Giardinetto Giuseppe matricola n°. 0512114655

Progetto di CPSM

Famiglie con accesso ad Internet per regioni italiane

Contents

Giardinetto Giuseppe matricola n°. 0512114655

1. Introduzione

Lo scopo dell'indagine statistica presentata in questo documento è analizzare i dati forniti dall'Istituto Nazionale di Statistica (ISTAT) sull'accesso ad internet e sul non accesso ad Internet e le sue cause, i dati sono divisi per regione. I dati presentati sono in percentuale, di seguito il Dataset usato

1.1 Dataset

La fonte del dataset utilizzato è il sito dell'Istituto nazionale di statistica. Il dataset colleziona dati raccolti su campioni di migliaia di famiglie per ciascuna regione riguardo il loro accesso ad Internet. Per ciascuna regione è indicata, nell'ordine di presentazione: Famiglie che dispongono di accesso a Internet da casa, Accede a Internet da altro luogo, Internet non è utile,non è interessante. Alto costo del collegamento, Nessuno sa usare Internet. I dati risultano essere relativi all'anno 2022.

Di seguito la tabella.

Tempo	2022				
Indicatore	Famiglie che dispongono di accesso a Internet da casa	Accede a Internet da altro luogo	Internet non è utile,non è interessante	Alto costo del collegamento	Nessuno sa usare internet
Territorio					
Piemonte	83,3	10,0	26,8	5,8	57,1
Valle d'Aosta / Vallée d'Aoste	79,4	12,6	27,5	1,0	50,2
Liguria	82,9	4,6	24,7	9,5	57,0
Lombardia	86,1	6,6	19,6	6,2	67,2
Trentino Alto Adige / Südtirol	88,9	5,8	33,0	5,8	51,8
Veneto	83,8	7,6	22,6	8,1	61,5
Friuli-Venezia Giulia	84,7	10,5	33,5	5,6	54,3
Emilia-Romagna	83,8	4,1	21,7	6,5	64,4
Toscana	84,3	7,1	24,4	7,5	54,9
Umbria	82,7	13,1	21,9	2,9	57,3
Marche	84,2	5,7	23,2	4,0	62,9
Lazio	84,6	12,9	15,0	9,8	50,6
Abruzzo	82,3	9,7	18,3	3,2	61,5
Molise	80,5	7,8	18,7	8,5	66,1
Campania	82,0	8,7	15,7	10,6	59,5
Puglia	78,2	6,5	23,4	9,4	62,5
Basilicata	77,5	8,1	33,6	10,0	45,1
Calabria	73,6	7,0	15,8	10,6	57,8
Sicilia	80,2	4,3	23,6	5,4	63,1
Sardegna	81,6	15,1	23,1	5,8	56,5

Per la realizzazione del progetto si è scelto di utilizzare il **linguaggio di programmazione** e ambiente di sviluppo **R** poiché adatto allo specifico obiettivo dell'analisi statistica dei dati.

1.1.1 Inizializzazione dati matrice

L'azione preliminare necessaria per iniziare la nostra analisi è quella di inizializzare la matrice che utilizzeremo durante la nostra indagine statistica. Per ciascuna colonna della tabella creiamo un vettore:

```
InternetACasa <- c(83.3,79.4,82.9,86.1,88.9,83.8,84.7,83.8,
                   84.3,82.7,84.2,84.6,82.3,80.5,82.0,78.2,77.5,73.6,80.2,81.6)
InternetDaUnAltroLuogo <- c(10.0,12.6,4.6,6.6,5.8,7.6,10.5,4.1,7.1,13.1,
                            5.7,12.9,9.7,7.8,8.7,6.5,8.1,7.0,4.3,15.1)
InternetNonInteressa <- c(26.8,27.5,24.7,19.6,33.0,22.6,33.5,21.7,24.4,21.9,
                          23.2,15.0,18.3,18.7,15.7,23.4,33.6,15.8,23.6,23.1)
AltoCostoCollegamento <- c(5.8, 1.0, 9.5, 6.2, 5.8, 8.1, 5.6, 6.5, 7.5, 2.9, 4.0,
                           9.8, 3.2, 8.5, 10.6, 9.4, 10.0, 10.6, 5.4, 5.8)
NonSaUsareInternet <- c(57.1,50.2,57.0,67.2,51.8,61.5,54.3,64.4,54.9,
                        57.3,62.9,50.6,61.5,66.1,59.5,62.5,45.1,57.8,63.1,56.5)
```

A questo punto è possibile costruire la matrice utilizzando gli array appena creati per definire le colonne della matrice:

```
matriceFamiglie <- cbind(InternetACasa, InternetDaUnAltroLuogo,
          InternetNonInteressa, AltoCostoCollegamento, NonSaUsareInternet)
```

Per rendere la matrice più leggibile associamo alle righe e alle colonne dei nomi, creiamo quindi 2 array: uno per le righe:

```
regioni<- c("Piemonte", "Valle d'Aosta", "Liguria", "Lombardia",
"Trentino Alto Adige", "Veneto", "Friuli-Venezia Giulia", "Emilia-Romagna",
"Toscana", "Umbria", "Marche", "Lazio", "Abruzzo", "Molise", "Campania",
"Puglia", "Basilicata", "Calabria", "Sicilia", "Sardegna")
```

E uno per le colonne:

```
categorie <- c("Famiglie che dispongono di accesso a Internet da casa",
               "Accede a Internet da altro luogo",
               "Internet non è utile, non è interessante",
               "Alto costo del collegamento",
               "Nessuno sa usare internet")
```

E le associamo alla matrice:

```
rownames(matriceFamiglie) <- regioni
colnames(matriceFamiglie) <- categorie
```

2. Distribuzioni di Frequenza

Il primo passo da seguire ora è di rendere ***significativi*** i dati prima elencati rendendoli anche più facilmente leggibili. Il primo metodo di cui ci serviamo è la **distribuzione di frequenza** che attraverso una vista tabulare (che può essere un istogramma o un diagramma a torta) riesce a rappresentare i dati significativi di un dataset. Attraverso queste rappresentazioni possiamo effettuare una lettura immediata del dataset.

Prima di costruire una tabella di frequenza si deve conoscere la natura dei dati con i quali si sta avendo a che fare. Considerando la tipologia di dati nel dataset preso in analisi, ad esempio, la scelta migliore è quella di dividere i vari insiemi in classi. Nel nostro dataset, infatti, è immediato notare che tutti i dati a nostra disposizione sono percentuali numeriche e il contenuto di ciascuna colonna della tabella non può assumere valori classificabili all'interno di una ben precisa modalità. La soluzione ideale è quella di raccogliere le informazioni in classi e poi calcolare le frequenze con cui gli elementi del campione cadono in ciascuna di esse.
La scelta del range da usare non è standard e varia di caso in caso. Questa dev'essere effettuata in modo corretto in quanto con un range troppo largo verrebbe sacrificata la rappresentazione utile e significativa dei dati, mentre un range troppo stretto comporterebbe un grafico poco leggibile, andando così contro il proposito principale della creazione dei range. Generalmente i range sono fatti di step di 5 o 10.
Osservando i dati conviene utilizzare un range di step pari a 5, ovvero diciotto classi così formate:

(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]

In R:

```
classiScelte <- c(0,5,10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90)
```

2.1 Distribuzione di frequenza assoluta

Utilizziamo i grafici a barre(istogrammi) per mostrare la frequenza

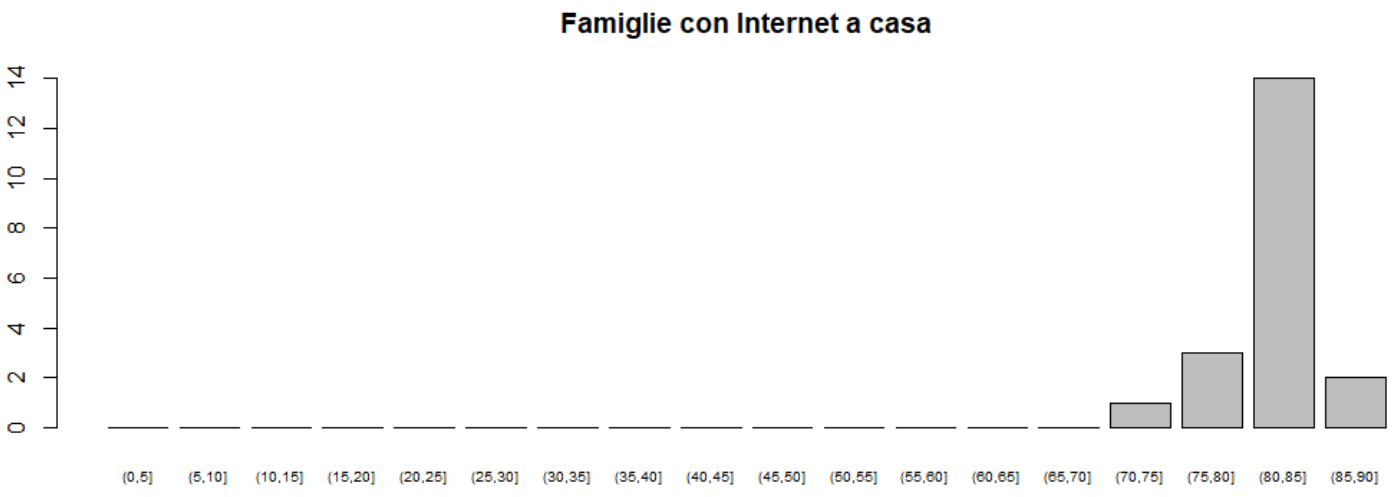
2.1.1 Famiglie che dispongono dell'accesso ad Internet da casa

```
table(cut(InternetACasa,classiScelte))
```

```
##
```

(0,5]	(5,10,]	(10,15]	(15,20]	(20,25]	(25,30]	(30,35]	(35,40]	(40,45]	(45,50]	(50,55]	(55,60]	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]	(85,90]
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	14	2

R:



```
barplot(table(cut(InternetACasa,classiScelte)), cex.names = 0.65, main="Famiglie
con Internet a casa")
```

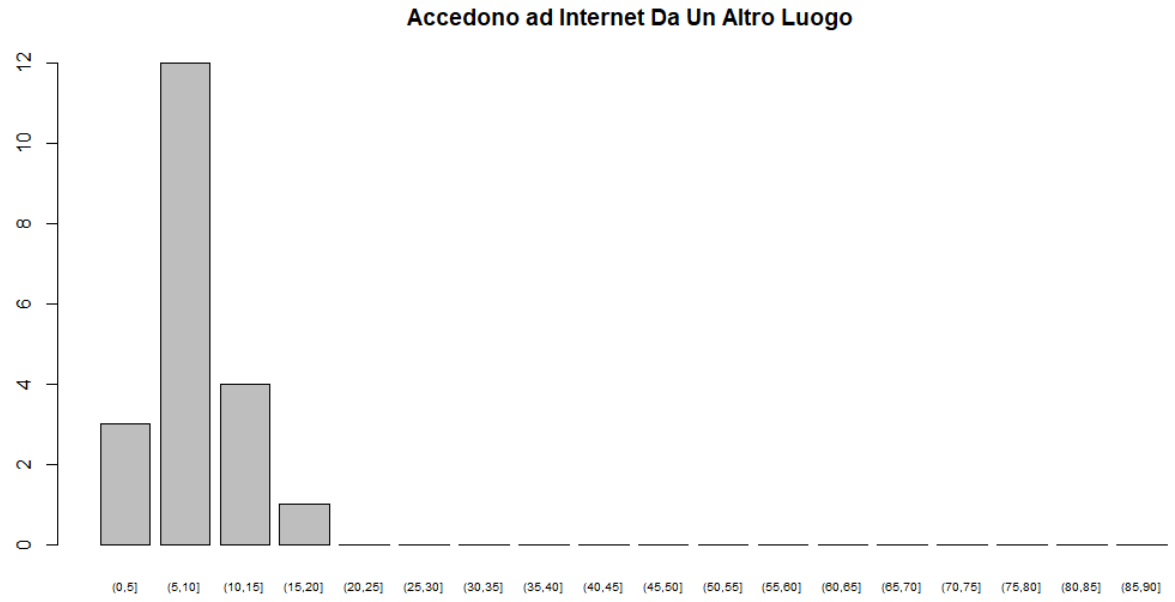
Da questo risultato possiamo notare che:

- La maggior parte delle famiglie ha la possibilità di accedere ad Internet
- Un dato molto più basso degli altri è sorprendentemente quello della Calabria, d'altro canto il più alto possiamo trovarlo in Trentino Alto Adige

2.1.2 Accede ad Internet da un altro luogo

```
table(cut(InternetDaUnAltroLuogo,classiScelte))
```

```
##
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
    3    12     4      1      0      0      0      0      0      0      0      0      0      0      0      0      0      0
```



R:

```
barplot(table(cut(InternetDaUnAltroLuogo,classiScelte)), cex.names = 0.65,
main="Accedono ad Internet Da Un Altro Luogo ")
```

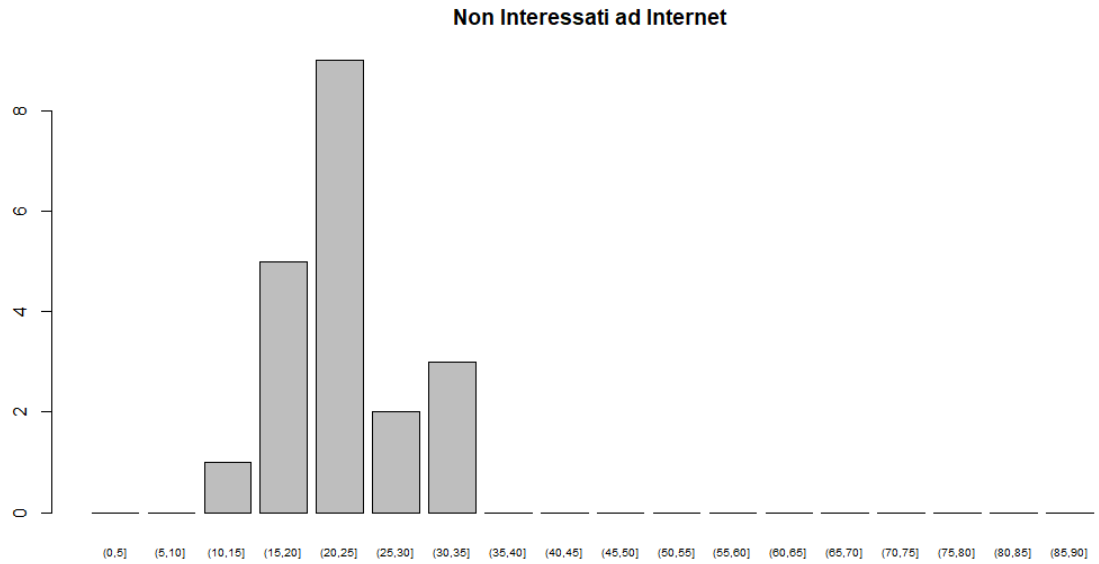
Da questo risultato possiamo notare che:

- La percentuale di famiglie che accedono ad internet da un altro luogo è molto bassa, specialmente in Liguria, Sicilia ed Emilia-Romagna in cui la percentuale non supera il 5%
- È interessante notare come da questi risultati nessuna regione del sud abbia una percentuale altissima.
- È interessante anche notare come la Sardegna possegga una percentuale del 15%

2.1.3 Non interessati ad Internet

```
table(cut(InternetNonInteressa,classiScelte))
```

```
##
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
    0       0        1        5        9        2        3        0        0        0        0        0        0        0        0        0        0        0
```



R:

```
barplot(table(cut(InternetNonInteressa,classiScelte)), cex.names = 0.65, main="Non Interessati ad Internet ")
```

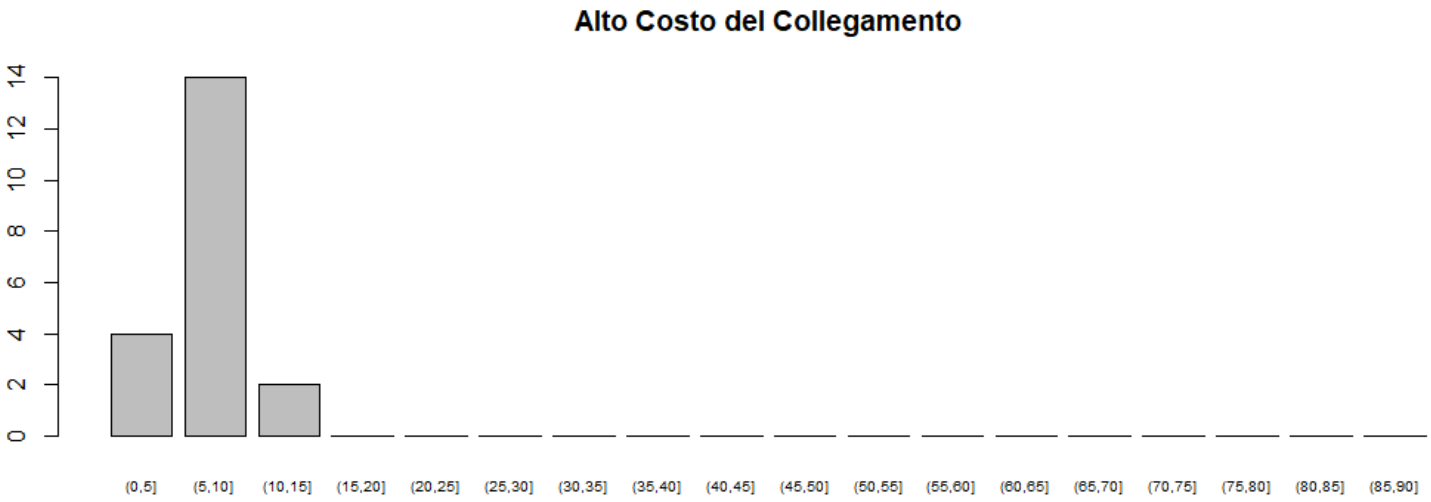
Da questo risultato possiamo notare che:

- La percentuale di famiglie a cui non interessa internet è molto simile nelle varie regioni attestandosi intorno al 23%, ad eccezione di: Basilicata, Trentino-Alto Adige e Friuli-Venezia Giulia, in cui la percentuale supera di non poco il 30%, avvicinandosi molto al 35%
- È interessante notare come in Trentino-Alto Adige pur avendo la percentuale di famiglie che dispongono di Internet a casa, posseggono anche una delle percentuali più alte di famiglie a cui non interessa Internet

2.1.4 Alto Costo del Collegamento

```
table(cut(AltoCostoCollegamento,classiScelte))
```

```
##
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
   4    14     2      0      0      0      0      0      0      0      0      0      0      0      0      0      0      0
```



R:

```
barplot(table(cut(AltoCostoCollegamento,classiScelte)), cex.names = 0.65,
main="Alto Costo del Collegamento")
```

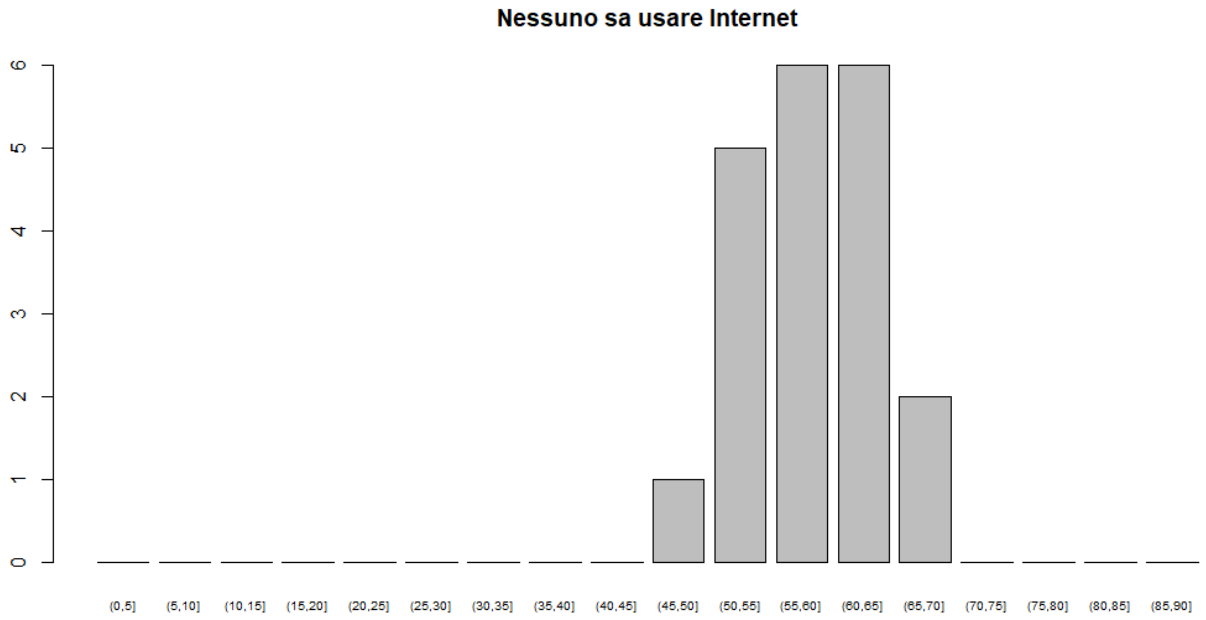
Da questo risultato possiamo notare che:

- Generalmente la percentuale di famiglie che riscontrano un alto costo del collegamento è molto bassa, attestandosi di media circa al 7%

2.1.5 Nessuno sa usare Internet

```
table(cut(NonSaUsareInternet,classiScelte))
```

```
##
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
    0       0       0       0       0       0       0       0       0       1       5       5       6       6       2       0       0       0
```



R:

```
barplot(table(cut(NonSaUsareInternet,classiScelte)), cex.names = 0.5, main="Nessuno sa usare Internet")
```

Da questo risultato possiamo notare che:

- Come ci si potrebbe aspettare, la percentuale di famiglie in cui nessuno sa usare Internet è di molto maggiore nelle regioni dove l'interesse per Internet è più basso
- È interessante notare come la percentuale di famiglie in cui nessuno sa usare Internet si riscontra in Lombardia, nonostante l'alta percentuale di famiglie con accesso ad Internet a casa

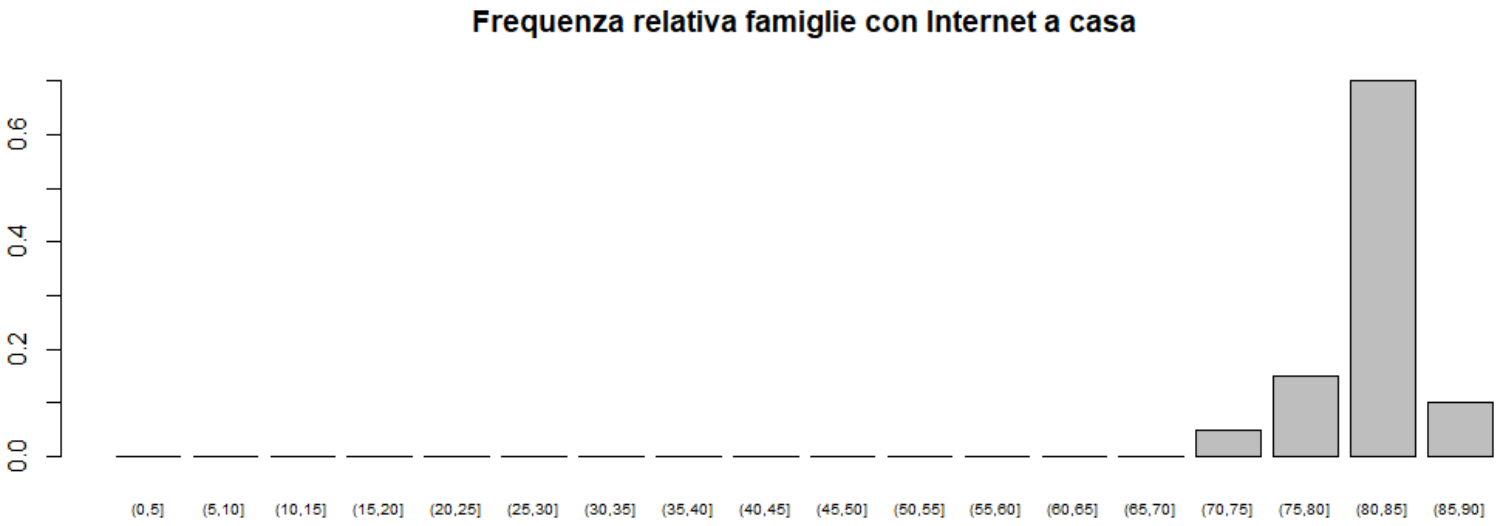
2.2 Distribuzioni di frequenza relative

Per calcolare le frequenze relative, invece, è necessario dividere l'output della frequenza assoluta per la lunghezza dei vettori presi in considerazione tramite la funzione lenght(). Vediamo dunque quali percentuali otteniamo calcolando le frequenze relative delle famiglie che dispongono di accesso ad Internet usando come scala la suddivisione in classi definita in precedenza.

2.2.1 Nessuno sa usare Internet

```
FRelativeInternetACasa<- table(cut(InternetACasa,classiScelte))/
    length(InternetACasa)
```

(0,5]	(5,10,]	(10,15]	(15,20]	(20,25]	(25,30]	(30,35]	(35,40]	(40,45]	(45,50]	(50,55]	(55,60]	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]	(85,90]
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.15	0.70	0.10



R:

```
barplot(FRelativeInternetACasa , cex.names = 0.65, main="Frequenza relativa
famiglie con Internet a casa")
```

Da questo risultato possiamo notare che:

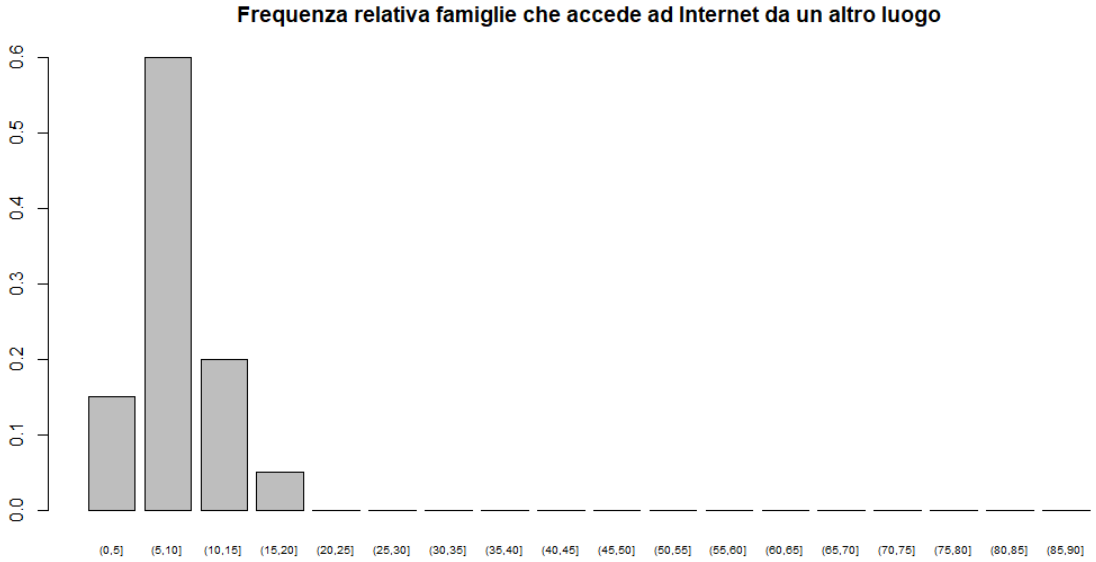
- Come detto in precedenza, la maggior parte delle famiglie divise per regioni ha la stessa percentuale di avere una connessione Internet a casa. Nello specifico il 70% delle regioni si trova nel range 80-85 e solo un 5% si trova ad avere una percentuale tra i 70 e i 75 (ovvero la sola calabria)
- Infine solo un 10% delle regioni (Lombardia e Trentino Alto Adige) hanno una percentuale di famiglie con accesso Internet a casa molto più alto delle altre regioni

2.2.2 Accede ad Internet da un altro luogo

```
FRelativeInternetDaUnAltroLuogo<- table(cut(InternetDaUnAltroLuogo,classiScelte))/
    length(InternetDaUnAltroLuogo)
```

##

(0,5]	(5,10,]	(10,15]	(15,20]	(20,25]	(25,30]	(30,35]	(35,40]	(40,45]	(45,50]	(50,55]	(55,60]	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]	(85,90]
0.15	0.6	0.2	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0



R:

```
barplot(table(cut(InternetDaUnAltroLuogo,classiScelte)), cex.names = 0.65,
main="Accedono ad Internet Da Un Altro Luogo ")
```

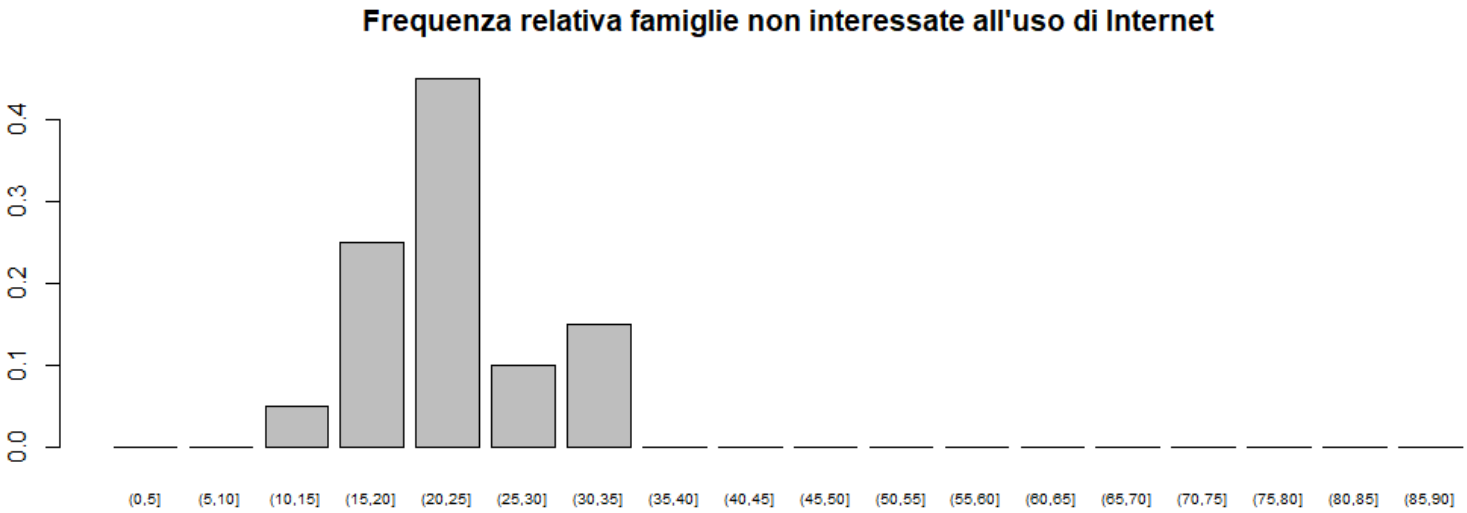
Da questo risultato possiamo notare che:

- Come detto in precedenza, solo nello 0,15% delle regioni (Liguria, Sicilia ed Emilia-Romagna) la percentuale di famiglie che non ha internet a casa ma accede da un altro luogo è più bassa rispetto alle altre
- Le regioni al 60% hanno quindi una percentuale di famiglie che accedono ad Internet da un altro posto compresa tra il 5% e il 10%
- Come osservato in precedenza, solo lo 0,05% delle regioni ha una percentuale di famiglie che accedono ad internet da un altro posto compresa tra il 15% e il 20%

2.2.3 Non interessati ad Internet

```
FRelativeInternetNonInteressa<- table(cut(InternetNonInteressa,classiScelte))/
    length(InternetNonInteressa)
```

```
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
    0      0    0.05    0.25    0.45    0.10    0.15       0       0       0       0       0       0       0       0       0       0       0
```



R:

```
barplot(FRelativeInternetNonInteressa , cex.names = 0.65, main="Frequenza relativa
famiglie non interessate all'uso di Internet")
```

Da questo risultato possiamo notare che:

- Come osservato in precedenza, molte delle regioni, ovvero il 45%, hanno una percentuale compresa tra il 20% e il 25%.

2.2.4 Alto costo del collegamento

```
FRelativeAltoCostoCollegamento<- table(cut(AltoCostoCollegamento,classiScelte))/
    length(AltoCostoCollegamento)
```

```
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
 0.20   0.70    0.10       0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
```

R:

```
barplot(FRelativeAltoCostoCollegamento , cex.names = 0.65, main="Frequenza relativa
famiglie che riscontrano un alto costo del collegamento")
```

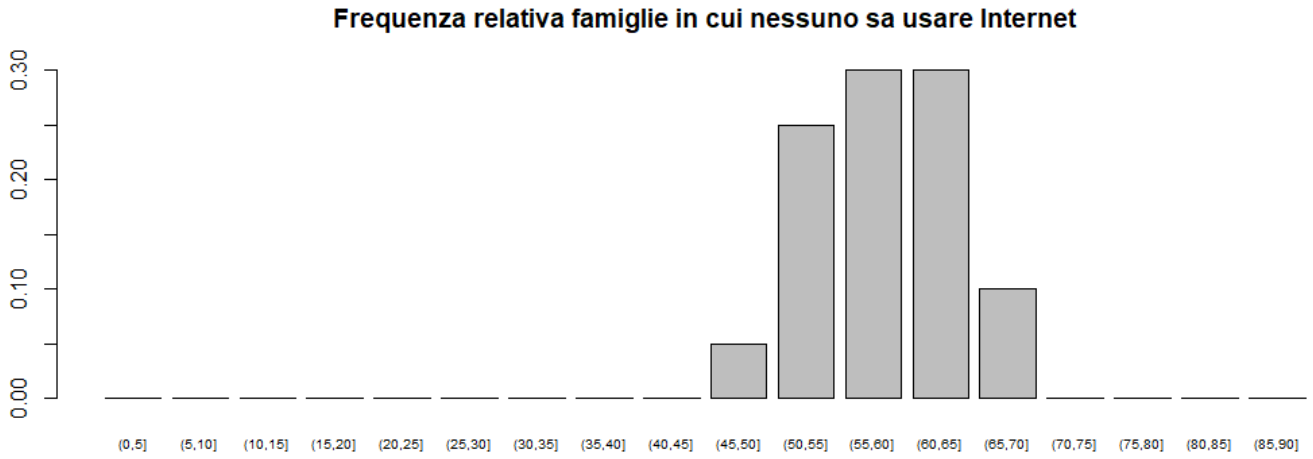
Da questo risultato possiamo notare che:

- Come osservato in precedenza, fortunatamente, nella maggior parte delle regioni, **70%**, le famiglie che riscontrano un alto costo per il collegamento sono “soltanto” comprese tra il 5% e il 10%
- Mentre il 20% ancora più fortunate da avere una percentuale di famiglie che riscontrano questo problema compresa tra lo **0%** e il **5%**
- Mentre sfortunatamente nel **10%** ci sono percentuali leggermente più alte della media nazionale, in queste due regioni la percentuale di famiglie a riscontrare un costo alto sono comprese tra il 10% e il 15%

2.2.5 Nessuno sa usare Internet

```
FRelativeNonSaUsareInternet<- table(cut(NonSaUsareInternet,classiScelte))/
    length(NonSaUsareInternet)
```

```
(0,5] (5,10,] (10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50] (50,55] (55,60] (60,65] (65,70] (70,75] (75,80] (80,85] (85,90]
    0       0        0       0       0       0       0       0       0    0.05    0.25    0.30    0.30    0.10       0       0       0       0
```



R:

```
barplot(FRelativeNonSaUsareInternet , cex.names = 0.65, main="Frequenza relativa
famiglie in cui nessuno sa usare Internet")
```

Da questo risultato possiamo notare che:

- Come osservato in precedenza, spaventosamente c'è il **30%** ha una percentuale di famiglie in cui nessuno sa usare Internet con una percentuale compresa tra 55% e 60%
- Altrettanto spaventoso è l'altro **30%** di regioni in cui la percentuale di famiglie in cui nessuno sa usare Internet è ancora più alta compresa tra 60% e 65%
- È, inoltre, preoccupante quel **10%** di regioni in cui le famiglie che non sanno usare Internet con una percentuale compresa tra il 65% e il 70%

3. Analisi tramite rappresentazione grafiche

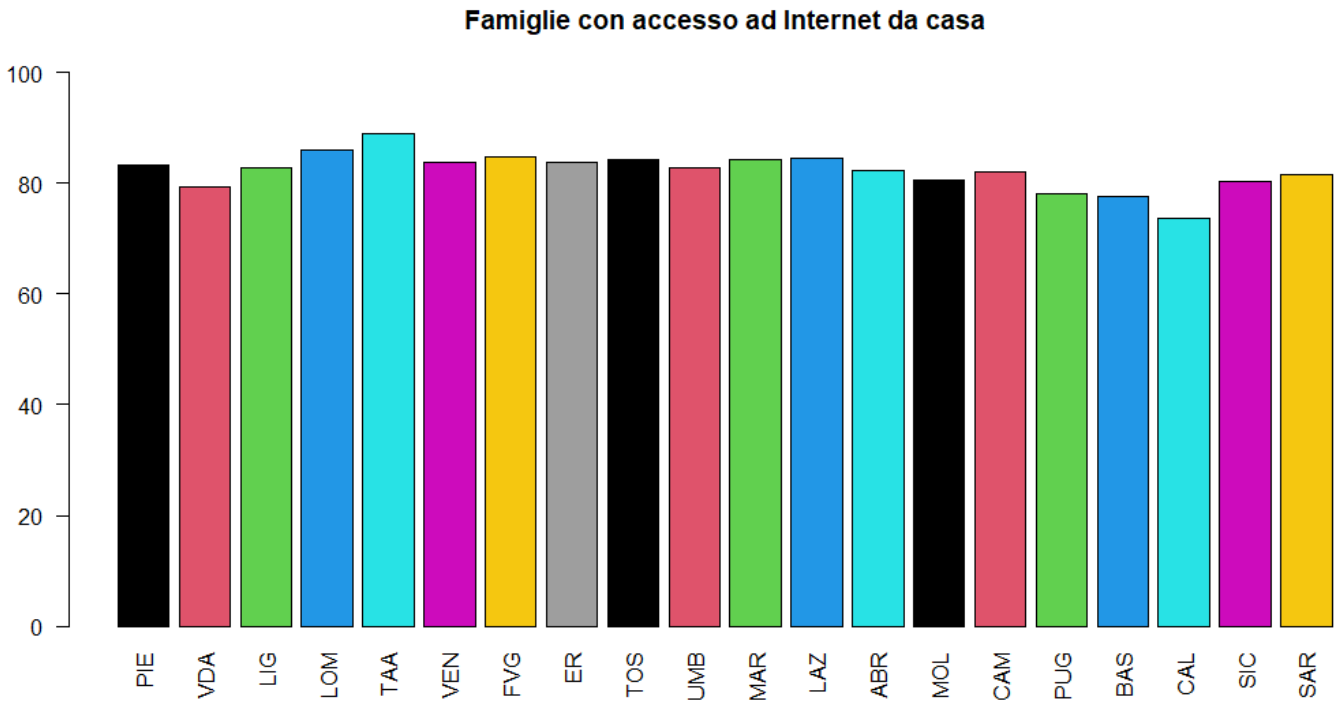
La rappresentazione grafica, come usata in precedenza, rende molto più immediato e semplice leggere i dati e le analisi effettuate, ci serviremo di molte di queste per rappresentare efficientemente i dati

Creiamo un vettore per racchiudere tutti i nomi abbreviati delle regioni:

```
regioniAbbreviate <- c("PIE","VDA","LIG","LOM","TAA","VEN","FVG","ER","TOS",
"UMB","MAR","LAZ","ABR","MOL","CAM","PUG","BAS","CAL", "SIC","SAR")
```

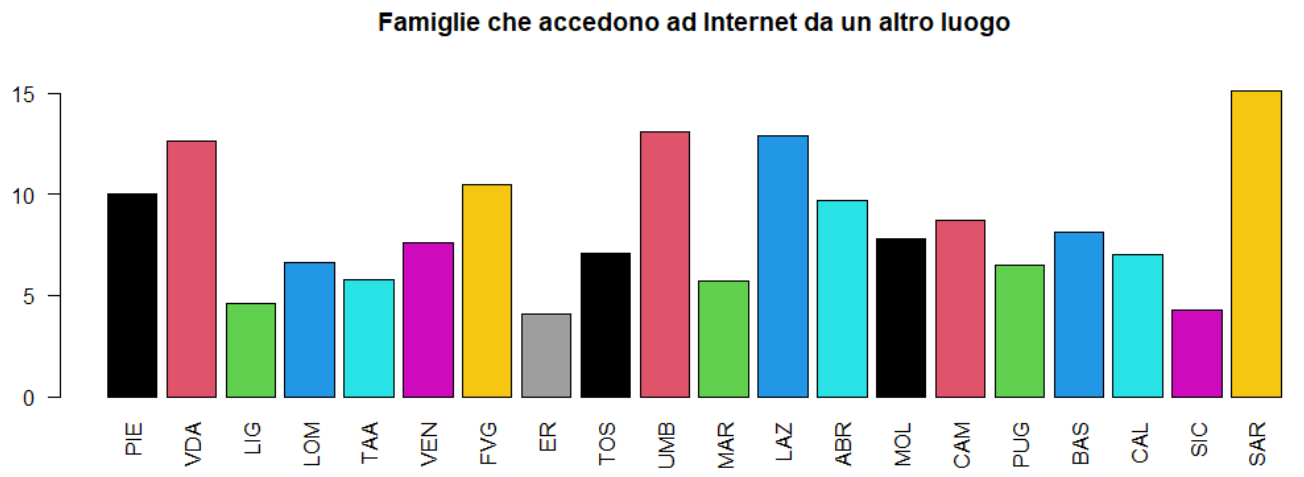
3.1 Grafici a barre

Per ogni colonna della matrice creiamo un grafico a barre, ovvero un grafico con tante linee (barre) quante sono le righe, della stessa larghezza e lunghezza proporzionata alla percentuale per ogni riga

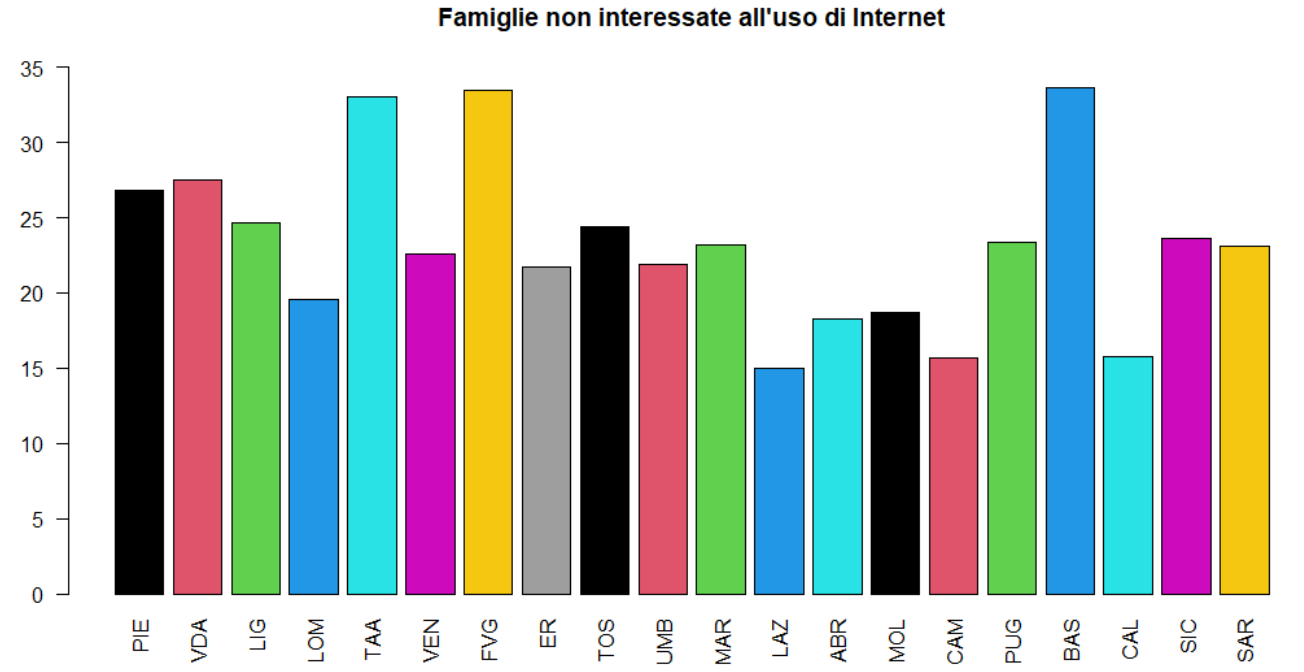



```
barplot(matriceFamiglie[,1], col = 1:13,
        main = "Famiglie con accesso ad Internet da casa",
        las=2, names.arg=regioniAbbreviate, ylim = c(0,100))
```

Giardinetto Giuseppe matricola n°. 0512114655

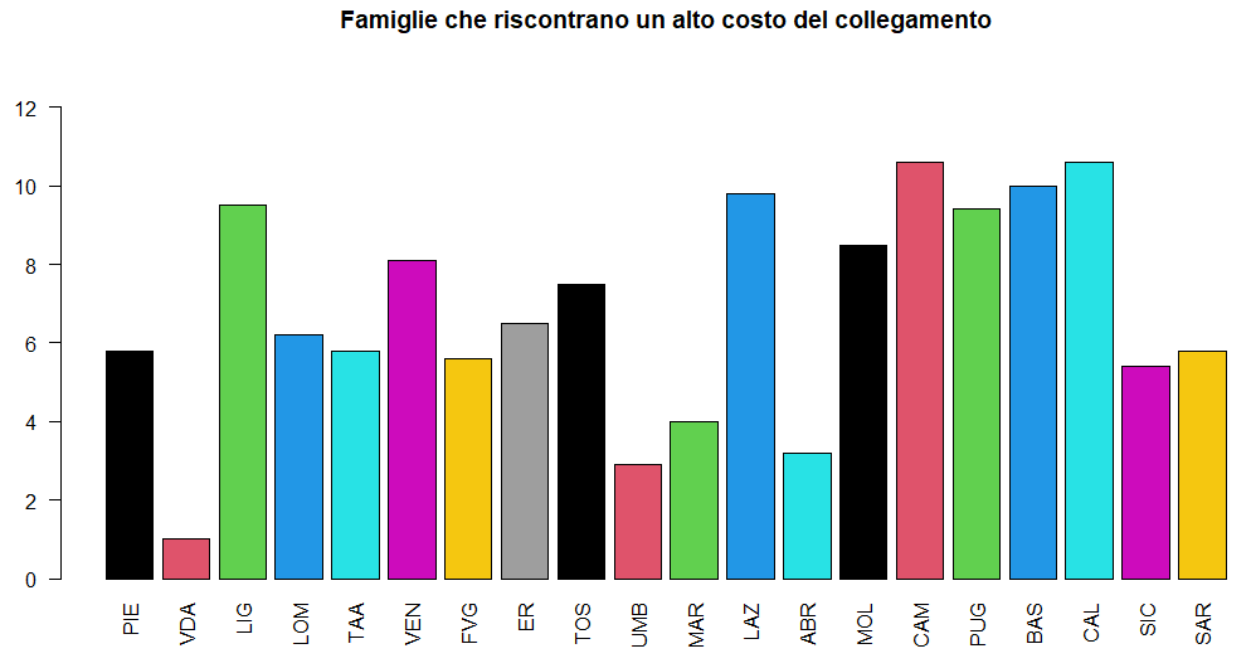



```
barplot(matriceFamiglie[,2], col = 1:13,
        main = "Famiglie che accedono ad Internet da un altro luogo",
        las=2, names.arg=regioniAbbreviate, ylim = c(0,16))
```

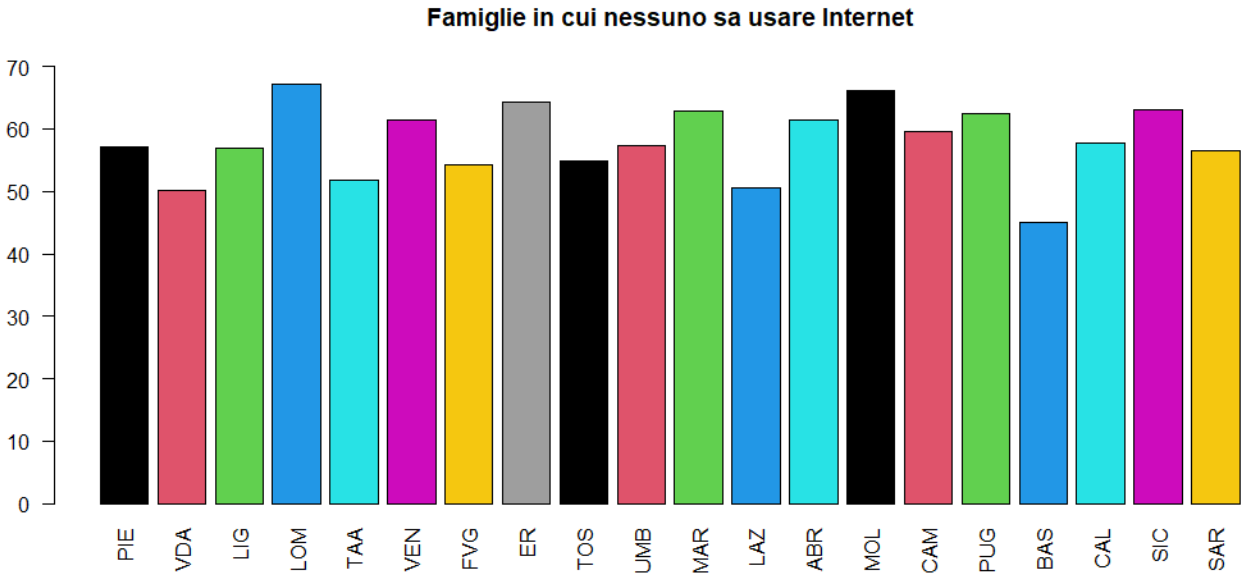



```
barplot(matriceFamiglie[,3], col = 1:13,
        main = "Famiglie non interessate all'uso di Internet",
        las=2, names.arg=regioniAbbreviate, ylim = c(0,35))
```

Giardinetto Giuseppe matricola n°. 0512114655




```
barplot(matriceFamiglie[,4], col = 1:13,
        main = "Famiglie che riscontrano un alto costo del collegamento",
        las=2, names.arg=regioniAbbreviate, ylim = c(0,13))
```




```
barplot(matriceFamiglie[,5], col = 1:13,
        main = "Famiglie in cui nessuno sa usare Internet",
        las=2, names.arg=regioniAbbreviate, ylim = c(0,70))
```

3.2 Grafico a torta

Un grafico a torta è una rappresentazione dei dati dalla forma circolare, suddiviso in “fette” (archi alla circonferenza) grandi proporzionatamente alla quantità da rappresentare. Questi, come i grafici a barre, permettono la rappresentazione e la resa immediata dei dati.

Spesso viene preferito l’uso dei grafici a barre per la loro semplicità e didascalicità.

Ci avvarremo di un grafico a torta per la rappresentazione della situazione delle famiglie italiane per l’accesso ad Internet.

Iniziamo con il definire i dati da rappresentare:

usiamo l’operazione **mean()** per calcolare la media aritmetica di ogni colonna della matrice, quindi:

```
mean(InternetACasa)
```
82.23

```
mean(InternetDaUnAltroLuogo)
```
8.39

```
mean(InternetNonInteressa)
```
23.305

```
mean(AltoCostoCollegamento)
```
6.81

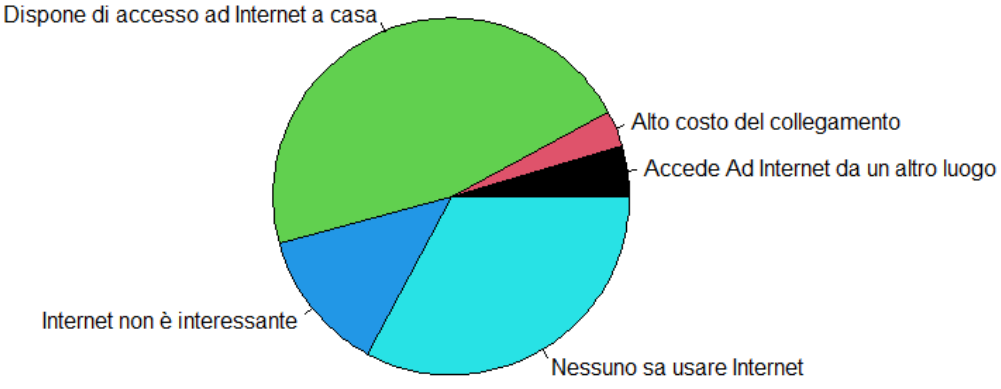
```
mean(NonSaUsareInternet)
```
58.065

Racchiudiamo queste informazioni in un vettore chiamato italia

```
italia <- c(
rep("Dispone di accesso ad Internet a casa",82.23),
rep("Accede Ad Internet da un altro luogo",8.39),
rep("Internet non è interessante",23.305),
rep("Alto costo del collegamento",6.81),
rep("Nessuno sa usare Internet",58.065)
)
```

Usiamo la funzione **pie()** di R per rappresentare il grafico:

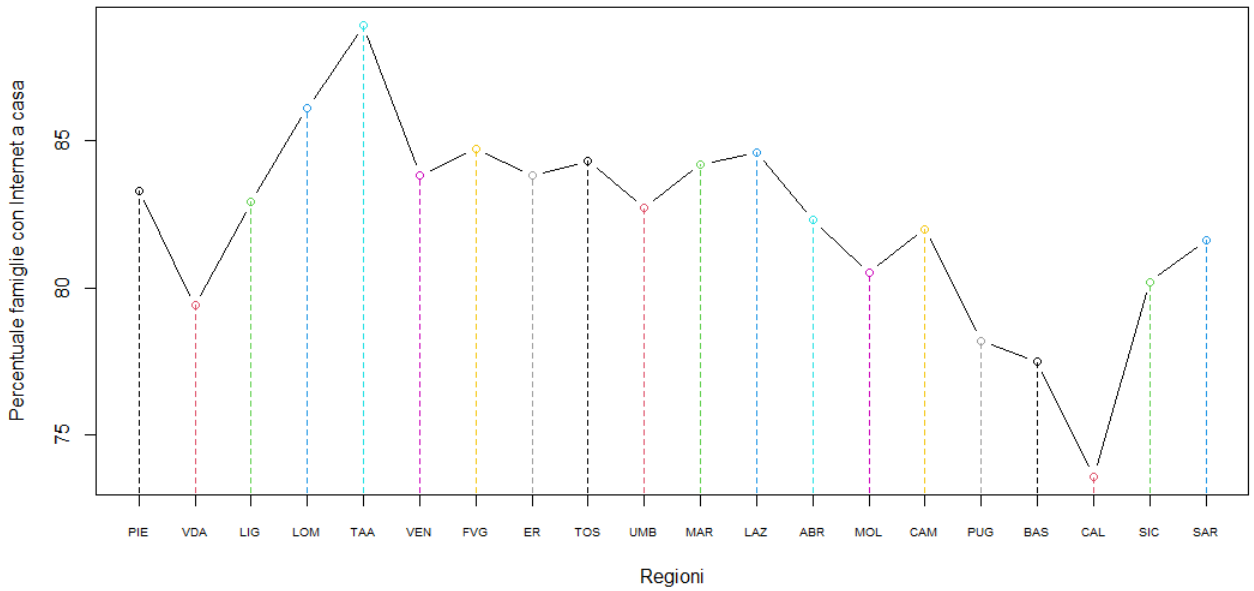
```
pie(table(italia), col=1:5)
```
Risultato:



3.3 Plot per vettori

La funzione **plot()** è utile per rappresentare l'andamento dei valori assunti dai vettori che costituiscono la matrice del nostro dataset. Andiamo a rappresentare uno ad uno i vettori (colone) del dataset(matrice) mostrando i valori che vengono assunti da ogni regione:

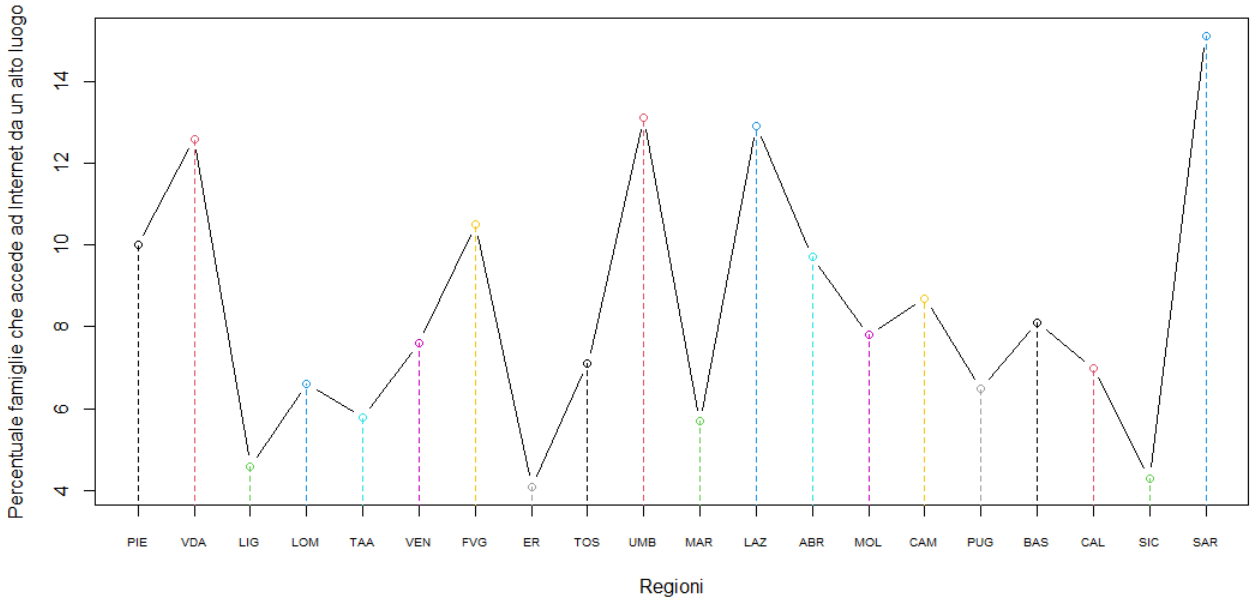
- **Grafico relativo alle famiglie che dispongono di accesso ad Internet da casa**




```
plot(InternetACasa,
            ylab="Percentuale famiglie con Internet a casa",
            xlab="Regioni",
            col =1:20,type = "b",axes = FALSE)

box(which = "plot", lty = "solid")
axis(side=2)
axis(side=1, at=1:20, labels=regioniAbbreviate,cex.axis=0.65)

for (i in 1:length(InternetACasa)) {
        lines(x=i, y=InternetACasa[i], "h", lty = 2, col=i)
    }
```

Giardinetto Giuseppe matricola n°. 0512114655

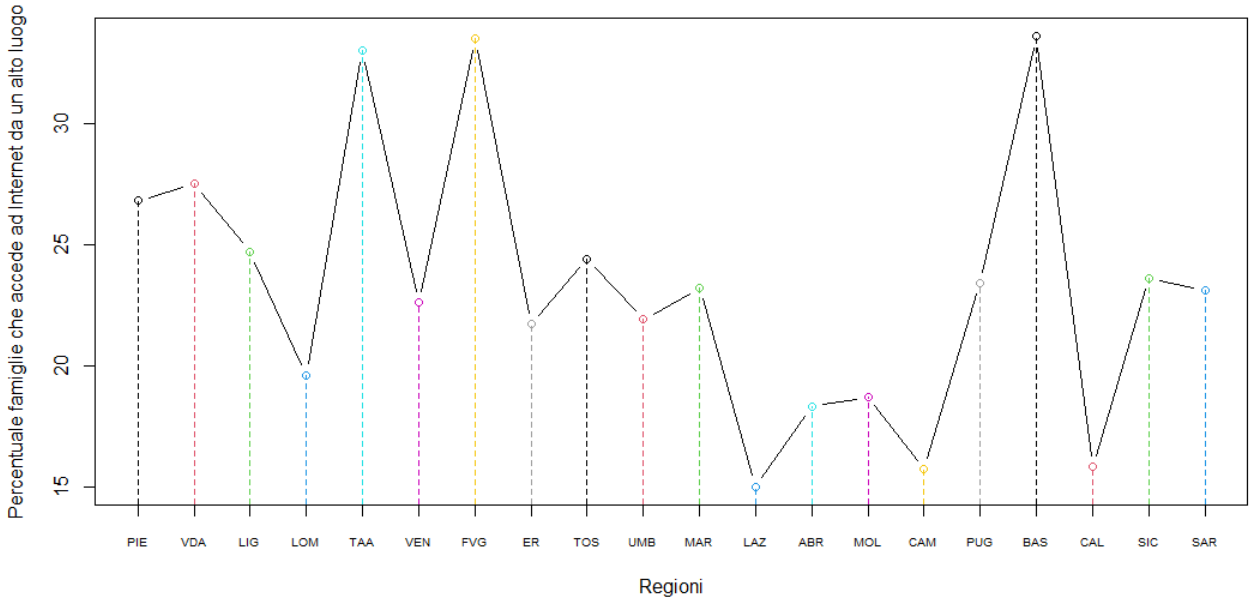
- **Grafico relativo alle famiglie che accedono ad Internet da un altro luogo**




```
plot(InternetDaUnAltroLuogo,
            ylab="Percentuale famiglie che accedono ad Internet da un altro luogo",
            xlab="Regioni",
            col =1:20,type = "b", axes = FALSE)

box(which = "plot", lty = "solid")
axis(side=2)
axis(side=1, at=1:20, labels=regioniAbbreviate,cex.axis=0.65)

for (i in 1:length(InternetDaUnAltroLuogo)) {
        lines(x=i, y=InternetDaUnAltroLuogo[i], "h", lty = 2, col=i)
    }
```

Giardinetto Giuseppe matricola n°. 0512114655

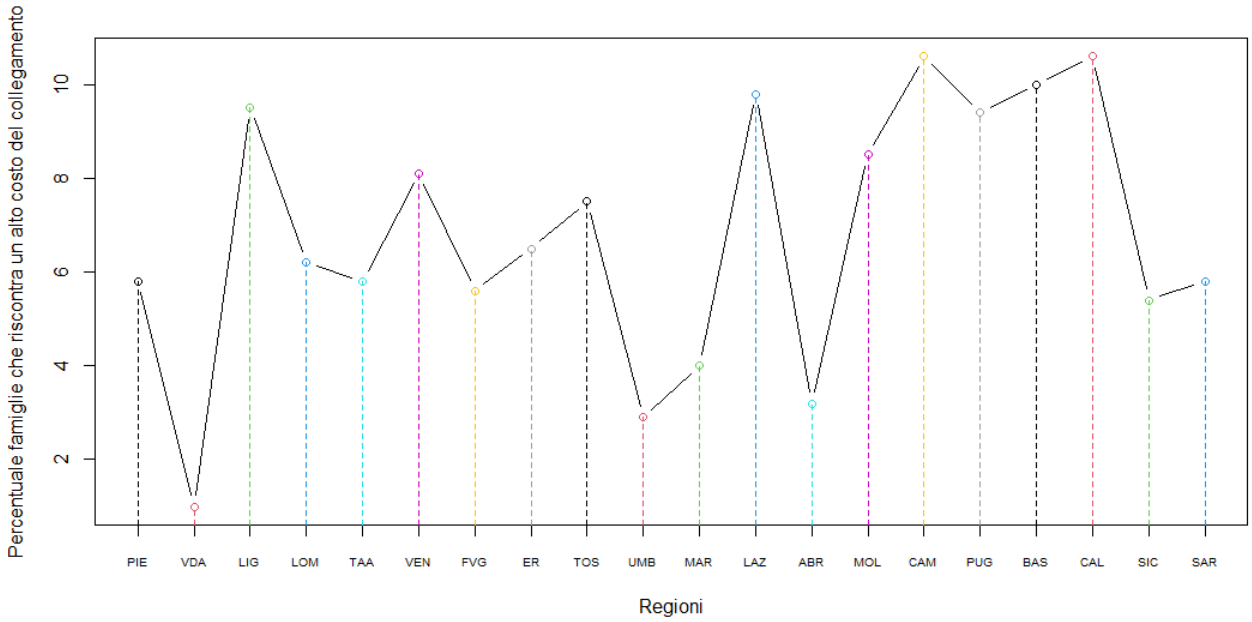
- **Grafico relativo alle famiglie che non sono interessate all'uso di Internet**




```
plot(InternetNonInteressa,
     ylab="Percentuale famiglie che accede ad Internet da un alto luogo",
     xlab="Regioni",
     col =1:20,type = "b",axes = FALSE)

box(which = "plot", lty = "solid")
axis(side=2)
axis(side=1, at=1:20, labels=regioniAbbreviate,cex.axis=0.65)

for (i in 1:length(InternetNonInteressa)) {
         lines(x=i, y=InternetNonInteressa[i], "h", lty = 2, col=i)
     }
```

Giardinetto Giuseppe matricola n°. 0512114655

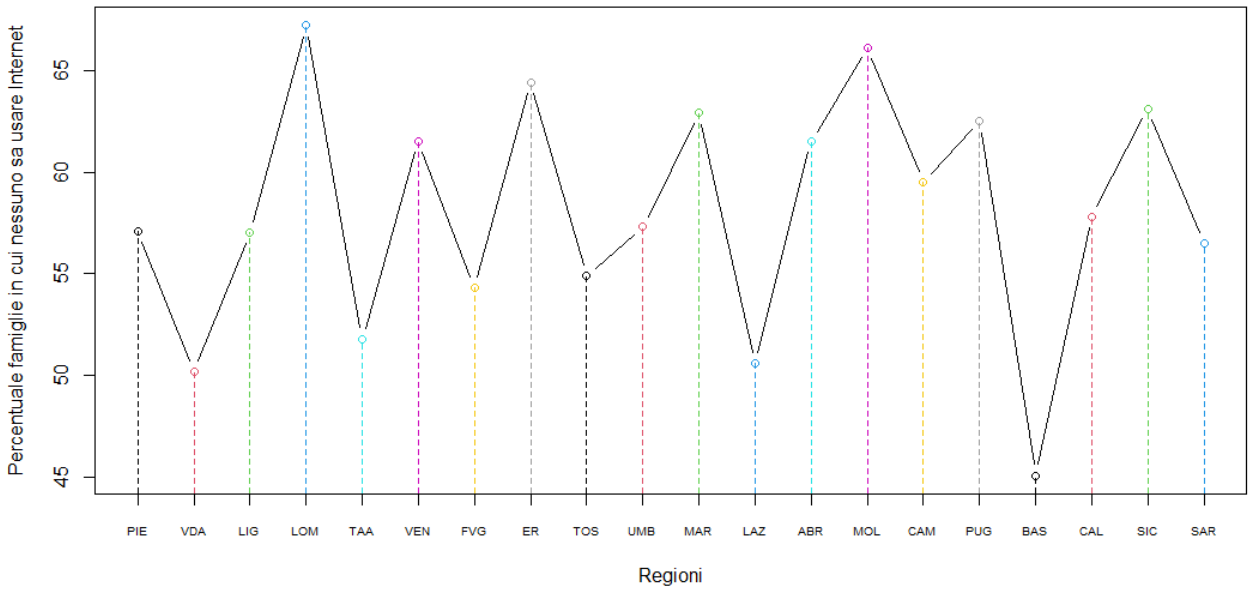
- **Grafico relativo alle famiglie che riscontrano un alto costo del collegamento**




```
plot(AltoCostoCollegamento,
     ylab="Percentuale famiglie che riscontra un alto costo del collegamento",
     xlab="Regioni",
     col =1:20,type = "b",axes = FALSE)

box(which = "plot", lty = "solid")
axis(side=2)
axis(side=1, at=1:20, labels=regioniAbbreviate,cex.axis=0.65)

for (i in 1:length(AltoCostoCollegamento)) {
         lines(x=i, y=AltoCostoCollegamento[i], "h", lty = 2, col=i)
     }
```

Giardinetto Giuseppe matricola n°. 0512114655

- **Grafico relativo alle famiglie in cui nessuno sa usare Internet**




```
plot(NonSaUsareInternet,
     ylab="Percentuale famiglie in cui nessuno sa usare Internet",
     xlab="Regioni",
     col =1:20,type = "b",axes = FALSE)

box(which = "plot", lty = "solid")
axis(side=2)
axis(side=1, at=1:20, labels=regioniAbbreviate,cex.axis=0.65)

for (i in 1:length(NonSaUsareInternet)) {
         lines(x=i, y=NonSaUsareInternet[i], "h", lty = 2, col=i)
     }
```

3.4 Istogrammi

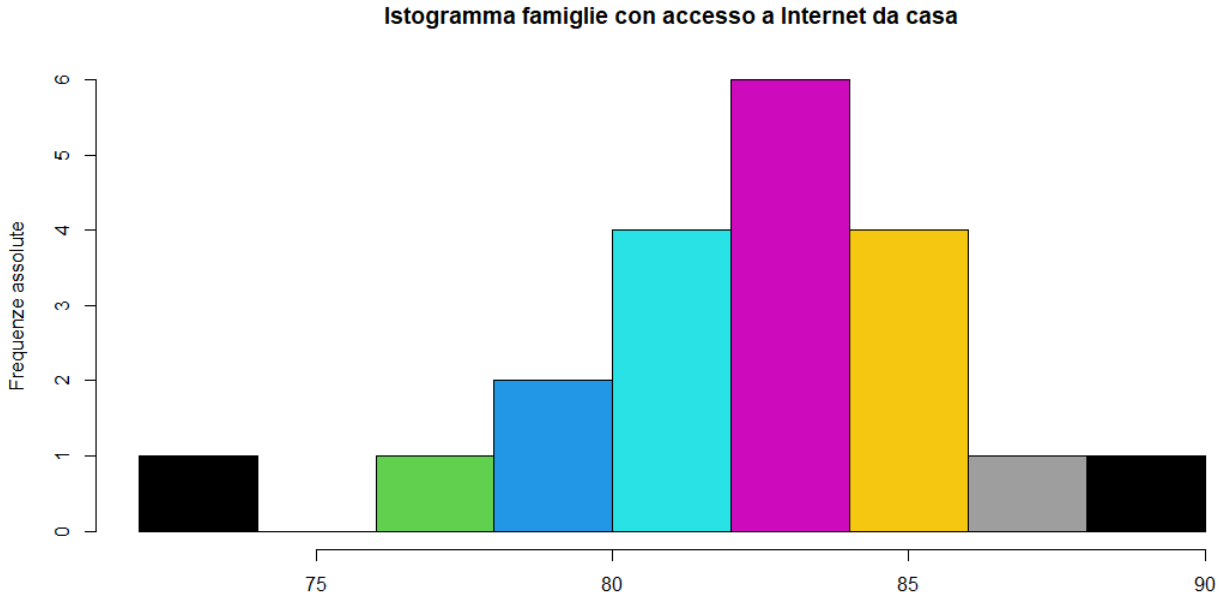
Un istogramma è un tipo di grafico formato da un asse cartesiano su cui sono disposte barre verticali in corrispondenza dei valori dei dati da rappresentare, alte proporzionatamente alla loro frequenza. La costruzione di un istogramma inizia dalla definizione delle classi nelle quali rappresentare i valori.

In R è possibile realizzare un istogramma tramite il comando **hist()** in cui è possibile definire le classi o lasciare al programma decidere le più appropriate per dimensione e numero.

R, oltre al grafico genera anche i **breaks** usati per la generazione dell'istogramma, le **frequenze assolute** delle classi, la **densità** delle classi e il loro **punto centrale**

Possiamo vedere un esempio pratico con l'istogramma riferito alle **famiglie che dispongono di accesso ad Internet da casa**:

```
HistInternetACasa <- hist(InternetACasa, freq=TRUE, main="Istogramma famiglie con
                  accesso a Internet da casa", ylab="Frequenze assolute", col=1:8)
```




```
>str(HistInternetACasa)
##List of 6
## $ breaks  : int [1:10] 72 74 76 78 80 82 84 86 88 90
## $ counts  : int [1:9] 1 0 1 2 4 6 4 1 1
## $ density : num [1:9] 0.025 0 0.025 0.05 0.1 0.15 0.1 0.025 0.025
## $ mids    : num [1:9] 73 75 77 79 81 83 85 87 89
## $ xname   : chr "InternetACasa"
## $ equidist: logi TRUE
## - attr(*, "class")= chr "histogram"
HistInternetACasa$breaks
## [1] 72 74 76 78 80 82 84 86 88 90
HistInternetACasa$counts
## [1] 1 0 1 2 4 6 4 1 1
HistInternetACasa$density
## [1] 0.025 0 0.025 0.05 0.1 0.15 0.1 0.025 0.025
HistInternetACasa$mids
## [1] 73 75 77 79 81 83 85 87 89
```

In questo caso R ha effettuato la divisione nelle seguenti classi:

(72,74] (74,76] (76,78] (78,80] (80,82] (82,84] (84,86] (86,88] (88,90]

Come indicato dal valore **couts**, dei 20 valori:

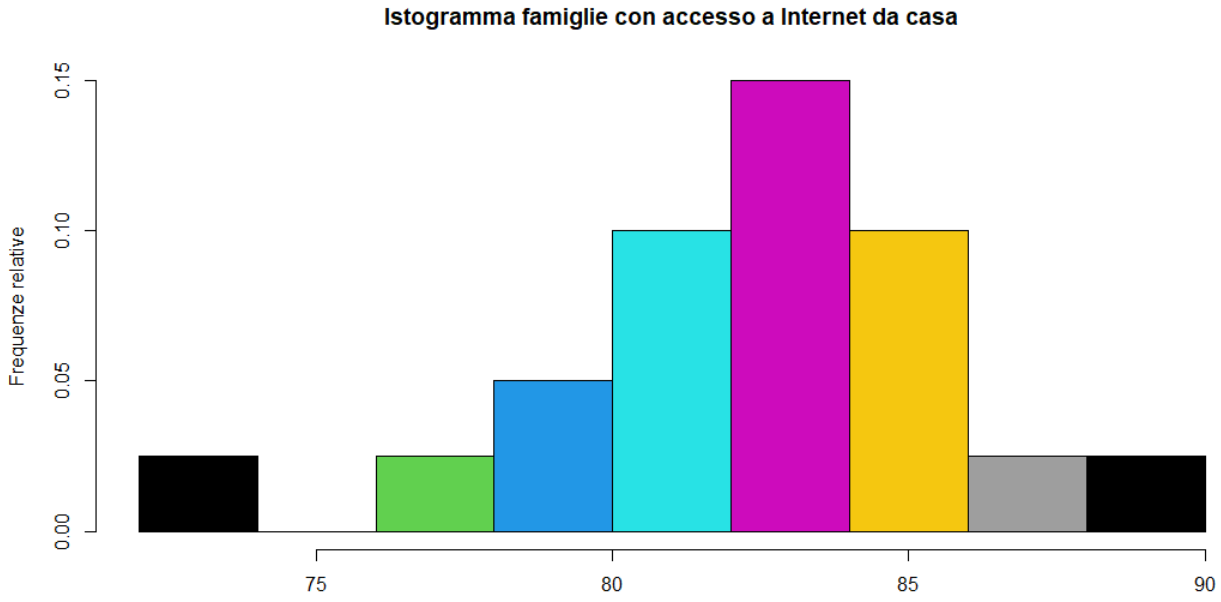
- Nella prima classe ne cade 1
- Nella seconda 0
- Nella terza 1
- Nella quarta 2
- Nella quinta 4
- Nella sesta 6
- Nella settima 4
- Nell'ottava 1
- Nella nona 1

Conoscendo la densità e l'ampiezza delle classi, possiamo ricavarci la frequenza relativa moltiplicando la densità di ogni classe per la sua ampiezza (2), quindi:

```
> HistInternetACasa$density*2
## [1] 0.05 0.00 0.05 0.10 0.20 0.30 0.20 0.05 0.05
```

Il calcolo delle frequenze relative è possibile anche attraverso l'attributo **freq=FALSE** del comando hist, producendo così un istogramma con altezza delle barre pari alle frequenze relative di ogni classe:

```
HistRelInternetACasa <- hist(InternetACasa, freq=FALSE, main="Istogramma famiglie
                      con a Internet da casa", ylab="Frequenze relative", col=1:8)
```

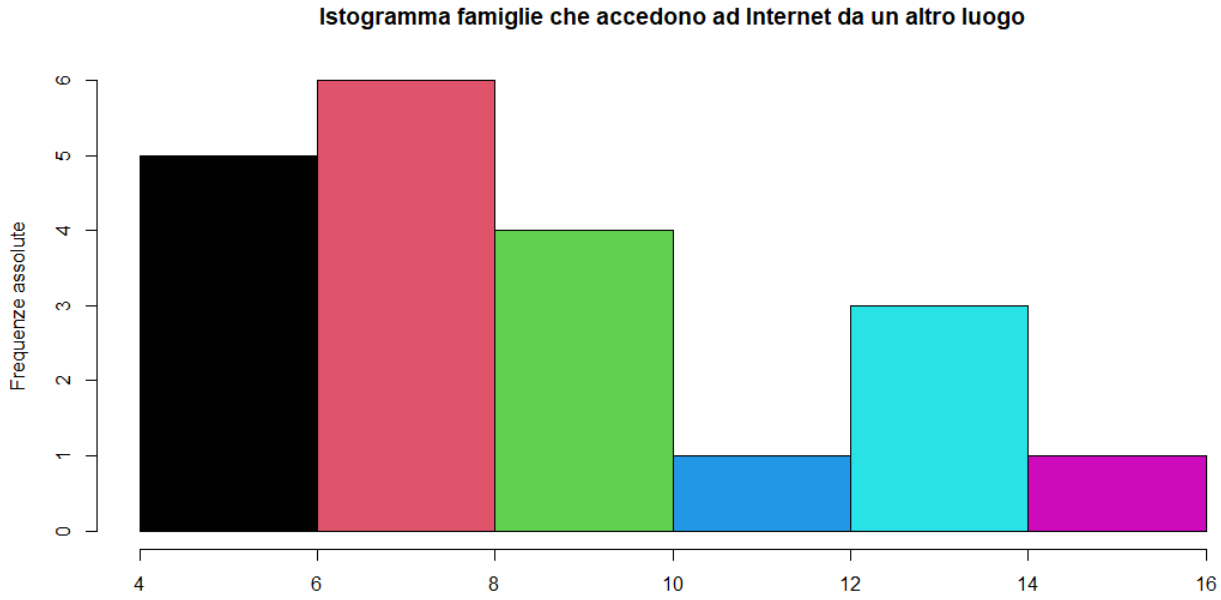



```
str(HistRelInternetACasa)
## List of 6
##  $ breaks  : int [1:10] 72 74 76 78 80 82 84 86 88 90
##  $ counts  : int [1:9] 1 0 1 2 4 6 4 1 1
##  $ density : num [1:9] 0.025 0 0.025 0.05 0.1 0.15 0.1 0.025 0.025
##  $ mids    : num [1:9] 73 75 77 79 81 83 85 87 89
##  $ xname   : chr "InternetACasa"
##  $ equidist: logi TRUE
##  - attr(*, "class")= chr "histogram"
```

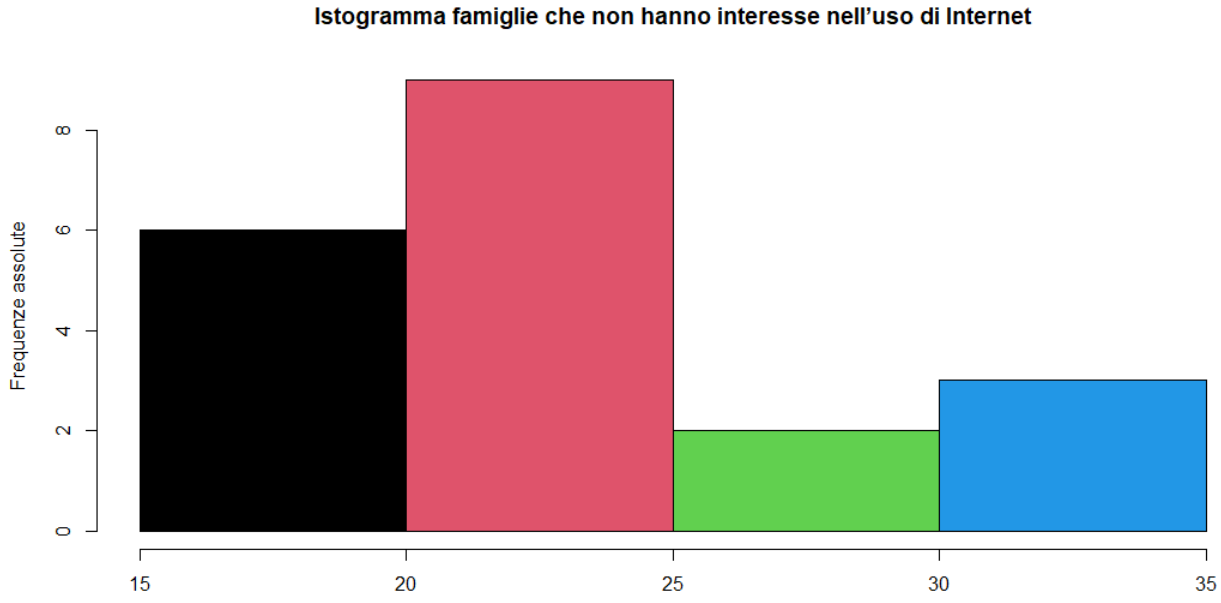
Analogamente al metodo precedente, le frequenze relative vengono calcolate moltiplicando la densità per l'ampiezza delle classi.

Giardinetto Giuseppe matricola n°. 0512114655

Istogramma relativo alle famiglie che accedono ad Internet da un altro luogo:

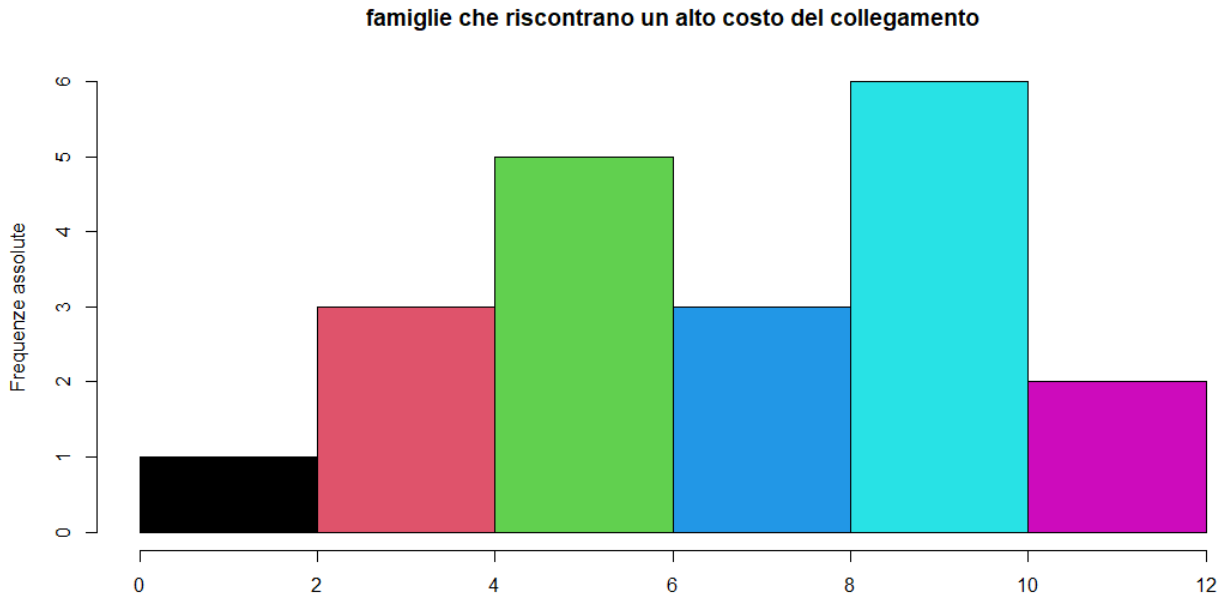


Istogramma relativo alle famiglie che non hanno interesse nell'uso di Internet:

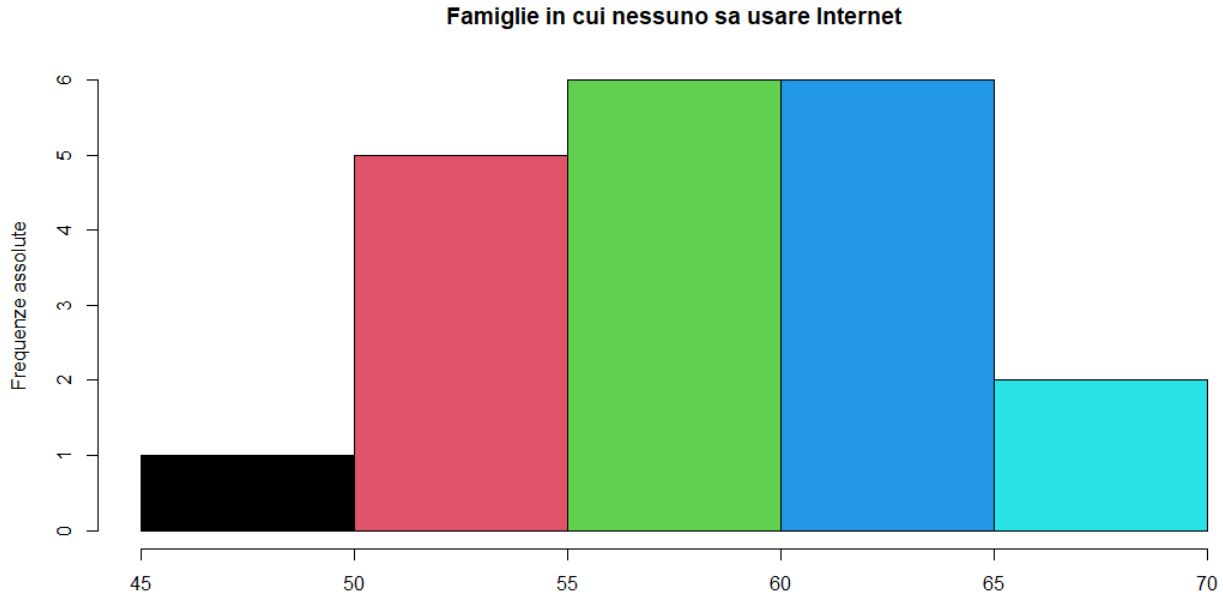


Giardinetto Giuseppe matricola n°. 0512114655

Istogramma relativo alle famiglie che riscontrano un alto costo del collegamento:



Istogramma relativo alle famiglie in cui nessuno sa usare Internet:



Giardinetto Giuseppe matricola n°. 0512114655

3.5 Boxplot

Per parlare di Boxplot dobbiamo introdurre il concetto di **percentile** e **quartile**:

Si definisce **percentile k-esimo** un numero h, con $k \leq h \leq 100-k$. Più in generale, preso un insieme numerico, il percentile k-esimo di questo insieme è un numero (o esattamente due) che rispettano la condizione sopracitata (nel caso siano due i numeri che rispettano questa condizione, il percentile k-esimo è pari alla media aritmetica dei due)

Più semplicemente, il **percentile k-esimo** è un numero che è maggiore del k% degli altri numeri dell'insieme e minore degli altri 100-k% valori

In statistica un boxplot è una rappresentazione grafica dei dati attraverso l'uso dei **quartili**.

Un **quartile** non è altro che un particolare tipo **percentile**:

1. Il **primo quartile** (**Q1**) è definito come il **25° percentile**, ovvero il numero k tale che k è maggiore del 25% degli altri dati e minore degli altri 75%
2. Il **secondo quartile**(**Q2**) è definito come la mediana, ovvero il valore centrale dell'insieme di valori
3. Il **terzo quartile**(**Q3**) è definito come il **75° percentile**, ovvero il numero k tale che k è maggiore del 75% degli altri dati ma minore del restante 25%

Rispettivamente **Q0** e **Q4** poi sono il valore minimo e massimo dell'insieme di dati (estremi)

In R, i quartili vengono calcolati attraverso il comando **quantile()** mentre con **summary()** vengono elencati i valori precisi dei quartili:
Un boxplot è un grafico composto da un rettangolo le cui basi sono poste perpendicolarmente all'asse delle y, in corrispondenza rispettivamente del Q1 e Q3 separati da una linea che individua due sotto-rettangoli in corrispondenza del valore di Q2, questa figura rettangolare viene definita **scarto interquartile (IQR)**, e oltre questo, sono posti dei segmenti, detti **baffi**, che insieme allo scarto interquartile vengono definiti **range**

Le differenze Q3-Q2 e Q2-Q1 possono indicare una simmetria o asimmetria nei dati, poiché Q2 vale come mediana, se la differenza tra Q3 e Q2 è molto più grande o più piccola rispetto a quella tra Q2 e Q1, allora possiamo affermare che i dati sono asimmetrici

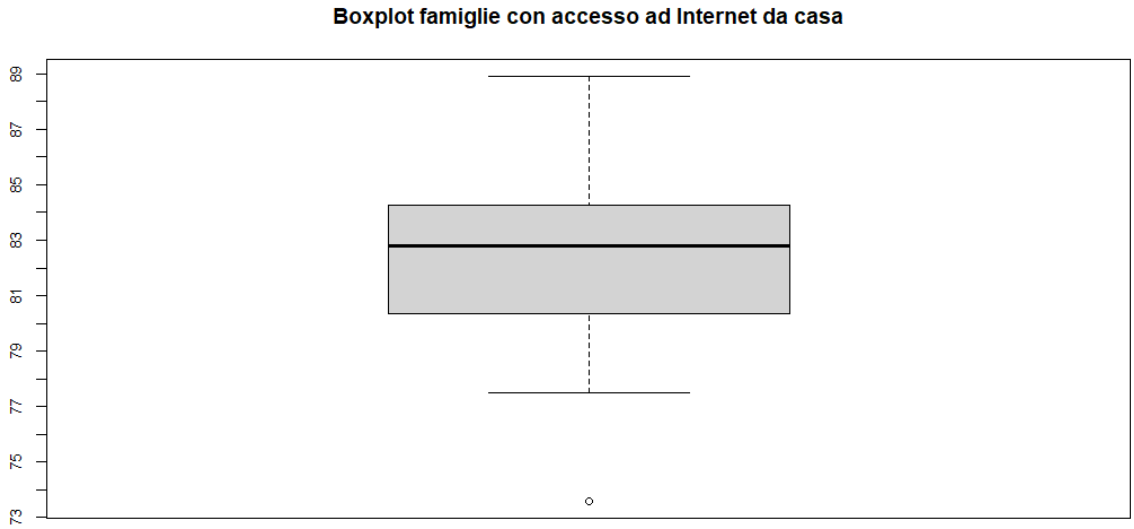
I baffi del boxplot si estendono per 1,5 volte l'IQR, quindi:

- **Baffo inferiore:**
  - **Q1 - 1.5*IQR -> Q1 – 1.5*(Q3-Q1)**
- **Baffo Superiore:**
  - **Q3 + 1.5*IQR -> Q3 + 1.5*(Q3-Q1)**

I boxplot sono molto utili per poter rappresentare i dati appena descritti, inoltre sono fondamentali per l'individuazione degli **outliers**, ovvero valori "anomali" che possono essere dovuti ad errori di misurazione, presenza di eventi molto rari che risultano "anormali". Questi spesso sono causati da un'estrema **curtosi**

3.5.1 Famiglie che hanno accesso ad Internet da casa:

```
boxplot(InternetACasa, main = "Boxplot famiglie con accesso ad Internet da casa")
```




```
> summary(InternetACasa)
##  Min.   1st Qu.  Median   Mean   3rd Qu.  Max.
## 73.60   80.42   82.80   82.23   84.22   88.90
```

Da questo possiamo ricavare che:

Q0= 73.60, **Q1**=80.42, **Q2**= 82.80, **Q3**=84.22, **Q4**=88.90

In R è anche possibile ottenere informazioni sull'esecuzione del comando boxplot attraverso **boxplot.stats()** che produce un dataframe con un vettore di lunghezza 5 composto rispettivamente da baffo inferiore, Q1, Q2, Q3, baffo superiore e inoltre il dataframe contiene un campo numerico che indica la presenza o meno di valori anomali (**outliers**)

```
BPstatsInternetACasa <- boxplot.stats(InternetACasa)
```

Quindi per trovare baffo inferiore e superiore:

```
> BPstatsInternetACasa$stats[1]
## 77.5
> BPstatsInternetACasa$stats[5]
## 88.9
```

Mentre il campo che indica gli outliers:

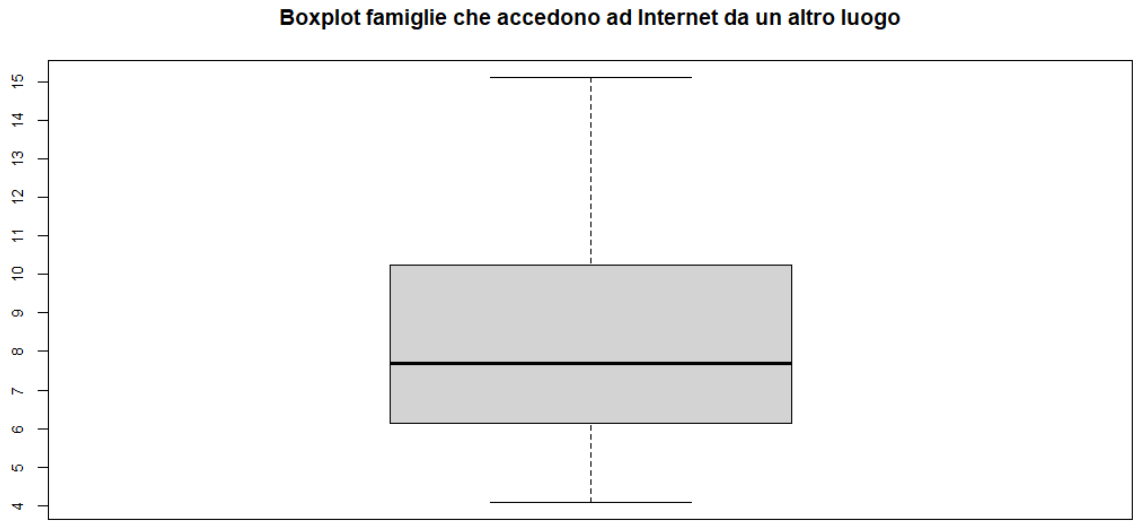
```
> BPstatsInternetACasa$out
## 73.6
```

Possiamo vedere che 73.6 corrisponde con il valore più basso della colonna, ovvero quello della Calabria che come mostrato più volte in passato è molto più basso degli altri.

Possiamo affermare che i dati sono *leggermente* asimmetrici in quanto Q3-Q2 = 1.42 < 2.38 = Q2-Q1 (una verifica più formale della simmetria avverrà più avanti tramite indice di asimmetria (skew))

3.5.2 Famiglie che accedono ad Internet da un altro luogo:

```
boxplot(InternetDaUnAltroLuogo, main = "Boxplot famiglie che accedono ad Internet
      da un altro luogo", axes=FALSE)
axis(2, at = seq(3, 15, by = 1), labels = seq(3, 15, by = 1), cex.axis=0.8)
box()
```




```
> summary(InternetDaUnAltroLuogo)
##   Min.   1st Qu.   Median    Mean   3rd Qu.   Max.
## 4.100   6.325    7.700    8.390  10.125  15.100
```

Da questo possiamo ricavare che:

Q0= 4.1, **Q1**=6.325, **Q2**= 7.7, **Q3**=10.125, **Q4**=15.1

```
BPstatsInternetDaUnAltroLuogo <- boxplot.stats(InternetDaUnAltroLuogo)
```

Quindi per trovare baffo inferiore e superiore:

```
> BPstatsInternetDaUnAltroLuogo$stats[1]
## 4.1
> BPstatsInternetDaUnAltroLuogo$stats[5]
## 15.1
```

Mentre il campo che indica gli outliers:

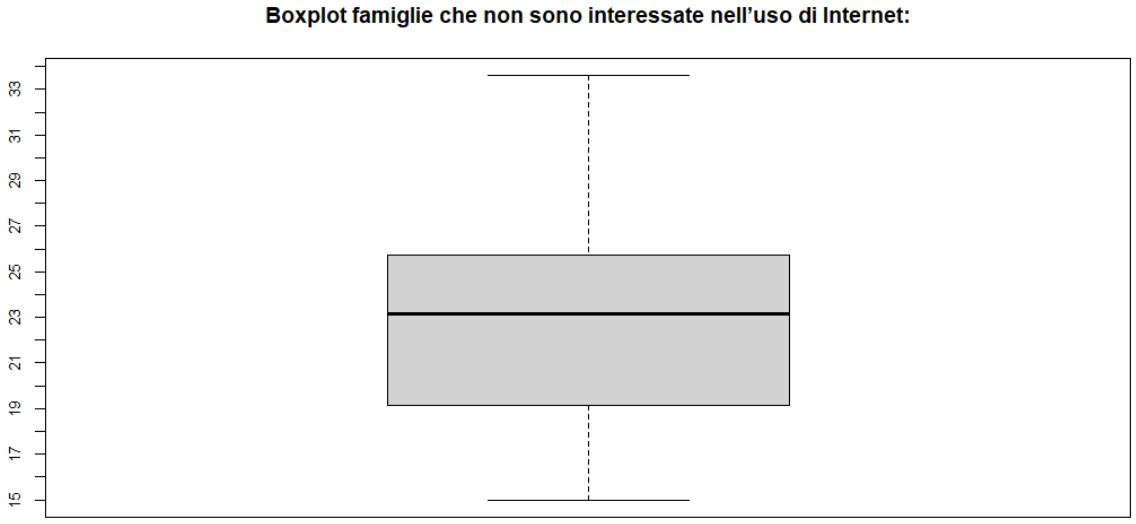
```
> BPstatsInternetACasa$out
## numeric(0)
```

Possiamo notare che i baffi corrispondono rispettivamente con Q0 e Q4; quindi, in questa tabella non sono presenti **outliers**

Possiamo affermare che i dati sono *leggermente* asimmetrici in quanto Q3-Q2 = 2.425 > 1.375 = Q2-Q1 (una verifica più formale della simmetria avverrà più avanti tramite indice di asimmetria (skew))

3.5.3 Famiglie che non sono interessate nell'uso di Internet:

```
boxplot(InternetNonInteressa, main = "Boxplot famiglie che non sono interessate
      nell'uso di Internet:", axes=FALSE)
axis(2, at = seq(3, 15, by = 1), labels = seq(3, 15, by = 1), cex.axis=0.8)
box()
```




```
> summary(InternetNonInteressa)
##   Min.   1st Qu.   Median    Mean   3rd Qu.   Max.
## 15.00    19.38    23.15    23.30    25.23    33.60
```

Da questo possiamo ricavare che:

Q0= 15, **Q1**=19.38, **Q2**=23.15, **Q3**=25.23, **Q4**=33.6

```
BPstatsInternetNonInteressa <- boxplot.stats(InternetNonInteressa)
```

Quindi per trovare baffo inferiore e superiore:

```
> BPstatsInternetNonInteressa$stats[1]
## 15
> BPstatsInternetNonInteressa$stats[5]
## 33.6
```

Mentre il campo che indica gli outliers:

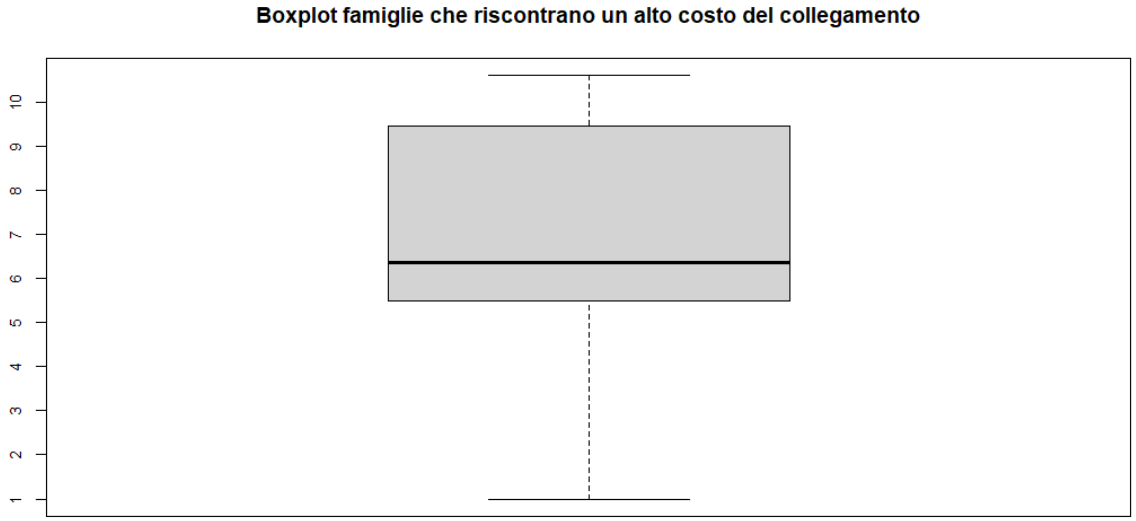
```
> BPstatsInternetNonInteressa$out
## numeric(0)
```

Possiamo notare che i baffi corrispondono rispettivamente con Q0 e Q4; quindi, in questa tabella non sono presenti **outliers**

Possiamo affermare che i dati sono *molto* asimmetrici in quanto Q3-Q2 = 2.08 < 3.77= Q2-Q1 (una verifica più formale della simmetria avverrà più avanti tramite indice di asimmetria (skew))

3.5.4 Famiglie che riscontrano un alto costo del collegamento:

```
boxplot(AltoCostoCollegamento, main = "Boxplot famiglie che riscontrano un alto
      costo del collegamento", axes=FALSE)
axis(2, at = seq(0, 11, by = 1), labels = seq(0, 11, by = 1), cex.axis=0.8)
box()
```




```
> summary(AltoCostoCollegamento)
##   Min.   1st Qu.   Median    Mean    3rd Qu.   Max.
## 1.000    5.550    6.350    6.810    9.425    10.600
```

Da questo possiamo ricavare che:

Q0= 1, **Q1**=5.55, **Q2**=6.35, **Q3**=9.425, **Q4**=10.6

```
BPstatsAltoCostoCollegamento <- boxplot.stats(AltoCostoCollegamento)
```

Quindi per trovare baffo inferiore e superiore:

```
> BPstatsAltoCostoCollegamento$stats[1]
## 1
> BPstatsAltoCostoCollegamento$stats[5]
## 10.6
```

Mentre il campo che indica gli outliers:

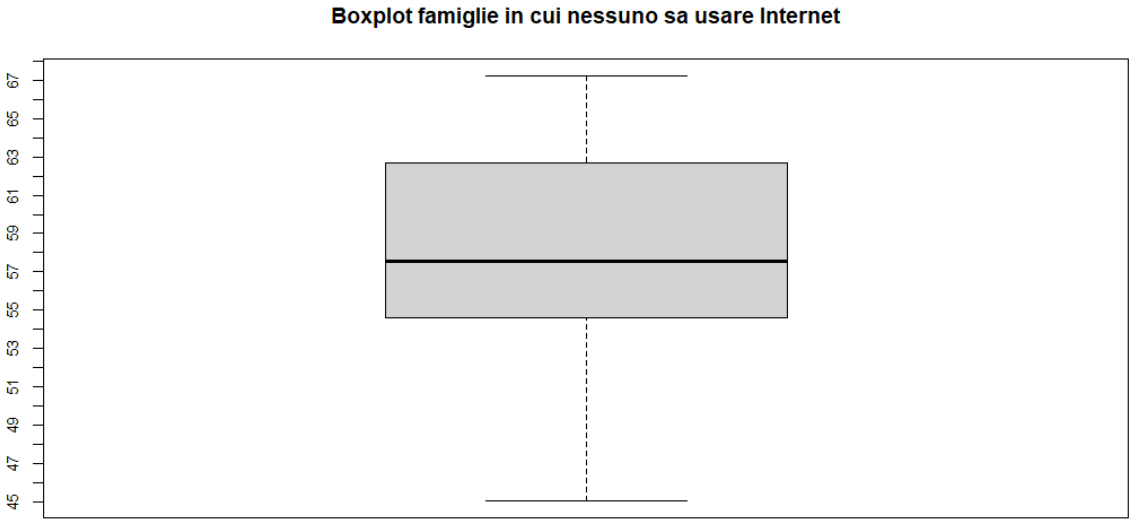
```
> BPstatsAltoCostoCollegamento$out
## numeric(0)
```

Possiamo notare che i baffi corrispondono rispettivamente con Q0 e Q4; quindi, in questa tabella non sono presenti **outliers**

Possiamo affermare che i dati sono *molto* asimmetrici in quanto Q3-Q2 = 3.075 > 0.8= Q2-Q1 (una verifica più formale della simmetria avverrà più avanti tramite indice di asimmetria (skew))

3.5.5 Famiglie in cui nessuno sa usare Internet:

```
boxplot(NonSaUsareInternet, main = "Boxplot famiglie in cui nessuno sa usare
      Internet", axes=FALSE)
axis(2, at = seq(40, 70, by = 1), labels = seq(40, 70, by = 1), cex.axis=0.8)
box()
```




```
> summary(NonSaUsareInternet)
##   Min.   1st Qu.   Median    Mean   3rd Qu.   Max.
## 45.10    54.75    57.55    58.06    62.60    67.20
```

Da questo possiamo ricavare che:

Q0= 45.10, **Q1**=54.75, **Q2**=57.55, **Q3**=62.60, **Q4**=67.20

```
BPstatsNonSaUsareInternet <- boxplot.stats(NonSaUsareInternet)
```

Quindi per trovare baffo inferiore e superiore:

```
> BPstatsNonSaUsareInternet$stats[1]
## 45.10
> BPstatsNonSaUsareInternet$stats[5]
## 67.20
```

Mentre il campo che indica gli outliers:

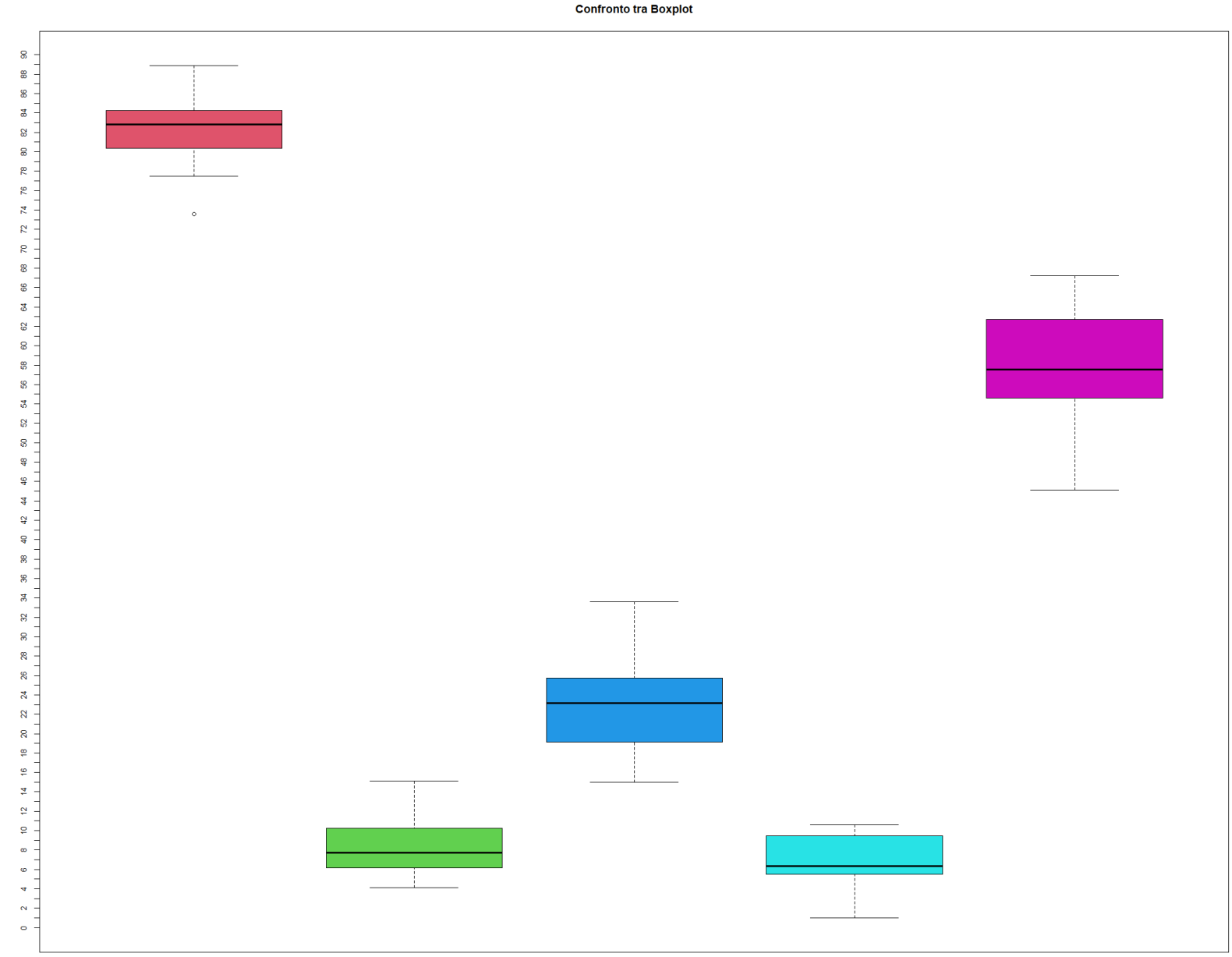
```
> BPstatsNonSaUsareInternet$out
## numeric(0)
```

Possiamo notare che i baffi corrispondono rispettivamente con Q0 e Q4; quindi, in questa tabella non sono presenti **outliers**

Possiamo affermare che i dati sono *molto* asimmetrici in quanto Q3-Q2 = 5.05 > 2.8 = Q2-Q1 (una verifica più formale della simmetria avverrà più avanti tramite indice di asimmetria (skew))

3.5.6 Confronto tra boxplot:

```
boxplot(InternetACasa, InternetDaUnAltroLuogo, InternetNonInteressa,
      AltoCostoCollegamento, NonSaUsareInternet, main = "Confronto tra Boxplot",
      axes=FALSE, col=2:6)
axis(2, at = seq(0, 90, by = 1), labels = seq(0, 90, by = 1), cex.axis=0.8)
box()
```

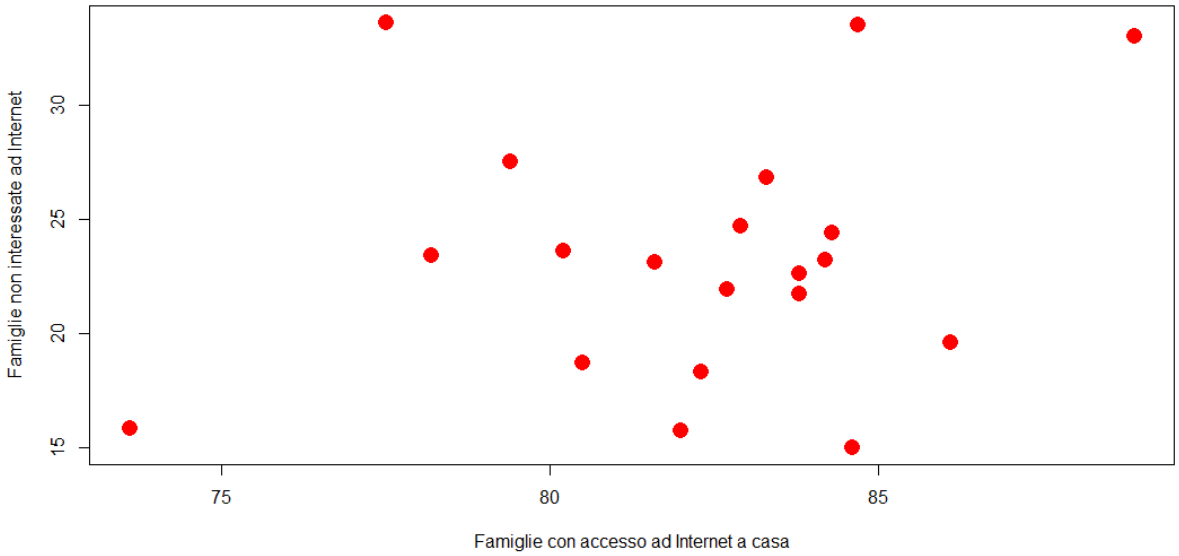


3.6 Grafico a dispersione

Un grafico a dispersione o **scatterplot** in statistica è un grafico che mostra come vengono distribuiti i valori di un vettore in un'asse cartesiano. Nel caso di uno scatterplot di un solo vettore, sull'asse delle x generalmente sono indicati gli indici e sull'asse y i valori. Mentre un uso più proficuo degli scatterplot è poter rappresentare le eventuali relazioni di due vettori.

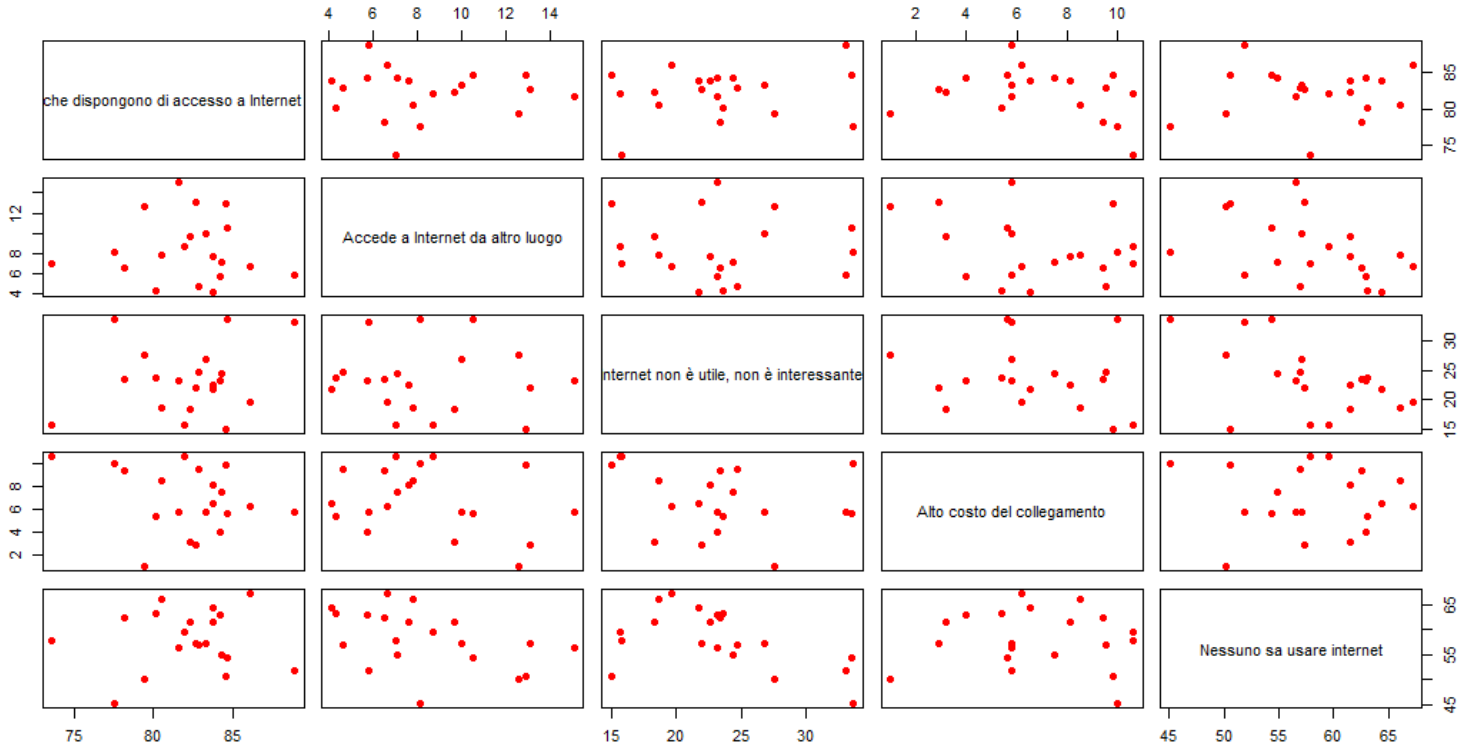
In R è possibile ottenere uno scatterplot attraverso l'uso della funzione **plot()**

Es:



Nel capitolo successivo vedremo come sfruttare questa rappresentazione per indicare la relazione tra i due vettori

Intanto, è possibile rappresentare tutti gli scatterplot possibili del dataset attraverso la funzione **pairs()**:



4. Statistica descrittiva

La statistica descrittiva è una branca della statistica che ha come obiettivo quello di sintetizzare quantitativamente le informazioni di un dataset.

Le misure utilizzate dalla statistica descrittiva sono principalmente:

- Indici di centralità
  - Media campionaria
  - Mediana campionaria
  - Moda
- Misure di dispersione (o variabilità)
  - Varianza
  - Deviazione standard
  - Indice di Asimmetria (Skewness)
  - Indice di Curtosi

Nel seguente capitolo procediamo all'analisi della statistica descrittiva univariata dei vari vettori del nostro dataset:

4.1 Statistica descrittiva univariata

La statistica descrittiva univariata descrive la distribuzione di una singola variabile e include gli indici di posizione centrali (media, mediana, moda) e non centrali (quantili: quartili, decili, percentali) e gli indici di dispersione (varianza, deviazione standard, coefficiente di variazione) che misurano quanto si disperdono i dati rispetto alla media. Descriviamo la forma della distribuzione invece attraverso gli indici di skewness e curtosi. Nel corso della trattazione dei vari indici di sintesi, questi saranno accompagnati con opportuni grafici.

4.1.1 Indici di sintesi

Gli indici di sintesi sono uno degli elementi principali della statistica descrittiva univariata, in quanto rendono possibile condensare i dati significativi di un intero data set in pochi valori significativi

4.1.1.1 Media campionaria

Si definisce media campionaria di un vettore, il valore ottenuto dalla media aritmetica del suddetto vettore.

Siano x1,x2,…,xn i valori di un vettore, la media campionaria è pari a:

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

In R è possibile calcolare la media campionaria di un vettore tramite il comando **mean():**

```
> mean(InternetACasa)
## 82.23
> mean(InternetDaUnAltroLuogo)
## 8.39
> mean(InternetNonInteressa)
## 23.305
> mean(AltoCostoCollegamento)
## 6.81
> mean(NonSaUsareInternet)
## 58.065
```

4.1.1.2 Mediana campionaria

Si definisce mediana campionaria di un vettore, il valore ottenuto dalla media aritmetica dei valori centrali del suddetto vettore.

Siano x1,x2,…,xn i valori di un vettore, la media campionaria è pari a:

$$\text{(n pari): } \tilde{x} = \frac{x_n + x_{n+1}}{2} \qquad \text{(n dispari): } \tilde{x} = x_{n+1}$$

In R è possibile calcolare la mediana campionaria di un vettore tramite il comando **median():**

```
> median(InternetACasa)
## 82.8
> median(InternetDaUnAltroLuogo)
## 7.7
> median(InternetNonInteressa)
## 23.15
> median(AltoCostoCollegamento)
## 6.35
> median(NonSaUsareInternet)
## 57.55
```

	Media	**Mediana**
Famiglie che dispongono di accesso Internet a casa	**82.23**	**82.8**
Famiglie che accedono ad Internet da un altro luogo	**8.39**	**7.7**
Famiglie non interessate all'uso di Internet	**23.305**	**23.15**
Alto costo del collegamento	**6.81**	**6.35**
Famiglie in cui nessuno sa usare Internet	**58.065**	**57.55**

Possiamo notare che in alcuni casi la media e la mediana sono molto simili, e generalmente quando queste due stime sono molto vicine la distribuzione di ogni vettore è simmetrica

4.1.1.2 Moda campionaria

Si definisce moda campionaria di un vettore, il valore con frequenza massima del vettore.

Se il valore è unico si dice **unimodale**, se sono due **bimodale**, se sono di più si dice **multimodale**

In R è possibile calcolare la mediana campionaria di un vettore tramite il comando **Mode():**

```
> Mode(InternetACasa)
## 83.8
> Mode(InternetDaUnAltroLuogo)
## 4.1
> Mode(InternetNonInteressa)
## 15
> Mode(AltoCostoCollegamento)
## 5.8
> Mode(NonSaUsareInternet)
## 61.5
```

4.1.2 Varianza e Deviazione standard campionaria

I soli dati della media, moda e in generale gli indici di sintesi permettono un'analisi scarna dalla precisione. Infatti, non tengono conto della variabilità dei dati. Serve quindi aggiungere degli indici ci quantificare la variabilità dei dati: **varianza campionaria** e **deviazione standard campionaria.**

La varianza campionaria si indica con s^2 ed è definita:

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

La deviazione standard campionaria invece si indica con **s** ed è definita come:

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

In R è possibile calcolare la varianza con **var()** e la deviazione standard con **sd()**, allora:

Varianza Campionaria:

```
> var(InternetACasa)
## 11.30326
> var(InternetDaUnAltroLuogo)
## 9.947263
> var(InternetNonInteressa)
## 30.79945
> var(AltoCostoCollegamento)
## 7.565158
> var(NonSaUsareInternet)
## 33.49924
```

Deviazione Standard Campionaria:

```
> sd(InternetACasa)
## 3.362033
> sd(InternetDaUnAltroLuogo)
## 3.153928
> sd(InternetNonInteressa)
## 5.549725
> sd(AltoCostoCollegamento)
## 2.750483
> sd(NonSaUsareInternet)
## 5.787853
```

Variazione e deviazione standard indicano quanto i dati si discostano dal valore medio del vettore. Però di per sé hanno senso solo nel contesto del proprio vettore, per rendere più significativo e utilizzabile si adopera il coefficiente di variazione, utile per fare confronti tra insiemi.

4.1.3 Coefficiente di variazione

Il coefficiente di variazione, come introdotto prima, è utile per effettuare confronti tra gli insiemi. È un numero puro adimensionale, ed è definito come il rapporto tra la deviazione standard e il modulo della media campionaria (non nulla):

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

Il coefficiente più è piccolo e meno sono dispersi i dati, mentre più è grande e più il valore medio è una misura poco significativa poiché c'è una grande dispersione dei dati nell'insieme.

In R non esiste una funzione per calcolare il coefficiente di variazione, però è possibile creare delle procedure ad hoc, in questo caso:

```
## cv <- function(x){
##       return( sd(x)/abs( mean(x) ) )
## }
```

```
cv(InternetACasa)
## 0.04088572
> cv(InternetDaUnAltroLuogo)
## 0.3759152
> cv(InternetNonInteressa)
## 0.2381345
> cv(AltoCostoCollegamento)
## 0.4038889
> cv(NonSaUsareInternet)
## 0.09967885
```

Utilizziamo una visualizzazione tabellare:

	Media	Varianza	Deviazione Standard	Coefficiente di Variazione
Famiglie con internet a casa	82.23	11.30326	3.362033	0.04088572
Famiglie che accedono ad internet da un altro luogo	8.39	9.947263	3.153928	0.3759152
Famiglie non interessate all'uso di internet	23.305	30.79945	5.549725	0.2381345
Famiglie che riscontrano un alto costo del collegamento	6.81	7.565158	2.750483	0.4038889
Famiglie in cui nessuno sa usare Internet	58.065	33.49924	5.787853	0.09967885

In quasi tutti i casi possiamo notare come il coefficiente di variazione sia abbastanza basso, escluse le famiglie che riscontrano un alto costo del collegamento. Il valore minore lo troviamo riguardante le famiglie che hanno disponibile l'accesso ad Internet da casa, ciò significa che quasi tutti i dati sono molto vicini al valore medio.

4.1.4 Indici di forma

Gli indici di forma sono utili, insieme ai valori di media, moda e mediana per definire la probabile forma della distribuzione. Prenderemo in analisi la **skewness** (**indice di asimmetria**) e la **indice di curtosi**

4.1.4.1 Indici asimmetria (Skewness)

L'indice di simmetria è un dato che indica quanto sia "sbilanciata" la distribuzione o quanto è centrata, una skewness maggiore di 0 indica una distribuzione con una coda più lunga a destra (asimmetria positiva)

Una skewness minore di 0 indica invece una distribuzione con una coda più luna a sinistra (asimmetria negativa)

Mentre più il valore di skewness è vicino allo 0 e più la distribuzione è simmetrica

L'indice di asimmetria γ definito su un insieme (x1,x2,…,xn), con media $\bar{x}$, e deviazione standard s:

$$\gamma = \frac{1}{s^3}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

In R è possibile calcolare l'indice di asimmetria attraverso **skewness()** della libreria **moments**

```
skewness(InternetACasa)
## -0.6273681
> skewness(InternetDaUnAltroLuogo)
## 0.5716145
> skewness(InternetNonInteressa)
## 0.4631557
> skewness(AltoCostoCollegamento)
## -0.3136
> skewness(NonSaUsareInternet)
## -0.4099556
```

Possiamo notare che nessuno di questi insiemi ha una simmetria perfetta, non si distaccano tanto da 0 quindi l'asimmetria non è così estremamente marcata. Il valore con l'indice di asimmetria più vicino a 0 è quello delle famiglie che riscontrano un alto costo del collegamento, che si "allontana" di soli 0,3 dal valore di simmetria perfetta

In generale in questa analisi possiamo notare che la maggior parte delle distribuzioni presenta un'asimmetria negativa

4.1.4.2 Indice di forma (curtosi)

L'indice di curtosi indica la densità dei dati attorno alla media, e quindi la "piattezza" o meno della distribuzione

L'indice può assumere 3 tipi di valori:

- >0
  - È presente un eccesso di dati nelle classi centrali (**leptocurtica**)
- <0
  - La distribuzione risulta piatta, carenza di dati nelle classi centrali (**platicurtica**)
- =0
  - Segue la distribuzione normale (**normocurtica**)

Si indica con k e per un insieme (x1,x2,…,xn), avente media $\bar{x}$ e deviazione standard campionaria s, è definita come segue:

$$k := \left[\frac{1}{s^4}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4\right] - 3$$

In R è possibile calcolare l'indice di forma attraverso **kurtosis()** della libreria **moments**

```
kurtosis(InternetACasa)
## 3.781581
> kurtosis(InternetDaUnAltroLuogo)
## 2.367809
> kurtosis(InternetNonInteressa)
## 2.563107
> kurtosis(AltoCostoCollegamento)
## 2.245235
> kurtosis(NonSaUsareInternet)
## 2.549307
```

Possiamo notare che nessuno di questi insiemi ha un indice di curtosi pari a 0. Possiamo affermare che tutti gli insiemi analizzati sono del tipo di distribuzione leptocurtica, avendo un eccesso di dati nelle classi centrali, specialmente nel caso delle famiglie che dispongono di accesso ad Internet da casa

4.2 Statistica descrittiva bivariata

Si definisce statistica bivariata la branca della statistica in cui si paragonano due insiemi per cercare relazioni tra loro

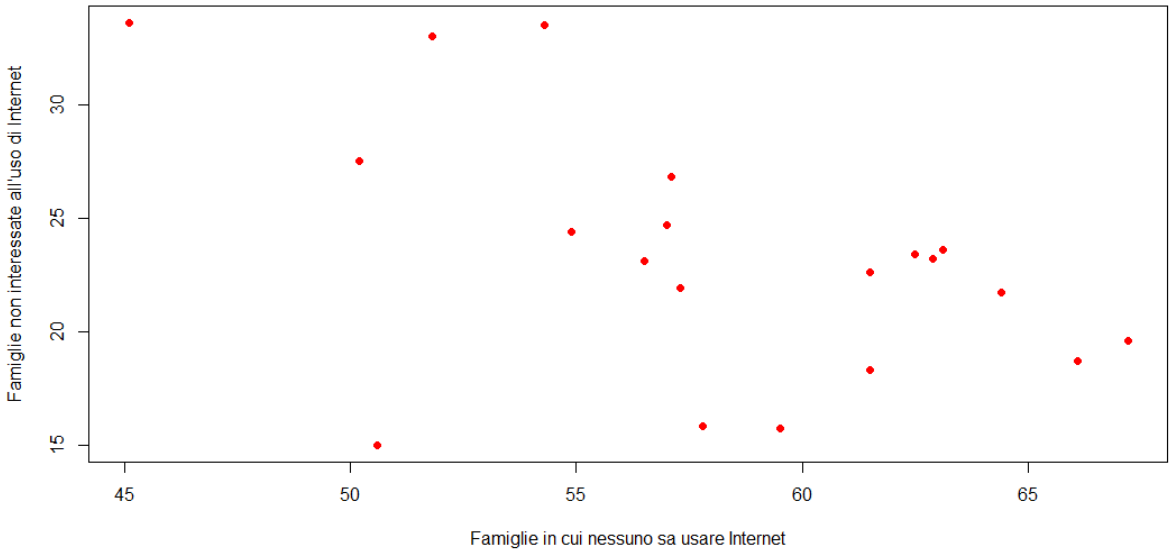
Per questa analisi è innanzitutto necessario definire uno scatterplot in cui sull'asse delle x verrà posta la variabile indipendente e sull'asse y quella dipendente.

L'analisi in questo caso verrà fatta tra: **famiglie in cui nessuno sa usare Internet** e **famiglie non interessate all'uso di Internet**

Di questi due insiemi sono noti i loro indici di sintesi:

	Famiglie in cui nessuno sa usare Internet	**Famiglie non interessate all'uso di Internet**
Media	58.065	23.305
Mediana	57.55	23.15
Deviazione Standard	5.787853	5.549725

E il loro scatterplot è:



4.2.1 Coefficiente di correlazione campionario

Il coefficiente di correlazione campionario è un numero adimensionale, necessario per quantificare la "forza" della relazione, e della disposizione lungo una retta dei valori dei due insiemi.

Può assumere valore tra -1 e 1

È definita come:

$$r := \frac{1}{(n-1)s_x s_y} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$
$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Per:

- $r > 0$
  - I dati si dicono correlati positivamente
- $r < 0$
  - I dati si dicono correlati negativamente

Sul coefficiente di correlazione c'è da dire che:

1. se esistono due numeri reali a e b, con $a > 0$, tali che $y_i = a_{xi} + b$ per ogni $i = 1, 2, .., n$ allora $r_{xy} = 1$
2. se esistono due numeri reali a e b, con $a < 0$, tali che $y_i = a_{xi} + b$ per ogni $i = 1, 2, .., n$ allora $r_{xy} = -1$
3. se esistono quattro numeri reali a, b, c e d, tali che $z_i = a_{xi} + b$ e $w_i = c_{yi} + d$ per $i = 1, 2, .., n$, allora $r_{zw} = r_{xy}$ se $ac > 0$ e $r_{zw} = -r_{xy}$ se invece $ac < 0$.

Il punto 1 e il 2 dimostrano che i valori limite -1 e +1 sono effettivamente raggiunti solo quando tra X e Y sussiste una relazione lineare, ossia quando i punti dello scatterplot giacciono tutti su di una retta.

Il punto 3 invece ci dice che il quadrato del coefficiente non cambia se sommiamo o moltiplichiamo costanti a tutti i valori di x e/o y.

In R è possibile calcolare il coefficiente di correlazione campionario eseguendo la funzione **cor()**:

```
cor(NonSaUsareInternet,InternetNonInteressa)
## -0.5221647
```

Da questo risultato possiamo affermare che tra i due vettori esiste una correlazione negativa mediamente forte

La disposizione dei valori lungo una retta verrà approfondita nel paragrafo successivo

4.2.2 Regressione lineare

Si definisce regressione lineare l'approccio predittivo basato su una variabile dipendente e una o più variabili indipendenti. In presenza di una sola variabile indipendente si parla di **regressione lineare semplice**

4.2.2.1 Regressione lineare semplice

Il modello di regressione lineare semplice è esprimibile attraverso un'equazione di una retta:

$$\boldsymbol{Y = g(x) + \varepsilon \; = \; a + bX + \varepsilon}$$

In cui:

- a è l'**intercetta**, ovvero, l'ordinata con cui l'asse di regressione si interseca
- b è il **coefficiente angolare**
  - >0 regressione crescente
  - <0 regressione decrescente
  - =0 retta orizzontale
- ε è una **variabile aleatoria avente**
  - Valore medio E[ε] = 0
  - Varianza Var[ε] = σ^2

Per ottenere la retta di regressione si fa ricorso al metodo dei minimi quadrati (in inglese Ordinary Least Squares (OLS) method), che consiste nell'attribuire ai coefficienti a e b i valori che rendono minima la somma

$$Q(a,b) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + b\,x_i)]^2$$

$$\text{con } e_i = y_i - (a + b\,x_i),\ i = 1, 2, \ldots, n.$$

Derivando Q rispetto α e β e ponendo le derivate parziali ottenute a 0, il metodo dei minimi quadrati conduce a:

$$b = \frac{s_{xy}}{s_{xx}} r_{xy} \qquad\qquad a = \bar{y} - b\bar{x}$$

segue che la retta ai minimi quadrati $y = \hat{a} + \hat{b}x$ è data da

$$y \; = \; \bar{y} \; + \; \hat{b}\,(x \; - \; \bar{x}).$$

In R:

```
beta <-d(InternetNonInteressa)/sd(NonSaUsareInternet))*cor(NonSaUsareInternet,
                                                          InternetNonInteressa)
alpha <-mean(InternetNonInteressa)-beta*mean(NonSaUsareInternet)

c(alpha,beta)
[1] 52.3770699 -0.5006815
```

Possiamo notare come beta essendo negativo, la retta sarà discendente, inoltre, conoscendo alfa possiamo dire dove la retta di regressione interseca l'asse delle ordinate

In R è possibile anche riassumere il calcolo di alfa e beta attraverso la funzione **lm(y~x)** in cui specifichiamo che y dipenda da x, allora:

```
linearModel <- lm(InternetNonInteressa ~ NonSaUsareInternet)
## Coefficients:
##        (Intercept)  NonSaUsareInternet
##            52.3771             -0.5007
```

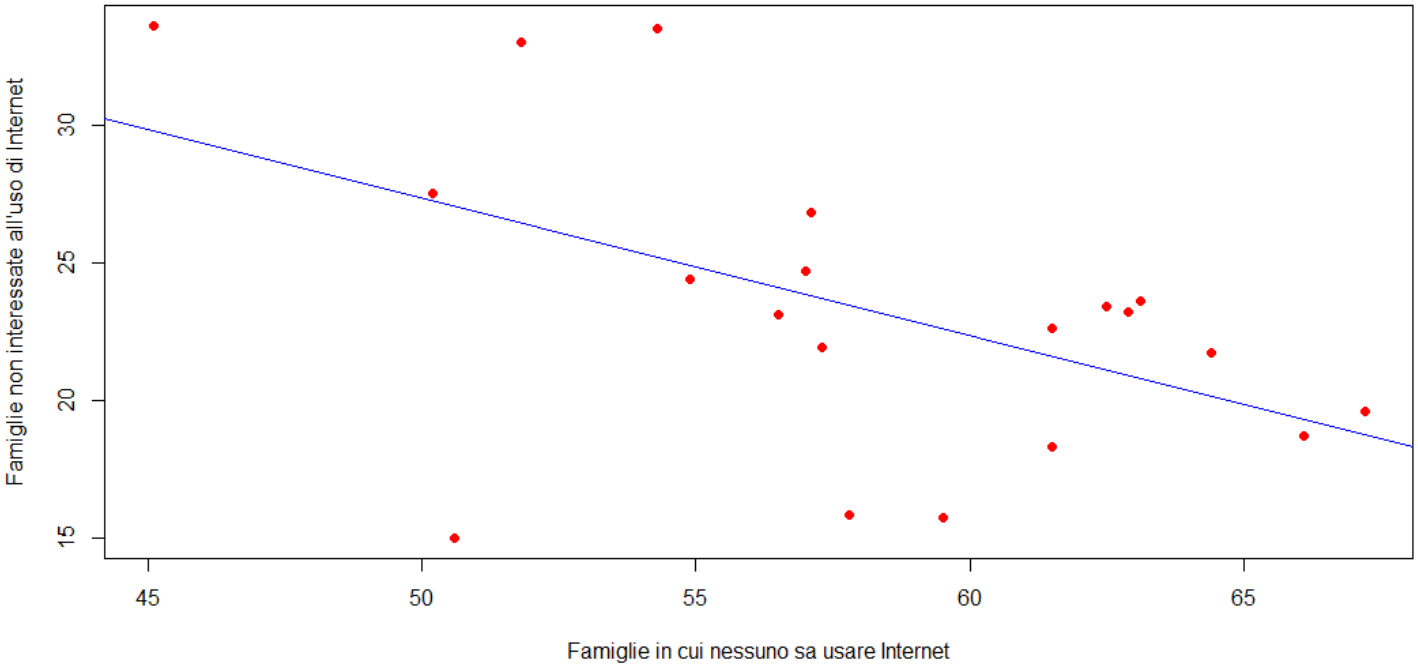
Memorizzando il risultato di lm in linearModel, i coefficienti che compongono la retta di regressione sono memorizzati in linearModel$coefficients, allora:

```
linearModel$coefficients
       (Intercept) NonSaUsareInternet
         52.3770699         -0.5006815
```

La retta di regressione allora è:

$$y = 52.3770699 - 0.5006815x$$

Possiamo disegnarla attraverso la funzione **abline()**, per rendere più chiaro il grafico della retta di regressione la disegniamo nello scatterplot dei vettori studiati:



4.2.2.1.1 Residui

Si definisce **residuo** la differenza tra valore osservato e valore stimato, sono utilizzati per valutare l'accuratezza del modello statistico.

I valori stimati sono espressi tramite l'equazione

$$\hat{y} = a + bX_i$$

Ottenuti tramite la retta di regressione. Risulta inoltre che il valore medio dei valori stimati è lo stesso dei valori osservati.

I residui sono quindi definiti:

$$E_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

La varianza dei residui è $\boldsymbol{S_E^2 = \frac{1}{n-1}\sum_{i=1}^{n} E_i^2}$ in quanto la media campionaria è pari a 0

In R il vettore dei residui è contenuto nel risultato di **lm(y~x)** sotto l'alias di $residuals

Giardinetto Giuseppe matricola n°. 0512114655

Quindi:

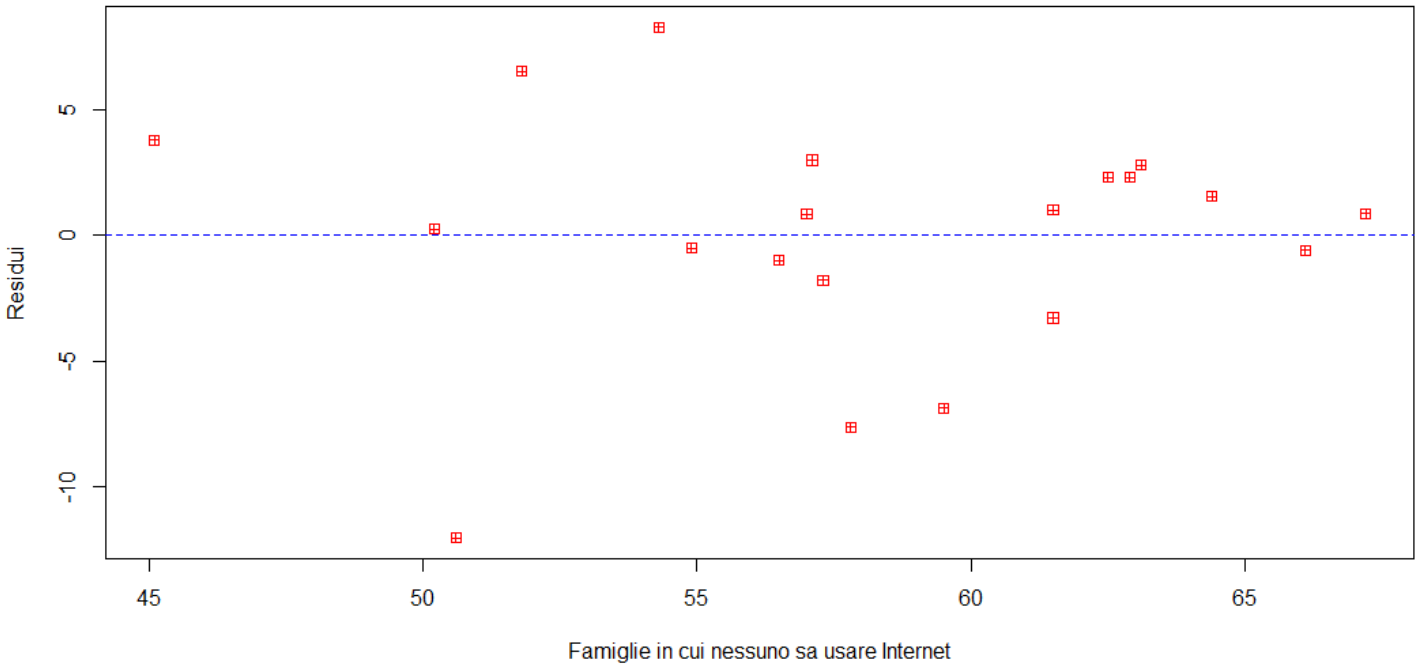
```
linearModel$residuals
```

```
        1         2         3         4         5         6         7 
3.0118424 0.2571402 0.8617742 0.8687253 6.5582305 1.0148409 8.3099342 

        8         9        10        11        12        13        14 
1.5668172 -0.4896569 -1.7880213 2.3157949 -12.0425872 -3.2851591 -0.5820243 

       15        16        17        18        19        20 
-6.8865221 2.3155223 3.8036647 -7.6376806 2.8159312 -0.9885665
```

Dei residui è possibile calcolare **mediana**, **varianza**, **deviazione standard**:

```
median(linearModel$residuals)
## [1] 0.8652498
var(linearModel$residuals)
## [1] 22.40179
sd(linearModel$residuals)
## [1] 4.733053
```

Possiamo rappresentare i residui in un grafico dei residui, sull'asse x poniamo i valori della variabile indipendente e sull'asse y i valori dei residui:



In R:

```
plot(NonSaUsareInternet,linearModel$residuals, pch=12, col="red", xlab="Famiglie in
cui nessuno sa usare Internet",ylab="Residui")
abline(lm(linearModel$residuals~NonSaUsareInternet), col="blue", lty=2)
```

Per rendere significativi e comparabili i residui, è necessario calcolarne la controparte standardizzata, definita come segue:

$$E_i^{(s)} = \frac{E_i - \bar{E}}{s_E}$$

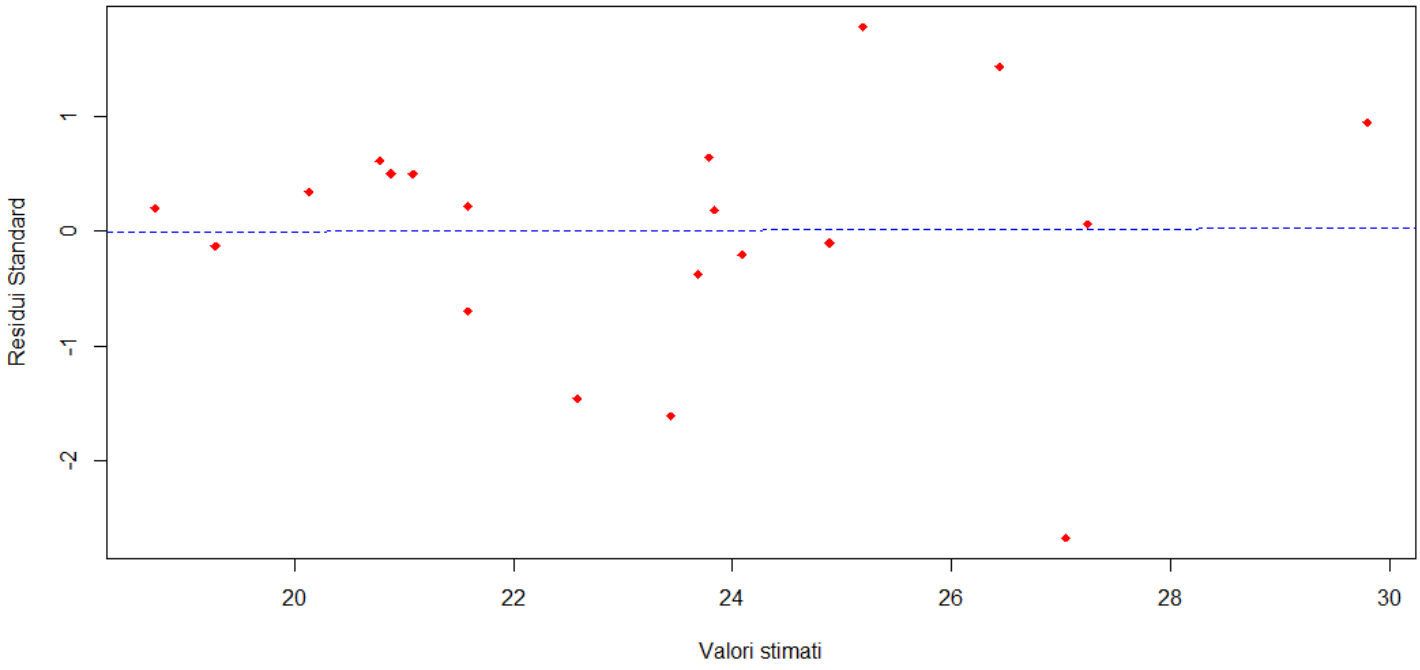
Il residuo standardizzato è caratterizzato da:

- Hanno media pari a 0
- Variazione standard pari ad 1

In R è possibile calcolare i residui attraverso la funzione **rstandard()**:

```
rstandard(linearModel)
     1          2          3          4          5          6          7
0.63595030  0.05726129  0.18199416  0.19741819  1.43092498  0.21623872  1.77421030
     8          9         10         11         12         13         14
0.34212712 -0.10417791 -0.37743243  0.49833137 -2.66668460 -0.69999013 -0.12993207
    15         16         17         18         19         20
-1.45544894  0.49669104  0.94446874 -1.61154696 0.60698624 -0.20899866
```

Rappresentiamo in un grafico, sull'asse delle x i valori stimati e sull'asse y i residui standard:



4.2.2.1.2 Coefficiente di determinazione

Il coefficiente di determinazione è utile per comprendere quanto è precisa l'approssimazione attraverso una rappresentazione tramite retta di regressione, questo assume valori tra 0 e 1, più si avvicina all'1 e meglio vengono rappresentati i valori per via di una retta, mentre più si avvicina a 0 e più è difficile rappresentare i valori per via di una retta di regressione. È inoltre un valore adimensionale, definito come segue:

$$D^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

In R è possibile calcolare il coefficiente di determinazione come segue:

```
summary(linearModel)$r.square
## 0.272656
```

Possiamo notare che il coefficiente di determinazione è molto vicino allo 0, quindi serve una rappresentazione più precisa

4.2.3 Regressione lineare multivariata

In molti casi, come questo, è utile utilizzare più variabili indipendenti per ottenere una rappresentazione molto più affidabile.

Utilizziamo le funzioni **cov()** e **cor()** di R per poter calcolare, rispettivamente, le **covarianze** e i **coefficienti di correlazione** di tutte le possibili coppie della matrice:

```
cov(matriceFamiglie)
```

produce la seguente tabella:

	Famiglie che dispongono di accesso Internet a casa	Famiglie che accedono ad Internet da un altro luogo	Famiglie non interessate all'uso di Internet	Famiglie che riscontrano un alto costo del collegamento	Famiglie in cui nessuno sa usare Internet
Famiglie che dispongono di accesso Internet a casa	11.303263	-0.456000	3.736158	-2.685579	1.450053
Famiglie che dispongono di accesso Internet a casa	-0.4560000	9.9472632	-0.9599474	-2.6172632	-7.8908947
Famiglie che dispongono di accesso Internet a casa	3.7361579	-0.9599474	30.7994474	-4.0432105	-16.7724474
Famiglie che dispongono di accesso Internet a casa	-2.6855789	-2.6172632	-4.0432105	7.5651579	-0.6996316
Famiglie che dispongono di accesso Internet a casa	1.4500526	-7.8908947	-16.7724474	-0.6996316	33.4992368

In questa tabella possiamo notare che sulla diagonale principale sono presenti le varianze di ogni vettore

```
cor(matriceFamiglie)
```

produce la seguente tabella:

	Famiglie che dispongono di accesso Internet a casa	Famiglie che accedono ad Internet da un altro luogo	Famiglie non interessate all'uso di Internet	Famiglie che riscontrano un alto costo del collegamento	Famiglie in cui nessuno sa usare Internet
Famiglie che dispongono di accesso Internet a casa	1	-0.04300422	0.20024046	-0.29042032	0.07451855
Famiglie che accedono ad Internet da un altro luogo	-0.04300422	1	-0.05484337	-0.30170784	-0.43227188
Famiglie non interessate all'uso di Internet	0.20024046	-0.05484337	1	-0.26487800	-0.52216471
Famiglie che riscontrano un alto costo del collegamento	-0.29042032	-0.30170784	-0.26487800	1	-0.04394839
Famiglie in cui nessuno sa usare Internet	0.07451855	-0.43227188	-0.52216471	-0.04394839	1

Possiamo notare come sulla diagonale principale siano disposti solo 1, inoltre, possiamo notare che i coefficienti (in valore assoluto) più alti sono quelli tra **famiglie non interessate all'uso di Internet** e **Famiglie in cui nessuno sa usare Internet** (caso studiato) e **famiglie che accedono ad Internet da un altro luogo** e **famiglie in cui nessuno sa usare Internet** mentre tutti gli altri casi sono coefficienti relativamente bassi. È interessante notare come la maggior parte dei coefficienti di correlazione è minore di 0

Il modello di regressione multivariata è definito dalla seguente equazione:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

in cui:

- a è l'intercetta
- b1, b2, …, bp sono i **regressori**

Anche in questo caso è possibile utilizzare il metodo dei minimi quadrati per ricavare i coefficienti dell'intercetta e dei vari regressori, minimizzando la seguente quantità:

$$Q = \sum_{i=1}^{n} [y_i - (\alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p})]^2$$

In R è possibile calcolare i coefficienti attraverso la funzione **lm()**:

```
linearModelMultiplo <- lm(InternetNonInteressa ~ NonSaUsareInternet+
                                                 InternetACasa+
                                                 InternetDaUnAltroLuogo+
                                                 AltoCostoCollegamento)
```

Specificando che InternetNonInteressa dipende da tutte le altre variabili indipendenti

```
linearModelMultiplo
##
## Call:
## lm(formula = InternetNonInteressa ~ NonSaUsareInternet + InternetACasa +
##     InternetDaUnAltroLuogo + AltoCostoCollegamento)
##
## Coefficients:
##         (Intercept)    NonSaUsareInternet    InternetACasa   InternetDaUnAltroLuogo
##             63.9238               -0.7374           0.1881                  -0.8953
##       AltoCostoCollegamento
##             -0.8456
```

L'equazione della curva è quindi:

$$y = 63.9238 - 0,7374x_1 - 0,1881x_2 - 0,8953x_3 - 0,8456x_4$$

Possiamo notare che tutti i regressori sono negativi significando che tutte le altre variabili sono legate negativamente alla percentuale alle famiglie a cui non interessa usare Internet

4.2.3.1 Residui (regressione multivariata)

Analogamente alla definizione precedente della regressione lineare semplice, mostrano di quanto si discostano i valori dalla curva di regressione, I residui sono quindi definiti:

$$E_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

La varianza dei residui è $S_E^2 = \frac{1}{n-1}\sum_{i=1}^{n} E_i^2$ in quanto la media campionaria è pari a 0

In R i valori stimati e i residui sono entrambi contenuti in **linearModelMultiplo**, rispettivamente sotto, **$fitted.values** e **$residuals**

```
linearModelMultiplo$residuals
    1          2          3          4          5          6          7          8
 3.1695484 -2.2159166 -0.6348853  0.1845726  0.6477243  1.9161277  7.8200762 -1.3320708
    9          10         11         12         13         14         15         16
-2.1996651 -1.1469170 -1.6949159 -7.6890652 -4.3650755  2.5461245 -3.0211573  4.6213292
    17         18         19         20
 4.0624568 -4.1166927 -0.4646153  3.9130172
```

Conoscendo i valori dei residui, ci è possibile calcolarne **mediana**, **varianza** e **deviazione standard**

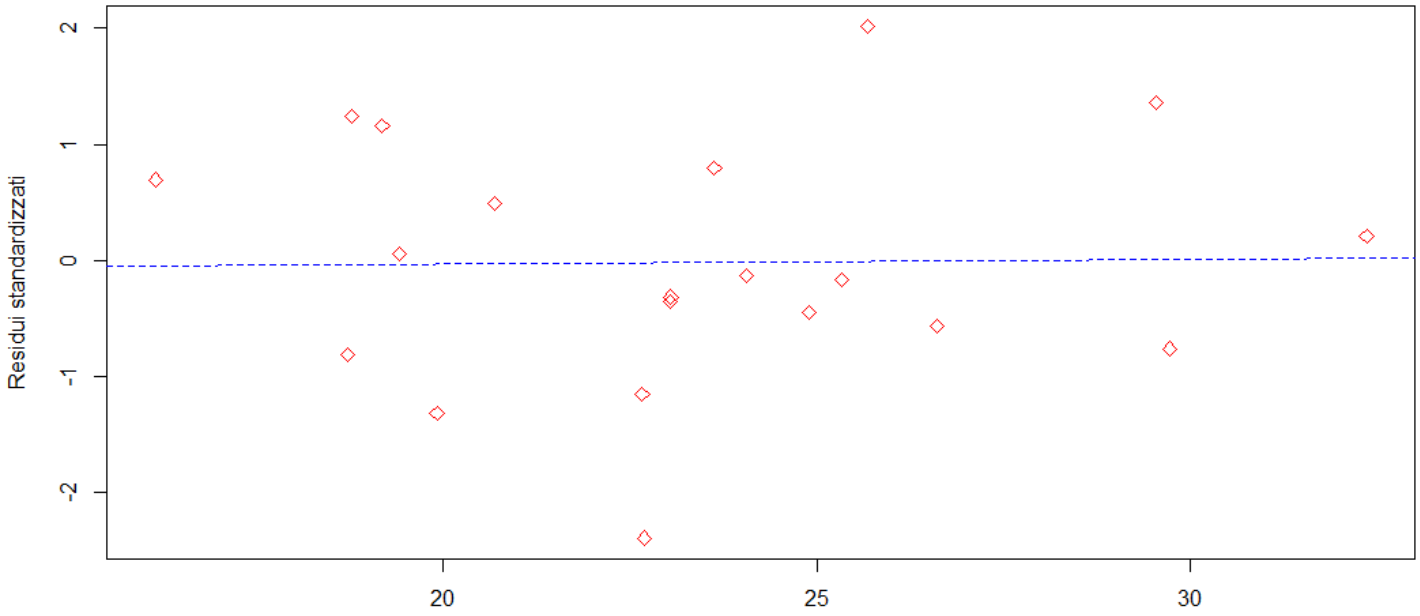
```
median(linearModelMultiplo$residuals)
## [1] -0.5497503
var(linearModelMultiplo$residuals)
## [1] 13.45052
sd(linearModelMultiplo$residuals)
## [1] 3.667496
```

Anche in questo caso è utile calcolare i residui standardizzati:

```
rstandard(linearModelMultiplo)
    1          2          3          4          5          6          7
0.79652059 -0.76259545 -0.16968380  0.05133502  0.21123613  0.49038201  2.01433265
    8          9          10         11         12         13         14
-0.35586518 -0.56701732 -0.31666000 -0.45546440 -2.39667460 -1.15952571  0.69462307
     15         16         17         18         19         20
-0.81637522  1.24115763  1.35857741 -1.31684624 -0.12922364  1.15460814
```

Giardinetto Giuseppe matricola n°. 0512114655

Disegniamo quindi il grafico dei residui:

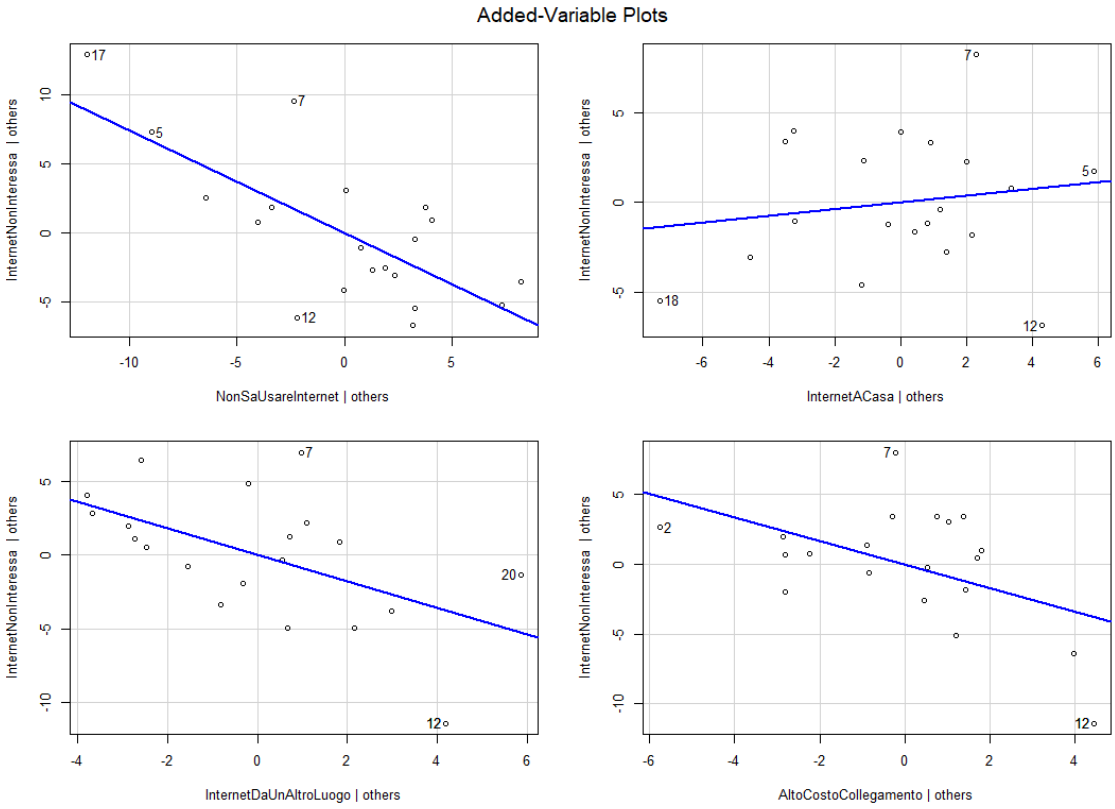



```
plot(linearModelMultiplo$fitted.values,rstandard(linearModelMultiplo), xlab="Valori
stimati", ylab="Residui standardizzati", col="red", pch=5)

abline(lm(rstandard(linearModelMultiplo)~linearModelMultiplo$fitted.values), col="b
lue", lty=2)
```

Calcoliamo quindi il coefficiente di correlazione per verificare la precisione della rappresentazione:

```
summary(linearModelMultiplo)$r.square
## 0.5632868
```

Possiamo notare che la rappresentazione in regressione multivariata offre una rappresentazione molto più affidabile di quella in regressione lineare semplice, di lato i grafici:



4.2.4 Regressione non lineare

Spesso una rappresentazione lineare non è adatta per la rappresentazione di un set di dati, riguardando la rappresentazione lineare semplice e il suo coefficiente di correlazione, possiamo affermare che è relativamente basso, e quindi inadatto ad una rappresentazione affidabile:

```
summary(linearModel)$r.square
## 0.272656
```

Attraverso alcune trasformazioni è possibile però linearizzare modelli che sembrano non lineari, questo ci permette di usare comunque un modello lineare. Consideriamo dunque il modello non lineare:

$$y = a + bX + \gamma X^2$$

Possiamo ricorrere alla regressione multivariata considerando $X_1 = X$ e $X_2 = X^2$

$$y = a + bX_1 + \gamma X_2$$

In R:

```
RegressionePolinomiale <- lm(InternetNonInteressa~NonSaUsareInternet+
                                                  I((NonSaUsareInternet)^2) )
```

```
RegressionePolinomiale
## Call:
## lm(formula = InternetNonInteressa ~ NonSaUsareInternet + 
##                                       I((NonSaUsareInternet)^2) )
##
## Coefficients:
##                (Intercept)         NonSaUsareInternet  I((NonSaUsareInternet)^2)
##                  119.07476                   -2.86896                    0.02081
```

Giardinetto Giuseppe matricola n°. 0512114655

Quindi l'equazione della curva è:

$$y = 119.07876 - 2.86896X + 0.02081X^2$$

Calcoliamo il coefficiente di correlazione:

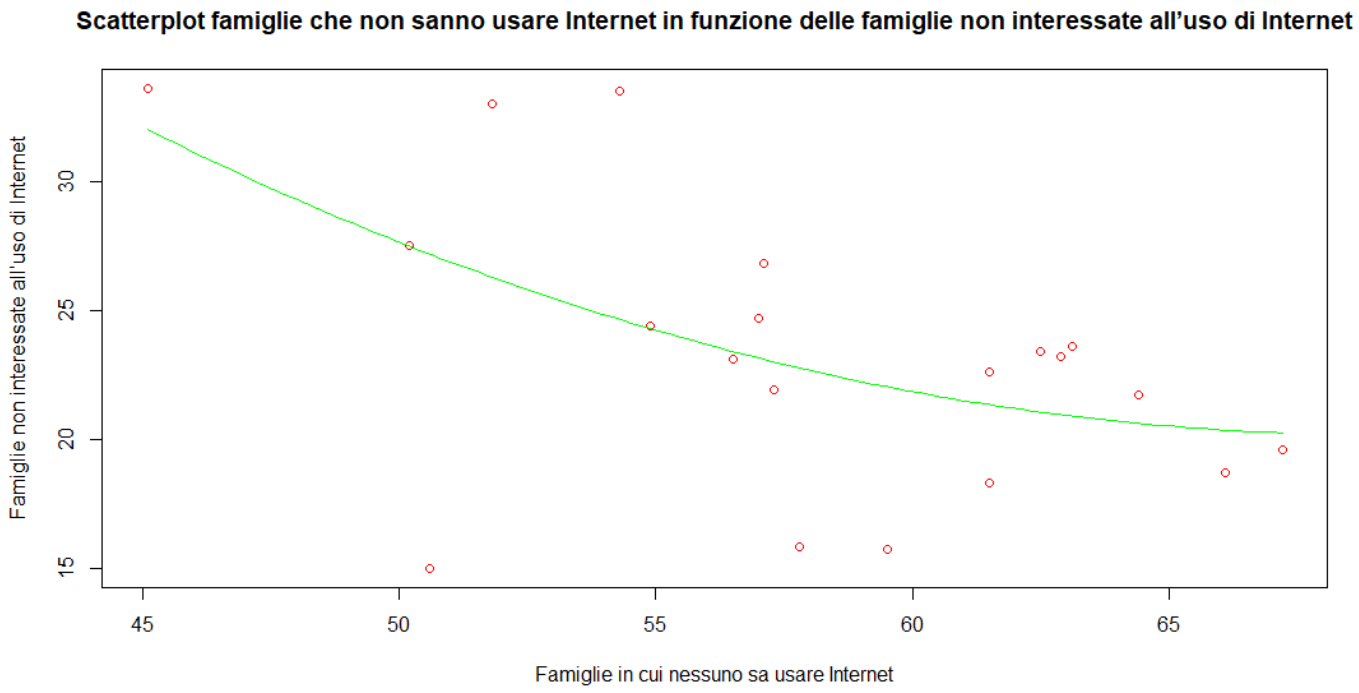
```
summary(RegressionePolinomiale)$r.square
## 0.2933559
```

Notiamo che è comunque un risultato di bassa affidabilità ma leggermente migliore di quello della rappresentazione lineare semplice

Giardinetto Giuseppe matricola n°. 0512114655

Disegniamo il grafico in R:

```
plot(NonSaUsareInternet,InternetNonInteressa,main="Scatterplot famiglie che non san
no usare Internet in funzione delle famiglie non interessate all'uso di Internet" ,
xlab="Famiglie in cui nessuno sa usare Internet",ylab="Famiglie non interessate all
'uso di Internet" , col = "red")
alpha <- regressionePolinomiale$coefficients[[1]]
beta <- regressionePolinomiale$coefficients[[2]]
gamma <- regressionePolinomiale$coefficients[[3]]
curve (alpha+beta*x+gamma*x^2, add=TRUE, col = "green")
```



Ora, il grafico di regressione lineare semplice per un confronto: