PCTO in Coding & Data Science

Modulo 2

Liceo Morgagni, Roma, 02/02/2022

CD:50/50
CODING DIVERSITY

Dati, dati, dati!

Regione	AGGIORNAMENTO 03/04/2020 ORE 17.00							
	POSITIVI AL nCoV				DIMESSI/ GUARITI	DECEDUTI	CASI TOTALI	TAMPONI
	Ricoverati con sintomi	Terapia intensiva	Isolamento domiciliare	Totale attualmente positivi				
Lombardia	11.802	1.381	13.006	26.189	13.020	8.311	47.520	135.051
Emilia Romagna	3915	364	7899	12.178	1.852	1.902	15.932	63.682
Piemonte	3.300	452	5.378	9.130	723	1.043	10.896	34.281
Veneto	1714	335	6812	8.861	1.031	572	10.464	126.490
Toscana	1.149	288	3.472	4.909	300	290	5.499	44.460
Marche	982	158	2491	3.631	42	557	4.230	13.678
Liguria	1147	173	1426	2.746	700	519	3.965	12.934
Lazio	1194	188	1627	3.009	392	199	3.600	43.776
Campania	532	115	1705	2.352	144	181	2.677	19.237
Trento	343	80	1.236	1.659	246	204	2.109	8.993
Puglia	648	123	1178	1.949	69	164	2.182	17.924
Friuli V.G.	201	61	1.062	1.324	419	136	1.879	19.985
Sicilia	535	73	1.056	1.664	94	101	1.859	18.686
Abruzzo	361	76	864	1.301	116	146	1.563	11.890
Bolzano	291	60	858	1.209	211	139	1.559	13.976
Umbria	165	48	707	920	220	39	1.179	10.614
Sardegna	122	24	598	744	40	41	825	6.478
Calabria	183	17	462	662	26	45	733	11.608
Valle d'Aosta	85	25	450	560	89	70	719	2.106
Basilicata	41	19	187	247	3	11	261	2.622
Molise	31	8	105	144	21	11	176	1.378
TOTALE	28.741	4.068	52.579	85.388	19.758	14.681	119.827	619.849

ATTUALMENTE POSITIVI	85.388
TOTALE GUARITI	19.758
TOTALE DECEDUTI	14.681
CASI TOTALI	119.827

I dati grezzi sono difficili da trasformare in informazione utile.

La nostra attenzione e capacità di comprendere è limitata: cerchiamo quindi forme di aggregazioni che:

- Riducano la complessità
- Facciano emergere l'informazione che ci interessa.

Il **totale** è una quantità aggregata che ci dà una informazione sul fenomeno globale

La media

MEDIA

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

età = [10, 14, 17, 19, 21, 16, 18]

Media = 16.42 anni

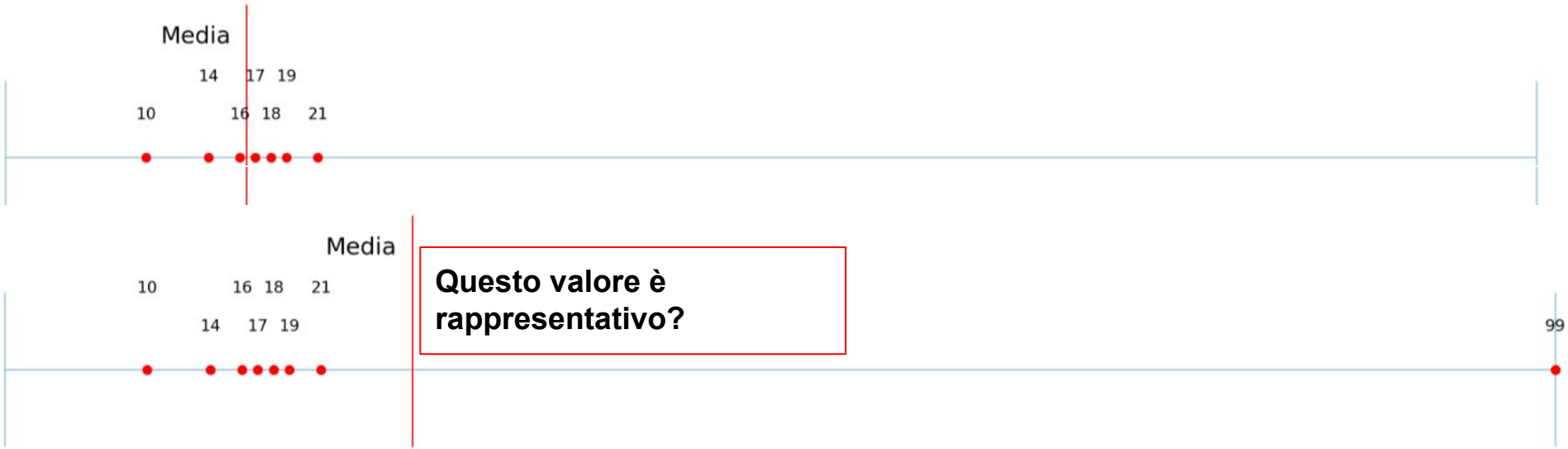
La media

$$\bar{x} = \frac{\sum_1^N x_i}{N}$$

La media vuole essere il valore più rappresentativo di un insieme di osservazioni, il “valore tipico”.

età = [10, 14, 17, 19, 21, 16, 18] ← Media: 16.42 anni

età = [10, 14, 17, 19, 21, 16, 18, 99] ← Media 26.75



La mediana

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.

età = [10, 14, 17, 19, 21, 16, 18]

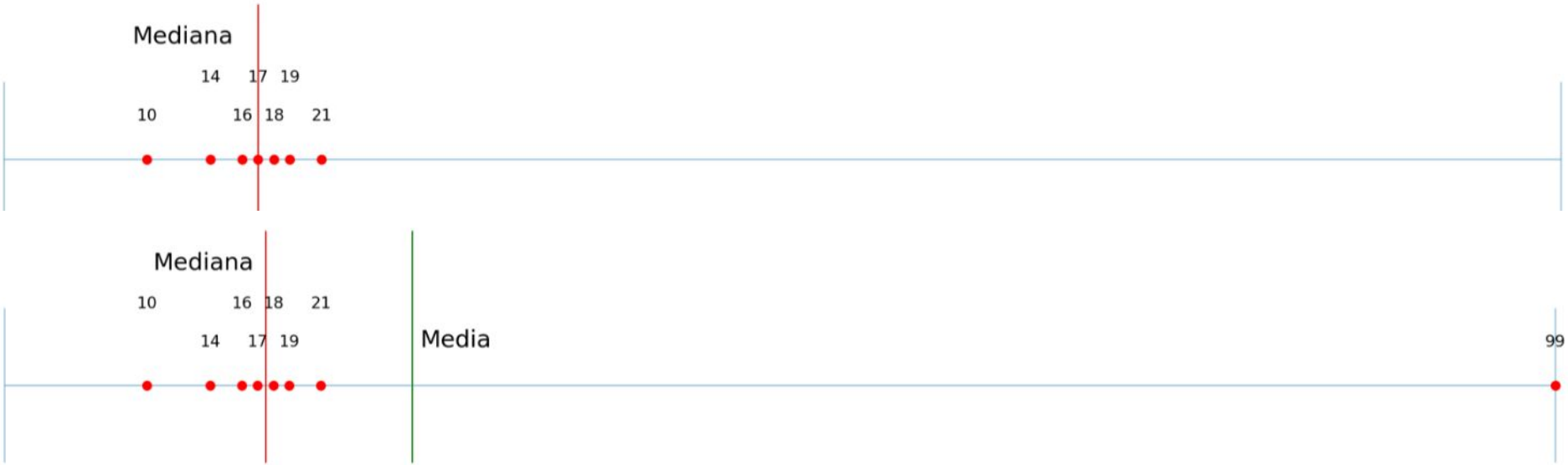
10, 14, 16, **17**, 18, 19, 21

La mediana

La mediana “ignora” il valore effettivo degli elementi; non ci importa di quanto stanno sopra o sotto, ma solo separarli in due parti uguali: non siamo sensibili a valori molto grandi o piccoli, che sposterebbero la media.

età = [10, 14, 17, 19, 21, 16, 18] ← Mediana 17

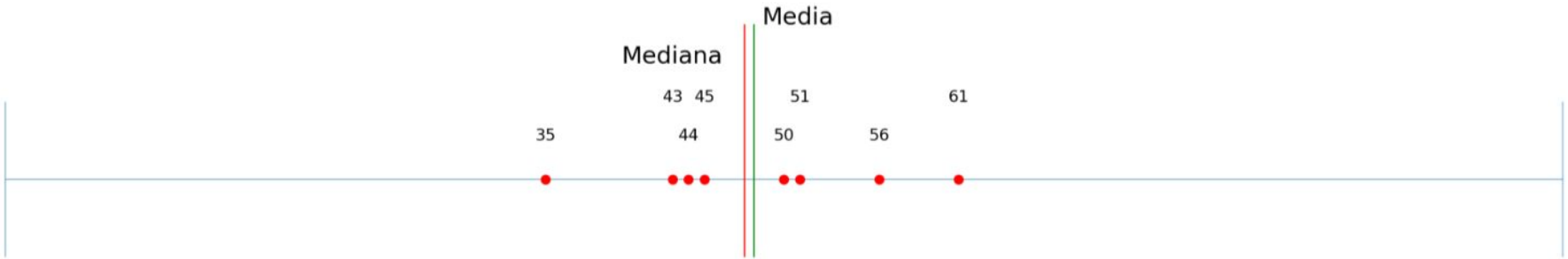
età = [10, 14, 17, 19, 21, 16, 18, 99] ← Mediana 17.5

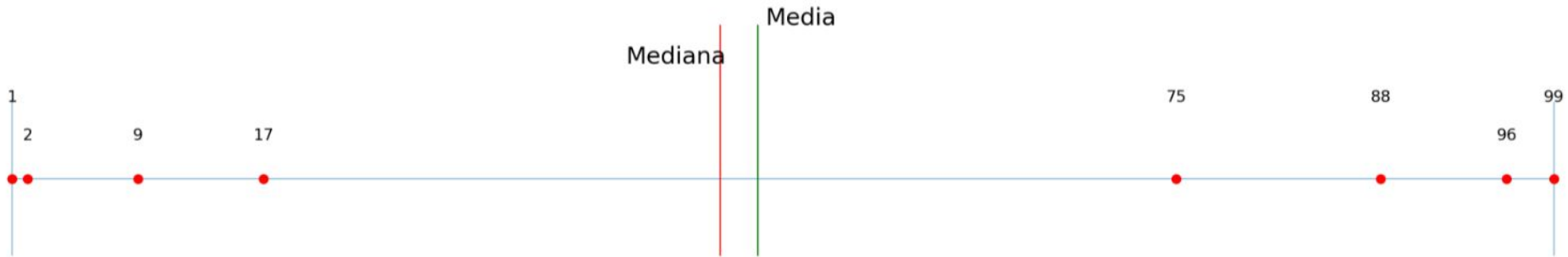


Un valore può arrivare da popolazioni molto diverse

età = [50, 56, 45, 61, 35, 43, 44, 51] <- Media: 48.1, mediana: 47.5

età = [1, 2, 9, 17, 88, 75, 99, 96] <- Media: 48.4, mediana: 46





Un valore può arrivare da popolazioni molto diverse

età = [50, 56, 45, 61, 35, 43, 44, 51] ← Media: 48.1, mediana: 47.5

età = [1, 2, 9, 17, 88, 75, 99, 96] ← Media: 48.4, mediana: 46

Abbiamo preso un insieme di 8 osservazioni, e abbiamo “compresso” la loro informazione in un unico numero.
Questa misura ci ha fatto perdere delle informazioni importanti!

Vorremmo sapere ad esempio come i valori si discostano da questi aggregatori. Ad esempio quanto sono spostati rispetto alla media.

Deviazione standard

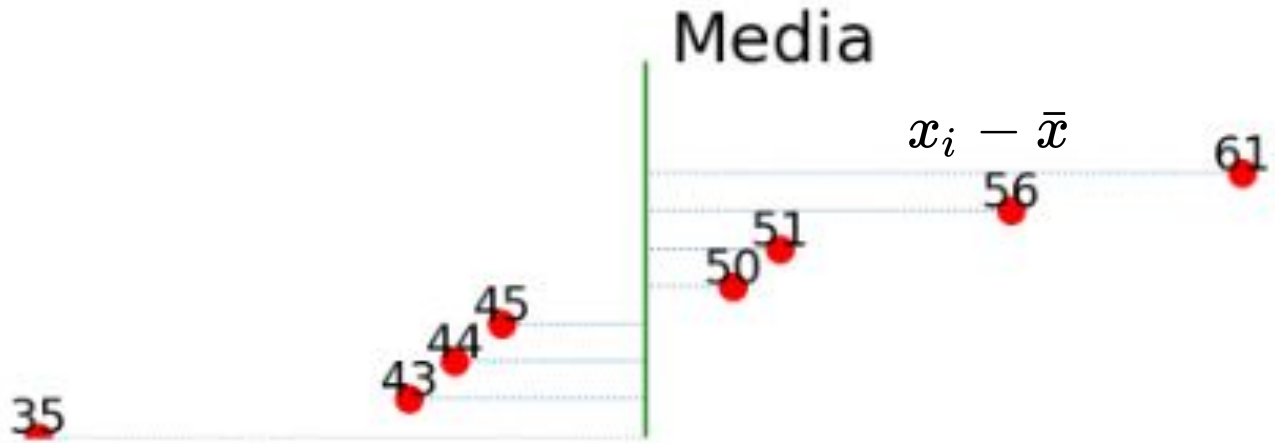
Quanto si discostano i valori dalla media? Quanto variano rispetto alla media?

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

età = [50, 56, 45, 61, 35, 43, 44, 51]
media(età) = 47.5

$$\sigma = \sqrt{\sigma^2}$$

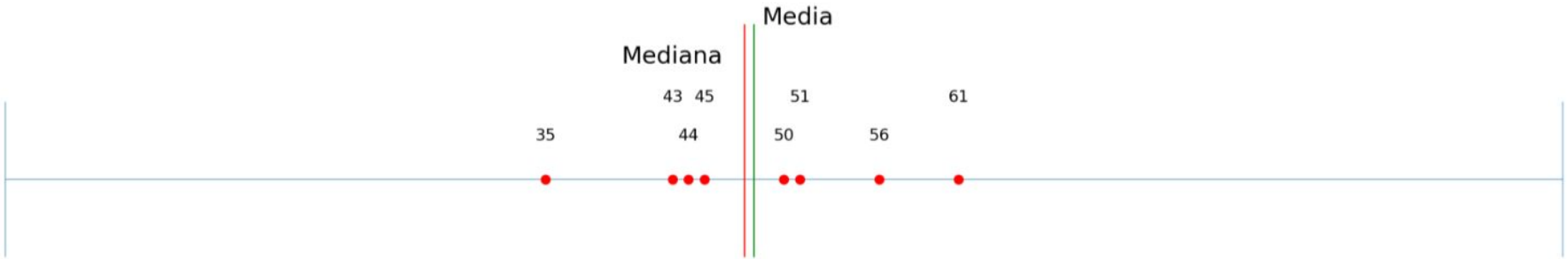
$$\sigma = \frac{(50-47.5)^2 + (56-47.5)^2 + \cdots + (51-47.5)^2}{7} = 7.6$$

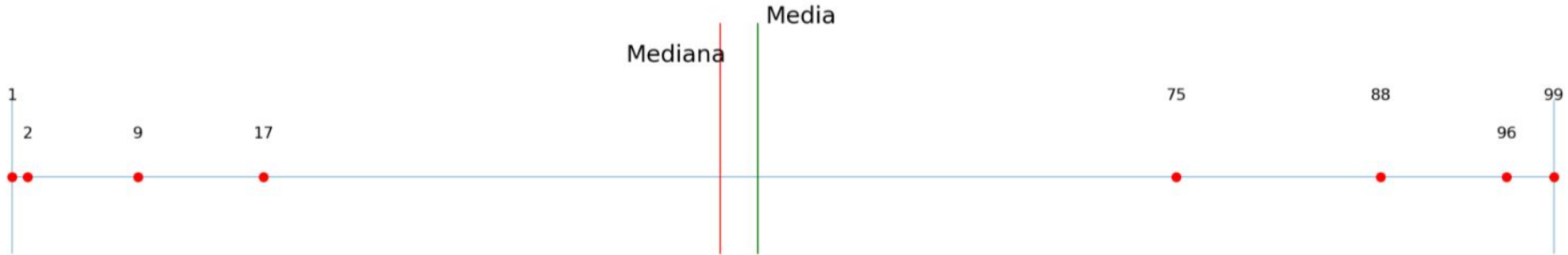


Nel caso precedente….

età = [50, 56, 45, 61, 35, 43, 44, 51] <- Media: 48.1, mediana: 47.5, **σ=7.6**

età = [1, 2, 9, 17, 88, 75, 99, 96] <- Media: 48.4, mediana: 46, **σ=41.9**





Deviazione standard

età = [50, 56, 45, 61, 35, 43, 44, 51]

σ = 7.6 **E’ tanto? E’ poco?**

Deviazione standard

Confrontando due diverse varianze, è facile capire quale insieme di osservazioni è più variabile.
Ma in senso assoluto?
Per avere una misura, possiamo calcolare il rapporto tra media e varianza:
il Coefficiente di variazione:

$$CV = \frac{\sigma}{\bar{x}}$$

età = [50, 56, 45, 61, 35, 43, 44, 51]
σ = 7.6
$\bar{x}$ = 48.1
CV = 0.2

Media e varianza sono nella **unità di misura del dominio**.

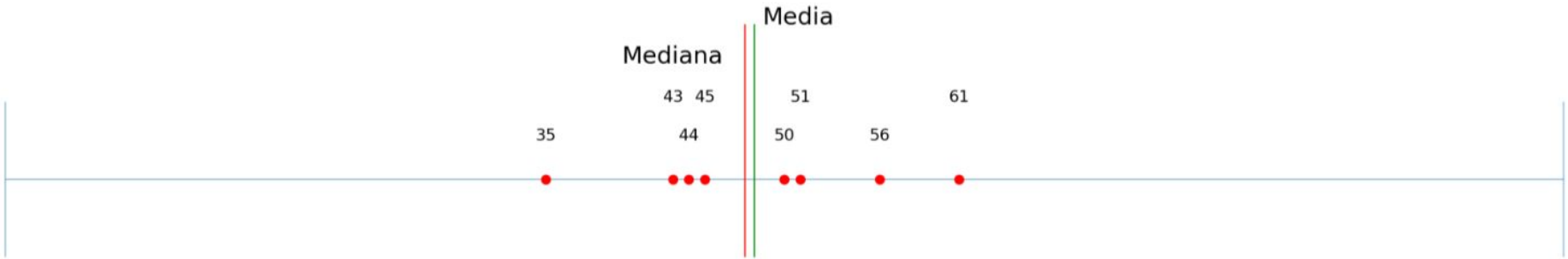
Ad esempio, l'età media è di 33.5 **anni**. Se avessimo convertito tutti i valori in mesi, avremmo ottenuto un'età media di 402 mesi (33.5x12).
Leggere questi valori richiede di avere un'idea di "tanto" o "poco" nel dominio di studio.

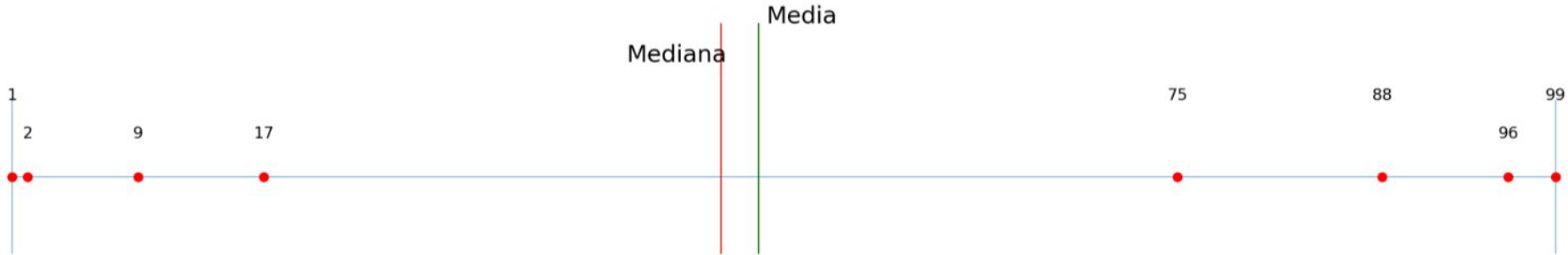
Il coefficiente di variazione è un valore senza unità di misura (**adimensionale**). Questo permette di farsi un'idea senza preoccuparsi del contesto.Viene spesso espresso come percentuale

Nel caso precedente….

età = [50, 56, 45, 61, 35, 43, 44, 51] <- Media: 48.1, mediana: 47.5, σ: 7.6, **CV=0.2**

età = [1, 2, 9, 17, 88, 75, 99, 96] <- Media: 48.4, mediana: 46, σ: 41.9 **CV=0.9**





Come possiamo “sintetizzare” un insieme di dati?

Un altro modo per trasformare un insieme di dati in informazione “comprensibile” è quello di visualizzarli

Perché visualizzare i dati?

- Mantenimento delle informazioni

Dopo 3 giorni:

- solo testo: 10% info
- Testo + visualizzazione: 65% info

[Fonte: Madina, Brain Rules, 2008]

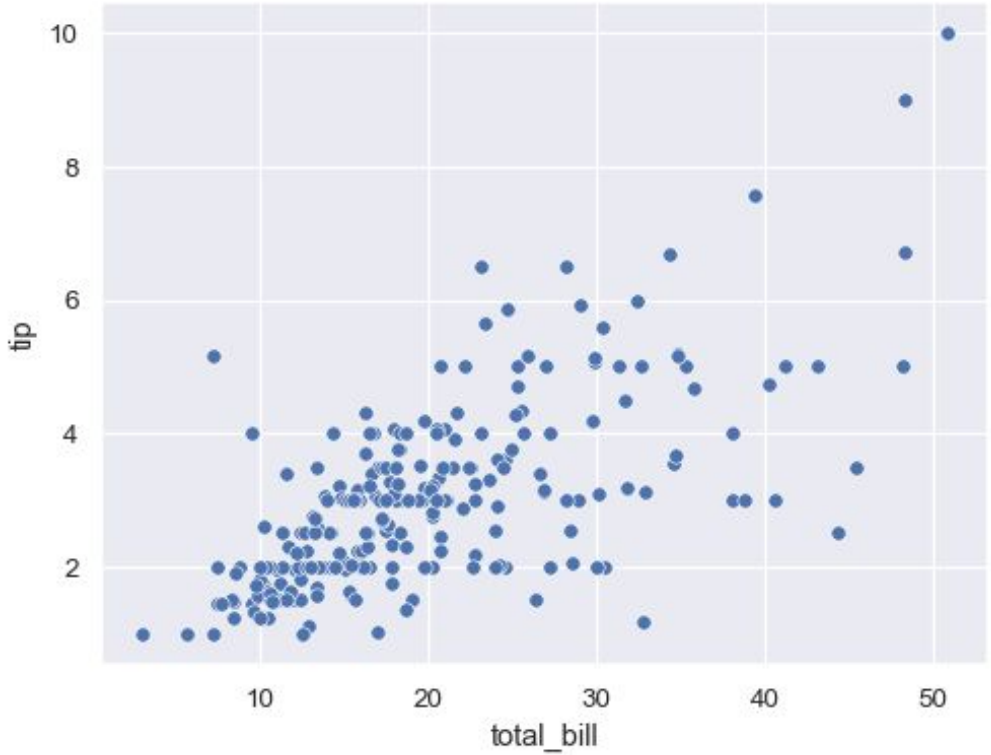
Perché visualizzare i dati?

- Mantenimento delle informazioni
- Densità delle informazioni

Perché visualizzare i dati?

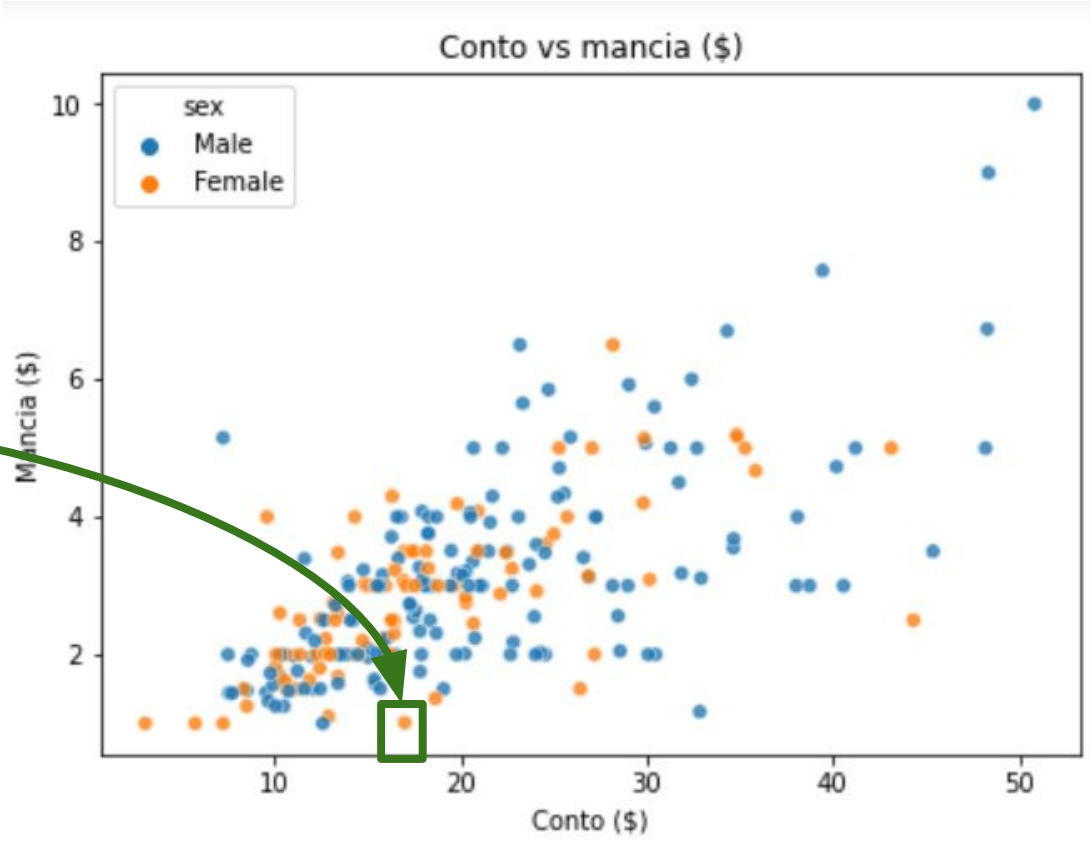
- Mantenimento delle informazioni
- Densità delle informazioni
- Confronto tra valori e contesto

Scatter Plot



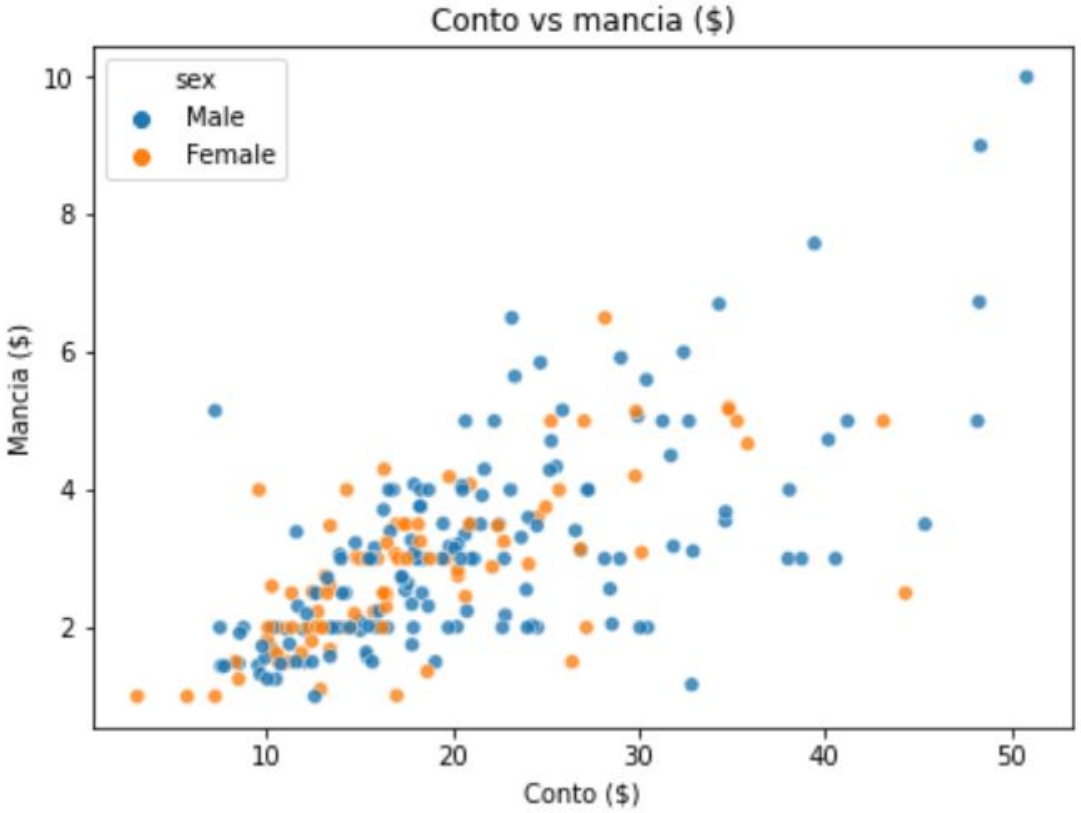
Scatterplot

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

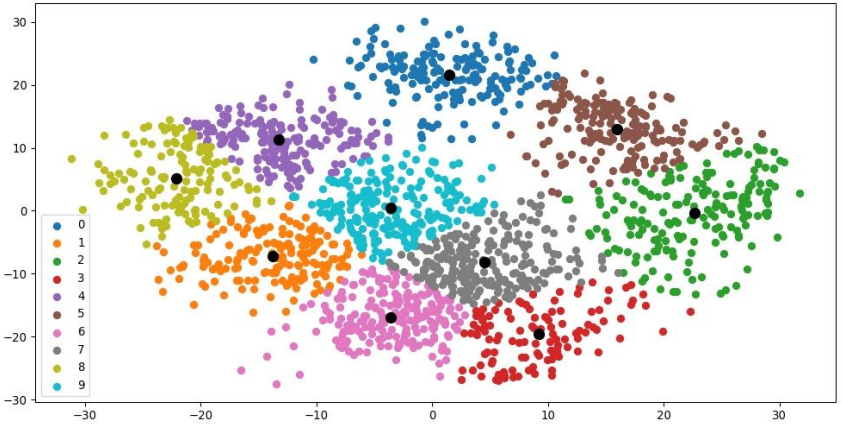


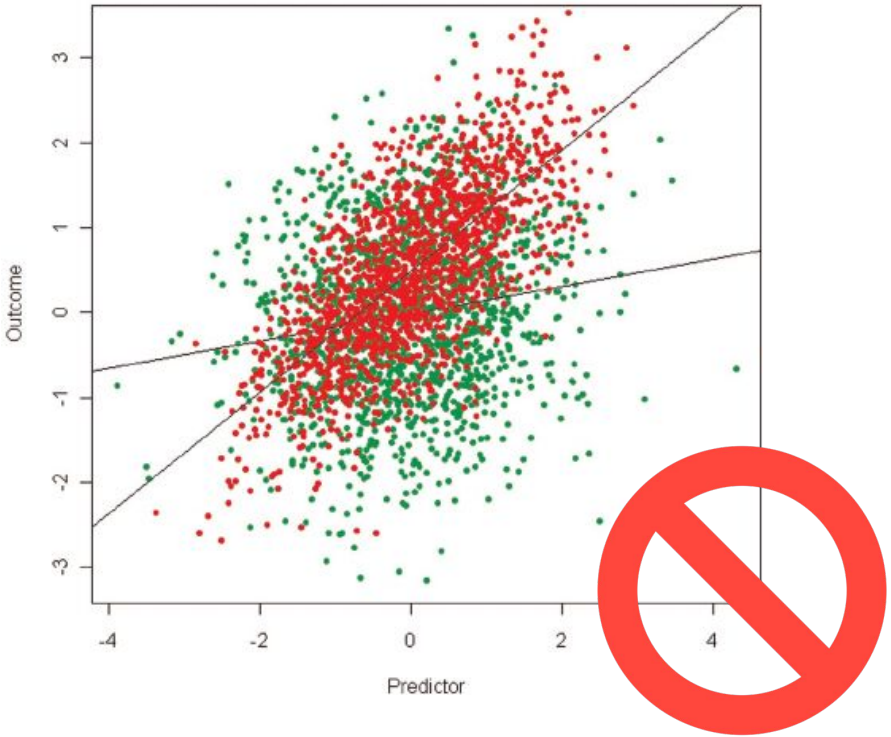
Uno scatterplot

- rappresentata ogni osservazione con un punto
- ha per assi due variabili del dataset (conto, mancia)
- è utile per rappresentare la relazione tra queste variabili
  - in questo caso, a conti più alti corrispondono in genere mance più alte
- Può rappresentare altre variabili (per esempio il sesso), col colore o la forma dei punti.

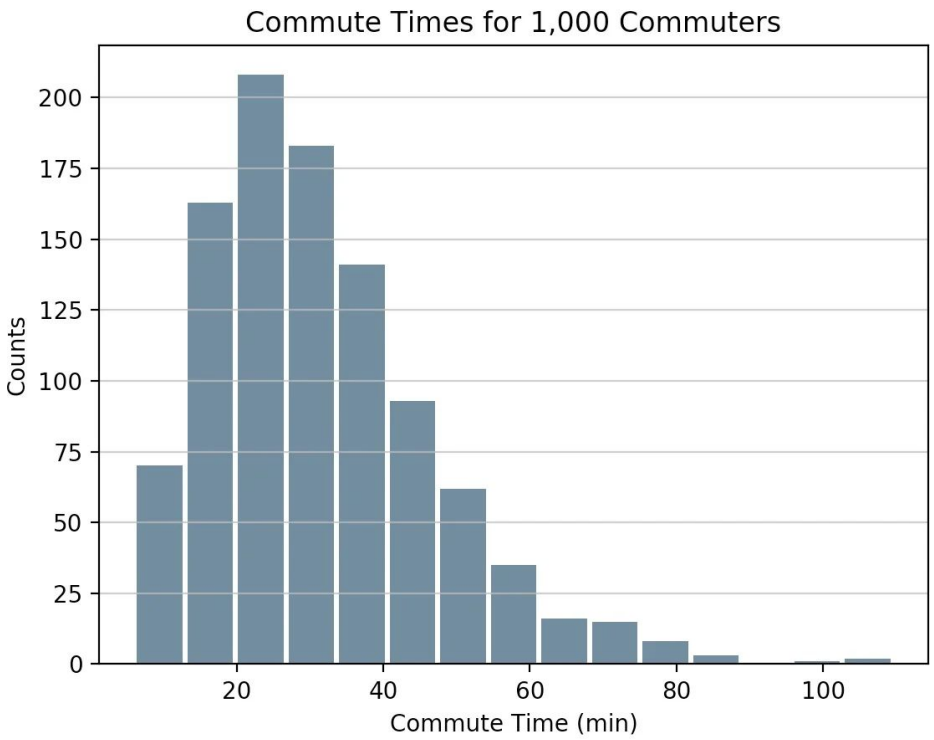


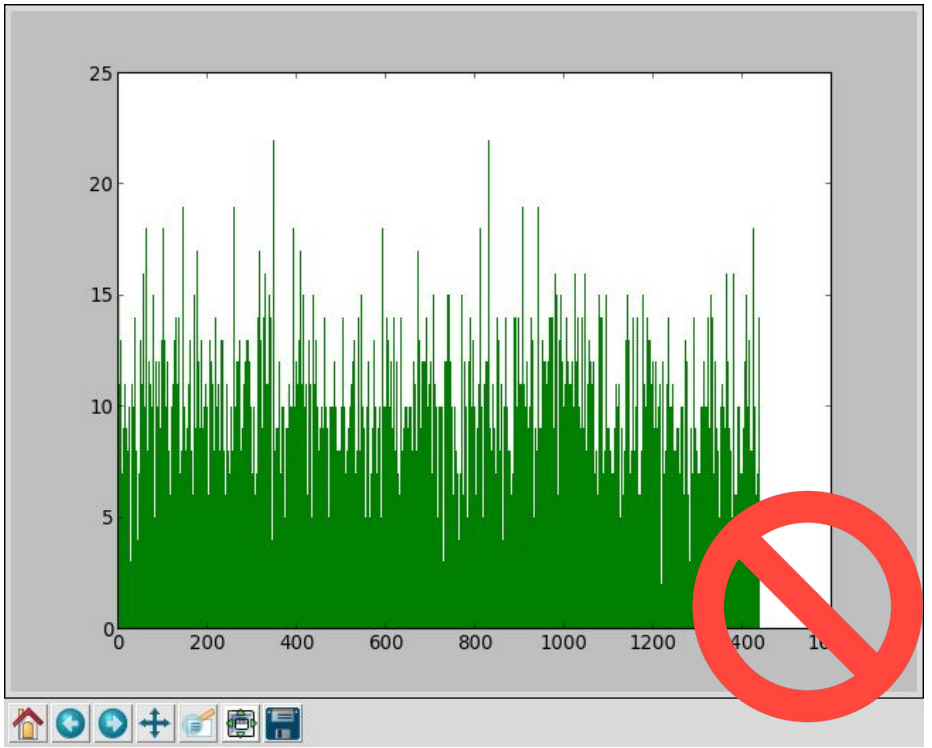
Scatter Plot



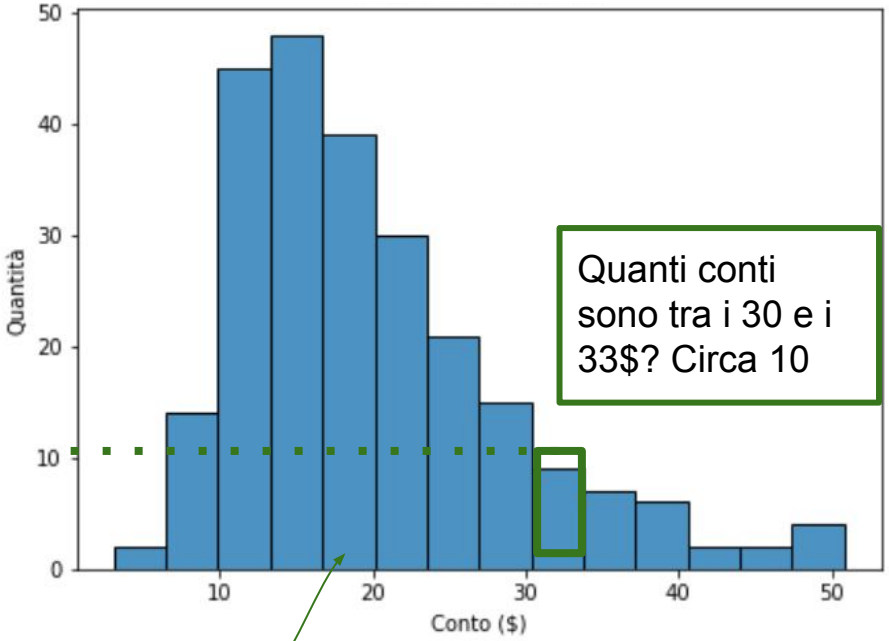


Istogrammi


Commute Times for 1,000 Commuters
Counts
0
25
50
75
100
125
150
175
200
20
40
60
80
100
Commute Time (min)


0
5
10
15
20
25
0
200
400
600
800
1000
1200

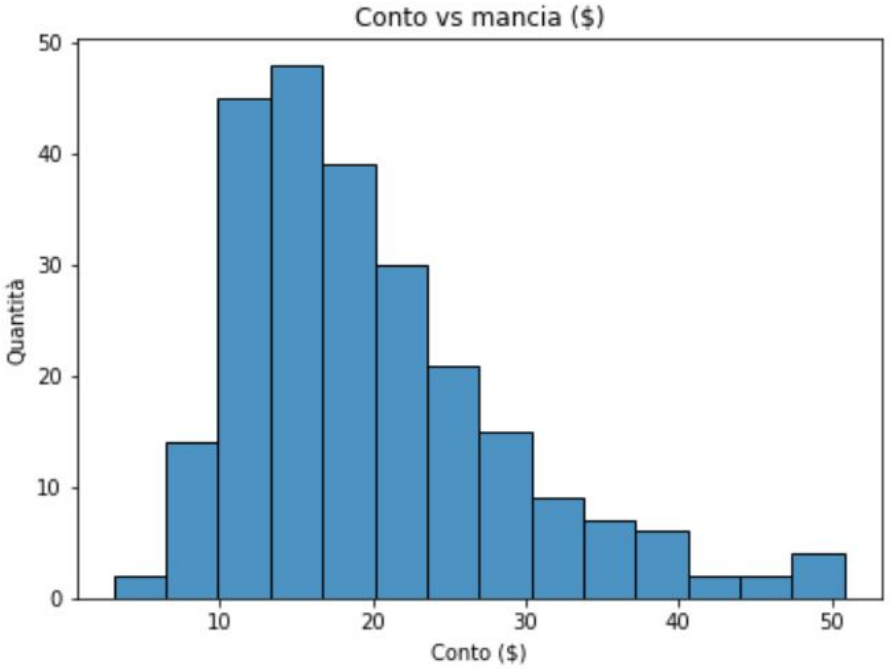
	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

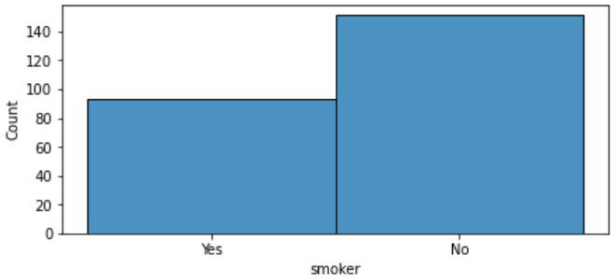


Quanti conti sono tra i 30 e i 33$? Circa 10

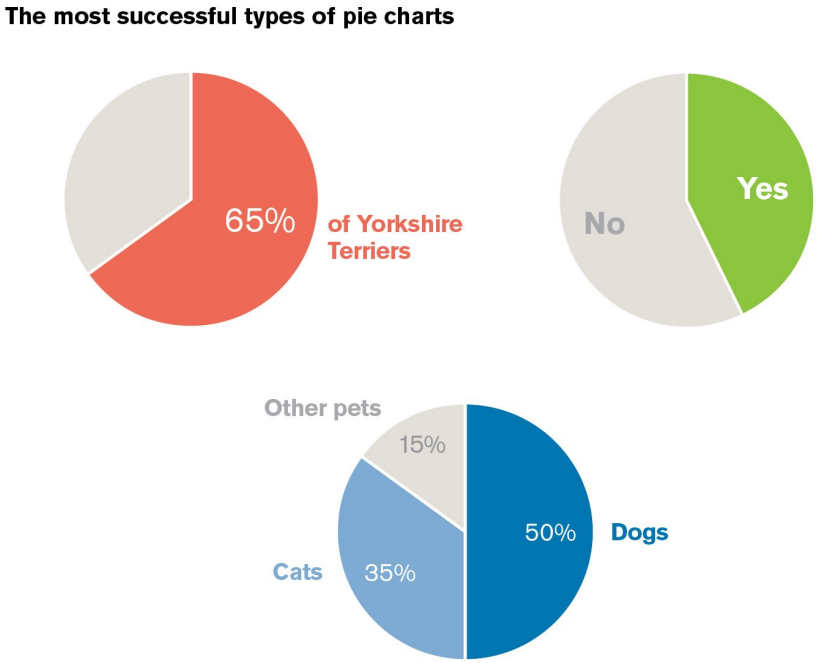
L'istogramma viene generalmente usato per rappresentare una sola variabile

- Per variabili numeriche (per esempio il valore del conto), serve a capire come è distribuito questo valore
  - Dividiamo tutti i valori in N intervalli (x)
  - Sulla y, segnamo quante volte una osservazione in tale intervallo è osservata nel nostro dataset
- Per variabili categoriche (per esempio fumatore / non fumatore), facciamo un confronto del numero di elementi in ogni categoria

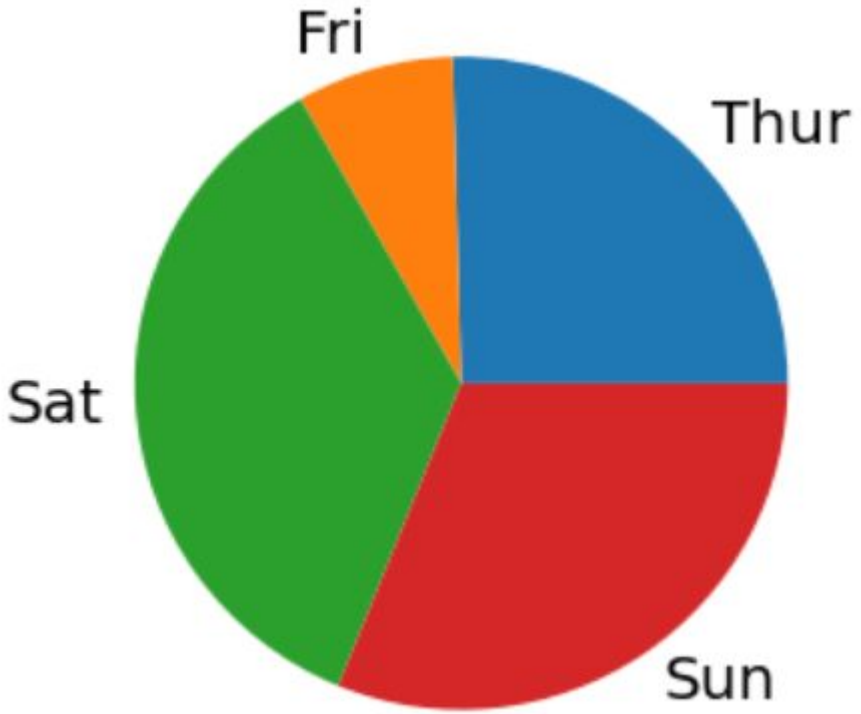


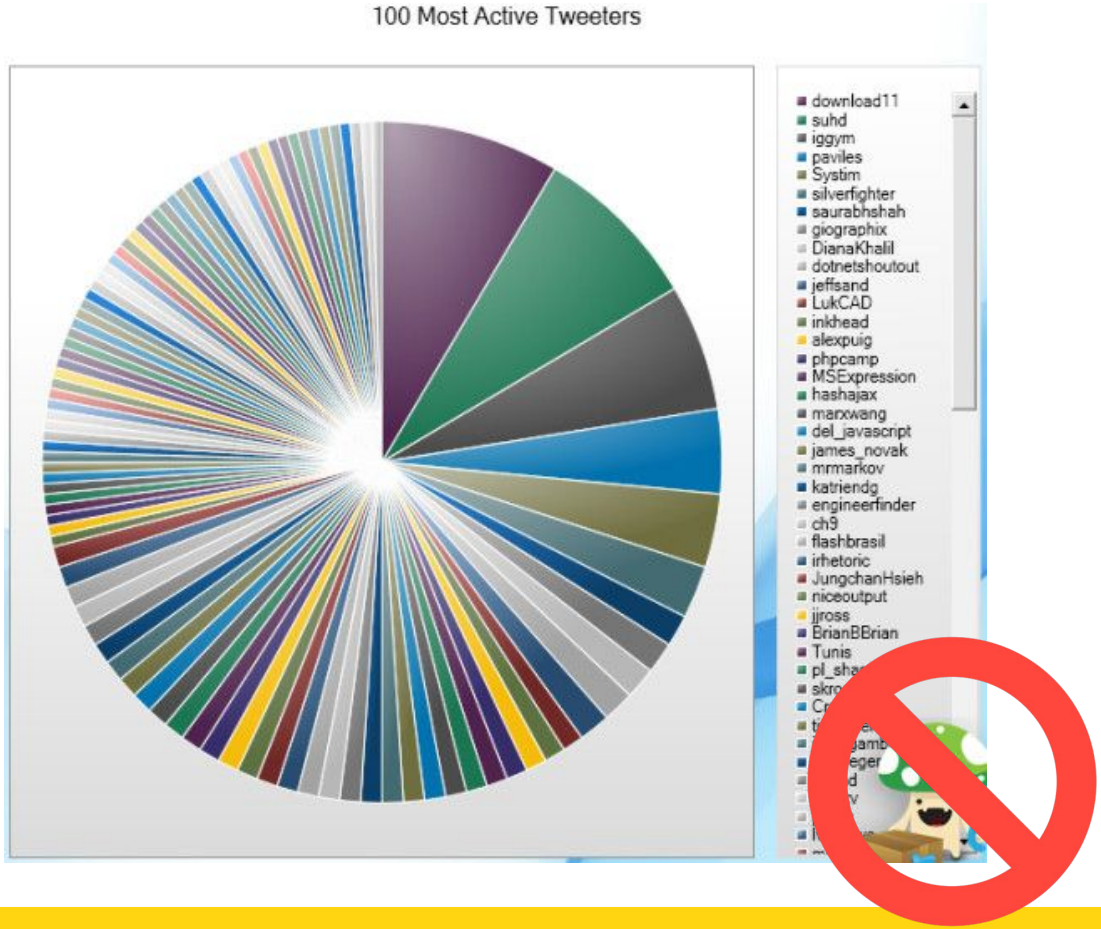


Pie Chart

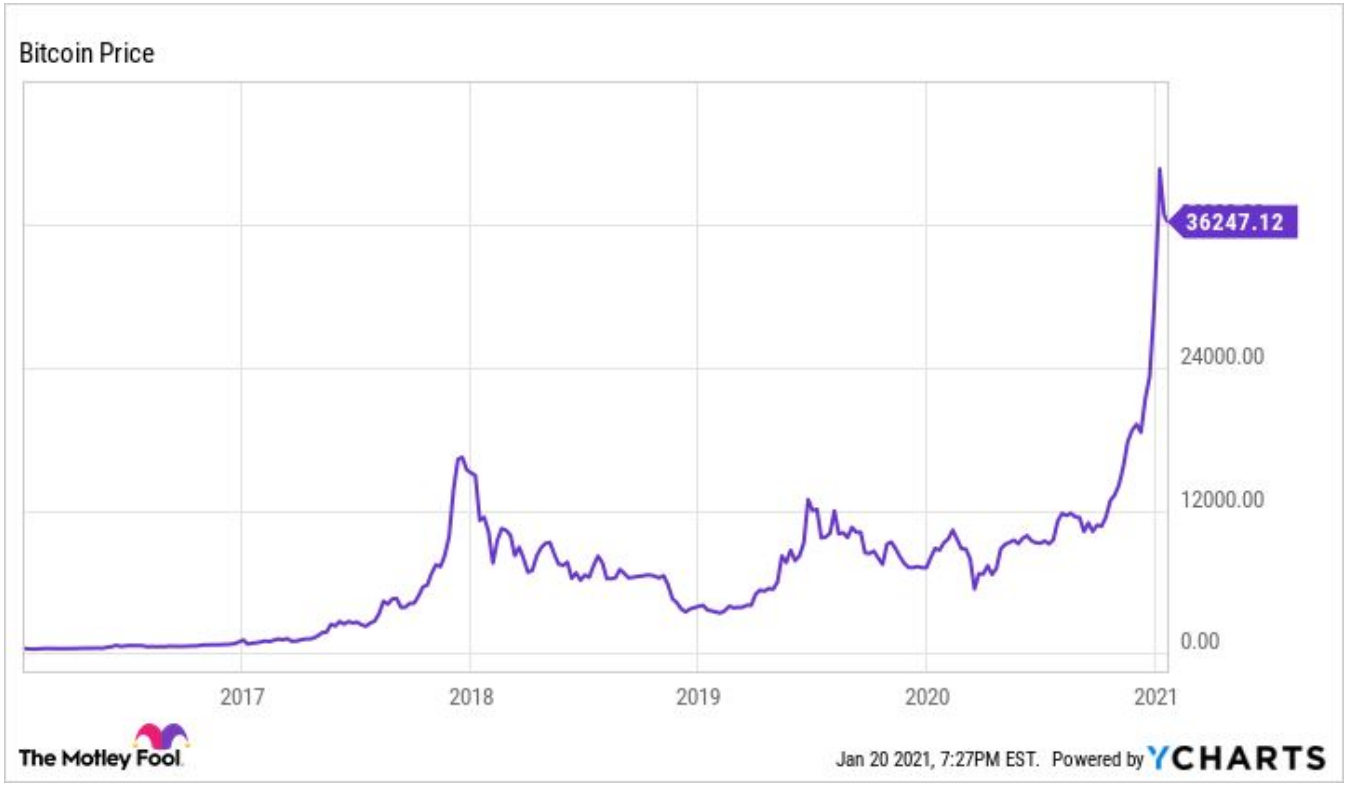

The most successful types of pie charts
65%
of Yorkshire Terriers
No
Yes
Other pets
15%
50%
Dogs
Cats
35%

Un pie chart (diagramma a torta) viene utilizzato per rappresentare una variabile in relazione al totale.

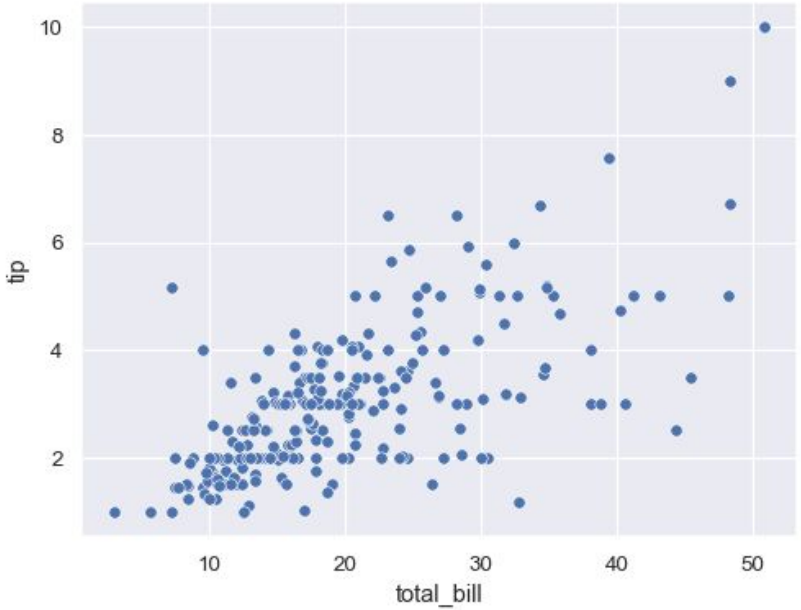



100 Most Active Tweeters
download11
suhd
iggym
paviles
Systim
silverfighter
saurabhshah
giographix
DianaKhalil
dotnetshoutout
jeffsand
LukCAD
inkhead
alexpuig
phpcamp
MSExpression
hashajax
marxwang
del_javascript
james_novak
mrmarkov
katriendg
engineerfinder
ch9
flashbrasil
irhetoric
JungchanHsieh
niceoutput
jjross
BrianBBrian
Tunis

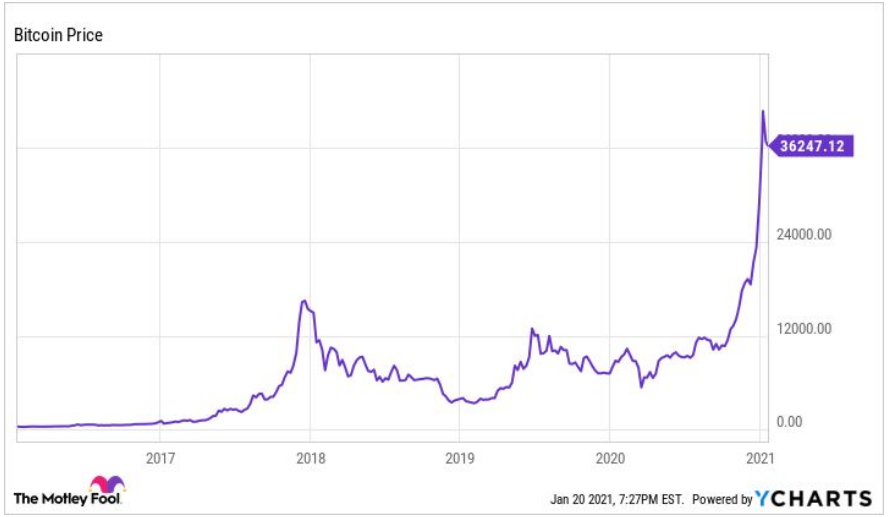
Line plot


Bitcoin Price
36247.12
24000.00
12000.00
0.00
2017
2018
2019
2020
2021
The Motley Fool
Jan 20 2021, 7:27PM EST. Powered by YCHARTS

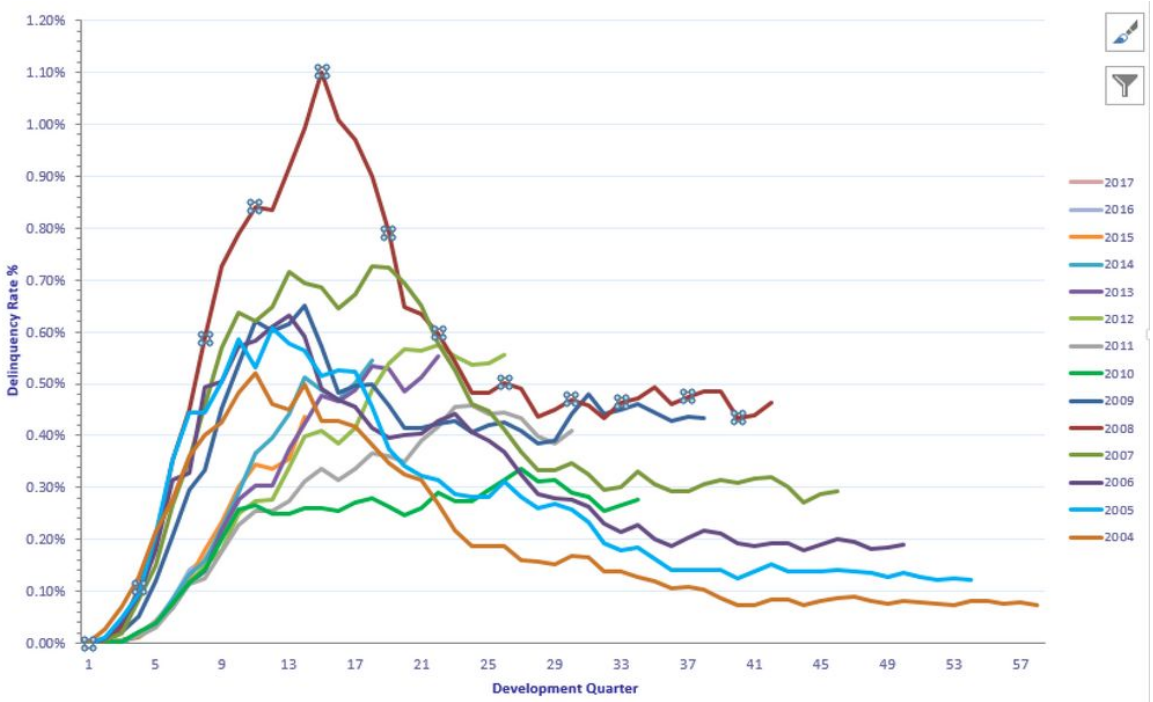
Line plot

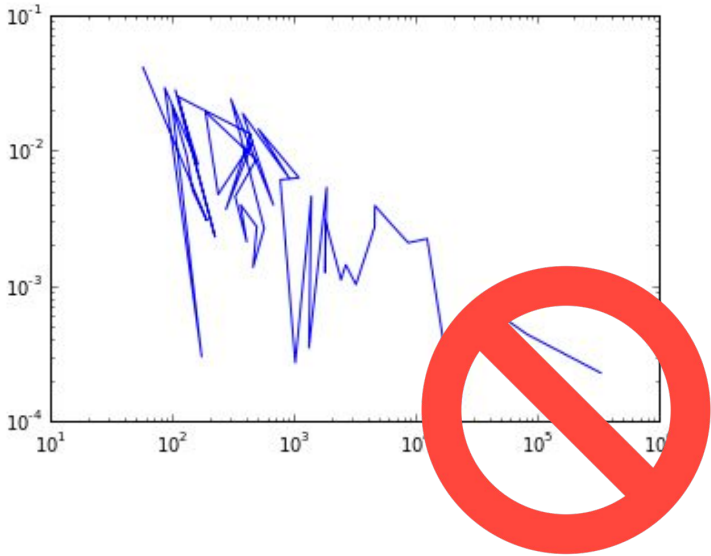

10
8
6
4
2
tip
10
20
30
40
50
total_bill

VS


Bitcoin Price
36247.12
24000.00
12000.00
0.00
2017
2018
2019
2020
2021
The Motley Fool
Jan 20 2021, 7:27PM EST. Powered by YCHARTS

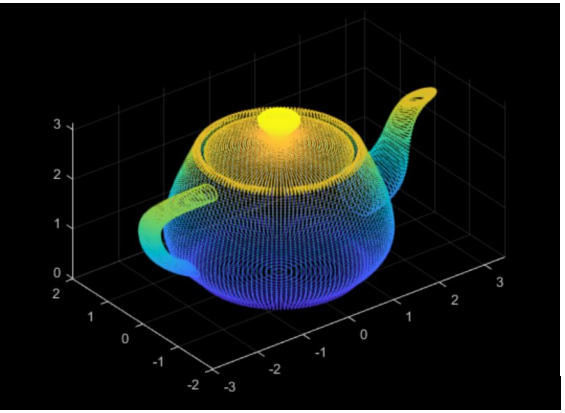
Line plot

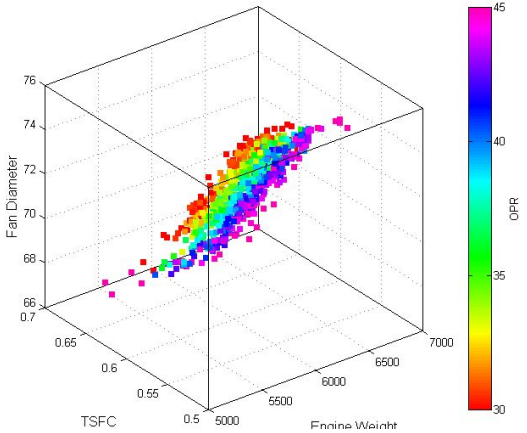

1.20%
1.10%
1.00%
0.90%
0.80%
0.70%
0.60%
0.50%
0.40%
0.30%
0.20%
0.10%
0.00%
Delinquency Rate %
1
5
9
13
17
21
25
29
33
37
41
45
49
53
57
Development Quarter
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004

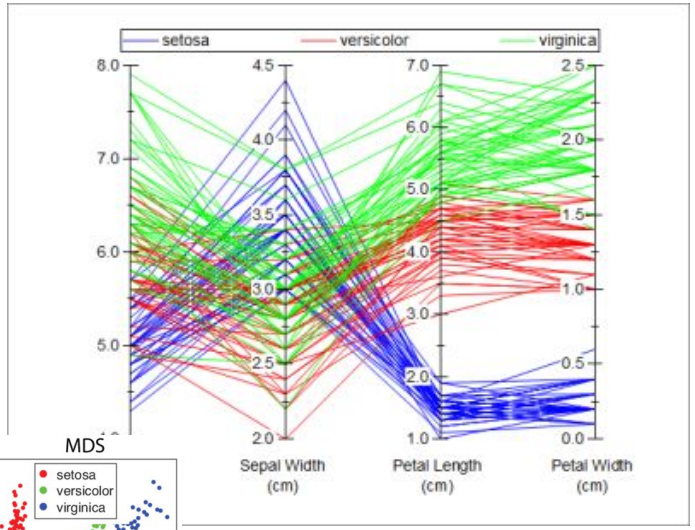

10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{1}
10^{2}
10^{3}
10^{5}

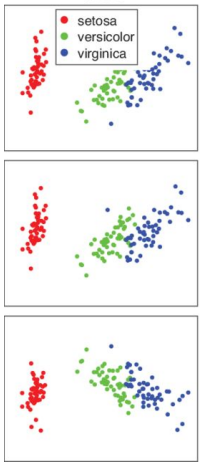
Più di 2 variabili?

Visualizzare più di due variabili per volta può essere molto difficile, ed è oggetto di studio!







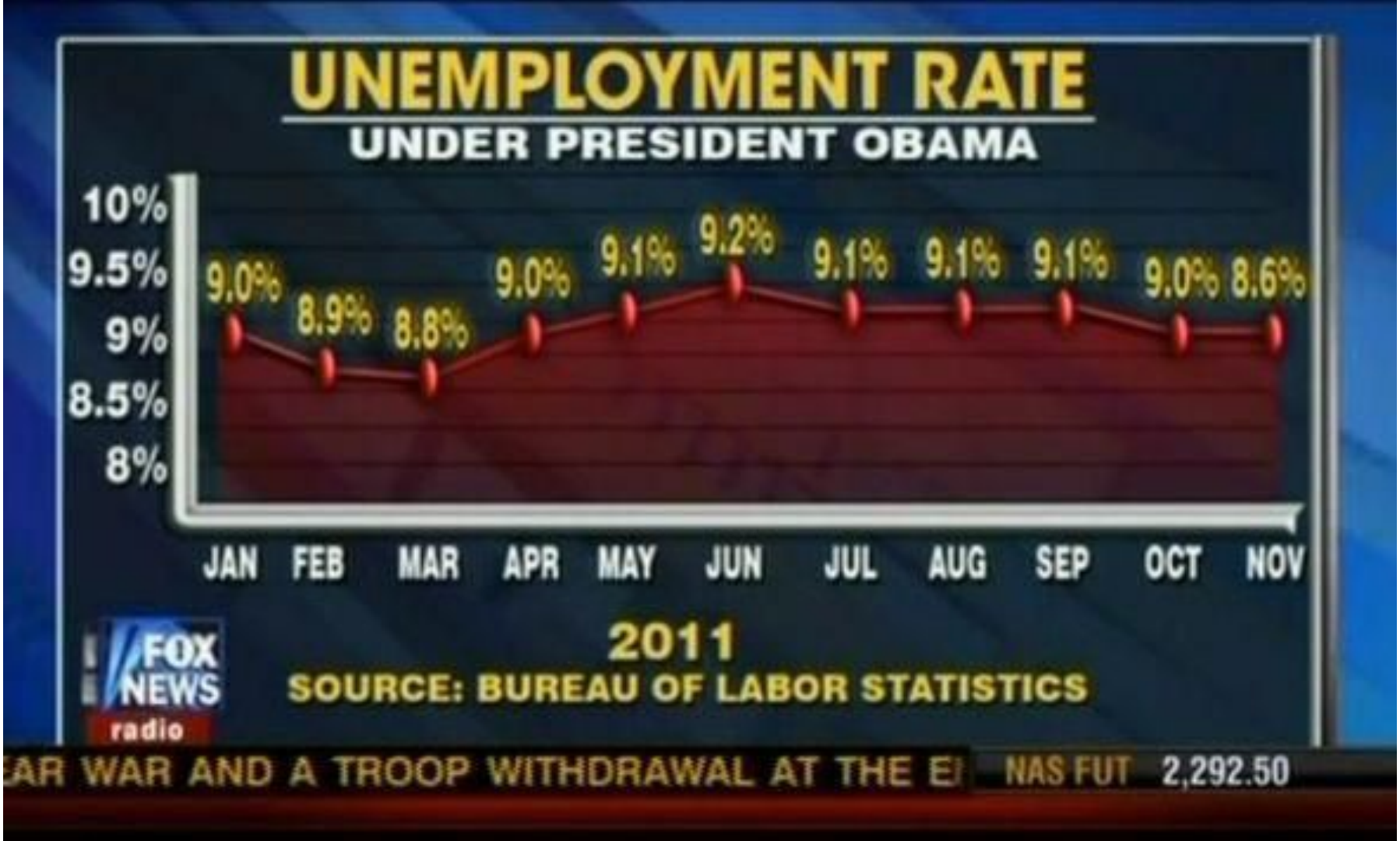


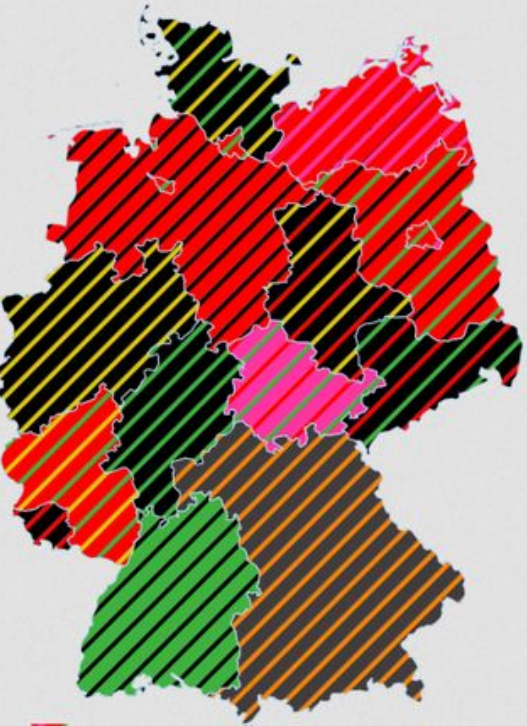
I colori aiutano a percepire altre dimensioni

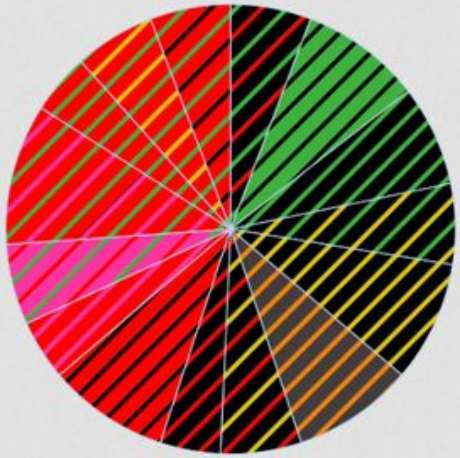
Tecniche più sofisticate cercano di mantenere i dati nel 2D

COME **NON** VISUALIZZARE I DATI

https://www.reddit.com/r/dataisugly/


UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.0%
9.1%
9.2%
9.1%
9.1%
9.1%
9.0%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
2011
SOURCE: BUREAU OF LABOR STATISTICS
FOX NEWS radio
AR WAR AND A TROOP WITHDRAWAL AT THE E
NAS FUT 2,292.50



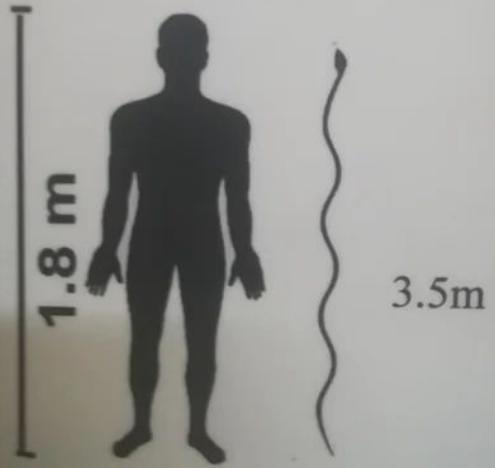


SPD + GRÜNE + LINKE	7
GRÜNE + CDU	6
CDU + FDP	6
CSU + FW	6
SPD + CDU	6
CDU + GRÜNE	5
CDU + SPD + GRÜNE	4
CDU + GRÜNE + FDP	4
CDU + SPD + FDP	4
LINKE + SPD + GRÜNE	4
SPD + FDP + GRÜNE	4
SPD + CDU + GRÜNE	4
CDU + SPD	3
SPD + LINKE	3
SPD + GRÜNE	3

Yellow Anaconda

Eunectes notaeus

SIZE

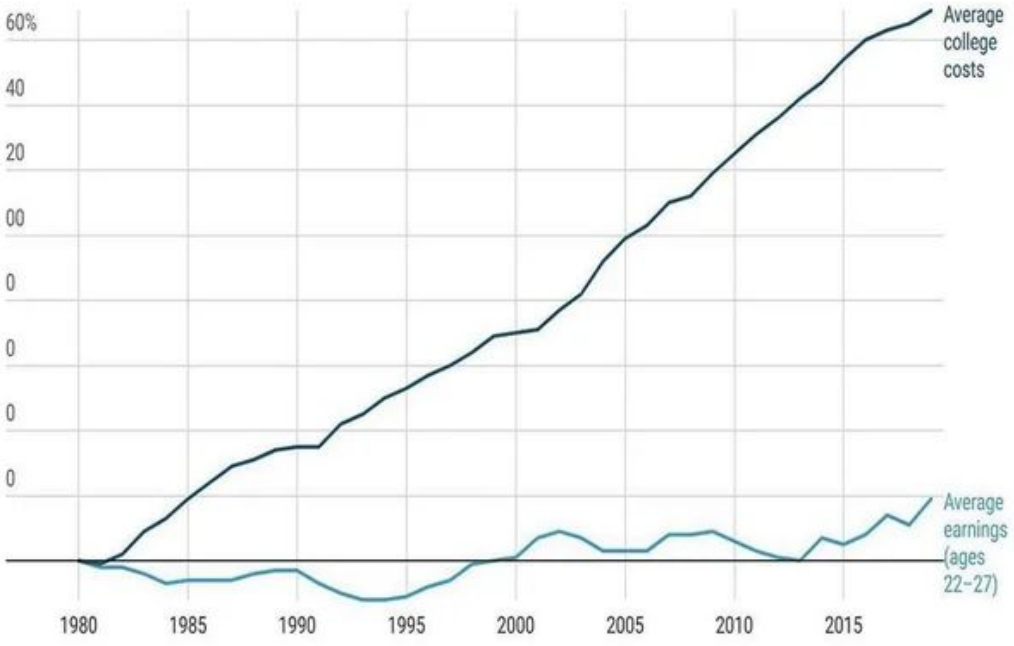


IMDb

TOP 10 MOVIES OF 2021

1	 Dune	★8.2
2	 The Suicide Squad	★7.3
3	 Eternals	★6.8
4	 Mortal Kombat	★6.1
5	 Zack Snyder's Justice League	★8.1
6	 Godzilla vs. Kong	★6.4
7	 Black Widow	★6.7
8	 Army of the Dead	★5.8
9	 Cruella	★7.4
10	 Shang-Chi and the Legend of the Ten Rings	★7.6

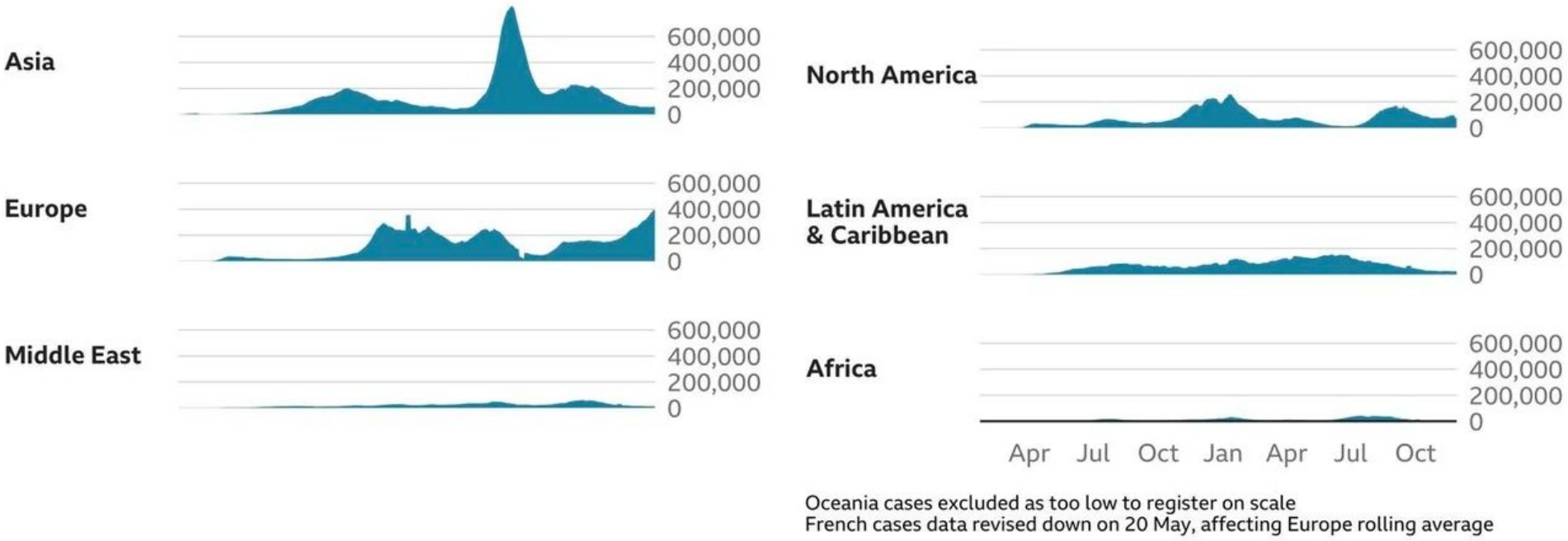
Percent change in college costs and earnings for young workers



Source: Georgetown University

Covid-19 cases compared by region

Number of cases per day, seven-day rolling average



Oceania cases excluded as too low to register on scale
French cases data revised down on 20 May, affecting Europe rolling average