
Dispensa del corso di Data Visualization

⁰Pietro Carmine Valenti, CDLM Data Science, Università di Milano-Bicocca

Introduzione

Il ciclo dei dati si compone di diverse fasi, che partono dall'elaborazione e arrivano all'utilizzo finale:



Dal momento che la comunicazione dei dati e dei risultati è un passaggio fondamentale di un'analisi, tanto quanto la parte di analisi quantitativa e computazionale, la *data visualization* gioca un ruolo fondamentale; in particolare essa consente l'analisi, l'utilizzo ma ancor di più la condivisione del dato, in un linguaggio visuale che consente di comunicare anche a chi non è esperto di dati. Lo **scopo principale** della dataviz è proprio questo: fornire uno strumento agli esperti analisti per comunicare i propri risultati in un linguaggio che possa convincere le persone. In questo senso la dataviz rende immediate alcune informazioni che risulterebbero di difficile comprensione se espresse con altri metodi (parole, report, formule, modelli, ...), i quali richiederebbero uno sforzo eccessivo dal fruttore.

Exit this room. **Turn right** and **walk 10 feet** to the end of the hallway, where you'll be facing a large conference room. **Turn left** and **walk another 12 feet** until you come to the end of that hallway. To your left is a fire alarm, near the elevator. To your right at the end of the hall is a stairwell. Do not go to the elevator. **Turn right** and **walk another 12 feet** to the end of the hall, **turn left and enter the stairwell**. **Go down two flights of stairs** and **exit the building** at the door at the bottom of the stairs.



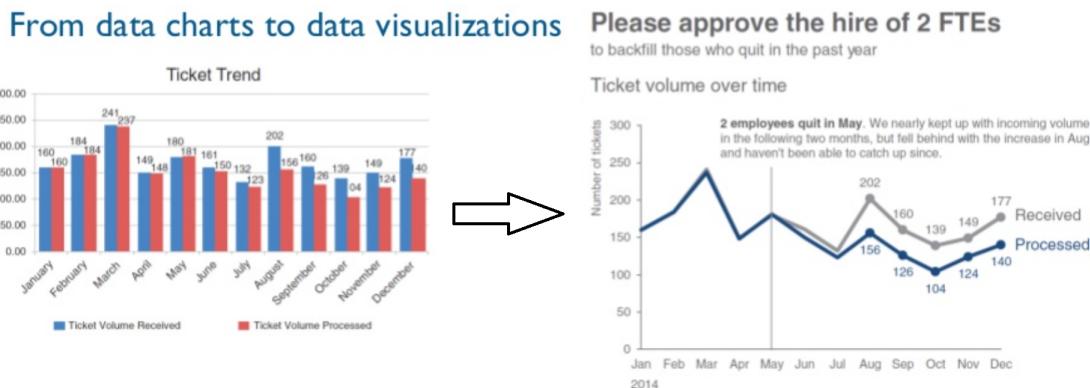
La dataviz condivide molti punti in comune con la *data science*, in particolare le principali fasi di raccolta e analisi dei dati:

- Formulano delle domande e provano a fornire una risposta.
- I dati vengono raccolti utilizzando tecniche adeguate.
- Vi è un'analisi dei dati raccolti.

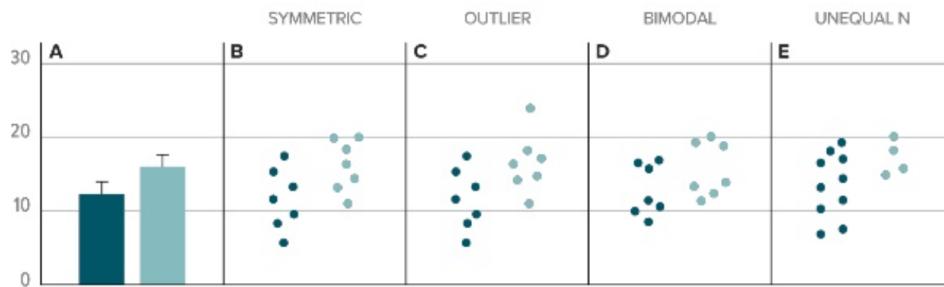
Esistono tuttavia una serie di metodi e intenti che rendono le due discipline fortemente distinte:

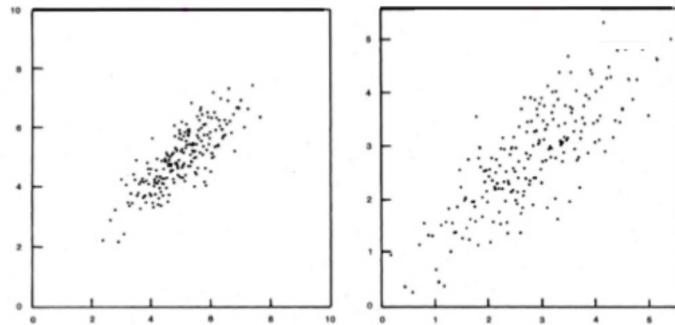
- La DS cerca di rendere esplicita un'ipotesi, la DV un punto di vista.
- DS sviluppa una procedura per dimostrare tale ipotesi, DV per dimostrare (o meglio mostrare) il suo punto di vista.
- DS testa l'adattamento delle ipotesi ai dati, DV visualizza i risultati.
- DS pubblica sotto forma di report i risultati, DV testa la qualità della percezione della dimostrazione (è arrivato il messaggio?).

La dataviz non si compone unicamente di grafici risultanti dall'analisi dei dati, bensì li inserisce in un contesto visuale (*dashboards*, commenti, immagini,) che consentono di creare un ecosistema visuale che renda il più esplicito possibile i risultati o il messaggio che si vogliono trasmettere.

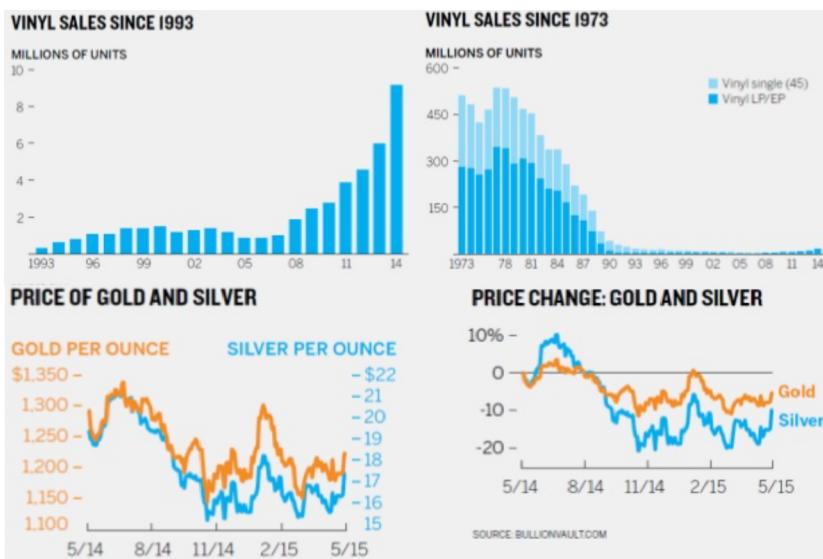


La *data visualization* ha la caratteristica di poter comunicare punti di vista differenti utilizzando gli stessi dati, di conseguenza bisogna sempre fare molta attenzione a comunicare il giusto messaggio, senza tuttavia distorcere i risultati. Bisogna inoltre utilizzare la giusta combinazione di indicatori, poiché spesso un semplice misura non comunica tutta la realtà dei dati (come nel grafico della media sottostante, o il successivo con R^2).

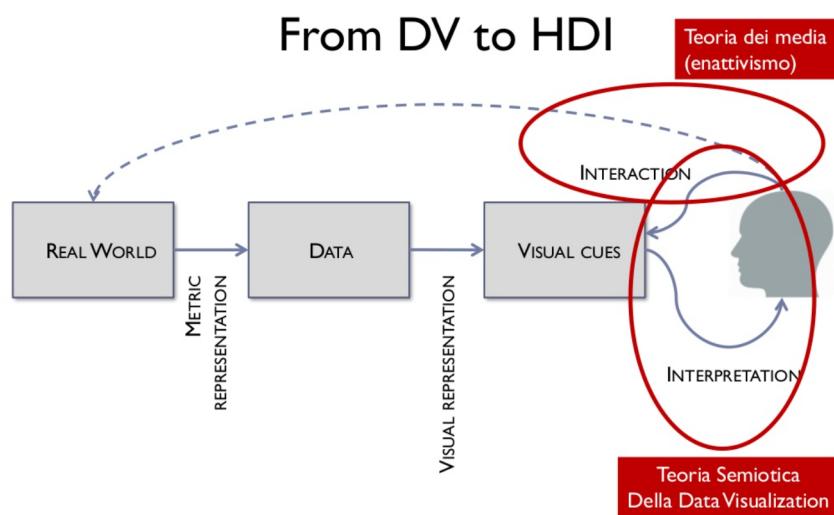




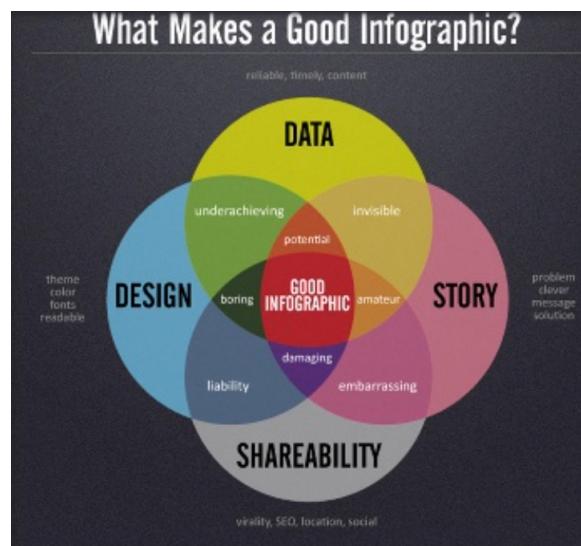
What's dataset is more correlated? Equally correlated ($r=0.8$) but point size different.



Si possono commettere numerosi errori quando si produce una dataviz, a volte per incompetenza, altre per mentire utilizzando i dati; infatti è possibile, utilizzando dati corretti, distorcere le visualizzazioni fino a far passare un determinato messaggio, il quale tuttavia cozza con la realtà dei fatti. Da questo punto di vista la dataviz è uno strumento con il quale l'uomo interpreta il mondo.



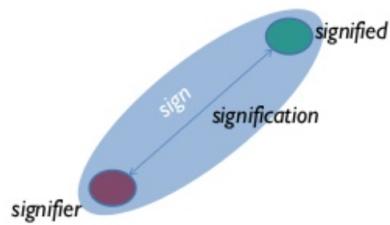
Un infografica per essere di qualità deve avere diverse caratteristiche: deve partire dai dati reali al fine di essere veritiera, deve raccontare una storia (un processo) e aver cura del *design* e della sua comunicabilità/comprensibilità.



Semiotica e *data visualization*

Una disciplina fortemente legata alla DV è la semiotica; la sua definizione più semplice ed immediata è lo studio dei segni. In epoca moderna si tende tuttavia a dare una definizione più ampia inserendo i segni non più in una realtà isolata, ma come un sistema semiotico dei segni: la semiotica diventa quindi lo studio di come viene costruito il significato e come viene comunicato. L'idea di fondo è che i segni di per sé non abbiano un significato intrinseco (segnali stradali, linguaggio corporeo, lettere, ...), e lo assumono solamente in relazione a chi li percepisce e alla sua esperienza; volendo fornire una definizione dei segno si potrebbe identificare come un elemento percepibile il quale acquista un significato se inserito in un contesto sociale.

A correlation of differences, the whole that results from the association of the signifier with the signified, which are inseparable as the two sides of a piece of paper (Saussure)

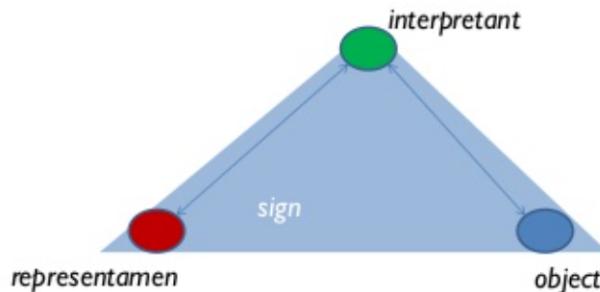


Un segno è composto da 2 elementi inscindibili, ed emerge solo dall'interazione di essi (**Sassurean Legacy**):

- **Significante** (*signifier*): è il veicolo del segno, materiale e fisico, il quale può essere percepito dai sensi.
- **Significato** (*signified*): rappresenta il significato inserito nel contesto sociale.

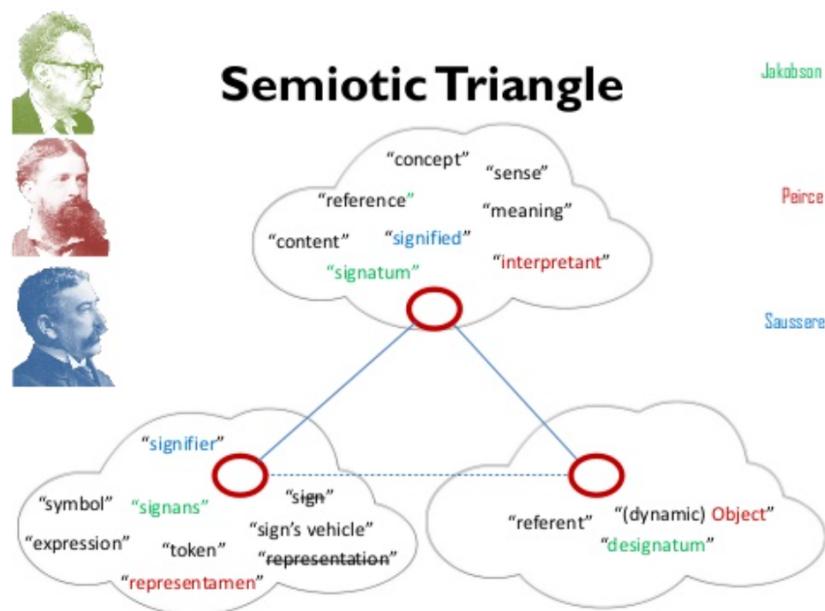
Il legame tra il significante ed il significato è quindi convenzionale, dipendente dalle convenzioni socio-culturali.

Un secondo schema interpretativo viene fornito da Hére Peirce, il quale introduce un terzo polo nella relazione significato-significante, ovvero l'**interprete**:



dove con **representamen** si intende il veicolo del segno, con **interpretant** si intende il senso di fatto del segno e infine con **object** si intende la realtà dietro il segno potrebbe essere ricondotto al significante, mentre gli altri due al significato. L'interazione tra i tre elementi viene definita **Semiosi**, ovvero un processo in cui viene applicato il ragionamento abduttivo al fine di generare ipotesi per spiegare elementi della realtà, testandoli attraverso l'evidenza. Quando l'ipotesi viene accettata il ragionamento viene accettato come principio generale per tutti i segni simili.

I principali concetti legati al **Triangolo della Semiotica** vengono riassunti nella seguente immagine:

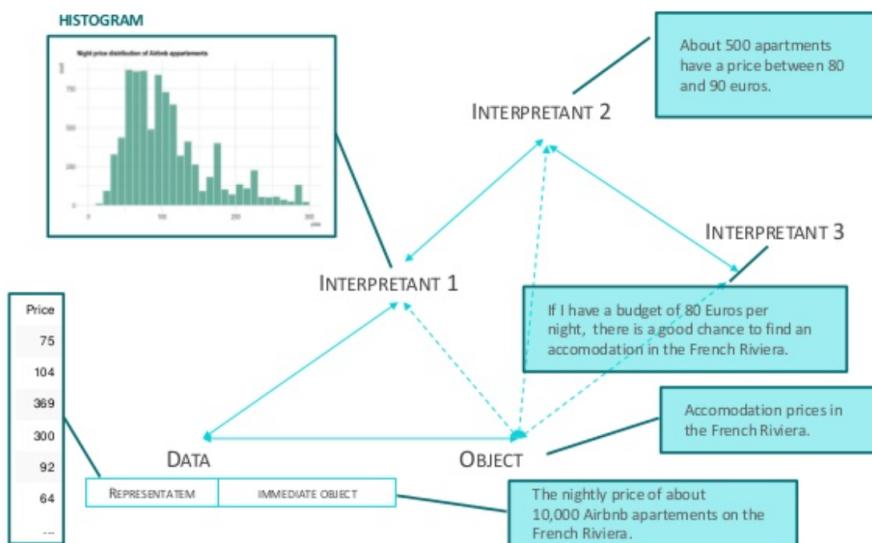


Possiamo distinguere alcuni concetti legati ai segni:

- **Simboli:** una modalità nel quale il significante non somiglia al significato, ovvero in cui è puramente convenzionale/arbitrario; in questo caso la relazione necessita di essere accettata e imparata.
- **Icône:** una modalità in cui il significante imita/ricorda il significato, ovvero in cui possiede alcune delle sue qualità.
- **Indice:** una modalità in cui il significante è direttamente connesso al significato (fisicamente o causalmente) attraverso un legame osservabile.

Un esempio di oggetto che contiene tutti questi elementi è la mappa, la quale è indicale nel puntare i luoghi, iconica nel rappresentare le relazioni direzionali e di distanza tra i punti e infine simbolica nell'utilizzare simboli convenzionali per identificare i luoghi.

Definito il concetto di semiotica è ora possibile legarlo alla dataviz: la *Data Visualization* comprende un'insieme di tecniche e metodi in cui la rappresentazione visuale dei dati ha lo scopo di concepire e rappresentare la realtà, al fine di trasmettere informazioni agli utilizzatori, supportando la loro interpretazione. In questa visione semiotica della DV possiamo identificare sia i dati, sia le visualizzazioni come costrutti, descrivibili più come risultato di un processo semiotico rispetto ad una loro natura intrinseca. Da questo punto di vista la **dataviz** non è altro che un **interpretante dei dati**, il quale aggiunge valore esplicando maggiormente i dati grezzi.



Tra i padri della Semiotica troviamo Umberto Eco, il quale afferma che il processo di semiosi è illimitato, in quanto ogni espressione necessita di essere tradotta in altri segni, in maniera che l'interpretante renda ragione dell'interpretato e al tempo stesso ne faccia conoscere qualcosa in più. Afferma inoltre che i testi non sono altro che enunciati legati da vincoli di coerenza, emessi contemporaneamente sulla base di più sistemi semiotici; sono più ampi di un singolo segno, possiedono dei confini che lo

separano da ciò che non ne fa parte, rispettano un certo grado di consistenza interna e infine possiedono una struttura narrativa sottostante. Tali aspetti riferiti ai testi sono tuttavia estendibili a tutti i sistemi semiotici che prevedano una relazione tra segni, di conseguenza anche la DV li rispetta.

Da Eco in poi si inizia a distinguere tra due concetti:

- **Dizionario:** si suddivide in segni con relazioni di equivalenza; il significato di un segno può essere descritto da un insieme finito di unità (minime) con significato. Presenta dei difetti, come il suo logocentrismo (centralità del logos), statico, incapace di catturare la creatività, insensibile al contesto e disconnesso dal resto della realtà.
- **Enciclopedia:** insieme registrato di tutte le interpretazioni, come libreria di librerie (archivio dell'info. non verbale registrata), non descrivibile nella sua totalità. Le sue principali caratteristiche sono la raffigurazione triadica del segno, il significato del segno è un altro segno che lo interpreta per certi aspetti (e in relazione all'oggetto) e infine la sua non rappresentabilità (derivante dalla natura infinita). Ha numerose qualità, come ad esempio la natura multimediale, dinamica, creativa, sensibile al contesto e inserita all'interno della realtà/mondo.

Per la sua natura infinita e non rappresentabile, è bene pensare al singole porzioni di enciclopedia, le quali vengono inserite in un contesto socio-culturale. Le DV dipendono quindi dal percorso interpretativo registrato nell'enciclopedia, di conseguenza è importante considerare le conoscenze enciclopediche al fine di interpretarle correttamente. Nei testi, così come nella dataviz, è necessario non solo identificare il proprio **lettore modello**, ma arrivare a costruirlo; per fare ciò è necessario immaginarsi in quale contesto di utilizzo pratico il lettore modello interpreterà le visualizzazioni. Per riassumere:

1. Any dataviz is **doubly constructed**.
2. Any dataviz **says something more about the world**.
3. Any dataviz **can be seen as a text**.
4. Any dataviz **tells a story**.
5. Any dataviz has to consider **encyclopedic knowledges**.
6. Any dataviz **presupposes a Model Reader**.

Quando si analizzano immagini è necessario distinguere tra due piani:

- **Piano Figurativo:** vengono riconosciute le figure del mondo naturale, percepite dai sensi, descritte dal linguaggio e con un significato riconoscibile nella cultura di riferimento.
- **Piano Plastico:** vengono identificati linee, colori e spazio, indipendentemente dal fatto rappresentato; a tale livello è importante considerare la topologia (distribuzione di elementi nello spazio), il livello eidetico (forme) e il livello cromatico.

Dark Side of DV

1. Assi troncati

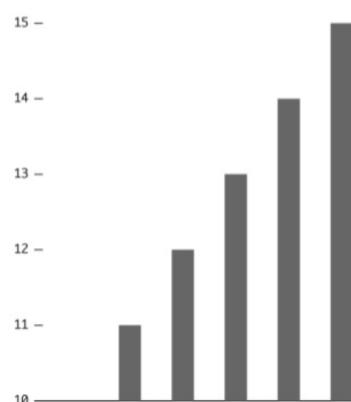
How to Spot Visualization Lies

Keep your eyes open.

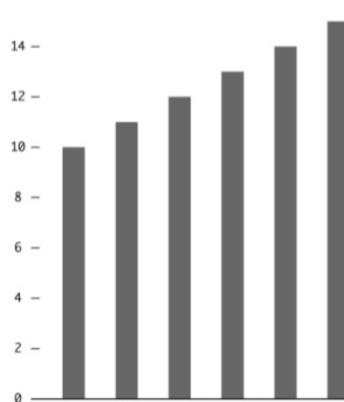
BY NATHAN YAU .

TRUNCATED AXIS

The value axis starts at ten. Liar, liar, pants on fire.

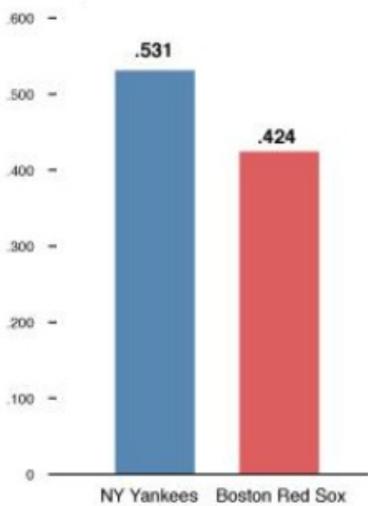
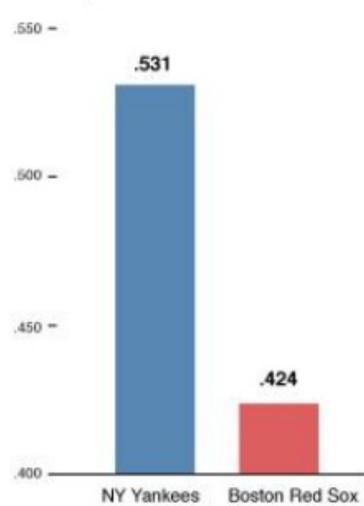


The value axis starts at zero. Good.

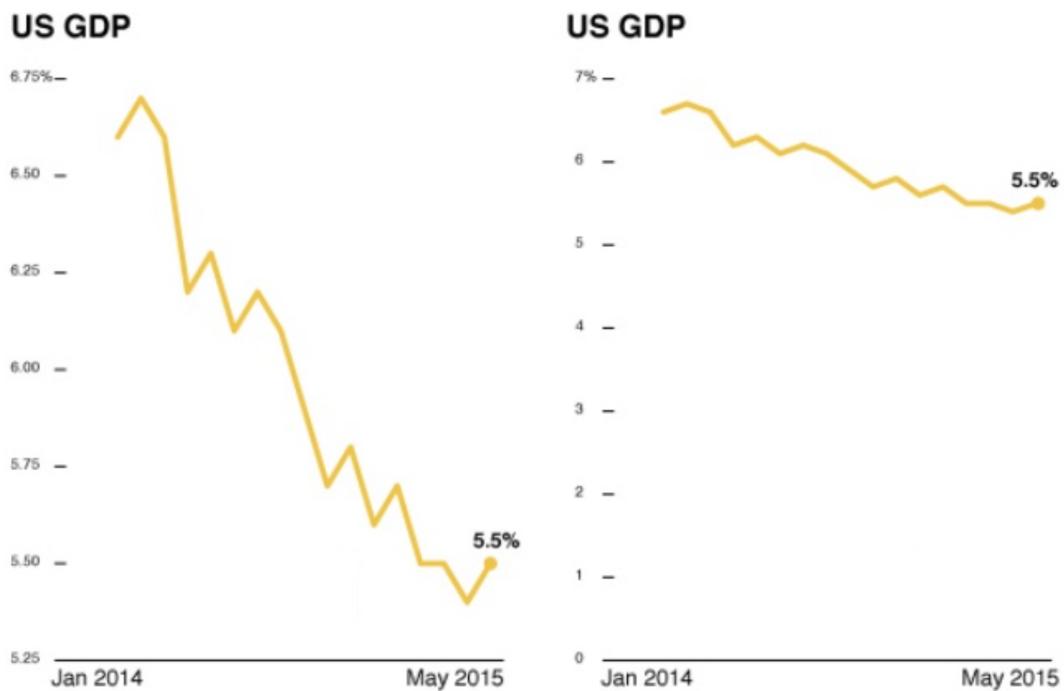
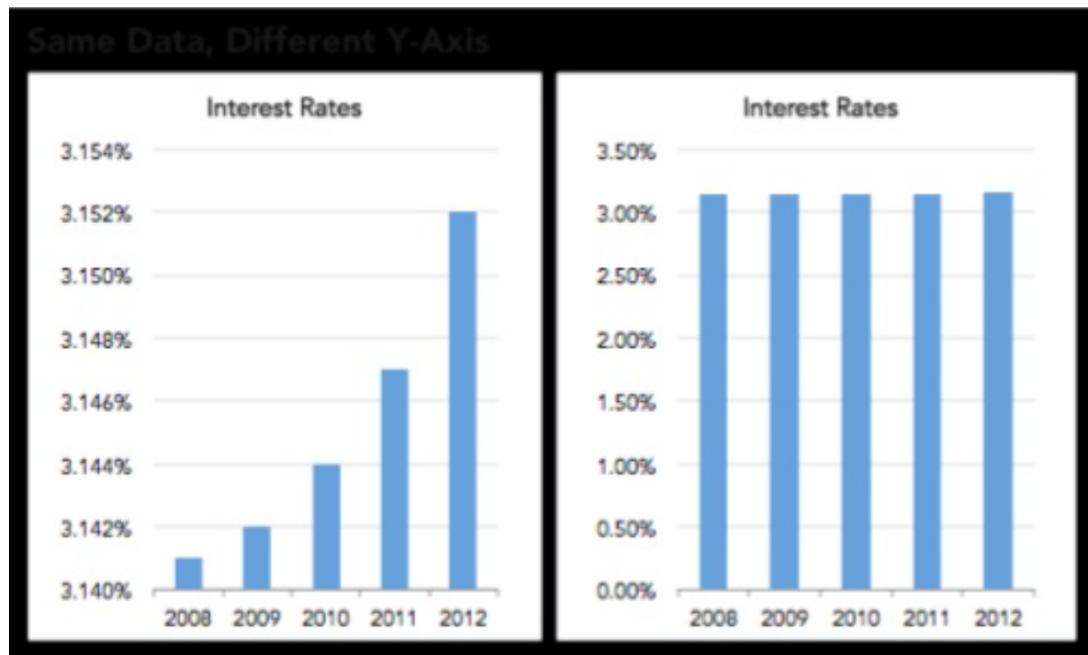


Percentage of victories

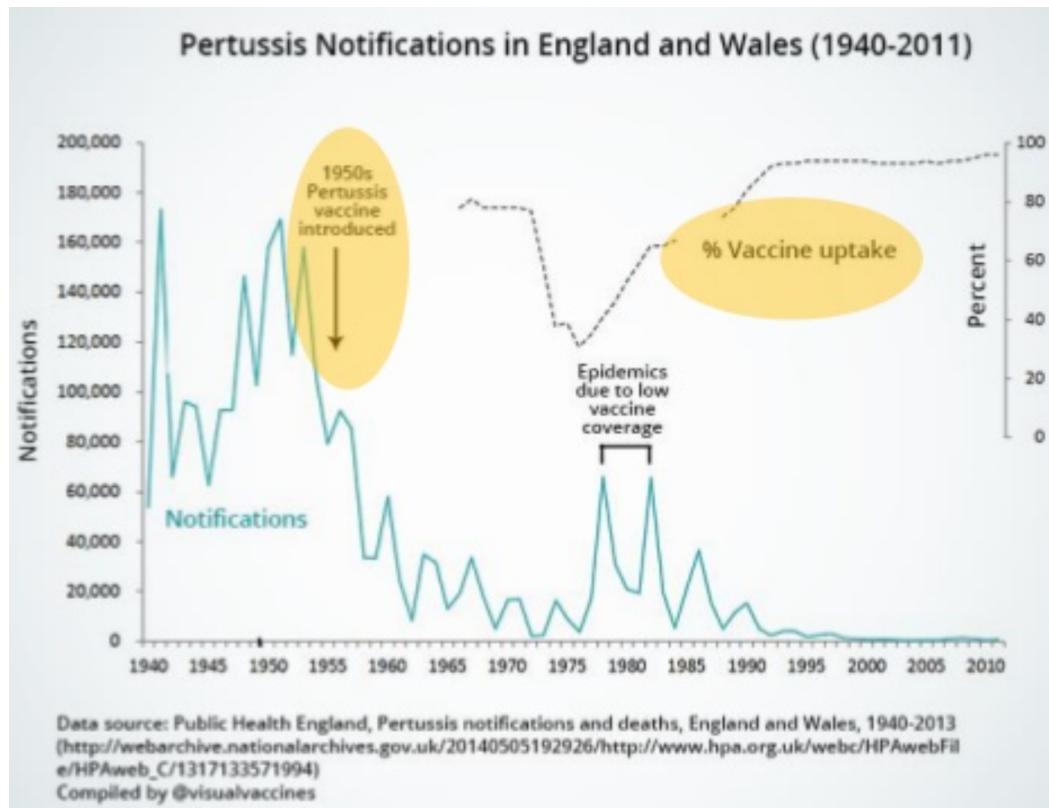
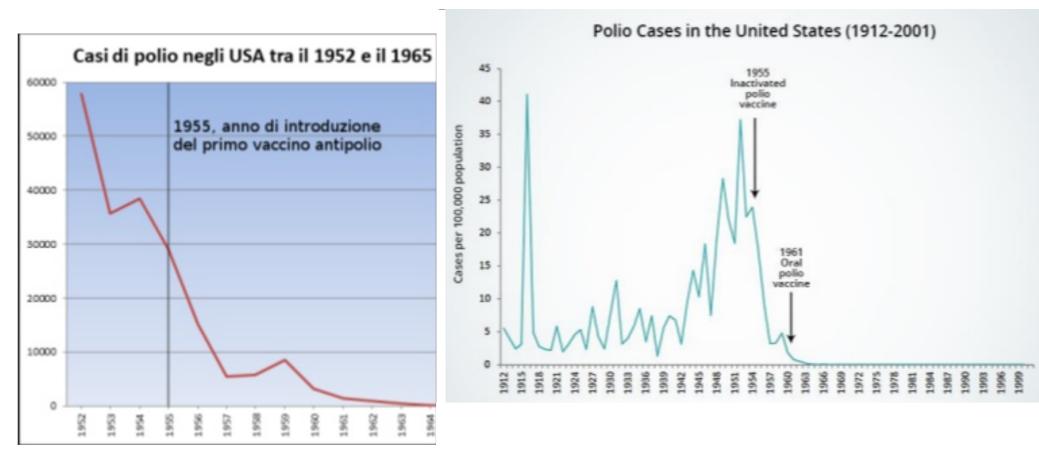
Percentage of victories



2. **Errori di Scala:** lie factor = $\frac{\text{size of effect shown in graphs}}{\text{size of effect in data}}$.



3. Inserire troppi elementi contestuali e/o suggerire collegamenti impropri (indicare eventi su serie storiche, doppi assi, ...):



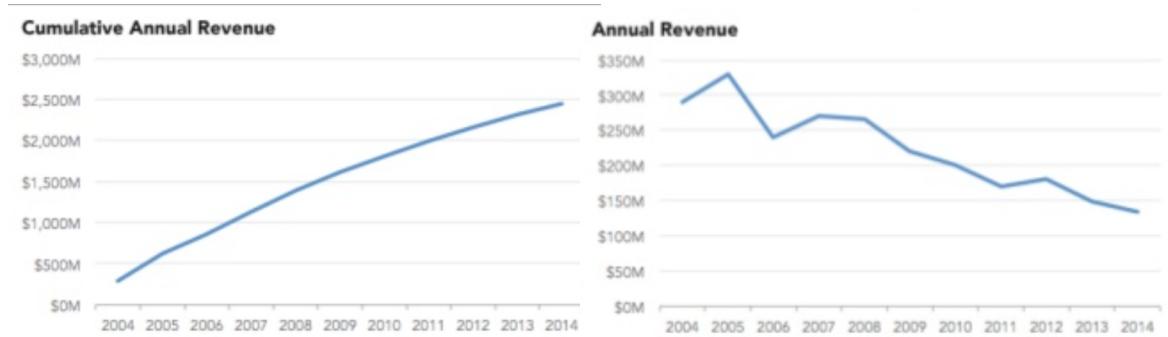
4. Mettere in relazione variabili non correlate:



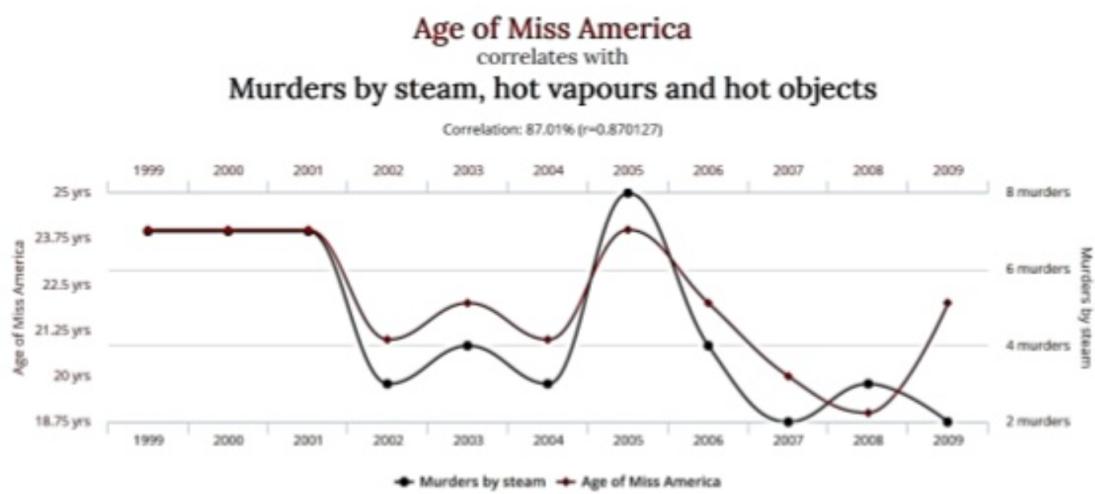
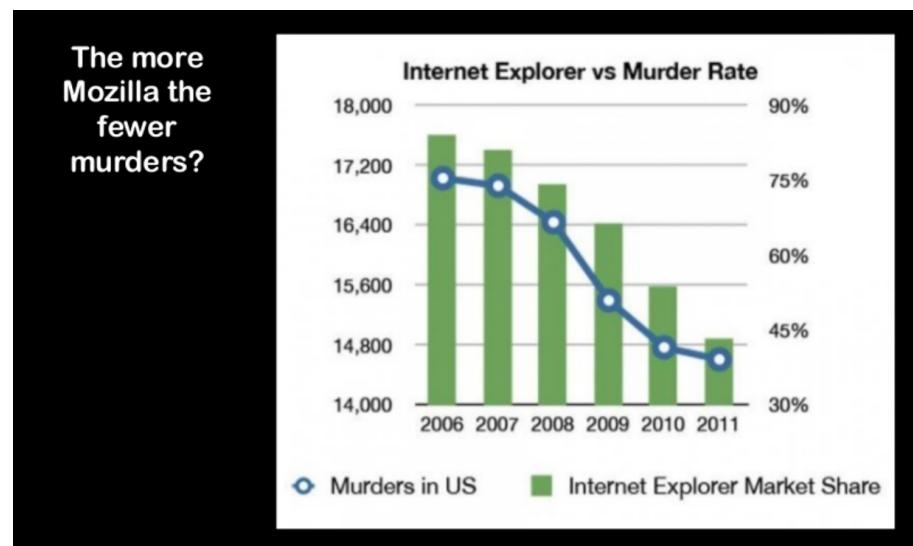
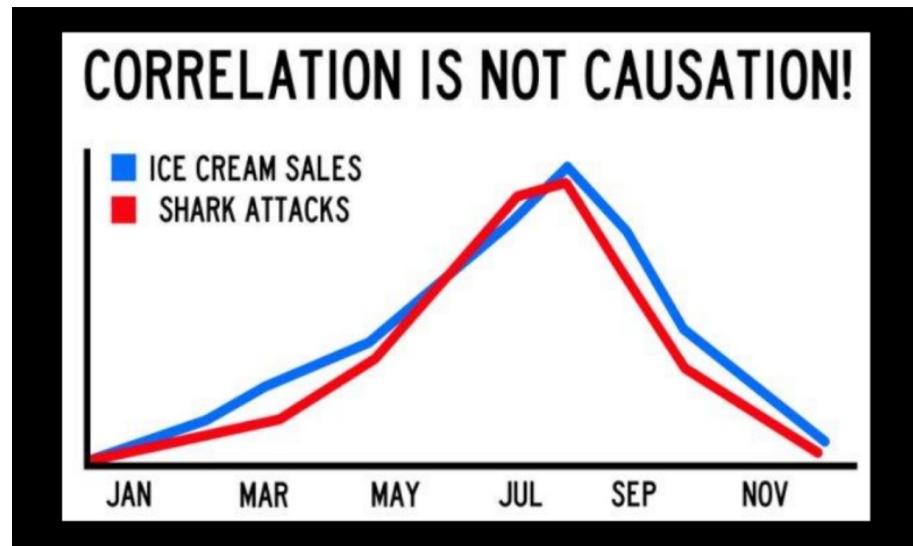
5. Errori di granularità



6. Uso improprio dei frequenze cumulative:



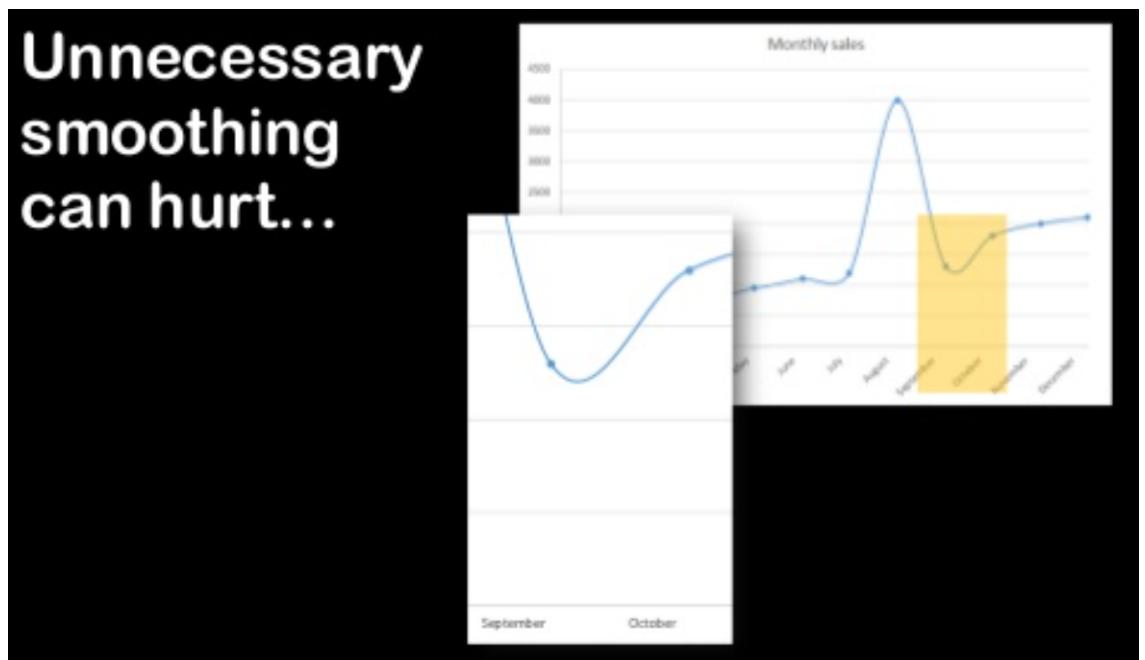
7. Mostrare correlazioni spurie:



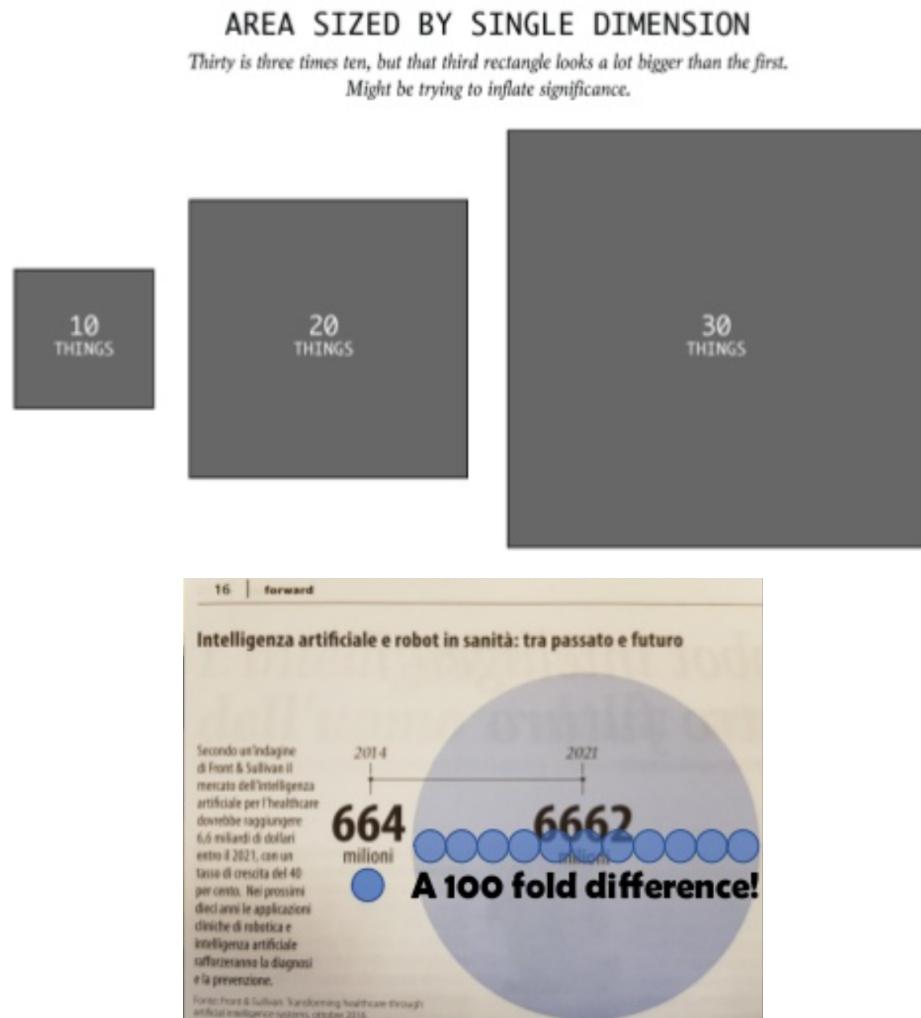
8. Non utilizzare **Pie Chart** quando le parti non sommano a 100%, quando vi sono troppe categorie o utilizzando 3 dimensioni.
9. Ignorare le principali convenzioni:



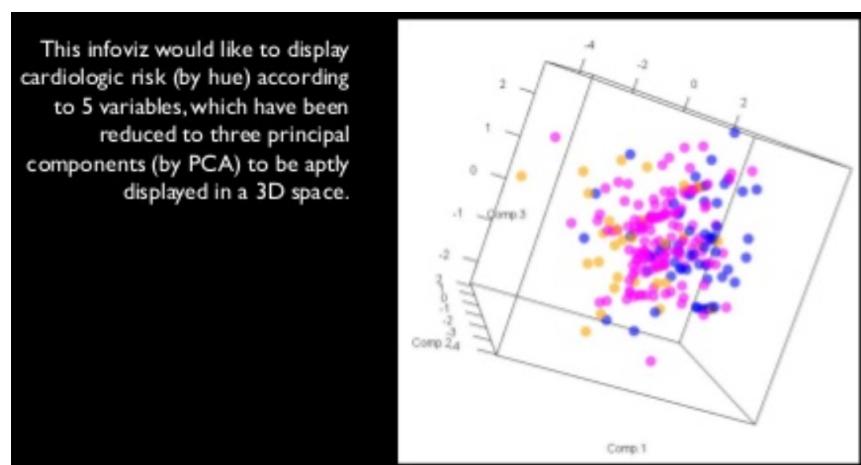
10. Utilizzare eccessivamente lo smoothing:



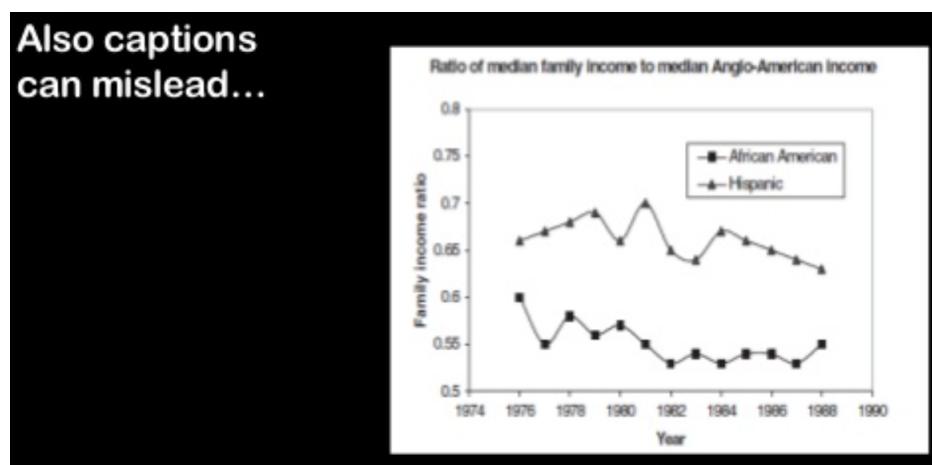
11. Non utilizzare correttamente i grafici in cui conta la scala delle aree (utilizzare raggio per i cerchi, lato per i quadrati porta a non poter confrontare le aree):



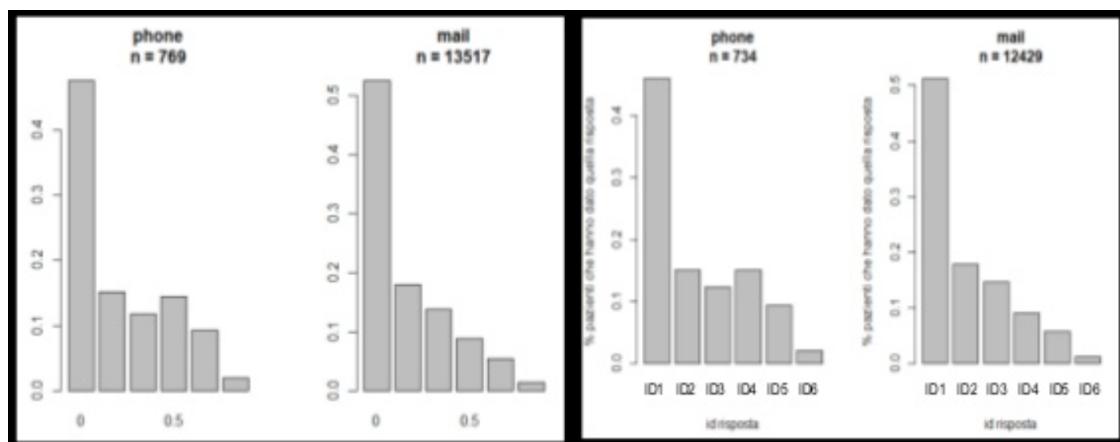
12. Non rendere le cose semplici più semplici e le cose complesse ancora più complicate.



13. Evitare quando possibile di impiegare colori che mettano in difficoltà i daltonici (scale di verde, arancione e rosso)
14. Mostrare solo le frequenze assolute, quando potrebbe essere utile mostrare quelle relative (densità abitativa).
15. Utilizzare la media come unico valore centrale; spesso conviene impiegare la mediana e/o accostare la deviazione standard alla media.
16. Fare attenzione alle Heatmap, in particolare alle interpolazioni tra i punti dove non vi sono dati, agli overlap ingannevoli e ad utilizzare scale di colori coerenti (che diano idea di crescendo).
17. Riportare sempre gli intervalli di confidenza se l'indicatore mostrato è campionario.
18. Le legende devono essere coerenti con il grafico (colore, posizione, ...)

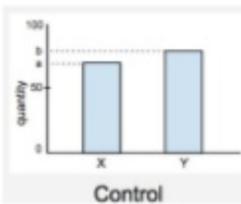


19. Riportare gli assi e le loro descrizioni:



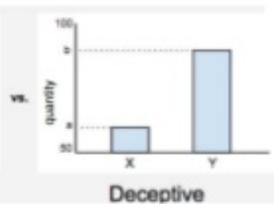
How to Spot Visualization Lies

Keep your eyes open.



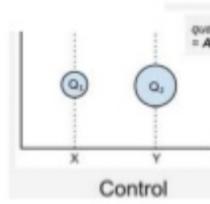
TRUNCATED AXIS DISTORTION

MESSAGE EXAGGERATION/UNDERSTATEMENT

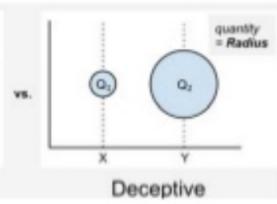


vs.

Deceptive



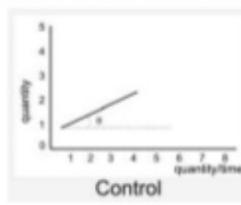
Control



Deceptive

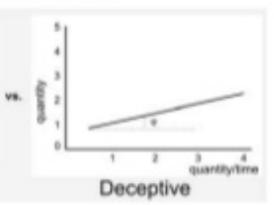
AREA AS QUANTITY DISTORTION

MESSAGE EXAGGERATION/UNDERSTATEMENT



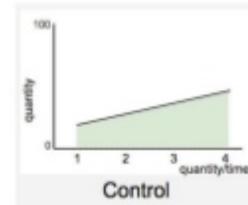
ASPECT RATIO DISTORTION

MESSAGE EXAGGERATION/UNDERSTATEMENT

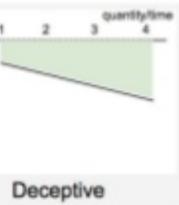


vs.

Deceptive



vs.



Deceptive

INVERTED AXIS DISTORTION

MESSAGE REVERSAL