 3.3693787 5.3501699 1.9275908 2.0801015 4.4257877
GR 3.4830636 3.2226873 1.1281762 0.9737617
IR 2.587562 2.3530360 3.3208343 3.67174795 2.3398789 3.2831152 2.1891496 3.4571412 2.0002538 3.2938948 2.3752530
IT 2.6635269 3.5026691 2.3786143 2.0231321 1.9755386 3.7754722 1.1432181 4.2623167 2.4804338 3.3418961 1.1588254 2.0790163
LE 4.4989676 3.8493973 2.1838734 3.3120975 1.3932610 4.6477961 2.6681938 5.1362700 2.8613418 2.2498883 2.9567289 2.1007880
LT 4.4989676 3.8493973 2.1838734 3.3120975 1.3932610 4.6477961 2.6681938 5.1362700 2.8613418 2.2498883 2.9567289 2.1007880
LU 3.6568596 2.8643608 4.7089389 4.8568221 4.8148282 4.2249193 4.4987861 4.949491 4.6997078 3.859147 2.8688374 2.7396362 3.2181156 4.2249193 2.8567277 5.3104969
MA 3.5336256 2.8643608 4.7089389 4.8568221 4.8148282 4.2249193 4.4987861 4.949491 4.6997078 3.859147 2.8688374 2.7396362 3.2181156 4.2249193 2.8567277 5.3104969
PA 1.6291143 0.9548612 3.8761632 1.6431699 3.6704927 2.5495858 3.1028760 3.1576338 2.6551182 2.0697337 2.9715971 3.0037328 1.9624114 3.6372983 4.3079367 2.1323046 2.7226746
PL 2.4218748 1.4366945 2.5223889 2.0634698 3.5084393 2.0801015 4.4257877 2.0801015 4.4257877 2.0801015 4.4257877 2.0801015 4.4257877 2.0801015 4.4257877 2.0801015 4.4257877
PR 2.9195325 2.2552661 2.6936899 2.9192146 1.7739883 3.7443027 1.6313842 4.0798238 1.8681388 3.7175513 1.5280866 0.9594354 1.5268763 2.4495280 2.1158131 4.3686165 1.9731998 2.9122368 2.3264389
RU 2.1.422315 2.0.0859314 2.7558209 3.2589552 1.9887935 3.1844108 1.0218344 3.2818502 2.0234553 0.957792 1.7310732 3.5201295 2.0061564 4.2238541 2.3588816 2.5389522 2.2763937 1.1138542
RC 2.3.5580511 2.1.134651 2.2181344 2.6837671 1.5416954 5.7084256 3.1414887 6.0975956 2.6683710 3.6438800 3.4914493 1.6518240 1.6676720 6.0196683 3.0862441 4.9114147 1.0180375 3.2883077 1.1936372 3.2961857
RO 4.981756 3.3508511 1.1314651 2.2181344 2.6837671 1.5416954 5.7084256 3.1414887 6.0975956 2.6683710 3.6438800 3.4914493 1.6518240 1.6676720 6.0196683 3.0862441 4.9114147 1.0180375 3.2883077 1.1936372 3.2961857
SC 3.7841756 3.3508511 1.1314651 2.2181344 2.6837671 1.5416954 5.7084256 3.1414887 6.0975956 2.6683710 3.6438800 3.4914493 1.6518240 1.6676720 6.0196683 3.0862441 4.9114147 1.0180375 3.2883077 1.1936372 3.2961857
SV 1.4989676 3.8493973 2.1838734 3.3120975 1.3932610 4.6477961 2.6681938 5.1362700 2.8613418 2.2498883 2.9567289 2.1007880
SP 3.6594635 2.5794156 1.3648614 2.1715736 1.1136888 3.9147141 1.6785789 4.5388202 2.4254598 4.2514359 1.1414164 1.799321 1.3364949 1.5697327 2.9862704 1.2227893 1.7311576 1.5824129 1.2954512 2.1525804 0.9125828 2.1302146
SZ 1.2463145 2.3598247 5.4462445 5.5119397 5.0462863 1.3668959 3.5433788 1.0488012 2.7168413 1.0629853 4.7185441 3.6369595 3.7177358 2.282288 5.1964284 3.9893043 4.6394907 2.2598947 5.2587886 4.0617193 3.3271077 3.5916388 6.4985851 5.0015218 2.4709823
UN 3.4411831 3.1929614 1.6861071 2.8911012 3.2424532 5.3114224 3.7285446 5.0083763 4.8817428 3.6311195 2.3000293 1.1659425 2.9962992 3.9929992 1.7139243 3.3182367 1.0015913 2.8853735 2.4421857 2.4774276 1.9149327 3.1430806 4.9520928
```

Ottenuta la matrice delle distanze, si procede alla determinazione dei cluster mediante un algoritmo di raggruppamento che produrrà m partizioni degli n dati. Gli algoritmi di raggruppamento possono essere divisi in 3 gruppi:

- Metodi di enumerazione completa
- Metodi gerarchici
- Metodi non gerarchici

I metodi di enumerazione completa sono costosi dal punto di vista computazionale, in quanto considerano ogni possibile suddivisione degli n dati in m cluster. Con i metodi gerarchici, invece, non deve essere definito a priori il numero di cluster ed è possibile ottenere una visione completa dei dati in termini di distanza o similarità, ma non è possibile però riassegnare una variabile che è stata già assegnata ad un cluster. Ciò è, invece, possibile mediante l'uso dei metodi non gerarchici, che permettono di specificare o meno il numero di cluster prodotti dall'analisi.

6.1. Metodi gerarchici

I metodi gerarchici di clustering si dividono in agglomerativi e divisivi. I metodi agglomerativi partono da una situazione in cui si hanno n cluster distinti, ognuno contenente un solo elemento, per concludere con una situazione dove vi è un unico cluster contenente tutti gli individui, ottenuto mediante unioni successive di cluster meno distanti tra loro. I metodi divisivi, invece, partono da una situazione in cui vi è un unico cluster con tutti gli elementi, per concludere con n cluster distinti ottenuti mediante divisioni successive dei cluster più distanti tra loro. L'obiettivo dei metodi gerarchici è quello di ottenere una sequenza di partizioni che possono essere rappresentate graficamente mediante una struttura ad albero chiamata dendrogramma dove sulle ordinate sono riportati i livelli di distanza mentre sulle ascisse i singoli individui. Ad ogni livello di distanza corrisponde una partizione, mentre ad ogni partizione corrispondono infiniti livelli di distanza compresi tra quelli che individuano due successive unioni o divisioni. I metodi gerarchici sono caratterizzati da un algoritmo generale composto dai seguenti passi:

- Passo 1: a partire dalla matrice originaria dei dati o quella scalata, calcolare la matrice delle distanze (o la matrice di similarità) tra gli individui considerati come cluster contenenti un solo individuo
- Passo 2: individuare i cluster meno distanti (o più simili) ed unirli in un unico cluster calcolando successivamente la distanza del nuovo cluster dagli altri
- Passo 3: costruire una nuova matrice delle distanze che avrà una riga ed una colonna in meno in quanto due cluster vengono raggruppati in uno unico
- Passo 4: operare sulla matrice ottenuta a partire dal Passo 2 fino ad esaurire la possibilità di raggruppamento. Equivale a raggiungere una matrice 2×2 richiedendo $n-1$ iterazioni
- Passo 5: rappresentare graficamente il processo di agglomerazione attraverso un dendrogramma

I metodi gerarchici di distinguono tra loro per il modo in cui viene calcolata la distanza tra cluster. Abbiamo i seguenti metodi:

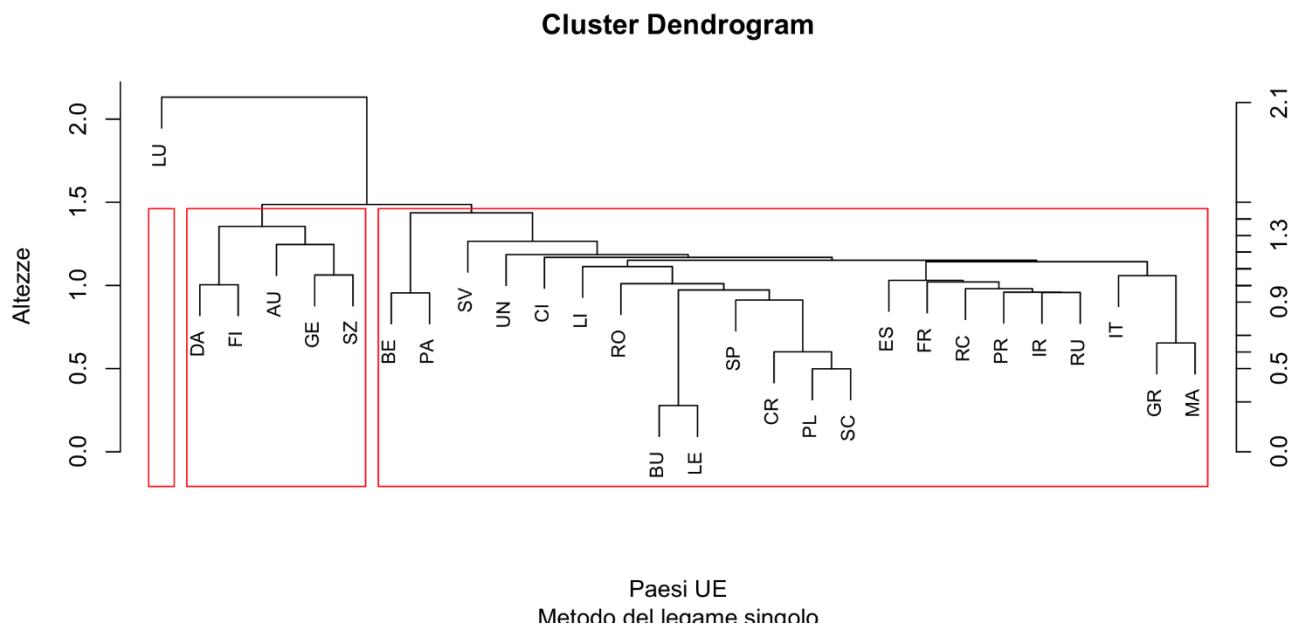
- Metodo del legame singolo
- Metodo del legame completo
- Metodo del legame medio
- Metodo del centroide
- Metodo della mediana

Nei sottoparagrafi successivi vengono mostrati i cluster ottenuti mediante i diversi metodi.

6.1.1 Metodo del legame singolo

Nel metodo del legame singolo la distanza tra cluster viene definita come la distanza minima tra tutte le distanze che collegano gli elementi dei cluster presi in considerazione. Il vantaggio del metodo è quello di riuscire a mettere in evidenza eventuali errori in modo migliore rispetto alle altre tecniche, ma presenta lo svantaggio di poter dare origine alla formazione di catene ponendo elementi dissimili nello stesso cluster. Di seguito i comandi:

```
> mytableScaled <- scale(mytable)
> d <- dist(mytableScaled,method="euclidean")
> c <- hclust(d,method="single")
> plot(c,xlab="Paesi UE",ylab="Altezze",sub="Metodo del legame singolo",cex=0.8)
> axis(side=4,at=round(c(0,c$height),1))
> rect.hclust(c,k=3,border="red")
```



Per ottenere una suddivisione in cluster in base ad un livello di altezza prefissato o ad un numero di cluster specifico viene utilizzata la funzione `cutree()`. Nel caso specifico abbiamo:

```
> cutree(c,k=3)
AU BE BU CI CR DA ES FI FR GE GR IR IT LE LI LU MA PA PL PR RU RC RO SC SV SP SZ UN
 1 2 2 2 2 1 2 1 2 1 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 1 2
```

6.1.2 Metodo del legame completo

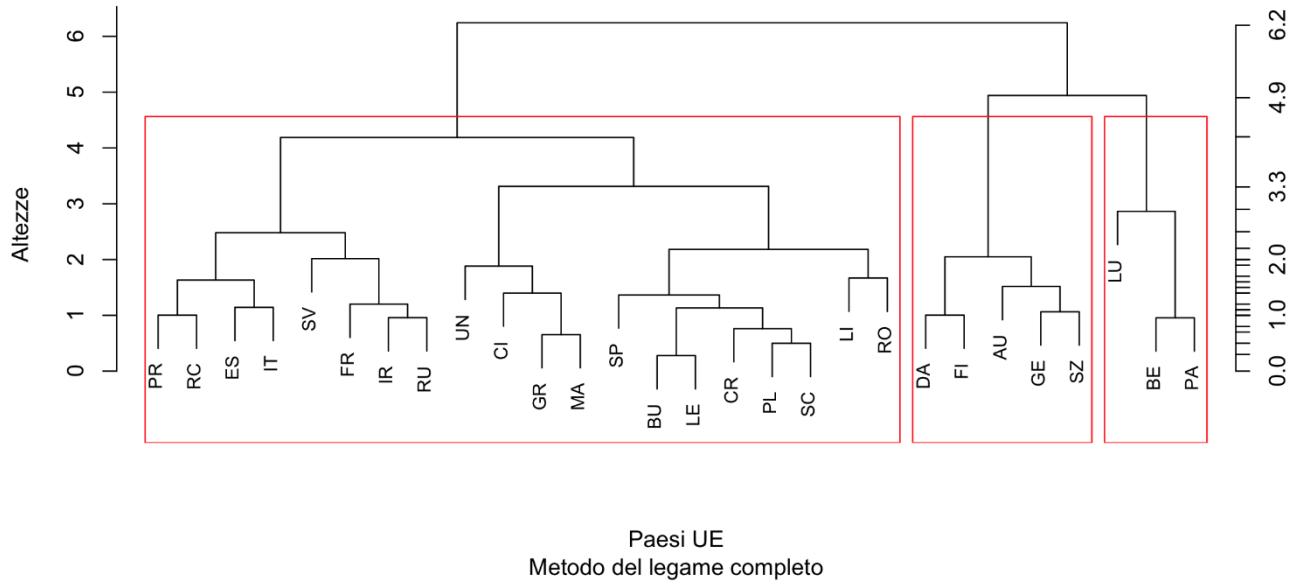
Con il metodo del legame completo la distanza tra due cluster viene definita come la distanza massima tra le distanze che possono essere calcolate tra ogni individuo di un cluster e ogni individuo dell'altro. Il metodo del legame completo individua cluster di forma elissoidale, ossia una serie di punti che si addensano intorno ad un nucleo centrale, privilegiando l'omogeneità tra gli elementi del cluster a scapito della differenziazione tra cluster.

```

> mytableScaled <- scale(mytable)
> d <- dist(mytableScaled,method="euclidean")
> c <- hclust(d,method="complete")
> plot(c,xlab="Paesi UE",ylab="Altezze",sub="Metodo del legame completo",cex=0.8)
> axis(side=4,at=round(c(0,c$height),1))
> rect.hclust(c,k=3,border="red")

```

Cluster Dendrogram



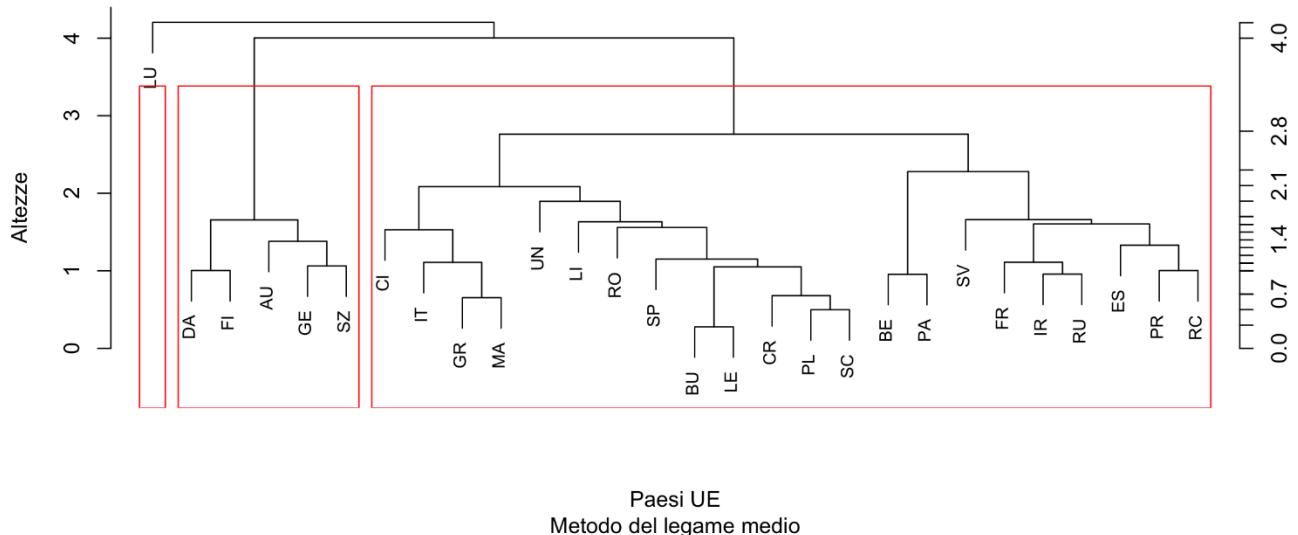
```
> cutree(c,k=3)
AU BE BU CI CR DA ES FI GE GR IR IT LE LI LU MA PA PL PR RU RC RO SC SV SP SZ UN
 1 2 3 3 3 1 3 1 3 1 3 3 3 3 3 2 3 2 3 3 3 3 3 3 3 3 3 1 3
```

6.1.3. *Metodo del legame medio*

In questo metodo la distanza tra due cluster viene definita mediante la media aritmetica delle distanze tra tutte le coppie di unità che compongono i due cluster. Uno svantaggio di questo metodo è che nel caso in cui si considerano cluster molto differenti, la distanza sarà molto vicina a quella del cluster più numeroso.

```
> mytableScaled <- scale(mytable)
> d <- dist(mytableScaled,method="euclidean")
> c <- hclust(d,method="average")
> plot(c,xlab="Paesi UE",ylab="Altezze",sub="Metodo del legame medio",cex=0.8)
> axis(side=4,at=round(c(0,c$height),1))
> rect.hclust(c,k=3,border="red")
```

Cluster Dendrogram



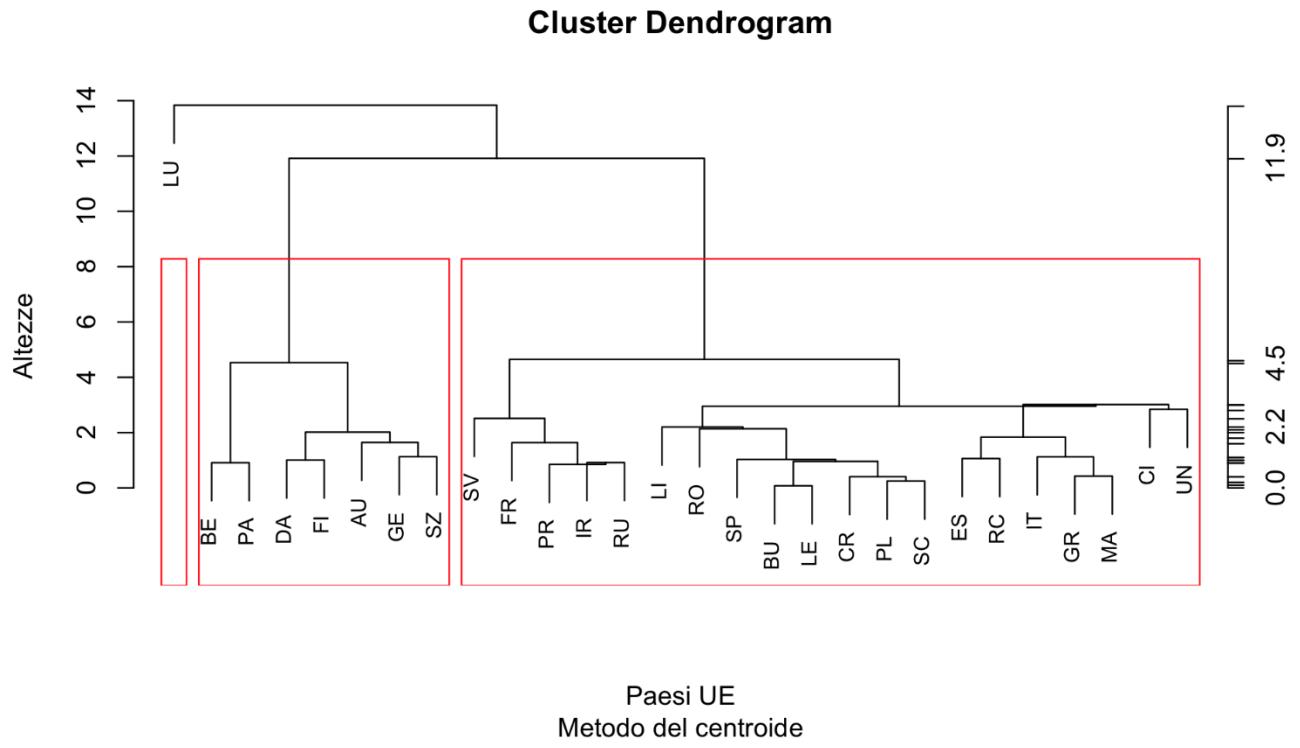
```
> cutree(c,k=3)
```

AU	BE	BU	CI	CR	DA	ES	FI	FR	GE	GR	IR	IT	LE	LI	LU	MA	PA	PL	PR	RU	RC	RO	SC	SV	SP	SZ	UN
1	2	2	2	2	1	2	1	2	1	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	1	2

6.1.4 Metodo del centroide

In questo metodo la distanza tra due cluster è definita come la distanza tra i centroidi, ossia tra le medie campionarie calcolate sugli individui appartenenti ai due cluster. Il metodo del centroide può dare origine a fenomeni gravitazionali, per cui i gruppi grandi tendono ad attrarre al loro interno i piccoli gruppi. Inoltre le distanze in cui si verificano le successive agglomerazioni possono essere non crescenti. Uno svantaggio di questo metodo è che se le misure dei due cluster da unire sono molto diverse, il centroide del nuovo cluster sarà molto vicino a quello del cluster più numeroso.

```
> mytableScaled <- scale(mytable)
> d <- dist(mytableScaled,method="euclidean")
> d2 <- d^2
> c <- hclust(d2,method="centroid")
> plot(c,xlab="Paesi UE",ylab="Altezze",sub="Metodo del centroide",cex=0.8)
> axis(side=4,at=round(c(0,c$height),1))
> rect.hclust(c,k=3,border="red")
```

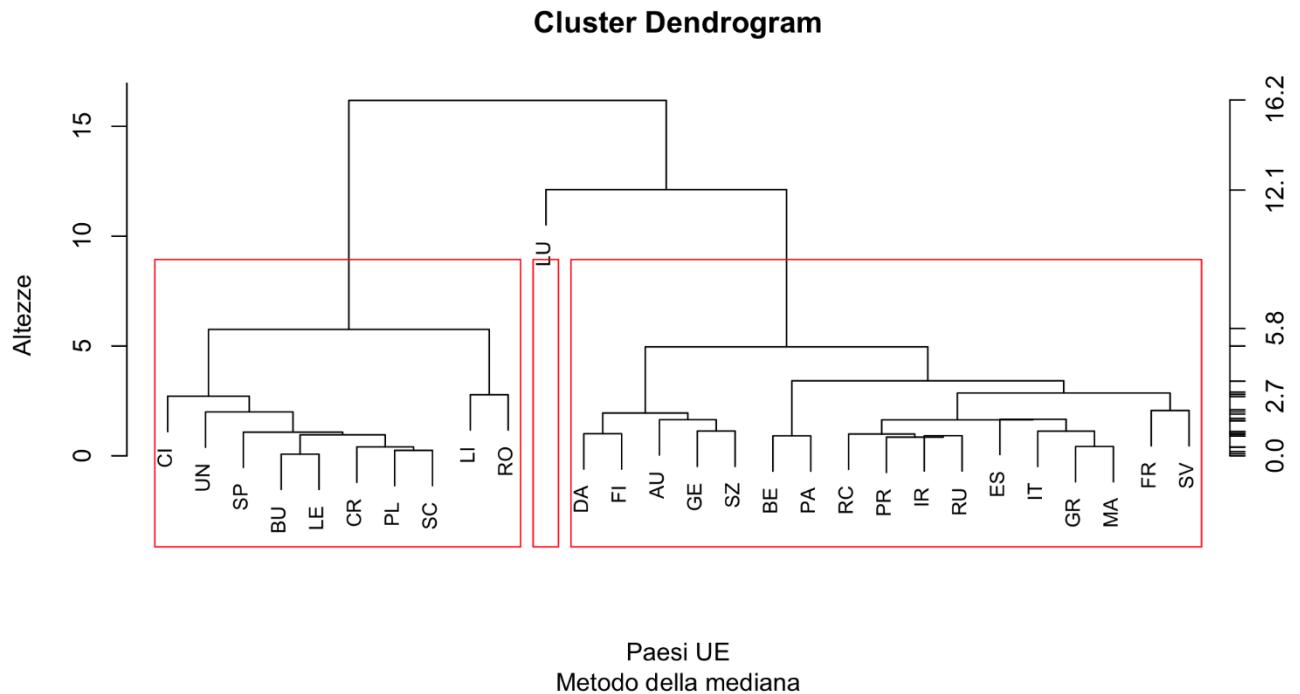


```
> cutree(c,k=3)
AU BE BU CI CR DA ES FI FR GE GR IR IT LE LI LU MA PA PL PR RU RC RO SC SV SP SZ UN
 1 1 2 2 2 1 2 1 2 1 2 2 2 2 2 2 3 2 1 2 2 2 2 2 2 2 1 2
```

6.1.5 Metodo della mediana

Simile al metodo del centroide, il metodo della mediana si differenzia da esso in quanto risulta essere più indipendente dalla numerosità dei cluster. Infatti, quando due cluster si aggregano, il nuovo centroide risulta essere la semisomma dei due centroidi precedenti. Come il metodo del legame singolo può comportare la formazione di una catena tra gli individui.

```
> mytableScaled <- scale(mytable)
> d <- dist(mytableScaled,method="euclidean")
> d2 <- d^2
> c <- hclust(d2,method="median")
> plot(c,xlab="Paesi UE",ylab="Altezze",sub="Metodo della mediana",cex=0.8)
> axis(side=4,at=round(c(0,c$height),1))
> rect.hclust(c,k=3,border="red")
```



```
> cutree(c,k=3)
```

AU	BE	BU	CI	CR	DA	ES	FI	GE	GR	IR	IT	LE	LI	LU	MA	PA	PL	PR	RU	RC	RO	SC	SV	SZ	UN
1	1	2	2	2	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	2	2	1	2	

6.1.6 Misure di non omogeneità statistiche

Dopo aver effettuato il taglio, siamo interessati a calcolare le misure di non omogeneità statistica relative all'insieme totale di individui (tr T), ai singoli cluster ottenuti effettuando il taglio e alla somma delle loro misure di non omogeneità (tr S) e alla misura di omogeneità tra i cluster (tr B):

$$\text{trT} = \text{trS} + \text{trB}$$

Poichè per ogni fissata matrice X dei dati si ha che la tr T è fissata, i cluster dovrebbero essere individuati in modo da minimizzare la misura di non omogeneità statistica all'interno dei cluster (within) e massimizzare la misura di non omogeneità statistica tra i gruppi (between).

Quindi, se due differenti metodi gerarchici conducono a due diverse partizioni con lo stesso numero di cluster, occorre scegliere quella partizione con misura di non omogeneità statistica all'interno dei cluster (tr S) più piccola, che corrisponde a maggiore omogeneità interna.

La funzione utilizzata per calcolare le misure di non omogeneità è la seguente:

```

homogeneity <- function(dataframe,metodo,myCut=3){

  #NB quando in un cluster ci sta un solo elemento da problemi
  n <- nrow(dataframe)
  #misura di non omogeneità statistica totale

  #In R utilizzando le funzione apply(X, 2, mean), apply(X, 2, var) e apply(X, 2, sd) è possibile calcolare la media campionaria, la varianza campionaria e la deviazione standard campionaria delle colonne di una matrice X
  trHI <- (n-1)*sum(apply(dataframe,2,var))

  print(c("Misura di omogeneità statistica totale: ",trHI))

  #misure di non omogeneità tra i gruppi

  dati <- scale(dataframe)
  matriceDistanze <- dist(dati,method="euclidean", diag=TRUE, upper=TRUE)

  if (metodo=="centroid")
    matriceDistanze <- matriceDistanze^2

  tree <- hclust(matriceDistanze,method=metodo)
  taglio <- cutree(tree,k=myCut,h=NULL)
  num <- table(taglio)
  tagliolist<-list(taglio)

  #Splits the data into subsets, computes summary statistics for each, and returns the result in a convenient form - by sono le suddivisioni e l'argomento
  #successivo è la funzione (nel nostro caso la varianza)
  avgvar <- aggregate(dataframe,tagliolist, var)[-1]
  #misura di omogeneità dei vari gruppi in base a quante partizioni sono le partizione
  sum<-0

  for(i in 1:(myCut)){
    value <- (num[i]-1)*sum(avgvar[i,])
    sum <- sum+value
    print(c("Misura omogeneità cluster",value))
  }

  print(c("Somma (within): ",sum))
  between <- trHI-sum
  print(c("Between: ",between))
  print(c("Between/total", between/trHI))

}

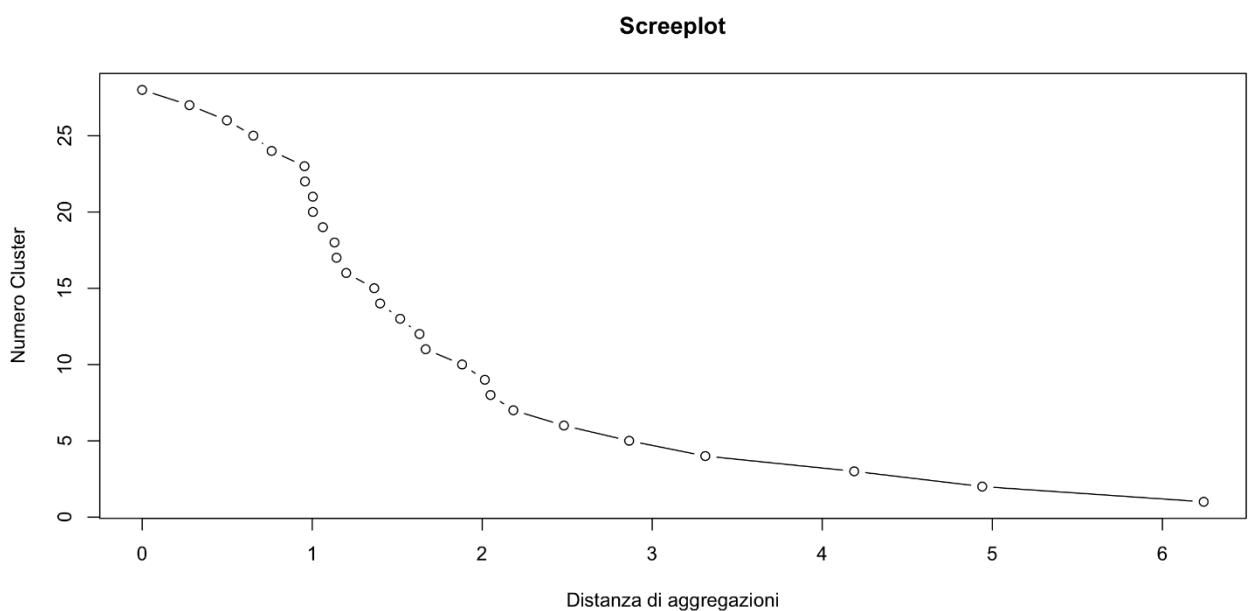
> homogeneity(mytable,"single",3)
[1] "Misura di omogeneità statistica totale: " "254366.616057143"
[1] "Misura omogeneità cluster" "5566.42172"
[1] "Misura omogeneità cluster" "58236.8440772727"
[1] "Misura omogeneità cluster" NA
[1] "Somma (within): " "63803.2657972727"
[1] "Between: " "190563.35025987"
[1] "Between/total" "0.749168083507706"
> homogeneity(mytable,"complete",3)
[1] "Misura di omogeneità statistica totale: " "254366.616057143"
[1] "Misura omogeneità cluster" "5566.42172"
[1] "Misura omogeneità cluster" "1083.5718"
[1] "Misura omogeneità cluster" "25555.13722"
[1] "Somma (within): " "32205.13074"
[1] "Between: " "222161.485317143"
[1] "Between/total" "0.873390890521714"
> homogeneity(mytable,"average",3)
[1] "Misura di omogeneità statistica totale: " "254366.616057143"
[1] "Misura omogeneità cluster" "5566.42172"
[1] "Misura omogeneità cluster" "58236.8440772727"
[1] "Misura omogeneità cluster" NA
[1] "Somma (within): " "63803.2657972727"
[1] "Between: " "190563.35025987"
[1] "Between/total" "0.749168083507706"
> homogeneity(mytable,"centroid",3)
[1] "Misura di omogeneità statistica totale: " "254366.616057143"
[1] "Misura omogeneità cluster" "19152.4404"
[1] "Misura omogeneità cluster" "25555.13722"
[1] "Misura omogeneità cluster" NA
[1] "Somma (within): " "44707.57762"
[1] "Between: " "209659.038437143"
[1] "Between/total" "0.824239602220613"
> homogeneity(mytable,"median",3)
[1] "Misura di omogeneità statistica totale: " "254366.616057143"
[1] "Misura omogeneità cluster" "223630.691817391"
[1] "Misura omogeneità cluster" "600.8366"
[1] "Misura omogeneità cluster" NA
[1] "Somma (within): " "224231.528417391"
[1] "Between: " "30135.0876397516"
[1] "Between/total" "0.118471079683593"

```

6.2 Screeplot metodo legame completo

Al fine di scegliere una buona partizione del dendrogramma, si può costruire lo screeplot in cui si pongono sull'asse delle ordinate i numeri di cluster ottenibili con il metodo gerarchico e sull'asse delle ascisse le distanze a cui avvengono le successive aggregazioni tra i cluster. Se nel passaggio da k cluster a k-1 cluster vi è un forte incremento della distanza di aggregazione è consigliabile tagliare il dendrogramma a k gruppi. Di seguito riportiamo lo screeplot utilizzando come metodo gerarchico quello del legame completo. Il grafico è costruito tramite i seguenti comandi:

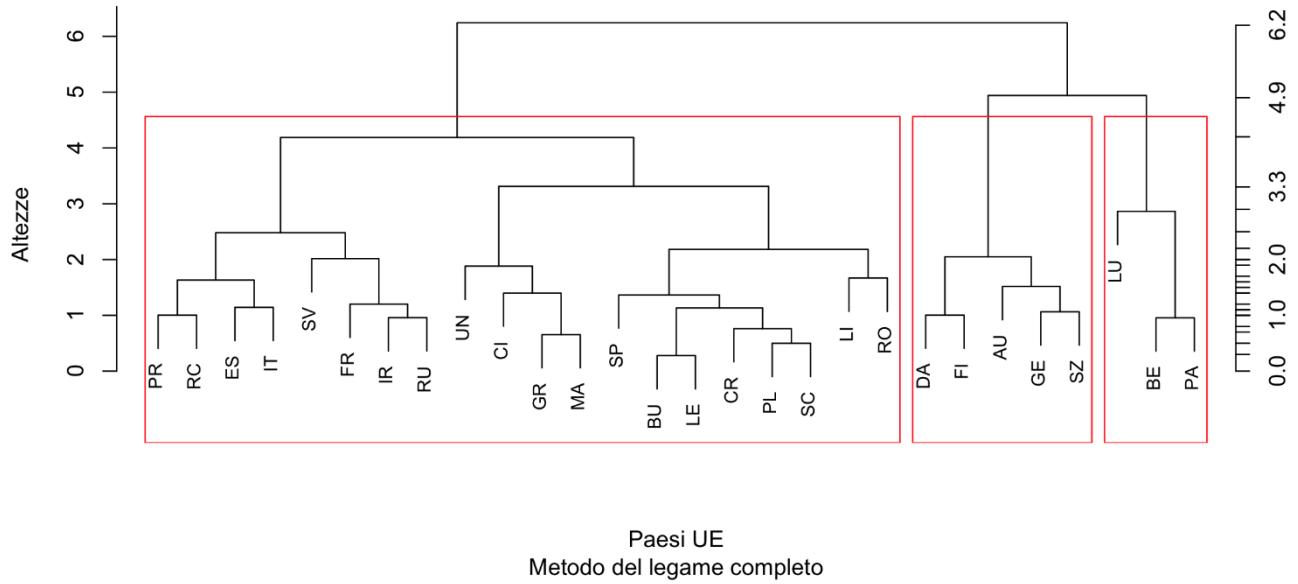
```
> z <- scale(myttable)
> d <- dist(z,method="euclidean")
> c <- hclust(d,method="complete")
> plot(rev(c$height)),seq(1,28),type="b",main="Screeplot",xlab="Distanza di aggregazione", ylab="Numero Cluster")
```



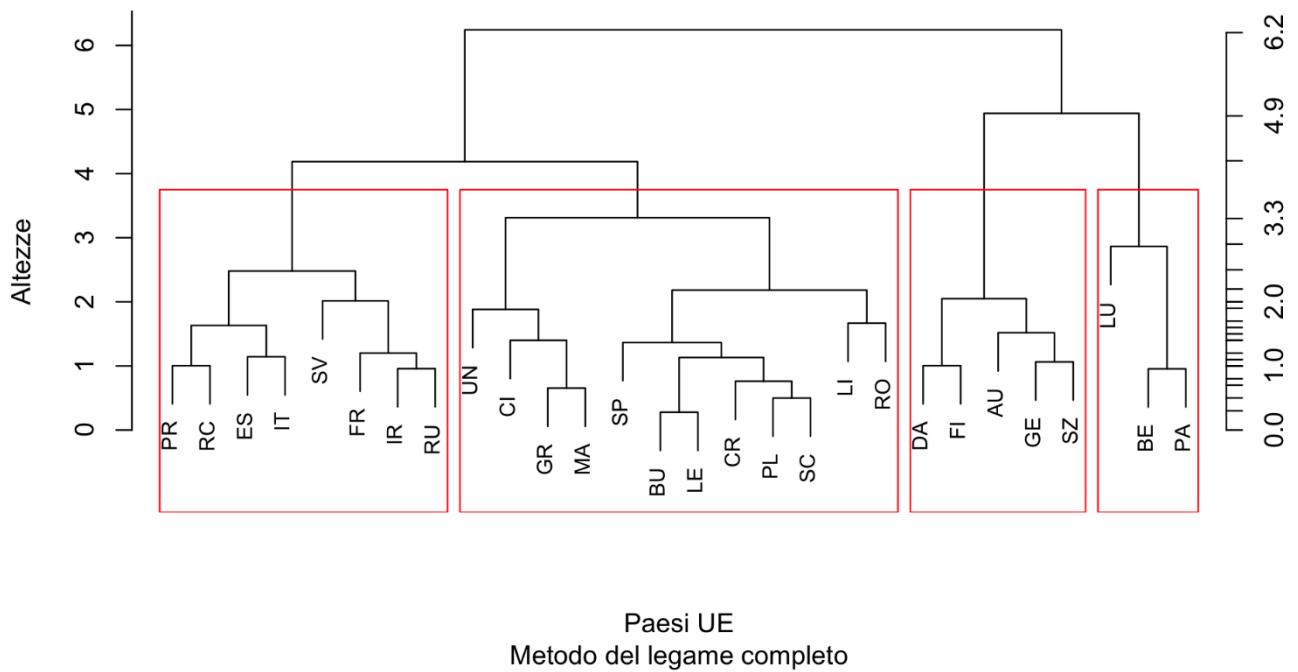
Alla luce di quanto detto quindi, realizzando tre cluster, tagliamo il dendrogramma ad una distanza di aggregazione vicino a 4. È interessante notare, però, come tra 3 e 4 vi sia una linea molto lunga e convenga quindi aggiungere un ulteriore partizione al fine di evitare che l’”insieme più grande” contenga nazioni troppo distanti tra loro.

Le figure successive mostrano la differenza con tre e quattro cluster:

Cluster Dendrogram



Cluster Dendrogram



6.2.1 Metodi non gerarchici

A differenza dei metodi gerarchici, nei metodi non gerarchici è possibile riallocare elementi già classificati a livelli precedenti di analisi. Gli algoritmi di tipo non gerarchico procedono, data una prima partizione, a riallocare gli individui nel gruppo con centroide più vicino, fino a che per nessuno individuo si verifica che esso sia

assegnato ad una partizione diversa da quella a cui apparteneva all'iterazione precedente. Il metodo più utilizzato è l'algoritmo k-means composto dai seguenti passi:

- Passo 1: Fissare a priori il numero k di cluster specificando k punti di riferimento che inducono ad una prima partizione
- Passo 2: Attribuire ogni unità al cluster per cui la distanza sia minore
- Passo 3: Calcolare il centroide di ognuno dei k cluster ottenuti che diventano i nuovi punti di riferimento
- Passo 4: Valutare la distanza di ogni unità da ogni centroide ottenuto al passo precedente. Se la distanza minima non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, allora si procede a spostare l'individuo presso il cluster che ha il centroide più vicino.
- Passo 5: Ricalcolare i centroidi dei cluster ottenuti.
- Passo 6: Ripetere il procedimento a partire dal Passo 4 fino a quando i centroidi non subiscono ulteriori variazioni rispetto all'iterazione precedente, raggiungendo così una configurazione stabile.

I vantaggi del metodo del k-means sono:

- Velocità di esecuzione dei calcoli
- Libertà degli individui di raggrupparsi e allontanarsi

Gli svantaggi riscontrabili si possono riassumere come segue:

- La classificazione è influenzata dalla scelta iniziale dei k vettori delle caratteristiche, dall'ordine in cui sono presi tali vettori e dalle proprietà geometriche del vettore
- L'algoritmo potrebbe convergere ad un ottimo locale e non globale, quindi la partizione finale dipende dalla scelta iniziale

Al fine di evitare ottimi locali, le partizioni generate possono essere osservate a partire da tre diverse configugazioni iniziali:

- Scelta casuale dei punti di riferimento
- Ripetizione della procedura di scelta casuale dei punti di riferimento
- Scelta dei centroidi come punti di riferimento

6.2.2. Scelta casuale dei punti di riferimento

Di seguito applicheremo il metodo k-means utilizzando come configurazione iniziale una configurazione dove i punti di riferimento sono scelti casualmente.

```

> k <- kmeans(mytable, centers=3, iter.max=50, nstart=100)
> k
K-means clustering with 3 clusters of sizes 16, 6, 6

Cluster means:
      STRS       NB       ARS       II        L
1 1.096250 13.8750 3.556250 25.62500 15.50000
2 2.900000 241.6000 8.516667 44.45000 16.61667
3 1.683333 107.5167 5.983333 41.56667 15.56667

Clustering vector:
AU BE BU CI CR DA ES FI FR GE GR IR IT LE LI LU MA PA PL PR RU RC RO SC SV SP SZ UN
 2 3 1 1 1 2 1 2 3 2 1 3 3 1 1 3 1 2 1 1 3 1 1 1 1 1 2 1

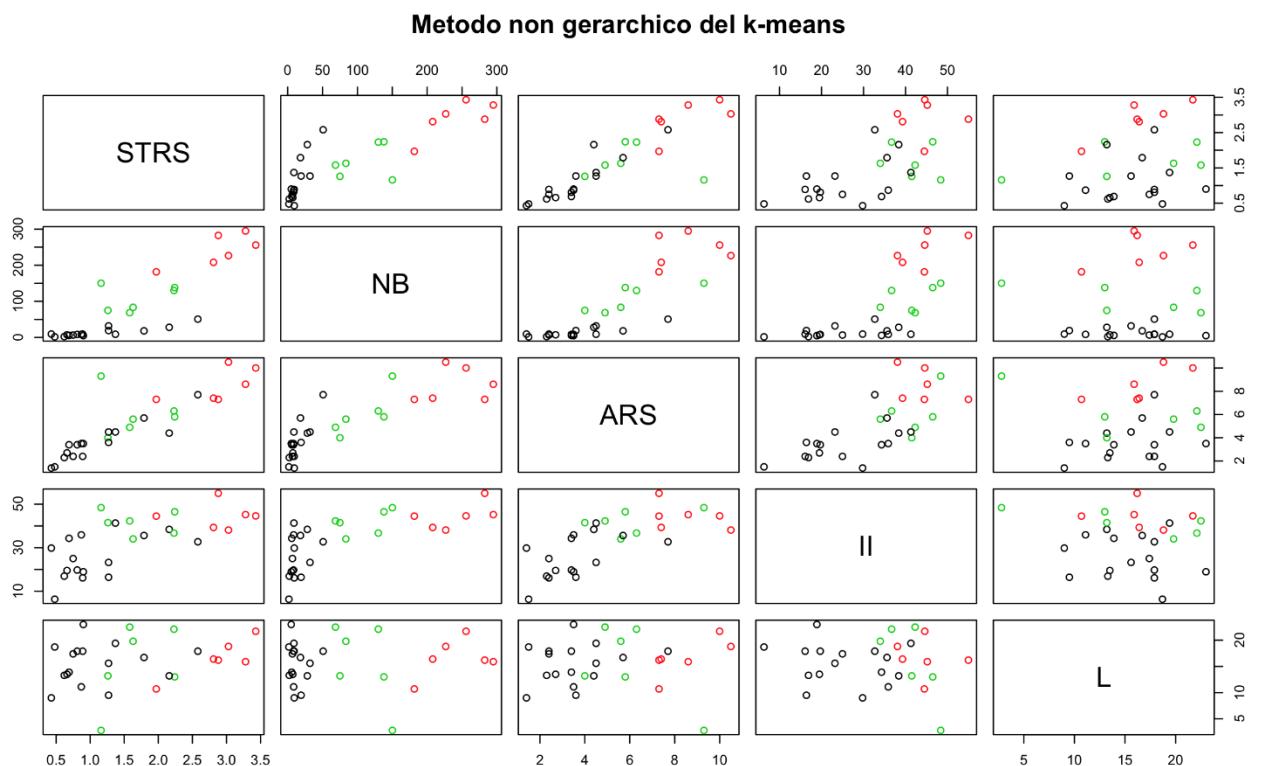
Within cluster sum of squares by cluster:
[1] 4351.689 9931.247 6873.379
(between_SS / total_SS =  91.7 %)

Available components:

[1] "cluster"      "centers"       "totss"         "withinss"      "tot.withinss" "betweenss"     "size"          "iter"          "ifault"

```

Rappresentando graficamente i cluster otteniamo:



6.2.3. Scelta dei centroidi come punti di riferimento

Di seguito applicheremo il metodo del k-means scegliendo noi i centroidi iniziali. È possibile notare come il risultato ottenuto sia il medesimo di quello ottenuto nel paragrafo 6.2.1.

```

> mytableScaled <- scale(mytable)
> d <- dist(mytableScaled,method="euclidean")
> tree <- hclust(d,method="centroid")
> taglio <- cutree(tree,k=3,h=NULL)
> tagliolist<-list(taglio)
> centroidiIniziali <- aggregate(mytable,tagliolist,mean)[,-1]
> k <- kmeans(mytable,centers=centroidiIniziali,iter.max=50)
> k
K-means clustering with 3 clusters of sizes 6, 16, 6

Cluster means:
      STRS      NB      ARS      II      L
1 2.900000 241.6000 8.516667 44.45000 16.61667
2 1.096250 13.8750 3.556250 25.62500 15.50000
3 1.683333 107.5167 5.983333 41.56667 15.56667

Clustering vector:
AU BE BU CI CR DA ES FI FR GE GR IR IT LE LI LU MA PA PL PR RU RC RO SC SV SP SZ UN
1 3 2 2 2 1 2 1 3 1 2 3 3 2 2 3 2 1 2 2 3 2 2 2 2 2 1 2

Within cluster sum of squares by cluster:
[1] 9931.247 4351.689 6873.379
(between_SS / total_SS =  91.7 %)

Available components:

[1] "cluster"      "centers"       "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"

```