Università di Genova | MaLGa

Esplorazione e visualizzazione dei dati

Introduzione alla Data Science

Nicoletta Noceti

Oggi discutiamo di…

- **Esplorazione dei dati: il primo passo per capire qualcosa di più riguardo ai dati**
  - OLAP
  - Analisi statistica
- **Comunicare visivamente dati e risultati**
  - Anche «non esperti» devono poter fruire i risultati
  - L’efficacia della data science dipende anche da come viene spiegata

Più precisamente parliamo di...

- Semplici statistiche sui dati (media, mediana, deviazione standard, ...)
- Tecniche di visualizzazione (istogrammi, plot, ...)
- OLAP (esplorazione multi-dimensionale)

UniGe | MaLGa

(Ancora) qualche semplice statistica da cui partire

Statistiche

- Frequenza, media, moda, mediana, deviazione standard sono semplici quantità che possono darci interessanti informazioni sulla distribuzione dei valori all'interno dei nostri dati
- Abbiamo già discusso che, a seconda della tipologia di dato, può essere utile o necessario utilizzare solo alcune di esse
- Rivediamole brevemente...

Statistiche

Frequenza e moda

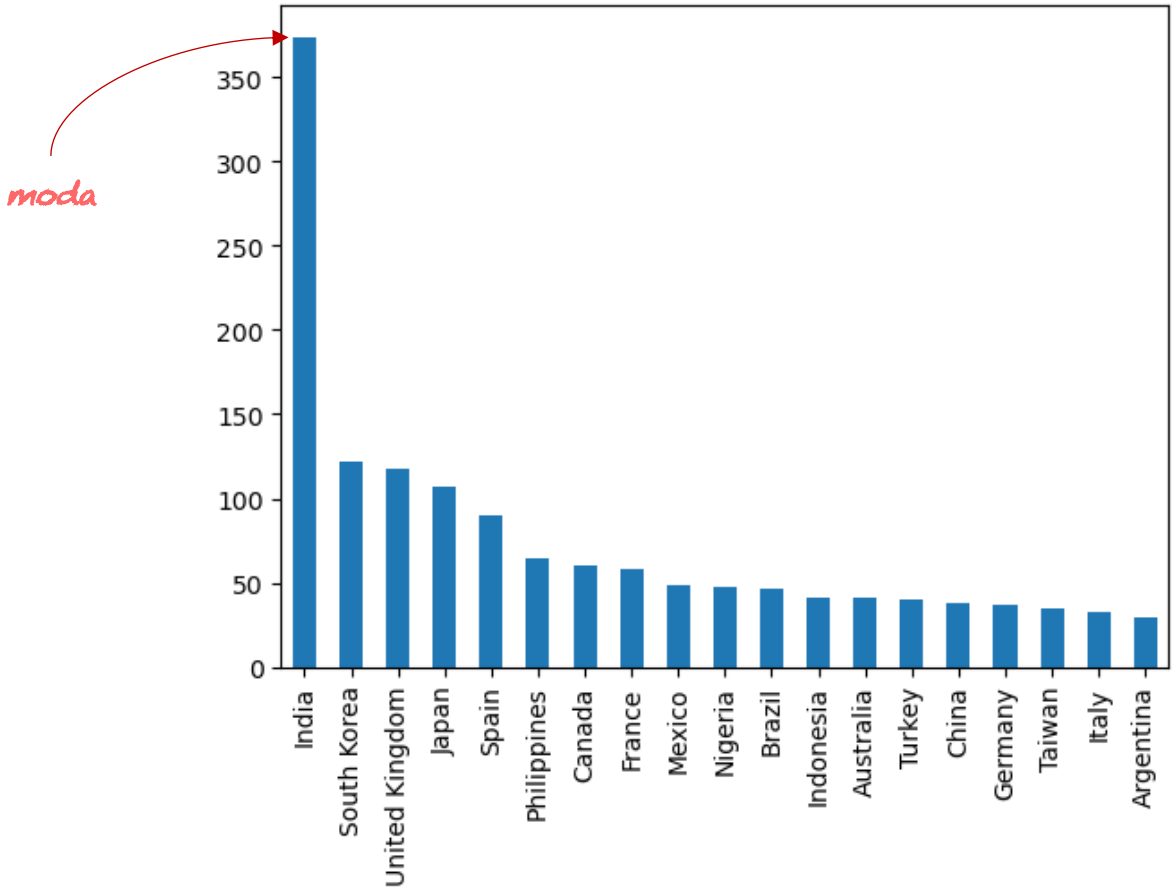
- Si riferisce ad un possibile valore assunto da un attributo categorico, e rappresenta la percentuale di volte che quel valore appare nel data set
- Dato un attributo categorico x, che può assumere valori $\{v_1, \dots, v_i, \dots, v_k\}$ ed un insieme di m dati, la frequenza di ogni valore v_i è

$$frequenza(v_i) = \frac{numero\ di\ dati\ con\ valore\ dell'attributo\ v_i}{m}$$

- La moda di un attributo categorico è il valore che ha la frequenza massima

Esempio sui dati Netflix

Uno degli attributi nella tabella era **country**




```
rank = netflix_titles['country'].value_counts()
rank[1:20].plot(kind='bar')
```

Statistiche

Percentili

- Dato un attributo ordinale o continuo x, ed un valore p compreso tra 0 e 100, il p-esimo percentile x_p è un valore di x tale che il $p\%$ dei valori osservati di x sia più basso di x_p

Statistiche

Media

Se consideriamo m dati che contengono un attributo x definiamo la media come

$$media(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Statistiche

Mediana

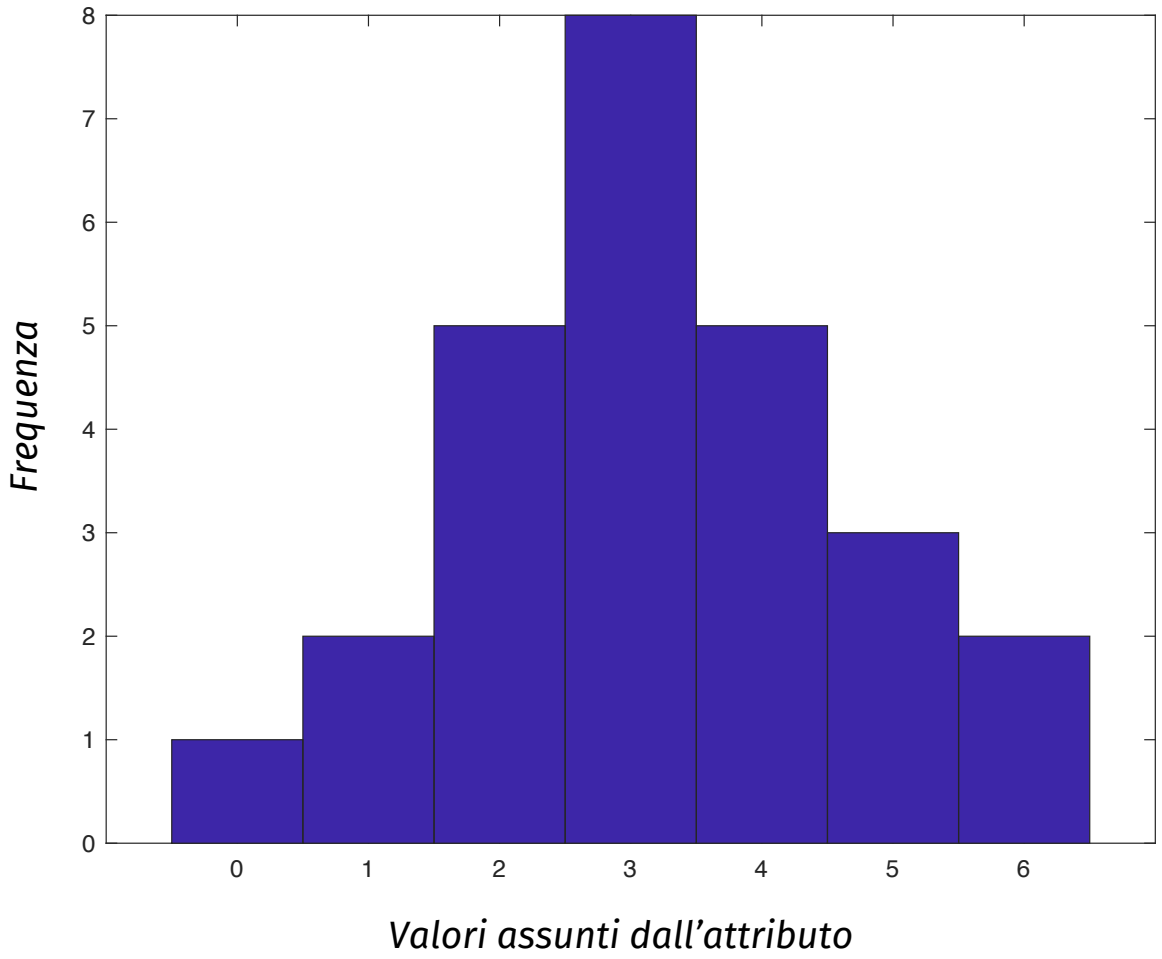
Se consideriamo m dati che contengono un attributo x definiamo $\{x_{(1)}, \dots, x_{(m)}\}$ la sequenza dei valori di x ordinata in modo crescente. La mediana è definita come

$$mediana(x) = \begin{cases} x_{(r+1)} \; se \; m \; dispari \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) \; se \; m \; pari \end{cases}$$

dove $r = \frac{m}{2}$

Differenza tra media e mediana

{0 1 1 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 5 5 5 6 6 }

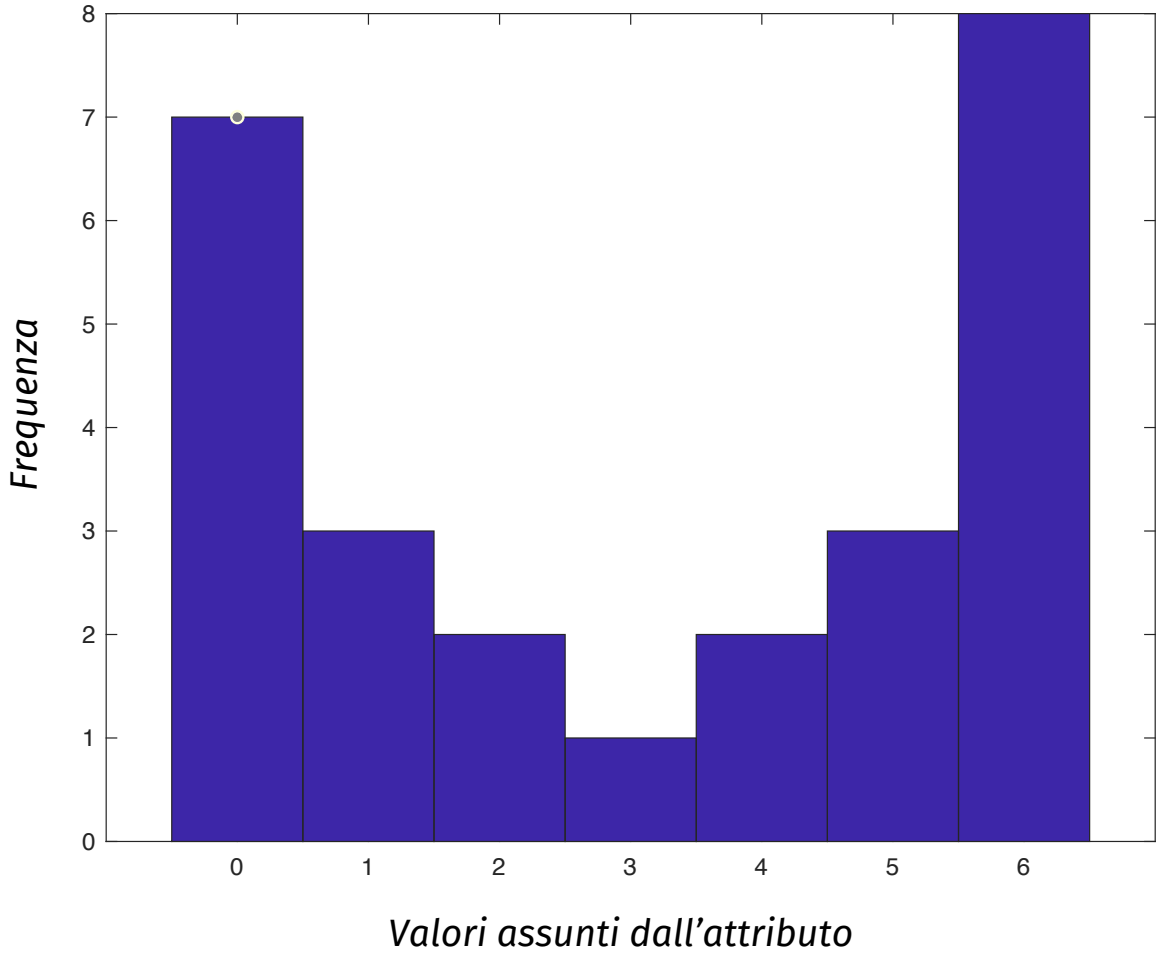


Media 3.19

Mediana 3

Differenza tra media e mediana

{0 0 0 0 0 0 0 1 1 1 2 2 3 4 4 5 5 5 6 6 6 6 6 6 6}

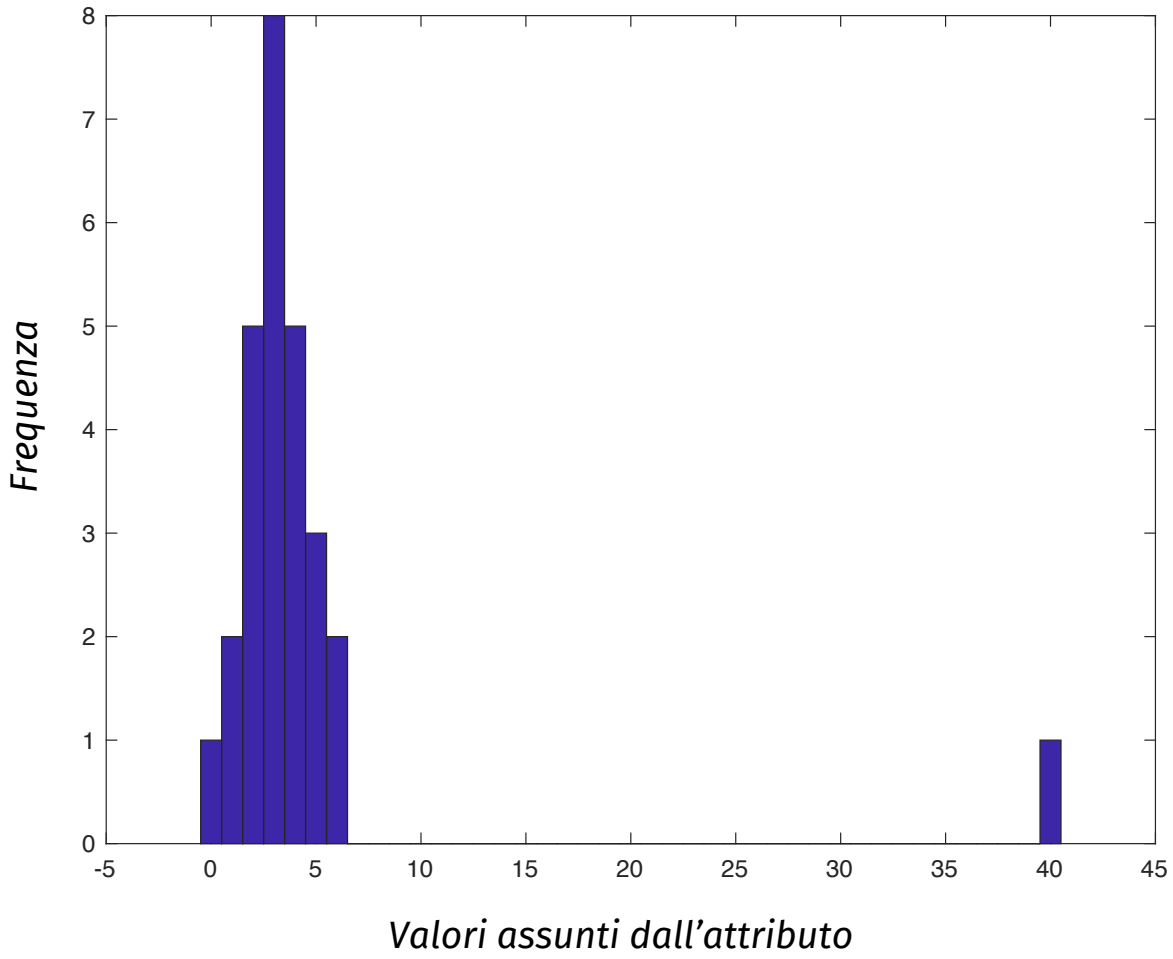


Media 3.12

Mediana 3.5

Differenza tra media e mediana

0 1 1 2 2 2 2 2 3 3 3 3 3 3 40 3 3 4 4 4 4 4 5 5 5 6 6

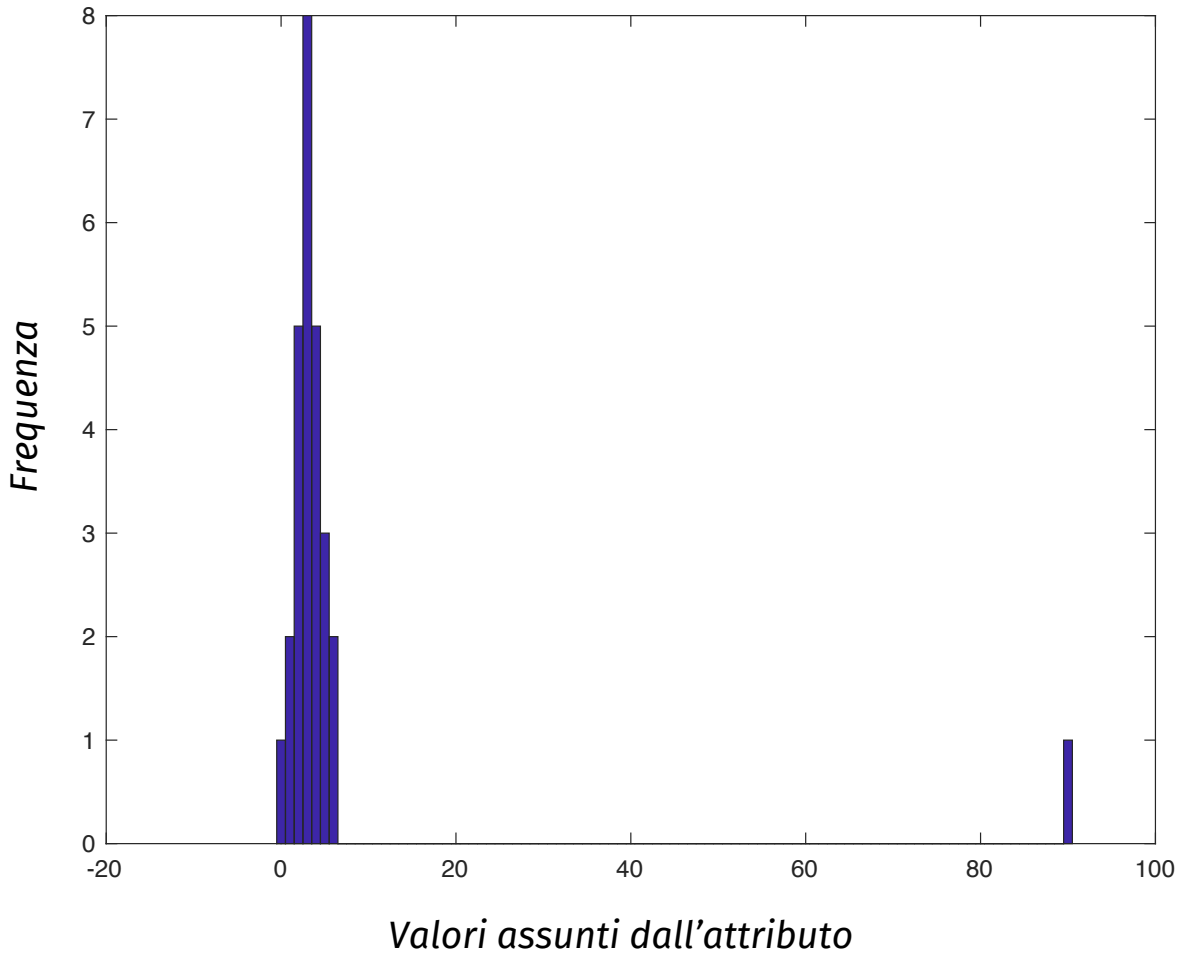


Media 4.6

Mediana 3

Differenza tra media e mediana

0 1 1 2 2 2 2 2 3 3 3 3 3 3 90 3 3 4 4 4 4 4 5 5 5 6 6

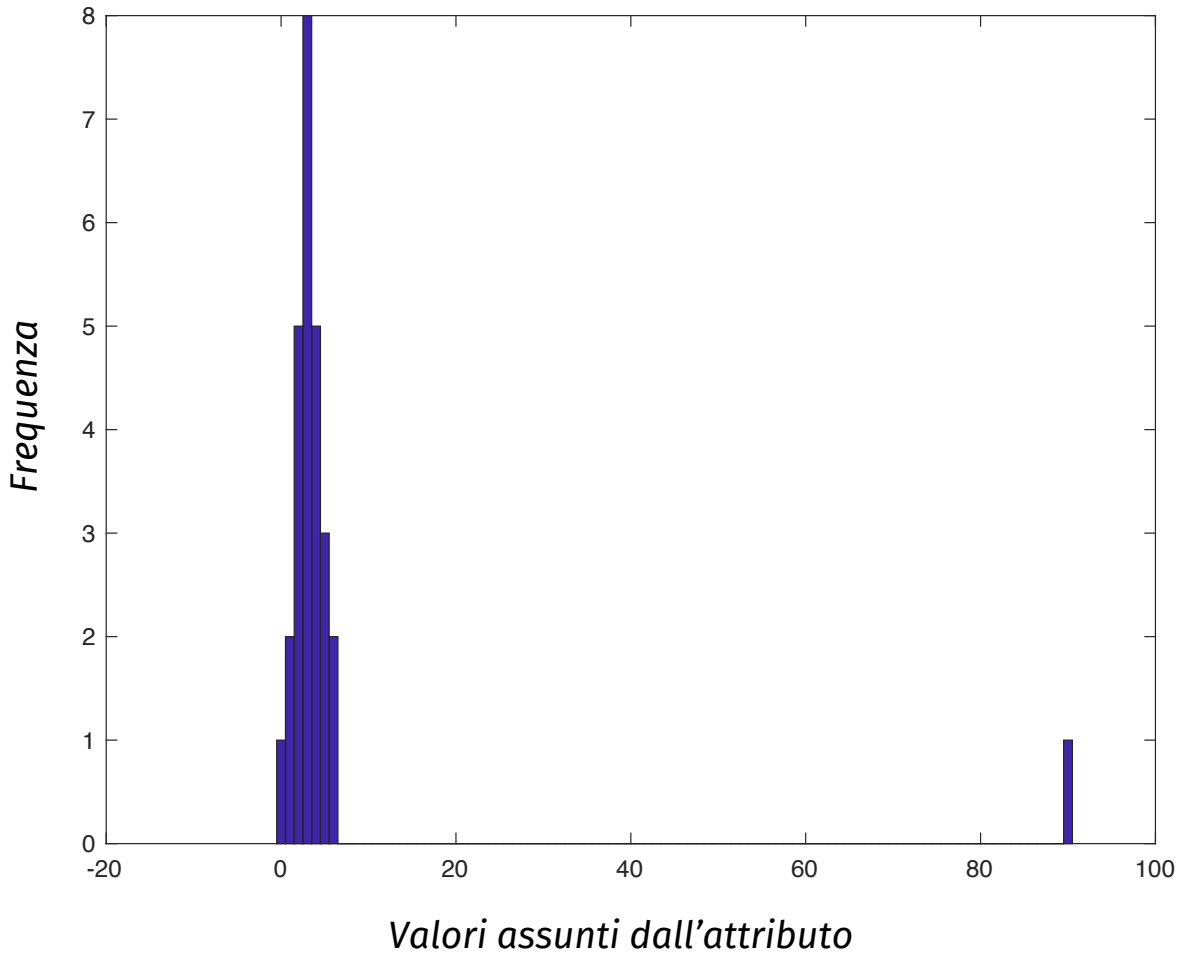


Media 6.41

Mediana 3

Differenza tra media e mediana

0 1 1 2 2 2 2 2 3 3 3 3 3 3 **90** 3 3 4 4 4 4 4 5 5 5 6 6



Media 6.41

Mediana 3

La media risulta maggiormente affetta negativamente dagli outliers

Statistiche

Indici di dispersione

- Ci dicono se i valori di un certo attributo sono “sparpagliati” tra il minimo ed il massimo oppure concentrate intorno ad un valore
- Semplice - Intervallo

$$range(x) = \max(x) - \min(x)$$

- Più informativa – Varianza

$$varianza(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

La deviazione standard s_x è la radice quadrata della varianza

Statistiche

Indici di dispersione

- Anche la varianza risente degli outliers, essendo basata sulla media
- Esistono altre soluzioni più robuste:

  Deviazione media assoluta (AAD) $AAD(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{x}|$

  Deviazione mediana assoluta (MAD) $AAD(x) = median(\{\ |x_1 - \bar{x}|, \ldots, |x_m - \bar{x}|)\}$

  Intervallo interquartile $InterquartileRange(x) = x_{75\%} - x_{25\%}$

Analisi multivariata

- Fino ad ora abbiamo ragionato su una unica variabile/attributo, ma in generale ne abbiamo diversi

$$x = (x_1, \dots, x_n)$$

- Possiamo calcolare separatamente la media per ogni attributo

$$x = (\overline{x_1}, \dots, \overline{x_n})$$

così come la sua varianza, ma è più interessante valutare la covarianza

Covarianza

- La matrice di covarianza S contiene in ogni suo elemento di posizione (i, j) la covarianza tra l'attributo i e l'attributo j

$$cov(x_i, x_j) = \frac{1}{m-1} \sum_{k=1}^{m} (x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})$$

  con m dati/oggetti (cosa c'è sulla diagonale?)

- La covarianza ci dice quanto due variabili variano insieme e dipende dalla loro magnitude
- Una quantità più descrittiva è la correlazione

Correlazione

- La correlazione viene definita come

$$corr(x_i, x_j) = \frac{cov(x_i, x_j)}{s_i s_j}$$

- Sulla diagonale ci sono degli 1
- Più il valore si avvicina a 1 e più le due variabili variano insieme
- Più il valore si avvicina a -1 e più le due variabili variano in modo inverso

UniGe | MaLGa

Visualizzazione

Principi ACCENT

- **Apprehension**: Capacità di percepire correttamente le relazioni tra variabili
- **Clarity**: Capacità di distinguere visivamente tutti gli elementi di un grafico
- **Consistency**: capacità di interpretare un grafico per confronto (similarità) con grafici precedenti

Principi ACCENT

- **Efficiency**: capacità di rappresentare una relazione anche complessa in modo semplice

- **Necessity**: la necessità che si ha di usare il grafico (ci sono modi migliori per rappresentare la stessa informazione?)

- **Truthfulness**: capacità di determinare il valore rappresentato da ogni elemento del grafico osservando la sua magnitude relativamente ad una scala implicita o esplicita

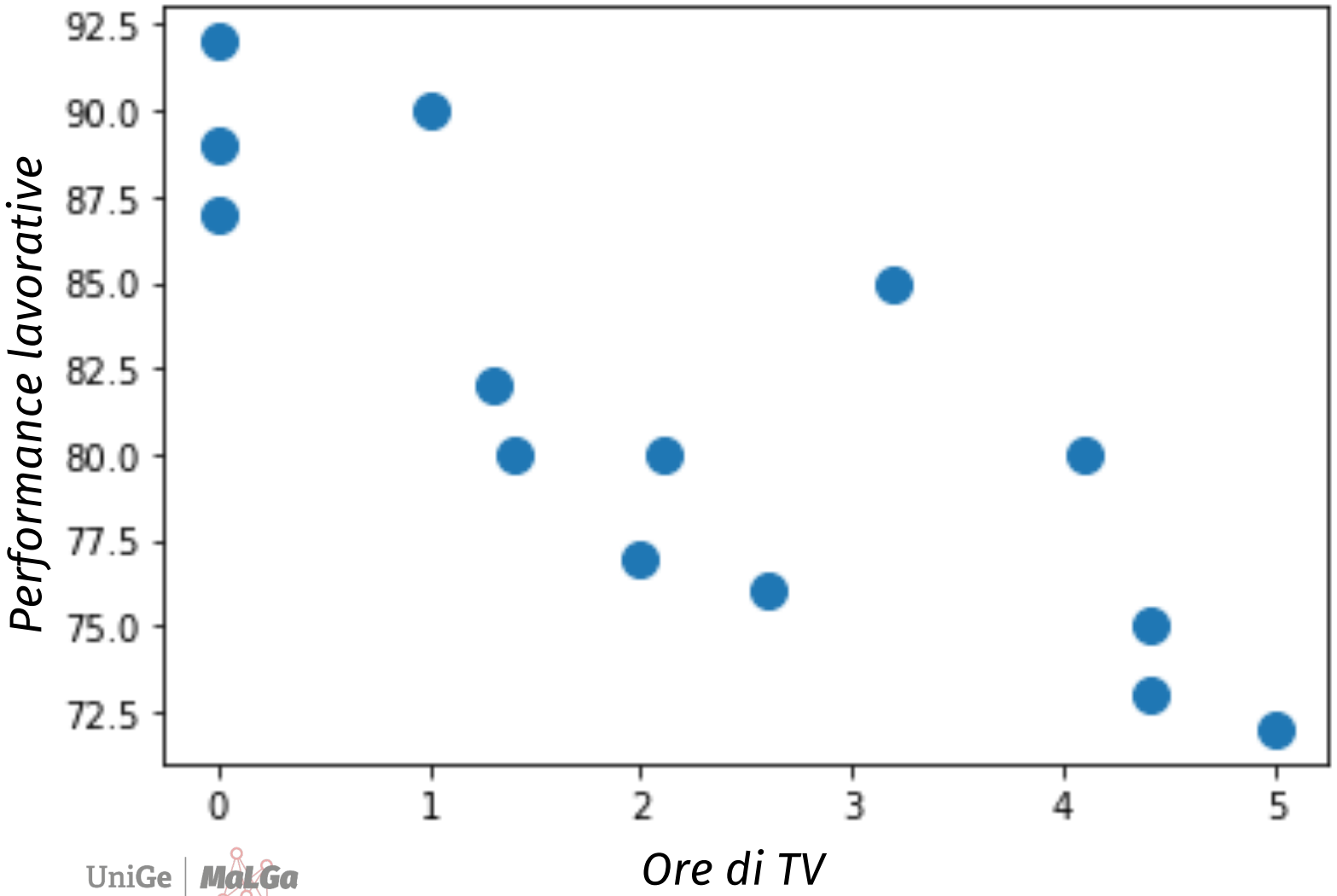
Scatter plot (grafico a dispersione)

- In un bit: una rappresentazione dei punti su un piano cartesiano
- Obiettivo: evidenziare visivamente relazioni esistenti tra variabili e se possibile rivelarne una correlazione
- Viene usato per variabili quantitative

ESEMPIO

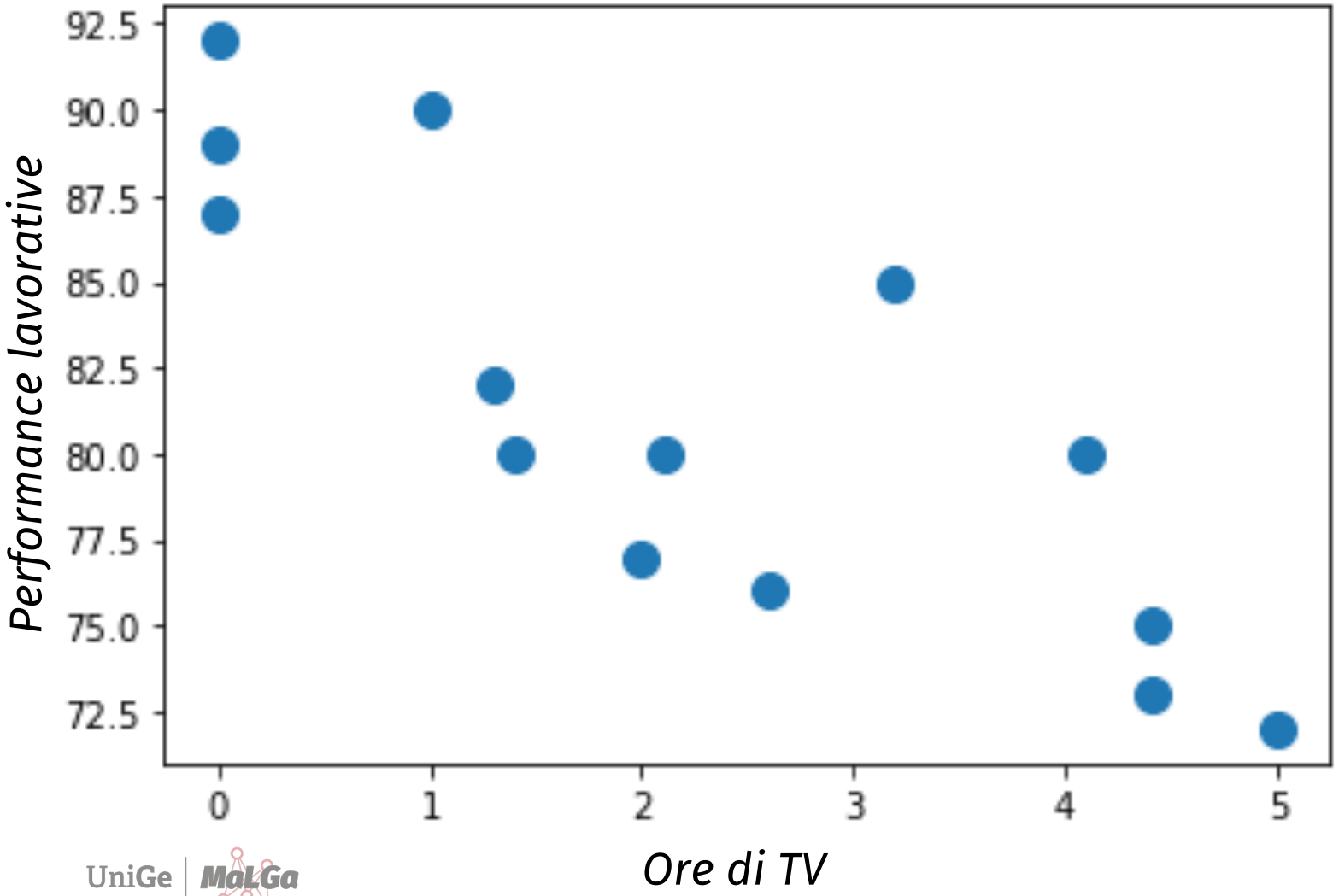
Osserviamo la relazione tra quantità di ore passate davanti alla TV e performance lavorative...

Un esempio


92.5
90.0
87.5
85.0
82.5
80.0
77.5
75.0
72.5
0
1
2
3
4
5
Performance lavorative
Ore di TV

Un esempio

Il grafico sembra suggerire che più tempo passiamo al giorno davanti alla TV più le nostre prestazioni lavorative risultano pregiudicate



Significa che esiste una relazione di causa-effetto tra queste due variabili?

NON E' DETTO... correlation is not causation

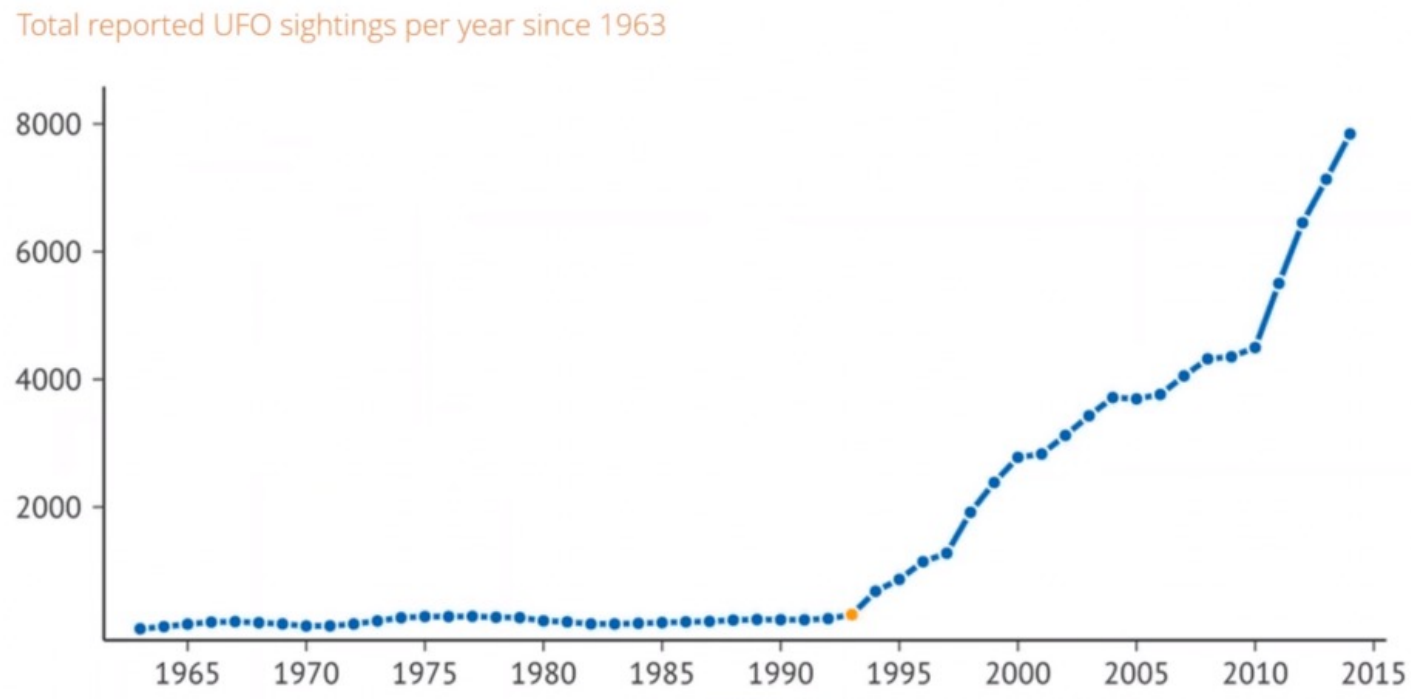
Line plot (grafico «a linea»)

- Vengono di solito usati per mostrare le variazioni nelle variabili con il trascorrere del tempo
- Viene usato per variabili quantitative

ESEMPIO

Analizziamo la quantità di avvistamenti di UFO negli anni, dal 1963

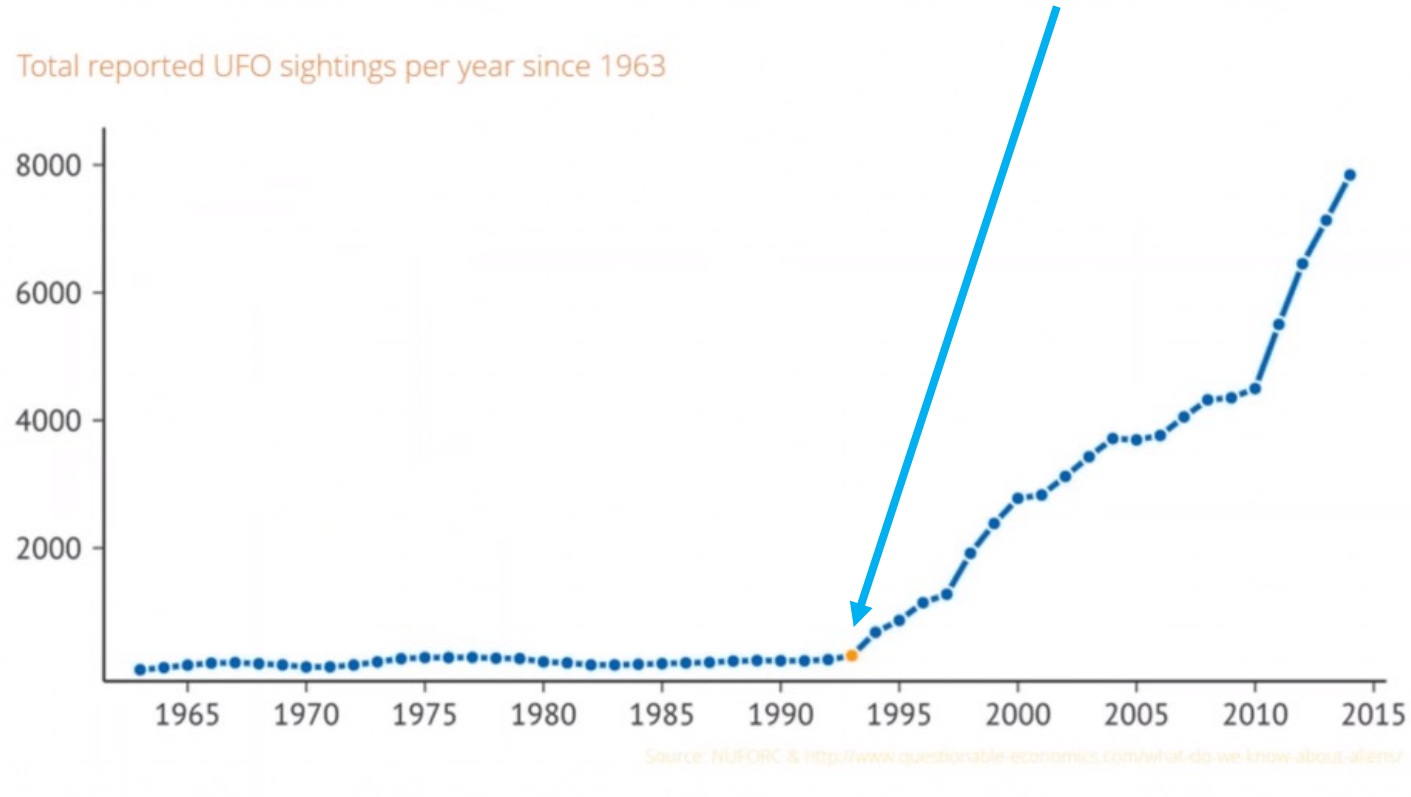
Analizzare le variabili nel tempo



Da https://rapidminer.com/blog/3-ways-ruin-business-data-science/

Analizzare le variabili nel tempo

Cosa è successo nel 1993?



Da https://rapidminer.com/blog/3-ways-ruin-business-data-science/

Diagrammi a barre

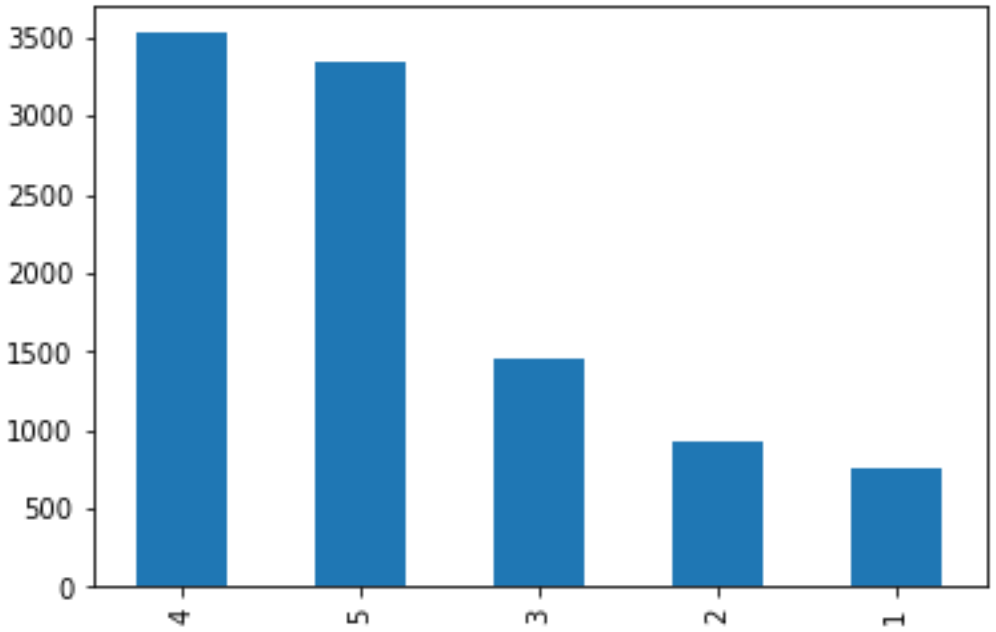
- Si utilizzano quando dobbiamo confrontare le variabili di vari gruppi
- In genere sull'asse x troviamo una variabile categorica mentre sull'asse y una variabile quantitativa o un conteggio

ESEMPIO: ricordate i dati delle recensioni da Yelp?

	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny
0	9yKzy9PApeiPPOUJEtnvkg	2011-01-26	fWKvX83p0-ka4JS3dc6E5A	5	My wife took me here on my birthday for breakf...	review	rLtl8ZkDX5vH5nAx9C3q5Q	2	5	0
1	ZRJwVLyzEJq1VAihDhYiow	2011-07-27	IjZ33sJrzXqU-0X6U8NwyA	5	I have no idea why some people give bad review...	review	0a2KyEL0d3Yb1V6aivbluQ	0	0	0
2	6oRAC4uyJCsJl1X0WZpVSA	2012-06-14	IESLBzqUCLdSzSqm0eCSxQ	4	love the gyro plate. Rice is so good and I als...	review	0hT2KtfLiobPvh6cDC8JQg	0	1	0
3	_1QQZuf4zZOyFCvXc0o6Vg	2010-05-27	G-WvGaISbqqaMHlNnByodA	5	Rosie, Dakota, and I LOVE Chaparral Dog Park!!...	review	uZetl9T0NcROGOyFfughhg	1	2	0
4	6ozycU1RpktNG2-1BroVtw	2012-01-05	1uJFq2r5QfJG_6ExMRCaGw	5	General Manager Scott Petello is a good egg!!!...	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0

Diagrammi a barre

Esempi

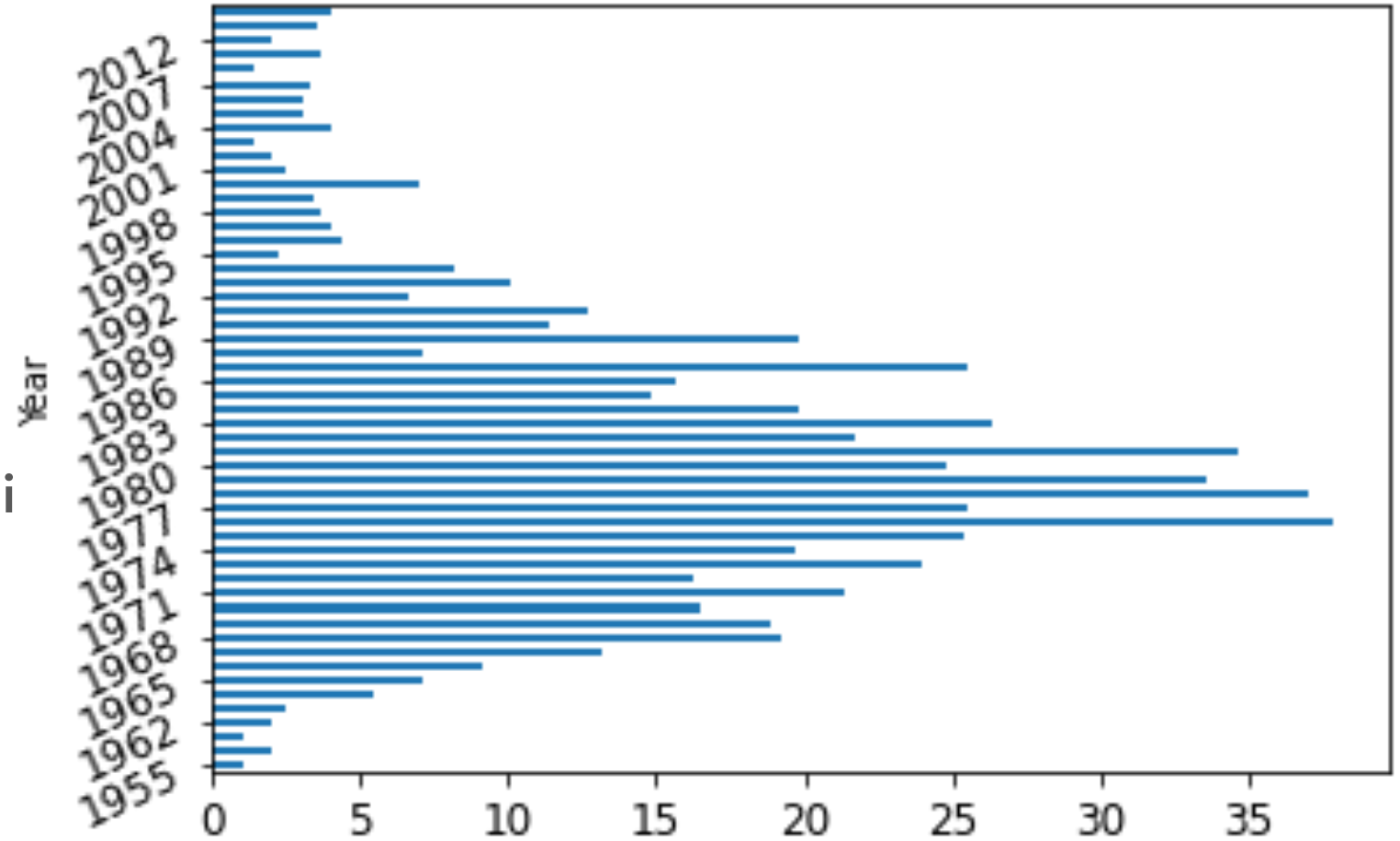


Abbiamo usato un grafico a barre per rappresentare graficamente il numero di review con 5, 4, 3, 2, o 1 stella

Un altro esempio

- Consideriamo un insieme di dati che raccoglie informazioni circa il numero di volte che una canzone è stata passata alla radio, incluso l'anno di pubblicazione

- Possiamo rappresentare il numero medio di volte che una canzone è stata passata in radio per anno di pubblicazione



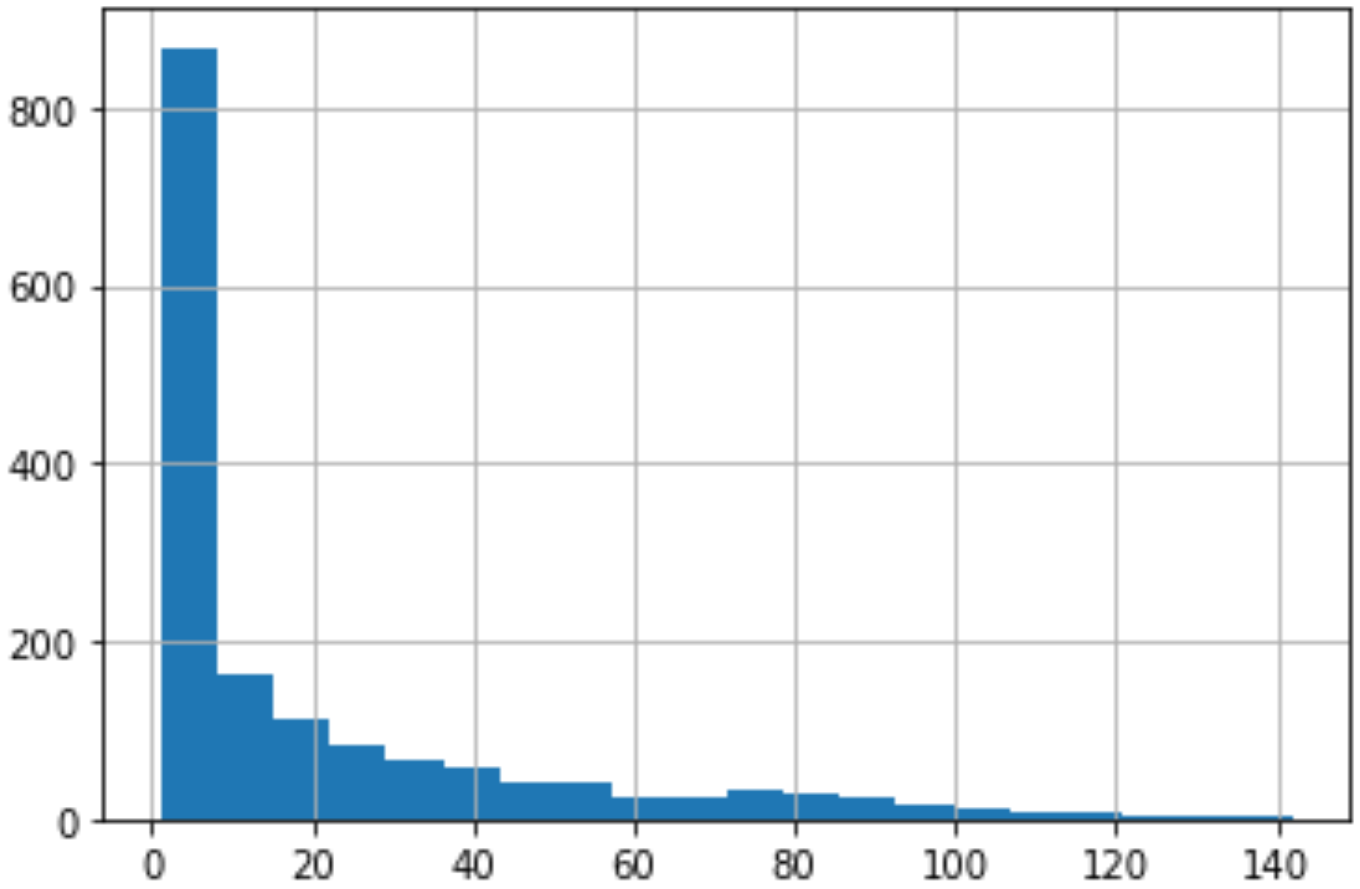
Istogrammi

- Servono per rappresentare e visualizzare la distribuzione di frequenza di una variabile quantitativa, raggruppando i dati in intervalli (bin) equidistanti
- Possono anche essere bidimensionali

- ESEMPIO: Possiamo costruire un istogramma che rappresenti la distribuzione di frequenza della variabile che conteggia il numero di volte che una canzone è stata trasmessa alla radio. Ne costruiamo poi un altro focalizzandoci sull'anno

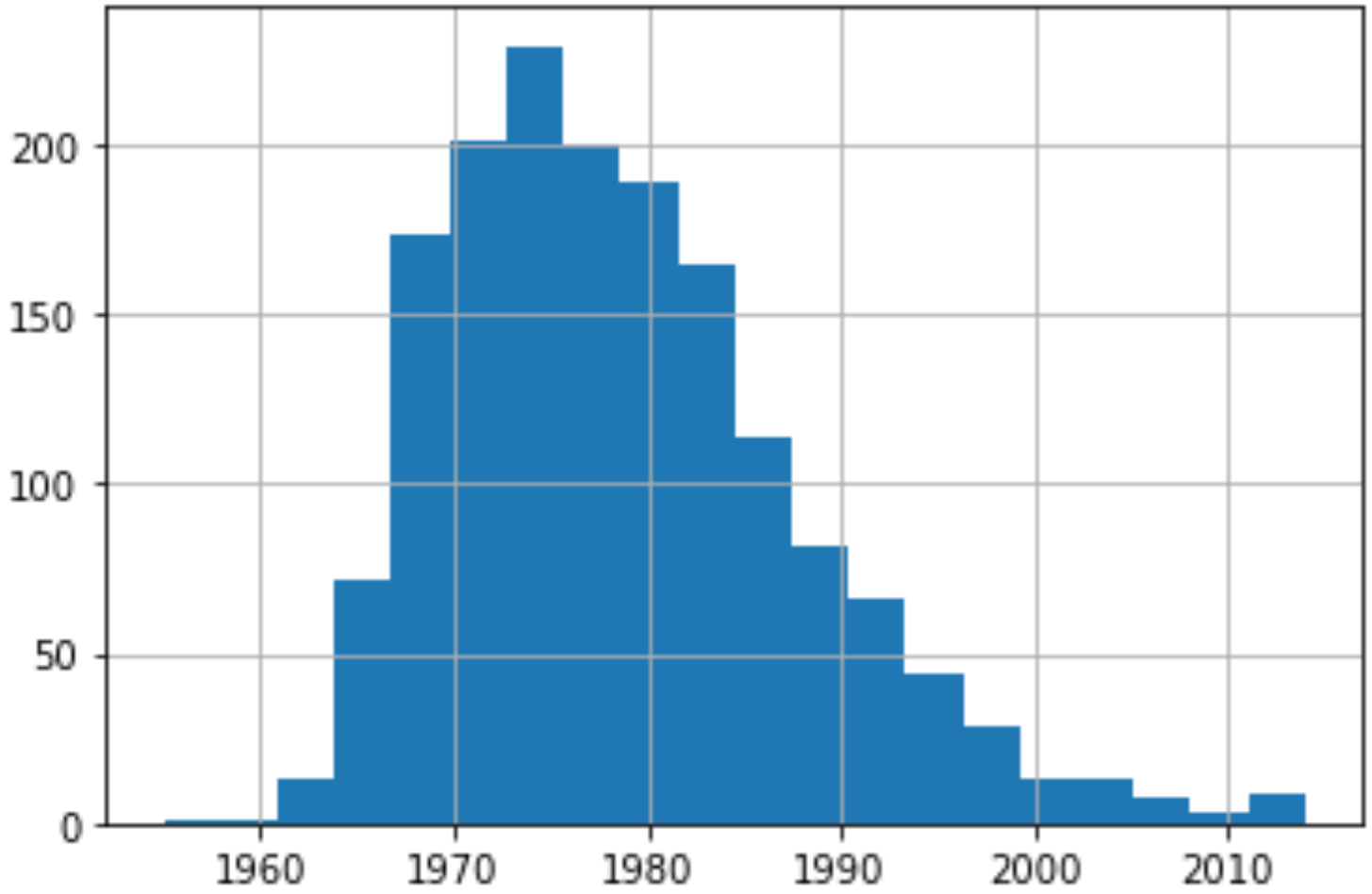
Istogrammi e canzoni

```
df1[PC'].describe()

count 1622.000000

mean 20.373613

std 27.644964

min 1.000000

25%  2.000000

50%  7.000000

75%  28.000000

max 142.000000

Name: PC, dtype: float64
```



Istogrammi e canzoni

```
df1['Year'].describe()

count    1622.000000
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
```



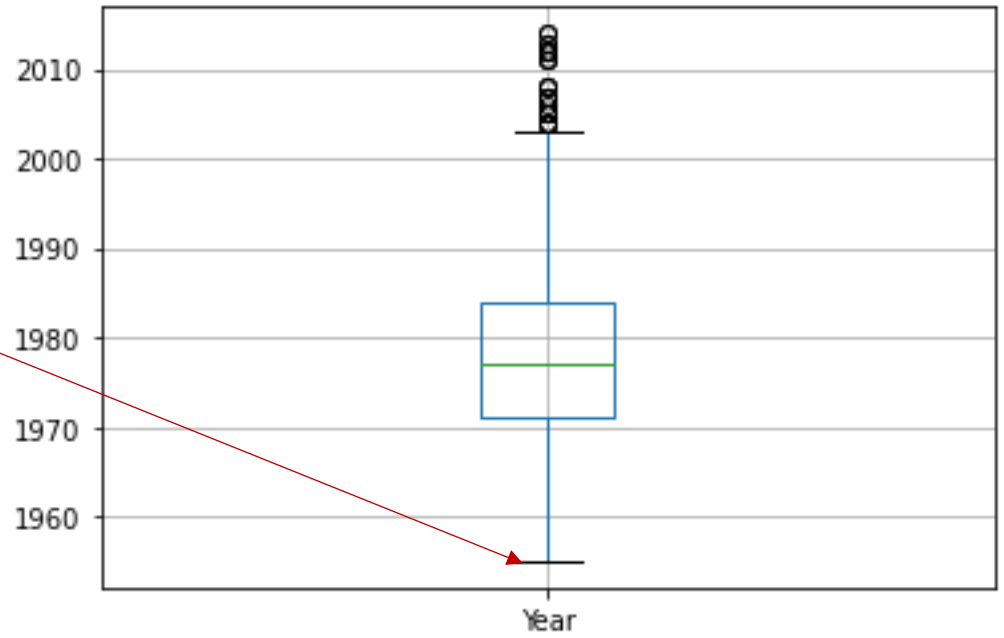
Grafici box-plot

- **Vengono utilizzati per mostrare una distribuzione di valori e mettono in evidenza cinque diverse quantità**
  - Il valore minimo
  - Il primo quartile → Valore che separa il 25% di valori più bassi da tutti gli altri → E' detto anche 25esimo percentile
  - La mediana → E' il secondo quartile
  - Il terzo quartile → Valore che separa il 25% di valori più alti da tutti gli altri → E' detto anche 75esimo percentile
  - Il valore massimo

Esempio: ancora canzoni

- Vogliamo rappresentare con box-plot la distribuzione del numero canzoni proposte per anno (ovviamente rispetto al nostro campione: il numero di volte che un anno compare corrisponde al numero di canzoni che stimiamo siano state pubblicate quell'anno)

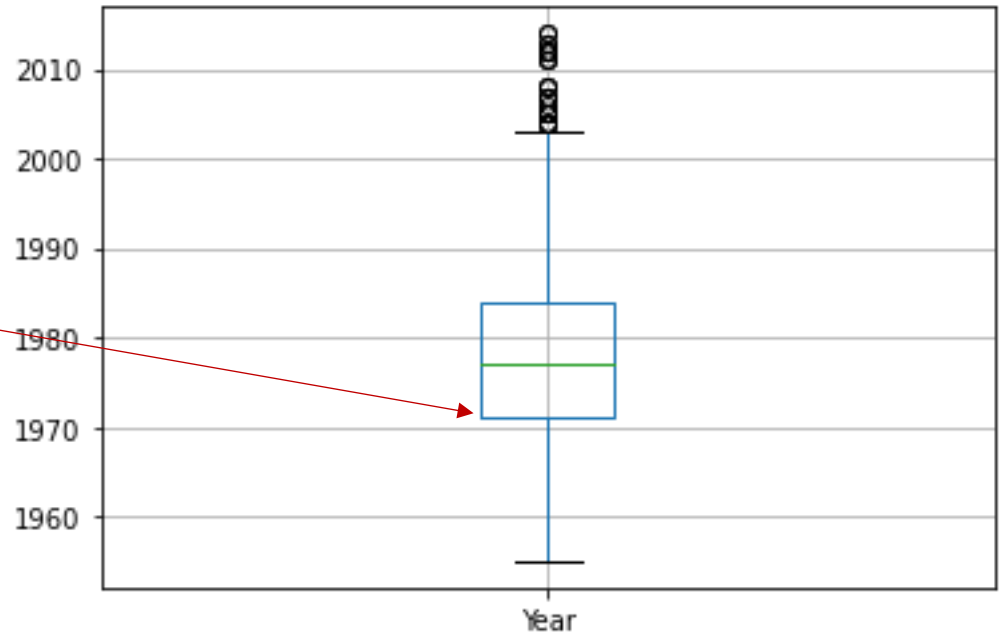
```
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
      Median  1977.0
```



Esempio: ancora canzoni

- Vogliamo rappresentare con box-plot la distribuzione del numero canzoni proposte per anno (ovviamente rispetto al nostro campione: il numero di volte che un anno compare corrisponde al numero di canzoni che stimiamo siano state pubblicate quell'anno)

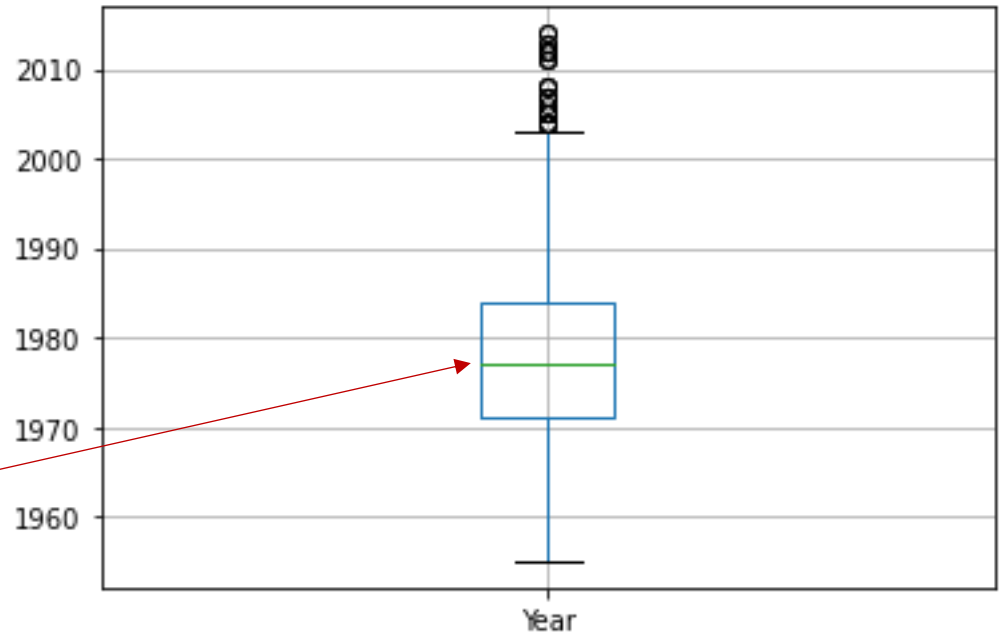
```
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
      Median  1977.0
```



Esempio: ancora canzoni

- Vogliamo rappresentare con box-plot la distribuzione del numero canzoni proposte per anno (ovviamente rispetto al nostro campione: il numero di volte che un anno compare corrisponde al numero di canzoni che stimiamo siano state pubblicate quell'anno)

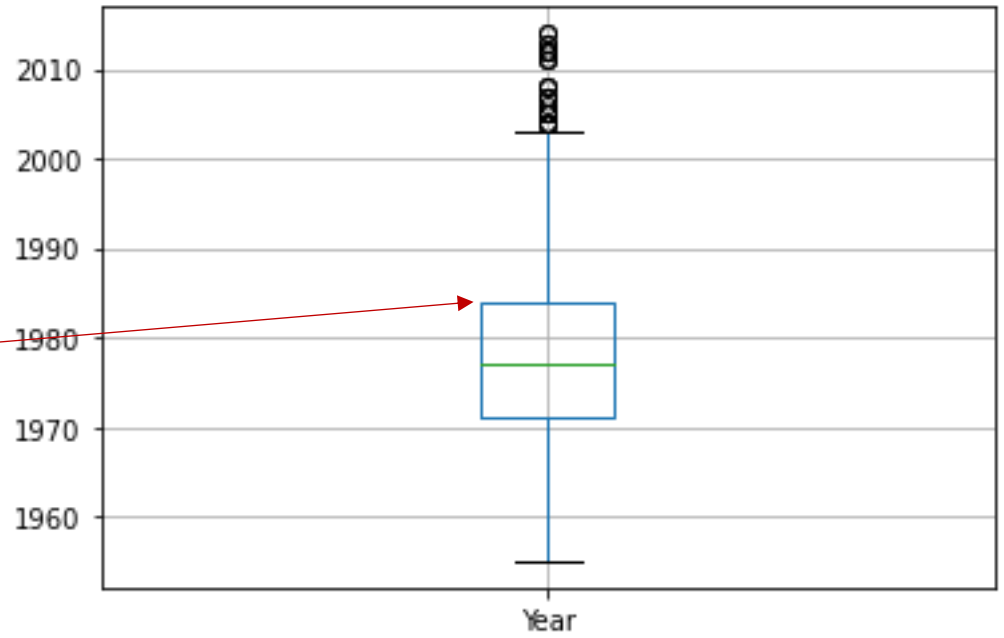
```
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
Median  1977.0
```



Esempio: ancora canzoni

- Vogliamo rappresentare con box-plot la distribuzione del numero canzoni proposte per anno (ovviamente rispetto al nostro campione: il numero di volte che un anno compare corrisponde al numero di canzoni che stimiamo siano state pubblicate quell'anno)

```
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
      Median  1977.0
```

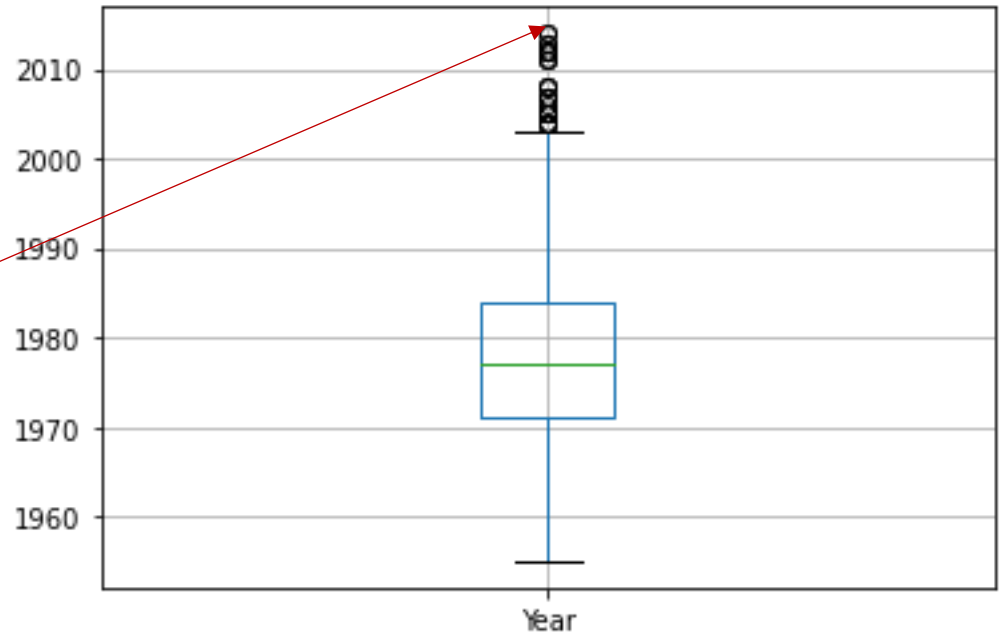


Esempio: ancora canzoni

- Vogliamo rappresentare con box-plot la distribuzione del numero canzoni proposte per anno (ovviamente rispetto al nostro campione: il numero di volte che un anno compare corrisponde al numero di canzoni che stimiamo siano state pubblicate quell'anno)

```
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
```

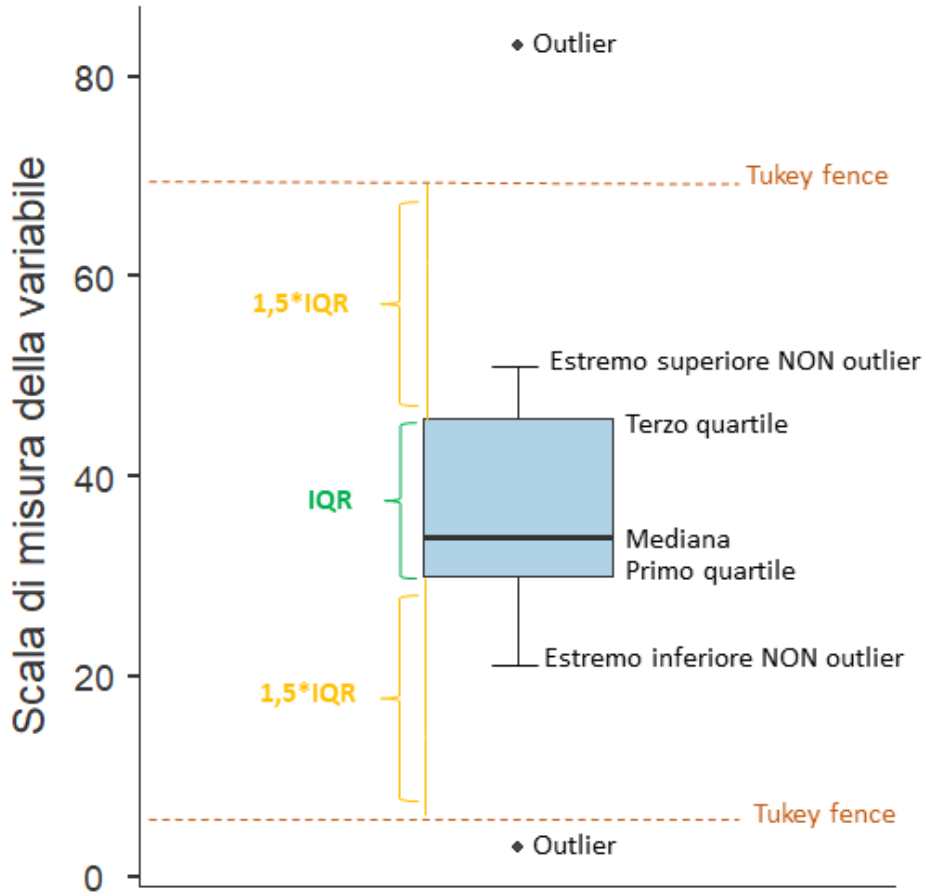
Median 1977.0



Cerchiamo di capire meglio

range interquartile (IQR):
contiene il 50% centrale delle osservazioni effettuate

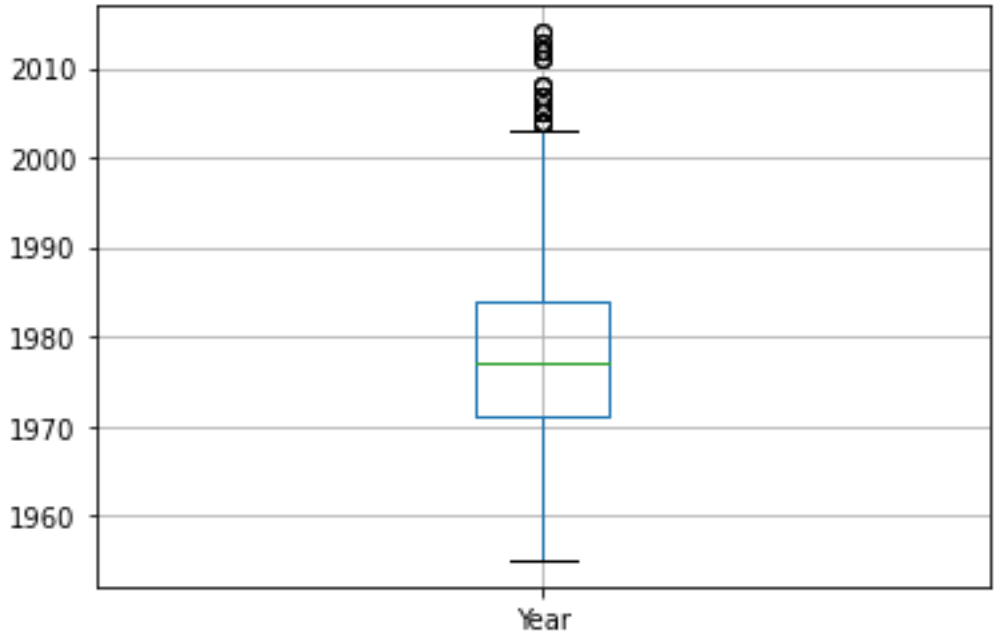
I baffi superior ed inferior sono
- Il valore Massimo/minimo
  OPPURE
- Il valore pari al 1,5 dell'IQR



Torniamo al nostro esempio

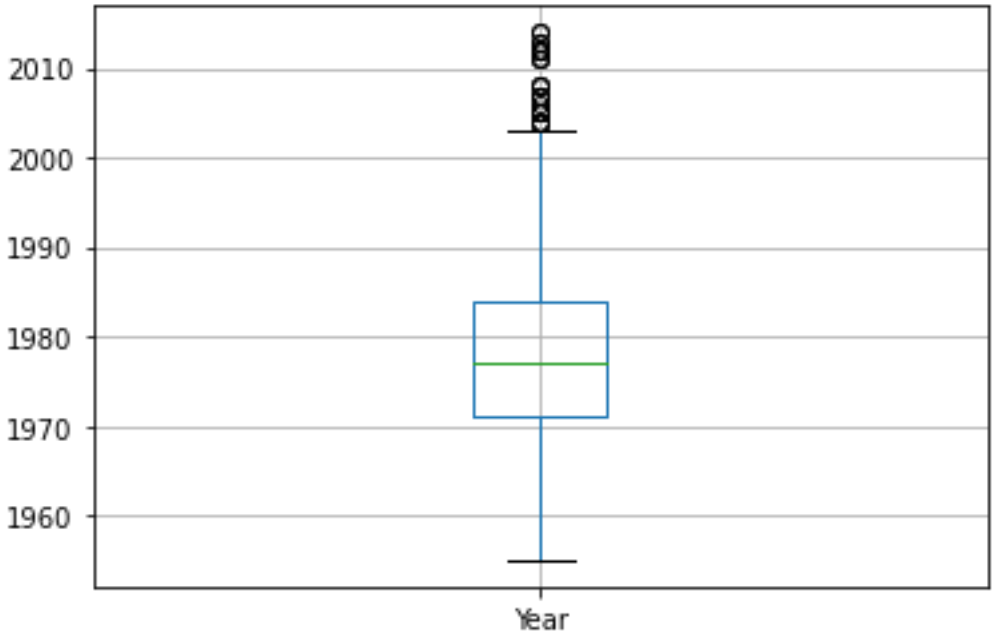
```
mean     1978.627004
std         9.345627
min      1955.000000
25%      1971.000000
50%      1977.000000
75%      1984.000000
max      2014.000000
Name: Year, dtype: float64
```

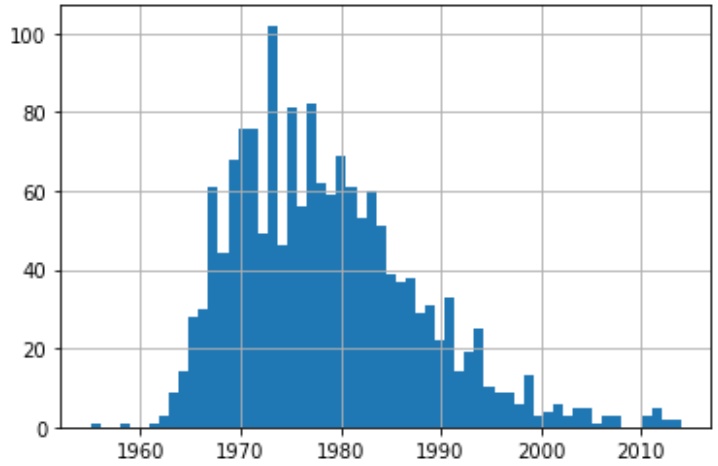
Median 1977.0

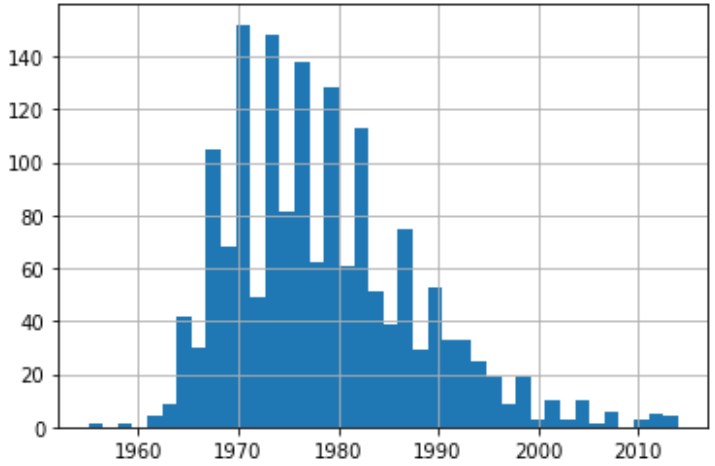


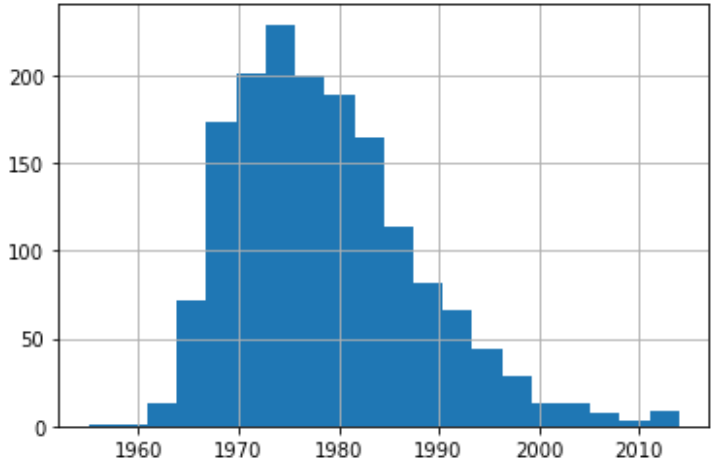
IQR = 1984-1971 = 13
13 * 1,5 = 19,5
1984 + 19,5 = 2003,5 → Minore del valore Massimo (2014) → Il baffo superiore è il terzo quartile + 1,5 * IQR
1971 – 19,5 = 1951,5 → Minore del valore minimo (1955) → Il baffo inferiore è il valore minimo

Box-plot e istogrammi



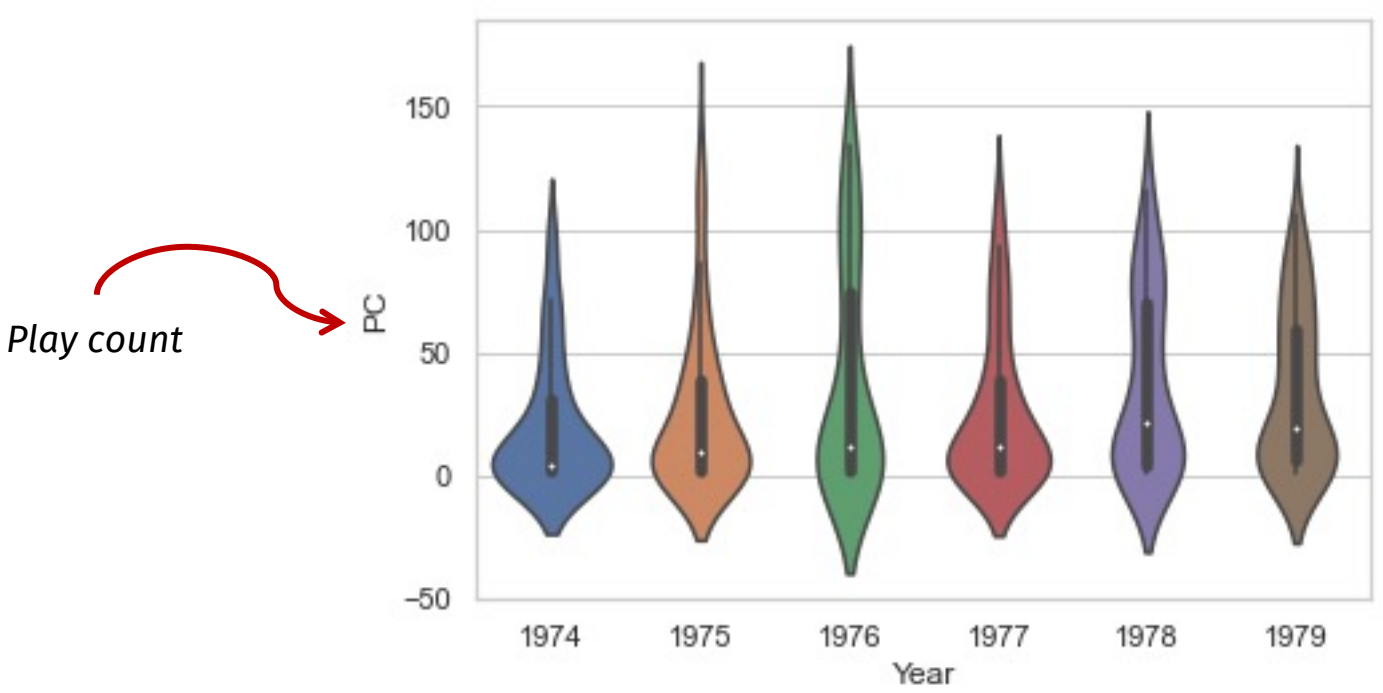






Qualche informazioni in più: violin-plot

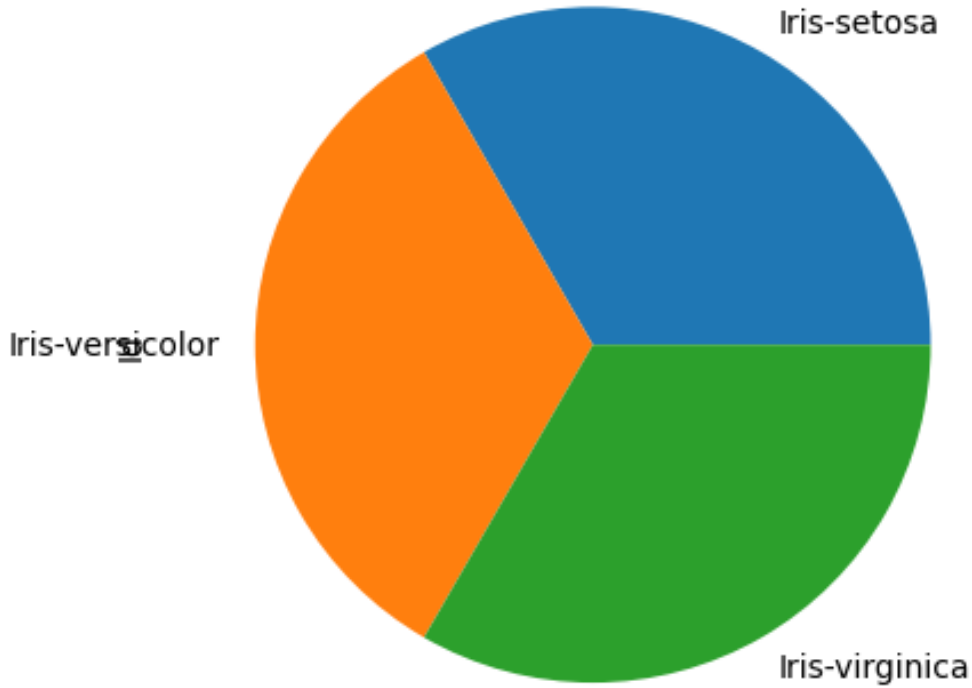
- Serve per rappresentare la distribuzione dei dati e la loro densità



Pie chart

- Possono essere usati in alternativa ai grafici a barre per variabili categoriche quando il numero di categorie è limitato

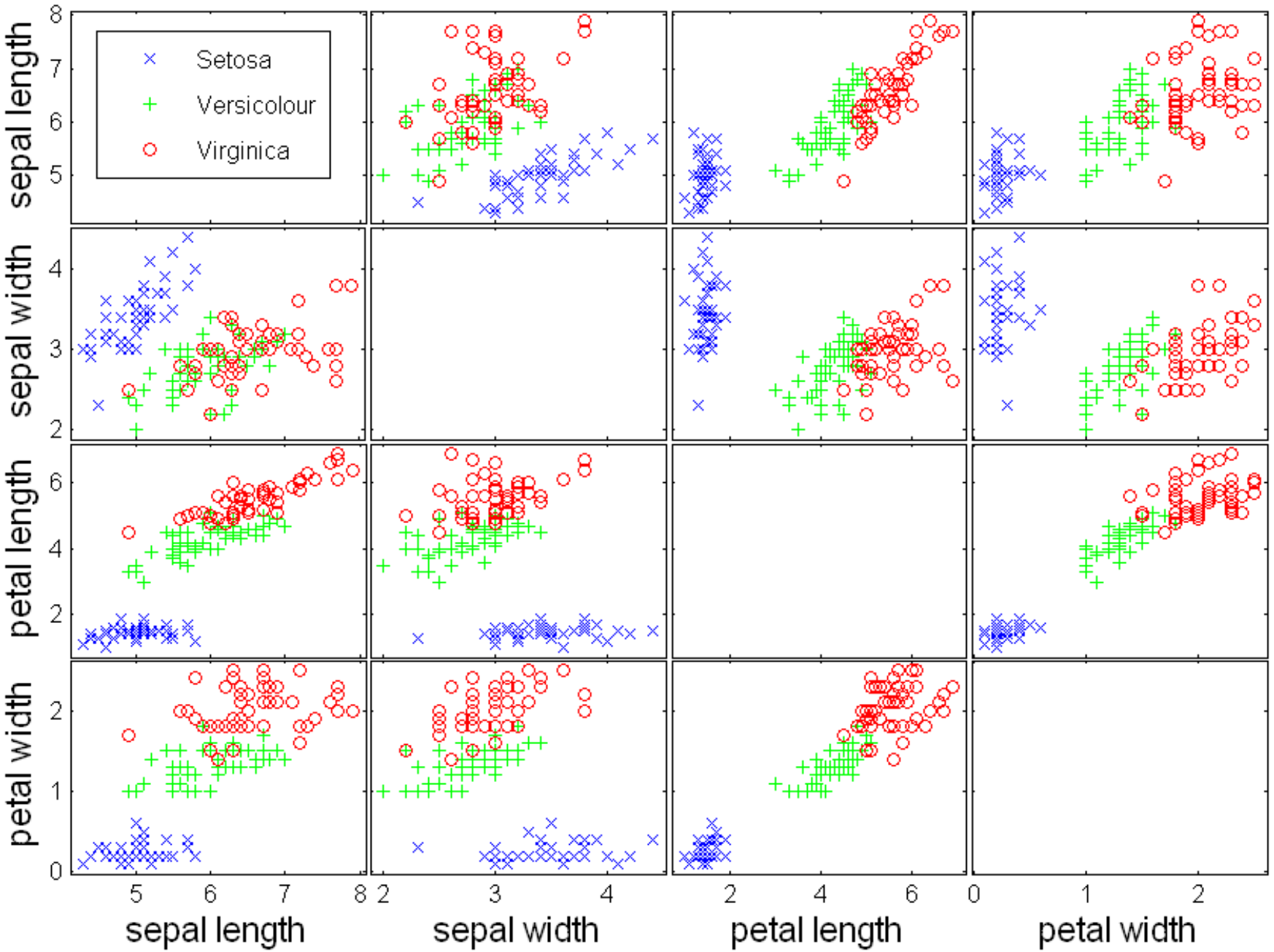
- Esempio: il dataset Iris contiene misure di petali e sepali di tre diverse categorie di Iris. Cosa si capisce immediatamente dal pie chart?



Un altro Scatter plot

Usiamo I valori degli attributi come posizioni

Di solito si tratta di grafici 2D ma possiamo aggiungere informazioni usando colore e/o forma dei marker



Matrici

Utili quando dobbiamo rappresentazione similarità/distanze tra dati

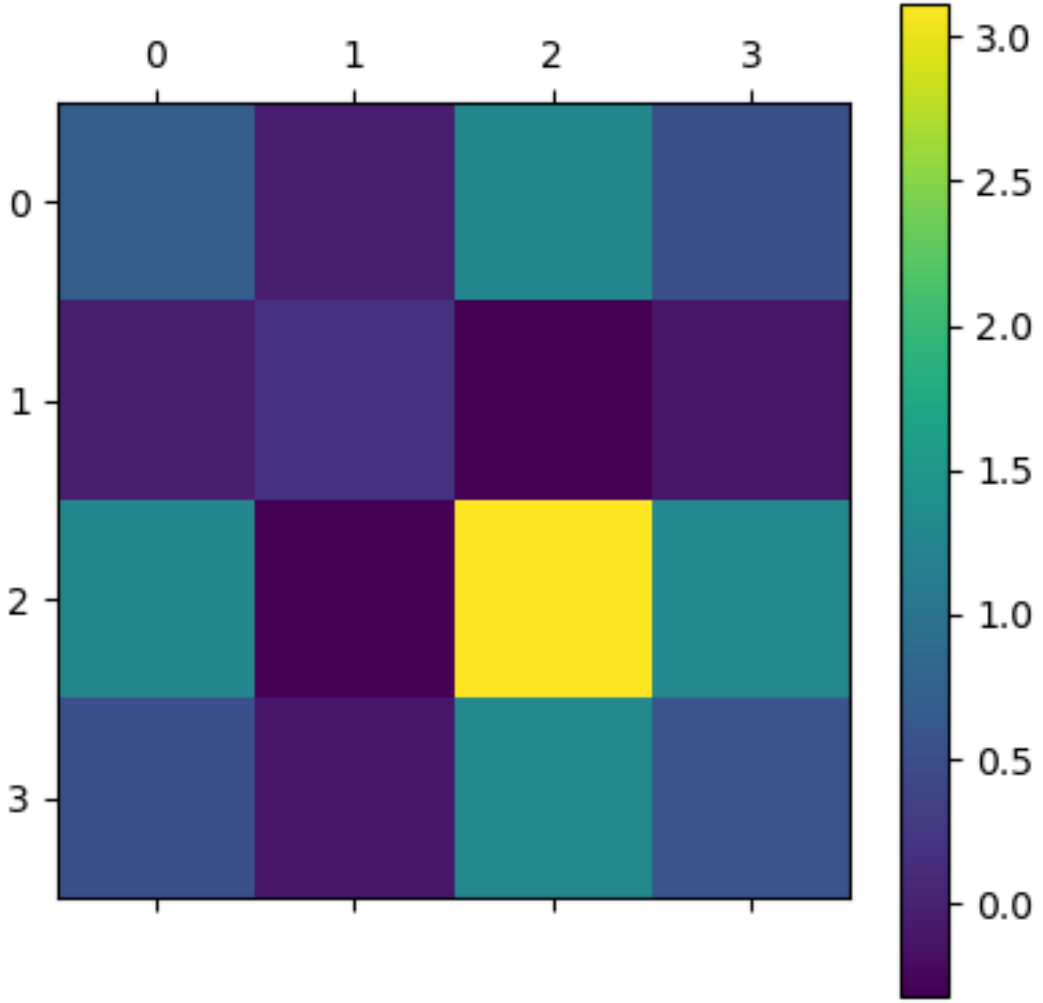
Esempio: matrice di covarianza

0: SepalLengthCm

1: SepalWidthCm

2: PetalLengthCm

3: PetalWidthCm



Matrici

Utili quando dobbiamo rappresentazione similarità/distanze tra dati

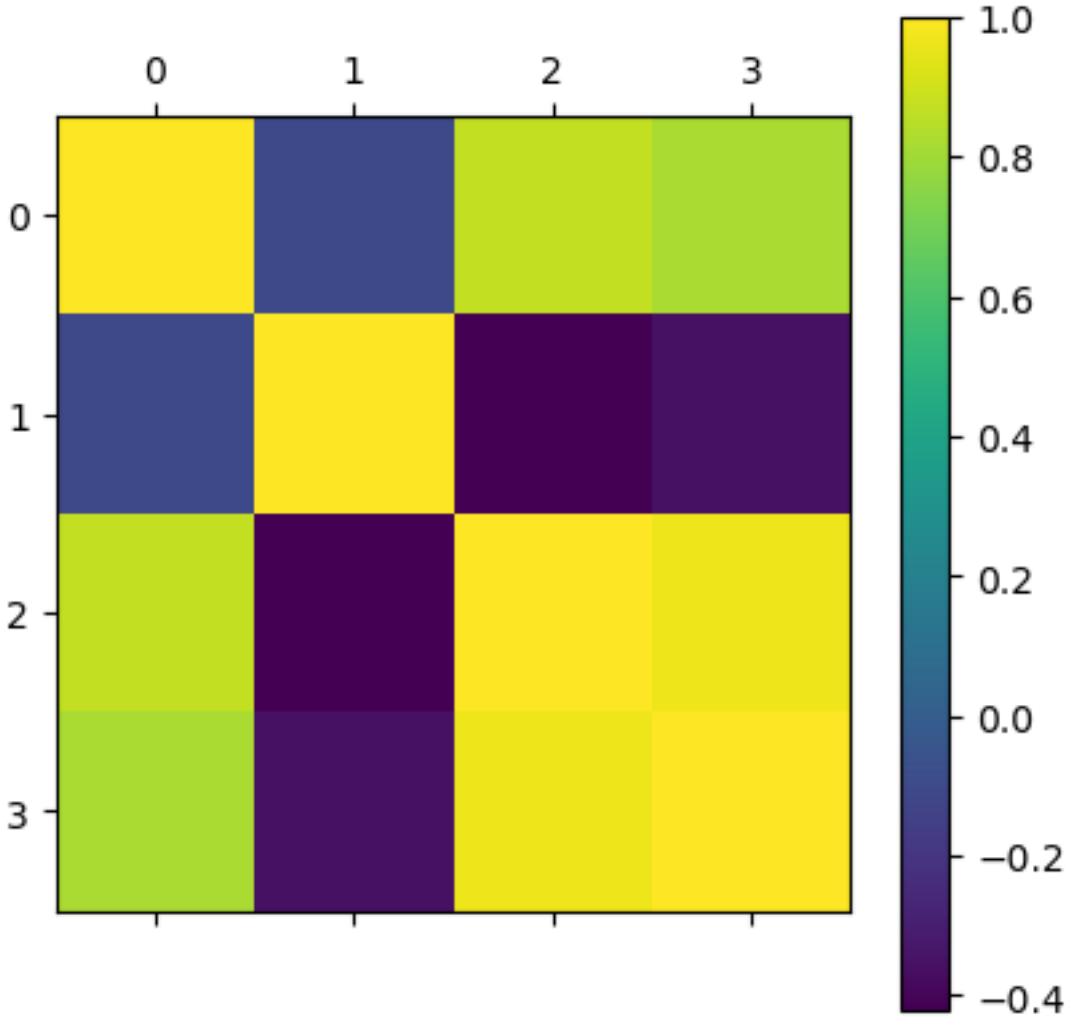
Esempio: matrice di correlazione

0: SepalLengthCm

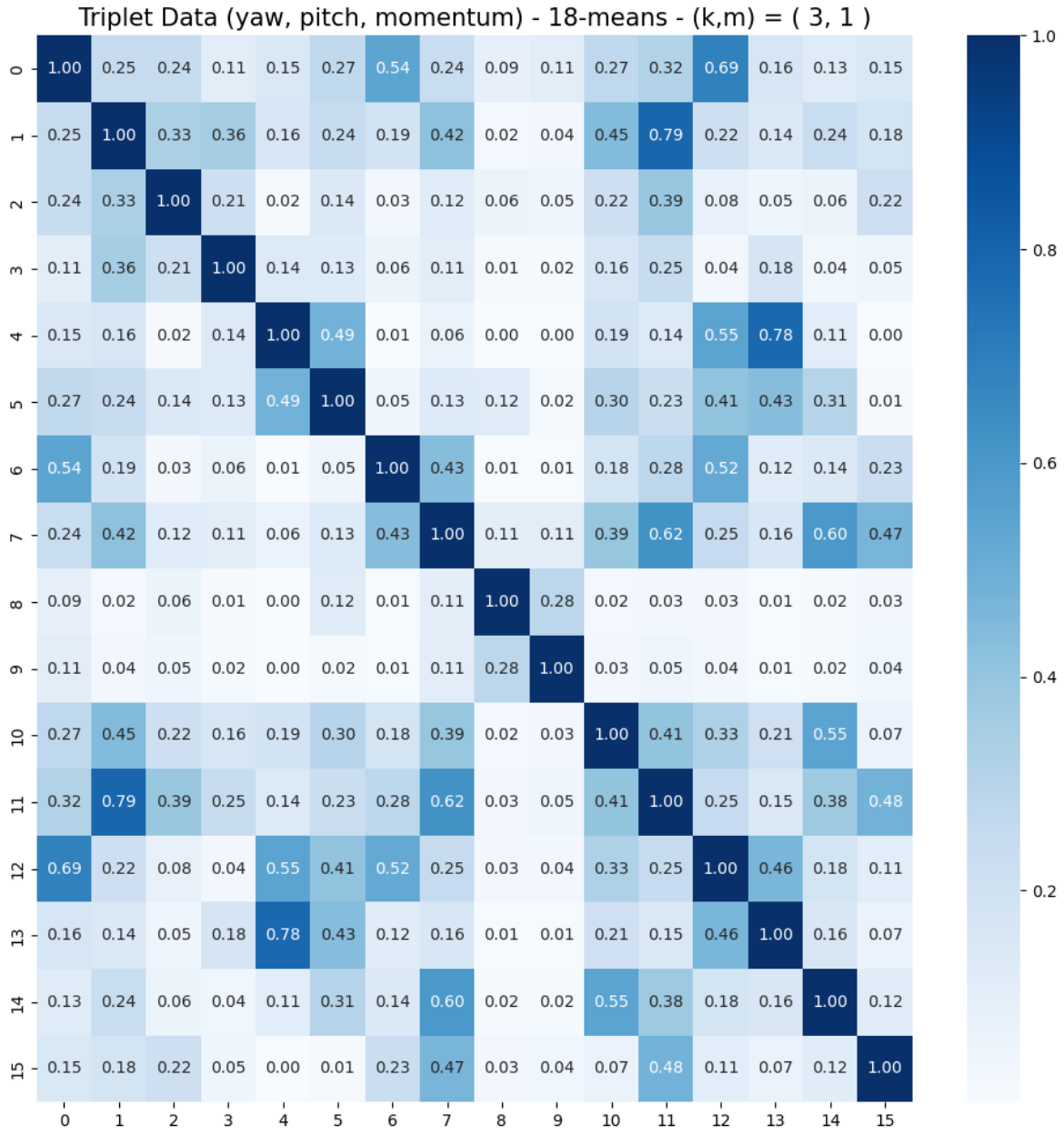
1: SepalWidthCm

2: PetalLengthCm

3: PetalWidthCm

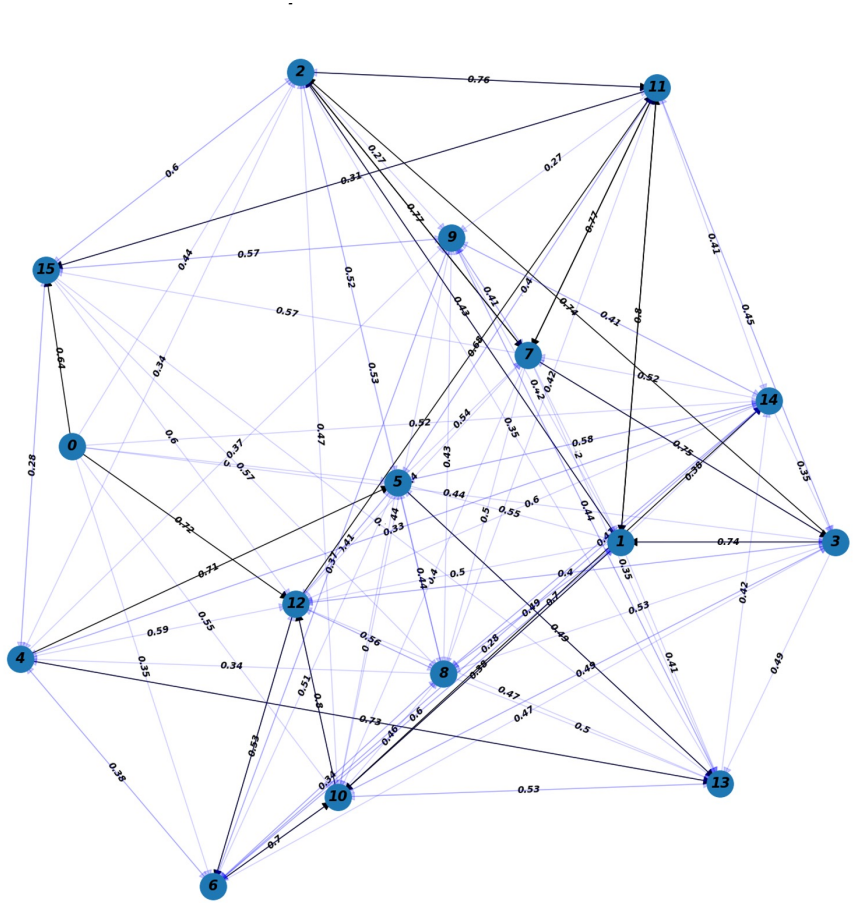


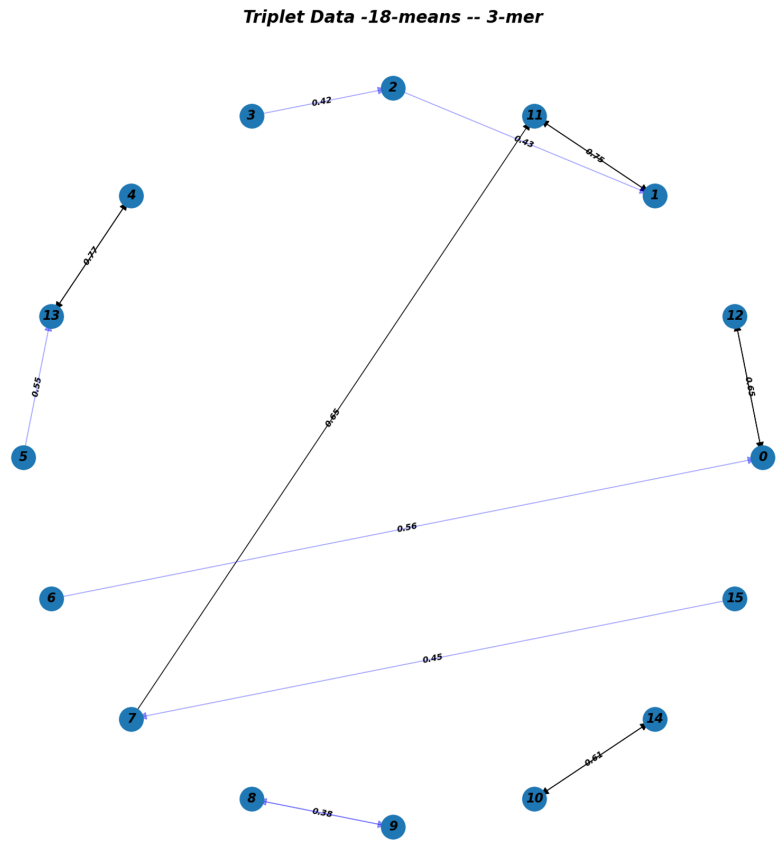
Matrici di similarità



Matrici di similarità e grafi

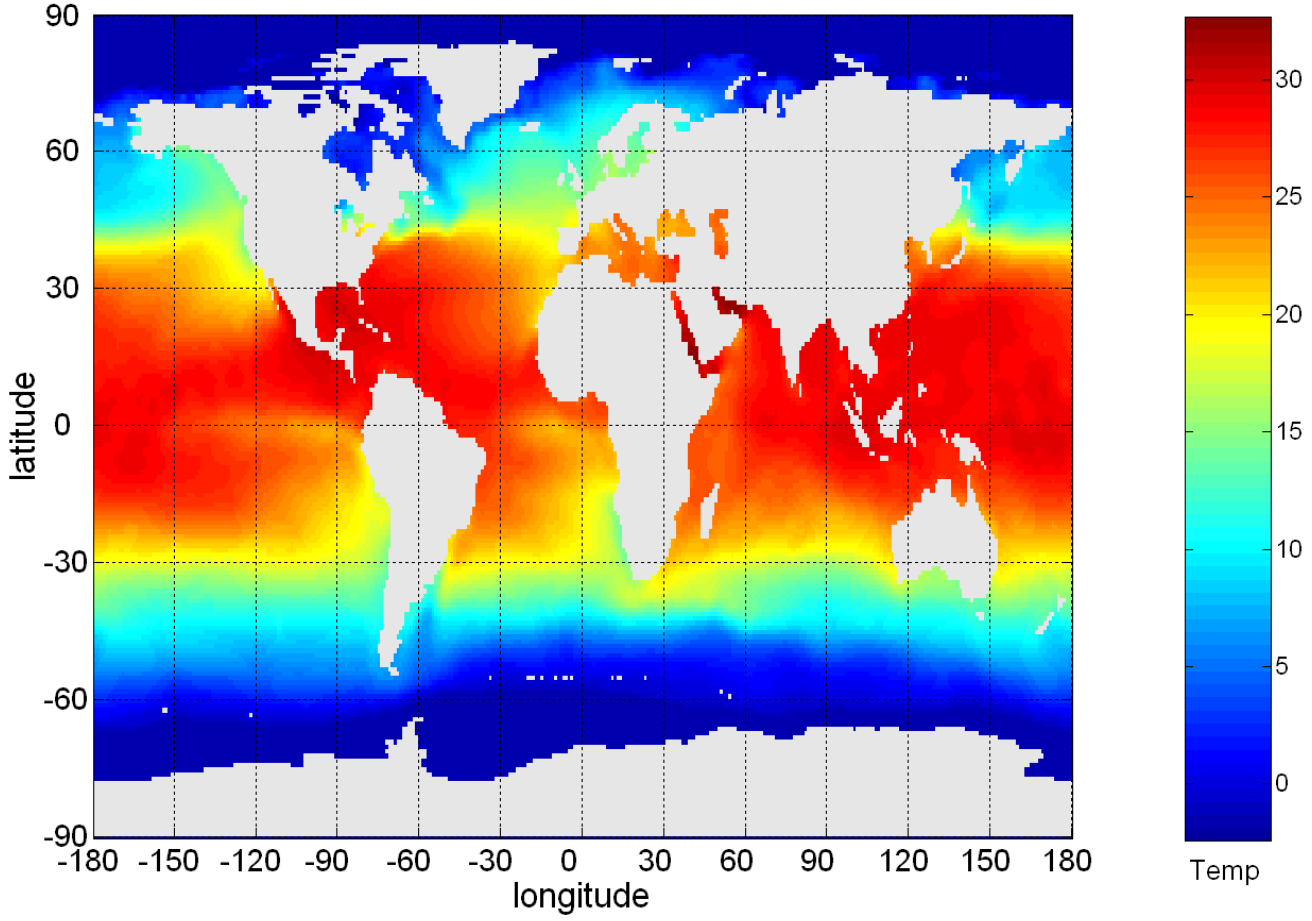
- Una modalità di visualizzazione che risulta essere più intuitiva ed immediata passa attraverso l'uso di grafi





Possiamo usare la libreria NetworkX di Python

Mappe di salienza/calore



Sono molto efficienti per rappresentare tanti dati numerici che siano “geograficamente” collocati o espressi rispetto ad uno stesso sistema di riferimento (es. su immagini)

Esempio: rappresentazione della temperature sulla supeficie marina a luglio 1982... migliaia di dati in una singola figura!

Ridurre la dimensionalità dei dati

Quando sospettiamo che la dimensione dei dati sia troppo alta rispetto alla loro quantità possiamo affidarci a tecniche di riduzione della dimensionalità, che hanno ulteriori effetti positivi

- Ci permettono di poter visualizzare i dati
- Ci permettono di interpretare meglio i dati

UniGe | MaLGa

OLAP

Rappresentazione multidimensionale

- I dati ci arrivano di solito in forma tabulare
- Per alcune tipologie di analisi ci conviene usare una rappresentazione alternativa e multidimensionale, su cui è più “facile” operare

Rappresentazione multidimensionale

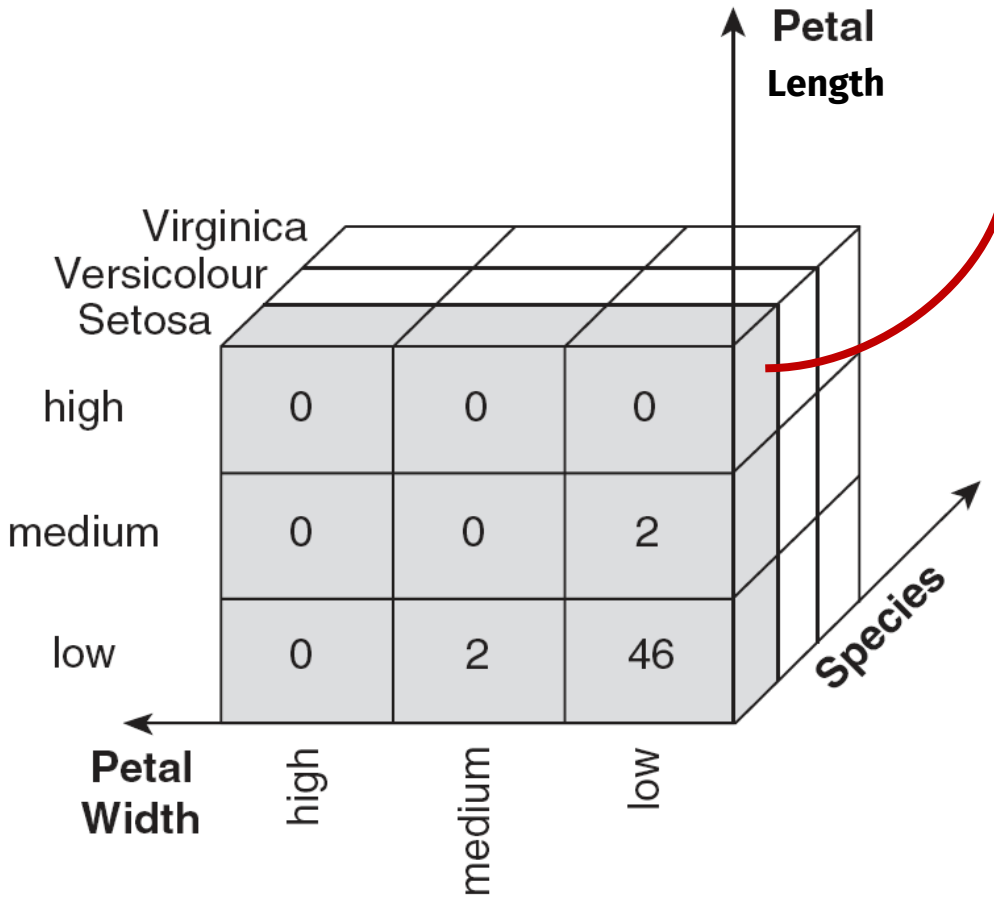
- Dobbiamo prima identificare quali attributi rappresentano le dimensioni (cioè i parametri) dell'analisi e quali le misure da analizzare
  - Gli attributi utilizzati come dimensioni hanno in genere valori discreti
  - Il valore della misure è sempre numerico
- Dobbiamo poi calcolare il valore di ogni entry dell'array multidimensionale aggregando i valori (della misura considerata) di tutti gli oggetti che hanno come valori per le dimensioni i valori corrispondenti a quell'entry

Esempio

Discretizzando i valori di petal_length e petal_width nel dataset Iris e conteggiando il numero di dati per ogni specie, otteniamo la tabella a lato

Petal Length	Petal Width	Species Type	Coun
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Esempio



Length \ Width	low	medium	high
low	46	2	0
medium	2	0	0
high	0	0	0

Length \ Width	low	medium	high
low	0	0	0
medium	0	43	3
high	0	2	2

Length \ Width	low	medium	high
low	0	0	0
medium	0	0	3
high	0	3	44

OLAP

Operazioni

- Slicing: significa selezionare un gruppo di celle dalla struttura multidimensionale selezionando uno specifico valore per una dimensione

- Dicing: significa selezionare un sottoinsieme di celle specificando una combinazione di condizioni per le diverse dimensioni

OLAP

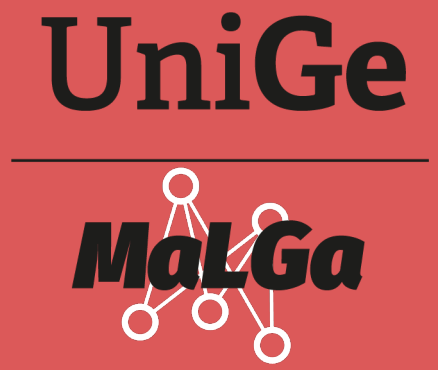
Roll-up e Drill-down

- I valori degli attributi hanno a volte una struttura gerarchica
  - Es. Data→ anno, mese, settimana,…
  - Es. Luogo → continente, stato, città,…
- I livelli di una gerarchia sono collegati da una associazione uno a molti
  - Un anno corrisponde a tanti mesi, ciascun mese corrisponde ad alcune settimane, ogni settimana a 7 giorni
  - Un continente corrisponde a più stati, ciascuno dei quali corrisponde a molte città

OLAP

Roll-up e Drill-down

- **Tali strutture gerarchiche ci consentono di aggregare o dividere i dati**
  - Roll-up: significa aumentare il livello di aggregazione dei dati
  - Drill-down: significa ridurre il livello di aggregazione dei dati


UniGe
MaLGa