Università degli Studi di Salerno

Corso di Statistica e Analisi dei Dati

Persone di 3 anni e più che svolgono/non svolgono attività sportiva

Progetto di Selice Andrea e Nisivoccia Giuseppe

INDICE

1 - INTRODUZIONE

La seguente indagine statistica analizza i dati forniti dall'ISTAT sulla popolazione italiana in merito allo sport e all'attività fisica.
L'obiettivo di questa indagine è quello di analizzare le percentuali, per ogni regione, di chi pratica sport e attività fisica, con quale frequenza e di chi invece non pratica suddette attività.
Questa indagine fa parte del gruppo "Aspetti della vita quotidiana" che a sua volta fa parte di un sistema integrato di indagini sociali che rileva le informazioni fondamentali relative alla vita quotidiana degli individui e delle famiglie.
Le informazioni raccolte consentono di conoscere le abitudini dei cittadini e i problemi che essi affrontano ogni giorno.
L'indagine rientra tra quelle comprese nel Programma statistico nazionale, che raccoglie l'insieme delle rilevazioni statistiche necessarie al Paese.

1.1 - Dataset

La fonte del dataset analizzato è il sito dell'Istituto Nazionale di Statistica.
Al suo interno troviamo i dati raccolti su un campione di migliaia di individui per ciascuna regione italiana, per capire la frequenza con cui praticano sport e attività fisica oppure se non praticano nessuno delle due opzioni.
Il campione considera le persone che hanno dai tre anni in su.
Per ciascuna regione è indicata la percentuale delle persone che:

- praticano sport in modo continuativo
- praticano sport in modo saltuario
- praticano solo qualche attività fisica
- non praticano sport né attività fisica.

I dati sono relativi al periodo del 2021.

	in modo continuativo	in modo saltuario	solo qualche attività fisica	non praticano sport né attività fisica
Piemonte	26	12,2	31,5	30,2
Valle d'Aosta	32,5	14	33,9	19,6
Liguria	23,1	13,1	38,6	25,2
Lombardia	28	13	36,4	22,6
Trentino Alto Adige	39,8	14,4	32,2	13,5
Veneto	27,8	14,4	34,4	23,4
Friuli-Venezia Giulia	24,2	13,4	37,4	25
Emilia-Romagna	28	11	34,2	26,8
Toscana	26,5	12,5	34	26,8
Umbria	23,7	10,7	33	32,6
Marche	25,5	10,7	33,2	30,5
Lazio	25,8	9,8	31,8	32,6
Abruzzo	23,5	11,8	31	33,7
Molise	15,4	8,2	30,2	46,2
Campania	14,3	6,5	26,3	52,8
Puglia	17,9	10,6	24,4	47,2
Basilicata	16	8,7	25,3	50
Calabria	15,8	6,7	27,9	49,7
Sicilia	15,6	7,6	24,6	52,2
Sardegna	22,3	9,6	36,4	31,6

1.1.1 - Inizializzazione della matrice

Per realizzare il seguente progetto è stato utilizzato il linguaggio di programmazione R e l'ambiente di sviluppo R Studio, in quanto è il più adatto per la Data Analysis. Per poter iniziare l'analisi il primo passo da compiere è l'inizializzazione della matrice che verrà usata nel corso dell'indagine statistica.
Iniziamo con la creazione di vettori per ciascuna colonna presente nel dataset:

```
sport_in_modo_continuativo<-c(26, 32.5, 23.1, 28, 39.8, 27.8, 24.2, 28, 26.5, 23.7, 25.5, 25.8, 23.5, 15.4, 14.3, 17.9, 16, 15.8, 15.6, 22.3)

sport_in_modo_saltuario <- c(12.2, 14, 13.1, 13, 14.4, 14.4, 13.4, 11, 12.5, 10.7, 10.7, 9.8, 11.8, 8.2, 6.5, 10.6, 8.7, 6.7, 7.6, 9.6)

sport_solo_qualche_attivita <- c(31.5, 33.9, 38.6, 36.4, 32.2, 34.4, 37.4, 34.2, 34, 33, 33.2, 31.8, 31, 30.2, 26.3, 24.4, 25.3, 27.9, 24.6, 36.4)

non_praticano_sport <- c(30.2, 19.6, 25.2, 22.6, 13.5, 23.4, 25, 26.8, 26.8, 32.6, 30.5, 32.6, 33.7, 46.2, 52.8, 47.2, 50, 49.7, 52.2, 31.6)
```

A questo punto viene creata la matrice inserendo i vettori sopra citati, successivamente si aggiungono altri due vettori rispettivamente con i nomi delle regioni e con le frequenze relative alla pratica di sport, per definire i nomi delle righe e delle colonne

```
matrice_analisi_sport <- cbind(sport_in_modo_continuativo, sport_in_modo_saltuario, sport_solo_qualche_attivita, non_praticano_sport)

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")
row.names(matrice_analisi_sport) <- regioni

frequenza_sport <- c("In modo continuativo", "In modo saltuario", "Solo qualche attività fisica", "Non praticano sport nè attività fisica")

colnames(matrice_analisi_sport) <- frequenza_sport
```

2 - DISTRIBUZIONI DI FREQUENZA

Dopo aver predisposto i dati da analizzare, bisogna organizzarli in maniera che siano significativi, in modo da poter estrapolare le informazioni non immediatamente evidenti.
Il primo metodo per organizzare i dati è la distribuzione di frequenza.
La distribuzione di frequenza è un'organizzazione dei dati in maniera tabulare che permette di rappresentare i valori individuali dei dati in una categoria su una determinata scala.
Attraverso la distribuzione di frequenza si ha uno sguardo all'intero dataset e si intuisce se determinate osservazioni sono più o meno significative e come esse sono distribuite sulla scala considerata.
Nel campione preso in analisi, i dati a disposizione sono percentuali numeriche il cui contenuto nella tabella non può assumere valori che possano essere classificabili all'interno di una ben precisa modalità.
La soluzione ideale è quella di raccogliere le informazioni in classi e calcolare le frequenze con cui gli elementi del campione cadono in ciascuna di esse.
Analizzando i dati presenti nel campione sono state scelte le seguenti classi:

(0,11] (11,22] (22,33] (33,44] (44,55]

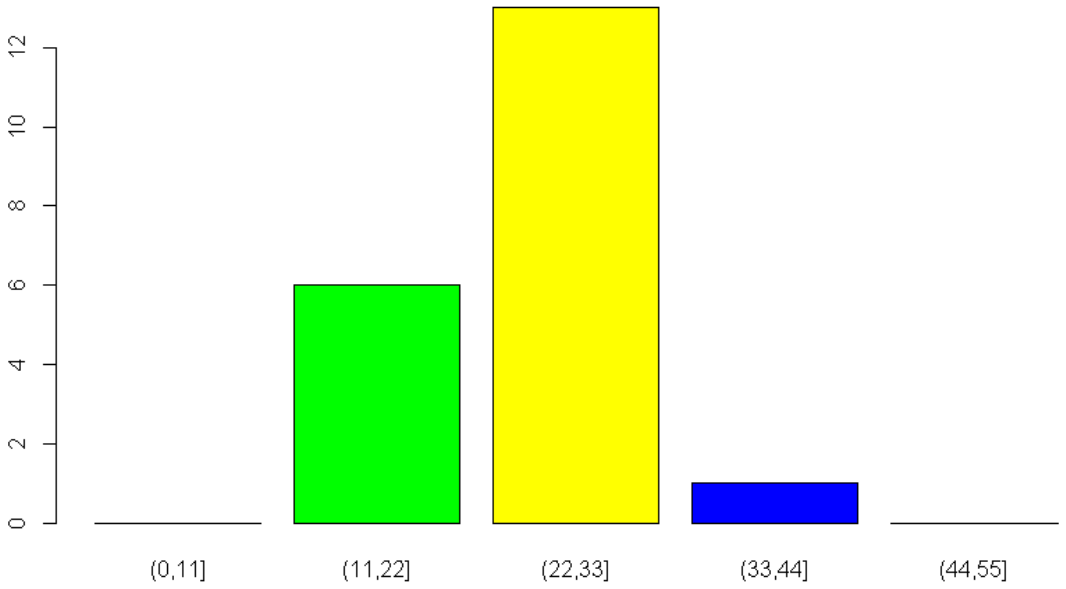
```
classi <- c(0,11,22,33,44,55)
```

2.1 Distribuzione di frequenza assoluta

2.1.1 Persone che praticano sport in modo continuativo

```
table(cut(sport_in_modo_continuativo,classi))
```

```
##   (0,11]   (11,22]   (22,33]   (33,44]   (44,55]
##        0          6         13          1          0
```



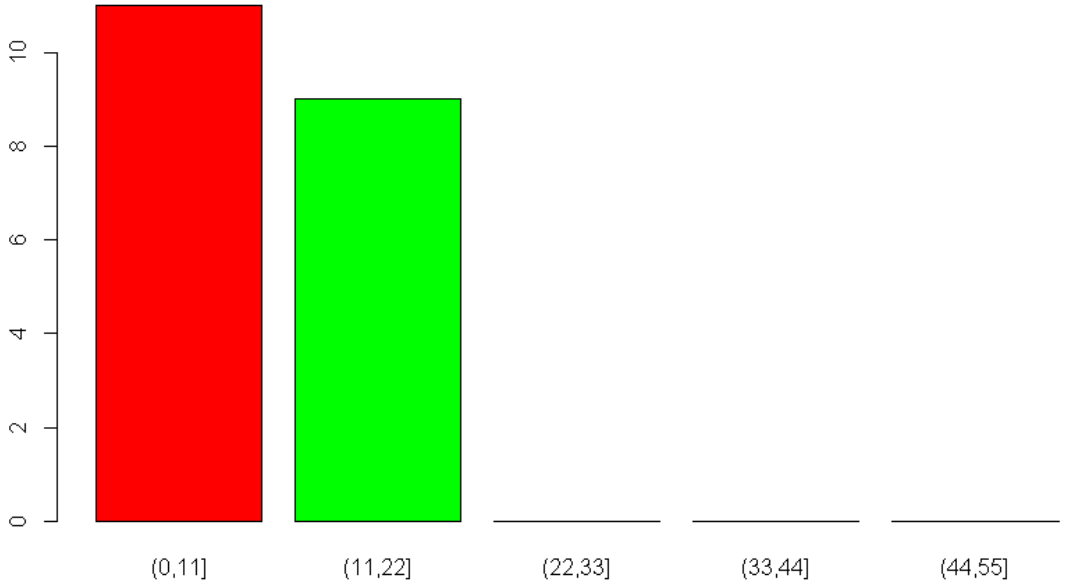
Da questo risultato si capisce che:

- Si hanno valori abbastanza diversificati in tutte le regioni riguardo le persone che praticano sport in modo continuativo
- Le regioni settentrionali hanno una percentuale maggiore rispetto a quelle del sud, in particolare il Trentino Alto Adige, Valle d'Aosta e Lombardia, che presentano dati molto al di sopra della media
- Anche altre regioni si discostano dalla media ma in modo meno evidente.

2.1.2 Persone che praticano sport in modo saltuario

```
table(cut(sport_in_modo_saltuario,classi))
```

```
##   (0,11]   (11,22]   (22,33]   (33,44]   (44,55]
##       11          9          0          0          0
```



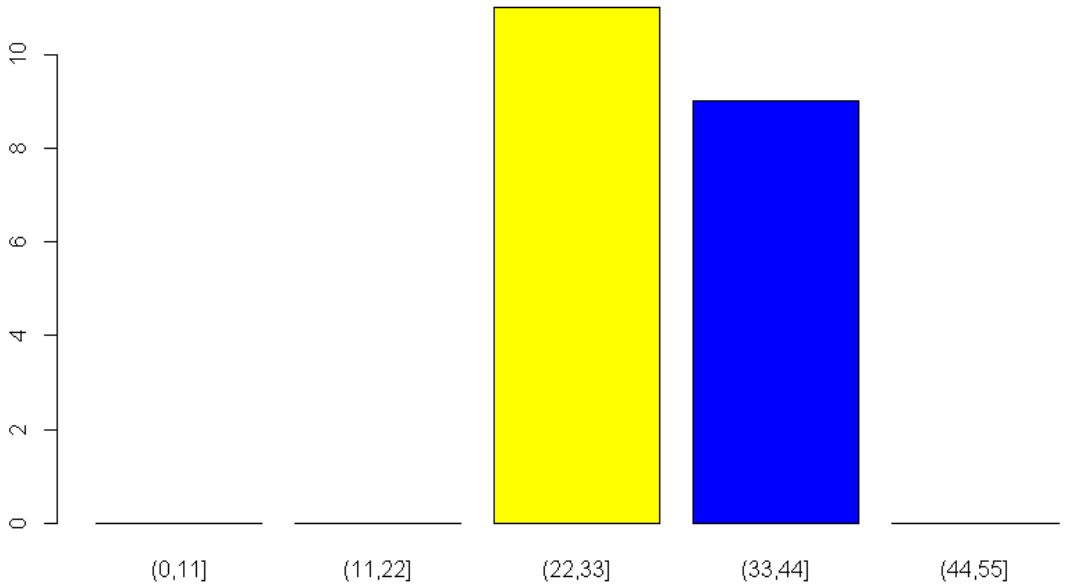
Da questo risultato si capisce che:

- La percentuale di persone che praticano sport in modo saltuario è omogenea in tutte le regioni italiane.
- Risultano essere poche le regioni con una percentuale minore della media, come la Campania, la Calabria e la Sicilia
- Le regioni settentrionali continuano ad avere una percentuale maggiore rispetto a quelle del sud

2.1.3 Persone che praticano solo qualche attività fisica

```
table(cut(sport_solo_qualche_attivita,classi))
```

```
##   (0,11]   (11,22]   (22,33]   (33,44]   (44,55]
##        0          0        11         9         0
```



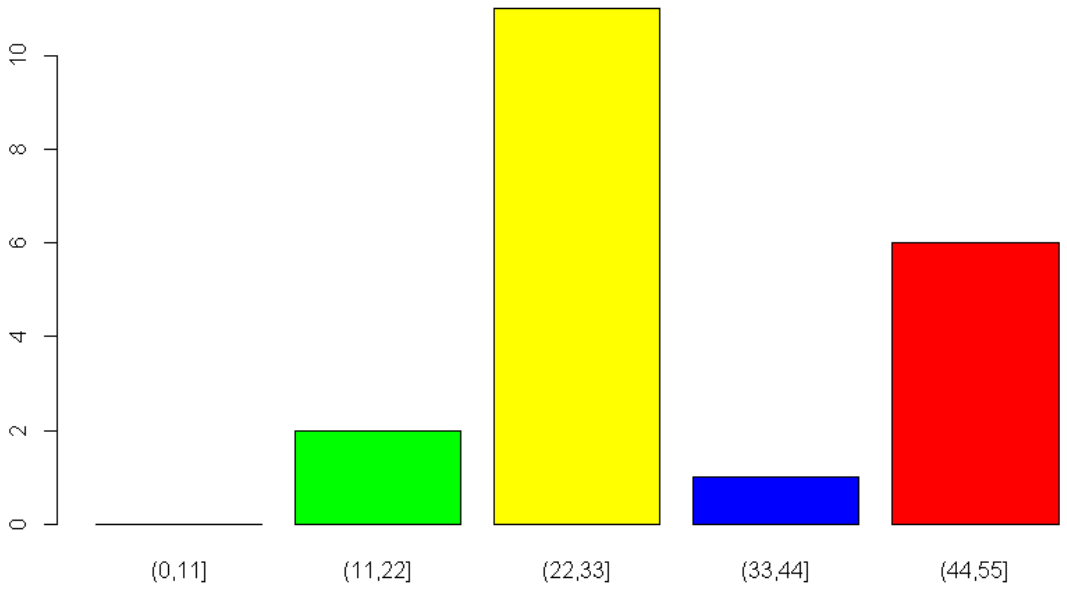
Da questo risultato si capisce che:

- La percentuale di persone che praticano solo qualche attività fisica è omogenea in tutte le regioni italiane.
- Risultano essere poche le regioni con una percentuale minore della media, come la Campania e la Sicilia

2.1.4 Persone che non praticano sport né attività fisica

```
table(cut(non_praticano_sport,classi))
```

```
##   (0,11]   (11,22]   (22,33]   (33,44]   (44,55]
##        0          2         11          1          6
```



Da questo risultato si capisce che:

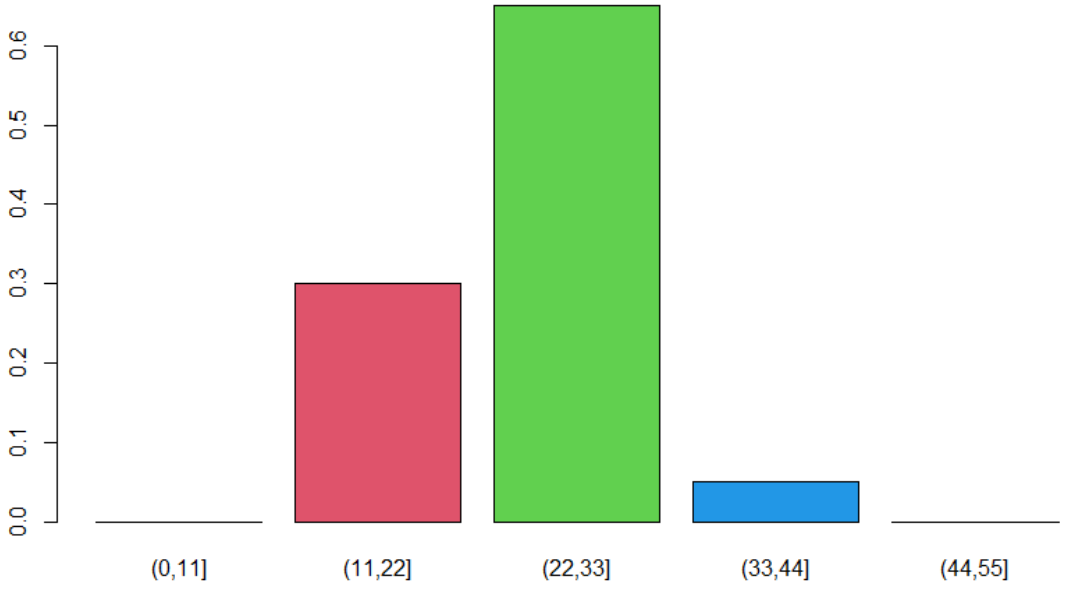
- Si hanno valori abbastanza diversificati in tutte le regioni riguardo le persone che non praticano sport né attività fisica
- Anche altre regioni si discostano dalla media in modo abbastanza evidente, ad esempio il Trentino Alto Adige, i cui individui che non praticano sport né attività fisica sono solo il 13,5%.
- Una regione su tutte, la Campania, ha una percentuale altissima pari al 52,8%

2.2 Distribuzione di frequenza relativa

2.2.1 Persone che praticano sport in modo continuativo

```
table(cut(sport_in_modo_continuativo, classi))/length(sport_in_modo_continuativo)
```

```
##  (0,11] (11,22] (22,33] (33,44] (44,55]
##    0.00     0.30     0.65     0.05     0.00
```



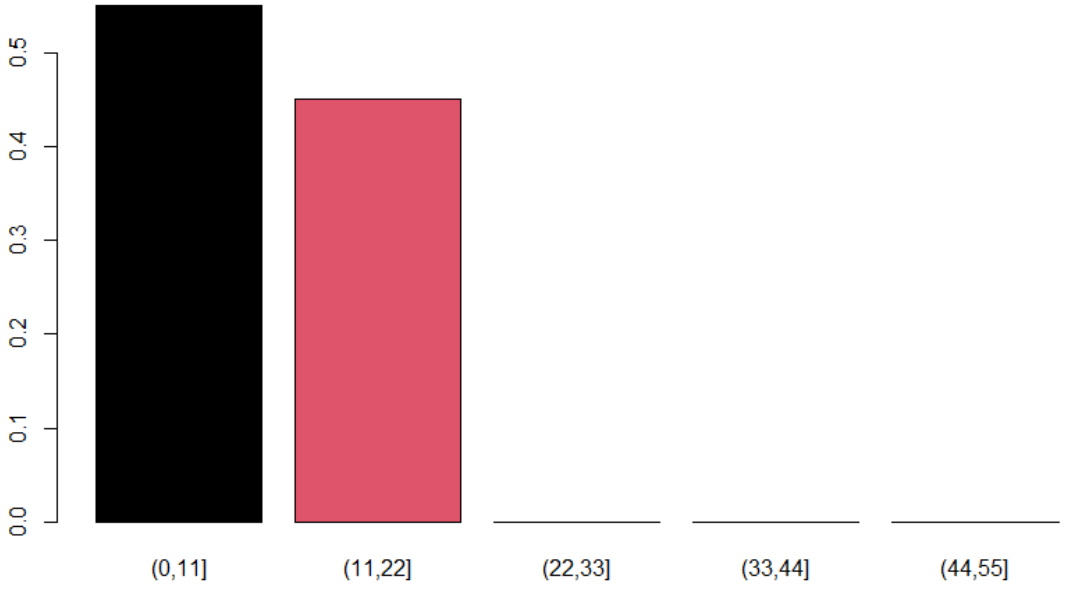
Da questo risultato si evince che:

- Il 30% delle regioni italiane presenta valori compresi tra l'11% e il 22%
- Una più corposa fetta di regioni ha valori che oscillano tra il 22% e il 33%
- Il 5% delle regioni che ha valori tra il 33% e il 44% è rappresentato dal Trentino Alto Adige, con una percentuale di persone che praticano sport in modo continuativo del 39,8%

2.2.2 Persone che praticano sport in modo saltuario

```
table(cut(sport_in_modo_saltuario, classi))/length(sport_in_modo_saltuario)
```

```
##  (0,11] (11,22] (22,33] (33,44] (44,55]
##    0.55     0.45     0.00     0.00     0.00
```



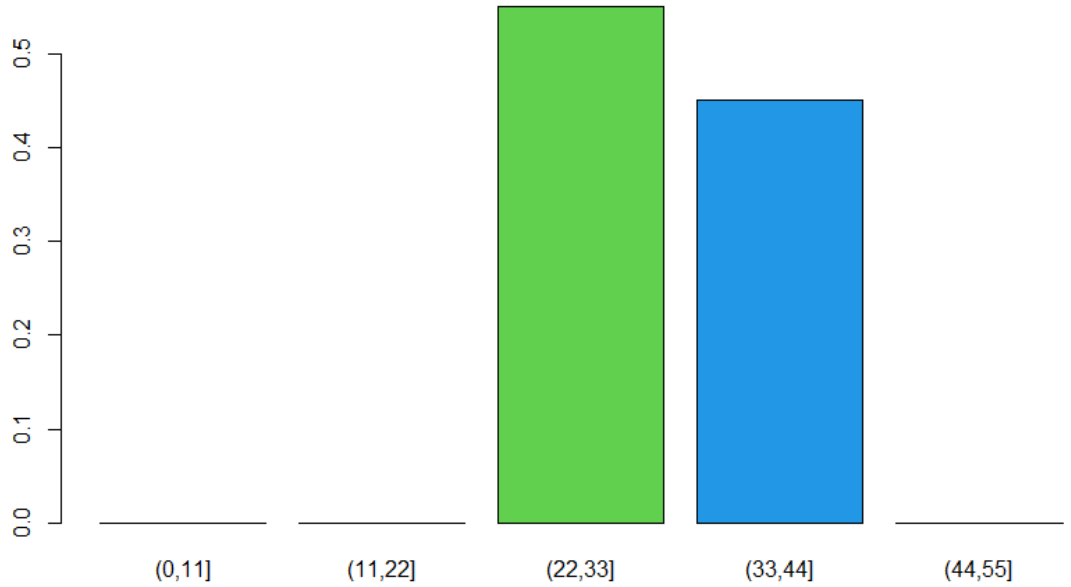
Da questo risultato si evince che:

- Le percentuali risultano essere omogenee con una percentuale del 55% delle regioni comprese tra 0% e 11%
- Il restante 45% rappresentato maggiormente, come detto in precedenza, da regioni settentrionali, presenta invece valori compresi tra il 22% e il 33%

2.2.3 Persone che praticano solo qualche attività fisica

```
table(cut(sport_solo_qualche_attivita, classi))/length(sport_solo_qualche_attivita)
```

```
##  (0,11] (11,22] (22,33] (33,44] (44,55]
##    0.00     0.00     0.55     0.45     0.00
```



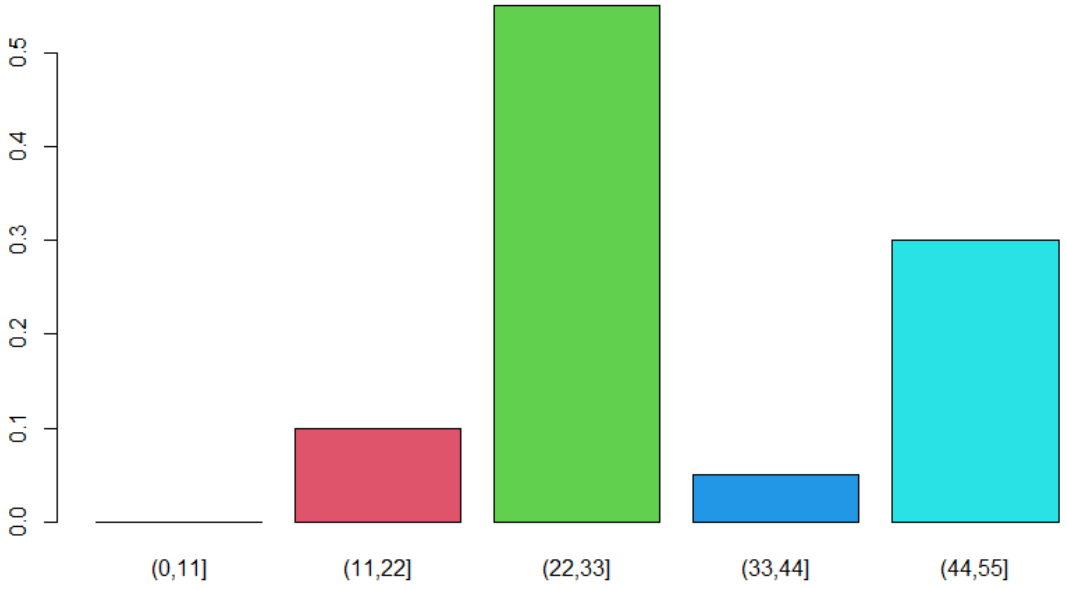
Da questo risultato si evince che:

- Anche qui le percentuali risultano essere omogenee, con una fetta di regioni del 55% che presenta percentuali comprese tra il 22% e il 33%
- Il restante 45% presenta invece percentuali comprese tra il 33% e il 44%

2.2.4 Persone che non praticano sport né attività fisica

```
table(cut(non_praticano_sport, classi))/length(non_praticano_sport)
```

```
##  (0,11] (11,22] (22,33] (33,44] (44,55]
##    0.00    0.10    0.55    0.05    0.30
```



Da questo risultato si evince che:

- I valori tra le regioni sono maggiormente diversificati rispetto alle precedenti analisi, in particolar modo ritroviamo Trentino Alto Adige e Valle d'Aosta che rientrano nel 10% delle regioni che hanno valori compresi tra l'11% e il 22%
- Il 55% delle regioni presenta invece valori che rientrano tra il 22% e il 33%
- Il 5% delle regioni che ha valori tra il 33% e il 44% è rappresentato dall'Abruzzo, con una percentuale di persone che non praticano sport né attività fisica del 33,7%
- Le regioni meridionali presentano una percentuale decisamente più elevata rispetto a quelle settentrionali, una buona fetta di esse presenta valori che rientrano tra il 44% e il 55%

3 - ANALISI TRAMITE RAPPRESENTAZIONI GRAFICHE

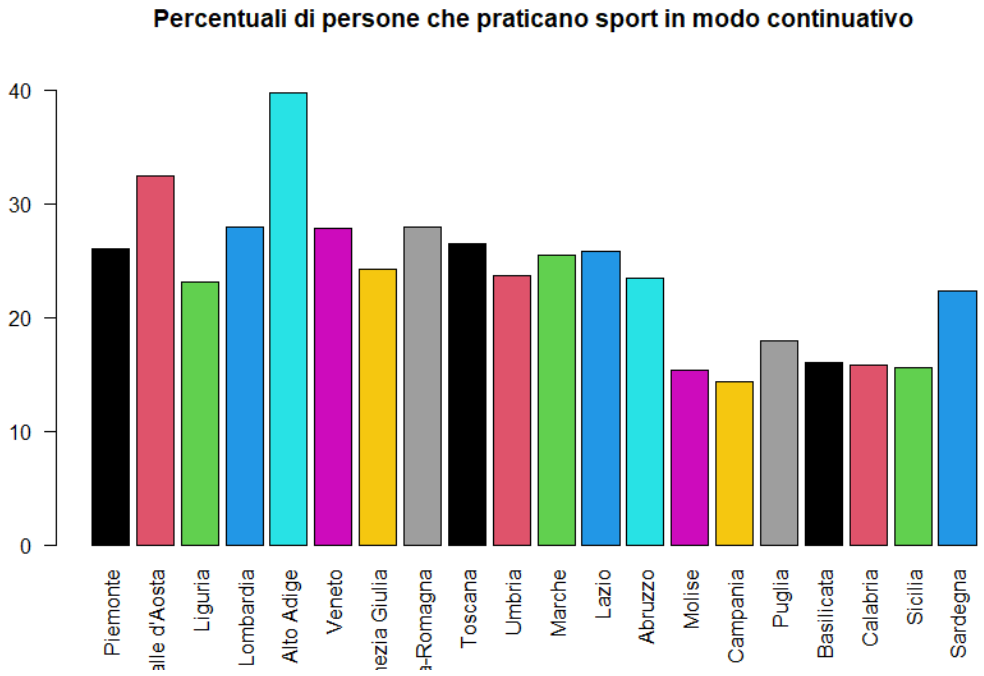
Le rappresentazioni grafiche permettono di avere una raffigurazione dei dati.
La statistica produce i propri output in forma numerica o in tabelle.
Ci sono tantissime tipologie di grafici attraverso cui esprimiamo il nostro lavoro di analisi.

3.1 Grafici a barre

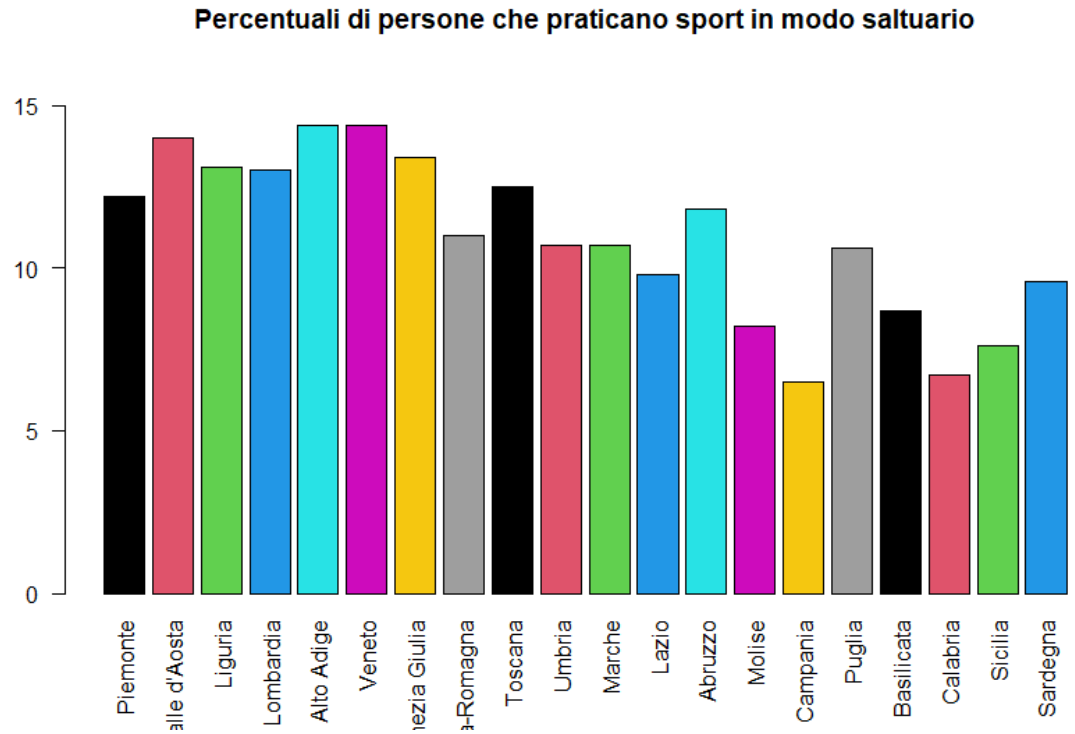
Un grafico a barre permette di riassumere una serie di dati relativi alle categorie che vengono analizzate.
Esso visualizza i dati usando più barre della stessa larghezza, ciascuna delle quali rappresenta una specifica categoria.
L'altezza di ciascuna barra è proporzionale al valore individuato per ogni categoria presa in considerazione.
In questo caso le categorie sono le regioni italiane e i valori riscontrati per ciascuna di esse sono le percentuali associate allo stile di vita degli Individui.
Per ciascuna delle colonne della matrice si va a realizzare un grafico a barre in cui sull'asse delle X vi sono le regioni, mentre sull'asse delle Y, le percentuali individuate per ogni tipologia di frequenza di attività fisica.
In R per realizzare un grafico a barre si utilizza la funzione barplot().

```
barplot(matrice_analisi_sport[,1], col=1:20, main = "Percentuali di persone che praticano sport in modo continuativo", las=2, ylim = c(0,42))
```

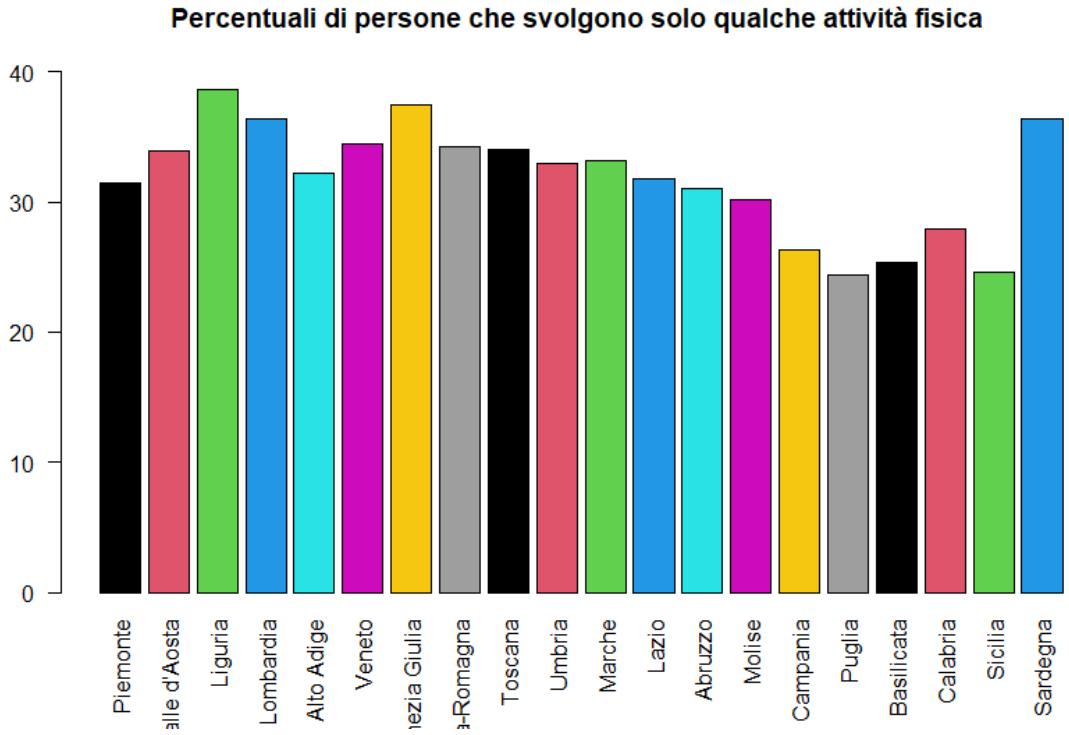
- Grafico riguardante le percentuali di persone che praticano sport in modo continuativo



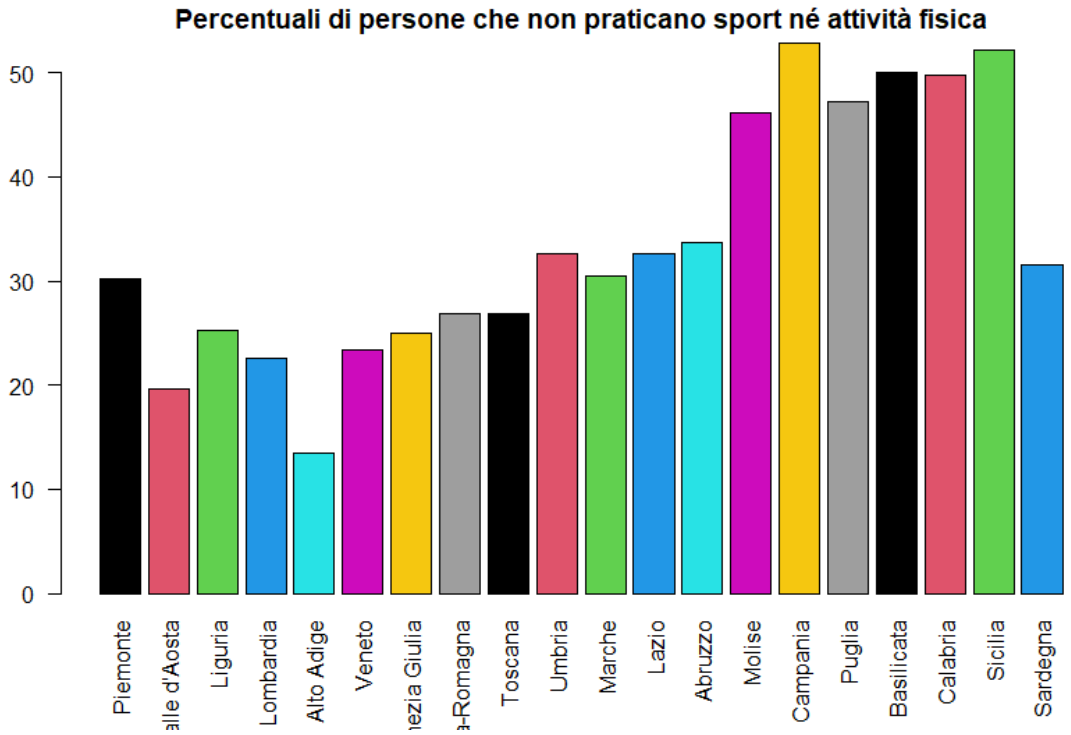
- Grafico riguardante le percentuali di persone che praticano sport in modo saltuario



- Grafico riguardante le percentuali di persone che svolgono solo qualche attività fisica



- Grafico riguardante le percentuali di persone che non praticano sport né attività fisica


Percentuali di persone che non praticano sport né attività fisica
50
40
30
20
10
0
Piemonte
alle d'Aosta
Liguria
Lombardia
Alto Adige
Veneto
ezia Giulia
a-Romagna
Toscana
Umbria
Marche
Lazio
Abruzzo
Molise
Campania
Puglia
Basilicata
Calabria
Sicilia
Sardegna

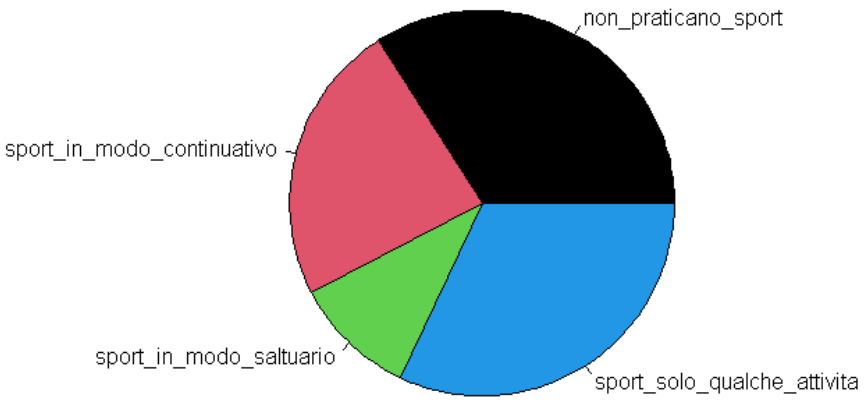
3.2 Grafico a torta

Un grafico a torta è un tipo di grafico utilizzato in statistica per rappresentazioni di variabili quantitative.
Viene creato un diagramma circolare e viene diviso in spicchi dove le ampiezze sono proporzionali alle classi di frequenza.
Si userà un grafico a torta per riassumere la situazione italiana relativa allo svolgimento e non di attività sportive introducendo il vettore corrispondente ricavato dal dataset iniziale:

```
italia <-c(
   rep("sport_in_modo_continuativo",23.56),
   rep("sport_in_modo_saltuario",10.95),
   rep("sport_solo_qualche_attivita",31.83),
   rep("non_praticano_sport",33.61)
)
```

In R per rappresentare un grafico a torta si usa il comando pie():

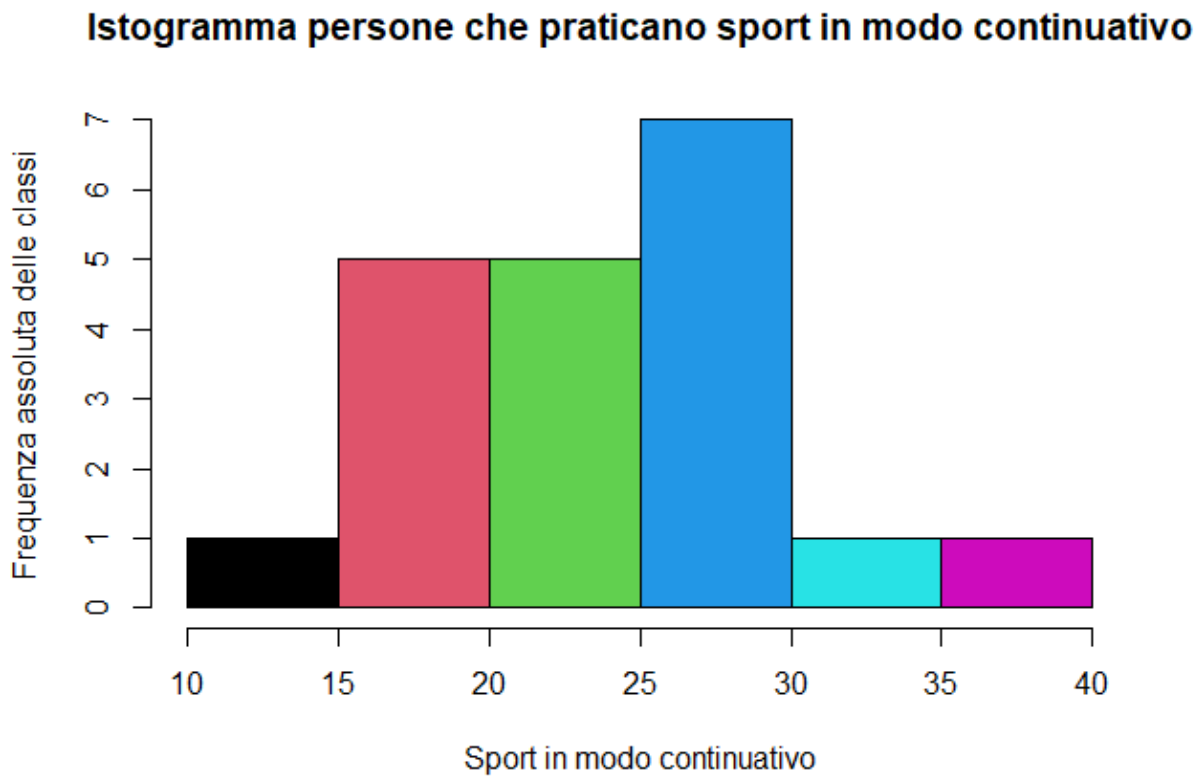
```
pie(table(italia),col=1:4)
```



3.3 Istogrammi

Un istogramma è una particolare rappresentazione grafica di una distribuzione di frequenza in classi.
Esso è ottenuto attraverso rettangoli adiacenti aventi per base dei segmenti i cui estremi corrispondono agli estremi delle classi.
Le altezze dei rettangoli sono tali che l'area risulta uguale alla frequenza della classe selezionata.
In R la funzione che realizza un istogramma è hist(). Il numero di classi può essere definito dall'utente mediante il parametro breaks, ma può anche essere lasciato a discrezione di R che deciderà il numero di classi che ritiene più adeguato.
La funzione hist genera oltre al grafico anche altre informazioni utili: i punti di suddivisione in classi, le frequenze assolute delle classi, la densità delle classi e i punti centrali delle classi.
Di seguito viene riportata l'analisi relativa all'istogramma delle percentuali di persone che praticano sport in modo continuativo.

```
continuativo <- hist(sport_in_modo_continuativo, freq = TRUE, main="Istogramma persone che praticano sport in modo continuativo", ylab = "Frequenza assoluta delle classi", xlab = "Sport in modo continuativo", col=1:6)
```



```
str(continuativo)
List of 6
 $ breaks  : int [1:7] 10 15 20 25 30 35 40
 $ counts  : int [1:6] 1 5 5 7 1 1
 $ density : num [1:6] 0.01 0.05 0.05 0.07 0.01 0.01
 $ mids    : num [1:6] 12.5 17.5 22.5 27.5 32.5 37.5
 $ xname   : chr "sport_in_modo_continuativo"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

```
continuativo$breaks
[1] 10 15 20 25 30 35 40
```

R ha effettuato la seguente suddivisione in classi: (10,15] (15,20] (20,25] (25,30] (30,35] (35,40].

```
continuativo$counts
[1] 1 5 5 7 1 1
```

Dei 20 valori considerati

- 1 è presente all'interno della prima classe
- 5 sono presenti all'interno della seconda classe
- 5 sono presenti all'interno della terza classe
- 7 sono presenti all'interno della quarta classe
- 1 è presente all'interno della quinta classe
- 1 è presente all'interno della sesta classe

```
continuativo$density
[1] 0.01 0.05 0.05 0.07 0.01 0.01
```

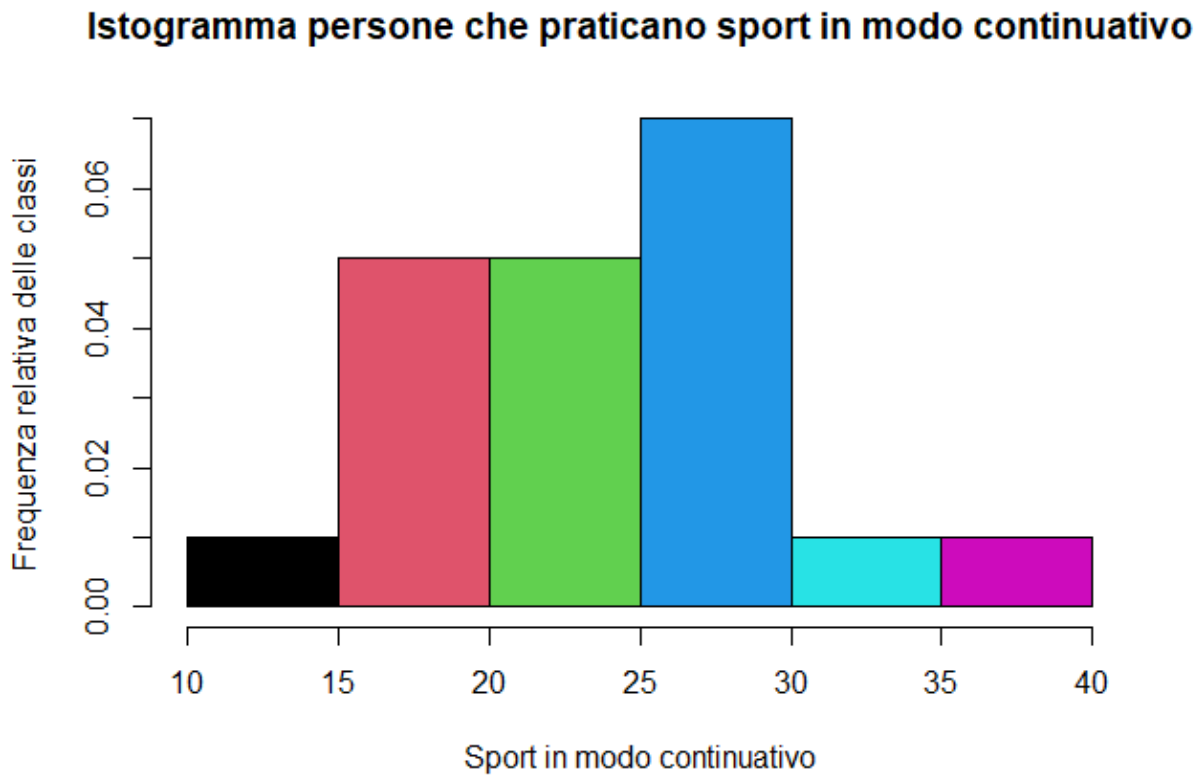
```
continuativo$mids
[1] 12.5 17.5 22.5 27.5 32.5 37.5
```

Le frequenze relative associate alle sei classi possono essere ottenute moltiplicando gli elementi del vettore continuativo$dentisy per l'ampiezza effettiva di ogni classe, ovvero 5

```
continuativo$density * 5
[1] 0.05 0.25 0.25 0.35 0.05 0.05
```

All'interno della funzione hist() impostando il parametro freq = FALSE si ottiene un istogramma in base alle frequenze relative

```
continuativo_rel <- hist(sport_in_modo_continuativo, freq = FALSE,
main="Istogramma persone che praticano sport in modo continuativo", ylab =
"Frequenza assoluta delle classi", xlab = "Sport in modo continuativo", col=1:6)
```

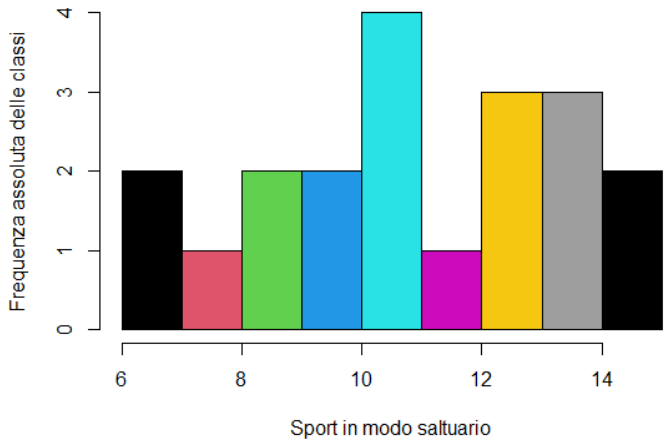


Anche in questo caso il numero di classi è scelto automaticamente da R e l'area totale è unitaria

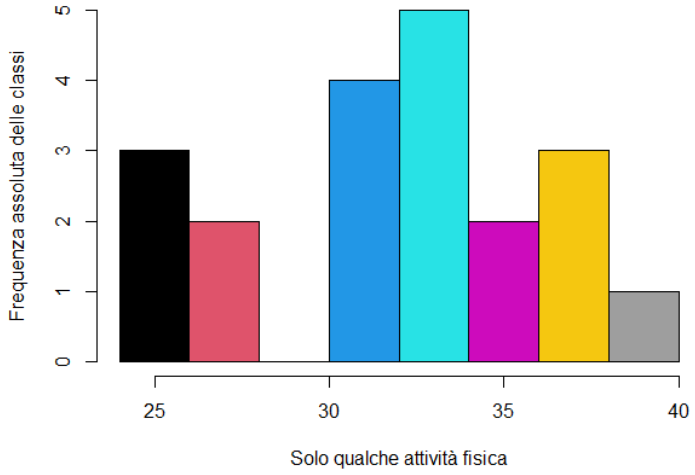
```
str(continuativo_rel)
List of 6
 $ breaks  : int [1:7] 10 15 20 25 30 35 40
 $ counts  : int [1:6] 1 5 5 7 1 1
 $ density : num [1:6] 0.01 0.05 0.05 0.07 0.01 0.01
 $ mids    : num [1:6] 12.5 17.5 22.5 27.5 32.5 37.5
 $ xname   : chr "sport_in_modo_continuativo"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

Di seguito sono riportati gli istogrammi relativi alle percentuali di persone che praticano sport in modo saltuario, che praticano solo qualche attività fisica e che non praticano sport né attività fisica

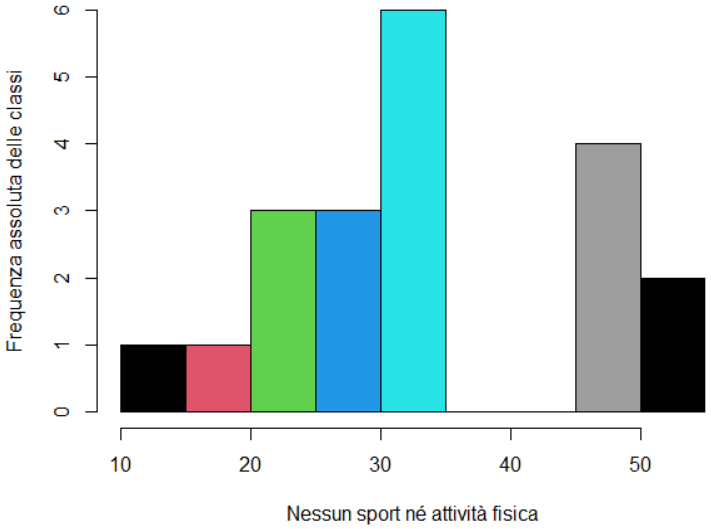
Istogramma persone che praticano sport in modo saltuario



Istogramma persone che praticano solo qualche attività fisica



Istogramma persone che non praticano sport né attività fisica



3.4 Boxplot

Il boxplot è un grafico statistico che viene utilizzato per variabili quantitative e per illustrare alcune caratteristiche di una distribuzione di frequenza quali la centralità, la forma, la dispersione e la presenza di eventuali valori anomali.
Dato un campione di valori, ordinati in modo crescente, assunti da una variabile quantitativa, si indica con Q1 il primo quartile, ovvero il valore per il quale il 25% dei dati sono sulla sinistra e il restante 75% dei dati sulla destra; con Q3 si indica il terzo quartile, ovvero il valore per il quale il 75% dei dati sono sulla sinistra e il restante 25% sulla destra; con Q2 si indica il secondo quartile ovvero il valore per il quale 50% dei dati sono sulla sinistra e 50% sulla destra, quest'ultimo è detto mediana.
Q0 e Q4 forniscono invece il minimo e il massimo dei valori assunti dal campione.
In R i quartili si calcolano grazie alla funzione quantile() e grazie alla funzione summary() è possibile determinare i valori di minimo, massimo, media, mediana, primo e terzo quartile.
Il boxplot, definito anche scatola con baffi, è il disegno di una scatola i cui estremi sono Q1 e Q3, tagliata dalla mediana in corrispondenza di Q2.
In basso e in alto sono presenti i baffi, ovvero due linee tratteggiate. L'estremo del baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale di:

$$Q1 - 1.5 \cdot (Q3 - Q1)$$

L'estremo del baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a:

$$Q3 + 1.5 \cdot (Q3 - Q1)$$

La distanza tra il primo e il terzo quartile è chiamata intervallo interquartile o scarto interquartile.
Se tutti i dati del campione rientrano nel seguente intervallo:

$$(Q1 - 1.5 \cdot (Q3 - Q1), Q3 + 1.5 \cdot (Q3 - Q1))$$

gli estremi dei baffi sono posti in corrispondenza del minimo valore e del massimo valore del campione. La presenza di eventuali valori al di fuori di questo intervallo sono visualizzati nel grafico sotto forma di valori anomali.
Individuarli è necessario per poter analizzare le caratteristiche e le eventuali cause che li hanno determinati.
All'interno del boxplot la centralità è espressa dalla mediana, la forma simmetrica o asimmetrica è individuata esaminando le distanze del primo e terzo quartile dalla

mediana. I baffi forniscono informazioni sulla dispersione e sulla forma della distribuzione ed anche sulle code della distribuzione.
La dispersione è deducibile esaminando le distanze dell'estremo del baffo superiore da Q3 e dell'estremo del baffo inferiore da Q1.
In R un boxplot si ottiene tramite la funzione boxplot().
Di seguito viene riportata l'analisi dei boxplot sui valori assunti dal dataset.

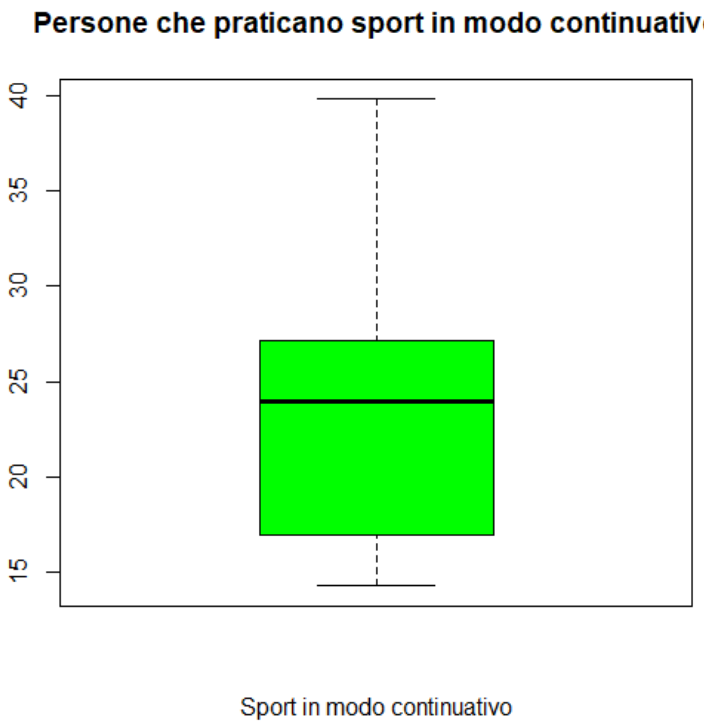
- **Persone che praticano sport in modo continuativo**

```
quantile(sport_in_modo_continuativo)

   0%    25%      50%      75%     100%
14.300 17.425  23.950 26.825   39.800
```

Si deduce che Q0 = 14.3, Q1 = 17.425, Q2 = 23.950, Q3 = 26.825, Q4 = 39.8

```
summary(sport_in_modo_continuativo)

  Min.  1st Qu.  Median   Mean  3rd Qu.  Max.
  14.30   17.43   23.95   23.59   26.82   39.80
```

```
boxplot(sport_in_modo_continuativo, xlab = "Sport in modo continuativo", main =
"Persone che praticano sport in modo continuativo", col = "green")
```



Grazie alla funzione boxplot.stats() è possibile raccogliere le statistiche necessarie per la creazione di un boxplot.

```
boxplot.stats(sport_in_modo_continuativo)
$stats
[1] 14.30 16.95 23.95 27.15 39.80

$n
[1] 20

$conf
[1] 20.34635 27.55365

$out
numeric(0)
```

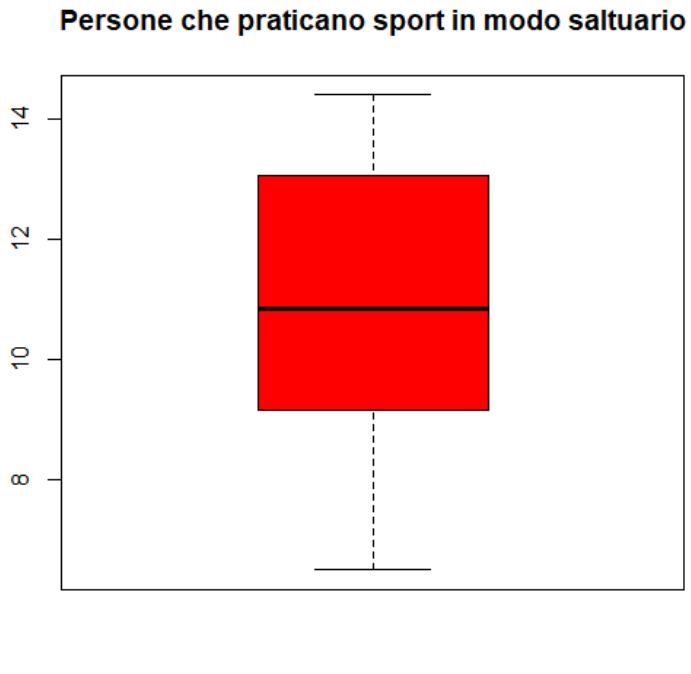
Da queste statistiche si comprende che il baffo inferiore corrisponde a 14.30 ovvero il valore minimo presente nel vettore, mentre il baffo superiore corrisponde a 39.80 ovvero il valore massimo presente nel vettore.
Non sono presenti valori anomali in quanto tutti i valori sono compresi tra i due baffi.
La distanza tra Q3 - Q2 e Q2 - Q1 non corrisponde, ed è possibile affermare che vi è un'asimmetria dei dati, visibile dal grafico in quanto la mediana è più vicina al terzo quartile.

- **Persone che praticano sport in modo saltuario**

```
quantile(sport_in_modo_saltuario)

   0%    25%    50%    75%   100%
 6.500  9.375 10.850 13.025 14.400
```

Si deduce che Q0 = 6.5, Q1 = 9.375, Q2 = 10.85, Q3 = 13.025, Q4 = 14.4

```
summary(sport_in_modo_saltuario)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.500   9.375  10.850  10.945  13.025  14.400
```




```
boxplot.stats(sport_in_modo_saltuario)
$stats
[1]  6.50  9.15 10.85 13.05 14.40

$n
[1] 20

$conf
[1]  9.472135 12.227865

$out
numeric(0)
```

Da queste statistiche si comprende che il baffo inferiore corrisponde a 6.50 ovvero il valore minimo presente nel vettore, mentre il baffo superiore corrisponde a 14.40 ovvero il valore massimo presente nel vettore.
Non sono presenti valori anomali in quanto tutti i valori sono compresi tra i due baffi. La distanza tra Q3 - Q2 e Q2 - Q1 non corrisponde, ed è possibile affermare che vi è un'asimmetria dei dati, visibile dal grafico in quanto la mediana è più vicina al primo quartile.

- **Persone che praticano solo qualche attività fisica**

```
quantile(sport_solo_qualche_attivita)

    0%     25%     50%     75%     100%
24.400 29.625 32.600 34.250  38.600
```

Si deduce che Q0 = 24.4, Q1 = 29.625, Q2 = 32.6, Q3 = 34.25, Q4 = 38.6

```
summary(sport_solo_qualche_attivita)

   Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
  24.40   29.62   32.60   31.84   34.25   38.60
```


```
boxplot.stats(sport_solo_qualche_attivita)
$stats
[1] 24.40 29.05 32.60 34.30 38.60

$n
[1] 20

$conf
[1] 30.74518 34.45482
```

```
$out
numeric(0)
```

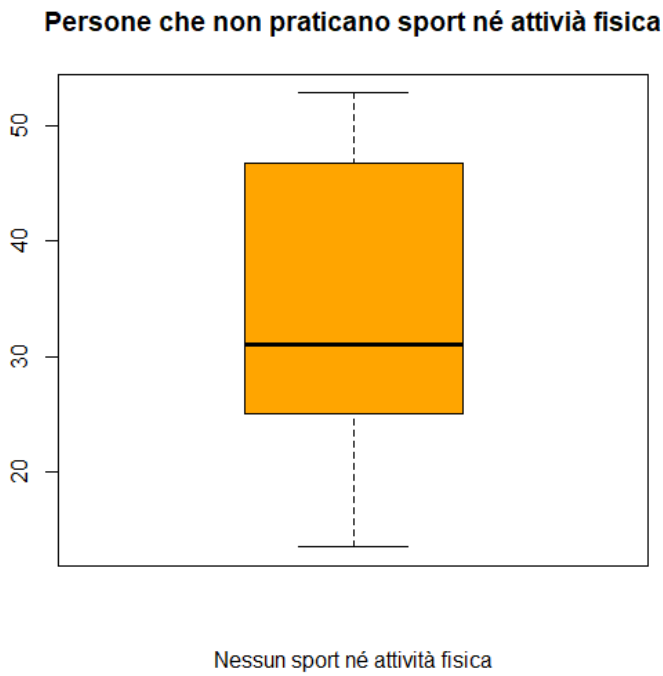
Da queste statistiche si comprende che il baffo inferiore corrisponde a 24.40 ovvero il valore minimo presente nel vettore, mentre il baffo superiore corrisponde a 38.60 ovvero il valore massimo presente nel vettore.
Non sono presenti valori anomali in quanto tutti i valori sono compresi tra i due baffi.
La distanza tra Q3 - Q2 e Q2 - Q1 non corrisponde, ed è possibile affermare che vi è un'asimmetria dei dati, visibile dal grafico in quanto la mediana è più vicina al terzo quartile.

- **Persone che non praticano sport né attività fisica**

```
quantile(non_praticano_sport)

   0%   25%   50%   75%  100%
13.50 25.15 31.05 46.45  52.80
```

Si deduce che Q0 = 13.50, Q1 = 25.15, Q2 = 31.05, Q3 = 46.45, Q4 = 52.80

```
summary(non_praticano_sport)

   Min. 1st Qu.  Median  Mean 3rd Qu.  Max.
  13.50   25.15   31.05   33.61   46.45   52.80
```

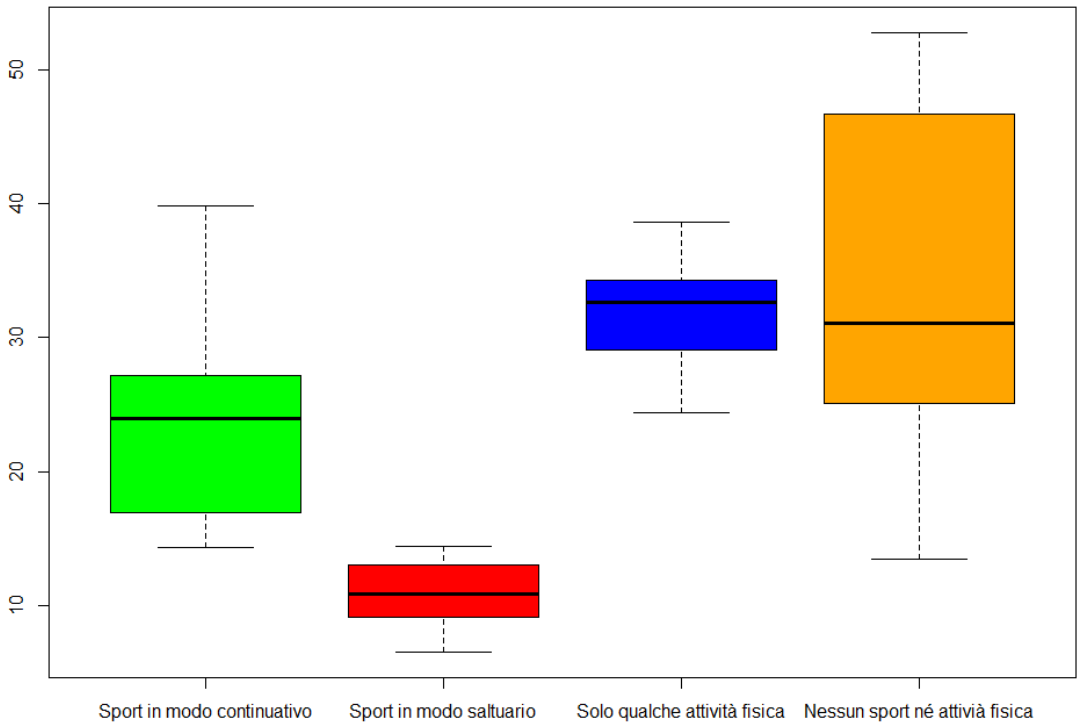


```
boxplot.stats(non_praticano_sport)
$stats
[1] 13.50 25.10 31.05 46.70 52.80

$n
[1] 20

$conf
[1] 23.41875 38.68125

$out
numeric(0)
```

Da queste statistiche si comprende che il baffo inferiore corrisponde a 13.50 ovvero il valore minimo presente nel vettore, mentre il baffo superiore corrisponde a 52.80 ovvero il valore massimo presente nel vettore.
Non sono presenti valori anomali in quanto tutti i valori sono compresi tra i due baffi.
La distanza tra Q3 - Q2 e Q2 - Q1 non corrisponde, ed è possibile affermare che vi è un'asimmetria dei dati, visibile dal grafico in quanto la mediana è più vicina al primo quartile.

Di seguito viene riportato il confronto dei boxplot visti in precedenza.
Grazie alla funzione boxplot() è possibile confrontare vari boxplot.

```
 boxplot(sport_in_modo_continuativo, sport_in_modo_saltuario,
sport_solo_qualche_attivita, non_praticano_sport, names = c("Sport in modo
continuativo", "Sport in modo saltuario", "Solo qualche attività fisica", "Nessun
sport né attività fisica"), col = c("green", "red", "blue", "orange"))
```



Da questa analisi si evince che vi è un’asimmetria tra i vettori del dataset e che non vi è presenza di valori anomali.

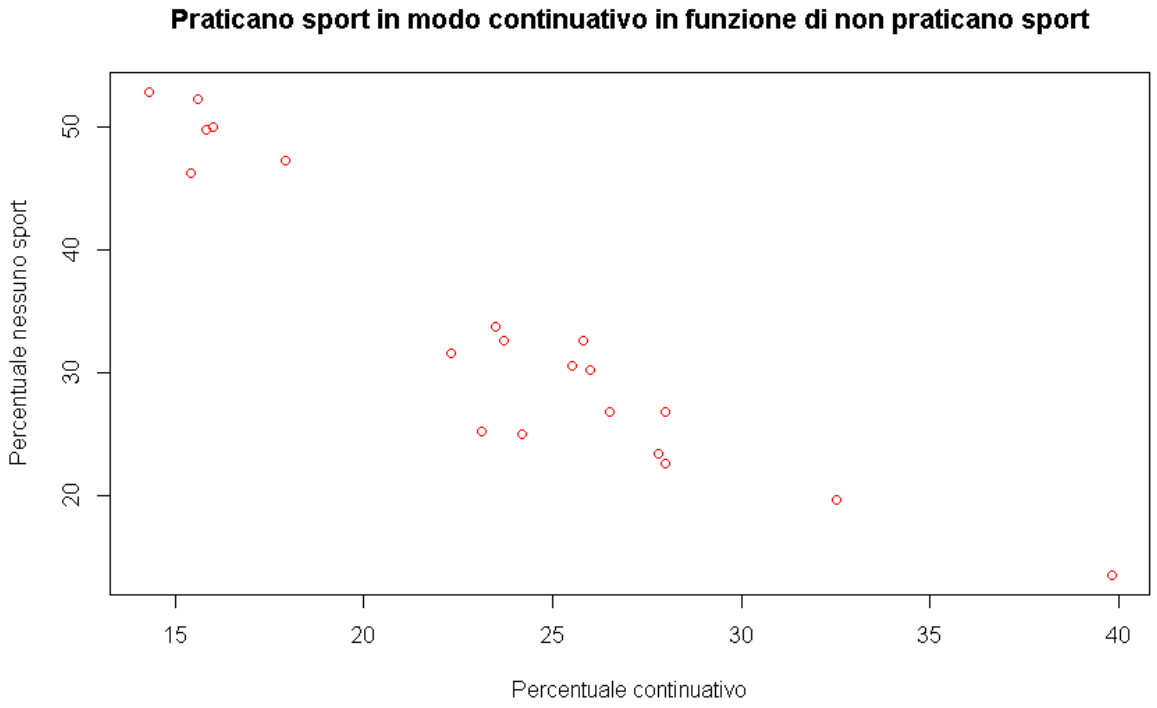
3.5 Grafico di dispersione

Un grafico di dispersione, anche chiamato scatterplot, è un tipo di diagramma che permette di visualizzare la relazione tra due variabili quantitative.
Le due variabili sono riportate su uno spazio cartesiano dove ogni unità è rappresentata da un punto posizionato sul grafico in base alle sue coordinate.
La variabile posta sull'asse delle ascisse viene definita variabile indipendente, mentre quella posta sull'asse delle ordinate viene chiamata variabile dipendente.

In R uno scatterplot di due vettori viene definito tramite il comando:

```
plot(vettore1, vettore2)
```

Di seguito viene riportato il grafico di dispersione prendendo due categorie del dataset:



In R è possibile visualizzare uno scatterplot per ogni coppia di variabili del dataset. Per farlo basta utilizzare la funzione pairs() e passare la matrice_analisi_sport.
(Vedi pag.51)

4 - STATISTICA DESCRITTIVA UNIVARIATA

La statistica descrittiva è costituita da un insieme di metodi di natura logica e matematica il cui obiettivo è quello di ricavare da un insieme di dati raccolti in tabelle e grafici alcune informazioni significative per il problema studiato.
Nella statistica descrittiva, si possono descrivere i dati qualitativamente o quantitativamente.

- Variabile qualitativa: Si riferisce a una qualità.
- Variabile quantitativa: Si riferisce a una misura quantitativa.

Pertanto, su queste variabili possono essere calcolati alcuni parametri, soprattutto sulle variabili quantitative.
La statistica descrittiva si differenzia dalla statistica inferenziale in quanto nel primo caso i risultati ottenuti si basano su valori effettivi mentre nel secondo su stime.
Vi sono due tipi di statistica descrittiva:

- Statistica descrittiva univariata: dove si analizza una sola variabile dell'intera popolazione con:
    - Indici di posizione centrali(media, mediana, moda)
    - Indici non centrali(quantili, quartili, decili, percentili)
    - Indici di dispersione(varianza, deviazione standard, coefficiente di variazione)
    - la forma della distribuzione attraverso gli indici di skewness e curtosi
- Statistica descrittiva bivariata: dove si analizza due variabili dell'intera popolazione con:
    - Regressione lineare semplice
    - Regressione lineare multipla
    - Regressione non lineare

Di seguito viene riportata l'analisi relativa alla statistica descrittiva univariata.

4.1 Funzione di Distribuzione Empirica

La funzione di distribuzione empirica è una funzione che viene usata per descrivere fenomeni quantitativi.

4.1.1 Funzione di distribuzione empirica discreta

Considerando una variabile quantitativa i quali valori distinti che può assumere sono z_1,z_2, …, z_k, assumiamo che siano ordinati in ordine crescente e considerando anche le frequenze relative e quelle cumulate:

$$F_i = f_1 + f_2 + \ldots + f_i = \frac{n_1 + n_2 + \ldots + n_i}{n} \qquad (i = 1, 2, \ldots, k)$$

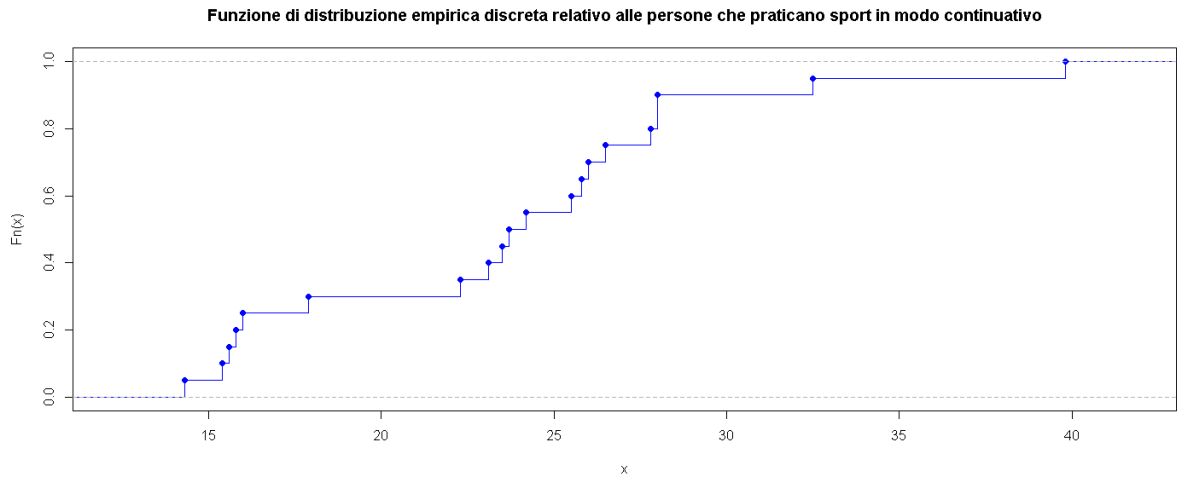
dove F_i rappresenta la proporzione dei dati del campione minori o uguali di z_i,
La funzione di distribuzione empirica discreta è descritta nel seguente modo:

$$F(x) = \frac{\#\{x_i \leq x, i = 1, 2, \ldots, n\}}{n} = \begin{cases} 0, & x < z_1 \\ F_1, & z_1 \leq x < z_2 \\ \ldots & \\ F_i, & z_i \leq x < z_{i+1} \\ \ldots & \\ 1, & x \geq z_k \end{cases}$$

La funzione di distribuzione empirica discreta è una funzione a scalini non decrescente dove ad ogni salto assume il valore a sinistra sull'asse delle ascisse.
La funzione vale:

- 0 per ogni valore minore dell'osservazione minima
- 1 per ogni valore maggiore dell'osservazione massima

In R per creare un grafico di una funzione di distribuzione empirica discreta si utilizza la funzione ecdf() (empirical cumulative distribution function)
Di seguito viene riportato un esempio di distribuzione empirica discreta relativo alle persone che praticano sport in modo continuativo:



4.1.2 Funzione di distribuzione empirica continua

La funzione di distribuzione empirica continua si utilizza con dati raccolti in classi. Anche questa viene definita a partire dalle frequenze relative cumulate e infatti la funzione di distribuzione empirica continua è così definita:

$$F(x) = \begin{cases} 0, & x < z_1 \\ \dots & \\ F_i, & x = z_i \\ \dfrac{F_{i+1} - F_i}{z_{i+1} - z_i}\, x + \dfrac{z_{i+1} F_i - z_i F_{i+1}}{z_{i+1} - z_i}, & z_i < x < z_{i+1} \\ F_{i+1}, & x = z_{i+1} \\ \dots & \\ 1, & x \geq z_{k+1} \end{cases}$$

La funzione di distribuzione empirica continua vale:

- 0 per ogni x minore di z_1
- 1 per ogni x maggiore o uguale di z_k+1
- Coincide con il segmento che passa per i punti (z_i,F_i) e (z_i+1,F_i+1) se x è compreso tra z_i e z_{i+1}

$$\frac{y - F_i}{x - z_i} = \frac{F_{i+1} - F_i}{z_{i+1} - z_i}$$

Di seguito viene riportato un esempio di funzione di distribuzione empirica continua relativa alle persone che praticano sport in modo continuativo, introducendo le seguenti classi:

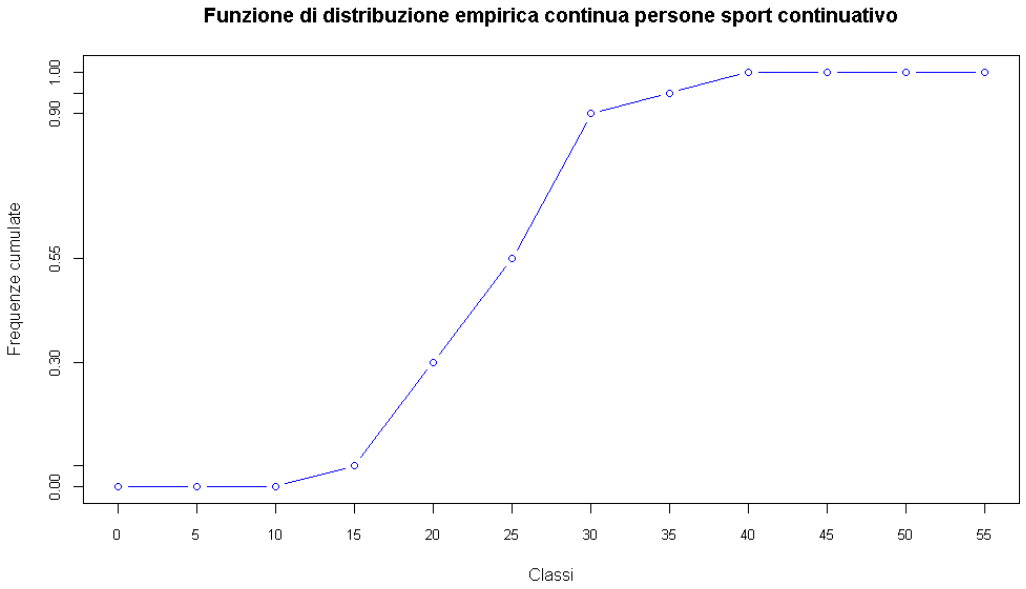
[0,5) [5,10) [10,15) [15,20) [20,25) [25,30) [30,35) [35,40) [40,45) [45,50) [50,55)

```
classEmpCont<-c(0,5,10,15,20,25,30,35,40,45,50,55)
```

Di seguito viene riportato come visualizzare il grafico della funzione di distribuzione empirica in R relativo alle persone che praticano sport in modo continuativo:

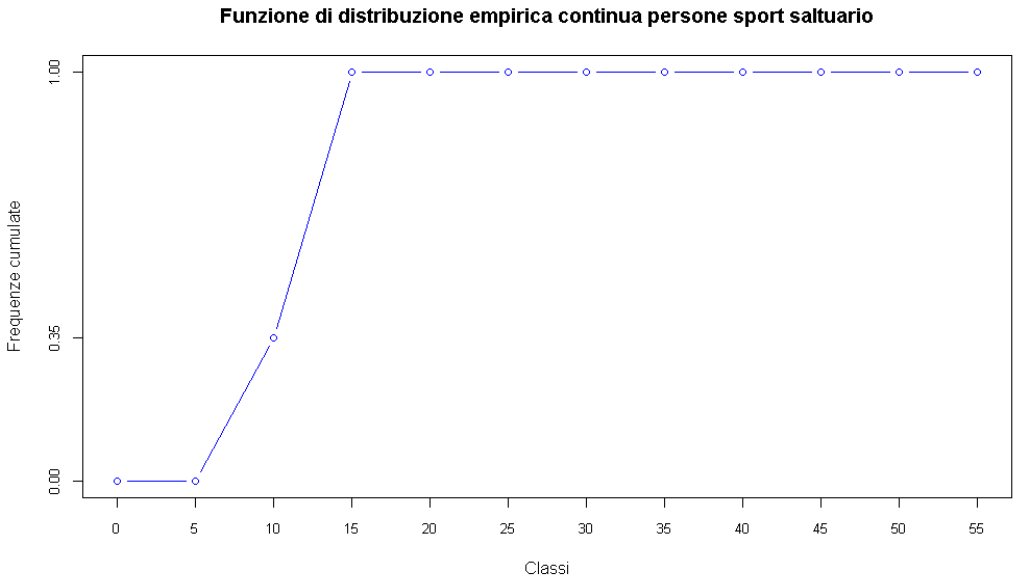
```
FcumSportContinuativo <- cumsum(table(cut(sport_in_modo_continuativo,
breaks=classEmpCont, right=FALSE)))/length(sport_in_modo_continuativo)

FcumSportContinuativo<-c(0,FcumSportContinuativo)
```

```
plot(classEmpCont, FcumSportContinuativo, type="b", axes=FALSE,
main="Funzione di distribuzione empirica continua persone sport continuativo",
col="blue",xlab="Classi",ylab="Frequenze cumulate")

axis(1, classEmpCont, cex.axis=0.8)

axis(2, FcumSportContinuativo, cex.axis=0.80)

box()
```

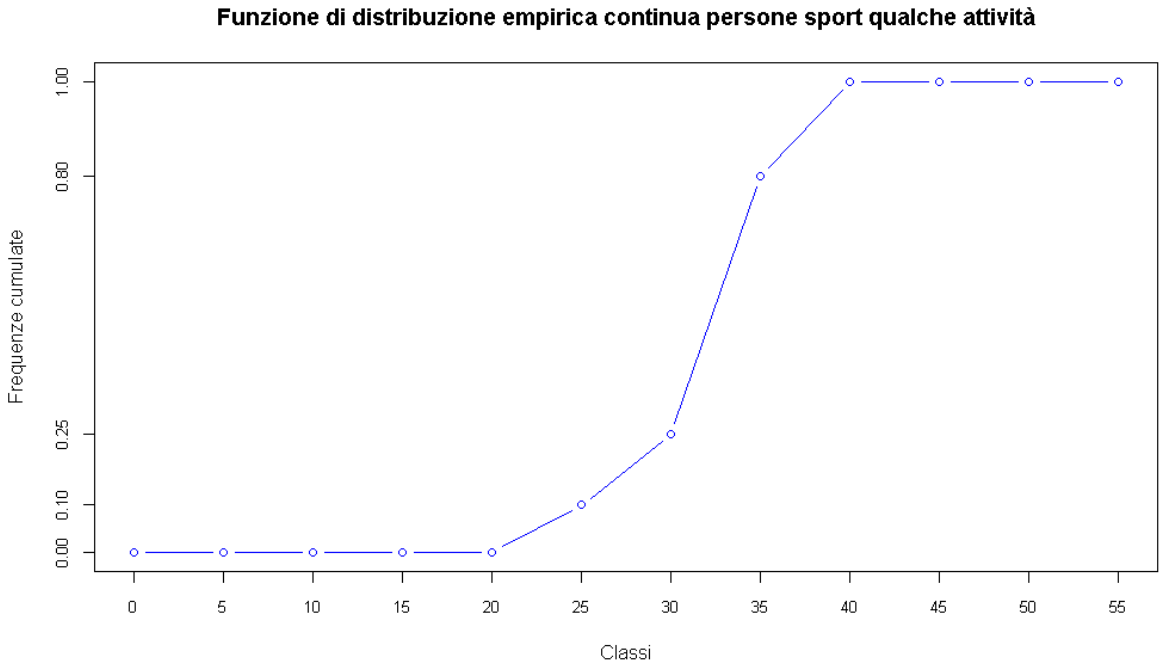


La maggior parte dei dati è concentrata nelle classi [20,25) e [25,30)

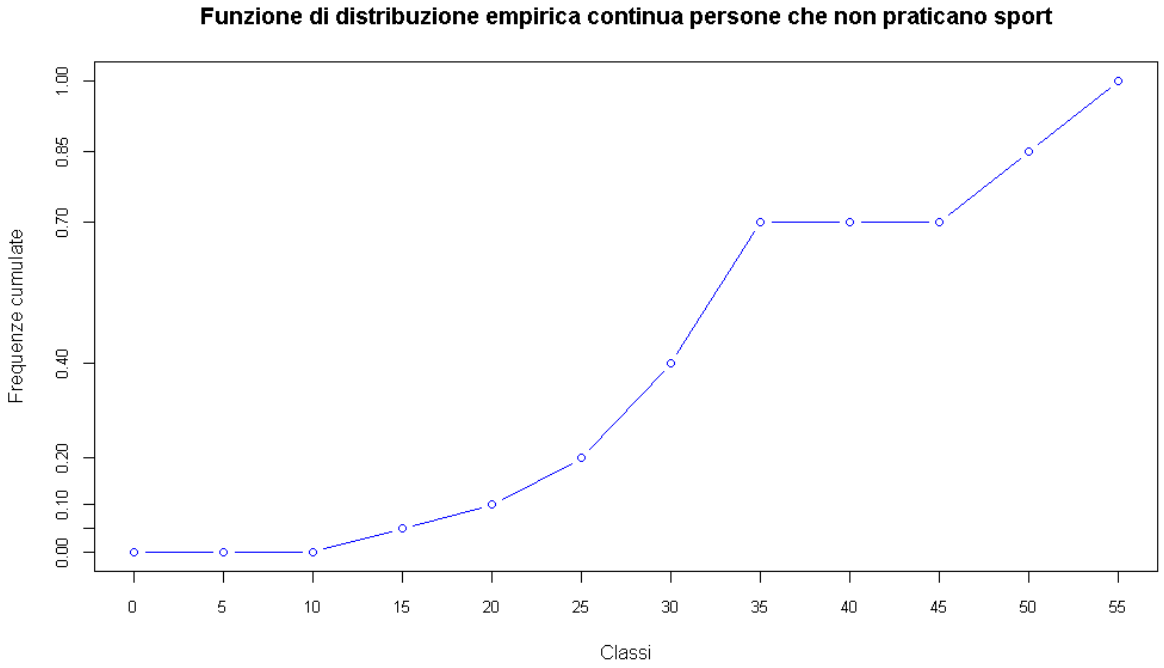
Di seguito verranno rappresentati i grafici per tutte le altre variabili.



Tutti i dati, com'è possibile notare, sono concentrati nelle prime classi



Qui invece è l'opposto, in quanto tutti i dati si concentrano nelle ultime classi, specialmente nella classe [30,35)



I dati sono molto centrati e ben distribuiti tra le classi.

4.2 Indici di Sintesi

Alcuni indici di sintesi sono utili a descrivere dei dati numerici.
Come detto in precedenza gli indici di posizione centrali sono:

- la media: si usa nel caso di variabili quantitative, è quel valore che corrisponde alla somma di tutti i valori diviso il numero dei valori stessi.
- la mediana: anche questa si usa nel caso di variabili quantitative, una sua caratteristica è quella di dividere l'insieme dei dati in due parti di uguale numerosità.
- la moda: si usa nel caso di variabili qualitative ed indica la modalità di un carattere che compare con la massima frequenza.

Gli indici di posizione non centrali dividono l'insieme dei dati ordinati in un fissato numero di parti uguali e sono:

- I quantili (percentili): si utilizzano per insiemi numerosi di dati.
- I quartili: sono un caso particolare dei percentili e si ottengono dividendo l'insieme dei dati in quattro parti uguali.
- I decili: anche questi sono un caso particolare dei percentili e si ottengono dividendo l'insieme dei dati in dieci parti uguali

Gli indici di dispersione sono misure utilizzate per confrontare la variazione tra due serie di dati misurati in unità diverse e sono:

- la varianza: che fornisce la misura di quanto i dati si discostino quadraticamente dalla media aritmetica
- la deviazione standard: è un modo per esprimere la dispersione dei dati intorno ad un indice di posizione, ha la stessa unità di misura dei valori osservati a differenza della varianza
- il coefficiente di variazione: permette di confrontare misure di fenomeni con unità di misura differenti, in quanto il risultato che si ottiene è un numero puro (cioè privo di alcuna unità di misura)

4.2.1 Media Campionaria

Per calcolare la media campionaria si sommano i dati e si dividono per il numero degli stessi.

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

La media campionaria ha diverse proprietà:

- Proprietà di linearità

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} (a\, x_i + b) = a\, \overline{x} + b.$$

- È una media pesata dei valori distinti assunti dai dati dove ogni valore distinto usa come peso la frequenza

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{k} z_i\, n_i = \sum_{i=1}^{k} \frac{n_i}{n}\, z_i = \sum_{i=1}^{k} f_i\, z_i,$$

- La somma algebrica degli scarti della media campionaria è sempre nulla

$$\sum_{i=1}^{n} s_i = \sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\, \overline{x} = \overline{x} - \overline{x} = 0$$

- La media è influenzata in particolar modo da valori particolarmente grandi o piccoli, cioè da valori anomali

In R per calcolare la media campionaria si utilizza la funzione mean().
Di seguito sono riportate le medie per ogni vettore del dataset:

```
mean(sport_in_modo_continuativo)
[1] 23.585
```

```
mean(sport_in_modo_saltuario)
[1] 10.945
```

```
mean(sport_solo_qualche_attivita)
[1] 31.835
```

```
mean(non_praticano_sport)
[1] 33.61
```

4.2.2 Mediana Campionaria

Si chiama mediana di una serie di n dati ordinati $x_1,x_2,...,x_n$ il valore centrale della serie, cioè quel valore che occupa il posto (n+1)/ 2 se n è dispari o la media dei valori che occupano i posti n/2 e n/2+1 se n+1 / 2 è pari.
La mediana dipende solo da uno o due valori e non risente dei valori estremi.
Per calcolare la mediana:

- Si ordinano gli n elementi in ordine crescente
- Se il numero di dati è dispari la mediana corrisponde al valore (n+1)/ 2
- Se il numero di dati è pari allora la mediana è calcolata come la media aritmetica dei valori alla posizione n/2 e alla posizione n/2+1

In R per calcolare la mediana campionaria si utilizza la funzione median().
Di seguito è riportata la mediana campionaria relativa ad ogni vettore del dataset:
Di seguito vengono visualizzate le mediane per ogni variabile del dataset.

```
median(sport_in_modo_continuativo)
[1] 23.95
```

```
median(sport_in_modo_saltuario)
[1] 10.85
```

```
median(sport_solo_qualche_attivita)
[1] 32.6
```

```
median(non_praticano_sport)
[1] 31.05
```

Dai seguenti risultati si ottiene:

	Sport in modo continuativo	Sport in modo saltuario	Sport qualche attività	Non praticano sport
Media	23.585	10.945	31.835	33.61
Mediana	23.95	10.85	32.6	31.05

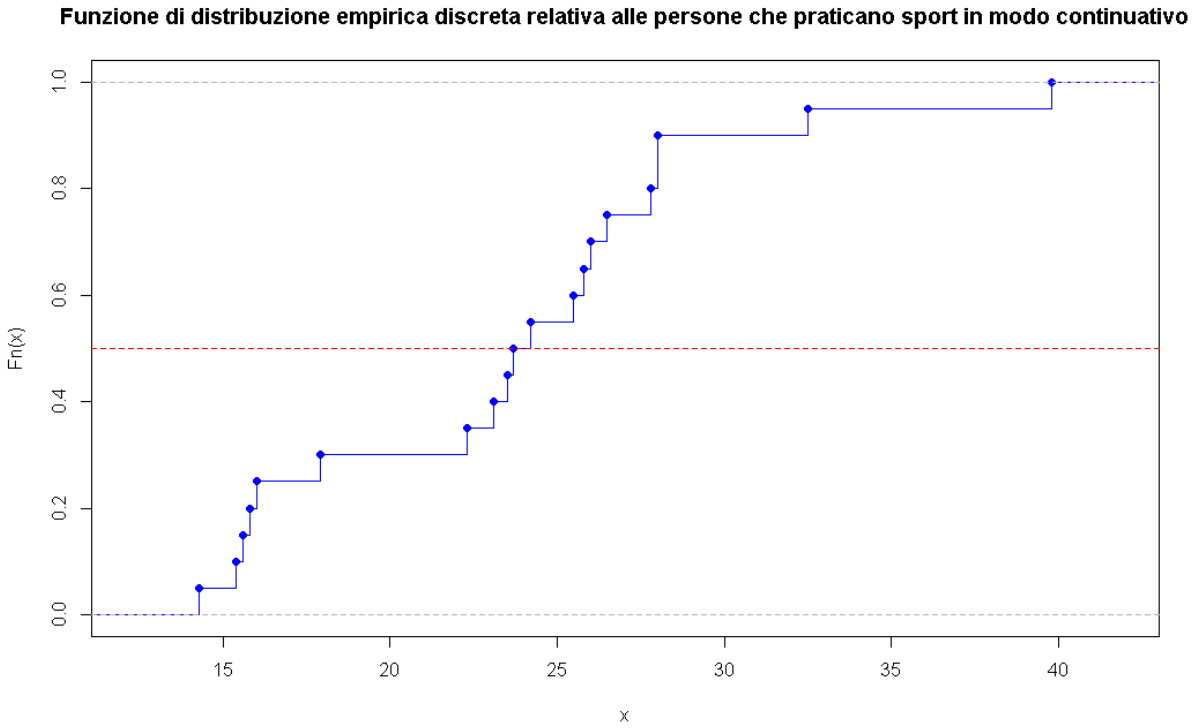
Dalla seguente tabella si nota che la media e la mediana hanno prodotto risultati pressoché simili per ciascuna categoria, la differenza maggiore è per l'ultimo vettore.

4.2.2.1 Mediana per una distribuzione di frequenza

Si può definire la mediana anche attraverso le frequenze cumulative.
La mediana per una distribuzione di frequenze è definita come la modalità i-esima che soddisfa la doppia disuguaglianza:

$$F_{i-1} < 0.5, \qquad F_i \geq 0.5$$

La mediana di una distribuzione di frequenza può essere realizzata graficamente utilizzando una funzione di distribuzione empirica discreta, ad esempio:



4.2.3 Moda Campionaria

La moda campionaria è una statistica usata per descrivere la centralità di una distribuzione di dati.
Rappresenta il valore prevalente in un insieme di dati, ossia il valore che si presenta con maggiore frequenza, assoluta o relativa.
La moda campionaria può non esistere e non essere unica. Quando è unica la distribuzione viene detta unimodale, quando ci sono due o più mode diverse viene detta bimodale o multimodale.
In una distribuzione di frequenze non esiste la moda campionaria se nessuna modalità ha una frequenza superiore alle altre.
La moda campionaria è anche usata quando si trattano dati di tipo qualitativo, per i quali non è possibile calcolare media e mediana.
Essendo presenti nel dataset percentuali, non risulta utile effettuare il calcolo della moda campionaria.

4.2.4 Quantili

I quantili dividono l'insieme dei dati ordinati in un fissato numero di parti uguali.
Come detto in precedenza, si suddividono in percentili, decili e quartili.
In R esistono nove differenti algoritmi per calcolare i quantili e vengono usati nella funzione quantile(), specificando nel parametro type l'algoritmo scelto.
R utilizza di default l'algoritmo di tipo 7, basato su tecniche di interpolazione tra i punti. Se il numero di osservazioni è elevato i valori dei quantili tendono a coincidere qualsiasi sia l'algoritmo scelto.
I percentili maggiormente usati sono il 25-esimo, ovvero il primo quartile Q_1, il 50-esimo, ovvero il secondo quartile Q_2, il 75-esimo, ovvero il terzo quartile Q_3.
Di seguito viene riportato il calcolo dei quartili, in base all'algoritmo scelto, per il vettore relativo alle persone che praticano sport in modo continuativo.
Omettendo all'interno della funzione quantile() i parametri probs e type, la funzione restituisce il minimo, massimo, i tre quartili e usa di default l'algoritmo di tipo 7.

```
quantile(sport_in_modo_continuativo)

   0%    25%     50%     75%    100%
14.300 17.425 23.950 26.825  39.800
```

```
quantile(sport_in_modo_continuativo, type = 2)

   0%   25%    50%   75%  100%
14.30 16.95  23.95  27.15 39.80
```

L’algoritmo di tipo 1 definisce i quantili considerando le frequenze relative cumulative:

```
quantile(sport_in_modo_continuativo, type = 1)

  0%  25%  50%  75% 100%
14.3 16.0   23.7   26.5   39.8
```

4.2.5 Varianza e Deviazione Standard Campionaria

Gli indici di posizione contengono informazioni che non tengono conto della variabilità dei dati, possono infatti esistere distribuzioni di frequenza che sono diverse tra loro pur avendo la stessa media campionaria.
Varianza campionaria e deviazione standard campionaria, detta anche scarto quadratico medio campionario, sono indici di dispersione, significativi per misurare la variabilità di una distribuzione di frequenza.
La varianza campionaria è così definita:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad (n = 2, 3, \ldots),$$

La deviazione standard campionaria è invece la radice quadrata della varianza campionaria:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad (n = 2, 3, \ldots).$$

La varianza e la deviazione standard sono tanto più grandi quanto più i dati si discostano dalla media.
I valori della varianza campionaria e della deviazione standard campionaria dipendono dall’unità di misura dei dati.
La varianza possiede le seguenti proprietà:

- Non gode della proprietà di linearità
- Sommare una costante a ciascuno dei dati non fa cambiare la varianza campionaria
- Moltiplicare ciascuno dei dati per un fattore costante fa sì che la varianza campionaria dell’insieme iniziale dei dati risulta moltiplicata per il quadrato di tale fattore.

In R la varianza campionaria si calcola tramite la funzione var() mentre la deviazione standard campionaria si calcola tramite la funzione sd().

Di seguito viene riportato il calcolo della varianza campionaria e della deviazione standard per i vettori presenti nel dataset.

```
var(sport_in_modo_continuativo)
[1] 41.67187

sd(sport_in_modo_continuativo)
[1] 6.455375
```

Per il vettore relativo alle persone che praticano sport in modo continuativo la varianza è più elevata in quanto i dati si discostano maggiormente dalla media.

```
var(sport_in_modo_saltuario)
[1] 6.196289

sd(sport_in_modo_saltuario)
[1] 2.489235
```

```
var(sport_solo_qualche_attivita)
[1] 18.04345

sd(sport_solo_qualche_attivita)
[1] 4.247758
```

```
var(non_praticano_sport)
[1] 140.8273

sd(non_praticano_sport)
[1] 11.86707
```

Per il vettore relativo alle persone che non praticano sport né attività fisica è ancora più evidente che la varianza è più elevata in quanto i dati si discostano maggiormente dalla media.

4.2.6 Coefficiente di Variazione

Il coefficiente di variazione è utile per confrontare le variazioni esistenti tra diversi campioni di dati, ed è definito come il rapporto tra la deviazione standard campionaria e il modulo della media campionaria:

$$CV = \frac{s}{|\overline{x}|}.$$

Il coefficiente di variazione è un numero puro, ed è un indice di dispersione che ha senso soltanto per campioni aventi la media campionaria non nulla.
Il coefficiente di variazione è sempre maggiore di 0, ma se compreso tra 0 e 1 allora la media è più grande della deviazione standard. Più è grande più c'è dispersione nei dati e quindi il valore medio non è molto significativo.
In R non è presente una funzione per calcolare il coefficiente di variazione, ma può essere implementata nel seguente modo:

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

Di seguito vengono riportati i valori del coefficiente di variazione dei vettori presenti nel dataset

```
cv(sport_in_modo_continuativo)
[1] 0.2737068

cv(sport_in_modo_saltuario)
[1] 0.2274312

cv(sport_solo_qualche_attivita)
[1] 0.1334304

cv(non_praticano_sport)
[1] 0.3530814
```

Di seguito vengono riportati i dati relativi alla media, varianza, deviazione standard e coefficiente di variazione dei vari vettori:

	sport_in_modo_continuativo	sport_in_modo_saltuario	sport_solo_qualche_attività	non_praticano_sport
Media	23.59	10.945	31.84	33.61
Varianza	41.67187	6.196289	18.04345	140.8273
Deviazione Standard	6.455375	2.489235	4.247758	11.86707
Coefficiente di Variazione	0.2737068	0.2274312	0.1334304	0.3530814

Da questa tabella è possibile notare come il coefficiente di variazione più alto si ottiene per il vettore relativo alle persone che non praticano sport né attività fisica, all'interno del quale i dati si discostano maggiormente dalla media campionaria.

4.3 Forma della Distribuzione di Frequenza

Media, mediana e moda sono indici utili per comprendere la forma delle distribuzioni di frequenze, differenze sostanziali tra questi indici indicano uno sbilanciamento eccessivo della distribuzione di frequenze verso destra o sinistra.
Esistono indici che permettono di misurare quando una distribuzione di frequenze presenta simmetria o asimmetria, oppure se essa è più o meno piccata.
Nel dettaglio questi indici sono:

- Skewness
- Curtosi

4.3.1 Skewness

La skewness è definita come il valore:

$$\gamma_1 = \frac{m_3}{m_2^{3/2}}$$

dove m_3 è il momento centrato campionario di ordine 3, definito come:

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^j \qquad (j = 1, 2, \ldots).$$

La skewness è un indice adimensionale.
Se la skewness assume valore:

- = 0 allora la distribuzione di frequenze è simmetrica
- > 0 allora la distribuzione di frequenze ha la coda di destra più allungata per l'asimmetria positiva
- < 0 allora la distribuzione di frequenze ha la coda di sinistra più allungata per l'asimmetria negativa

Ogni distribuzione di frequenze simmetrica ha una skewness nulla, ma esistono distribuzioni di frequenze non simmetriche con skewness nulla, inoltre se la skewness è uguale a 0 non significa che mediana, media e moda coincidono.
In R la skewness campionaria può essere implementata nel seguente modo:

```
skw <- function (x){n <-length (x)
        m2 <-(n -1) *var (x)/n
        m3 <- (sum ( (x-mean(x))^3) )/n
        m3/(m2 ^1.5)}
```

Di seguito vengono riportate le skewness dei vari vettori:

```
 skw(sport_in_modo_continuativo)
[1] 0.4664549

> skw(sport_in_modo_saltuario)
[1] -0.3042624

> skw(sport_solo_qualche_attivita)
[1] -0.4078287

> skw(non_praticano_sport)
[1] 0.366378
```

L'indice non è mai pari a 0, non è presente simmetria.
Il vettore relativo alle persone che praticano sport in modo saltuario e quello relativo alle persone che praticano solo qualche attività fisica presentano un'asimmetria negativa e quindi una distribuzione di frequenza con la coda di sinistra più allungata.
Il vettore relativo alle persone che praticano sport in modo continuativo e quello relativo alle persone che non praticano sport né attività fisica presentano invece un'asimmetria positiva e quindi una distribuzione di frequenza con la coda di destra più allungata.

4.3.2 Curtosi

La curtosi è un indice che permette di misurare la densità dei dati intorno alla media, si definisce come il valore:

$$\gamma_2 = \beta_2 - 3,$$

dove

$$\beta_2 = \frac{m_4}{m_2^2},$$

è l'indice di Pearson, che è un indice adimensionale.
I seguenti indici permettono di confrontare la distribuzione di frequenze dei dati con una densità di probabilità normale standard, caratterizzata da $\beta_2 = 3$ e e indice di curtosi $\gamma_2 = 0$.
Se risulta:

- $\beta_2 < 3$, $\gamma_2 < 0$: la distribuzione di frequenze è definita platicurtica, ossia la distribuzione di frequenze è più piatta di una normale.
- $\beta_2 > 3$, $\gamma_2 > 0$: la distribuzione di frequenze è definita leptocurtica, ossia la distribuzione di frequenze è più piccata di una normale.
- $\beta_2 = 3$, $\gamma_2 = 0$: la distribuzione di frequenze è definita normocurtica, ossia piatta come una normale.

In R il calcolo della curtosi può essere implementata nel seguente modo:

```
curt <-function (x){
      n<-length (x)
      m2 <-(n-1)*var (x)/n
      m4 <- (sum( (x-mean(x))^4) )/n
      m4/(m2^2) -3}
```

Di seguito vengono riportate le curtosi dei vari vettori:

```
curt(sport_in_modo_continuativo)
[1] 0.2121162

curt(sport_in_modo_saltuario)
[1] -1.002682

curt(sport_solo_qualche_attivita)
[1] -0.8136908
```

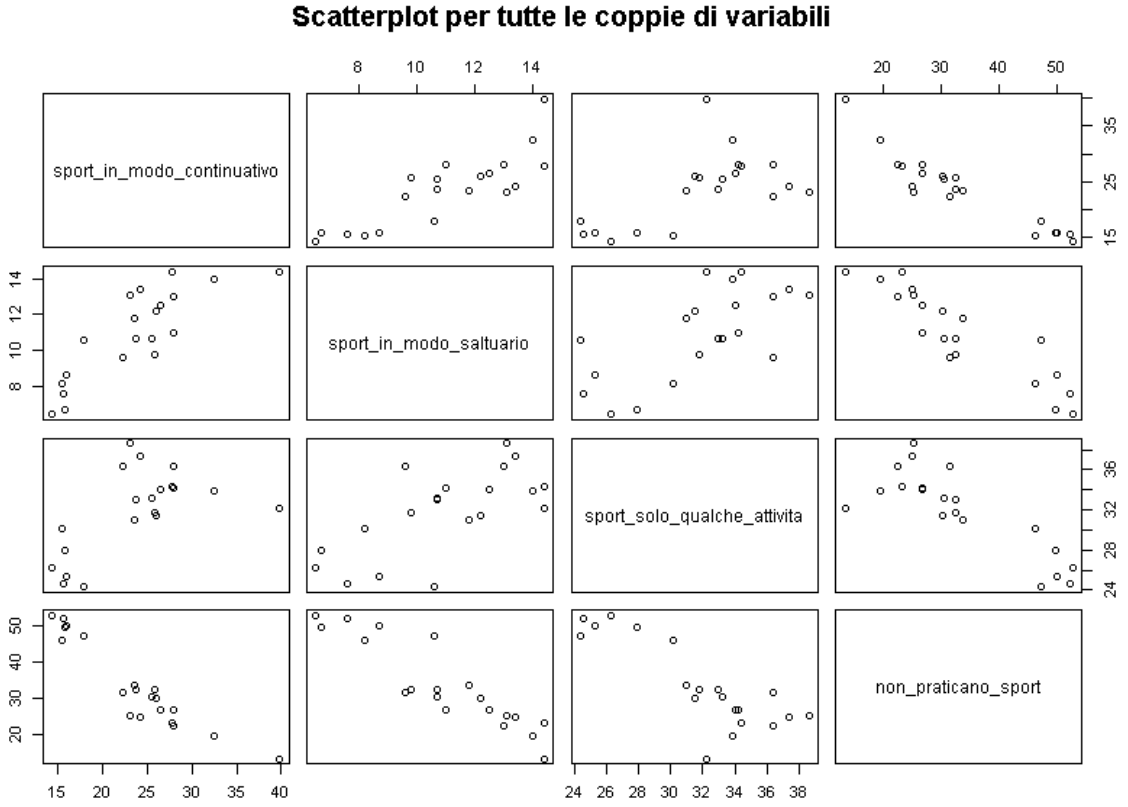
```
curt(non_praticano_sport)
[1] -1.05531
```

La curtosi risulta negativa per i vettori relativi alle persone che praticano sport in modo saltuario, che praticano solo qualche attività fisica e che non praticano sport né attività fisica; risulta invece positiva solo per il vettore relativo alle persone che praticano sport in modo continuativo.

5 - STATISTICA DESCRITTIVA BIVARIATA

La statistica descrittiva bivariata è quel ramo della statistica che si occupa di metodi grafici e statistici volti a descrivere le relazioni che intercorrono tra due variabili.
Le relazioni tra variabili quantitative possono essere rappresentate tramite scatterplot in cui ogni coppia di osservazioni viene rappresentata sotto forma di punti in un piano euclideo, in cui sulle ascisse è presente la variabile indipendente e sulle ordinate la variabile dipendente. Lo scopo è quello di evidenziare se le coppie di punti presentano qualche forma di regolarità e se esiste una relazione tra le variabili e di quale tipo è tale relazione.
Di seguito viene riportata l'analisi prendendo in considerazione come variabile indipendente quella relativa alle persone che praticano sport in modo continuativo e come variabile dipendente quella relativa alle persone che non praticano sport né attività fisica.

Di seguito viene riportato lo scatterplot per tutte le coppie di variabili



5.1 Covarianza campionaria

Assegnato un campione bivariato $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ di una variabile quantitativa bidimensionale (X,Y) siano $\overline{x}$ e $\overline{y}$ rispettivamente le medie campionarie di $x_1,x_2,...,x_n$ e di $y_1,y_2,...,y_n$.
La covarianza campionaria tra le due variabili X e Y è così definita:

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) \qquad (n = 2, 3, \ldots).$$

Si divide per n-1 per normalizzare la sommatoria in quanto se x e y sono uguali, si ottiene la varianza campionaria.
Se la covarianza campionaria è positiva indica che all'aumentare della prima variabile ci si aspetta anche l'aumentare della seconda e viceversa.

La covarianza campionaria può avere segno positivo, negativo o nullo:

- Se $C_{xy} > 0$, allora le variabili sono correlate positivamente, quindi le due variabili hanno un comportamento concorde.
- Se $C_{xy} < 0$, allora le variabili sono correlate negativamente, quindi le due variabili hanno comportamenti mediamente discordi
- Se $C_{xy} = 0$, allora le variabili non sono correlate, quindi le due variabili non sono in relazione diretta tra loro

La covarianza rispetta le seguenti proprietà:

- Cov(X,Y) = Cov(Y,X)
- Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)
- Cov(X,Y) = var(x) dove X=Y

Di seguito viene riportata la covarianza tra le persone che non praticano sport né attività fisica e le persone che praticano sport in modo continuativo:

```
cov(non_praticano_sport, sport_in_modo_continuativo)
[1] -71.80195
```

La covarianza risulta esageratamente negativa, di conseguenza le due variabili sono correlate negativamente.

5.2 Coefficiente di correlazione campionario

Il coefficiente di correlazione è un indice statistico adimensionale, serve a capire se esiste un legame lineare tra due variabili, indipendentemente dalle unità di misura scelte.
Assegnato un campione bivariato $(x_1,y_1),(x_2,y_2),...,(x_n, y_n)$ di una variabile quantitativa bidimensionale (X,Y), siano $\overline{x}$ e s_x la media campionaria e la deviazione standard campionaria di $x_1,x_2,...,x_n$ ed inoltre siano $\overline{y}$ e s_y la media campionaria e la deviazione standard campionaria di $y_1, y_2,..., y_n$.
Il coefficiente di correlazione campionario tra le due variabili X e Y è così definito:

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

Nel caso di due variabili, la formula confronta la distanza di ogni punto di dati dalla media della variabile(quindi è fortemente influenzato da valori anomali), definendo quanto la relazione tra le due variabili si posizionerebbe vicino a una linea immaginaria tracciata tra i dati.
Per questo si dice che le correlazioni sono relazioni lineari.
Il coefficiente di correlazione campionario gode delle seguenti proprietà:

- $-1 \leq r_{xy} \leq 1$
- se esistono due numeri reali a e b, con a>0, tali che $y_i = ax_i+b$ per ogni i=1,2,...,n, allora $r_{xy}=1$
- se esistono due numeri reali a e b, con a<0, tali che $y_i = ax_i +b$ per ogni i=1,2,...,n, allora $r_{xy}=-1$
- se esistono quattro numeri reali a, b, c, d e se risulta $z_i = ax_i+b$ e $w_i=cy_i+d$ per i=1,2,...,n, allora $r_{zw}=r_{xy}$ se ac>0 e $r_{zw}=-r_{xy}$ se invece ac<0
- Può essere calcolato solo se entrambe le variabili sono quantitative

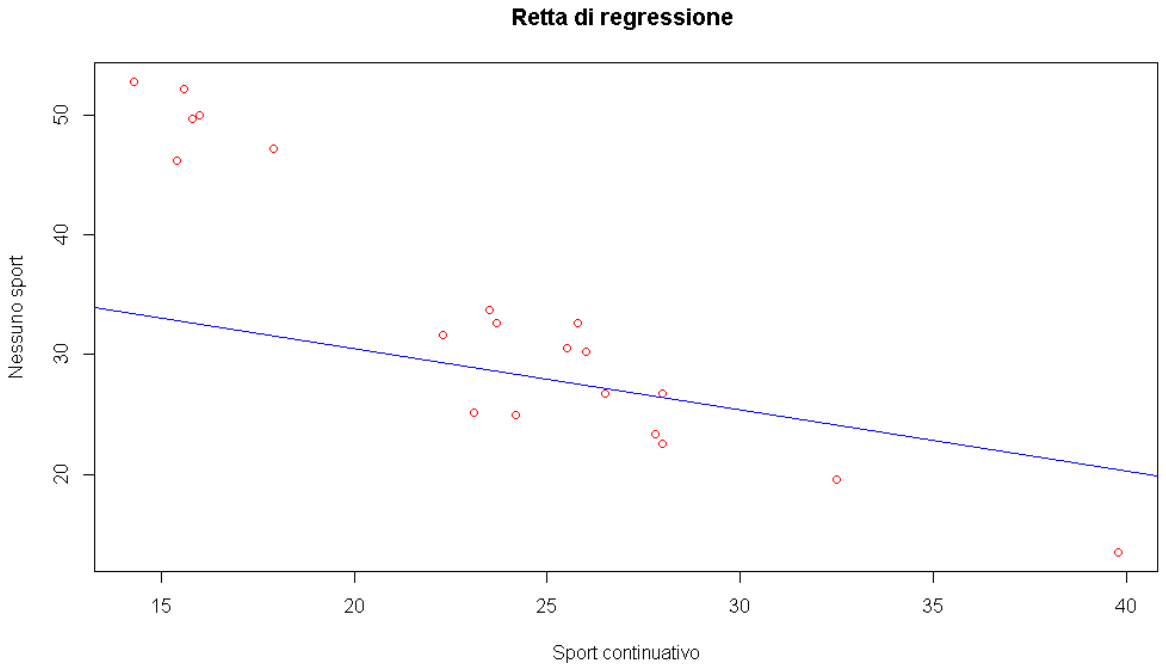
Il coefficiente di correlazione campionario r_{xy} misura la forza del legame di natura lineare esistente tra due variabili quantitative e il segno di r_{xy} indica la direzione di una retta che si crea con eventuali relazioni tra le variabili:

- $r_{xy}=1$ (correlazione perfetta positiva) tutti i punti sono allineati su una linea retta ascendente
- r_{xy} compreso tra 0 e 1 estremi esclusi (correlazione positiva) i punti (x_i,y_i) sono posizionati in una nuvola attorno ad una linea retta ascendente
- $r_{xy}=0$ (nessuna correlazione) i punti sono completamente dispersi in una nuvola che non presenta alcuna direzione lineare
- r_{xy} compreso tra -1 e 0 estremi esclusi (correlazione negativa) i punti (x_i,y_i) sono posizionati in una nuvola attorno ad una linea retta discendente
- $r_{xy}=-1$ (correlazione perfetta negativa) tutti i punti sono allineati su una linea discendente

Di seguito viene riportata la correlazione tra le persone che non praticano sport né attività fisica e le persone che praticano sport in modo continuativo:

```
cor(non_praticano_sport, sport_in_modo_continuativo)
[1] -0.9372844
```

Essendo il risultato prossimo al valore -1 è possibile comprendere che tra le due variabili c'è una forte correlazione negativa con una retta discendente



5.3 Regressione Lineare

Il modello lineare viene usato per spiegare, descrivere o prevedere un andamento futuro sulla base di una relazione tra una variabile Y detta variabile dipendente e una o più variabili indipendenti X_1, X_2, … , X_p.
Quando è presente una sola variabile indipendente l'analisi prende il nome di regressione lineare semplice.
Nel caso di più variabili indipendenti l'analisi prende il nome di regressione lineare multipla.

5.3.1 Regressione lineare semplice

La regressione è una tecnica statistica che si utilizza per modellare le relazioni tra una variabile risposta ed una o più regressori, quando c'è un solo regressore si parla di modello di regressione lineare semplice e si studia quindi la relazione tra due variabili.
Il modello di regressione lineare semplice è esprimibile attraverso l'equazione di una retta che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte le altre possibili rette.
Data l'equazione della retta:

$$Y = \alpha + \beta X$$

dove:

- α è l'intercetta e corrisponde all'ordinata del punto di intersezione della retta di regressione con l'asse delle ordinate
- β è il coefficiente angolare (esprime la pendenza della retta)
    - $\beta > 0$ indica una retta di regressione crescente
    - $\beta < 0$ indica una retta decrescente
    - $\beta = 0$ indica una retta orizzontale

Per identificare questa retta si utilizza il metodo dei minimi quadrati dove i coefficienti di regressione sono α e β dove la somma Q dei quadrati degli errori è minima.

$$Q = \sum_{i=1}^{n} \Big[y_i - (\alpha + \beta\, x_i) \Big]^2$$

dove n è il numero di osservazioni (x_1,x_2,...,x_n) sono i valori osservati della variabile X e (y_1,y_2,...,y_n) sono i valori osservati della variabile Y.
Derivando rispetto ai parametri α e β risulta che:

$$\beta = \frac{s_y}{s_x}\, r_{xy}, \qquad\qquad \alpha = \overline{y} - \beta\, \overline{x}.$$

Di seguito viene riportato il calcolo di α e β tra le persone che praticano sport in modo continuativo e le persone che non praticano sport né attività fisica:

```
beta<-(sd(non_praticano_sport)/sd(sport_in_modo_continuativo))*cor(non_pratica
no_sport,sport_in_modo_continuativo)

alpha<-mean(non_praticano_sport)-beta*mean(sport_in_modo_continuativo)

c(alpha,beta)
[1] 74.247701 -1.723032
```

Dato che β risulta negativa, la retta è discendente.

In R per ottenere gli stessi risultati in minor tempo si utilizza la funzione

lm(y~x)

```
linearmodel<-lm(non_praticano_sport~sport_in_modo_continuativo)
linearmodel

Call:
lm(formula = non_praticano_sport ~ sport_in_modo_continuativo)

Coefficients:
            (Intercept)  sport_in_modo_continuativo
                 74.248                      -1.723
```

La retta di regressione ha come equazione:

$$y=74.248-1.723x$$

5.3.1.1 Residui

I residui sono la differenza tra i valori osservati (coppie(x_i,y_i)) e stimati (coppie (x_i,$\widehat{y_i}$)) in un'analisi di regressione e mostrano di quanto si discostano i valori osservati dai valori stimati con la retta di regressione.
I valori osservati che si trovano al di sopra della curva di regressione hanno un valore residuo positivo e i valori osservati che scendono al di sotto della curva di regressione hanno un valore residuo negativo.
I valori stimati ottenuti mediante la retta di regressione si denotano con

$$\widehat{y_i} = \alpha + \beta x_i$$

I residui sono così definiti

$$E_i = y_i - \widehat{y_i} = y_i - (\alpha + \beta\, x_i) \qquad (i = 1, 2, \ldots, n)$$

I residui godono delle seguenti proprietà:

- La media campionaria dei residui risulta sempre nulla
- Dato che $\overline{E} = 0$ La varianza dei residui è

$$s_E^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(E_i - \overline{E}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} E_i^2,$$

In R per calcolare i valori stimati si utilizza la funzione fitted()
Di seguito viene riportato il vettore dei valori stimati tra le persone che praticano sport in modo continuativo e le persone che non praticano sport né attività fisica:

```
fitted(lm(non_praticano_sport~sport_in_modo_continuativo))
       1        2        3        4        5
29.448879 18.249173 34.445670 26.002815  5.671042
       6        7        8        9       10
26.347422 32.550336 26.002815 28.587363 33.411851
      11       12       13       14       15
30.310394  29.793485 33.756458 47.713014 49.608349
      16       17       18       19       20
43.405435 46.679195 47.023801 47.368408 35.824096


#Si possono ottenere anche da un attributo di linear model

linearmodel$fitted.values
       1        2        3        4        5
29.448879 18.249173 34.445670 26.002815  5.671042
       6        7        8        9       10
26.347422 32.550336 26.002815 28.587363 33.411851
      11       12       13       14       15
```

```
30.310394  29.793485 33.756458 47.713014 49.608349
     16         17         18         19         20
43.405435 46.679195 47.023801 47.368408 35.824096
```

I residui in R si calcolano utilizzando la funzione resid()

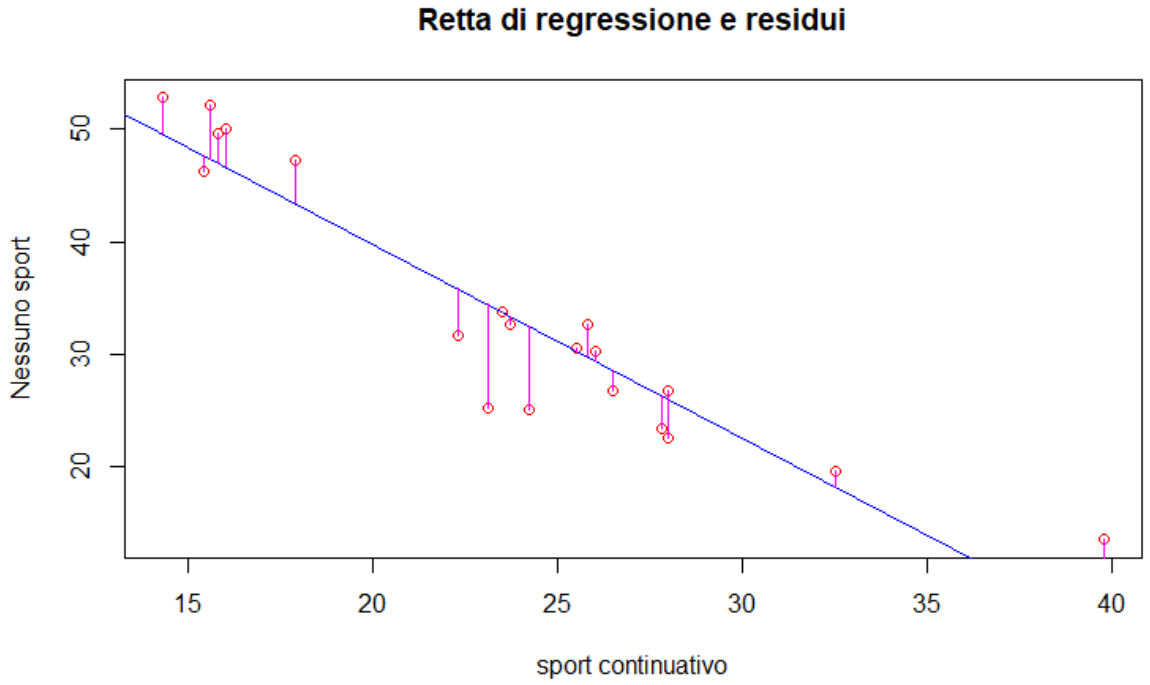
```
resid(lm(non_praticano_sport ~ sport_in_modo_continuativo))
1           2           3           4           5
0.75112139  1.35082698  -9.24567034  -3.40281535 7.82895788
6           7           8           9           10
-2.94742168 -7.55033555 0.79718465 -1.78736280 -0.81185136
11          12          13          14          15
0.18960557  2.80651506 -0.05645769 -1.51301389  3.19165132
16          17          18          19          20
3.79456518  3.32080509  2.67619876  4.83159243 -4.22409564

#Si possono ottenere anche da un attributo di linear model

linearmodel$residuals
1           2           3           4           5
0.75112139  1.35082698  -9.24567034  -3.40281535 7.82895788
6           7           8           9           10
-2.94742168 -7.55033555 0.79718465 -1.78736280 -0.81185136
11          12          13          14          15
0.18960557  2.80651506 -0.05645769 -1.51301389  3.19165132
16          17          18          19          20
3.79456518  3.32080509  2.67619876  4.83159243 -4.22409564
```

Dai residui si può calcolare la mediana, la varianza e la deviazione standard

```
median(linearmodel$residuals)
[1] 0.4703635

var(linearmodel$residuals)
[1] 17.11024

sd(linearmodel$residuals)
[1] 4.136452
```

Di seguito vengono rappresentati i residui in R:

```
plot(sport_in_modo_continuativo,non_praticano_sport,main = "Retta di
regressione e residui", xlab="sport continuativo",ylab="Nessuno sport",col="red")

abline(lm(non_praticano_sport ~ sport_in_modo_continuativo),col="blue")

stime<-fitted(lm(non_praticano_sport~sport_in_modo_continuativo))

segments(sport_in_modo_continuativo,stime,sport_in_modo_continuativo,non_pra
ticano_sport,col="magenta")
```



5.3.1.2 Coefficiente di determinazione

Il coefficiente di determinazione è un indice che misura il legame tra la variabilità dei dati e la correttezza del modello statistico utilizzato ed è definito come il rapporto tra la varianza dei valori stimati e la varianza dei valori osservati:

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

Il coefficiente di determinazione coincide con il quadrato del coefficiente di correlazione nel caso della regressione lineare semplice:

$$D^2 = r_{xy}^2$$

I valori del coefficiente di determinazione variano tra 0 ed 1, quando è più vicino ad 1 significa che i punti tenderanno ad allinearsi lungo la retta di regressione, vicino allo 0 invece non esiste una retta capace di interpolare i punti

Di seguito viene riportato il coefficiente di determinazione tra le persone che praticano sport in modo continuativo e le persone che non praticano sport né attività fisica:

```
(cor(non_praticano_sport,sport_in_modo_continuativo))^2
[1] 0.87850

#Si possono ottenere anche con linearmodel

summary(linearmodel)$r.square
[1] 0.87850
```

5.3.2 Regressione lineare multipla

Il modello di regressione lineare multipla viene usato per descrivere la relazione tra una variabile quantitativa Y, detta variabile dipendente, e più variabili quantitative indipendenti X_1, X_2, … , X_p.
Con le funzioni cov() e cor() si ottengono due matrici simmetriche, in questo caso di dimensione 4x4, i cui elementi sono le covarianze e le correlazioni tra coppie di variabili.
La matrice della covarianza contiene sulla diagonale principale la varianza delle singole colonne del dataframe, mentre la matrice delle correlazioni presenta il numero 1 sulla diagonale principale.

cov(matrice_analisi_sport)

	In modo continuativo	In modo saltuario	Solo qualche attività fisica	Non praticano sport né attività fisica
In modo continuativo	41.67187	13.488079	16.487921	-71.80195
In modo saltuario	13.48808	6.196289	7.206763	-26.92468
Solo qualche attività fisica	16.48792	7.206763	18.043447	-41.82300
Non praticano sport né attività fisica	-71.80195	-26.924684	-41.823000	140.82726

cor(matrice_analisi_sport)

	In modo continuativo	In modo saltuario	Solo qualche attività fisica	Non praticano sport né attività fisica
In modo continuativo	1.0000000	0.8393880	0.6012909	-0.9372844
In modo saltuario	0.8393880	1.0000000	0.6815766	-0.9114679
Solo qualche attività fisica	0.6012909	0.6815766	1.0000000	-0.8296827
Non praticano sport né attività fisica	-0.9372844	-0.9114679	-0.8296827	1.0000000

Esiste una forte correlazione tra le seguenti variabili:

- la variabile relativa alle persone che non praticano sport né attività fisica e la variabile relativa alle persone che praticano sport in modo continuativo

- la variabile relativa alle persone che praticano sport in modo saltuario e la variabile relativa alle persone che non praticano sport né attività fisica

Esiste una correlazione sufficientemente alta tra le seguenti variabili:

- la variabile relativa alle persone che praticano sport in modo saltuario e la variabile relativa alle persone che praticano sport in modo continuativo
- la variabile relativa alle persone che praticano solo qualche attività fisica e la variabile relativa alle persone che non praticano sport né attività fisica
- la variabile relativa alle persone che praticano solo qualche attività fisica e la variabile relativa alle persone che praticano sport in modo continuativo e in modo saltuario
- la variabile relativa alle persone che praticano solo qualche attività fisica e la variabile relativa alle persone che praticano sport in modo saltuario

Le variabili sono principalmente correlate positivamente.
Il modello di regressione lineare multipla con p variabili indipendenti è esprimibile tramite la seguente equazione:

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

dove:

- α è l'intercetta, ovvero il valore di Y quando $X_1 = X_2 = \ldots = X_p = 0$
- $\beta_1, \beta_2, \ldots, \beta_p$ sono i regressori. β_p rappresenta l'inclinazione di Y rispetto alla variabile X_p tenendo costanti le variabili $X_1, X_2, \ldots, X_{p-1}$

Per determinare le stime di $\alpha, \beta_1, \ldots, \beta_p$ si utilizza il metodo dei minimi quadrati ed occorre minimizzare la quantità

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

dove n è il numero di osservazioni, $(x_{1,j}, x_{2,j}, \ldots, x_{n,j})$ sono i valori osservati della variabile X_j (j = 1,2,...,p) e $(y_1, y_2, ..., y_n)$ i valori osservati della variabile Y.
Derivando rispetto ai parametri $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ risulta:

$$\alpha = \overline{y} - \beta_1 \overline{x}_1 - \beta_2 \overline{x}_2 - \ldots - \beta_p \overline{x}_p.$$

In R per la regressione lineare multipla si usa la funzione $lm(y \sim x_1 + x_2 + \ldots + x_p)$, il cui argomento indica che y dipende da $x_1, x_2, \ldots, x_p$

```
multiplelinearmodel
<-lm(non_praticano_sport~sport_in_modo_continuativo+sport_in_modo_saltuario+sport_solo_qualche_attivita)

Call:
lm(formula = non_praticano_sport ~ sport_in_modo_continuativo +
    sport_in_modo_saltuario + sport_solo_qualche_attivita)

Coefficients:
    (Intercept)   sport_in_modo_continuativo   sport_in_modo_saltuario   sport_solo_qualche_attivita
        100.1134                    -1.0063                   -0.9868                    -1.0042
```

L'intercetta α è uguale a 100.1134, i regressori sono β_1 = -1.0063, β_2 = -0.9868, β_3 = -1.0042.
Il modello di regressione multipla stimato è

$$y = 100.1134 - 1.0063\, x_1 - 0.9868\, x_2 - 1.0043\, x_3$$

Tutti i regressori sono negativi, per cui tutte le altre variabili sono legate negativamente alla percentuale relativa alle persone che non praticano sport né attività fisica.

5.3.2.1 Residui

I residui mostrano di quanto si discostano i valori osservati dai valori stimati con la regressione lineare multipla:

$$E_i = y_i - \widehat{y}_i = y_i - (\alpha + \beta_1\, x_{i,1} + \beta_2\, x_{i,2} + \ldots + \beta_p\, x_{i,p}) \qquad (i = 1, 2, \ldots, n)$$

Così come per la regressione lineare semplice, la media campionaria dei residui è nulla e la varianza dei residui è:

$$s_E^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(E_i - \overline{E}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} E_i^2$$

dato che la media campionaria è nulla.
In R per calcolare il vettore dei valori stimati tramite regressione lineare multipla si usa la funzione fitted()

```
stimemult <-
fitted(lm(non_praticano_sport~sport_in_modo_continuativo+sport_in_modo_saltua
rio+sport_solo_qualche_attivita))

      1          2          3          4          5
30.27778 19.55045 25.17800 22.55506 13.51687

      6          7          8          9          10
23.38328 24.98010 26.73789 26.96804 32.56611

     11         12         13         14         15
30.55391 32.54603 33.69037 46.19723 52.89815

     16         17         18         19         20
47.13769 50.02075 49.58457 52.21167 31.64604
```

In R per calcolare il vettore dei residui si usa la funzione resid()

```
residmult <-
resid(lm(non_praticano_sport~sport_in_modo_continuativo+sport_in_modo_saltua
rio+sport_solo_qualche_attivita))

       1                2              3                4                5
-0.077781304  0.049546372  0.021999116  0.044943344 -0.016872162

       6                7              8                9               10
0.016717000  0.019902783  0.062110244 -0.168039212  0.033885661

      11               12             13               14               15
-0.053914760  0.053971610 0.009627795  0.002773677 -0.098145450

      16               17             18               19               20
0.062305973 -0.020745942  0.115427541 -0.011673993 -0.046038292
```

Di seguito vengono riportate media, varianza campionaria e deviazione standard campionaria dei residui

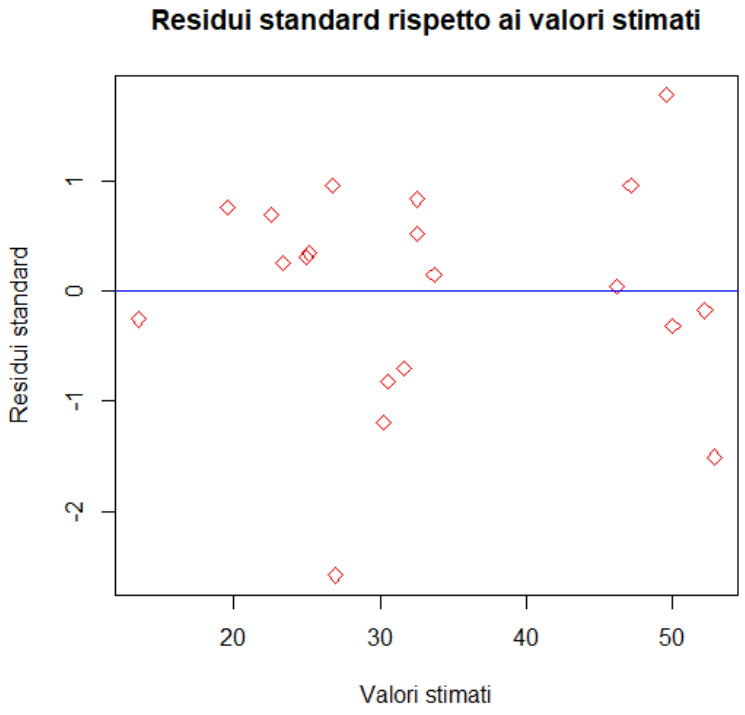
```
median(multiplelinearmodel$residuals)
[1] 0.0131724

var(multiplelinearmodel$residuals)
[1] 0.004245073

sd(multiplelinearmodel$residuals)
[1] 0.06515423
```

Anche nel caso multivariato è interessante calcolare i residui standardizzati:

```
residuimultstandard <- residmult/sd(residmult)

        1          2          3          4          5
-1.19380289  0.76044756  0.33764680  0.68979937 -0.25895729

        6          7          8          9         10
 0.25657583  0.30547186  0.95328036 -2.57909919  0.52008385

       11         12         13         14         15
-0.82749444  0.82836698  0.14776931 0.04257094 -1.50635586

       16         17         18         19         20
 0.95628444 -0.31841284  1.77160481 -0.17917477 -0.70660485
```

Di seguito viene riportato il grafico in cui i residui standardizzati sono disegnati in funzione dei valori stimati:

```
 plot(stimemult, residuimultstandard, main= "Residui standard rispetto ai valori stimati", xlab = "Valori stimati", ylab = "Residui standard", pch=5, col="red")

abline(h=0, col="blue")
```



I punti indicano la posizione dove si collocano i residui standardizzati rispetto ai valori stimati con la retta di regressione. La retta orizzontale posizionata nel punto zero corrisponde alla media campionaria dei residui standardizzati.

5.3.2.2 Coefficiente di determinazione

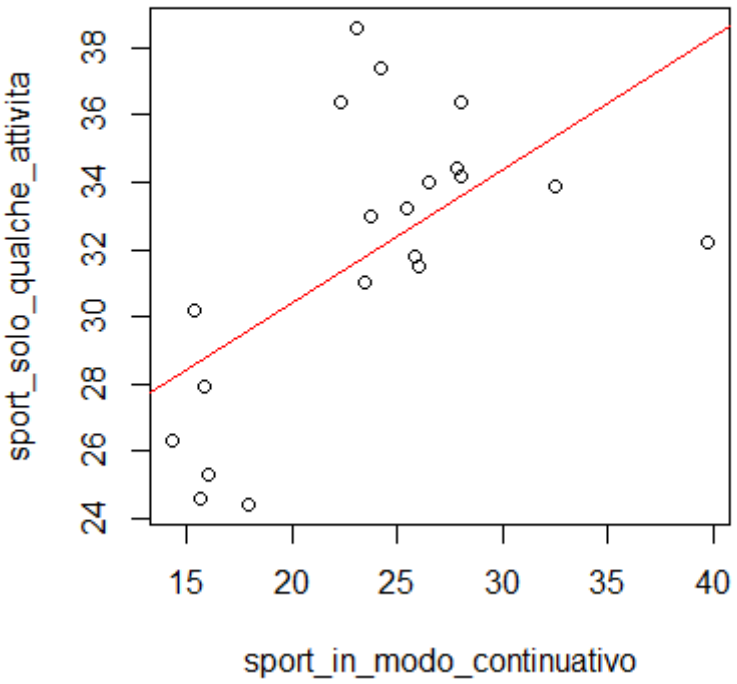
Il coefficiente di determinazione in un modello di regressione lineare multipla è definito come il rapporto tra varianza dei valori stimati tramite la funzione di regressione multipla e la varianza dei valori osservati della variabile dipendente. L'indice del coefficiente è adimensionale e risulta compreso tra 0 e 1, estremi inclusi. Quando l'indice è uguale a 0 il modello di regressione multipla usato non spiega per nulla i dati, quando è uguale a 1 il modello di regressione multipla usato spiega perfettamente i dati.
In R per calcolare l'indice del coefficiente di determinazione per la regressione lineare multipla si usa la funzione summary()

```
summary(multiplelinearmodel)$r.square
[1] 0.9999699
```

Il coefficiente di determinazione è molto vicino ad 1, il modello di regressione multipla usato spiega significativamente i dati.
Rispetto al coefficiente di determinazione relativo al modello di regressione semplice, il cui valore è D^2 = 0.87850, si è ottenuto un miglioramento.

5.3.3 Regressione non lineare

Quando l'ipotesi di linearità di un modello non risulta accettabile, in quanto i dati sperimentali non evidenziano una correlazione di tipo lineare, occorre ricorrere a modelli di regressione non lineare.
Dallo scatterplot contenente tutte le variabili è possibile notare come un modello lineare non sia del tutto adatto ad approssimare la coppia (sport_solo_qualche_attivita, sport_in_modo_continuativo)



La retta di regressione non approssima adeguatamente i dati.
Di seguito viene riportato il coefficiente di determinazione usando il modello lineare:

```
summary(lm(sport_solo_qualche_attivita~sport_in_modo_continuativo))$r.square
[1] 0.3615508
```

In alcuni casi, modelli che sembrano non lineari lo possono diventare usando opportune trasformazioni.
In questo caso la regressione quadratica può essere considerata quella più adatta per l'approssimazione dei dati.
Dal modello polinomiale del secondo ordine:

$$Y = \alpha + \beta X + \gamma X^2.$$

si può ricorrere alla regressione multipla per la stima dei parametri α, β, γ

$$Y = \alpha + \beta X_1 + \gamma X_2$$

con intercetta α, e regressori β e γ per le variabili $X_1 = X$ e $X_2 = X^2$.
In R per stimare i seguenti parametri si utilizza la funzione lm(y˜x + I(x ∧ 2))

```
regressione_quadratica <-lm(sport_solo_qualche_attivita~sport_in_modo_continuativo +
I((sport_in_modo_continuativo)^2))

regressione_quadratica

Call:
lm(formula = sport_solo_qualche_attivita ~ sport_in_modo_continuativo +
    I((sport_in_modo_continuativo)^2))

Coefficients:
                (Intercept)      sport_in_modo_continuativo I((sport_in_modo_continuativo)^2)
                    0.49050                         2.27820                         -0.03757
```

Da cui α = 0.49050, β = 2.27820 e γ = -0.03757, quindi:

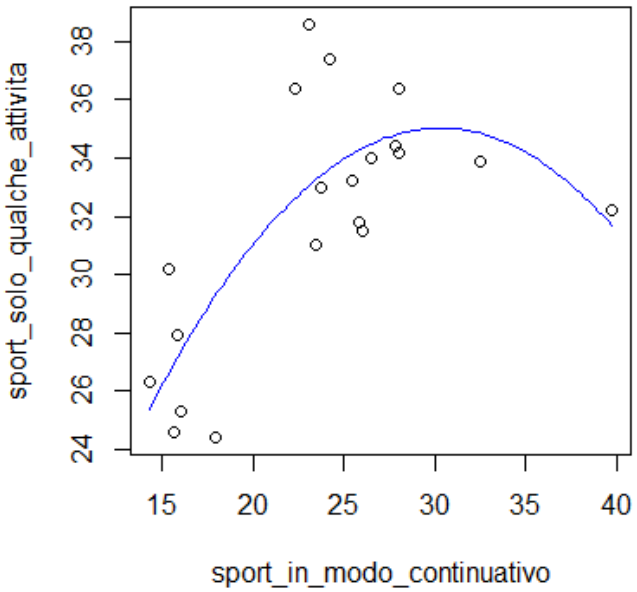
$$Y = 0.49050 + 2.27820X - 0.03757X^2$$

Di seguito viene riportato il coefficiente di determinazione per verificare la correttezza del modello statistico:

```
summary(regressione_quadratica)$r.square
[1] 0.6189822
```

Il risultato ottenuto è migliore rispetto a quello lineare, è stata migliorata la stima.
Di seguito viene riportato lo scatterplot con la curva appena stimata:

```
plot(sport_in_modo_continuativo,sport_solo_qualche_attivita)
alpha <- regressione_quadratica$coefficients[[1]]
beta <- regressione_quadratica$coefficients[[2]]
gamma <- regressione_quadratica$coefficients[[3]]
curve(alpha+beta*x+gamma*x^2, add=TRUE, col="blue")
```


sport_solo_qualche_attivita
24
26
28
30
32
34
36
38
15
20
25
30
35
40
sport_in_modo_continuativo

6 - ANALISI DEI CLUSTER

L’analisi dei cluster è un metodo statistico per processare i dati, raggruppando gli elementi di un insieme a seconda delle loro caratteristiche in classi, dette cluster. I cluster servono a mostrare relazioni fra i dati che a prima vista non risultano evidenti, in modo da creare insiemi omogenei utili per ulteriori analisi.
L’analisi dei cluster permette di raggiungere i seguenti obiettivi:

- individuazione di una reale tipologia
- previsioni basate su gruppi
- esplorazione dei dati
- generazione di ipotesi di ricerca
- verifica di ipotesi
- riduzione della complessità dei dati.

Le tecniche dell’analisi dei cluster possono essere:

- gerarchiche, i quali si suddividono a sua volta in:
    - agglomerativi
    - divisivi
- non gerarchiche

6.1 Distanza e Similarità

L'obiettivo dell'analisi dei cluster è quello di raggruppare gli elementi simili di un insieme in gruppi con un certo grado di omogeneità.
Per definire il termine "somiglianza" di due individui I_i e I_j e per risolvere i problemi di clustering è preferibile parlarne in modo quantitativo.
La somiglianza può essere definita tramite un coefficiente di similarità $s_{ij}=s(X_i,X_j)$ oppure mediante una misura di distanza $d_{ij}=d(X_i,X_j)$ tra due individui I_i e I_j $(i \neq j)$

Una funzione a valori reali $d(X_i,X_j)$ è detta funzione distanza se e soltanto se essa soddisfa le seguenti condizioni:

- $d(X_i , X_j) = 0$ se e solo se $X_i = X_j$, con X_I e X_J in E_P
- $d(X_I , X_J) \geq 0$ per ogni X_I e X_J in E_P
- $d(X_i , X_j) = d(X_j , X_i)$ per ogni X e X_j in E_p
- $d(X_i , X_j) \leq d(X_i , X_k) + d(X_k, X_j)$ per ogni X_i, X_j e X_k in E_p

Le funzioni distanza rispettano anche le seguenti proprietà:

- se d e d ′ sono due metriche anche d + d ′ è una metrica
- se d è una metrica e c un numero reale positivo allora anche cd è una metrica
- se d è una metrica e c un numero reale positivo allora anche d ′ = d/(c + d) è una metrica
- Il prodotto di due metriche (in particolare il quadrato di una metrica) non necessariamente soddisfa la disuguaglianza triangolare e quindi può non essere una misura di distanza

L'obiettivo è quello di costruire e ottenere una matrice delle distanze.
Vi sono molte metriche per calcolare la distanza, nella seguente analisi viene utilizzata la metrica euclidea.
Di seguito viene riportata il comando per calcolare la matrice delle distanze.

```
dist(matrice_analisi_sport,method="euclidean",diag=TRUE,upper=TRUE)
```

6.1.1 Metrica Euclidea

Per calcolare la distanza ci sono molte metriche, le più famose sono:

- Metrica Euclidea
- Metrica del valore asssoluto (Manhattan)
- Metrica del massimo (Chebychev)
- Metrica di Minkowski
- Distanza di Camberra
- Distanza di Jaccard

L'analisi tuttavia si concentrerà solo sulla metrica più famosa, ovvero quella Euclidea, così definita:

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

dove x_{ik} è il valore della k–esima caratteristica dell'individuo I_i.
La distanza Euclidea tuttavia è fortemente influenzata dall'unità di misura utilizzata.

6.1.2 Misure di similarità

Una funzione a valori reali $s_{ij}=(s(X_i,X_j)$ è detta misura di similarità se e soltanto se essa soddisfa le seguenti condizioni:

- $s(X_i , X_i) = 1$
- $0 \leq s(X_i, X_j) \leq 1$
- $s(X_i,X_j)=s(X_j,X_i)$ per ogni X_i e X_j

Una misura di similarità a differenza dalle misure di distanza fornisce un valore compreso tra 0 e 1, dove 0 indica l'assenza totale di similarità mentre 1 la massima presenza di somiglianza.
Tuttavia è inutile calcolare la misura di similarità nella seguente analisi in quanto i valori del dataset sono espressi in percentuali.

6.2 Misure di non omogeneità totale

Alla matrice delle misure e alla matrice delle distanze si può associare la matrice W_x di cardianlità p x p, definita come la matrice delle varianze e covarianze:

$$W_X = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1p} \\ w_{21} & w_{22} & \dots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \dots & w_{pp} \end{pmatrix},$$

dove il generico elemento w_{rl} è definito come:

$$w_{r\ell} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)\,(x_{i\ell} - \overline{x}_\ell) \qquad (r, \ell = 1, 2, \dots, p)$$

se r = ℓ l'elemento w_{rl} è la varianza campionaria relativa della caratteristica r-esima, altrimenti è la covarianza campionaria tra la caratteristica r-esima e la caratteristica ℓ-esima.
In R è possibile ottenere la matrice W_x delle varianze e covarianze campionarie tra le varie caratteristiche tramite la funzione cov().
La matrice di statistica di non omogeneità, di cardinalità p x p è definita come:

$$H_I = (n-1)W_I = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1p} \\ h_{21} & h_{22} & \dots & h_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p1} & h_{p2} & \dots & h_{pp} \end{pmatrix},$$

dove l'elemento generico h_{rl} è definito come:

$$h_{r\ell} = \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)\,(x_{i\ell} - \overline{x}_\ell) = (n-1)\,w_{r\ell} \qquad (r, \ell = 1, 2 \dots, p).$$

Quando r = ℓ allora h_{rr} corrisponde a:

$$h_{rr} = (n-1)\,\mathrm{Var}(C_r) = (n-1)\,s_r^2 \qquad (r = 1, 2 \dots, p).$$

Si definisce misura di non omogeneità statistica dell'insieme I di individui la traccia della matrice H_I :

$$\mathrm{tr}H_I = \sum_{r=1}^{p} h_{rr} = (n-1)\sum_{r=1}^{p} s_r^2.$$

Esprimibile anche in termine della somma dei quadrati delle distanze euclidee tra ogni vettore X_1, X_2, …, X_n e il vettore delle medie campionarie:

$$\mathrm{tr}H_I = \sum_{i=1}^{n} d_2^2(X_i, \overline{X}),$$

dove d_2 indica la distanza euclidea e il vettore $\overline{X}$ è un vettore di cardinalità p, il cui generico elemento rappresenta la media campionaria relativa alla caratteristica j-esima effettuata sugli n individui.
Inoltre, risulta che:

$$\mathrm{tr}H_I = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i}^{n} d_2^2(X_i, X_j),$$

ossia che la traccia della matrice di non omogeneità statistica corrisponde al rapporto tra la somma dei quadrati degli elementi al di sotto della diagonale principale della matrice delle distanze euclidee e il numero n di individui.
La distanza euclidea gioca un ruolo rilevante nel calcolo della misura di non omogeneità statistica, che dipende sia dall'omogeneità interna sia dalla numerosità del gruppo.

6.3 Misure di non omogeneità tra cluster

Al termine del procedimento di classificazione ciò che si vuole ottenere è che gli individui appartenenti allo stesso cluster siano il più possibile omogenei tra loro e il più possibile differenti da quelli che appartengono agli altri cluster.
Per arrivare a ciò si considerano una misura di non omogeneità interna ai cluster (within) e una misura di non omogeneità tra cluster (between).
Per arrivare ad ottenere la matrice di non omogeneità statistica totale, si considerano la somma delle matrici di non omogeneità statistica relative ai singoli cluster e la matrice di non omogeneità statistica tra i cluster considerati:

$$T = S + B$$

dove S è la somma delle matrici di non omogeneità statistica relative ai singoli cluster e B la matrice di non omogeneità statistica tra i cluster.
Per ogni partizione dell'insieme I degli n individui in m fissati cluster, si ottiene un'equazione matriciale come vista in precedenza, da cui segue:

$$\text{tr } T = \text{tr } S + \text{tr } B$$

o equivalentemente:

$$1 = \frac{\text{tr } S}{\text{tr } T} + \frac{\text{tr } B}{\text{tr } T}.$$

Poiché tr T è univocamente determinata per ogni matrice X di cardinalità n x p, i cluster dovrebbero essere individuati in modo da minimizzare la misura di non omogeneità all'interno dei cluster e massimizzare la misura di non omogeneità statistica tra i gruppi.

6.4 Metodi di raggruppamento

Una volta scelta la misura di distanza bisogna procedere alla scelta di un algoritmo di raggruppamento idoneo.
I metodi di raggruppamento si distinguono in tre tipi:

- metodi di enumerazione completa
- metodi gerarchici
- metodi non gerarchici

Le tecniche di enumerazione completa sono computazionalmente onerose poiché prevedono il calcolo delle funzioni di non omogeneità per ogni possibile partizione dell'insieme totale di n individui in m cluster.
Per questo motivo si usano maggiormente metodi di raggruppamento gerarchici e non gerarchici.

6.4.1 Metodi gerarchici

I metodi gerarchici di clustering eseguono una sequenza ordinata di operazioni della stessa natura.
Possono essere di due tipi:

- Agglomerativi
- Divisivi

I metodi gerarchici hanno due vantaggi ovvero quello di fornire una visione completa dell'insieme in termini di distanza e quello di non comportare la scelta a priori del numero di cluster. Uno svantaggio è invece quello che non è consentito riallocare gli individui che sono stati già classificati ad un livello precedente dell'analisi.
L'obiettivo finale dei metodi gerarchici è quello di ottenere una sequenza di partizioni che possono essere rappresentate mediante una struttura ad albero chiamata dendrogramma all'interno del quale sono riportate sulle ordinate i livelli di distanza mentre sulle ascisse i singoli individui.
Il dendrogramma fornisce un quadro completo della struttura dell'insieme in termini delle misure di distanza tra gli individui. Grazie ad esso è facile stabilire a quale stadio dell'analisi gerarchica occorre fermarsi ottenendo la partizione dell'insieme totale di individui in cluster.

6.4.1.1 Metodi gerarchici agglomerativi

I metodi gerarchici di tipo agglomerativi partono da una situazione in cui si hanno n cluster distinti ognuno contenente un solo individuo per giungere, attraverso unioni di cluster meno distinti tra loro, ad una situazione in cui si ha un solo cluster che contiene tutti gli n individui.
Molti metodi di analisi gerarchica hanno una struttura comune che si riflette in un algoritmo generale che ha i seguenti passi:

1. A partire dalla matrice X originaria dei dati o dalla matrice scalata si considera la matrice delle distanze D tra gli individui
2. Si individua la coppia di cluster meno distanti e si raggruppano in un unico cluster; inoltre, si calcola la distanza di questo nuovo cluster da tutti gli altri gruppi già esistenti
3. Si costruisce una nuova matrice di distanza che risulta ridotta di una riga e di una colonna
4. Operare sulla matrice ottenuta a partire dal passo 2 fino ad esaurire tutte le possibilità di raggruppamento, procedura che richiede n-1 iterazioni
5. Si rappresenta graficamente il processo di agglomerazione attraverso un dendrogramma

I vari metodi differiscono tra loro nel passo 1 e nel passo 2. Il passo 1 influenza il metodo richiedendo più o meno forti proprietà mentre il passo 2 caratterizza i metodi in base a come vengono individuati i cluster meno distanti e per il modo in cui si determinano le distanze con i cluster ottenuti.
In R l'analisi gerarchica di tipo agglomerativo viene effettuata attraverso la funzione hclust()

```
hclust(dist(matrice_analisi_sport), method = "complete")
```

Per method sono disponibili alcune opzioni:

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

Di default method è posto a "complete".
La funzione hclust produce come output una lista, i cui elementi sono:

- la sequenza di agglomerazione ($merge)
- un vettore che indica il livello di distanza alla quale è avvenuta l'unione tra due cluster ($height)
- la permutazione delle unità finalizzata alla costruzione del dendrogramma ($order)
- un vettore delle etichette che contrassegnano le varie unità ($labels)

6.4.1.1.1 Metodo del legame singolo

In questo metodo la distanza tra i gruppi è posta pari alla più piccola delle distanze istituibili a due a due tra tutti gli elementi dei due gruppi: se C e D sono due gruppi, la loro distanza, secondo questo metodo, è definita come la più piccola (il minimo) tra tutte le n_1n_2 distanze che si possono calcolare tra ciascuna unità i di C e ciascuna unità j di D
Un vantaggio del metodo del legame singolo è di consentire di individuare gruppi di qualsiasi forma e di evidenziare la presenza di eventuali valori anomali (meglio di altre tecniche gerarchiche).
Uno svantaggio è che invece esso può dare origine alla formazione di una catena tra
gli individui in quanto il legame singolo basa l'unione di due cluster G_1 e G_2, di numerosità n_1e n_2, su un solo legame, quello corrispondente alla distanza più piccola tra gli n_1n_2 individui esistenti.
Si può verificare che si vengano a trovare nello stesso cluster individui piuttosto dissimili e di conseguenza il metodo del legame singolo non è sempre affidabile

L'algoritmo funziona nel seguente modo:

- Inizialmente al livello 0 si considera un insieme di n cluster $\{I_1\},\{I_2\},...,\{I_n\}$,
    - si cerca nella matrice D delle distanze il coefficiente di distanza minima
    - si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
    - Nel caso i coefficienti di distanza minima siano più di uno si attua una scelta arbitraria tra esse
- Al livello 1
    - si modifica la matrice delle distanze valutando le distanze di G_{ij} da ogni altro individuo I_k non appartenente a G_{IJ}
    - la distanza dell'individuo I_k dal cluster G_{ij} si ottiene scegliendo la più piccola tra le due distanze d_{ik} e d_{jk}:

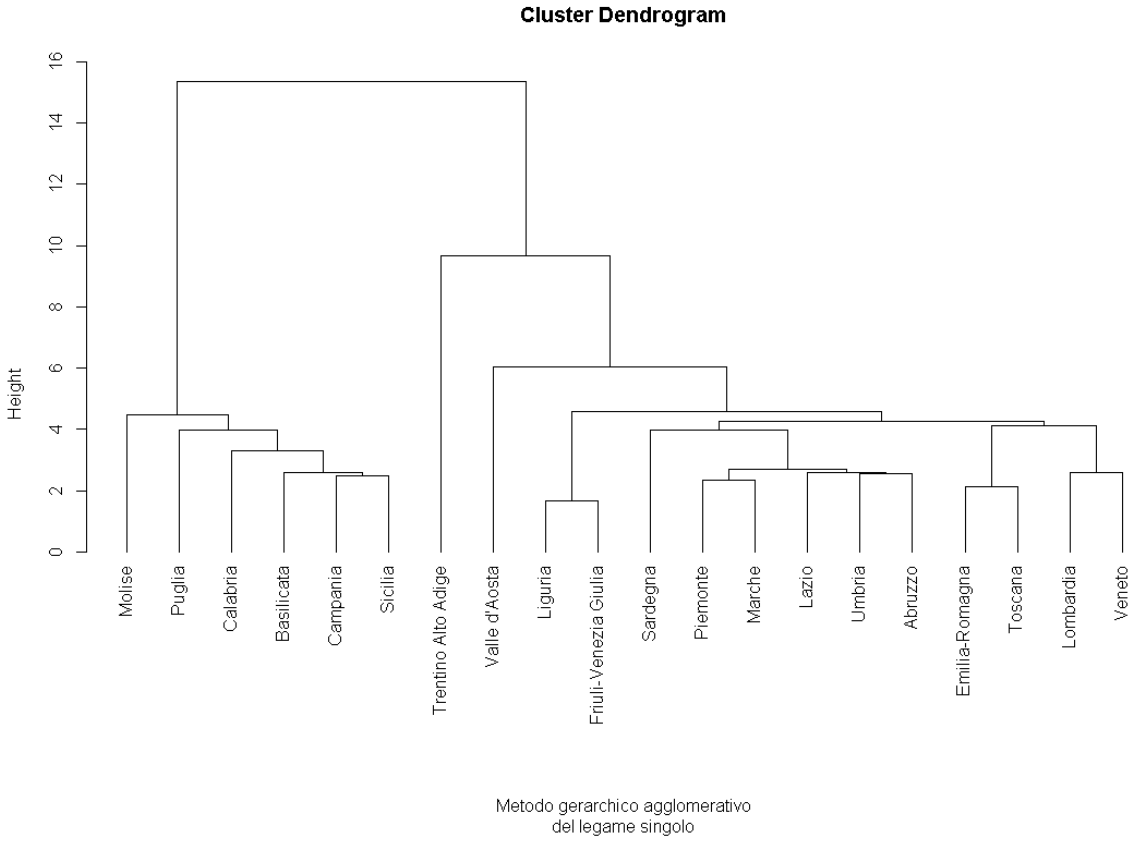
$$d_{(ij),k} = \min(d_{ik}, d_{jk}).$$

    - si costruisce una nuova matrice D_1 di cardinalità (n-1)x(n-1) costituita da G_{ij}, considerato come un unico elemento, e dagli n-2 individui esterni a G_{ij}
- Ad ogni passo successivo
    - dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

$$d_{(uv),z} = \min(d_{uz}, d_{vz}).$$

Di seguito viene riportato il calcolo del legame singolo in R e la costruzione del dendrogramma in R:

```
legameSingolo<-hclust(matriceDistanzeEuclidee,method="single")

plot(legameSingolo,hang=-1,xlab="Metodo gerarchico agglomerativo"+sub="del legame singolo")
```



Di seguito viene calcolata la misura di non omogeneità statistica con un partizionamento in 3 gruppi:

```
taglio <-cutree (legameSingolo , k =3, h = NULL)
num <-table (taglio )
tagliolist <- list(taglio)
agvar <- aggregate (matrice_analisi_sport, tagliolist , var)[, -1]

#Calcolo misura di non omogeneità totale
numeroRighe <- nrow(matrice_analisi_sport)
trHI <- (numeroRighe-1) *sum(apply(matrice_analisi_sport,2, var ))

#Calcolo misura di non omogeneità tra i cluster
trH1 <-(num [[1]] -1) * sum(agvar [1, ])
trH2 <-(num [[2]] -1) *sum(agvar [2, ])
trH3 <-(num [[3]] -1) *sum(agvar [3, ])

sum <- trH1 + trH2 +trH3
trB <- trHI - sum
trB/trHI

[1] 0.8723165
#La suddivisione va più che bene
```

6.4.1.1.2 Metodo del legame completo

In questo metodo la distanza tra i gruppi G_1 (contenente n_1 individui) e G_2 (contenente $n_2$individui) è definita come la massima tra tutte le n_1n_2 distanze che si possono calcolare tra ogni individuo di G_1 e ogni individuo di G_2

L'algoritmo funziona nel seguente modo:

- Inizialmente al livello 0 si considera un insieme di n cluster $\{I_1\},\{I_2\},...,\{I_n\}$,
    - si cerca nella matrice D delle distanze il coefficiente di distanza minima e si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
    - Nel caso i coefficienti di distanza minima siano più di uno si attua una scelta arbitraria tra esse
- Al livello 1
    - si modifica la matrice delle distanze valutando le distanze di G_{ij} da ogni altro individuo I_k non appartenente a G_{ij}
    - la distanza dell'individuo I_k dal cluster G_{ij} si ottiene scegliendo la più grande tra le due distanze d_{ik} e d_{jk}

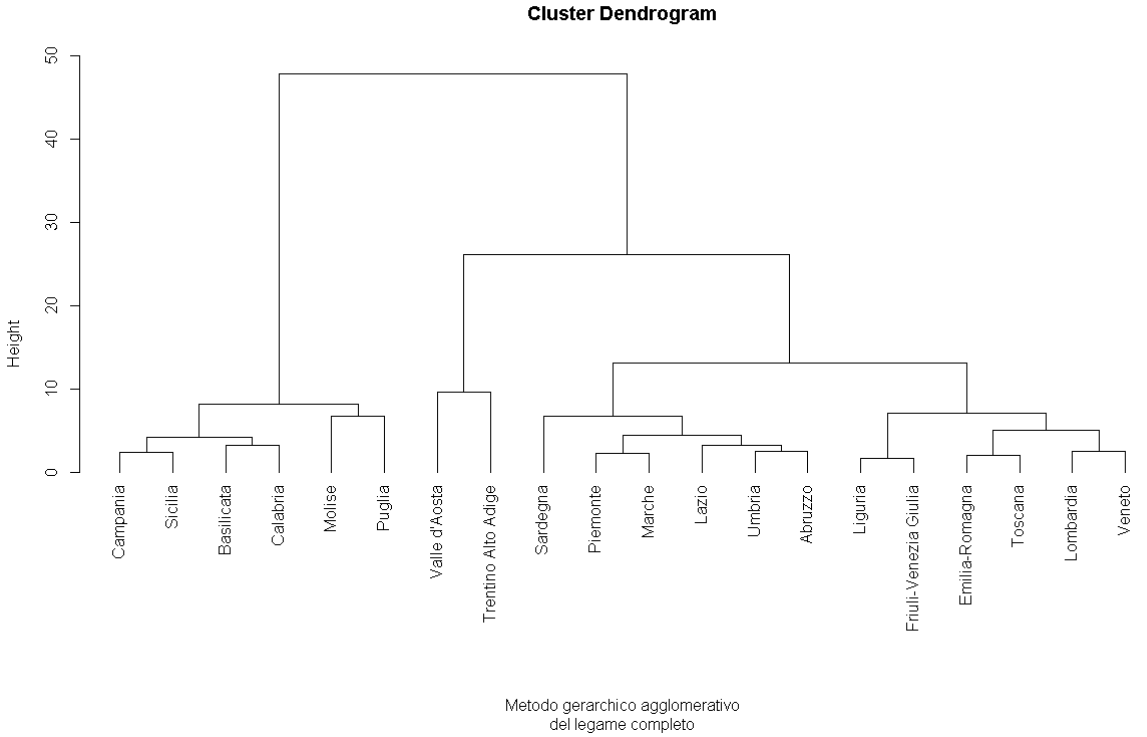
$$d_{(ij),k} = \max(d_{ik}, d_{jk}).$$

    - si costruisce una nuova matrice di cardinalità (n−1)x(n−1) costituita da G_{ij} , considerato come un unico elemento, e dagli n−2 individui esterni a G_{ij}
- Ad ogni passo successivo
    - dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

$$d_{(uv),z} = \max(d_{uz}, d_{vz})$$

Di seguito viene riportato il calcolo del legame completo in R e la costruzione del dendrogramma in R:

```
legameCompleto <- hclust(matriceDistanzeEuclidee,method="complete")
plot(legameCompleto,hang=-1,xlab="Metodo gerarchico agglomerativo",sub="del legame completo")
```



Di seguito viene calcolata la misura di non omogeneità statistica con un partizionamento in 3 gruppi:

```
taglio <-cutree (legameCompleto, k =3, h = NULL)
num <-table (taglio )
tagliolist <- list(taglio)
agvar <- aggregate (matrice_analisi_sport, tagliolist , var)[, -1]

#Calcolo misura di non omogeneità tra i cluster
trH1 <-(num [[1]] -1) * sum(agvar [1, ])
trH2 <-(num [[2]] -1) *sum(agvar [2, ])
trH3 <-(num [[3]] -1) *sum(agvar [3, ])

sum <- trH1 + trH2 +trH3
trB <- trHI - sum
trB/trHI
[1] 0.8917621
```

6.4.1.1.3 Metodo del legame medio

In questo metodo la distanza tra i gruppi è definita come la media aritmetica delle distanze tra tutte le coppie di unità che compongono i due gruppi.
L'algoritmo funziona nel seguente modo:

- Inizialmente al livello 0 si considera un insieme di n cluster $\{I_1\}$, $\{I_2\}$,…, $\{I_n\}$,
  - si cerca nella matrice D delle distanze il coefficiente di distanza minima
  - si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
- Al livello 1
  - si modifica la matrice delle distanze valutando le distanze di G_{ji} da ogni altro individuo I_k che non appartiene a G_{ij} attraverso la seguente relazione

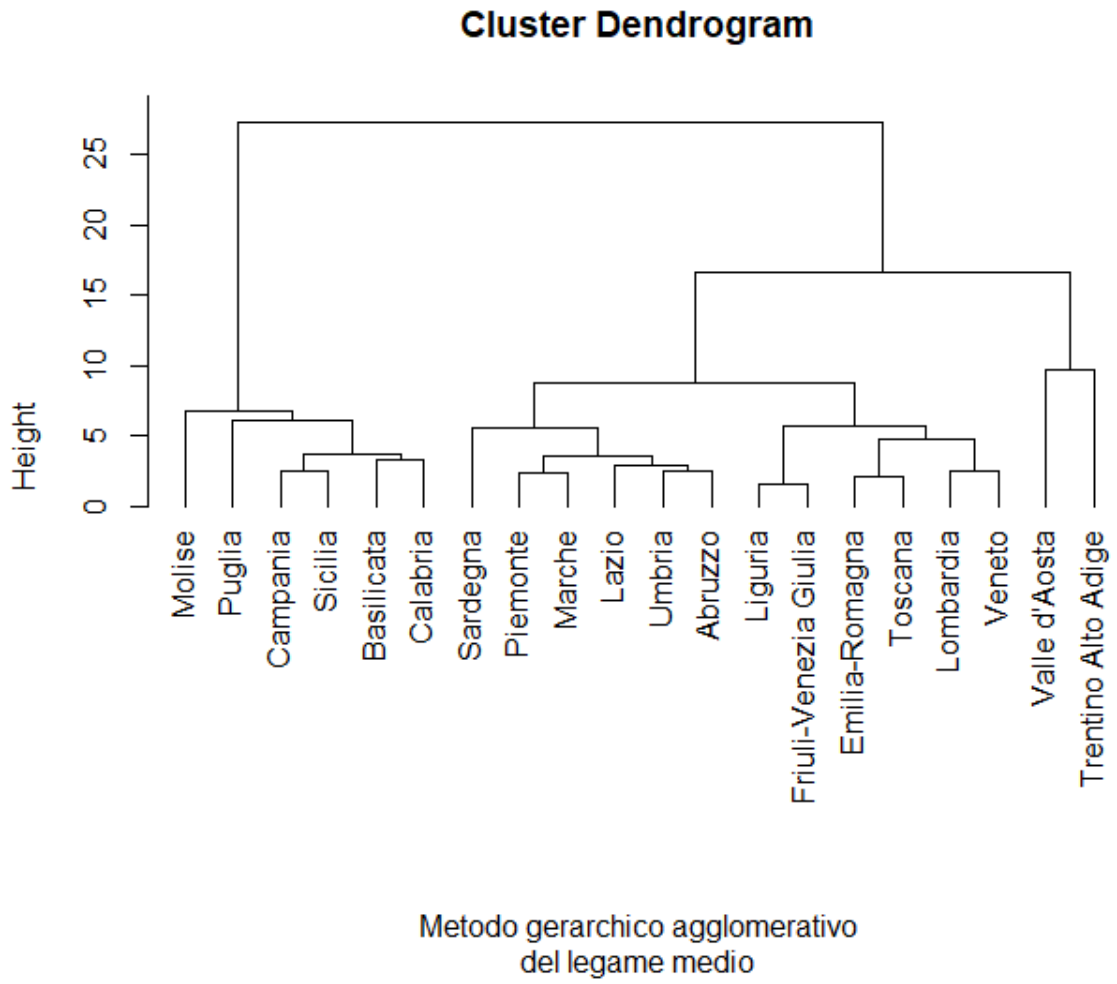
$$d_{(i,j),k} = \frac{1}{2}\left(d_{i,k} + d_{j,k}\right) \qquad (k = 1, 2, \ldots, n; k \neq i, j)$$

  - si costruisce una nuova matrice di cardinalità (n-1) x (n-1) costituita da G_{ij}, considerato come unico elemento, e dagli n-2 individui esterni a G_{ij}
- Ad ogni passo successivo
  - dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice delle distanze i cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

$$\begin{aligned} d_{(uv),z} &= \frac{1}{(N_u + N_v)\, N_z} \sum_{\{i:I_i \in G_{uv}\}} \sum_{\{j:I_j \in G_z\}} d_{ij} \\ &= \frac{1}{(N_u + N_v)\, N_z} \sum_{\{i:I_i \in G_u\}} \sum_{\{j:I_j \in G_z\}} d_{ij} + \frac{1}{(N_u + N_v)\, N_z} \sum_{\{i:I_i \in G_v\}} \sum_{\{j:I_j \in G_z\}} d_{ij} \\ &= \frac{N_u}{N_u + N_v} d_{uz} + \frac{N_v}{N_u + N_v} d_{vz}, \end{aligned}$$

Uno svantaggio di questo metodo è che se le misure dei due cluster da unire sono molto differenti la distanza $d_{(uv),z}$ sarà molto vicina a quella del cluster più numeroso.
Di seguito viene riportato il calcolo del legame medio in R e la costruzione del dendrogramma:

```
legameMedio <- hclust(matriceDistanzeEuclidee, method="average")

plot(legameMedio, hang=-1, xlab="Medoto gerarchico agglomerativo", sub="del legame medio")
```



Di seguito viene riportata la misura di non omogeneità statistica con un partizionamento in 3 gruppi:

```
taglio <- cutree(legameMedio, k=3, h=NULL)
num <-table(taglio)
tagliolist <- list(taglio)
agvar <- aggregate(matrice_analisi_sport, tagliolist, var)[,-1]

trH1 <- (num[[1]]-1)*sum(agvar[1,])
trH2 <- (num[[2]]-1)*sum(agvar[2,])
trH3 <- (num[[3]]-1)*sum(agvar[3,])

sum <- trH1 + trH2 + trH3
trB <- trHI - sum
trB/trHI
[1] 0.8917621
```

6.4.1.1.4 Metodo del centroide

A differenza dei metodi agglomerativi del legame singolo, completo e medio, che usano una qualsiasi misura di distanza, nel metodo del legame del centroide e della mediana si considera la distanza euclidea e si lavora con una matrice che contiene i quadrati delle singole distanze euclidee.
In questo metodo la distanza tra i gruppi è definita come la distanza tra i centroidi, ovvero tra le medie campionarie calcolate sugli individui appartenenti ai due gruppi. 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.
L'algoritmo funziona nel seguente modo:

- Inizialmente al livello 0 si considera un insieme di n cluster $\{I_1\}, \{I_2\},\ldots, \{I_n\}$,
  - si cerca nella matrice $D^{(2)}$, contenente i quadrati delle singole distanze euclidee, il coefficiente di distanza minima
  - si raggruppano nello stesso cluster G_{ij} i due individui I_i e I_j associati secondo tale coefficiente
- Al livello 1
  - si modifica la matrice dei quadrati delle distanze valutando i quadrati delle distanze di G_{ji} da ogni altro individuo I_k che non appartiene a G_{ij} attraverso la seguente relazione

$$d^2_{(ij),k} = \sum_{r=1}^{p} (\overline{x}_{(i,j),r} - \overline{x}_{k,r})^2 = \frac{1}{2}(d^2_{ik} + d^2_{jk}) - \frac{1}{4}d^2_{ij}, \qquad (k \neq i,j)$$

    dove

$$\overline{x}_{(i,j),r} = \frac{1}{2}(x_{i,r} + x_{j,r}) \qquad \overline{x}_{k,r} = x_{k,r} \qquad (r = 1,2,\ldots,p).$$

  - si modifica la matrice nel seguente modo

$$X_1 = \begin{array}{c} \\ I_1 \\ I_2 \\ \vdots \\ I_{i,j} \\ \vdots \\ I_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \ldots & C_p \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_{(i,j),1} & \overline{x}_{(i,j),2} & \ldots & \overline{x}_{(i,j),p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \end{array}$$

    ottenendo una matrice di cardinalità (n-1) x p.
- Ad ogni passo successivo
  - dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice dei quadrati delle distanze euclidee i due cluster più vicini, la

distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

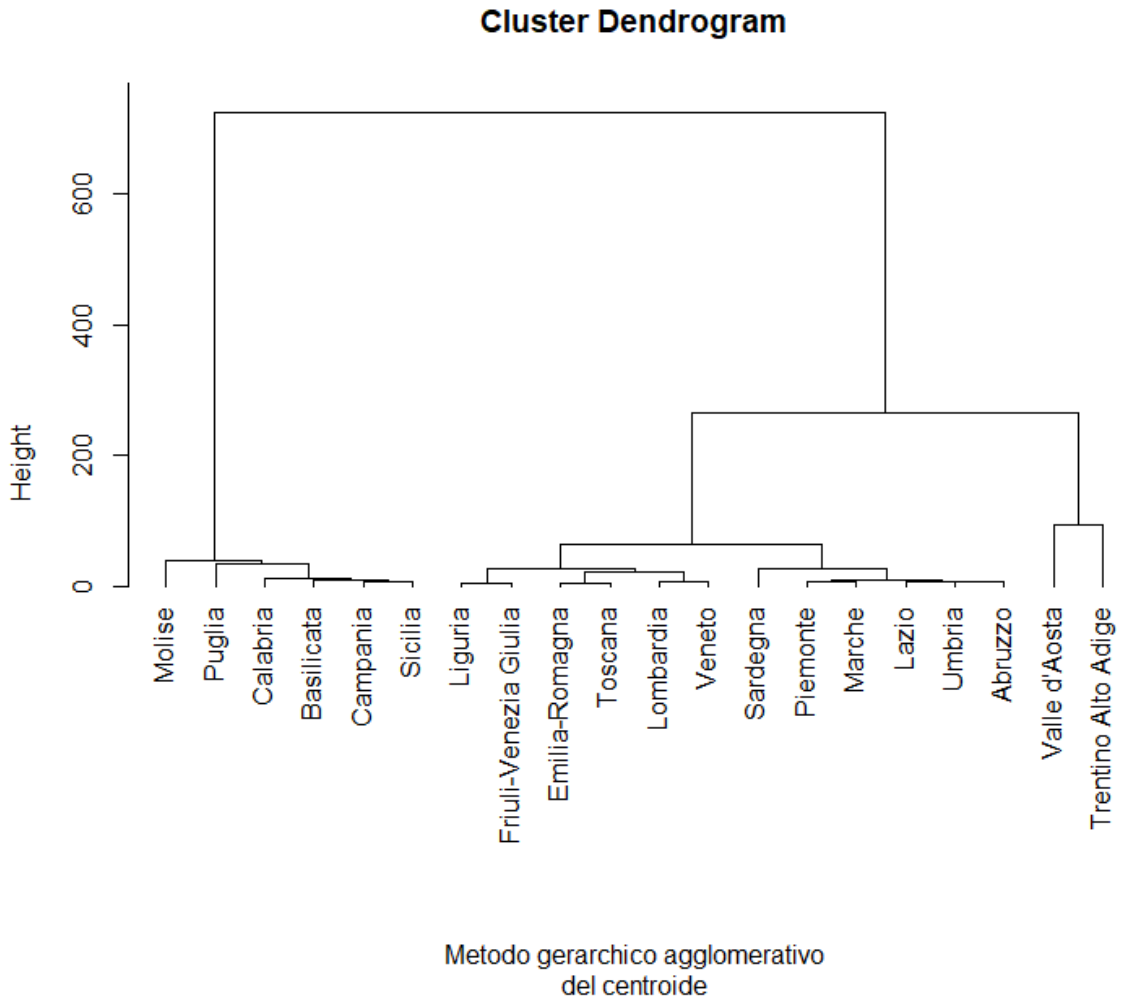
$$d^2_{(uv),z} = \sum_{k=1}^{p} (\overline{x}_{(u,v),k} - \overline{x}_{(z),k})^2 = \frac{N_u}{N_u + N_v} d^2_{uz} + \frac{N_v}{N_u + N_v} d^2_{vz} - \frac{N_u N_v}{(N_u + N_v)^2} d^2_{u,v},$$

dove

$$\begin{aligned}
\overline{x}_{(u,v),r} &= \frac{1}{N_u + N_v} \sum_{\{i:I_i \in G_{uv}\}} x_{i,r} \\
&= \frac{1}{N_u + N_v} \sum_{\{i:I_i \in G_u\}} x_{i,r} + \frac{1}{N_u + N_v} \sum_{\{i:I_i \in G_v\}} x_{i,r} \\
&= \frac{N_u}{N_u + N_v} \overline{x}_{(u),r} + \frac{N_v}{N_u + N_v} \overline{x}_{(v),r} \\
& \qquad\qquad\qquad\qquad\qquad\qquad (r = 1, 2, \ldots, p) \\
\overline{x}_{(z),r} &= \frac{1}{N_z} \sum_{k:I_k \in G_z} x_{kr}
\end{aligned}$$

Uno svantaggio di questo metodo è che se le misure dei due cluster da unire sono molto differenti il centroide del nuovo cluster sarà molto vicino a quello del cluster più numeroso.
Di seguito viene riportato il calcolo del metodo del centroide in R e la costruzione del dendrogramma:

```
#Calcolo della matrice dei quadrati delle distanze euclidee
matriceDistanzeEuclideeQuadrato <- matriceDistanzeEuclidee^2

metodoCentroide <- hclust(matriceDistanzeEuclideeQuadrato, method =
"centroid")

plot(metodoCentroide, hang=-1, xlab="Metodo gerarchico agglomerativo",
sub="del centroide")
```



Di seguito viene riportata la misura di non omogeneità statistica con un partizionamento in 3 gruppi:

```
taglio <- cutree(metodoCentroide, k=3, h=NULL)
num <-table(taglio)
tagliolist <- list(taglio)
agvar <- aggregate(matrice_analisi_sport, tagliolist, var)[,-1]

trH1 <- (num[[1]]-1)*sum(agvar[1,])
trH2 <- (num[[2]]-1)*sum(agvar[2,])
trH3 <- (num[[3]]-1)*sum(agvar[3,])

sum <- trH1 + trH2 + trH3
trB <- trHI - sum
trB/trHI
[1] 0.8917621
```

6.4.1.1.5 Metodo della mediana

Il metodo della mediana è simile a quello del centroide, con la differenza che la procedura è indipendente dalla numerosità dei cluster.
Quando due gruppi si aggregano, il nuovo centroide è calcolato come la semisomma dei due centroidi precedenti.
Il metodo della mediana, così come il metodo del legame singolo, può dare origine alla formazione di una catena tra gli individui.
L'algoritmo al livello 0 e al livello 1 coincide con il metodo del centroide

- Ad ogni passo successivo
    - dopo che i cluster G_u e G_v sono stati uniti scegliendo dalla precedente matrice dei quadrati delle distanze euclidee i due cluster più vicini, la distanza tra il nuovo cluster, denotato con G_{uv}, e un altro cluster G_z è così definita

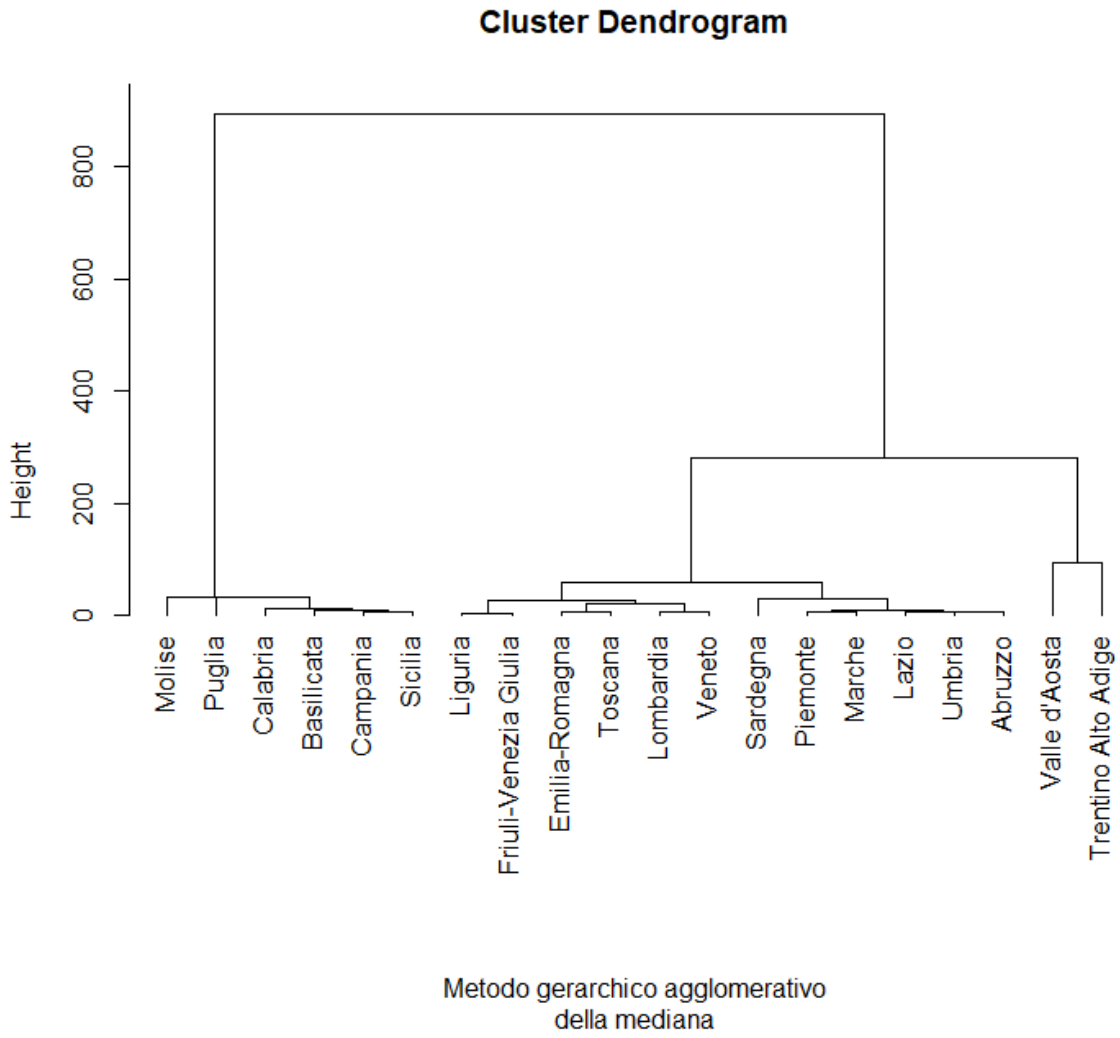
$$d^2_{(uv),z} = \sum_{k=1}^{p} (\overline{x}_{(u,v),k} - \overline{x}_{(z),k})^2 = \frac{1}{2} d^2_{u,z} + \frac{1}{2} d^2_{v,z} - \frac{1}{4} d^2_{u,v},$$

dove

$$\overline{x}_{(uv),r} = \frac{1}{2} (\overline{x}_{(u),r} + \overline{x}_{(v),r}) \qquad (r = 1, 2, \ldots, p).$$

Di seguito viene riportato il calcolo del metodo del centroide in R e la costruzione del dendrogramma:

```
metodoMediana <- hclust(matriceDistanzeEuclideeQuadrato, method = "median")

plot(metodoMediana, hang=-1, xlab="Metodo gerarchico agglomerativo",
sub="della mediana")
```



Di seguito viene riportata la misura di non omogeneità statistica con un partizionamento in 3 gruppi:

```
taglio <- cutree(metodoMediana, k=3, h=NULL)
num <-table(taglio)
tagliolist <- list(taglio)
agvar <- aggregate(matrice_analisi_sport, tagliolist, var)[,-1]

trH1 <- (num[[1]]-1)*sum(agvar[1,])
trH2 <- (num[[2]]-1)*sum(agvar[2,])
trH3 <- (num[[3]]-1)*sum(agvar[3,])

sum <- trH1 + trH2 + trH3
trB <- trHI - sum
trB/trHI
[1] 0.8917621
```

Nel metodo del legame singolo è stata ottenuta una misura di non omogeneità inferiore (0.8723165) rispetto ai metodi del legame completo, del legame medio, del centroide e della mediana che presentano un risultato maggiore (0.8917621) .

Di seguito viene riportato il clustering ottenuto:

Piemonte	Abruzzo	Liguria	Lombardia
1	1	1	1
Sardegna	Veneto	Friuli-Venezia Giulia	Emilia-Romagna
1	1	1	1
Toscana	Umbria	Marche	Lazio
1	1	1	1
Valle d'Aosta	Trentino Alto Adige	Campania	Puglia
2	2	3	3
Basilicata	Calabria	Sicilia	Molise
3	3	3	3

6.4.2 Metodi non gerarchici

Sono metodi di classificazione che forniscono una partizione dei dati in un numero di gruppi che bisogna decidere prima di classificare i dati.
Di solito anche in questo caso si tratta di algoritmi aggregativi che nei passi successivi cercano di migliorare la partizione ottenuta.

Il metodo non gerarchico più noto è quello delle k-medie (K-MEANS).

1. All'inizio bisogna specificare a priori il numero k di cluster e specificare m punti di riferimento iniziali per produrre una prima partizione provvisoria
2. Ad esse associa, formando k gruppi, le unità che si trovano più vicine a ciascuna, poi ricalcola le medie dei gruppi e le utilizza come centroidi a cui aggregare i gruppi.
3. In questo momento alcune unità che all'inizio erano state assegnate ad un gruppo possono cambiare gruppo rivalutando la distanza di ogni individuo da ogni centroide e nel caso in cui la distanza minima sia col centroide di un altro gruppo, l'individuo viene spostato in quel gruppo.
4. L'algoritmo continua fino a che nessuna unità cambia più gruppo, quindi fino a che non si raggiunge l'ottimo.

Da notare che talvolta questo ottimo non è l'ottimo globale, cioè l'algoritmo può fallire il suo obiettivo. Per questo talvolta si ripete più volte cambiando le prime k unità in modo da confrontare i risultati e prendere quelli che si verificano più volte

Di seguito viene riportato il calcolo della matrice delle distanze euclidee in R:

```
matriceDistanzeEuclidee <-
dist(matrice_analisi_sport,method="euclidean",diag=TRUE,upper=TRUE)
```

Di seguito invece si applica il k-means alla matrice appena ottenuta

km<-kmeans(matriceDistanzeEuclidee,centers=3,iter.max=20,nstart=10)
#3 Cluster 10 tentativi 20 iterazioni

K-means clustering with 3 clusters of sizes 12, 2, 6

Cluster means

	Piemonte	Valle d'Aosta	Liguria	Lombardia	Trentino Alto Adige
1	5.385002	12.029080	7.045059	7.46288	21.314116
2	17.288853	4.836062	16.676670	10.79331	4.836062
3	23.067019	35.776680	28.826806	31.79657	44.313340

	Veneto	Friuli-Venezia Giulia	Emilia-Romagna	Toscana	Umbria
1	7.026731	6.308127	5.446548	4.963001	5.929329
2	10.894645	15.305274	13.611176	14.241181	20.666163
3	30.716797	28.870357	27.328438	26.839418	20.247570

	Marche	Lazio	Abruzzo	Molise	Campania
1	4.951738	6.256022	7.134983	21.305748	28.583460
2	17.964518	19.743854	21.616392	36.839480	43.606284
3	22.813088	20.728987	18.852394	5.617855	4.388059

	Puglia	Basilicata	Calabria	Sicilia	Sardegna
1	22.71122	25.53212	24.886384	27.935867	6.746942
2	36.96511	40.25887	40.007660	42.592651	21.257139
3	5.21757	3.36543	3.664129	3.968804	21.822150

Clustering vector:

Piemonte	Valle d'Aosta	Liguria	Lombardia
1	2	1	1
Trentino Alto Adige	Veneto	Friuli-Venezia Giulia	Emilia-Romagna
2	1	1	1
Toscana	Umbria	Marche	Lazio
1	1	1	1
Abruzzo	Molise	Campania	Puglia
1	3	3	3
Basilicata	Calabria	Sicilia	Sardegna
3	3	3	1

Within cluster sum of squares by cluster:
[1] 3141.7121 829.9937 790.3811
 (between_SS / total_SS = 89.5 %)

Available components:

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

La misura di non omogeneità è calcolata (anche) con:

```
km$betweenss/km$totss

[1] 0.8948887
```