PCTO in Coding & Data Science

CD: 50/50 - Coding Diversity

Liceo Scientifico S. Cannizzaro
16/03/2022

CD:50/50
CODING DIVERSITY

Nelle ultime puntate…

1. Che cos’è un dato
2. Che cos’è la data science
3. Che cos’è un linguaggio di programmazione
4. Cosa sono le librerie
5. Cos’è un comando
6. Cos’è una variabile
7. Tipi di dati: stringhe, variabili numeriche, variabili logiche, liste, etc.
8. Cos’è una funzione
9. Comandi condizionali
10. Cos’è un database e i suoi elementi costitutivi

Modulo 1: Introduzione al coding

Modulo 2: Saper leggere e rappresentare i dati

Modulo 3: Basi di inferenza e analisi predittiva

Modulo 4: Basi di machine learning

Parte A:
Saper leggere i dati

Estrarre informazione dai dati: un esempio (1)

- Sono una studentessa iscritta al primo anno di università di un corso di laurea in Lettere.
- Nel primo semestre dovrò affrontare 4 esami: Letteratura, Storia, Inglese e Informatica.
- Vorrei farmi un'idea su quanto sia complicato passare ognuno di questi esami, ma non conosco nessuno degli studenti che ha sostenuto questi esami in passato.
- Decido di cercare informazioni sul sito dell'università.

Estrarre informazione dai dati: un esempio (2)

Trovo la media dei voti presi dagli studenti:

- Letteratura=20
- Informatica=20
- Storia=20
- Inglese=20

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{\sum_1^N x_i}{N}$$

Estrarre informazione dai dati: un esempio (3)

Trovo la mediana dei voti presi dagli studenti:

- Letteratura= 20
- Informatica= 20
- Storia= 20
- Inglese= 15

La mediana è un valore tale per cui al più metà degli elementi stanno al di sopra e al più metà stanno al di sotto di esso.

Estrarre informazione dai dati: un esempio (4)

Trovo la varianza dei voti presi dagli studenti:

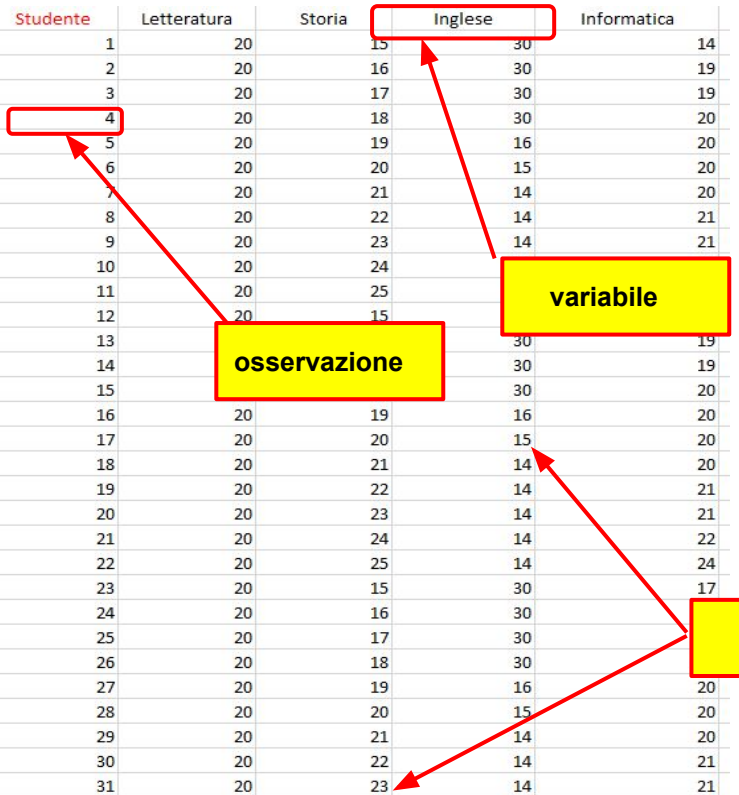
- Letteratura= 0
- Informatica= 5,45
- Storia= 10
- Inglese= 56.5

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{x})^2}{N-1}$$

I dati grezzi

Studente	Letteratura	Storia	Inglese	Informatica
1	20	15	30	14
2	20	16	30	19
3	20	17	30	19
4	20	18	30	20
5	20	19	16	20
6	20	20	15	20
7	20	21	14	20
8	20	22	14	21
9	20	23	14	21
10	20	24		
11	20	25		
12	20	15		
13			30	19
14			30	19
15			30	20
16	20	19	16	20
17	20	20	15	20
18	20	21	14	20
19	20	22	14	21
20	20	23	14	21
21	20	24	14	22
22	20	25	14	24
23	20	15	30	17
24	20	16	30	
25	20	17	30	
26	20	18	30	
27	20	19	16	20
28	20	20	15	20
29	20	21	14	20
30	20	22	14	21
31	20	23	14	21
32	20	24	14	22
33	20	25	14	24
34	20	15	30	17
35	20	16	30	19
36	20	17	30	19
37	20	18	30	20
38	20	19	16	20
39	20	20	15	20
40	20	21	14	20

40	20	21	14	20
41	20	22	14	21
42	20	23	14	21
43	20	24	14	22
44	20	25	14	24
45	20	15	30	17
46	20	16	30	19
47	20	17	30	19
48	20	18	30	20
49	20	19	16	20
50	20	20	15	20
51	20	21	14	20
52	20	22	14	21
53	20	23	14	21
54	20	24	14	22
55	20	25	14	24
56	20	15	30	17
57	20	16	30	19
58	20	17	30	19
59	20	18	30	20
60	20	19	16	20
61	20	20	15	20
62	20	21	14	20
63	20	22	14	21
64	20	23	14	21
65	20	24	14	22
66	20	25	14	24
67	20	15	30	17
68	20	16	30	19
69	20	17	30	19
70	20	18	30	20
71	20	19	16	20
72	20	20	15	20
73	20	21	14	20
74	20	22	14	21
75	20	23	14	21
76	20	24	14	22
77	20	25	14	24



Semplifichiamo un po’

Studente	Letteratura	Storia	Inglese	Informatica
1	20	15	30	17
2	20	16	30	19
3	20	17	30	19
4	20	18	30	20
5	20	19	16	20
6	20	20	15	20
7	20	21	14	20
8	20	22	14	21
9	20	23	14	21
10	20	24	14	22
11	20	25	14	24

Ragioniamo insieme

Dal punto di vista dell'utente:

- Quale sarebbe potuto essere un modo più intuitivo di sintetizzare l'informazione, rispetto ai tre indici di media, moda e varianza?

Dal punto di vista del data analyst:

- Come possiamo capire in quali casi sarebbe opportuno utilizzare un indicatore piuttosto che un altro?

- Se la studentessa avesse trovato SOLO il dato sulla media, quale "scenario" avrebbe ritenuto più verosimile?

- In quale dei quattro "scenari" sarebbe stato "deontologicamente corretto" fornire SOLO il dato sulla media?

Grafico: quale?

Dai valori assoluti alle “frequenze”

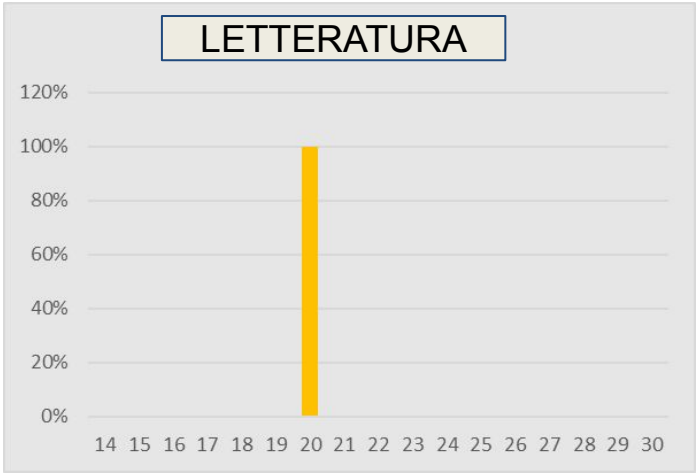
Studente	Letteratura	Storia	Inglese	Informatica
1	20	15	30	17
2	20	16	30	19
3	20	17	30	19
4	20	18	30	20
5	20	19	16	20
6	20	20	15	20
7	20	21	14	20
8	20	22	14	21
9	20	23	14	21
10	20	24	14	22
11	20	25	14	24

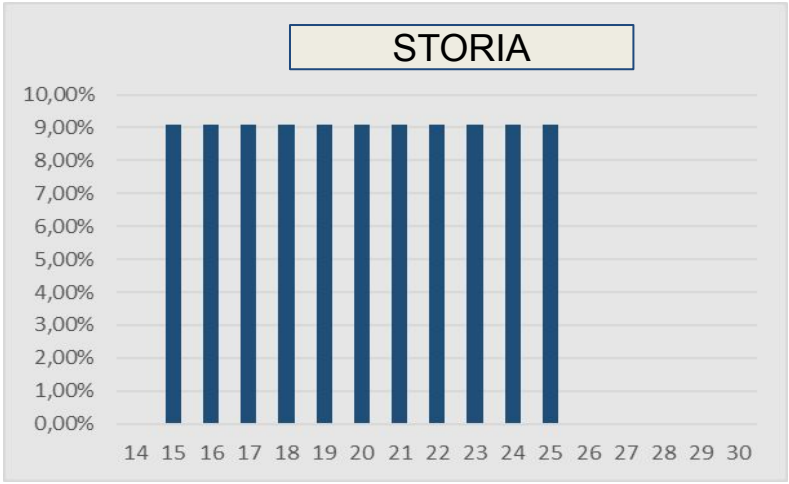
Tutti i possibili valori che la variabile può assumere: *ASSE X*

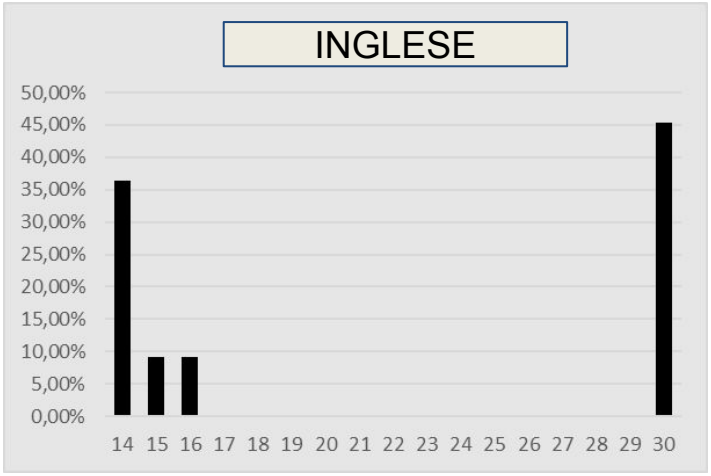
Voti possibili	Letteratura	Storia	Inglese	Informatica
14	0%	0,00%	36,40%	0,00%
15	0%	9,09%	9,09%	0,00%
16	0%	9,09%	9,09%	0,00%
17	0%	9,09%	0%	9,09%
18	0%	9,09%	0%	0,00%
19	0%	9,09%	0%	18,20%
20	100%	9,09%	0%	36,40%
21	0%	9,09%	0%	18,20%
22	0%	9,09%	0%	9,09%
23	0%	9,09%	0%	0,00%
24	0%	9,09%	0%	9,09%
25	0%	9,09%	0%	0,00%
26	0%	0,00%	0%	0,00%
27	0%	0,00%	0%	0,00%
28	0%	0,00%	0%	0,00%
29	0%	0,00%	0%	0,00%
30	0%	0,00%	45,40%	0,00%

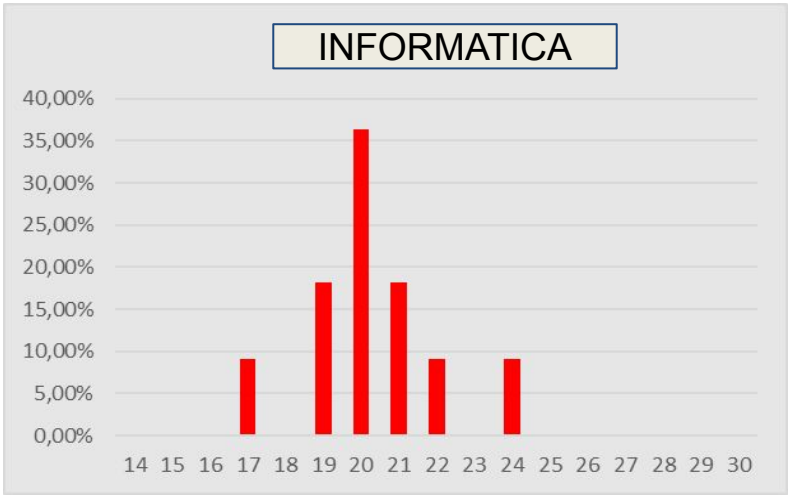
Frequenza = quante volte la variabile assume uno specifico valore/il totale delle osservazioni (11 in questo caso): *ASSE Y*

Un tipo di grafico molto esplicativo









Ragioniamo insieme

Dal punto di vista dell'utente:

- Quale sarebbe potuto essere un modo più intuitivo di sintetizzare l'informazione, rispetto ai tre indici di media, moda e varianza?
- Se la studentessa avesse trovato SOLO il dato sulla media, quale "scenario" avrebbe ritenuto più verosimile?

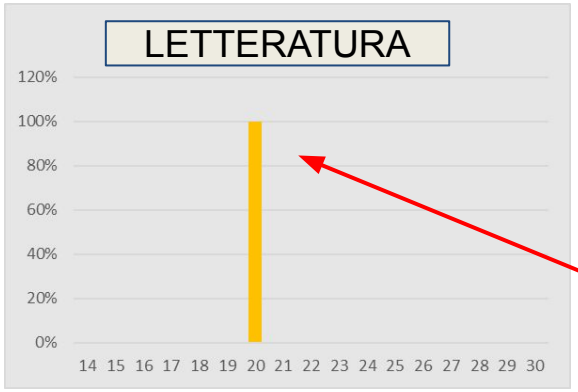
Dal punto di vista del data analyst:

- Come possiamo capire in quali casi sarebbe opportuno utilizzare un indicatore piuttosto che un altro?
- In quale dei quattro "scenari" sarebbe stato "deontologicamente corretto" fornire SOLO il dato sulla media?

Indici e informazione: un esempio

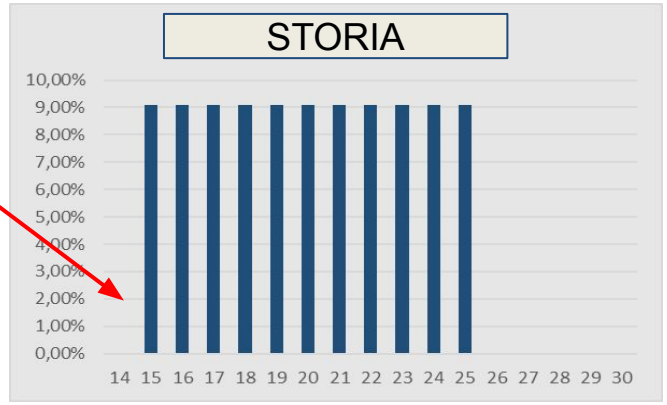
Studente	Letteratura	Storia	Inglese	Informatica
1	20	15	30	15
2	20	16	30	17
3	20	17	30	19
4	20	18	30	20
5	20	19	16	20
6	20	20	15	20
7	20	21	14	20
8	20	22	14	21
9	20	23	14	21
10	20	24	14	22
11	20	25	14	25

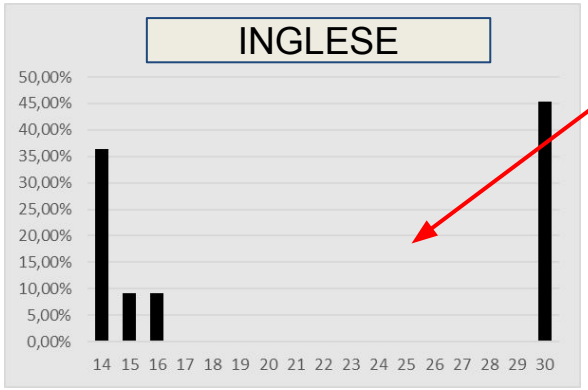
Ogni “scenario” ha un nome: le distribuzioni



COSTANTE

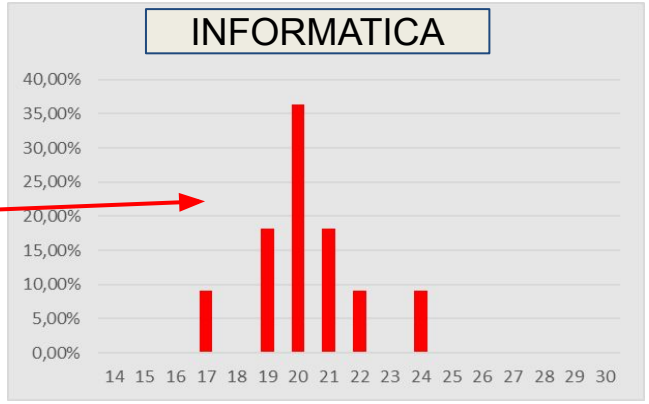
UNIFORME



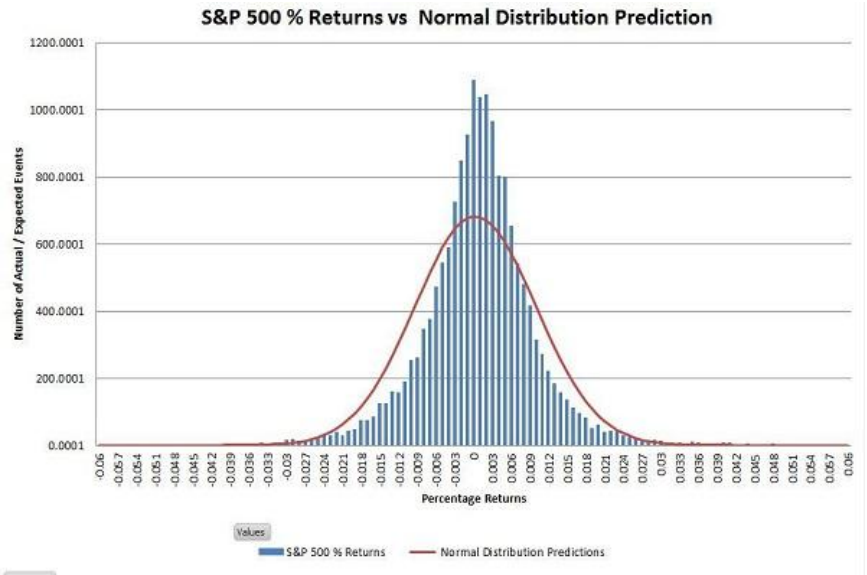


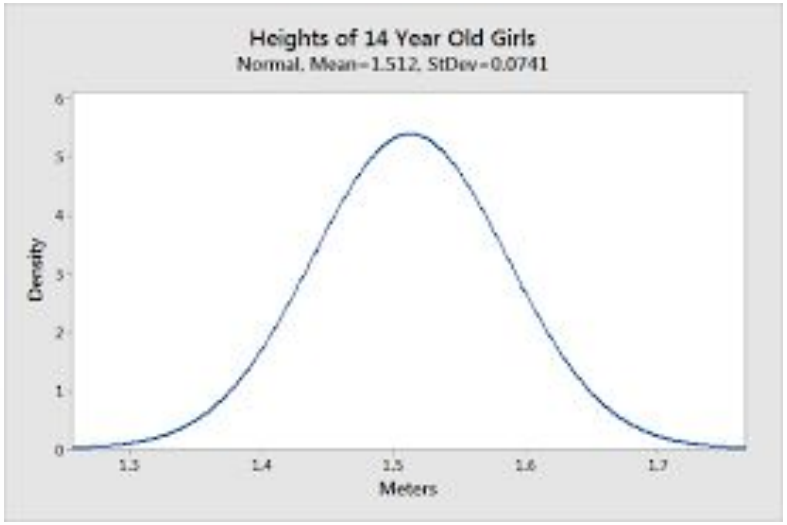
BIMODALE

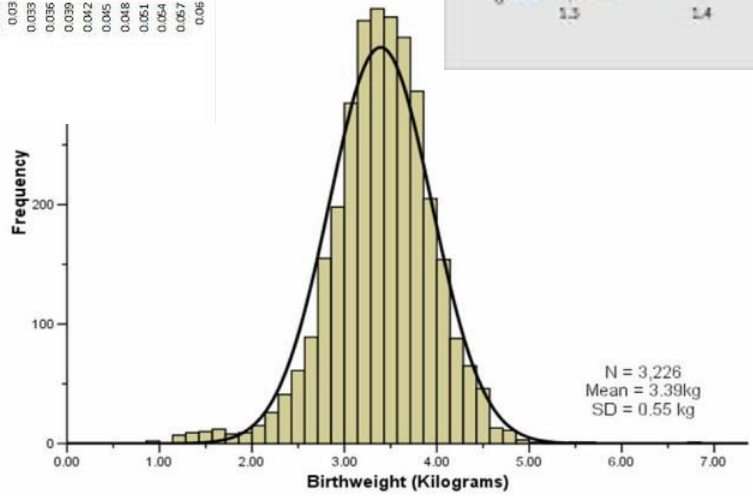
NORMALE



Una distribuzione molto particolare: la distribuzione normale (o gaussiana)

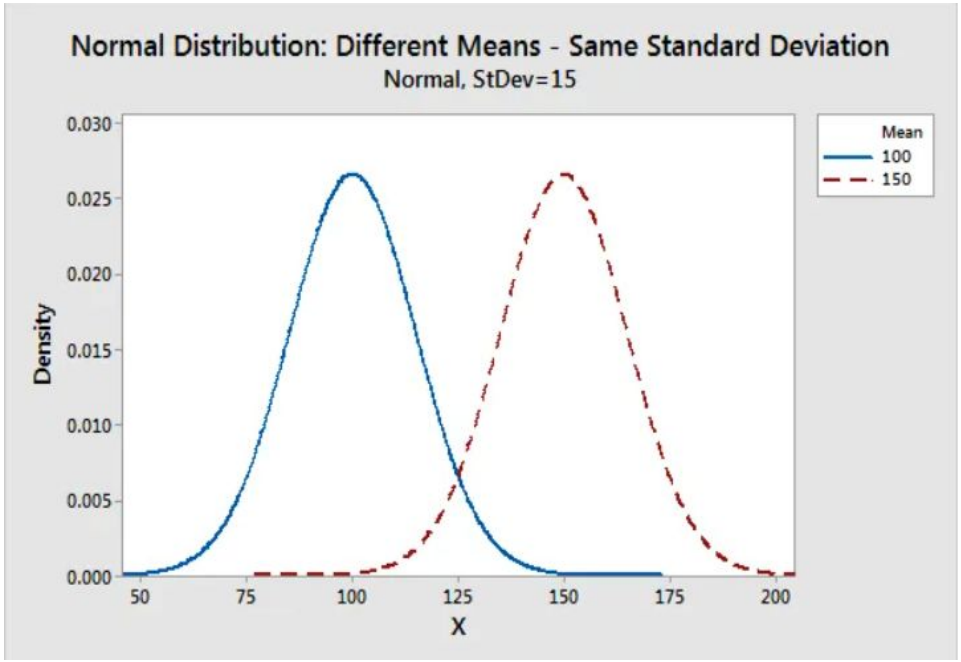


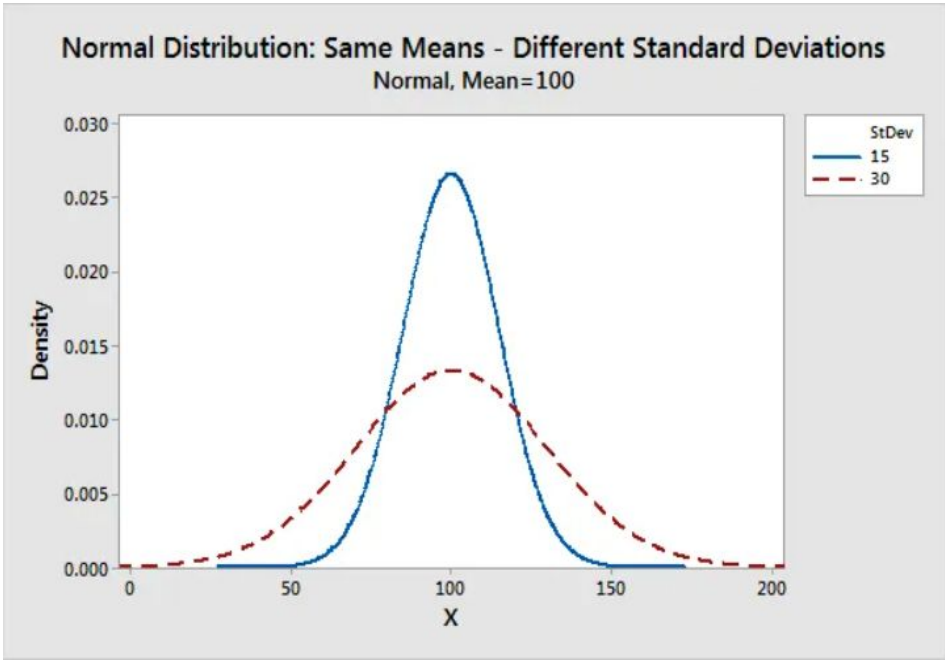




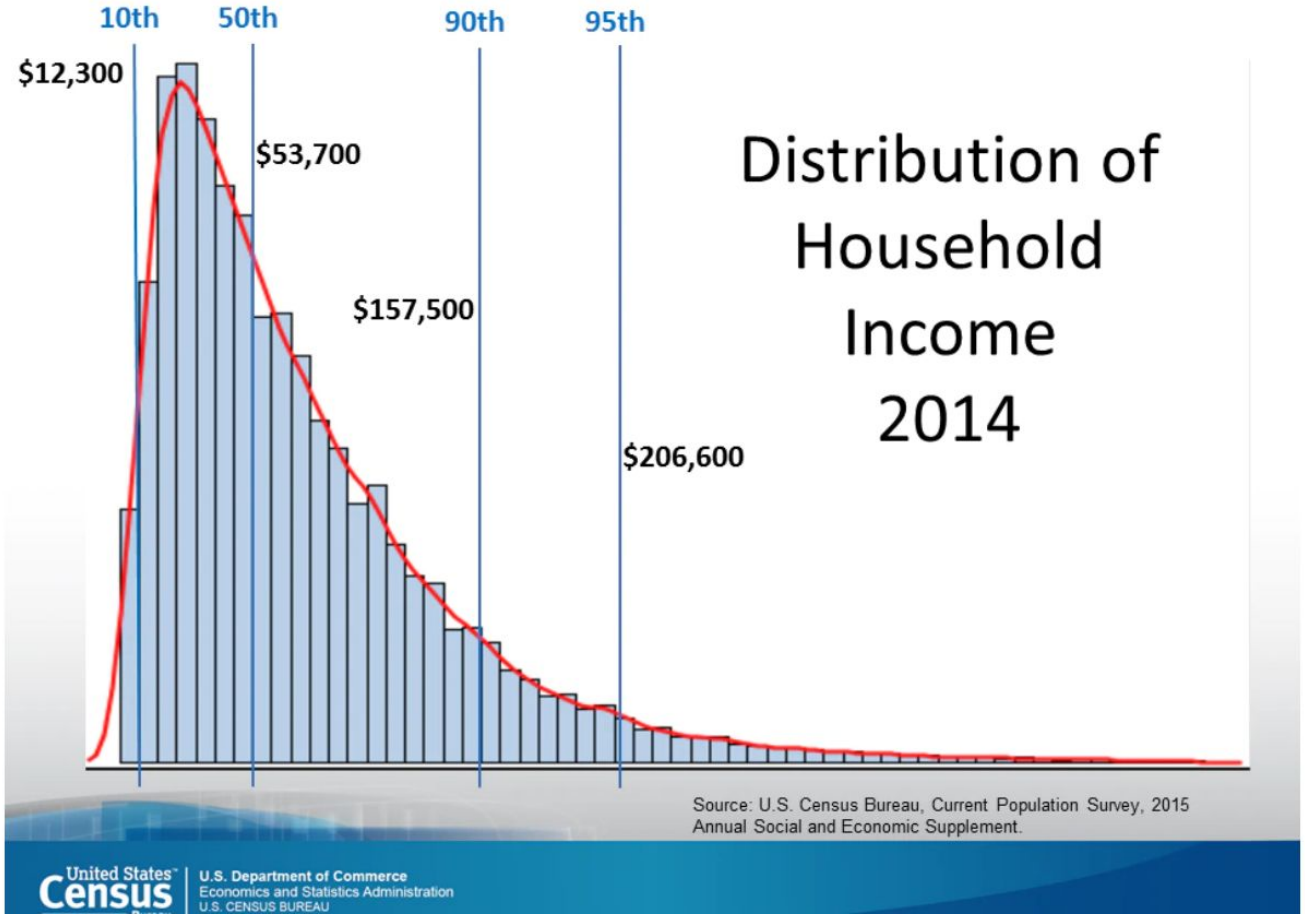
CD:50/50
CODING DIVERSITY

Media e varianza nella distribuzione gaussiana



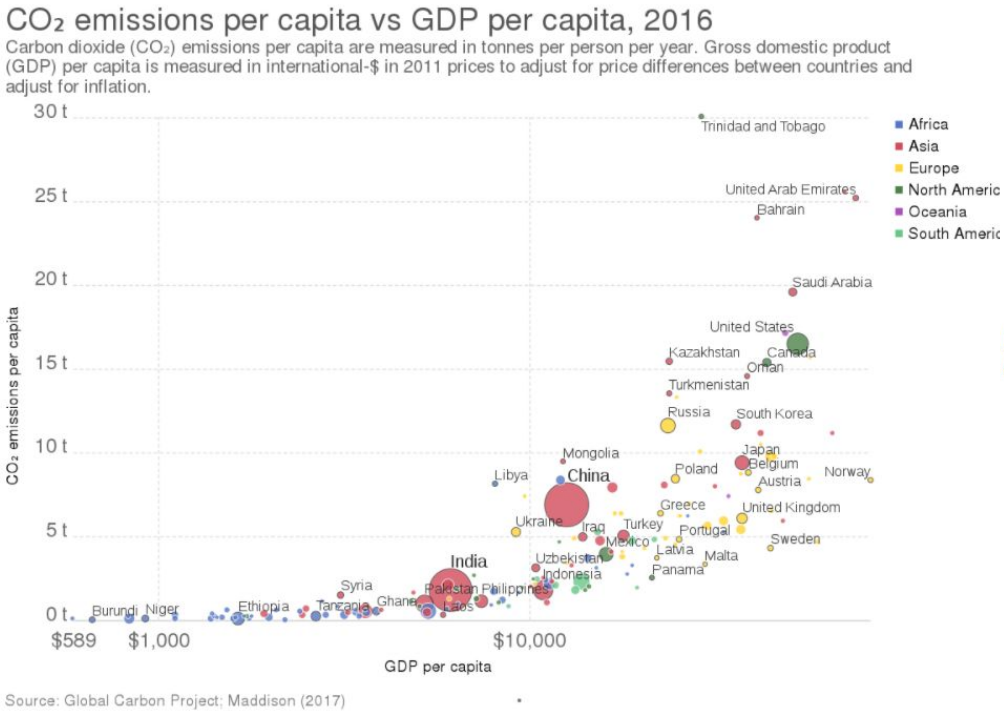


Non sempre una determinata serie di dati ha una distribuzione asimmetrica

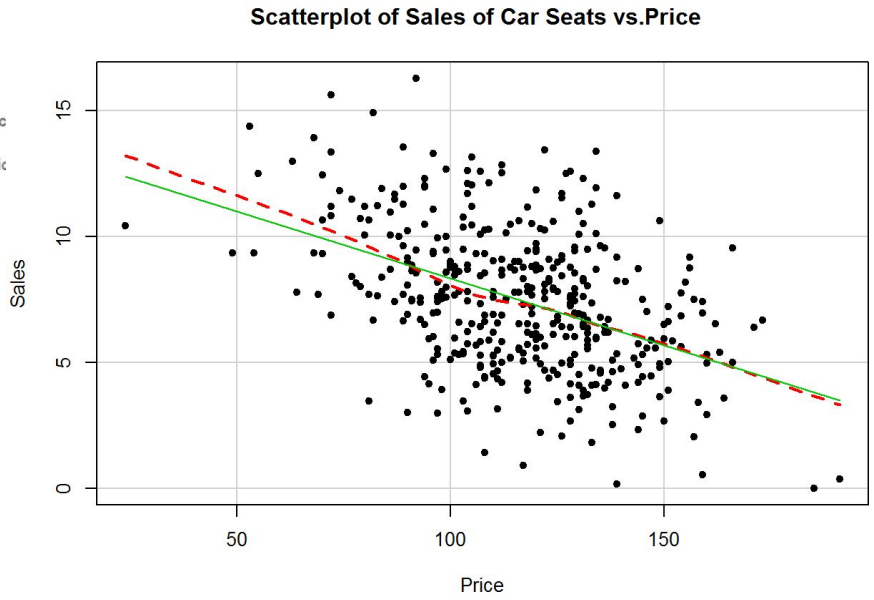


Parte B:
Saper rappresentare i dati

Esplorare la correlazione tra due serie di dati



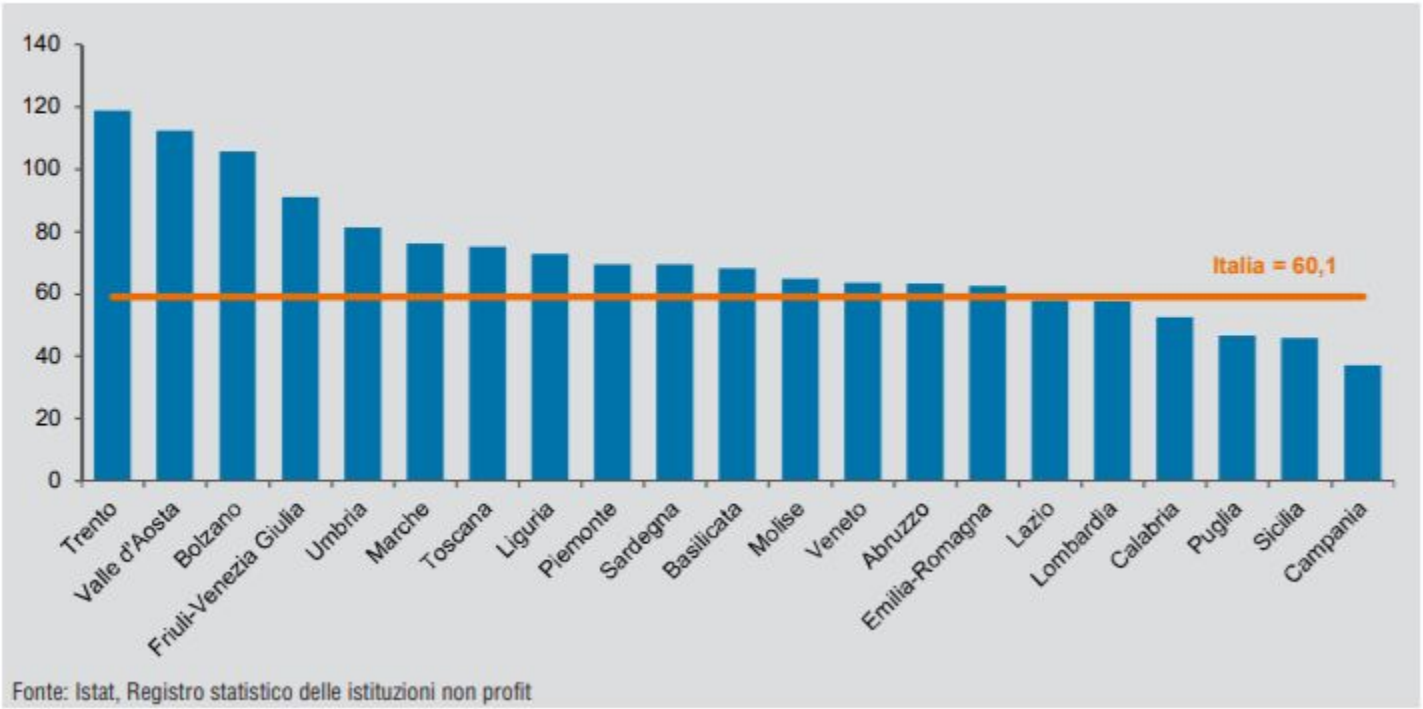
(a) correlazione positiva



(b) correlazione negativa

Rappresentare congiuntamente dati numerici e non: i grafici a barre

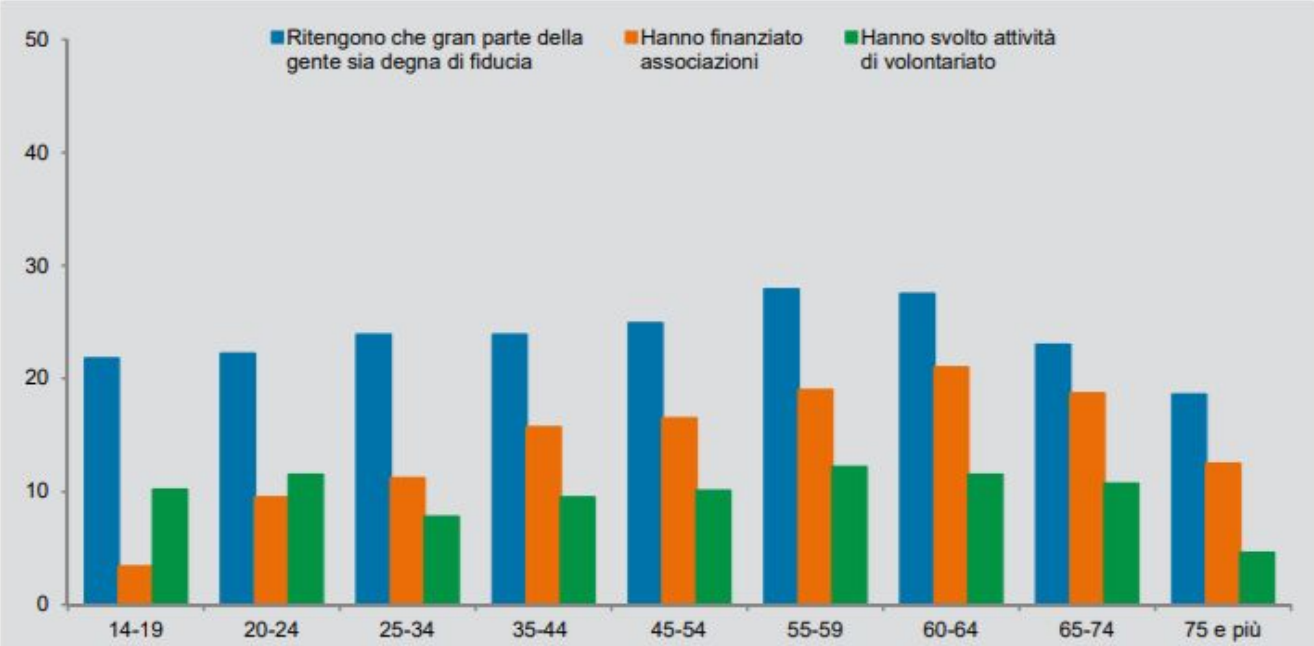
Figura 11. Numero di istituzioni non profit ogni 10.000 abitanti per regione. Anno 2018



Fonte: Istat, Registro statistico delle istituzioni non profit

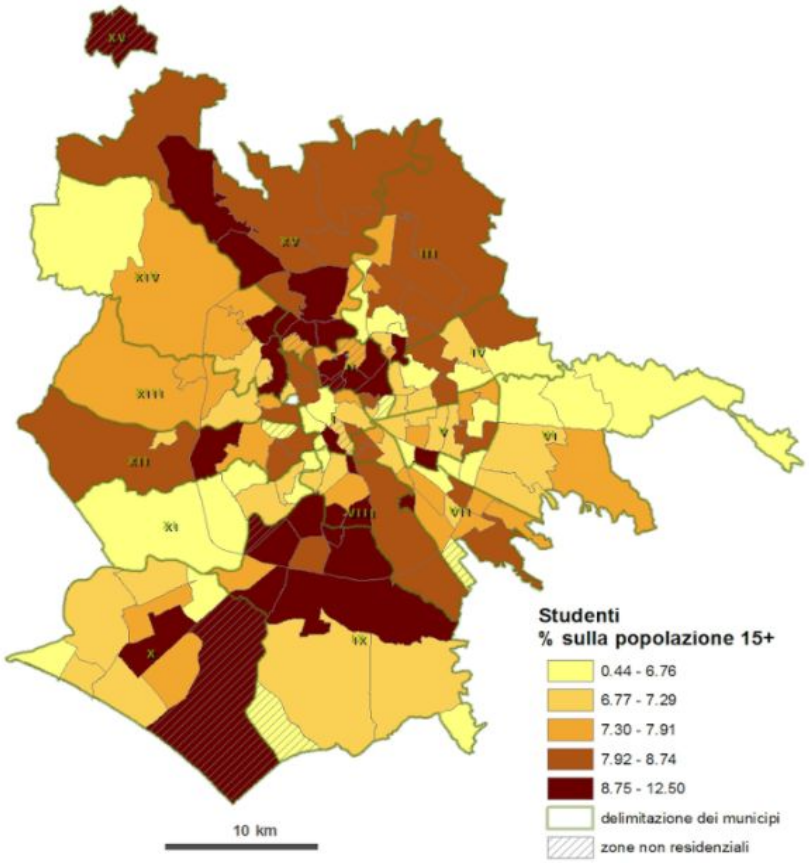
Descrivere più di due variabili congiuntamente

Figura 4. Persone di 14 anni e più che ritengono che gran parte della gente sia degna di fiducia, che negli ultimi 12 mesi hanno finanziato associazioni o che hanno svolto attività gratuita per associazioni o gruppi di volontariato per classe di età. Anno 2020 (a). Per 100 persone di 14 anni e più della stessa classe di età



Fonte: Istat, Indagine Aspetti della vita quotidiana
(a) Dato provvisorio.

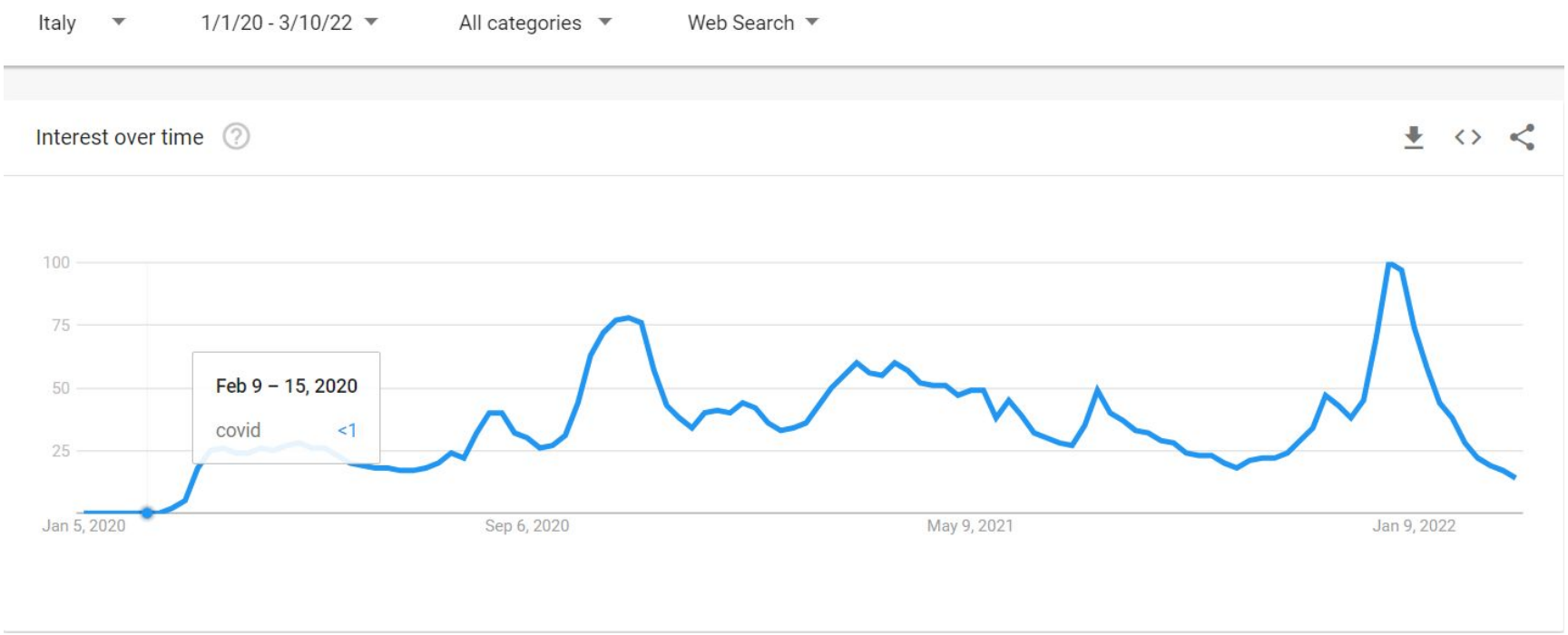
Gli studenti a Roma



Prima variabile: numero di studenti

Seconda variabile: ?

Le serie storiche: uno sguardo a Google Trends



Prima variabile: numero di casi Covid

Seconda variabile: ?

Cosa abbiamo visto oggi

1. Descrivere i dati: media, mediana, varianza
2. Dai valori assoluti alle frequenze
3. Il concetto di distribuzione e la distribuzione normale
4. Grafici bivariato di tipo “scatter”
5. Grafici bivariato di tipo “a barre”
6. Grafici a barre raggruppate
7. Grafici nel tempo: le serie storiche
8. Grafici nello spazio: le mappe
9. Alcuni errori comuni nella data analysis e come evitarli